Digital Humanities:
multimedial &
multimodal

DHd 2019

6. Jahrestagung

Frankfurt & Mainz

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

DHd 2019

Digital Humanities: multimedial & multimodal

Konferenzabstracts

Universitäten zu Mainz und Frankfurt
25. bis 29. März 2019

Partner

Unterstützer

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.

Herausgeber: Patrick Sahle
Redaktion und Korrektur der Auszeichnungen:
Rüdiger Gleim, Attila Kett, Nikolas Hilger
Konvertierung TEI nach PDF: Attila Kett
https://github.com/texttechnologylab/DHd2019BoA
Historie der Autorinnen und Autoren sowie Versionen der Konversionsskripte:
Claes Neuefeind (2018)
https://github.com/GVogeler/DHd2018
Aramís Concepción Durán (2016)
https://github.com/aramiscd/dhd2016-boa.git
Karin Dalziel (2013)
https://github.com/karindalziel/TEI-to-PDF
Konferenz-Logo: Vanessa Liebler

Online verfügbar: https://dhd2019.org

ISBN 978-3-00-062166-6 (gedruckte Ausgabe)

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

Vorwort

„Wer verstehen will, worum es in den Digital Humanities geht, was der Stand der Kunst ist, was die wichtigen Forschungsthemen und wer die Akteure sind, der oder die soll in die Books of Abstracts der DH-Konferenzen schauen." Das ist ein stehender Satz in allen meinen Einführungsvorträgen zum Thema Digital Humanities. Nirgendwo sonst bekommt man die ganze Breite der Forschungsthemen der DH in einer so vollständigen und zugleich wunderbar kondensierten Form dargeboten. Books of Abstracts als durch peer review-Verfahren gefilterte und qualitätsgesicherte Summen der aktuellen Forschungen definieren das Feld, sind ein äußerst nützliches Instrument der Fachkommunikation und wertvolle Dokumente zum Beleg der Entwicklung über die Zeit.

Dafür ist es wichtig, dass die Abstracts-Bände gut aufbereitet sind und dauerhaft erreichbar bleiben. Beides konvergiert mit der auch in diesem Jahr fortgesetzten Entwicklung zu immer größerer Professionalität bei der Erstellung des Abstractsbandes, wie auch der eingereichten Beiträge. Die Digital Humanities sind stark von Interdisziplinarität geprägt: Beitragende mit verschiedenen fachlichen Hintergründen haben ein sehr unterschiedliches Verständnis davon, was überhaupt ein „Abstract" ist und welchen inhaltlichen und formalen Kriterien es zu genügen hat. Als Gemeinschaft der digitalen Geisteswissenschaften im deutschsprachigen Raum sind wir immer noch auf dem Weg, ein gemeinsames Verständnis dieser Abstracts als ernstzunehmender wissenschaftlicher Publikationsform zu finden. Ich freue mich zu sehen, dass wir hierbei auch in diesem Jahr wieder einen Schritt weitergekommen sind. Um das Ziel ganz klar zu formulieren: die hier vorgelegten Abstracts sind wissenschaftliche Texte eigenen Rechts, die auch bibliografisch fassbar sein sollen, um die eigenen Forschungsgebiete und die gewonnenen Erkenntnisse sichtbar machen zu können. Diese Anspruchshöhe impliziert unter anderem, dass die Beiträge über *neue*, bisher *unveröffentlichte* Forschungsergebnisse berichten.

Für die entsprechende Qualität und Originalität tragen die AutorInnen die Verantwortung. Sie werden aber auch von der großen Gruppe der GutachterInnen sichergestellt, die durch Bewertung und Rückmeldung die Filterung durch das Programmkomitee und die Verbesserung der Beiträge durch die AutorInnen ermöglichen. Bedingt durch die neue DSGVO musste für diese Tagung der Pool der GutachterInnen neu aufgebaut werden. Er setzte sich schließlich aus 50 „AltgutachterInnen" und 100 neuen KollegInnen zusammen, denen insgesamt für Ihre gewissenhafte und gute Arbeit nicht genug gedankt werden kann. Von den 150 GutachterInnen sind insgesamt fast 600 Gutachten erstellt worden, so dass zu allen Einreichungen schließlich mindestens jeweils vier (bei Postern jeweils mindestens drei) Bewertungen vorlagen.

Auf der Grundlage der Punktevergabe der GutachterInnen und ihrer Annahmeempfehlungen hat das Programmkomitee Qualitätsschwellen definiert und schließlich die Beiträge zur Präsentation ausgewählt. Selbst bei maßvollen Ablehnungsquoten zwischen, je nach Format, 20% und 30% wird es enttäuschte Einreichende gegeben haben. Auch diesen sei aber ausdrücklich für ihre Einreichungen und ihren Beitrag gedankt.

Ebenso ist dem Programmkomitee zu danken. Heike Zinsmeister als stellvertretender Programmkomiteevorsitzenden, Lars Wieneke, Georg Vogeler, Christof Schöch, Stefan Schmunk, Andreas Münzmay, Mareike König, Andreas Henrich, Petra Gehring, Lisa Dieckmann und Alexander Czmiel als ProgrammkomiteemitgliederInnen und Kai Christian Bruhn als Vertreter des Organisationskomitees im Programmkomitee haben bei den vielen Prozessen der Entscheidungsfindung zu den verschiedensten Aufgaben und aufkommenden Sonderproblemen stets mit Rat und Tat und der Erfahrung aus den vorherigen Konferenzen zu konstruktiven Lösungen gefunden. Auf der organisatorischen und technischen Seite, der beständigen Bändigung des ConfTools ist Erik-Lân Do Dinh eine unverzichtbare Hilfe gewesen, der nicht nur immer wieder alle Einstellungsoptionen ausgereizt, sondern auch alle eingehenden Anfragen und Kommentare vorsortiert und ggf. beantwortet oder an die richtigen Stellen weitergeleitet hat.

Für die Realisierung des BoA danke ich schließlich Rüdiger Gleim (Redaktion), Attila Kett (Redaktion, Layout-Anpassung der Beiträge), Nikolas Hilger (Redaktion, Layout-Anpassung der Beiträge) und Vanessa Liebler (Cover-Gestaltung).

Viel Vergnügen mit den „Digital Humanities im deutschsprachigen Raum, Stand der Kunst 2019".

Köln, Februar 2019
Patrick Sahle
Vorsitzender des Programmkomitees

Inhaltsverzeichnis

Keynotes

Workshops

Panels

Vorträge

Posterpräsentationen

Anhang

Keynotes

Avancierte Methoden der computer-gestützten ästhetischen Filmanalyse

Flückiger, Barbara
Universität Zürich, Schweiz

Prof. Dr. Barbara Flückiger, Seminar für Filmwissenschaft, Universität Zürich

Audio-visuelle Bewegtbilder stellen hohe Anforderung an computer-gestützte Verfahren, einerseits wegen der Komplexität das Materials, andererseits wegen der hohen Dichte an audio-visuellen Informationen, die sie enthalten. Diese Anforderungen werden noch entschieden gesteigert, wenn es um ästhetische Dimensionen und die Identifikation von stilistischen Nuancen geht.

Genau eine solche Zielsetzung steht im Zentrum des Forschungsprojekts ERC Advanced Grant *FilmColors*, und zwar mit einem Fokus auf die Untersuchung der Farbfilmästhetik, die exemplarisch für Bewegtbildforschung insgesamt untersucht und entwickelt wird, um den Zusammenhang zwischen technischen Verfahren des Farbfilms und stilistischen Mustern systematisch anhand eines großen Korpus zu untersuchen.

Ziel war es einen Workflow aus Video-Annotation und manueller Analyse zu entwickeln, auf dessen Basis anschließend digitale Werkzeuge entstanden, welche diese Aufgaben möglichst weitgehend übernehmen und die Ergebnisse in überzeugender Art und Weise visualisieren. Zusätzlich entsteht eine Online-Plattform, auf die auch externe Nutzer ihre Analyse-Ergebnisse hochladen können, um sie mit bereits bestehenden Analysen zu vergleichen.

Grundlagenforschung, Primär- und Sekundärquellen zur Technik, Ästhetik, Analyse, Messung und Restaurierung / Digitalisierung von Filmfarben werden seit 2012 in der interaktiven Web-Plattform *Timeline of Historical Film Colors* (Flueckiger 2012 ff.) publiziert, einschließlich der fotografischen Dokumentation von historischen Farbfilmkopien aus Archiven in Europa, den USA und Japan, die inzwischen mehr als 20‘000 Fotos umfasst.

Entwicklung eines computer-gestützten Workflows

Traditionelle Analysen von Filmfarben und Filmstil generell gründen auf verbalen Beschreibungen, die insbesondere in der Domäne der Farben einen problematischen Hang zur hermeneutischen Interpretation haben. Mit dem Einbezug von Deep Learning Tools, datenbankgestützter Analyse und einem breiten Arsenal von Visualisierungen erarbeiten wir einen umfassenderen Ansatz, der den mannigfaltigen ästhetischen Phänomenen und Bedeutungsdimensionen von Filmfarben gerecht wird.

Grundlagen dazu wurden bereits 2011 in der Studie „Die Vermessung ästhetischer Erscheinungen“ in der *Zeitschrift für Medienwissenschaft* publiziert (Flückiger 2011), dies umfasste auch eine kritische Evaluation der experimentellen Ästhetik und ihrer wissenschaftsgeschichtlichen Verortung sowie der epistemologischen und wahrnehmungstheoretischen Grundlagen von Visualisierungen und standardisierten Konzepten zur Analyse von a priori wenig standardisierten ästhetischen Phänomenen. In den letzten Jahren sind mehrere Studien veröffentlicht worden, die sich in fundierter Weise mit Ansätzen der Digital Humanities für die Filmanalyse auseinander setzen (Gruber et al. 2009, Heftberger 2016, Stutz 2016, Olesen 2017). Weitere Grundlagen und neuere Ansätze unserer Forschung zur Farbfilmanalyse im Besonderen haben wir 2017 und 2018 publiziert (Flückiger 2017, Flückiger und Halter 2018).

Der von uns entwickelte Workflow verbindet die Analyse und Segmentierung der Filme durch eine Video-Annotations-Software mit einem Netzwerk von relationalen Datenbanken. Zur Analyse eines Korpus von mehr als 400 Filmen von 1895 bis 1995 hat das Analyse-Team zunächst – nach einer Evaluation der bestehenden Lösungen – die Video-Annotations-Software ELAN verwendet.

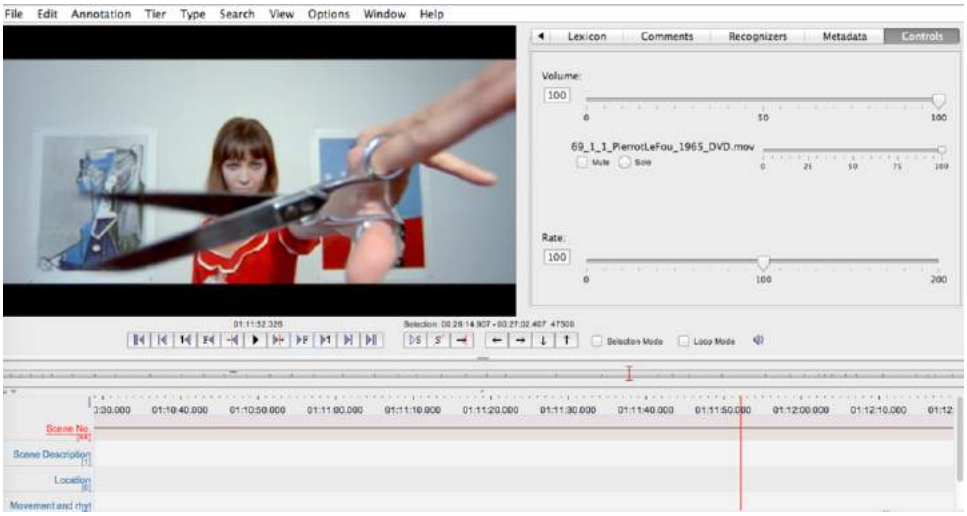


Abbildung 1: ELAN Interface und Template für die Filmanalysen,
Pierrot le fou (FRA 1965, Jean-Luc Godard)

Nach ca. einem Jahr haben wir basierend auf diesen Erfahrungen die Entwicklung eines eigenen Systems in die Wege geleitet: VIAN (*visual video annotation and analysis*), das der visuellen Analyse besser gerecht wird, wurde von Gaudenz Halter anhand der Forschungsdesiderate konzipiert und von ihm seit 2017 in Zusammenarbeit mit dem Visualization and MultiMedia Lab (VMML) der Universität Zürich umgesetzt.

Die Datenbanken für die manuelle Analyse wurden in FileMaker entwickelt, bestehend aus einer Korpus-DB mit den filmografischen Daten, einer Analyse-DB mit rund 1‘200 Konzepten, einer Auswertungs-DB, einer Glossar-DBs mit Definitionen und die Illustrationen dieser Konzepte in einer gesonderten Bilder-DB. Entstanden sind mehr als 17‘000 Segmente mit mehr als 170‘000 Screenshots und mehr als einer halben Million Aufsummierungen, eine Datenmenge, die hohe Anforderungen an die Standardisierung der Auswertung und die Performanz der Bildprozessierung durch die Analyse-Pipeline stellte. Alle DBs werden auf einem zentralen Server gehostet, ebenso die Filme, Screenshots und Resultate.

Entwicklung der visuellen Analyse- und Visualisierungsplattform VIAN

Video-Annotations-Systeme wurden bereits seit den frühen 2000er Jahren für die Filmanalyse entwickelt. 2016 führten wir eine umfassende Analyse der bestehenden Systeme durch, mit ernüchternden Ergebnissen. Viele Video-Annotations-Softwares wurden projektbasiert erstellt und danach nicht aktualisiert (Flückiger 2017). In den letzten Jahren sind besonders webbasierte Media-Suiten entstanden wie zum Beispiel in CLARIAH, die wegen der limitierten Ressourcen für die detaillierte ästhetische Analyse von Filmen in hoher Auflösung jedoch nicht geeignet sind. Grundlegende Untersuchungen von Video-Annotationen finden sich in Giunti (2010 und 2014), ein umfassendes Assessment in Melgar et al. (2017).

Unser Annotationssystem VIAN besteht aus mehreren Layern, die spezifisch auf visuelle Ausdrucksformen des Films ausgerichtet sind. Kernstück sind die temporale Segmentierung, die verbale Annotation sowie ein Screenshot-Manager zur Verwaltung und Visualisierung der Screenshots, siehe Abb. 2 und Screen-Video https://vimeo.com/287959722 . Zusätzlich sind avancierte Methoden der kolorimetrischen Analyse und Visualisierung von Farbschemata implementiert sowie das Vokabular als modulares Menu, ein Auswertungslayer.

Abbildung 2: VIAN Segmentierungslayer mit Screenshot-Manager
Blade Runner 2049 (USA 2017, Denis Villeneuve)

Die temporale Segmentierung von Filmen muss sich an den Forschungsfragen der Analyse ausrichten; sie ist wesentlich komplexer, als man vermuten könnte (Hahn 2009, Cutting et al. 2012). In den Analysen zur Farbe ging es darum, Segmente mit konsistenten Farbschemata zu extrahieren. Diese Aufgabe wird nun durch eine automatische Segmentierung basierend auf der kolorimetrischen Analyse unterstützt.

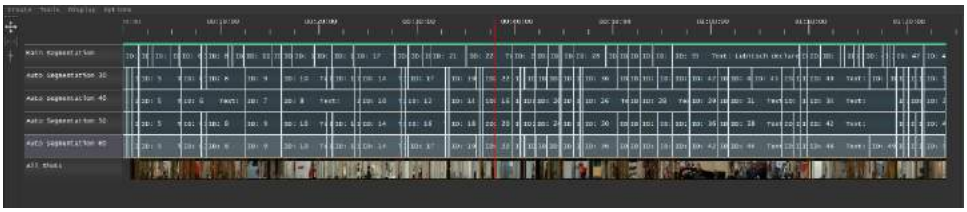


Abbildung 3: Vergleich manuelle (ganz oben) vs. vier Typen automatischer temporaler Segmentierung mit 30 bis 60 Segmenten in *Une femme est une femme* (FRA 1961, Jean-Luc Godard)

Im Laufe der Forschung haben sich die Screenshots zunehmend als heuristische Werkzeuge erwiesen. Einerseits dienen sie der visuellen Repräsentation der Konzepte in der Glossar-DB, andererseits werden sie in den kolorimetrischen Auswertungen selbst prozessiert und in den Visualisierungen verarbeitet. Pro Segment hat das Analyse-Team bis zu 32 Screenshots manuell ausgewählt, um die verschiedenen Einstellungen, Bildkompositionen und typischen Farbschemata abzubilden. Daher war die rasche Entnahme der Screenshots mit einem einzigen Befehl und einer standardisierten Nomenklatur ein Desiderat, das VIAN von bisherigen Tools unterscheidet. Die Screenshots werden in *bins* nach Segmenten geordnet und auch dann automatisch zugewiesen, wenn die Segmentierung nachträglich korrigiert wird.

VIAN erstellt unmittelbar eine kolorimetrische Analyse mit Farbhistogrammen, die für die Auto-Segmentierung wie auch für die Visualisierungen Verwendung finden. Die Farbschemata lassen sich als Paletten sortiert nach Farbverteilung, nach Häufigkeit oder basierend auf einer raumfüllenden Hilbert Kurve durch den Farbraum, organisieren sowie zunehmend verfeinert in Form von Baumdiagrammen darstellen, die der Nutzer anpassen kann (siehe Abb. 4 unten sowie Screen-Video: https://vimeo.com/299804415). Die Farbschemata widerspiegeln damit die ästhetische Konzeption wie auch die prozentualen Anteile der ermittelten Farbtöne im Unterschied zu den üblichen Farbpaletten, die auf K-Means mit fixen Einstellungen beruhen (siehe zum Beispiel Brodbeck 2011).

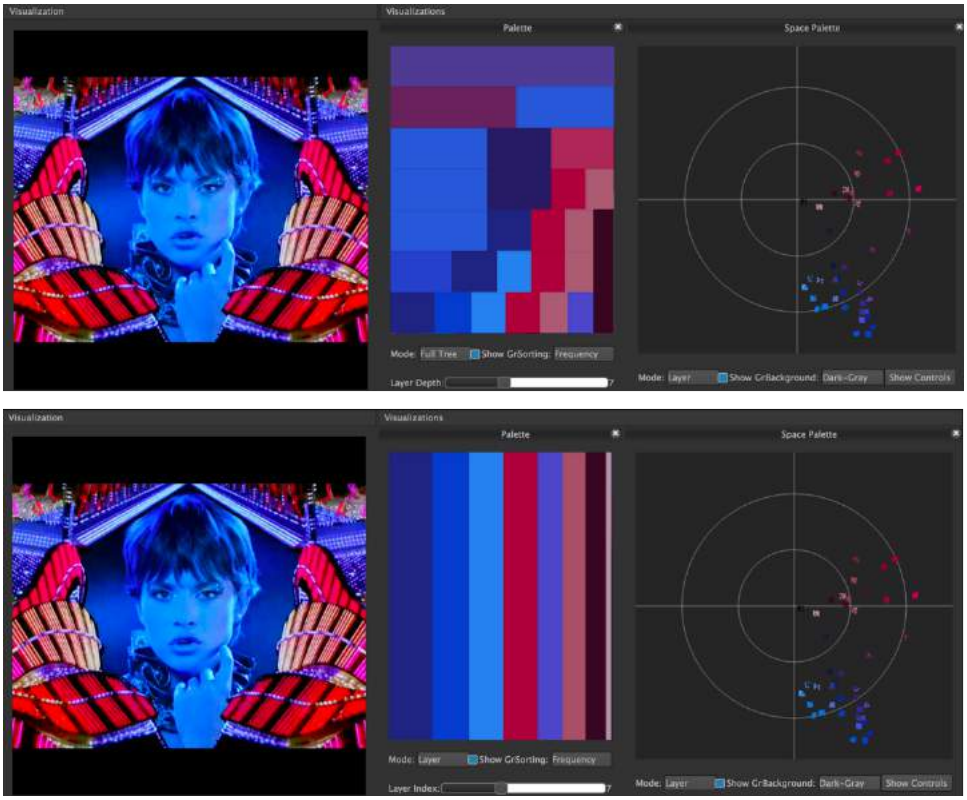


Abbildung 4: VIAN Farbpaletten *One from the Heart* (USA 1981, Francis Ford Coppola), Baumdiagramm (oben Mitte), Selektion von einer Palette mit sieben Farbabstufungen (unten Mitte), Darstellung im CIE L*a*a*-Farbraum (links)

Eine weitere Visualisierungsmethode ordnet die Screenshots oder auch die Farbwerte im wahrnehmungsgerechten CIE L*a*b*-Farbraum (im Folgenden als LAB bezeichnet) an, sodass die Farbverteilung eines ganzen Films oder eines Screenshots unmittelbar sichtbar wird. Schließlich sind die Color-dT-Plots zu nennen, mit denen die zeitliche Entwicklung der Farbschemata sichtbar wird, geordnet nach Sättigung, Chroma, Helligkeit (*luminance*) oder Farbton. Alle diese Visualisierungsmethoden kann der Nutzer nach seinem Erkenntnisinteresse skalieren, individuell anpassen und als Bilder exportieren. Die Plots lassen sich entweder bildbasiert erstellen, sodass man jederzeit in die Bilder hineinzoomen kann, oder als Punkt-Visualisierungen, wobei jeder Punkt einem Farbwert in LAB entspricht.

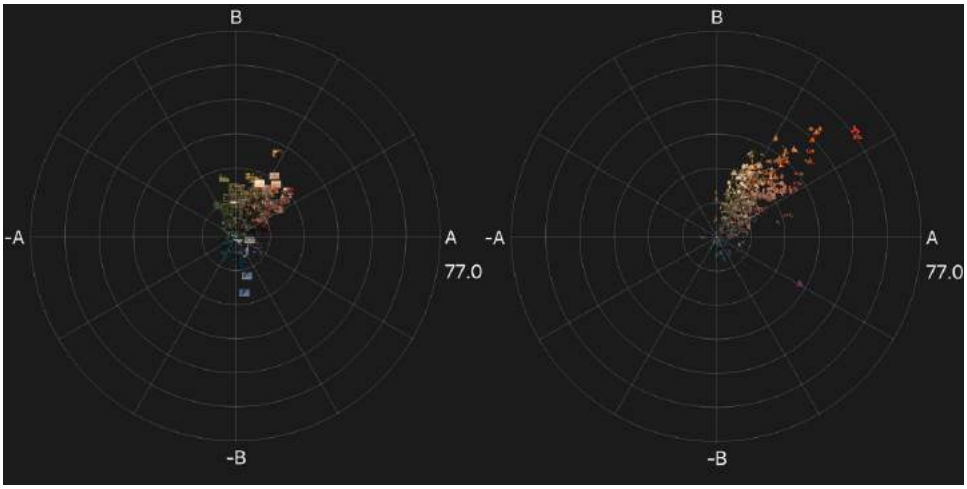


Abbildung 5: VIAN Visualisierungen der Farbverteilung in *Jigokumon* (JAP 1953, Teinosuke Kinugasa), Hintergrund (links), Figuren (rechts)

In der Vergangenheit wurden bereits verschiedene Visualisierungsmethoden für Gemälde und diverse Medien vorgeschlagen, unter anderem von Lev Manovich (2012 und 2015), von Everardo Reyes-García (2014, 2017), Lindsay M. King, und Peter S. Leonard (2017). Frederic Brodbeck (2011) hat Farbpaletten für ganze Filme mit K-Means berechnet und als Kreise dargestellt.

Kevin Ferguson (2013 und 2016) erstellte Z-Projections, indem er alle Bilder eines Films aufsummierte und anschließend normalisierte. Film-Barcodes sind am weitesten verbreitet, um die temporale Entwicklung von Filmen im Überblick darzustellen, siehe zum Beispiel Manuel Burghardt et al. (2016 und 2017). Michael Casey und Mark Williams haben im ACTION-Toolset Histogramme in einer Ähnlichkeitsmatrix visualisiert. Ebenfalls haben verschiedene Filmwissenschaftler bereits bestehende Ansätze wie ImageJ (Ross 2007) und ImagePlot (Manovich 2013) für Farbfilmvisualisierungen eingesetzt wie Adelheid Heftberger (2016) oder Christian Gosvig Olesen et al. (2016).

Abbildung 6: Movie Barcode erstellt in VIAN für *Hero* (HKG / CHN 2002, Yimou Zhang)

Figur-Grund-Trennung

Bereits zu Beginn des Forschungsprojekts stand die Hypothese im Raum, dass das Verhältnis von Figur und Grund ein wesentliches Element von Farbästhetiken sei. Dazu wurde in der manuellen Analyse eine Typologie erstellt, welche dieses Verhältnis anhand der Dimensionen Farbsättigung, Helligkeit, Kontrast sowie Figur-Grund-Inversionen verschiedener Stufen bis hin zur Silhouette systematisch erfasste.

Ab 2017 erarbeitete Noyan Evirgen, wiederum in Kooperation mit dem VMML der Universität Zürich, einen automatischen Workflow für die Figur-Grund-Trennung (Flueckiger et al. 2017). Mit Deep Learning Tools war es möglich, die Figuren aus dem Hintergrund auszuschneiden, indem sie zunächst mit der Objekterkennungssoftware YOLO identifiziert wurden (Redmon et al. 2015). YOLO erstellt *bounding boxes* um die Figuren und anderen identifizierten Objekte, welche die Objekte auch sprachlich benennen. Ein Tiefenerkennungs-Algorithmus (Ha 2016) löst die Figuren vom Hintergrund, die anschließend mit GrabCut ausgeschnitten wurden (Rother et al. 2004). In der Bearbeitung der riesigen Datenmengen hat sich dieser Workflow als zwar zuverlässig, aber zu wenig effizient erwiesen. Daher verwendet Gaudenz Halter nun in VIAN einen Deep-Learning-Ansatz, der die Figuren direkt identifiziert und mit semantischer Segmentation pixelweise markiert.

Abbildung 7: VIAN Figur-Grund-Trennung in *Jigokumon* (JAP 1953, Teinosuke Kinugasa)

Auswertung, Korpusvisualisierungen und Crowd-sourcing-Plattform

Wegen des Datenumfangs mussten die Auswertungen der Analysen zwar außerhalb von FileMaker ausgeführt werden, wurden als Summen anschließend die Auswertungs-DB importiert, sodass die Resultate in allen DBs zur Verfügung stehen, in der Korpus-DB pro Film, in der Glossar-DB pro Konzept, gefiltert nach filmografischen Daten und Subkorpora.

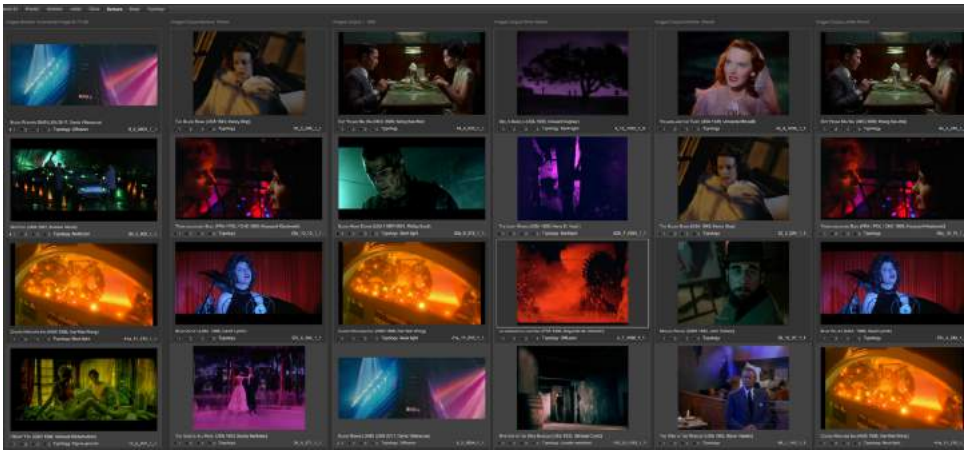


Abbildung 8: Ausschnitt des Eintrags farbiges Licht in der Glossar-DB,
Screenshots sortiert nach verschiedenen Korpora (Perioden, individuelle Auswahl) und typologischen Kriterien.

Zusätzlich ist der *Corpus Visualizer* als Teil von VIAN entstanden und verbindet die Auswertungen der manuellen Analysen mit den Visualisierungsmethoden.

Alle Daten aus den manuellen Analysen sind aus den FileMaker-DBs exportiert und in eine eigens entwickelte Datenstruktur in VIAN importiert worden, die einerseits aus einer visuell lesbaren JSON-Datei besteht, andererseits aus Gründen der Performanz numerische Daten in eine HDF5-Struktur integriert (Halter et al. 2019). Das JSON-Format ermöglicht auch die Interoperabilität mit anderen Video-Annotations-Systemen .

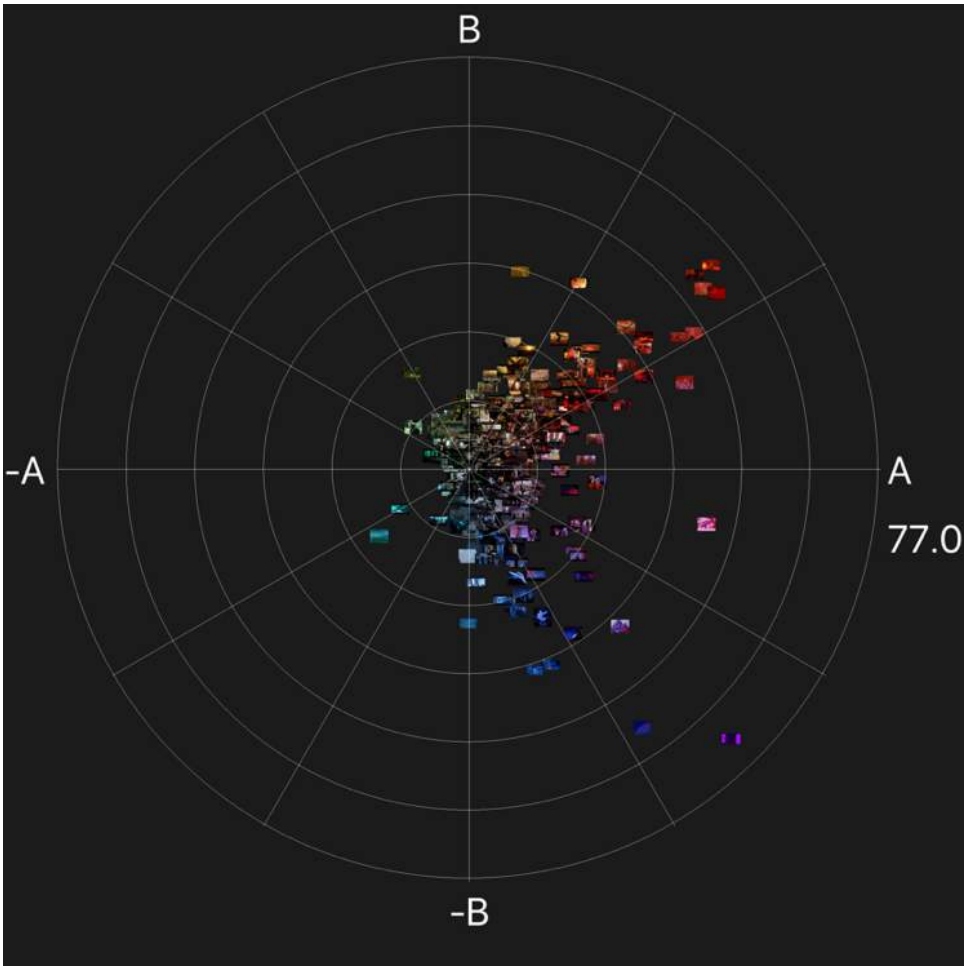


Abbildung 9: Korpusübergreifende Visualisierung des Analysekonzepts farbiges Licht, 1935–1995.

Es lassen sich daher – mit anderen Worten – alle 1‘200 Konzepte des Glossars sowie alle filmografischen Daten abfragen und auf alle unterschiedlichen Arten visualisieren, sowohl Personalstile von einzelnen Filmschaffenden der Bereiche Regie, Kamera, Ausstattung, Kostüme, Farbberatung, aber auch Genres, den Produktionskontext – Firmen, Länder, Perioden – sowie die für die hier beschriebene Forschung essenziellen technischen Farbverfahren.

Die Konzepte, die in der Glossar-DB erfasst, definiert und beschrieben sowie in der Analyse-DB in Form von Checkboxen integriert sind, umfassen ein großes Arsenal an analytischen Dimensionen, von narrativen Strukturen über Figurenemotionen zu verbalen Beschreibungen Farbwerten, Farbschemata und Farbkontrasten, die mit einer 8-stufigen Typologie basierend auf Johannes Itten (1970) systematisiert wurden.

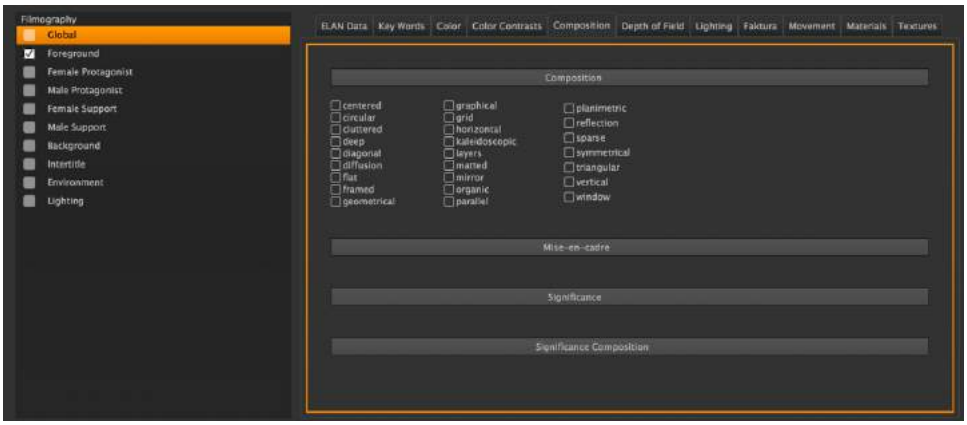


Abbildung 10: Übertragung der Analyse-DB mit allen Konzepten in VIANs Analyse-Widget.

Es sind weiter detaillierte ästhetische Analysekonzepte der Bildkomposition, der Schärfentiefe, Beleuchtung, Bewegung von Kamera und Figuren, Texturen und Muster sowie Materialisierung von in den Filmen dargestellten Kostümen, Objekten und Umgebungen.

Alle Visualisierungen auf Korpus-Ebene sind mit der Figur-Grund-Trennung umgesetzt. Es lassen sich ebenso alle Filme getrennt nach Figuren, Umgebung und Gesamtbild mit allen verschiedenen implementierten Visualisierungsmethoden darstellen.

Neben den Visualisierungen sind die Segmente, in denen die Konzepte vorkommen, mit einer Kurzbeschreibung aufgeführt, sodass sich unmittelbar eine Verbindung der Visualisierung mit der manuellen Analyse herstellen lässt.

Besonders ertragreich sind die Color_dT-Visualisierungen, die auf Filmebene die Farbschemata über die Zeit darstellen, wiederum getrennt für Figur, Grund und globales Bild, um die narrativen Entwicklungen der Filme zu untersuchen und das Verhältnis der Figuren zur Umwelt im Lauf dieser Entwicklung abzubilden.

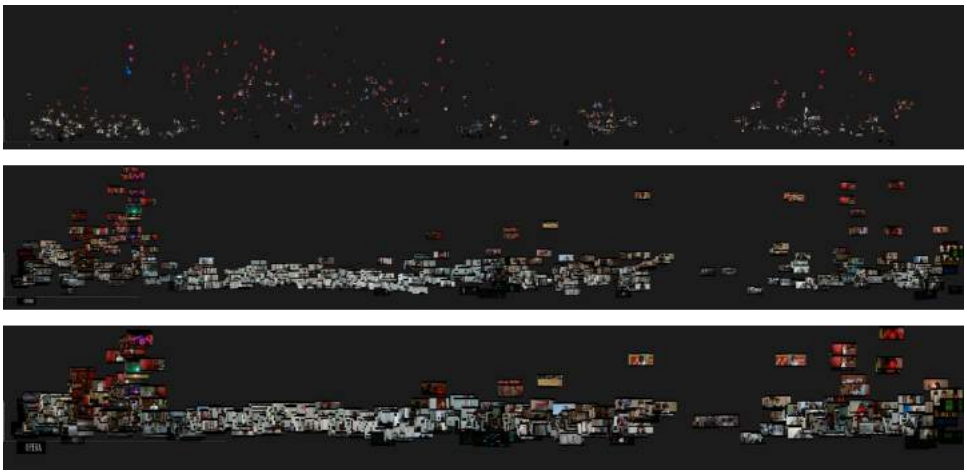


Abbildung 11: Color_dT von Une femme est une femme (FRA 1961, Jean-Luc Godard),
Figuren (oben), Hintergrund (Mitte), ganze Bilder (unten)

Dieses Konzept hat VIAN nun auch für die Visualisierungen auf Korpus-Ebene integriert, sodass die zeitlichen Entwicklungen innerhalb gewählter Perioden sichtbar werden, nun mit der Dimension Zeit in Jahren.

Konzepte werden in der Auswertung im *Features Tool* visualisiert (siehe Screen-Video https://vimeo.com/292861139) und Korrelationen in einer *Korrelationsmatrix* dargestellt.

Zusätzlich entwickelt Silas Weber ebenfalls in Zusammenarbeit mit dem VMML und Gaudenz Halter eine Crowd-sourcing-Plattform als Web-App, in der in Zukunft externe Nutzer ihre eigenen Analysen in VIAN deponieren können. Sie wird Ende Januar 2019 publiziert. Im Herbst 2019 findet zudem eine Ausstellung im Fotomuseum Winterthur statt, für die es eine App geben wird, welche die Exponate mit den Fotos und Quellen in der *Timeline of Historical Film Colors* verbindet und Zugriff auf einzelne Tools in VIAN und die Plattform erlauben wird, um den Besuchern einen spielerischen Zugang zur komplexen Materie zu ermöglichen, nach dem Motto „What is the color scheme of your favorite movie?".

Bibliografie

Brodbeck, Frederic (2011): *Cinemetrics. Film Data Visualization.* In: *Cinemetrics*, (= http://cinemetrics.fredericbrodbeck.de/, abgerufen 05/30/2016).

Cutting, James E./ Brunick, Kaitlin L./ Candan, Ayse (2012): *Perceiving Event Dynamics and Parsing Hollywood Films.* In: *Journal of Experimental Psychology*, Advance online publication, (= http://people.psych.cornell.edu/~jec7/pubs/jephppscenes.pdf, abgerufen 10/15/2016).

Ferguson, Kevin L. (2013): *Western Roundup.* (= http://typecast.qwriting.qc.cuny.edu/2013/10/07/western-roundup/, abgerufen 07/11/2016).

Ferguson, Kevin L. (2016): *The Slices of Cinema. Digital Surrealism as Research Strategy.* **In: Charles R. Acland and Eric Hoyt (eds.):** *The Arclight Guidebook to Media History and the Digital Humanities*. Reframe Books, pp. 270–299, (= http://projectarclight.org/book/).

Flückiger, Barbara (2011): *Die Vermessung ästhetischer Erscheinungen.* In: *Zeitschrift für Medienwissenschaft*, 5, pp. 44–60.

Flueckiger, Barbara (2012): *Timeline of Historical Film Colors.* (= http://zauberklang.ch/filmcolors/, retrieved 11/19/2017).

Flueckiger, Barbara (2017): *A Digital Humanities Approach to Film Colors.* In: *The Moving Image*, 17.2, S. 71–94.

Flueckiger, Barbara/ Evirgen, Noyan/ Paredes, Enrique G./ Ballester-Ripoll, Rafael/ Pajarola, Renato (2017): *Deep Learning Tools for Foreground-Aware Analysis of Film Colors.* In: *AVinDH SIG*, (= https://avindhsig.wordpress.com/deep-learning-tools-for-foreground-aware-analysis-of-film-colors/, abgerufen 04/10/2018).

Flueckiger, Barbara/ Halter, Gaudenz (2018): *Building a Crowdsourcing Platform for the Analysis of Film Colors.* In: *The Moving Image*, 18.1, S. 80–83.

Giunti, Livia (2010): *Problemi dell'analisi del testo di finzione audiovisivo. Verifica e sviluppo di un modello analitico e interpretativo con strumenti digitali*. Università degli Studi di Pisa, (= https://etd.adm.unipi.it/theses/available/etd-10172012-200229/unrestricted/TESI_caricamento.pdf).

Giunti, Livia (2014): *L'analyse du film a l'ère numérique. Annotation, geste analytique et lecture active.* In: *Cinéma & Cie*, 14,22/23, S. 127–143.

Gruber, Klemens/ Wurm, Barbara; Kropf, Vera (eds.) (2009): *Digital Formalism. Die kalkulierten Bilder des Dziga Vertov.* Wien: Böhlau Verlag.

Ha, H./ Im, S./ Park, J./ Jeon, H. G./ Kweon, I. S. (2016): *High-Quality Depth from Uncalibrated Small Motion Clip.* In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 5413–5421.

Hahn, Stefan (2009): *Filmprotokoll Revisited. Ground Truth in Digital Formalism.* In: **Klemens Gruber, Barbara Wurm und Vera Kropf (eds.):** *Digital Formalism: Die kalkulierten Bilder des Dziga Vertov*. Wien: Böhlau Verlag, S. 129–136.

Halter, Gaudenz/ Ballester-Ripoll, Rafael/ Flueckiger, Barbara/ Pajarola, Renato (2019): *VIAN. A Visual Annotation Tool for Film Analysis. Unveröffentlichtes Manuskript.*

Heftberger, Adelheid (2016): *Kollision der Kader. Dziga Vertovs Filme, die Visualisierung ihrer Strukturen und die Digital Humanities*. München: Edition Text + Kritik.

Itten, Johannes (1970): *Kunst der Farbe*. Ravensburg: Ravensburger Buchverlag.

King, Lindsay M./ Leonard, Peter S. (2017): *Processing Pixels.Towards Visual Culture Computation.* Montreal, Canada.

Manovich, Lev (2012): *How to Compare One Million Images?* In: **D. Berry (ed.):** *Understanding Digital Humanities*. London: Palgrave Macmillan UK, S. 249–278.

Manovich, Lev (2013): *Visualizing Vertov.* In: *Russian Journal of Communication*, 5,1, S. 44–55.

Manovich, Lev (2015): *Data Science and Digital Art History.* In: *International Journal for Digital Art History*, 1, (= https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631, abgerufen 08/13/2016).

Melgar Estrada, Liliana/ Hielscher, Eva/ Koolen, Marijn/ Olesen, Christian Gosvig/ Noordegraaf, Julia/ Blom, Jaap (2017): *Film Analysis as Annotation. Exploring Current Tools.* In: *The Moving Image: The Journal of the Association of Moving Image Archivists*, 17,2, S. 40–70.

Olesen, Christian Gosvig (2017): *Film History in the Making. Film Historiography, Digitised Archives and Digital Research Dispositifs*. Amsterdam: University of Amsterdam, (= https://dare.uva.nl/search?identifier=ad68a275-e968-4fce-b91e-4783cd69686c, abgerufen 10/07/2017).

Olesen, Christian Gosvig/ Gorp, Jasmijn van/ Fossati, Giovanna (2016): *Datasets and Colour Visualizations for 'Data-Driven Film History. A Demonstrator of EYE's Jean Desmet Collection.* In: *Creative Amsterdam. An E-Humanities Perspective. A Research Program at the University of Amsterdam*, (= http://www.create.humanities.uva.nl/results/desmetdatasets , abgerufen 11/11/2016).

Redmon, Joseph/ Divvala, Santosh/ Girshick, Ross/ Farhadi, Ali (2015): *You Only Look Once. Unified, Real-Time Object Detection.* In: *arXiv:1506.02640 [cs]*, Juni.

Reyes-García, Everardo (2014): *Explorations in Media Visualization.* New York: ACM, (= http://www.academia.edu/download/35860006/Reyes_2014-Explorations_in_Media_Visualization.pdf, http://ceur-ws.org/Vol-1210/datawiz2014_11.pdf, abgerufen 08/13/2016).

Reyes-García, Everardo (2017): *The Image-interface. Graphical Supports for Visual Information*. Hoboken, NJ: Wiley-ISTE.

Ross, Jacqui (2007): *Colour Analysis Tools in ImageJ*. (= https://www.unige.ch/medecine/bioimaging/files/3814/1208/6041/ColourAnalysis.pdf, abgerufen 02/10/2016).

Rother, Carsten/ Kolmogorov, Vladimir/ Blake, Andrew (2004): *GrabCut. Interactive Foreground Extraction using Iterated Graph Cuts.* In: *ACM Transactions on Graphics (SIGGRAPH)*, Aug.

Stutz, Olivia Kristina (2016): *Algorithmische Farbfilmästhetik. Historische sowie experimentell-digitale Notations- und Visualisierungssysteme des Farbfilms im Zeichen der Digital Humanities 2.0 und 3.0*. Zürich: Universität Zürich.

Danksagung

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No 670446 F*ilmColors.*

Zu danken ist Gaudenz Halter, dem Analyse-Team des Forschungsprojekts ERC Advanced Grant *FilmColors* sowie dem Visualization and MultiMedia Lab der Universität Zürich, geleitet von Prof. Dr. Renato Pajarola.

Understanding Social Structure and Behavior through Responsible Mixed-Methods Research: Bias Detection, Theory Validation, and Data Governance

Diesner, Jana
School of Information Sciences at the University of Illinois at Urbana Champaign

Dieser Vortrag wird auf Deutsch gehalten/This lecture will be given in German

Working with reliable data, metrics, and methods, as well as valid theories, is essential for advancing the computational humanities and social sciences. In this talk, I present our research on the following question: 1) How do limitations related to the provenance and quality of digital social data impact research results? I present on the impact of commonly used techniques for name disambiguation on the properties and dynamics of social networks, highlight measurement-induced biases in metrics and theories, and address means for mitigating these limitations. 2) How can we combine methods from natural language processing and network analysis to jointly consider the content and structure of social relations? I provide an example where we applied domain-adjusted text mining to enhance social networks to validate a classic social science theory in a contemporary setting. 3) How can we assess the impact of information and science on people and society beyond using bibliometric methods? I present our work on predicting the impact of media on individual behavior, cognition, and emotions, and measuring the long-term impact of scientific research on society. 4) When working with human-centered and online data, how can we comply with data governance regulations while still do innovative work? I discuss challenges and opportunities for using digital social data in responsible and practical ways. Overall, the work presented in this talk contributes to making sense of qualitative, distributed, and multi-modal data in a scalable way; and advancing the transparency, responsibility, and ethics of computing and technology as applied to trying to better understand society.

Workshops

Automatic Text and Feature Recognition: Mit READ Werkzeugen Texte erkennen und Dokumente analysieren

Hodel, Tobias
tobias.hodel@uzh.ch
Staatsarchiv des Kantons Zürich, Schweiz

Diem, Markus
diem@cvl.tuwien.ac.at
Computer Vision Lab, TU Wien

Oliveira Ares, Sofia
sofia.oliveiraares@epfl.ch
Digital Humanities Lab, EPF Lausanne

Weidemann, Max
max.weidemann@uni-rostock.de
Citlab, Universität Rostock

Dank *machine learning* und *computer vision* ist seit wenigen Jahren die automatisierte Handschriftenerkennung möglich. Obwohl aktuell einzelne Handschriften bzw. sehr ähnliche Handschriftentypen noch trainiert werden müssen, wird es in absehbarer Zeit allgemeine Modelle geben, die Rohtranskriptionen mit einer Fehlerquote unter 10% ausgeben. Paläographische Kenntnisse werden vor allem zur Korrektur und kritischen Begutachtung der Technik nötig sein.

Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Documents) werden unterschiedliche Aufgaben der Automatisierung (weiter-)entwickelt, um qualitativ gute Ergebnisse mit optimalem Ressourceneinsatz zu erhalten. Ein speziell dafür entwickeltes Tool ist die Software Transkribus und die Transkribus Weboberfläche (für Transkription, Tagging/Annotation und Korrektur in der Layouterkennung). Beide Ansätze verkoppeln auf unterschiedliche Weise die Arbeit von Expertinnen und maschinelle Erkennleistung. Software und Webservice sind frei verfügbar unter www.transkribus.eu . Darüber hinaus wurden im Rahmen von READ weitere Extraktions- und Annotationsmöglichkeiten entwickelt, die im Workshop zusammen mit Transkribus vorgestellt und durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden können.[1]

Transkribus unterstützt alle Prozesse vom Import der Bilder über die Identifikation der Textblöcke und Zeilen, die zu einer detaillierten Verlinkung zwischen Text und Bild führt, sowie die Transkription und Annotation der Handschrift bis zum Export der gewonnen Daten in standardisierten Formaten. Darüber hinaus wurden aber noch weitere Tools und Algorithmen entwickelt, die zur Erkennung von graphischen Features genutzt werden können und Tabellen als solche aufbereiten.

Transkribus als Arbeitsumgebung

Die Erkennung von Texten bedingt den Upload digitaler Bilder und Prozessierung mit Layouterkennungswerkzeugen. Upload und Layoutanalyse können grosse Batches verarbeiten. Die Nachkorrektur von Layoutanalyse ist nur noch in wenigen Fällen nötig.

Je nach Einsatzzweck könne Dokumente entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt oder händisch Transkriptionen erstellt werden. Um im erkannten Text und Variantenlesungen (sog. Keywordspotting) zu suchen reicht in den meisten Fällen die Anwendung von bestehenden Modellen.

Einzig im Umgang mit Tabellen sind weiterhin diverse manuelle Schritte möglich, um eine hochwertige Identifikation zu gewährleisten. Der Workshop wird einen Fokus auf die Bearbeitung und Erkennung von Tabellenstrukturen legen, die halbautomatisch erfolgen kann.

Korrekter Text – entweder durch Transkription oder Korrektur von erkanntem Text entstanden – kann danach zum Training von Handschriftenmodellen verwendet werden. Im Rahmen des Workshops wird das Trainieren von Handschriftenmodellen demonstriert und kann durch die Teilnehmenden selbst ausgetestet werden.

Aufbauend auf den Transkriptionen ist es möglich Entitäten (Personen, Orte, Verweise) auszuzeichnen und textuelle Annotationen (Titel, Marginalien, Fussnoten) innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Visuelle Features wie Seitenzahlen, Titel oder Marginalien lassen sich nach der Auszeichnung als Strukturmodelle trainieren und können für die Erkennung von grösseren Dokumentenmassen verwendet werden. Die Vorgehensweise wird im Rahmen des Workshops vorgeführt und kann selbst nachvollzogen werden. Daneben ist auch die Anreicherung der Dokumente mit *named entities* (Personen, Orten und Organisationen) möglich, sodass simple digitale Editionen grösstenteils in Transkribus erstellt werden können.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen (auch ist es möglich die Standardumformung abzuändern und den eigenen Bedürfnissen anzupassen). Weiter sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX, TXT) implementiert. Schließlich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Zielpublikum

Die Plattform Transkribus ist für unterschiedliche Gruppen konzipiert. Einerseits für Geisteswissenschaftler*innen, die selbst Transkriptionen und Editionen historischer Dokumente erstellen möchten. Andererseits richtet sich die Plattform an Archive, Bibliotheken und andere Erinnerungsinstitutionen, die handschriftliche Dokumente in ihren Sammlungen aufbewahren und ein Interesse an der Suchbarmachung des Materials haben. Angesprochen werden sollen auch Studierende der Geistes-, Archiv- und Bibliothekswissenschaften mit einem Interesse an der Transkription historischer Handschriften.

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout- und Handschriftenerkennung,

lässt sich nur durch die enge Zusammenarbeit zwischen Geisteswissenschaftler*innen und Informatiker*innen sowie anderen Computerspezialist*innen mit unterschiedlichen Voraussetzungen und Ansprüchen an Datenqualität und Herstellung von Transkriptionen erreichen. Die Algorithmen werden somit nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in größeren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die Informatiker*innen sowie Personen aus angrenzenden Fächern sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird.

Die Speicherung der Dokumente erfolgt in der Cloud, gehostet auf Servern der Universität Innsbruck. Die importierten Daten bleiben auch während der Bearbeitung unverändert im Dateisystem liegen und werden ergänzt durch METS und PAGE XML. Alle bearbeiteten Dokumente und Daten bleiben somit in den unterschiedlichen Bearbeitungsstadien nicht nur lokal verfügbar, sondern können für andere Transkribusnutzerinnen und -nutzer freigegeben werden. Dank elaboriertem *user-management* ist die Zuteilung von Rollen möglich.

Die eingespeisten Dokumente und Daten bleiben privat und vor dem Zugriff Dritter geschützt. Von Projektseite können vorgenommene Arbeitsschritte zwecks besserem Verständnis der ausgeführten Arbeiten und letztlich der Verbesserung der Produkte ausgewertet werden.

Die Erkennprozesse werden serverseitig durchgeführt, sodass die Ressourcen auf den lokalen Rechnern nicht strapaziert werden. Transkribus ist mit JAVA und SWT programmiert und kann daher plattformunabhängig (Windows, Mac, Linux) genutzt werden.

Ein- und Ausblicke im Workshop

Der Workshop richtet sich sowohl an Geisteswissenschaftler*innen als auch an Computerwissenschaftler*innen, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Drei zentrale Forschungsaspekte aus READ können im Rahmen des Workshops neben Transkribus *hands-on* ausgetestet werden:

1. Max Weidemann: Das Training von Handschriftenmodellen (HTR+);
2. Sofia Ares Oliveira (*in English*): Identifikation von visuellen Features mit dh-segment;
3. Markus Diem: Aufbereitung und Erkennung von Tabellen mit Transkribus und nomacs.

Programm/Ablauf des Workshops

- Begrüssung und Einführung in READ und Transkribus :45'
- Kurze Beschreibungen der vermittelten Forschungsaspekte (je 15'): 45'
- Kaffeepause: 30'
- Arbeit in Kleingruppen am jeweiligen Forschungsaspekt: 60' (nach ca. 40 Minuten besteht die Möglichkeit die Gruppe zu wechseln)
- Diskussion der Resultate, weiterer Ausblick und Evaluation: (15-)30'

Nach Interesse der Teilnehmenden können während der Gruppenarbeit weitere Tools und Ansätze, die im Rahmen von READ entwickelt wurden, kurz diskutiert werden: 1. Matching von Text und Bild (bspw. aus bestehenden Transkriptionen), 2. Transkribus Learn (e-Learningumgebung), 3. Crowdsourcing-Infrastruktur, 4. ScanTent und DocScan (Fotografieren von Dokumenten mit Android App).

Während des gesamten Workshops stehen vier wissenschaftliche Mitarbeitende des Projekts für Fragen und Auskünfte zur Verfügung.

Tobias Hodel nimmt bereits im Vorfeld gerne Dokumente oder Projektideen an, damit sich die Veranstalter bereits vor dem Workshop Gedanken zu möglichen technischen Umsetzungen machen können.

Das Projekt READ und somit die Weiterentwicklung von Transkribus werden finanziert durch einen Grant der Europäischen Union im Rahmen des Horizon 2020 Forschungs- und Innovationsprogramms (grant agreement No 674943).

Zahl der möglichen Teilnehmerinnen und Teilnehmer: Max. 30 Personen.

Benötigte technische Ausstattung: Beamer und Whiteboard.

Teilnehmende: Eigener Rechner (wenn möglich Installation von Transkribus; Hilfe zur Installation von Transkribus wird 15 Minuten vor der Veranstaltung angeboten).

Rückfragen bitte an tobias.hodel@ji.zh.ch

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

Sofia Ares Oliveira, École Polytechnique de Lausanne, CDH-DHLAB, INN 116 / Station 14 / CH-1015 Lausanne / Switzerland; sofia.oliveiraares@epfl.ch (Electrical engineering, signal processing, computer vision).

Markus Diem, Technische Universität Wien, Institute of Computer Aided Automation Computer Vision Lab, Favoritenstr. 9/183-2, A-1040 Vienna, Österreich; diem@caa.tuwien.ac.at (Computer Vision, Document Analysis, Layout Analysis/Page Segmentation, Cluster Analysis, Automated Flow Cytometry Analysis).

Tobias Hodel, Staatsarchiv des Kantons Zürich, Winterthurerstrasse 170, CH-8057 Zürich, Schweiz; tobias.hodel@ji.zh.ch (Digital Humanities; Automatic Text Recognition; eArchiving; Information Retrieval).

Max Weidemann, Institut für Mathematik, Ulmenstraße 69, Universität Rostock, 18051 Rostock, Deutschland; max.weidemann@uni-rostock.de; (Deep Learning, Information Retrieval und Natural Language Processing).

Fußnoten

1. Einführend siehe die online Tutorials: https://read.transkribus.eu/transkribus/ . Als *hands-on* Anleitung wird der Beitrag von Martin Prell empfohlen: »ps: ich bitt noch mahl umb ver gebung meines confusen und üblen schreibens wegen« – Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung. Online: https://doi.org/10.22032/dbt.34849.

Bibliographie

Leifert, Gundram / Strauß, Tobias / Grüning, Tobias / Wustlich, Welf / Labahn, Roger (2016): *„Cells in Multidimensional Recurrent Neural Networks"* in: *Journal of Machine Learning Research* 17, 1-37.

Prell, Martin (2018): *„»ps: ich bitt noch mahl umb ver gebung meines confusen und üblen schreibens wegen« – Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung“*. Online: https://doi.org/10.22032/dbt.34849.

READ (2018): *„Tutorials and How To Guides“*. Online: https://read.transkribus.eu/transkribus/.

Barcamp: Digitales Publizieren zwischen Experiment und Etablierung

Steyer, Timo
steyer@hab.de
Forschungsverbund Marbach Weimar Wolfenbüttel, Deutschland; Herzog August Bibliothek Wolfenbüttel

Neumann, Katrin
neumann@maxweberstiftung.de
Max Weber Stiftung

Seltmann, Melanie
melanie.seltmann@univie.ac.at
Universität Wien

Walter, Scholger
walter.scholger@uni-graz.at
Universität Graz

Einleitung: Digitales Publizieren in den Geisteswissenschaften

Mit der zunehmend selbstverständlichen Nutzung digitaler Ressourcen und der Etablierung der Digital Humanities rückt auch die Frage nach Formen des digitalen Publizierens in der Wissenschaft ins Blickfeld: Während zunehmend digitale Methoden der Erfassung, Erschließung und Analyse zur Anwendung kommen, bleiben jedoch die Publikationswege häufig noch traditionell und analog geprägt. Dabei bieten digitale Veröffentlichungsformen Potenziale für offene und innovative wissenschaftliche Erkenntnisprozesse sowie eine direktere Wissenschaftskommunikation. Die zunehmende Etablierung von (Open)-Peer-Review-Verfahren wirkt gegen das Vorurteil der vermeintlich geringeren Qualität von digitalen Publikationen; auch wissen die Wissenschaftlerinnen die freie und mobile Verfügbarkeit von digitalen Publikationen zunehmend zu schätzen.

Die sich im Wandel befindenden medialen Bedingungen wirken direkt auf die Akteurinnen im (digitalen) Publikationsprozess ein (DHd-Arbeitsgruppe 2016). Die Rolle und das Zusammenspiel von Urheberinnen, Autorinnen, Verlag und Rezipientinnen werden daher grundlegend in Frage gestellt (Fitzpatrick 2011: 50). Gleichfalls unterliegt die wissenschaftliche Publikation selbst einem Prozess der Neudefinition: Traditionelle Formen wie Monographie oder Zeitschriftenartikel verlieren ihren Ausschließlichkeitsanspruch, da zunehmend digitale Präsentationsformen im wissenschaftlichen Diskurs als vollwertige wissenschaftliche Publikationen angesehen werden (Kohle 2017: 199). Digitale Publikationen interagieren weit mehr als ihre analogen Vorbilder mit anderen mediale Formen, sei es durch die Einbettung von multimedialen Inhalten (Maciocci 2017), Social Media und Forschungsdaten oder durch Verweise auf andere online verfügbaren Ressourcen im Sinne von Linked Open Data (W3C 2017). Die Integration von crossmedialen Inhalten ist technisch bereits möglich, es fehlen allerdings noch Anwendungskonzepte und Best Practice Beispiele.

Die oftmals ungefilterte Offenheit digitaler Medien wirft jedoch auch kritische Fragen der Qualitätssicherung auf, da nicht alle aus dem Kontext der gedruckten Publikation gewohnte Mechanismen greifen (Herb 2012). Dennoch sind Vorteile und Mehrwert des Digitalen evident: Digitale Texte sind leicht aufzufinden, durchsuchbar und im Idealfall schrankenlos kopierbar. Sie begünstigen damit die breite Distribution und Rezeption sowie die Nachnutzung durch digitale (z B. analytische) Verfahren. Anders als im Druck erschienene Publikationen können digitale Publikationen fortgeschrieben werden, ohne ihre Referenzierbarkeit verlieren zu müssen (durch Versionierung). Sie lassen sich mit anderen Texten verknüpfen (Hypertext) und können auf der Basis geeigneter Vokabulare bzw. Ontologien in eine maschinell auswertbare semantische Beziehung mit anderen Dokumenten und Gegenständen treten (Semantic Web). Die bei digitalen Dokumenten favorisierte Trennung von Struktur- und Layoutschicht ermöglicht es, Texte nicht mehr einem starren Präsentationsregime zu unterwerfen, sondern nach Wünschen der BenutzerInnen neue Ansichten oder überhaupt Präsentationsformen jenseits traditioneller Textbegriffe zu generieren. Die kollaborative Text- und Datenpublikation wird im digitalen Raum begünstigt, zieht aber auch Probleme bezüglich der Autorinnenschaft und der Differenzierung der Rollen im digitalen Publikationsprozess nach sich (SoSciSo Redaktion: 2017). Abschließend gilt es, die Schlüsselfunktion von Open Access (OA) und freien Lizenzmodellen (z.B. nach Creative Commons) in digitalen Publikationsprozessen zu betonen: Sie schaffen die Voraussetzungen für ungehindertes Forschen und werden damit zu zentralen Bedingungen wissenschaftlichen Publizierens.

Veranstaltungsformat Barcamp

Mit dem Format eines halbtägigen Barcamps möchte die DHd-AG »Digitales Publizieren« der interessierten Community die Möglichkeit bieten, die soeben skizzierten Themen und Fragen, aber auch andere Aspekte rund um das digitale Publizieren gemeinsam zu diskutieren und sich dazu auszutauschen (Dogunke 2018). Das Format bedingt, dass das Programm maßgeblich von den Teilnehmerinnen gestaltet wird und sowohl dynamisch als auch interaktiv entwickelt werden kann. Das Barcamp möchte Expertinnen und interessierte Wissenschaftlerinnen aus unterschiedlichen Disziplinen zusammenbringen und wird ausreichend Raum bieten, sich in ausgewählte Bereiche der Thematik zu vertiefen, aber auch grundlegende Fragen zu thematisieren. Ziel ist es, gleichermaßen die Ansprüche einer

Informationsveranstaltung mit impulsgebenden Statements zu kombinieren.

Es bestehen zwei Möglichkeiten, Themen für die Veranstaltung zu benennen: Zum einen wird im Vorfeld der Tagung DHd2019 eine Umfrage über den DHd-Blog, Twitter und Mailinglisten stattfinden. Hier entscheidet die Quantität der Nennung einzelner Themen über ihre Annahme. Ähnlich gelagerte Themen werden dabei zusammengefasst bzw. gruppiert. Am Barcamp Interessierte haben dabei auch die Möglichkeit, eine Gestaltungsform für den genannten Vorschlag zu nennen und ihre Rolle zu definieren (s.u.). Spontan können zum anderen aber auch Themen direkt innerhalb des Workshops platziert werden. Die endgültige Tagungsordnung für das Barcamp wird gemeinsam mit dem Plenum zu Beginn des Workshops festgelegt. Die DHd-AG »Digitales Publizieren« möchte die Ergebnisses des Barcamps erstens zur Überarbeitung des Arbeitspapieres »Digitales Publizieren« nutzen und damit einen Beitrag zur Klärung des aktuellen Selbstverständnisses in der Gemeinschaft leisten. Zweitens soll die Veranstaltung der weiteren Vernetzung der Interessierten innerhalb der Community dienen. Drittens soll aber auch das gewählte Format auf seine Eignung geprüft werden, die Kommunikation zwischen der AG und der Community aktiver zu gestalten, woraus sich bei positivem Befund auch weitere Veranstaltungen ergeben könnten.

Potentielle Themen und Fragen

Um eine Vorstellung von potentiellen Themen und der inhaltlichen Gestaltung der Veranstaltung zu bekommen, seien im Folgenden einige Aspekte und zentrale Fragen zum digitalen Publizieren genannt, welche die Verfasserinnen der Einreichung auf der Grundlage eigener Erfahrung und der aktuellen Forschung im Rahmen der AG Digitales Publizieren identifiziert haben:

- Aktuelle und zukünftige Publikationsformate: Welche Rolle wird PDF als Publikationsformat in Zukunft haben? Werden Beiträge direkt in XML verfasst werden können?
- *Data Publications als Publikationsformat* : Wie und in welcher Form können (Forschungs)daten publiziert werden? Welche Formate existieren bereits und gibt es Best Practice Beispiele? Wie können Datenpublikationen als wissenschaftliches Publikationsformat etabliert werden?
- (darauf aufbauend): *Was zählt eigentlich als digitale Publikation und welche Abgrenzungen zu anderen Publikationsformen sind notwendig?* Welche technischen und inhaltlichen Kriterien müssen beispielsweise Blogbeiträge erfüllen, um als wissenschaftliche Publikation zu gelten?
- *Kollaboratives Schreiben* : Wie kann die Rolle der beteiligten Personen kenntlich gemacht werden und welche Rolle gibt es außer der der Autorinnen bei einer Publikation?
- *Infrastrukturen für digitale Publikationen* : Welche Repositorien und Publikationsumgebungen existieren und sind für Forscherinnen im deutschsprachigen Raum zugänglich? Welche Standards haben sich etabliert?
- *Wie kann die Qualität von digitalen Publikationen gemessen werden?* Welche Bedeutung könnte der Impactfaktor in Zukunft haben? Wie kann die Zitationshäufigkeit von digitalen Publikationsformen gesteigert werden?
- *Welche Bedeutung kommt dem traditionellen Intermediären im digitalen Publikationszyklus zu* ? Sind Bibliotheken die neuen Verlage? Ist das hybride Publizieren nur eine Übergangserscheinung oder ein langfristiges Erfolgsmodell?
- *Wie ist der aktuelle Stand bei den Lizenzen und Rechten im Kontext vom digitalen Publizieren?* Wie stark hat sich Open Access wirklich durchgesetzt?
- Bedarf es genuiner Gutachterkulturen für digitale Publikationen?
- Wie gestalten sich digitale Publikationsworkflows?
- Hat beim digitalen Publizieren die wissenschaftliche Kommunikation einen direkteren Einfluss auf die Publikation?
- Warum hat sich bisher trotz der stetig wachsenden Bedeutung von Forschungsdaten das Modell der *enhanced publication* noch nicht durchgesetzt und welche Chancen bestehen für dieses Format (Degwitz 2015: 52)?
- Welche Rolle kommt im Sinne des Titels der *Tagung cross- bzw. intermedialen Inhalten* bei digitalen Publikationen zu *?*

Durchführung

Wer ein Thema vorschlägt, hat gleichzeitig die Möglichkeit, auch ein Durchführungsformat zu wählen. Die unterschiedlichen Formate werden mit der Umfrage zusammen vorgeschlagen. Der Grund für diese flexible und Teilnehmerinnen-gesteuerte Auswahl des Barcamps ist es, dass einige der oben genannten Themen sich eher für ein Expertengespräch eignen, während andere eher in einer gemeinsamen Diskussion thematisiert werden könnten oder Gegenstand eines Impulsreferats sein könnten. Es soll daher weder bei den Inhalten noch bei den Formaten fest Vorgaben geben.

Die ein Thema vorschlagenden Personen können selber angeben, ob sie a) sich für das Thema grundsätzlich interessieren oder sich b) als ExpertIn für das Thema im Rahmen des Workshops zur Verfügung stellen. Zusätzlich werden die Organisatorinnen im Vorfeld des Workshops Expertinnen zu den einzelnen Themen einladen, bzw. Themen als gemeinsame Diskussionen mit dem Plenum planen und vorbereiten. Die Moderation und Durchführung der Veranstaltung wird von Mitgliedern der AG bedient.

Folgende Formate von circa jeweils 30 Minuten Dauer sind denkbar:

1. Expertinnenformat: Eine Expertin bzw. ein Experte hält ein impulsgebendes Referat, danach findet eine moderierte Diskussion statt.
2. Thementische (abhängig vom Raum): Es gibt unterschiedliche Thementische, an denen Expertinnen Rede und Antwort stehen.
3. Diskussionen: Mehrere Expertinnen diskutieren zu einem Thema, danach folgt eine Diskussion mit dem Plenum.
4. Gruppenformat: Kleinere Gruppen diskutieren gemeinsam ausgewählte Themen und präsentieren die Ergebnisse danach dem Plenum.

Vor allem der letzte Punkt scheint für das Tagungsformat gut geeignet zu sein, da dadurch alle Beteiligten involviert

werden. Für die Durchführung dieser unterschiedlichen Formate wäre ein gut unterteilbarer Raum ebenso sinnvoll wie der Einsatz von Moderationsmaterialen (Flipcharts etc.). Eine laufende Dokumentation der Ergebnisse des Barcamps wird während des Workshops über ein Etherpad erfolgen. Des Weiteren sollen zentrale Ergebnisse in die neue Version des Workingspapers "Digitales Publizieren" integriert werden.Abhängig vom Verlauf des Barcamps können weitere Formate, wie z. B. ein Blogbeitrag, möglich sein.

Organisatorisches

Das Barcamp wird von der DHd-AG Digitales Publizieren veranstaltet. Die Planung und Durchführung wird organisiert von:

Katrin Neumann, (Max-Weber-Stiftung), neumann@maxweberstiftung.de , Forschungsinteressen: Digitales Publizieren, Publikationsplattformen, Wissenschaftliches Bloggen

Melanie Seltmann, (Universität Wien), melanie.seltmann@univie.ac.at , Forschungsinteressen: Digitales Publizieren, Natural Language Processing, Citizen Science

Walter Scholger, (Universität Graz), walter.scholger@uni-graz.at , Forschungsinteressen: Digitales Publizieren, Digitale Editionen, Open Access und Lizenzen

Timo Steyer, (Forschungsverbund Marbach Weimar Wolfenbüttel/Herzog August Bibliothek Wolfenbüttel), steyer@hab.de , Forschungsinteressen: Digitales Publizieren, Digitale Editionen, Metadaten und Datemmodellierung

Die Teilnehmerzahl ist auf 40 Personen begrenzt. Der Workshop sollte eine Dauer von einem halben Tag haben. Benötigt werden ein Beamer, Moderationsmaterial und eine Raumgröße, welche die Bildung mehrere Arbeitsgruppen ermöglicht.

Bibliographie

Degkwitz, Andreas (2015): *"Enhanced Publications Exploit the Potential of Digital Media"*, in: Evolving Genres of ETDs for Knowledge Discovery. Proceedings of ETD 2015 18th International Symposium on Electronic Theses and Dissertations 51-59.

DHd-Arbeitsgruppe (2016): *"Digitales Publizieren"*, in: DHd-Arbeitsgruppe (eds.): Working Paper "Digitales Publizieren" http://diglib.hab.de/ejournals/ed000008/startx.htm [letzter Zugriff: 21.09.2018]

Dogunke, Swantje / Steyer, Timo / Mayer, Corinna (2018): *"Barcamp Data and Demons: von Bestands- und Forschungsdaten zu Services. Treffen sich ein Bibliothekar, eine Archäologin, ein Informatiker, ..."*, in: LIBREAS. Library Ideas 33 https://libreas.eu/ausgabe33/dogunke/ [letzter Zugriff: 21.09.2018].

Fitzpatrick, Kathleen (2011): *Planned Obsolescence Publishing, Technology, and the Future of the Academy*. New York: New York Univ. Press.

Herb, Ulrich (2012): *"Offenheit und wissenschaftliche Werke: Open Access, Open Review, Open Metrics, Open Science & Open Knowledge"*, in: **Herb, Ulrich (eds)**: *Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft*. Saarbrücken Universaar 11-44.

Kohle, Hubertus (2017): *"Digitales Publizieren"* in: **Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.)**: *Digital Humanities. Eine Einführung.* Stuttgart: Metzler Verlag 199-205.

Maciocci, Giuliano (2017): *"Designing Progressive Enhancement Into The Academic Manuscript: Considering a design strategy to accommodate interactive research articles"*, in: Blogpost auf eLife Sciences https://elifesciences.org/labs/e5737fd5/designing-progressive-enhancement-into-the-academic-manuscript [letzter Zugriff: 21.09.2018].

Penfold, Naomi (2017): *"Reproducible Document Stack – supporting the next-generation research article"*, in: Blogpost auf eLife Sciences https://elifesciences.org/labs/7dbeb390/reproducible-document-stack-supporting-the-next-generation-research-article [letzter Zugriff: 21.09.2018].

SoSciSo Redaktion (2017): *"Kollaboratives Schreiben mit webbasierten Programmen"*, in: Blogpost auf Social Science Software https://www.sosciso.de/de/2017/kollaboratives-schreiben/ [letzter Zugriff: 21.09.2018].>

W3C (2017): *"W3C Data Activity. Building the Web of Data"* https://www.w3.org/2013/data/ > [letzter Zugriff: 21.09.2018].

DHd 2019 Book of Abstracts Hackathon

Andorfer, Peter

peter.andorfer@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Cremer, Fabian

Cremer@MaxWeberStiftung.de
Max Weber Stiftung, Bonn

Steyer, Timo

steyer@hab.de
Forschungsverbund Marbach Weimar Wolfenbüttel / Herzog August Bibliothek Wolfenbüttel

Einleitung

"Auf der Konferenz [DHd 2018; d. Verf.] war zu jeder Zeit an jedem Ort genug Kompetenz versammelt, um in einer Kaffeepause eine digitale Publikation der Abstracts zu bauen, mit Inhaltsverzeichnis, Volltext- und Schlagwortsuche und anderen netten Features, die sonst zum Standard jeder Webpräsentation in den DH gehören." So postuliert Fabian Cremer in seinem Blog-Post "Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?" (Cremer 2018). Nur, wie Cremer weiter ausführt, niemand habe es gemacht oder irrt der Autor in dieser Aussage, da die Aufgabe nicht so trivial ist, wie auf dem ersten Blick erscheint? Der hier vorgeschlagene Workshop "DHd 2019

Book of Abstracts Hackathon" soll der DHd-Community den Raum bieten, dieser Frage nachzugehen, eine gemeinsame digitale Publikation der Konferenz-Abstracts zu realisieren und so einen Diskussionsimpuls zur Zukunft des Digitalen Publizierens in den Digital Humanities zu geben.

Digitales Publizieren und Digital Humanities

Das digitale Publizieren hat sich im Rahmen des Kanonisierungsprozesses zu einem etablierten Bestandteil der Digital Humanities entwickelt. Dies umfasst sowohl methodische Überlegungen (Kohle 2017) wie auch die praktische Umsetzungen (DHd-AG 2016). Längst existieren etablierte digitale Publikationsorgane der Digital Humanities, welche die Vielfalt und Potentiale der digitalen Publikationsformate demonstrieren und als Vorbild der Publikationspraxis und Wissenschaftskommunikation gelten.[1] Diesen Entwicklungen zum Trotz werden in der Breite und Spitze der Digital Humanities Forschung diese Potentiale nicht ausgeschöpft und traditionelle Publikationspraktiken weiterhin gepflegt (Stäcker 2012). Dies gilt für die Ebene der Technologie, so liegt das Book of Abstracts der DHd 2018 als PDF ohne Strukturdaten vor (Vogeler 2018). Aber gleichfalls auch für die offene Zugänglichkeit, so wurde das deutschsprachige Standardwerk zu den DH nicht als Open Access publiziert (Jannidis et al. 2017 vgl. dazu Stäcker 2017). Die Anforderungen an digitale Publikationen sind schon seit längerem formuliert. "Digital publishing is not simply repackaging a book or article as a computer file, although even a searchable pdf has advantages over paper", bemerkt Borgman in ihrem "Call to Action for the Humanities" und adressiert hier auch dezidiert die Digital Humanities (Borgman 2010: #p16).

Neben den wissenschaftsökonomischen und wissenschaftspolitischen Vorteilen digitalen Publizierens eröffnen digitale Formate neue wissenschaftliche Methoden zur Weiterverarbeitung. Diese Potentiale sind sich insbesondere die mit digitalen Quellen und Daten arbeitenden Geisteswissenschaftler*innen bewusst und Formulieren entsprechende Ansprüche an die Digitalisierung und Bereitstellung der Untersuchungsgegenstände, wie etwa Volltexte mit standardisierter Strukturierung und Interoperabilität sowie mit Entitäten und komplexen Strukturmerkmalen angereichert (Klaffki et al. 2018: 19-20).[2] Diese Ansprüche müssten in den Geisteswissenschaften, in der die eigenen Texte in Form einer kritischen Rezeption und Iteration Teil der Informationsquellen sind, auch an die eigene Textproduktion gestellt werden. Folgerichtig lautet die Empfehlung der DHd-AG "Digitales Publizieren", die semantischen Strukturen zu kodieren, die Dokumente maschinenlesbar und prozessierbar zu machen und PDF nicht als primäres Publikationsformat zu verwenden (DHd-AG 2016). Die TEI-basierten Veröffentlichungen der Digital Humanities Community demonstrieren das Potential der Publikationen als Untersuchungsgegenstand des Faches (Sahle/Henny-Krahmer 2018 und Hannesschläger/Andorfer 2018). Allein die Zusammenführung der strukturierten Datenbasis der Texte mit den vorhandenen Technologien und Methoden neuer Publikationsformate steht häufig noch aus. Diese Lücke adressiert das vorliegende Konzept.

Konzeption des Workshops

Ziele

Der Hackathon liefert einerseits ein Proof-of-Concept für die Implementierung digitaler Technologien in einen Publikationsprozess (der digitalen Geisteswissenschaften) und möchte andererseits ein partizipatives Format zur Unterstützung der DHd-Tagung durch die DHd-Community in kollaborativer Arbeitsform (Hackathon) darstellen. Der Einsatz vorhandener Frameworks aus der Community demonstriert die Leistungsfähigkeit und das vorhandene Potential. Die erarbeiteten Transformationsskripte und Workflows können die Basis für eine Weiterentwicklung und Nachnutzung in institutionellen Publikationsprozessen darstellen. Eine Analyse des Workshops, der Ergebnisse und der damit verbundenen Rezeption liefert die Grundlage für die Formulierung von Empfehlungen zum möglichen zukünftigen Umgang mit den DHd-Abstracts und deren stetig steigende Relevanz als Publikationsformat. Der Hackathon kann dabei sowohl technologische wie methodische Impulse für das Digitale Publizieren in der DHd-Community liefern.

Vorbereitungen

Die Datengrundlage für den Workshop bilden die zur DHd 2019 eingereichten Abstracts im dhc-Format und idealerweise in einer harmonisierten TEI-Version. Die Workshopleiter vertrauen hier auf die wertvolle Redaktionsarbeit der DHd-Konferenzorganisation, wie es die DHd 2018 vorbildlich umgesetzt hat.[3] Um die hier angestrebten Verarbeitungsprozesse und Methoden des Digitalen Publizierens umsetzen zu können, müssen die Daten weiter vorbereitet und angereichert werden. Die vorbereitende Prozessierung soll umfassen:

- Harmonisierung (Sichtung; Vereinheitlichung von Ambiguitäten, Setzen der Mindeststandards)
- Strukturdatenauszeichnung (Überschriften, Absatznummerierung, Metadaten)
- Verknüpfung mit Normdaten (Erkennung und Auszeichnung von zentralen Entitäten wie Personen, Orte, Institutionen, Werke)
- Rahmenwerke (Generierung von Listen und Registern für Schlagworte, Titel, Namen)

Die Aufbereitung und Anreicherung der Datenbasis wird von den Organisatoren in Zusammenarbeit mit dem Austrian Centre for Digital Humanities (ACDH) im Vorfeld der DHd 2019 vorgenommen. Die Workshopleitung wird dahingehend mit der Konferenzorganisation und dem Programmkomitee eine enge Kooperation anstreben.

Themen und Arbeitsgruppen

Die Entwicklungsphasen des Hackathons finden als Gruppenarbeit statt. Die parallel arbeitenden Gruppen widmen sich verschiedenen Repräsentationen und Verarbeitungsprozessen der gemeinsamen Datenbasis. Die hier skizzierten Themen und Arbeitsgruppen sind als Vorschläge von Seiten der Organisatoren zu verstehen und

werden abhängig von den Kompetenzen und Interessen der Teilnehmenden zu Beginn des Workshops adaptiert. Die beiden Arbeitsbereiche “Transformation” und “Präsentation” werden für die Ziele des Workshops als zwingend notwendig erachtet und sind daher gesetzt. Weitere Themen und Arbeitsgruppen können von den Teilnehmenden ausgestaltet werden. Die Organisatoren moderieren die Bildung der Arbeitsgruppen und begleiten beratend die Entwicklungsphasen.

AG “Style und Sheet”: Transformation

Die Transformation der Ausgangsdaten (XML/TEI) in verschiedene Zielformate bildet die Grundlage für die verschiedenen Nutzungs- und Rezeptionsformen. Als Zielformate der Stylesheets bieten sich u.a. an: HTML, LaTex, PDF, MS-Word, Markdown, JATS. Idealerweise können die Teilnehmenden eigene, bereits genutzte Transformationen in der Gruppe diskutieren, erweitern und optimieren.

AG “Web und App”: Präsentation

Der Zeitrahmen des Workshops erlaubt keine Neuentwicklung eines Frontends. Für die Realisierung verschiedener Präsentationschichten müssen die Basisdaten in vorhandene Publikationsframeworks integriert werden. Dabei sind generische Ansätze (z.B. eXist Webapp, GitHub pages, eLife Lens Viewer, Jupyter) ebenso möglich wie spezifische oder instiutionelle Lösungen der DH-Community.

AG “Maschine und Modell”: Schnittstellen

Die Bereitstellung der Daten über standardisierte maschinenlesbare Schnittstellen (API) ist eine grundlegende Repräsentationsform zukünftiger Publikationspraktiken. Idealerweise wird hier prototypisch ein Protokoll oder eine Spezifikation implementiert (z.B. OAI-PMH, FCS, DTS).

AG “Wolke und Vektor”: Textanalyse

Durch Tokenisieren, Lemmatisieren, Vektorisieren werden Wortlisten, Wortfrequenzen, Wortwolken, Topicmodelling realisiert. Die Strukturierungs- und Visualisierungsformen sind die Grundlage für alternative Rezeptionsformen und textinterne Analysen bis zu (korpus)linguistischen Methoden.

AG “Beziehung und Geflecht”: Netzwerke

Die mit Normdaten angereicherte Datenbasis bietet die Möglichkeit, die Beziehungen zwischen den Entitäten (Personen, Orte, Institutionen, Werke) zu extrahieren, zu visualisieren und zu analysieren. Zu den Anwendungsfällen gehören Netzwerkanalysen und explorative Navigationsformen.

AG “Medial und Modular”: Multimedia

Die Verwaltung, Adressierung und Einbettung multimedialer Inhalte die sowie Erzeugung und Integration interaktiver Elemente, wie dynamische Visualisierungen, gehören zu den großen Herausforderungen des Digitalen Publizierens. Prototypische Umsetzungen können hierfür Impulse liefern oder Lücken in bestehenden Frameworks aufzeigen.

Agenda:

15’ Begrüßung, Organisatorisches, Vorstellungsrunde,
15’ Vorstellung der Daten und Aufteilung auf Arbeitsgruppen
60’ Entwicklungsphase I
30’ Pause
45’ Entwicklungsphase II
30’ Vorstellung der Ergebnisse
15’ Abschlussdiskussion

Organisation

Ergebnissicherung und Outreach

Die ausgearbeiteten Ergebnisse der Arbeitsgruppen (Daten, Skripte, Anwendungen) werden offen lizensiert und frei zur Verfügung gestellt. Ein Workshopbericht fasst die Ergebnisse übersichtlich als Blogpost zusammen. Gemeinsam mit interessierten Teilnehmenden plant die Workshopleitung eine Ausarbeitung der Ergebnisse als Empfehlung für mögliche Umsetzungen der DHd-Abstracts. Die konkreten Implementierungen sollen während der Konferenz online verfügbar gehalten werden und werden von der Workshopleitung (und freiwillig Teilnehmenden) über soziale Medien bekannt gegeben und zur Diskussion gestellt. Weitere erwünschte Kommunikationskanäle (Website, Email) werden mit der Konferenzorganisation besprochen.

Datenmanagement und Infrastruktur

Die Einrichtung einer dezidierten Organisation auf GitHub.com für den Workshop erlaubt das Management der verschiedenen Entwicklungen und kollaborative Arbeitsformen. Die Dokumentation und das Projektmanagement werden ebenfalls über git umgesetzt (GitHubWiki/ GitHub-Projects/Issues). Die Publikation der Ergebnisse erfolgt über GitHub-Repositories via Zenodo. Grundsätzlich stehen auch nichtkommerzielle Gitlab-Instanzen und das DARIAH-DE-Repository zur Sicherung zur Verfügung. Das ACDH kann für das kurz- und mittelfristige Hosting der im Rahmen des Workshops entwickelten Applikationen und Services die erforderliche Server-Infrastruktur zur Verfügung stellen. Weiter ist die Bereitstellung von mehreren Instanzen der Applikationen Voyant und eXistDB zur Nutzung im Workshop über DARIAH-DE vorgesehen.

Teilnehmer*innen

Die praktischen Arbeitsphasen der Gruppenarbeit erlauben nur kleine Gruppengrößen. Die Zahl der möglichen Teilnehmer*innen beträgt daher maximal 25 Personen. Für die Workshopteilnahme werden zwar keine spezifischen technischen oder methodischen Grundlagen (jenseits des Umgangs mit den TEI/XML-Basisdaten) vorausgesetzt, jedoch erfordert ein Hackathon von den Teilnehmenden selbständiges Arbeiten und die Anwendung vorhandener Kompetenzen auf den Gegenstandsbereich. Der Workshop bietet den Rahmen und Raum für die gemeinsames Arbeiten und offenen Austausch.

Fußnoten

1. Dies zeigt sich am Beispiel der Zeitschrift für digitale Geisteswissenschaften (ZfdG), die als Vorbild für die Zeitschrift Medieval and Early Modern Material Culture Online (MEMO) fungierte.
2. Siehe die Digitalisierungsklassen IV und V für Texte.
3. Die vielgestaltigen TEI-Kodierungen aus den Einreichungen und Transformationen wurden einheitlich angeglichen, dokumentiert auf: https://github.com/GVogeler/DHd2018 .

Bibliographie

Borgman, Christine L. (2010): *"The Digital Future is Now. A Call to Action for the Humanities"*, in: Digital Humanities Quarterly 3 (4): #p16, http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html [letzter Zugriff 26. September 2018].

Cremer, Fabian (2018): *"Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?"*, in: Digitale Redaktion (Blog), 21.03.2018, https://editorial.hypotheses.org/113 [letzter Zugriff 26. September 2018].

DHd-AG "Digitales Publizieren" (2016): *"Digitales Publizieren"*, Working Paper, 01.03.2016, http://diglib.hab.de/ejournals/ed000008/startx.htm [letzter Zugriff 26. September 2018].

Hannesschläger, Vanessa / Andorfer, Peter (2018): *Menschen gendern? Datenmodellierung zur Erhebung von Geschlechterverteilung am Beispiel der TEI2016 Abstracts App*, DHd 2018, Köln, https://doi.org/10.5281/zenodo.1182576

Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (2017)(eds.): *Digital Humanities. Eine Einführung*, Stuttgart: J. B. Metzler.

Kohle, Hubertus (2017): *"Digitales Publizieren"*, in: **Jannidis, Fotis et al. (eds.)**: *Digital Humanities. Eine Einführung*, Stuttgart: J. B. Metzler 199-205.

Klaffki, Lisa / Schmunk, Stefan / Stäcker Thomas (2018): *"Stand der Kulturgutdigitalisierung in Deutschland. Eine Analyse und Handlungsvorschläge des DARIAH-DE Stakeholdergremiums 'Wissenschaftliche Sammlungen'"*, DARIAH-DE Working Papers 26, Göttingen, URN: urn:nbn:de:gbv:7-dariah-2018-1-3 .

Stäcker, Thomas (2012): *"Wie schreibt man digital humanities?"*, in: DHd-Blog, 19.08.2012, https://dhd-blog.org/?p=673 [letzter Zugriff 26. September 2018].

Stäcker, Thomas (2017): *"Digital Humanities : eine Einführung / Fotis Jannidis, Hubertus Kohle, Malte Rehbein (Hg.)"*, in: O-Bib. Das Offene Bibliotheksjournal 4(3): 142-148, https://doi.org/10.5282/o-bib/2017H3S142-148

Vogeler, Georg (2018)(ed.): *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts*, Köln: Universität zu Köln 2018.

Distant Letters: Methoden und Praktiken zur quantitativen Analyse digitaler Briefeditionen

Dumont, Stefan
dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

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

Henny-Krahmer, Ulrike
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Deutschland

Krautter, Benjamin
benjamin.krautter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Neuber, Frederike
neuber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Beschreibung

Briefeditionen sind ein Typus der digitalen Edition, in dem die Vorteile des digitalen Mediums bereits mit am intensivsten fruchtbar gemacht werden.[1] In der alltäglichen Arbeit des Edierens sowie der Software-Entwicklung richtet sich der Blick zum großen Teil meist auf den einzelnen Brief und seine Tiefenerschließung, weniger auf die Menge an Briefen eines Korrespondenzkorpus. Weiterführende quantitative Analysen auf Basis der Tiefenerschließung (vollständige Transkription, Modellierung in XML/TEI, Normdaten etc.) und mit digitalen Methoden, die gerade auch für korpusübergreifende Untersuchungen den Weg ebnen würden, sind traditionellerweise in Editionsprojekten (noch) nicht vorgesehen.[2] Mit dem eintägigen Workshop „Distant Letters" möchten wir ein Panorama an quantitativ orientierten Analysemethoden und -praktiken für Daten digitaler Briefeditionen vorstellen,

vermitteln und diskutieren, um so neue Perspektiven auf Briefkorpora zu erproben.[3] Der Workshop gliedert sich in vier Abschnitte:

Auswertung von Metadaten und Entitäten

Auf der Grundlage von standardisiert kodierten Briefmetadaten in XML/TEI sollen mit der Abfragesprache XQuery zunächst Fragen formuliert werden wie: „Wie viel hat Sender A an Empfänger B insgesamt geschrieben? Wie viel in einem bestimmten Jahr?“ Im Anschluss sollen vergleichende Untersuchungen angestellt werden, denen Fragen wie „Wie viel hat Sender A an Empfänger B und Empfänger C geschrieben?“ oder „Wie gestaltet sich das Verhältnis von gesendeten und empfangenen Briefen in der Korrespondenz von A und B?“ zugrunde liegen. Auch Entitäten aus dem Brieftext können in die Untersuchung mit einbezogen werden („Wie häufig wird Person X im Verlauf der Korrespondenz erwähnt?“). Das Ergebnis von derlei Fragen sind statistische Werte, die, um sie interpretatorisch zugänglicher zu machen, weiter aufbereitet werden müssen, z.B. als Visualisierungen in Diagrammen, Kreisen und Kurven.

Analyse linguistischer Merkmale

Im zweiten Teil wendet sich die Untersuchung den Volltextdaten zu. In den Blick genommen werden dabei linguistische Merkmale auf Token-Ebene (z.B. Lemma und Wortart), die einfachen oder komplexen Abfragen (z.B. nach typischen Adjektiv-Anbindungen bestimmter Substantive, Häufungen einer bestimmten Wortart, festen Wendungen, Kollokationen) an den Text zugrunde gelegt werden können und so u.a. Aufschluss über inhaltliche und stilistische Gegebenheiten ermöglichen. Im Workshop werden Werkzeuge gezeigt und benutzt, die zum einen die automatische linguistische Analyse von Texten, z.B. deren Lemmatisierung und POS-Tagging, erlauben und zum anderen Möglichkeiten der Auswertung annotierter linguistischer Merkmale bieten, z.B. mittels leistungsstarker Suchanfragesprachen oder Möglichkeiten der Visualisierung. Genauer in den Blick genommen und z.T. benutzt werden TXM, Corpus Workbench, DTA und WebLicht.[4]

Topic Modeling

Im dritten Teil des Workshops rücken die Inhalte der Briefkommunikation stärker in das Zentrum des Interesses, wenn Fragen aufgegriffen werden wie „Welche Themen werden behandelt und wie sind diese zeitlich verteilt?“ oder „Gibt es bestimmte Themen, die in einer bestimmten Personengruppe stärker verhandelt werden als in einer anderen?“. Analysiert wird dabei der Volltext der Briefe, zusätzlich können jedoch auch die Briefmetadaten in die Interpretation der Analyseergebnisse einfließen. Für die Modellierung der Topics wird das Tool „Mallet“ verwendet,[5] und es wird im Workshop gemeinsam ein Topic Model für ein Briefkorpus erstellt. Für die Auswertung in Kombination mit Metadaten und Visualisierungen wird der „Topic Modeling Workflow“ (TMW) verwendet.[6] Diskutiert werden soll außerdem, wie sich die Konzepte ‚Topic‘ und ‚Thema‘ zueinander verhalten.[7]

Stilometrie

Im letzten Teil des Workshops soll mit Methoden und Tools der Stilometrie der Sprach- bzw. Schreibstil eines Briefkorpus genauer untersucht werden. Analysiert wird dabei erneut der Volltext, diesmal in orthografisch normalisierter Form. Mögliche Fragestellungen der Analyse sind: „Welche Rückschlüsse erlauben stilometrische Analysen hinsichtlich Sender und Empfänger der Briefe?“, „Korrelieren die stilistische Nähe bzw. Distanz mit Faktoren wie Zeit, Raum oder Empfänger?“. Auch Stilvergleiche werden beispielhaft auf Grundlage der Fragen „Ändert sich der Stil von Sender A in seinen Briefen an die Empfänger B und C?“ und „Variiert der Stil zwischen Geschäfts- und Privatkorrespondenz?“ unternommen. Für die stilometrischen Analysen nutzen wir das „Stylo“-Paket für R.[8] Auch für die stilistischen Analysen ist zu diskutieren, welches Konzept von Stil hinter den gewählten Methoden steht und wie es sich zu anderen Definitionen von Stil verhält.[9]

Ziele

Ziel des Workshops ist es, ein Panorama quantitativer Analysemöglichkeiten für Briefkorpora vorzustellen, das eine Ergänzung zu den traditionellen ‚close reading‘-Verfahren in wissenschaftlichen Editionen darstellt und die Digitalität der Editionsdaten mit Methoden der Digital Humanities noch stärker für quellenimmanente Forschungsfragen fruchtbar macht. Die Teilnehmerinnen und Teilnehmer sollen den Workshop am Ende des Tages mit einem Set an Skripten und Tools verlassen und in der Lage sein, diese auf andere (ggf. eigene) Datensätze anzuwenden. Neben der Vermittlung von technischen Fertigkeiten ist die Diskussion der Methoden und Ergebnisse mit den Teilnehmerinnen und Teilnehmern fester Bestandteil des Workshops. Es soll dabei gemeinsam eruiert werden, auf welchen theoretischen Annahmen die Methoden jeweils basieren, wo ihre Stärken und Schwächen liegen und auch inwieweit die vermittelten Praktiken eine Chance haben könnten, zukünftig ein Bestandteil bei der Erstellung und Nutzung wissenschaftlicher digitaler Briefeditionen zu werden.

Daten

Die Organisatorinnen und Organisatoren stellen XML/TEI und Plain Text Datensätze aus zwei verschiedenen Briefeditionen für den Workshop bereit: ca. 5500 Brieftexte und ebenso viele Metadatensätze aus „Jean Paul - Sämtliche Briefe digital“ (Bernauer / Miller / Neuber 2018) sowie ca. 400 Brieftexte und 3000 Metadatensätze der „edition humboldt digital“ (Ette 2017-). Darüber hinaus steht es den Teilnehmerinnen und Teilnehmer frei, ihre eigenen Datensets (XML/TEI-kodiert und Plain Text) zu verwenden.

Teilnehmerzahl und Vorkenntnisse

Die Anzahl der Teilnehmerinnen und Teilnehmer ist auf 25 begrenzt. Gewisse Grundkenntnisse in der Programmierung (z.B. XSLT/XQuery, Python, R) sind von Vorteil, die im Workshop verwendeten Skripte werden jedoch so vorbereitet, dass sich die Arbeit daran auf Modifikationen und Erweiterungen unter Anleitung der Lehrenden beschränkt. Im Vorfeld des Workshops werden Installationshinweise für die verwendeten Werkzeuge gegeben und die Übungsdaten zum Download bereitgestellt.

Lehrende

Stefan Dumont: Wissenschaftlicher Mitarbeiter bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften, dort u.a. zuständig für die „edition humboldt digital". Wissenschaftlicher Koordinator des DFG-Projekts „correspSearch - Briefeditionen vernetzen". Co-Convener der TEI Special Interest Group „Correspondence". Expertise u.a. mit Standardisierung von Briefkodierung und -metadaten und X-Technologien.

Susanne Haaf: Wissenschaftliche Mitarbeiterin im Projekt CLARIN-D an der Berlin-Brandenburgischen Akademie der Wissenschaften, u.a. beteiligt am Auf- und Ausbau des Deutschen Textarchivs. Doktorandin im Bereich korpusbasierter Untersuchung von Textsortenspezifika. Spezialisierung in Korpusaufbau, Korpuslinguistik, Standards für Text- und Metadaten (insbes. TEI) sowie Textedition.

Ulrike Henny-Krahmer: Wissenschaftliche Mitarbeiterin im Projekt „Computergestützte Literarische Gattungsstilistik" (CLiGS) an der Universität Würzburg. Studium der Regionalwissenschaften Lateinamerika in Köln und Lissabon, Doktorandin in Digital Humanities mit dem Thema „Topic and Style in Subgenres of the Spanish American Novel (1830-1910)".

Benjamin Krautter: Wissenschaftlicher Mitarbeiter im Projekt "Quantitative Drama Analytics" (QuaDramA) an der Universität Stuttgart. Studium der Literaturwissenschaft (Germanistik) und Politikwissenschaft in Stuttgart und Seoul (Südkorea), Doktorand im Bereich Digital Literary Studies mit dem Thema "Quantitative Dramenanalyse - Operationalisierung aristotelischer Kategorien" (Arbeitstitel).

Frederike Neuber: Wissenschaftliche Mitarbeiterin bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften, dort u.a. zuständig für die Briefedition "Jean Paul - Sämtliche Briefe digital". Studium der Italianistik und Editionswissenschaft in Berlin und Rom, Doktorandin in Digital Humanities. Spezialisierung in Editionsphilologie, Datenmodellierung und Programmierung mit X-Technologien.

Fußnoten

1. Der webservice „correspSearch" etwa illustriert die Bedeutung von standardisierter Metadatenerfassung mit Normdaten zur Vernetzbarkeit von Korrespondenzen, https://correspsearch.net/ .

2. Vereinzelt werden quantitative Analysemethoden bereits auf Editionsdaten angewandt: Etwa wird im Kontext des Projekts "Mapping the Republic of Letters" (Stanford University 2013) zur Erschließung der Briefkommunikation und -verbreitung mit verschiedenen statistisch- und/oder netzwerkanalytisch-basierten Visualisierungen experimentiert; Andorfer (2017) erprobt Topic Modelling mit dem Korrespondenzkorpus Leo von Thun-Hohensteins.

3. Nicht Teil dieses Panoramas ist die Netzwerkanalyse, auch wenn diese Form der Auswertung bzw. Visualisierung für Briefdatensätze oft die am naheliegendste scheint. Grundkompetenzen zur Netzwerkanalyse bzw. -visualisierung werden mittlerweile regelmäßig in Workshops vermittelt, z.B. im Rahmen der „Historical Network Research-Community" (http://historicalnetworkresearch.org/). Der Fokus des Workshops richtet sich daher auf bisher weniger berücksichtigte Formen der Analyse von Briefkorpora.

4. http://textometrie.ens-lyon.fr , http://www.deutschestextarchiv.de/ , https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page , http://cwb.sourceforge.net/cqpweb.php .

5. http://mallet.cs.umass.edu/topics.php

6. https://github.com/cligs/tmw

7. Zwar ist das Verfahren für die Ermittlung von Schlüsselwörtern und Themen entwickelt worden, je nach verwendetem Korpus ergeben sich aber auch andere Arten von Topics, z.B. sprachspezifische oder motivische. Vgl. dazu u.a. Rhody (2012) und Schöch (2017).

8. https://sites.google.com/site/computationalstylistics/stylo

9. Für einen Überblick zu verschiedenen Stilbegriffen in der Literatur- und Sprachwissenschaft siehe Herrmann et al. (2015).

Bibliographie

Andorfer, Peter (2017): *Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich*, in: Zeitschrift für digitale Geisteswissenschaften; doi: 10.17175/2017_002 [zuletzt abgerufen 7. Januar 2019].

Bernauer, Markus / Miller, Norbert / Neuber, Frederike (eds.) (2018): *Jean Paul – Sämtliche Briefe digital.* In der Fassung der von Eduard Berend herausgegebenen 3. Abteilung der Historisch-kritischen Ausgabe (1952-1964), im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften überarbeitet und herausgegeben von Markus Bernauer, Norbert Miller und Frederike Neuber; http://jeanpaul-edition.de [letzter Zugriff 7. Januar 2019].

Burrows, John (2002): *Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship*, in: Literary and Linguistic Computing 17/3, S. 267–287.

Dumont, Stefan (2016): *correspSearch – Connecting Scholarly Editions of Letters*, in: Journal of the Text Encoding Initiative [Online], Issue 10; http://journals.openedition.org/jtei/1742 [letzter Zugriff 7.Januar 2019].

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): *Stylometry with R: A Package for Computational Text Analysis*, in: The R Journal 8/1 (2016), S. 107–121.

Ette Ottmar (eds.) (seit 2016): *edition humboldt digital.* Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. Version 3 vom 14.09.2018; https://edition-humboldt.de/ [letzter Zugriff 7. Januar 2019].

Graham, Shawn / Weingart, Scott / Milligan, Ian (2012): *Getting Started with Topic Modeling and MALLET*, in: The Programming Historian 1; https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [letzter Zugriff 7. Januar 2019].

Heiden, Serge (2010): *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*, 24th Pacific Asia Conference on Language, Information and Computation, Nov 2010, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University, S.389–398.

Herrmann, Berenike J. / van Dalen-Oskam, Karina / Schöch, Schöch (2015): *Revisiting Style, a Key Concept in Literary Studies*, in: Journal of Literary Theory 9/1, S. 25–52.

Rhody, Lisa M. (2012): *Topic Modeling and Figurative Language*, in: Journal of Digital Humanities 2/1; http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ [letzter Zugriff 7. Januar 2019].

Schöch, Christof (2017): *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama*, in: Digital Humanities Quarterly 11/2; http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [letzter Zugriff 7. Januar 2019].

Walmsley, Priscilla (2009): *XQuery: Search Across a Variety of XML Data*. O'Reilly Media.

Graphentechnologien in den Digital Humanities: Methoden und Instrumente zur Modellierung, Transformation, Annotation und Analyse

Jarosch, Julian
julian.jarosch@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Kuczera, Andreas
andreas.kuczera@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Schrade, Torsten
torsten.schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Yousef, Tariq
tariq.yousef@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Text und Graph

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. Durch die standardkonforme Auszeichnung der Forschungsgegenstände in TEI werden diese formal und inhaltlich erschlossen. TEI-kodierte Daten beinhalten mannigfaltige semantische Bezüge – aus der Perspektive einer Graphmodellierung sind diese Bezüge jedoch zunächst nur implizit und nicht explizit in den Daten vorhanden (Schrade 2013).

Während die formale Erschließung geisteswissenschaftlicher Forschungsgegenstände mittels XML-basierter Annotationsmethoden mittlerweile als weit fortgeschritten gelten kann, kann die semantische Erschließung häufig noch verdichtet werden. 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 Vernetzung der Fachdaten hinter den Möglichkeiten zurück, die Graphentechnologien bieten (Iglesia u.a. 2015; Grüntgens/Schrade 2016).

Graphentechnologien sind hervorragend für die Modellierung, Speicherung und Analyse semantisch vernetzter Daten – auch verschiedener Modalitäten – geeignet. Einerseits sind in Graphen modellierte Daten hinreichend genau und berechenbar, andererseits bietet die Schema- und Hierarchiefreiheit dieser Datenstrukturierung eine ausreichend große Flexibilität zur Erfassung auch komplexer geisteswissenschaftlicher Sachverhalte (Kuczera 2017).

Eine gegenseitig ausschließende oder separate Modellierung von Forschungsgegenständen *entweder* in klassischen Strukturen – linearem Text (etwa in TEI-XML kodiert), hierarchischer Baumstruktur (Ontologien) – *oder* als Graph ist nicht mehr zwangsläufig. Möglich ist auch die Verknüpfung beider Technologien, so dass das geeignetste Datenmodell für jeden Aspekt der Daten zur Anwendung kommt. Dies erlaubt die Synthese und zusammenfassende Analyse verschiedener Daten- und Objekttypen.

Darüber hinaus werden die Grenzen zwischen den Technologien zunehmend überbrückt: Die Extraktion von Graphstrukturen aus (annotierten) Texten ist ebenso möglich wie die Modellierung von annotiertem Text als Graph in Form von *standoff property markup*.

Es breitet sich also ein Spektrum an Möglichkeiten aus: Von der Ableitung von (ephemeren) Graphen aus (führenden) XML-Texten, über die verlustlose Migration von XML zu Graph, bis zu *text-as-a-graph* als führendes Datenformat mit geeigneten Editionsumgebungen.

Die Gangbarkeit jeder dieser Möglichkeiten und ihre Unterstützung durch flexible Werkzeuge soll in diesem

Workshop nachvollziehbar gemacht werden, sowie die Grundlagen zur eigenständigen Anwendung gelegt werden.

Werkzeuge

Zur Abdeckung des oben dargestellten Spektrums stellen wir vier Werkzeuge vor, die in der Digitalen Akademie der Akademie der Wissenschaften und der Literatur entwickelt werden. Teilweise eng aufeinander aufbauend, bieten sie einen aktuellen Werkzeugkasten der Graphentechnologien.

XTriples

XTriples (http://xtriples.spatialhumanities.de) ist ein Webservice zur Extraktion von RDF-Statements aus XML-Daten zur Vernetzung von Ressourcen im *semantic web*. Dieses Werkzeug ist insbesondere geeignet zur (einmaligen oder wiederkehrenden) Ableitung von RDF-Graphen aus XML-Daten (*RDF-Lifting*).

Grundfunktion des generischen Dienstes ist das Crawling beliebiger XML-Datenbestände und die anschließende Generierung semantischer Aussagen aus den XML-Daten auf Basis definierter Aussagemuster. Wird eine Dateneinheit in einer 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.

eXGraphs

eXGraphs ist der Ausbau des Grundprinzips von XTriples zur Extraktion von *property graphs*, d.h. Graphstrukturen, die über die Subjekt-Prädikat-Objekt-Tripelstruktur von RDF hinausgehen. Der Dienst ist so weit generalisiert, dass grundsätzlich keine Einschränkungen der Komplexität der extrahierten Graphen bestehen.

eXGraphs basiert auf der Graphdatenbank neo4j, das heißt es importiert entweder die gewonnenen Graphen direkt in eine spezifizierte Datenbank, oder oder gibt sie als Cypher-Abfrage zurück. Das Tool ist somit geeignet, Datenbestände von XML in gerichtete Property-Graphen zu migrieren oder wiederkehrend zur Aktualisierung der Graph-Datenbank aufgerufen zu werden. Die Konfiguration der Extraktion und Transformation wird in einer unkomplizierten XML-Konfiguration spezifiziert, deren hierarchische Struktur direkt mit den notwendigen Extraktionsschritten korrespondiert. Die gesuchten Informationen werden mittels XPath angesteuert.

GRACE

GRACE (***gra**ph **c**ontent **e**ditor*) ist eine Web-App, die das Erstellen und Pflegen von Graphdaten in neo4j-Datenbanken über eine GUI anwenderfreundlich ermöglicht. Unterstützt ist die Suche nach bestehenden Daten, das Verknüpfen von bestehenden Knoten mittels neuer Kanten, das Bearbeiten von Knoten, und die Neuanlage von Knoten. Die Attribute (*properties*) der Knoten werden als Tabelle bzw. bei der Bearbeitung als Formular dargestellt; das Nutzererlebnis ist also durchaus einer klassischen Datenbankeingabe oder Registerpflege vergleichbar.

Gegenüber einer klassischen Sacherschließung innovativ ist, dass Kanten zur Modellierung von Beziehungen, Zusammengehörigkeiten – generell für semantische Relationen verwendet werden können. Die Flexibilität und Aussagekraft ist Querverweisen klar überlegen, da die Kanten grundsätzlich klassifiziert sind und durch Attribute weiter spezifiziert werden können. Die Darstellung und Verwaltung der Kanten ist in die Nutzeroberfläche integriert, so dass die Sacherschließung im Graphen ohne Kenntnisse von Datenbankabfragesprachen aufgebaut werden kann.

SPEEDy

SPEEDy (*standoff* p*roperty* edi*tor*, https://github.com/argimenes/standoff-properties-editor) ist ein Editor zur Bearbeitung von *text-as-a-graph* – sowohl zur nativen Erfassung wie auch zur Weiterpflege von Datenbeständen nach Konvertierung.

Bei *standoff properties* werden Text und Annotationen voneinander getrennt gespeichert. Im Unterschied zu Standoff XML sind die in SPEEDy verwendeten *standoff properties* resistent gegen nachträgliche Änderungen, da der Editor die Indizes nach jeder Bearbeitung neu berechnet.

Mit diesem Konzept sind überlappende und auch konkurrierende Annotationshierarchien möglich. Annotationen lassen sich auch in Layern organisieren und in SPEEDy ein- und ausblenden.

Gespeichert werden die Texte im json-Format, wobei in der json-Datei als erstes der reine Text und anschließend die verschiedenen Annotationen abgelegt sind.

Mit dem in SPEEDy realisierten Annotationskonzept mit *standoff properties* werden multiple Annotationshierarchien möglich, die perspektivisch auch den wissenschaftlichen Diskurs abbilden könnten.

Material

Für den Workshop werden Beispieldaten aus den Sozinianischen Briefwechseln[1] herangezogen, die derzeit erfasst und annotiert werden. Charakteristikum dieser Korrespondenzen ist die enge Verzahnung verschiedener Themengebiete – beispielsweise werden astronomische Beobachtungen von Person zu Person entlang akademischer und familiärer Verbindungen weiter berichtet und von Ort zu Ort weitergetragen, um politisch und theologisch interpretiert zu werden … Dieses komplexe Ineinandergreifen der Themenfelder in den Korrespondenzen erfordert eine entsprechend verzahnte Registerstruktur, die die diversen

Relationen zwischen Entitäten verschiedener Art adäquat abbilden kann.

Die Brieftexte werden im TEI-Subset DTABf kodiert, das eindeutige Kodierungen und verlässliche Extraktion von Informationen ermöglicht.

Als Gegenstand der Übungen stehen die Korrespondenzmetadaten im correspDesc-Format wie auch das *named entity tagging* und die Sacherschließung im Brieftext zur Verfügung. Ersteres bildet bereits das Netzwerk von Korrespondenten und Orten ab; zweiteres gewährt Einblick in die inhaltlichen und thematischen Verknüpfungen. Die zugehörigen Register, auf die die Annotationen verweisen, werden sowohl im Ausgangs-XML-Format wie auch in Form des Graphregisters Teil der Übungsdaten sein.

Ablauf

Nach einer kurzen einführenden Standortbestimmung der Graphentechnologien in den DH und einer Vorstellung der Beispieldaten werden in den zwei Workshoptagen die vier oben genannten Werkzeuge vorgestellt. Zu jedem Tool zeigen wir Beispielkonfigurationen bzw. -anwendungen, und bieten Übungen an, die praxisnah an die Forschungsziele des datengebenden Projekts angelehnt sind. Darüber hinaus steht es den TeilnehmerInnen frei, auch mit eigenen Daten zu experimentieren.

Am ersten Workshoptag gehen wir auf die Werkzeuge XTriples und SPEEDy ein. Damit zeigen wir Optionen auf, welche ohne eine Migration von XML zum Graph zur Verfügung stehen: die Beibehaltung von XML als führendes Format, oder die native Erfassung in *text-as-a-graph*. In der Übung demonstrieren wir die Erzeugung einer RDF-Datei zur Verknüpfung von Ortserwähnungen mit einer geographischen Normdatenbank.

Am zweiten Workshoptag liegt unser Fokus auf eXGraphs und GRACE. Am Ende dieses Tages sollen die Grundprinzipien des Zusammenspiels von XML und neo4j klar geworden sein, indem Forschungsdaten zu neo4j migriert und aktualisiert werden, und dort in einer für ein breites Nutzerspektrum zugänglichen GUI bearbeitet werden.

Beschlossen werden soll der Workshop neben dem ausleitenden Resümee durch eine Feedback-Runde im Plenum insbesondere zu den neu entwickelten Werkzeugen eXGraphs und GRACE, sowie zur weiteren Entwicklung von SPEEDy.

Lernziele

Ziel des Workshops ist, einen Einblick in die Durchlässigkeit zwischen traditionellen (linearen und hierarchischen) Datenstrukturen und der Modellierung im Graphen zu bieten. Die Verwendung von Transformations-, Migrations- und Bearbeitungswerkzeugen wird praktisch vermittelt und ihre Position im DH-Ökosystem umrissen. Die interaktive Demonstration wird anwendungsnah auf reale Forschungsdaten und Auswertungsziele aufgebaut. Neben der technischen Kompetenz wird das Bewusstsein für implizit vorhandene und semantisch auswertbare Graphstrukturen in bestehenden XML-Daten geschärft.

Die Übungen werden ausgehend von einem Einstiegsniveau konzipiert, mit der Option, zu höheren Komplexitätsstufen weiterzuarbeiten oder die Methoden auf eigene Daten und Fragestellungen zu transferieren.

Zwei der vorgestellten Werkzeuge sind Neuentwicklungen, so dass dies eine der ersten Gelegenheiten zur Schulung in der Anwendung sein wird.

Zahl der TeilnehmerInnen

Maximal 20.

Technische Voraussetzungen

Die Teilnehmenden benötigen nur Laptops. Es muss im Vorfeld keine Software installiert werden.

Beitragende

Julian Jarosch

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

2007–2014 Studium der Allgemeinen Sprachwissenschaft an der Johannes Gutenberg-Universität Mainz. 2011 Auslandssemester an der Bangor University, Wales. Magisterarbeit zu »Typography and Legibility: Do Typeface, Serifs and Justification influence Reading Behaviour?«. Seit 2015 wiederum an der JGU Mainz Promotionsprojekt »Empirical Typography« an der Schnittstelle Sprachwissenschaft–Buchwissenschaft, 2015–2017 als Stipendiat der Stipendienstiftung Rheinland-Pfalz. Seit 2018 wissenschaftlicher Mitarbeiter der Digitalen Akademie im DFG-Projekt »Die sozinianischen Briefwechsel«.

Andreas Kuczera

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

1993-1998 Studium der Physik und Geschichte an der Justus-Liebig-Universität Gießen (Staatsexamen für das Lehramt an Gymnasien), 2001 Promotion »Grangie und Grundherrschaft«. Zur Wirtschaftsverfassung des Klosters Arnsburg als Stipendiat der hessischen Graduiertenförderung. 2001-2006 Mitarbeiter im DFG-Projekt Regesta Imperii Online. Von 2007–2012 leitend in der Projektverwaltung der Akademie Mainz und der Digitalen Akademie tätig. Sachverständiger der IT-Kommission der Akademie Mainz. Seit 2015 Zuständigkeit im Bereich des Projektes Regesta Imperii.

Torsten Schrade

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

Historiker, Germanist und Anglist, Softwareentwickler und Digitaler Geisteswissenschaftler (seit 2002). Seit 2009 wissenschaftlicher Mitarbeiter der Akademie und Leiter der Digitalen Akademie.

Tariq Yousef

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

Bachelor in Computer Science (Softwareentwicklung) an der AlBaath Universität (Syrien), Master in Computer Science an der Universität Leipzig. Forschungsinteressen: NLP, Datenextraktion, Datenaufbereitung, data mining, Visualisierung und Webentwicklung.

Fußnoten

1. http://www.adwmainz.de/projekte/zwischen-theologie-fruehmoderner-naturwissenschaft-und-politischer-korrespondenz-die-sozinianischen-briefwechsel/informationen.html

Bibliographie

Grüntgens, Max / Schrade, Torsten (2016): *Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising*, in: **Adamou, A. / Daga, E. / Isaksen, L. (Hrsg.)** *Hrsg .): Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe), CEUR Workshop Proceedings* . Aachen , S. 53–64.

Iglesia, Martin de la / Moretto, Nicolas / Brodhun, Maximilian (2015): Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten. In: *Heike Neuroth, Andrea Rapp, Sibylle Söring (Hrsg.): TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften* . Göttingen, S. 91–102. DOI: http://dx.doi.org/10.3249/webdoc-3947 [Letzter Zugriff 09 . 01 .201 9]

Kuczera, Andreas (2017): Graphentechnologien in den Digitalen Geisteswissenschaften. abitech 37, S. 179 –196 . https://doi.org/10.1515/abitech-2017-0042 [Letzter Zugriff 09.01.2019]

Schrade, Torsten (2013): Datenstrukturierung. In: *Über die Praxis des kulturwissenschaftlichen Arbeitens. Ein Handwörterbuch.* Bielefeld: transcript, S. 91–97.

Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse

Kremer, Gerhard
gerhard.kremer@ims.uni-stuttgart.de
Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Einleitung

Das Ziel dieses Tutorials ist es, den Teilnehmerinnen und Teilnehmern konkrete und praktische Einblicke in einen Standardfall automatischer Textanalyse zu geben. Am Beispiel der automatischen Erkennung von Entitätenreferenzen gehen wir auf allgemeine Annahmen, Verfahrensweisen und methodische Standards bei maschinellen Lernverfahren ein. Die Teilnehmerinnen und Teilnehmer können beim Bearbeiten von lauffähigem Programmiercode den Entscheidungsraum solcher Verfahren ausleuchten und austesten. Es werden dabei keinerlei Vorkenntnisse zu maschinellem Lernen oder Programmierkenntnisse vorausgesetzt.

Es gibt keinen Grund, den Ergebnissen von maschinellen Lernverfahren im Allgemeinen und NLP-Tools im Besonderen blind zu vertrauen. Durch die konkreten Einblicke in den "Maschinenraum" von maschinellen Lernverfahren wird den Teilnehmenden ermöglicht, das Potenzial und die Grenzen statistischer Textanalysewerkzeuge realistischer einzuschätzen. Mittelfristig hoffen wir dadurch, den immer wieder auftretenden Frustrationen beim Einsatz automatischer Verfahren für die Textanalyse und deren teilweise wenig zufriedenstellender Ergebnis-Daten zu begegnen, aber auch die Nutzung und Interpretation der Ergebnisse von maschinellen Lernverfahren (d.h. in erster Linie von automatisch erzeugten Annotationen) zu fördern. Zu deren adäquater Nutzung, etwa in hermeneutischen Interpretationsschritten, ist der Einblick in die Funktionsweise der maschinellen Methoden unerlässlich. Insbesondere ist die Art und Herkunft der Trainingsdaten für die Qualität der maschinell produzierten Daten von Bedeutung, wie wir im Tutorial deutlich machen werden.

Neben einem Python-Programm für die automatische Annotierung von Entitätenreferenzen, mit und an dem während des Tutorials gearbeitet werden wird, stellen wir ein heterogenes, manuell annotiertes Korpus sowie die Routinen zur Evaluation und zum Vergleich von Annotationen zu Verfügung. Das Korpus enthält Entitätenreferenzen, die im "Center for Reflected Text Analytics" (CRETA)[1] in den letzten

drei Jahren annotiert wurden, und deckt Texte verschiedener Disziplinen und Sprachstufen ab.

Entitätenreferenzen

Als empirisches Phänomen befassen wir uns mit dem Konzept der Entität und ihrer Referenz. Das Konzept steht für verschiedene linguistische und semantische Kategorien, die im Rahmen der Digital Humanities von Interesse sind. Es ist bewusst weit gefasst und damit anschlussfähig für verschiedene Forschungsfragen aus den geistes- und sozialwissenschaftlichen Disziplinen. Auf diese Weise können unterschiedliche Perspektiven auf Entitäten berücksichtigt werden. Insgesamt werden in den ausgewählten Texten fünf verschiedene Entitätenklassen betrachtet: PER (Personen/ Figuren), LOC (Orte), ORG (Organisationen), EVT (Ereignisse) und WRK (Werke).

Unter Entitätenreferenzen verstehen wir Ausdrücke, die auf eine Entität in der realen oder fiktiven Welt referieren. Das sind zum einen Eigennamen (Named Entities, z.B. "Peter"), zum anderen Gattungsnamen (z.B. "der Bauer"), sofern diese sich auf eine konkrete Instanz der Gattung beziehen. Dabei wird als Referenzausdruck immer die maximale Nominalphrase (inkl. Artikel, Attribut) annotiert. Pronominale Entitätenreferenzen werden hingegen nicht annotiert.

In **literarischen Texten** sind vor allem Figuren und Räume als grundlegende Kategorien der erzählten Welt von Interesse. Über die Annotation von Figurenreferenzen können u.a. Figurenkonstellationen und -relationen betrachtbar gemacht sowie Fragen zur Figurencharakterisierung oder Handlungsstruktur angeschlossen werden. Spätestens seit dem *spatial turn* rückt auch der Raum als relevante Entität der erzählten Welt in den Fokus. Als "semantischer Raum" (Lotmann, 1972) übernimmt er eine strukturierende Funktion und steht in Wechselwirkung mit Aspekten der Figur.

In den **Sozialwissenschaften** sind politische Parteien und internationale Organisationen seit jeher zentrale Analyseobjekte der empirischen Sozialforschung. Die Annotation der Entitäten der Klassen ORG, PER und LOC in größeren Textkorpora ermöglicht vielfältige Anschlussuntersuchungen, unter anderem zur Sichtbarkeit oder Bewertung bestimmter Instanzen, beispielsweise der Europäischen Union.

Textkorpus

Die Grundlage für (überwachte) maschinelle Lernverfahren bilden Annotationen. Um die Annotierung von Entitätenreferenzen automatisieren zu können, bedarf es Textdaten, die die Vielfalt des Entitätenkonzepts abdecken. Bei diesem Tutorial werden wir auf Annotationen zurückgreifen, die im Rahmen von CRETA an der Universität Stuttgart entstanden sind (vgl. Blessing et al., 2017; Reiter et al., 2017a). Das Korpus enthält literarische Texte aus zwei Sprachstufen des Deutschen (Neuhochdeutsch und Mittelhochdeutsch) sowie ein sozialwissenschaftliches Teilkorpus.[2]

Der Parzival **Wolframs von Eschenbach** ist ein arthurischer Gralroman in mittelhochdeutscher Sprache, entstanden zwischen 1200 und 1210. Der *Parzival* zeichnet sich u.a. durch sein enormes Figureninventar und seine komplexen genealogischen Strukturen aus, wodurch er für Analysen zu Figurenrelationen von besonderem Interesse ist. Der Text ist in 16 Bücher unterteilt und umfasst knapp 25.0000 Verse.

Johann Wolfgang von Goethes Die Leiden des jungen Werthers ist ein Briefroman aus dem Jahr 1774. Unsere Annotationen sind an einer überarbeiteten Fassung von 1787 vorgenommen und umfassen die einleitenden Worte des fiktiven Herausgebers sowie die ersten Briefe von Werther an seinen Freund Wilhelm.

Das **Plenardebattenkorpus des deutschen Bundestages** besteht aus den von Stenografinnen und Stenografen protokollierten Plenardebatten des Bundestages und umfasst 1.226 Sitzungen zwischen 1996 und 2015.[3] Unsere Annotationen beschränken sich auf Auszüge aus insgesamt vier Plenarprotokollen, die inhaltlich Debatten über die Europäische Union behandeln. Hierbei wurde pro Protokoll jeweils die gesamte Rede eines Politikers bzw. einer Politikerin annotiert.

Ablauf

Der Ablauf des Tutorials orientiert sich an sog. *shared tasks* aus der Computerlinguistik, wobei der Aspekt des Wettbewerbs im Tutorial vor allem spielerischen Charakter hat. Bei einem traditionellen *shared task* arbeiten die teilnehmenden Teams, oft auf Basis gleicher Daten, an Lösungen für eine einzelne gestellte Aufgabe. Solch eine definierte Aufgabe kann z.B. *part of speech-tagging* sein. Durch eine zeitgleiche Evaluation auf demselben Goldstandard können die entwickelten Systeme direkt verglichen werden. In unserem Tutorial setzen wir dieses Konzept live und vor Ort um.

Zunächst diskutieren wir kurz die zugrundeliegenden Texte und deren Annotierung. Annotationsrichtlinien werden den Teilnehmerinnen und Teilnehmern im Vorfeld zur Verfügung gestellt. Im Rahmen der Einführung wird auch auf die konkrete Organisation der Annotationsarbeit eingegangen, so dass das Tutorial als Blaupause für zukünftige Tätigkeiten der Teilnehmenden in diesem und ähnlichen Arbeitsfeldern dienen kann.

Die Teilnehmerinnen und Teilnehmer versuchen selbständig und unabhängig voneinander, eine Kombination aus maschinellen Lernverfahren, Merkmalsmenge und Parametersetzungen zu finden, die auf einem neuen, vom automatischen Lernverfahren ungesehenen Datensatz zu den Ergebnissen führt, die dem Goldstandard der manuellen Annotation am Ähnlichsten sind. Das bedeutet konkret, dass der Einfluss von berücksichtigten Features (z.B. Groß- und Kleinschreibung oder Wortlänge) auf die Erkennung von Entitätenreferenzen empirisch getestet werden kann. Dabei sind Intuitionen über die Daten und das annotierte Phänomen hilfreich, da simplem Durchprobieren aller möglichen Kombinationen („brute force'“) zeitlich Grenzen gesetzt sind.

Wir verzichten bewusst auf eine graphische Benutzerschnittstelle (vgl. Reiter et al., 2017b) – stattdessen editieren die Teilnehmerinnen und Teilnehmer das (Python)-Programm direkt, nach einer Einführung und unter Anleitung. Vorkenntnisse in Python sind dabei nicht nötig: Das von uns zur Verfügung gestellte Programm ist so aufgebaut, dass auch Python-Neulinge relativ schnell die zu bearbeitenden Teile davon verstehen und damit experimentieren können.

Wer bereits Erfahrung im Python-Programmieren hat, kann fortgeschrittene Funktionalitäten des Programms verwenden.

Wie am Ende jedes maschinellen Lernprozesses wird auch bei uns abschließend eine Evaluation der automatisch generierten Annotationen durchgeführt. Hierfür werden den Teilnehmerinnen und Teilnehmern nach Ablauf einer begrenzten Zeit des Experimentierens und Testens (etwa 60 Minuten) die finalen, vorher unbekannten Testdaten zur Verfügung gestellt. Auf diese Daten werden die erstellten Modelle angewendet, um automatisch Annotationen zu erzeugen. Diese wiederum werden dann mit dem Goldstandard verglichen, wobei die verschiedenen Entitätenklassen sowie Teilkorpora getrennt evaluiert werden. Auch das Programm zur Evaluation stellen wir bereit.

Lernziele

Am hier verwendeten Beispiel der automatischen Annotation von Entitätenreferenzen demonstrieren wir, welche Schritte für die Automatisierung einer Textanalyseaufgabe mittels maschinellen Lernverfahren nötig sind und wie diese konkret implementiert werden können. Die Teilnehmenden des Workshops bekommen einen zusammenhängenden Überblick von der manuellen Annotation ausgewählter Texte über die Feinjustierung der Lernverfahren bis zur Evaluation der Ergebnisse. Die vorgestellte Vorgehensweise für den gesamten Ablauf ist grundsätzlich auf ähnliche Projekte übertragbar.

Das Tutorial schärft dabei das Verständnis für den Zusammenhang zwischen untersuchtem Konzept und den dafür relevanten Features, die in ein statistisches Lernverfahren einfließen. Durch Einblick in die technische Umsetzung bekommen die Teilnehmerinnen und Teilnehmer ein Verständnis für die Grenzen und Möglichkeiten der Automatisierung, das sie dazu befähigt, zum einen das Potenzial solcher Verfahren für eigene Vorhaben realistisch(er) einschätzen zu können, zum anderen aber auch Ergebnisse, die auf Basis solcher Verfahren erzielt wurden, angemessen hinterfragen und deuten zu können.

Zeitplan:

(Dauer in Minuten, ca.)

Im Vorfeld der Veranstaltung: Installationsanweisungen und Support

- (10) Lecture
  - Intro & Ablauf
- (15) Hands-On
  - Test der Installation bei allen
- (50) Lecture
  - Einführung in Korpus und Annotationen
  - Grundlagen maschinellen Lernens
  - Überblick über das Skript (where can you edit what?)
    - Grundlagen Python Syntax
    - Bereitgestellte Features
- (15) Hands-On
  - Erste Schritte
- (30) Kaffeepause
- (60) Hands-On
  - Hack
- (30) Evaluation

Beitragende (Kontaktdaten und Forschungsinteressen)

Der Workshop wird ausgerichtet von Mitarbeitenden des "Center for Reflected Text Analytics" (CRETA) an der Universität Stuttgart. CRETA verbindet Literaturwissenschaft, Linguistik, Philosophie und Sozialwissenschaft mit Maschineller Sprachverarbeitung und Visualisierung. Hauptaufgabe von CRETA ist die Entwicklung reflektierter Methoden zur Textanalyse, wobei wir Methoden als Gesamtpaket aus konzeptuellem Rahmen, Annahmen, technischer Implementierung und Interpretationsanleitung verstehen. Methoden sollen also keine "black box" sein, sondern auch für Nicht-Technikerinnen und -Techniker so transparent sein, dass ihr reflektierter Einsatz im Hinblick auf geistes- und sozialwissenschaftliche Fragestellungen möglich wird.

Gerhard Kremer

gerhard.kremer@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Der Interessenschwerpunkt Gerhard Kremers ist der reflektierte Einsatz von Werkzeugen der Computerlinguistik für geistes- und sozialwissenschaftliche Fragestellungen. Damit zusammenhängend gehören die Entwicklung übertragbarer Arbeitsmethoden und die angepasste, nutzerfreundliche Bedienbarkeit automatischer linguistischer Analysetools zu seinen Forschungsthemen.

Kerstin Jung

kerstin.jung @ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Kerstin Jungs Forschungsinteressen liegen im Bereich der Nachhaltigkeit von (computer)linguistischen Ressourcen und Abläufen sowie der Verlässlichkeitsbeschreibung von automatisch erzeugten Annotationen. Dabei verfolgt sie einen aufgabenbasierten Ansatz und arbeitet an der Schnittstelle zwischen Computerlinguistik und anderen sprach- und textverarbeitenden Disziplinen.

Zahl der möglichen Teilnehmerinnen und Teilnehmer

Zwischen 15 und 25.

Benötigte technische Ausstattung

Es wird außer einem Beamer keine besondere technische Ausstattung benötigt. Es sollte sich um einen Raum handeln, in dem es möglich ist, den Teilnehmenden über die Schulter zu blicken und durch die Reihen zu gehen.

Fußnoten

1. www.creta.uni-stuttgart.de
2. Aus urheberrechtlichen Gründen wird das Tutorial ohne das Teilkorpus zu Adornos ästhetischer Theorie stattfinden, das in den Publikationen erwähnt wird.
3. Die Texte wurden im Rahmen des PolMine-Projekts verfügbar gemacht: http://polmine.sowi.uni-due.de/polmine/

Bibliographie

Kuhn, Jonas / Reiter, Nils (2015): *"A Plea for a Method-Driven Agenda in the Digital Humanities"* in: Digital Humanities 2015: Conference Abstracts, Sydney.

Reiter, Nils / Blessing, André / Echelmeyer, Nora / Koch, Steffen / Kremer, Gerhard / Murr, Sandra / Overbeck, Maximilian / Pichler, Axel (2017a): *"CUTE: CRETA Unshared Task zu Entitätenreferenzen"* in Konferenzabstracts DHd2017, Bern.

Reiter, Nils / Kuhn, Jonas / Willand, Marcus (2017b): *"To GUI or not to GUI?"* in Proceedings of INFORMATIK 2017, Chemnitz.

Blessing, André / Echelmeyer, Nora / John, Markus / Reiter, Nils (2017): *"An end-to-end environment for research question-driven entity extraction and network analysis"* in Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Vancouver.

Lotman, Juri (1972): *Die Struktur literarischer Texte*, München.

Open Graph Space – Einsatz und Potenzial von Graphentechnologien in den digitalen Geisteswissenschaften

Diehr, Franziska
f.diehr@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz

Brodhun, Maximilian
brodhun@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

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

Kollatz, Thomas
thomas.kollatz@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Wübbena, Thorsten
wuebbena@kunst.uni-frankfurt.de
Goethe Universität Frankfurt am Main, Kunstgeschichtliches Institut, Deutschland; Deutsches Forum für Kunstgeschichte Paris, Frankreich

Efer, Thomas
efer@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland; Universität Leipzig, Deutschland

Die AG Graphentechnologie möchte im Rahmen der DHd 2019 zum offenen Austausch von Erfahrungen, Fragestellungen und Ideen rund um das Thema Graphentechnologien in den digitalen Geisteswissenschaften zu einem „Open Graph Space“ einladen.

Graph-basierte Technologien erfreuen sich eines breiten Anwendungsfeldes in verschiedenartigen Projekten und sind Thema zahlreicher Forschungsarbeiten und -vorhaben im Kontext der Digital Humanities. Die Möglichkeiten der Modellierung, Speicherung und vor allem auch der Datenanalyse überzeugen bei der Nutzung von Graphdatenbanken (Kuczera 2017:179). Durch die fein granulare Strukturierung der Daten in einzelne Knoten und deren Verbindung durch gerichtete oder ungerichtete Kanten können diese auch gezielt abgefragt werden (Rodriguez / Neubauer 2011:45). Die Graphenstruktur ist daher für die Analyse von Beziehungsstrukturen deutlich effizienter als andere Technologien: Im Gegensatz zu den dokument- und hierarchie-basierten Retrievalmöglichkeiten der X-Technologien wie XQuery und XPath oder auch den aufwändigen JOIN-Abfragen von tabellarischen Strukturen in SQL zeichnen sich graph-basierte Abfragen durch eine ressourcenschonende Vorgehensweise aus und liefern dabei präzise und ausdrucksfähige Ergebnisse (Rodriguez / Neubauer 2011:30). Projekte wie der Datums-Gazetteer GODOT und das Editionsprojekt Regesta Imperii zeigen eine hohe Performanz in der Datenanalyse und bei der Verbindung und Einbettung neuer Daten (Kuczera 2017:183, 188). Sie stehen damit beispielhaft für die breite Nutzung von Graphentechnologien in Digital Humanities-Projekten.[1]

Mit dem „Open Graph Space“ soll die Gelegenheit zum Erfahrungsaustausch geschaffen werden, um konkrete Frage- und Problemstellungen zu erörtern und mögliche Lösungsansätze zu diskutieren. Darüber hinaus können weitere Anwendungsfelder für graph-basierte Technologien erschlossen werden. Um einen breiten Kreis von Interessierten zu erreichen, lehnt sich das geplante Format an die Methode des „Open Space“ (Owen 2008) und des „BarCamp“ (Hellmann 2012) an. Beides sind freie, offene Veranstaltungsformate mit selbstorganisierendem Charakter bei dem die TeilnehmerInnen das Programm aktiv mitgestalten. Konkrete Themenvorschläge werden ad hoc vom Plenum entwickelt und anschließend in Kleingruppen bearbeitet (Owen 2008: 2). Ziel ist es, freie Assoziationen und spontane Impulse aufzunehmen und Raum für aktive

Debatten zu schaffen (Owen 2008: 15). Es entstehen kreative Ansätze, die das Potenzial haben sich zu konkreten Projekten und auch gemeinsamen Vorhaben zu entwickeln.

TeilnehmerInnen mit unterschiedlichen Interessengebieten und Kenntnisstufen begegnen sich in einer offenen, ungezwungenen Atmosphäre, die dazu anregt, in einen lebendigen Wissensaustausch zu treten (Owen 2010: 11). Dadurch wird ein multiperspektivischer Blick auf Themen und Problemstellungen geschaffen. Angeregt durch andere und auch neue Betrachtungsmöglichkeiten entstehen kreative Impulse, die die TeilnehmerInnen zur aktiven Beteiligung animiert. Durch das sich entfaltende breite Spektrum an Betrachtungsweisen entsteht ein fachlicher Austausch, bei dem Themen auf vielschichtigen Ebenen exploriert werden. Von allgemeinen Fragen zu Graphentechnologien über konkreten Fragen der Modellierung und Abfrage bis hin zu praxisorientierten Problemen spezifischer Anwendungen ist ein weites Themenspektrum vorstellbar.

Mitglieder der AG Graphentechnologie werden anwesend sein und können für bestimmte Fragestellungen und Problematiken als Experten dienen. So wäre es bspw. möglich eine ‚GraphClinic' für konkrete Fragen rund um den Einsatz von Graphentechnologien in spezifischen Anwendungszusammenhängen anzubieten. Neben theoretischen Ansätzen der Weiterentwicklung graph-basierter Technologie möchten wir auch kritische Fragen zum praktischen Einsatz und Nutzen von Graphentechnologien diskutieren.

Auch wenn das Format sich thematisch frei gestaltet, so ist sein Aufbau und seine Durchführung genau geplant. Nach einer Einführung in das Format durch eine/n qualifizierte/n ModeratorIn werden von den Teilnehmenden spontan Themenvorschläge, Fragestellungen oder auch konkrete Probleme als Programmpunkte formuliert. Nachdem alle Vorschläge gesammelt wurden, bekunden die TeilnehmerInnen ihr Interesse an einem Thema. Je nach Gruppengröße werden die Themenvorschläge in einem bereits vorbereiteten Zeit- und Raumplan eingeordnet (Hellmann 2007:107). Anschließend verteilen sich die TeilnehmerInnen auf die Räume und finden sich so mit anderen Interessierten zusammen. Sollte während der Diskussion jemand für sich selbst feststellen, dass er oder sie nichts beitragen kann oder auch für sich nichts Neues gewinnen kann, so steht es jedem frei auch spontan die Gruppe zu wechseln (zur Bonsen 1998:21). In den Sessions wird das Thema vom Vorschlaggebenden nochmals skizziert, damit wird eine gemeinsame Diskussionsgrundlage geschaffen anhand derer alle in den gemeinsamen kreativen Ideenaustausch eintreten können. Hierbei werden gemeinsam Ansätze skizziert, spontan relevante Beispiele herangezogen und tragfähige Konzepte entwickelt. Dies geschieht unter Zuhilfenahme unterschiedlicher Medien wie Whiteboard oder großformatigem Papier sowie auch bspw. Moderationskärtchen. Auch Live-Demos, Impulsvorträge und spontane Kurzpräsentationen von Beteiligten sind erwünscht. Zur Dokumentation der schnelllebigen Kreativprozesse dienen Fotos, Videos und auch das zu Papier gebrachte. Die AG Graphentechnologien möchte den „Open Graph Space" anhand der Aufzeichnungen aufbereiten und frei zugänglich publizieren und damit auch zur weiteren Diskussion sowie zur Vertiefung und Ausgestaltung der formulierten Konzepte anregen.[2]

Erfahrungen aus der Community zeigen, dass ein „Open Space" bzw. „BarCamp" erfolgreich und gewinnbringend für Themenfelder der Digital Humanities genutzt werden kann. So hat bspw. der MWW-Forschungsverbund ein „BarCamp" zum Thema „Data and Demons – Von Bestands- und Forschungsdaten zu Services" veranstaltet[3] und dabei den kreativen und interdisziplinären Austausch in der Digital Humanities Community gefördert. Durch seinen offenen, aktivierenden Charakter eignet sich der „Open Space" insbesondere auch als Pre-Conference Veranstaltung, da er ein breiten Interessentenkreis aus allen digital arbeitenden und forschenden Geisteswissenschaften anspricht und dabei aktive AnwenderInnen als auch Interessierte zusammenbringt.

Mit dieser etwas anderen Form des wissenschaftlichen Austauschs möchte die AG Graphentechnologien eine neue Art des fachlichen Diskurses anregen. Das Format des „Open Space" unterstreicht dabei den kollaborativen und interdisziplinären Charakter der Digital Humanities und dient dabei auch der Verfestigung der Graphentechnologie-Community.

Fußnoten

1. Beispiele für DH-Projekte und Forschungsvorhaben, die Graphentechnologien einsetzen finden sich im Literaturverzeichnis.
2. Als mögliche Publikationsform bietet sich der Hypotheses-Blog der AG Graphentechnologien an: https://graphentechnologien.hypotheses.org/
3. „Data and Demons – Von Bestands- und Forschungsdaten zu Services." Barcamp im Rahmen des Forschungsverbunds Marbach Weimar Wolfenbüttel, 27.- 28. November 2017, Wolfenbüttel.

Bibliographie

Hellmann, Kai-Uwe (2012): *„Barcamps als kommunikative Treffpunkte der Internetszene"* in: **Bieber, Christoph / Leggewie, Claus (eds.)**: *Unter Piraten. Erkundungen in einer neuen politischen Arena*. Bielefeld: transcript 127–136.

Hellmann, Kai-Uwe (2007): *„Die Barcamp Bewegung. Bericht über eine Serie von Unconferences"*, in: **Klein, Ansgar / Legrand, Hans-Josef / Leif, Thomas / Rohwerder, Jan (eds.)**: *Forschungsjournal Soziale Bewegungen 20* (4): 107-110 https://doi.org/10.1515/fjsb-2007-0415 [letzter Zugriff 1.10.2018].

Kuczera, Andreas (2017): „Graphentechnologien in den Digitalen Geisteswissenschaften" in: ABI Technik 37/3: 179–196 https://doi.org/10.1515/abitech-2017-0042 [letzter Zugriff 1.10.2018].

Owen, Harrison (2010): *Expanding Our Now: The Story of Open Space Technology*. Berrett-Koehler Publishers Inc.

Owen, Harrison (2008): *'Open Space' Technology – A User's Guide*. San Francisco: Berrett-Koehler Publishers Inc.

Rodriguez, Marko A. / Neubauer, Peter (2011): *"The Graph Traversal Pattern"* in: **Sakr, Sherif / Pardede, Eric (eds.)**: **Graph Data Management: Techniques and Applications** 29-46 https://arxiv.org/abs/1004.1001 [letzter Zugriff 1.10.2018].

zur Bonsen, Matthias (1998): *„Mit der Konferenzmethode Open Space zu neuen Ideen"* in: HARVARD BUSINESSmanager 3: 19-26.

Dekker, Ronald H. / Birnbaum, David J. (2017): *„It's more than just overlap: Text As Graph"* in: Proceedings of Balisage – The Markup Conference 2017 https://www.balisage.net/Proceedings/vol19/print/Dekker01/BalisageVol19-Dekker01.html [letzter Zugriff 1.10.2018].

Efer, Thomas (2017): *Graphdatenbanken für die textorientierten e-Humanities*. Leipzig. http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-219122 [letzter Zugriff 1.10.2018].

GODOT – Graph of Dated Objects and Texts: https://godot.date/home [letzter Zugriff 1.10.2018].

Jödden, Tim (2013): *Einsatz von Graphdatenbanken zur Repräsentation kultureller Metadaten*. Bachelorarbeit, Universität Osnabrück.

Regesta Imperii: http://www.regesta-imperii.de/ [letzter Zugriff 1.10.2018].

Standoff-Property Editor for Text Annotation: https://argimenes.github.io/standoff-properties-editor/ [letzter Zugriff 1.10.2018].

Stemmaweb: https://stemmaweb.net/stemmaweb/ [letzter Zugriff 1.10.2018].

Textdatenbank und Wörterbuch das Klassischen Maya: http://mayawoerterbuch.de [letzter Zugriff 1.10.2018].

van Zundert, Joris J. / Andrews, Tara L. (2016): *„Apparatus vs. Graph: New Models and Interfaces for Text"* in: **Hadler, F. / Haupt, J. (eds)**: *Interface Critique.* Berlin: Kulturverlag Kadmos 183-205.

Qualitätsstandards und Interdisziplinarität in der Kuration audiovisueller (Sprach-)Daten

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Data Center for the Humanities, Universität zu Köln, Deutschland

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Gorisch, Jan
gorisch@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Rau, Felix
f.rau@uni-koeln.de
Data Center for the Humanities, Universität zu Köln, Deutschland

Wörner, Kai
kai.woerner@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Workshop-Beschreibung

Audiovisuelle Sprachdaten - d.h. Audio- und Videoaufzeichnungen sprachlicher Interaktion mit zugehörigen Metadaten, Transkripten und Annotationen - sind ein Typ multimedialer Daten, der in vielen geisteswissenschaftlichen Disziplinen eine Rolle spielt. Audiovisuelle Korpora und Datensammlungen werden in verschiedenen Teildisziplinen der Sprachwissenschaften (z.B. Sprachdokumentation: Drude et al. 2014, Variationsforschung: Kehrein & Vorberger 2018, Gesprächsforschung: Schmidt 2018, Phonetik: Draxler & Schiel 2018), in der Geschichtswissenschaft (Oral History: z.B. Pagenstecher & Apostolopoulos 2013, Pagenstecher & Pfänder 2017, Leh 2018, Boyd & Larson 2014), in der Qualitativen Sozialforschung (Qualitative Interviews: z.B. Medjedovic 2011), in den Medien- und Kulturwissenschaften (z.B. Klausmann & Tschofen 2011) sowie in ethnologischen oder volkskundlichen Forschungsprojekten (Harbeck et al. 2018) erstellt und als Basis empirischer Forschung verwendet. Mit der zunehmenden Bedeutung, die das Forschungsdatenmanagement und die Bereitstellung von Daten für eine Nachnutzung aktuell und insbesondere im Kontext der Digital Humanities erfahren, stellen sich auch neue Fragen und Herausforderungen für Archive und Datenzentren, die audiovisuelle Sprachdaten bewahren, kuratieren und bereitstellen. Um diese soll es im Workshop gehen.

Der Workshop versteht sich als Fortsetzung einer thematischen Reihe, die mit dem Workshop "Nachhaltigkeit von Workflows zur Datenkuratierung" auf der FORGE 2016 begonnen und bei der DHd-Tagung 2018 in Köln mit dem Workshop "Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten" weitergeführt wurde. Er richtet sich sowohl an Personen, die im Rahmen von Archiven und Datenzentren mit dem Forschungsdatenmanagement audiovisueller Sprachdaten zu tun haben, als auch an Forschende, Lehrende und Studierende, die solche Daten in ihrer akademischen Tätigkeit erstellen, bearbeiten oder nachnutzen.

Der Fokus des Workshops liegt auf Qualitätsstandards für audiovisuelle Sprachdaten vor dem Hintergrund, dass bei deren Nachnutzung interdisziplinäre Perspektiven eine zunehmend wichtige Rolle spielen. So werden etwa Daten, die in den 1950er und 1960er Jahren zur sprachwissenschaftlichen Untersuchung dialektaler Variation des Deutschen erhoben wurden, nun unter kulturwissenschaftlicher Perspektive betrachtet (Klausmann & Tschofen 2011); die Dokumentation bedrohter Sprachen

hat von jeher in ihrem Selbstverständnis neben sprachtypologischen Fragestellungen auch das Bewahren kulturellen Erbes und die Sprachpolitik im Blick (Seifart et al. 2012); und auch audiovisuelle Daten, die ursprünglich der empirischen Fundierung nicht-geisteswissenschaftlicher (psychologischer und medizinischer) Studien dienten, können bei geeigneter Aufbereitung ein Nachleben als Quellen zu Oral History Studies oder linguistischer Forschung führen (von Hodenberg 2018, Möller/Schmidt 2017). Zudem rücken - im Sinne einer "Third Mission" - auch nicht unmittelbar wissenschaftliche Nutzungsszenarien (z.B. als Objekte für Museen und Ausstellungen, als Material für die (Fremd-)Sprachvermittlung oder im Schulunterricht) und das wirtschaftliche Verwertungspotential (insbesondere in der Sprachtechnologie) solcher Daten zunehmend in den Blickpunkt. Schließlich geht mit einer Loslösung von der ursprünglichen Forschungsdisziplin oft auch eine Internationalisierung des Nutzerkreises einher.

Archive und Datenzentren müssen sich der Herausforderung stellen, die Qualitätsstandards, die sie für Ihre Kurationsarbeit etabliert haben, vor diesem Hintergrund neu zu bewerten und geeignet weiter zu entwickeln. Der Wert einer gegebenen Ressource mag unter Berücksichtigung ihres Potentials zur interdisziplinären Nachnutzung anders beurteilt werden, und es kann notwendig werden, technische Standards der Datenrepräsentation oder Instrumente zur Dissemination an dieses Potential anzupassen. Im Einzelnen wird der Workshop daher den folgenden Fragen nachgehen:

- Welche Qualitätsmaßstäbe sollten an Primärobjekte (Bild- und Tonaufnahmen), Sekundärdaten (Transkripte, Annotationen) und Metadaten eines audio-visuellen Sprachkorpus angelegt werden? Inwieweit sind solche Maßstäbe disziplinspezifisch oder abhängig vom Nachnutzungsszenario? Welche disziplinübergreifenden Qualitätsmaßstäbe kommen als gemeinsamer Nenner in Frage?
- Welche intersubjektiven Methoden oder Maße (z.B. Audio Quality Assessment, Inter-Annotator-Agreement) können zur Qualitätsbeurteilung von audiovisuellen Daten herangezogen werden?
- Nach welchen Kriterien lassen sich Qualität, Kurationsaufwand und Nachnutzungswert für eine gegebene Ressource bewerten und zueinander in Bezug setzen? Wie können oder sollten ggf. Teilaufgaben der Kuration einer Ressource (z.B. Digitalisierung von Aufnahmen vs. Aufbereitung von Annotationen) untereinander priorisiert werden?
- Welche Herangehensweisen gibt es, um den Nachnutzungswert disziplinspezifisch entstandener Ressourcen in einem interdisziplinären Kontext zu beurteilen?
- Nach welchen Kriterien können oder sollten innerhalb von Datenzentren oder Verbünden Datenkurationen priorisiert werden? Welche Ansätze für Sammelstrategien gibt es zentrenintern und zentrenübergreifend?
- Welche Verfahren und Standards sind geeignet, um bei Kurationen den interdisziplinären Nachnutzungswert einer Ressource zu steigern?
- Wie sollten sprachspezifische Ressourcen aufbereitet werden, um Möglichkeiten einer Nachnutzung in einem internationalen und mehrsprachigen Umfeld zu verbessern?
- Wie können durch eine geeignete Kuration auch nicht-wissenschaftliche Nutzungsweisen für die Daten ermöglicht werden?

Workshop-Beiträge

Wir haben neun Beiträge zusammengestellt, die sich mehrerer dieser Fragen annehmen und zugehörige Lösungsansätze anhand eigener Arbeiten zur Erstellung, Kuration oder Nutzung von audiovisuellen Sprachdaten konkret illustrieren. Wir erhoffen uns, auf Basis einer Darstellung des Status Quo eine fruchtbare Diskussion über offene Fragen und Zukunftsperspektiven zu den genannten Themen führen zu können. Berichte und Diskussionsbeiträge zu "Work in Progress" sind daher ausdrücklich erwünscht. Die primäre Workshop-Sprache ist Deutsch, Beiträge, Fragen und Kommentare auf Englisch sind aber ebenfalls willkommen.

Der Workshop umfasst folgende Beiträge:

1. Cord Pagenstecher (Center für Digitale Systeme, CeDIS, FU Berlin) stellt die Digitalen Interview-Sammlungen an der Freien Universität Berlin vor und diskutieren Kurationsstrategien und Qualitätsstandards für die Oral History.
2. Frank Seifart (CNRS, laboratoire Dynamique Du Langage, Lyon; Zentrum für Allgemeine Sprachwissenschaft, Berlin) wird über Kurationsarbeiten im Rahmen von "DoReCo: Ein Projekt zur Erstellung von Referenzkorpora aus Dokumentationen 50 kleiner Sprachen" berichten.
3. Almut Leh (Institut für Geschichte und Biographie, Fernuni Hagen) wird vor dem Hintergrund ihrer Erfahrung mit der Archivierung und Analyse von Oral History-Interviews über Akquirierung, Kuratierung und Nachnutzung von qualitativen audio-visuellen Interviews sprechen.
4. Bernd Meyer (Fachbereich Sprach-, Translations- und Kulturwissenschaften, Johannes Gutenberg-Universität Mainz) wird die "Community Interpreting Database (ComInDat)" als ein Pilotprojekt zur Kuration von gedolmetschten Gesprächen aus unterschiedlichen institutionellen Kontexten vorstellen.
5. Hanna Hedeland (Hamburger Zentrum für Sprachkorpora, Universität Hamburg) wird in ihrem Beitrag Fragen zur Anwendbarkeit und Angemessenheit verschiedener Qualitätsstandards in Bezug auf verschiedene Typen von audiovisuellen Sprachdaten thematisieren.
6. Jonathan Blumtritt und Felix Rau (Data Center for the Humanities und Institut für Linguistik, Universität zu Köln) werden Qualitätsicherungsmaßnahmen im Übernahmeprozess am Language Archive Cologne mit Blick auf interdisziplinäre Nachnutzungs- und Discoveryszenarien vorstellen.
7. Thomas Schmidt, Jan Gorisch, Josef Ruppenhofer und Ulf-Michael Stift werden laufende Arbeiten am Archiv für Gesprochenes Deutsch (AGD) des Instituts für Deutsche Sprache in Mannheim unter den genannten Aspekten beleuchten und dabei insbesondere auf interdisziplinäre Schnittstellen zwischen Sprachwissenschaft einerseits und Kulturwissenschaft und Oral History andererseits eingehen.
8. Sabine Imeri (Fachinformationsdienst Sozial- und Kulturanthropologie, UB der HU Berlin) wird Einblicke in

Debatten der ethnologischen Fächer zur Archivierung und Nachnutzung von u.a. audio-visuellen Daten beitragen, und dabei insbesondere forschungsethische Probleme thematisieren.

9. Christoph Draxler (Institut für Phonetik und Sprachverarbeitung, LMU München) präsentiert eine Pilotstudie zur Transkription studentischer Kurzpräsentation. In dieser Studie werden zwei Transkriptionsverfahren verglichen: zum einen die rein manuelle Transkription, zum anderen die manuelle Korrektur von Rohtranskripten der automatischen Sprachverarbeitung.

Beiträge sollen in Slots von 25+15 Minuten präsentiert werden. Wir rechnen mit etwa 30 Workshop-Teilnehmer(inn)en.

Beitragende (Kontaktdaten und Forschungsinteressen)

Jonathan Blumtritt
Data Center for the Humanities
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln
jonathan.blumtritt@uni-koeln.de

Jonathan Blumtritt ist Mitarbeiter im Data Center for the Humanities an der Universität zu Köln und technischer Koordinator im BMBF-Verbundprojekt Kölner Zentrum Analyse und Archivierung von AV -Daten (KA^3).

Christoph Draxler
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
Schellingstr. 3
80799 München
draxler@phonetik.uni-muenchen.de

Christoph Draxler ist Leiter des Bayerischen Archivs für Sprachsignale, einem CLARIN-D Zentrum für gesprochene Sprache. Seine Forschungsinteressen umfassen Web-basierte Dienste und Werkzeuge für die Sprachverarbeitung, z. B. online Perzeptionsexperimente und Transkription per Crowdsourcing, Sprachdatenbanken sowie die Automatisierung des Workflows bei der Erstellung von Sprachdatenbanken.

Jan Gorisch
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
gorisch@ids-mannheim.de

Jan Gorisch ist wissenschaftlicher Mitarbeiter im Programmbereich Mündliche Korpora am Institut für Deutsche Sprache in Mannheim. Neben der Kuration von Korpora gesprochener Sprache liegen seine Forschungsinteressen in der Analyse von Prosodie, Gestik und Konversation.

Hanna Hedeland
Hamburger Zentrum für Sprachkorpora
Universität Hamburg
Max-Brauer-Allee 60
22765 Hamburg
hanna.hedeland@uni-hamburg.de

Hanna Hedeland ist Geschäftsführerin des Hamburger Zentrum für Sprachkorpora und Mitarbeiterin im Projekt CLARIN-D.

Sabine Imeri
Universitätsbibliothek der Humboldt-Universität zu Berlin
Fachinformationsdienst Sozial- und Kulturanthropologie
Jacob-und-Wilhelm-Grimm-Zentrum
Planckstr. 16
10117 Berlin
sabine.imeri.1@ub.hu-berlin.de

Sabine Imeri ist Europäische Ethnologin und führt als wissenschaftliche Mitarbeiterin am Fachinformationsdienst Sozial- und Kulturanthropologie Erhebungen zum Umgang mit Forschungsdaten in den ethnologischen Fächern durch.

Almut Leh
Fern-Universität in Hagen
Institut für Geschichte und Biographie
Feithstr. 152
58097 Hagen
almut.leh@fernuni-hagen.de

Almut Leh ist Historikerin und leitet als wissenschaftliche Mitarbeiterin am Institut für Geschichte und Biographie der Fernuniversität in Hagen das Archiv "Deutsches Gedächtnis", eine Sammlung von derzeit 3.000 Oral History-Interviews, und führt intverviewbasierte Forschungs- und Dokumentationsprojekte.

Bernd Meyer
Johannes-Gutenberg-Universität Mainz
Arbeitsbereich Interkulturelle Kommunikation
An der Hochschule 2
76726 Germersheim
meyerb@uni-mainz.de
0152-33982190

Bernd Meyer ist Sprachwissenschaftler und untersucht institutionelle und mehrsprachige Kommunikation. Insbesondere interessiert er sich für gedolmetschte Gespräche und hat hierzu umfangreiche Datensammlungen aufgebaut.

Cord Pagenstecher
Freie Universität Berlin
Center für Digitale Systeme/Universitätsbibliothek
Ihnestr. 24
14195 Berlin
cord.pagenstecher@cedis.fu-berlin.de

Cord Pagenstecher ist Historiker am Center für Digitale Systeme/Universitätsbibliothek der Freien Universität Berlin. Er betreut das Interview-Archiv "Zwangsarbeit 1939-1945" und weitere Oral-History-Projekte und koordiniert den Bereich E-Research & E-Publishing.

Felix Rau
Institut für Linguistik
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln
f.rau@uni-koeln.de

Felix Rau ist wissenschaftlicher Mitarbeiter am Institut für Linguistik – im Rahmen des BMBF-Verbundprojekts Kölner Zentrum Analyse und Archivierung von AV -Daten (KA^3) und CLARIN-D – und am Language Archive Cologne.

Josef Ruppenhofer
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim

ruppenhofer@ids-mannheim.de

Josef Ruppenhofer ist wissenschaftlicher Mitarbeiter im Programmbereich Mündliche Korpora und Ko-Koordinator des Archivs für Gesprochenes Deutsch (AGD). Neben der Kuration von Korpora gesprochener Sprache liegen seine Forschungsinteressen in den Bereichen Korpuslinguistik, Computerlexikographie, Konstruktionsgrammatik, Sentimentanalyse und Sprache in sozialen Medien.

Thomas Schmidt
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
thomas.schmidt@ids-mannheim.de

Thomas Schmidt ist Leiter des Programmbereichs Mündliche Korpora am Institut für Deutsche Sprache und in dieser Funktion verantwortlich für das Archiv für Gesprochenes Deutsch (AGD), die Datenbank für Gesprochenes Deutsch (DGD) und das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK). Seine Forschungsinteressen liegen in den Bereichen Korpustechnologie, Korpuslinguistik und Computer-Lexikographie.

Frank Seifart
PD Dr. Frank Seifart
CNRS, laboratoire Dynamique Du Langage
14 avenue Berthelot
F-69363 Lyon CEDEX 07
frank.seifart@cnrs.fr

Frank Seifarts Forschungsinteressen gehen aus von der Sprachdokumentation und daraus resultierenden multimedialen Korpora. Er untersucht vergleichend die Morphosyntax, Semantik und Prosodie menschlicher Sprachen, besonders Sprechtempovariation; weitere

Interessen liegen in der Sprachgeschichte und im Sprachkontakt, besonders auf dem Gebiet der morphologischen Entlehnungen. Er ist spezialisiert in den Amazonassprachen Bora und Resígaro.

Ulf-Michael Stift
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
stift@ids-mannheim.de

Ulf-Michael Stift ist Historiker und Mitarbeiter des Archivs für Gesprochenes Deutsch.

Bibliographie

Boyd, Douglas A. / Larson, Mary A. (2014): *Oral History and Digital Humanities – Voice, Access, and Engagement.* Palgrave Studies in Oral History. Basingstoke: Palgrave Macmillan. [10.1057/9781137322029]

Draxler, Christoph / Schiel, Florian (2018): *"Moderne phonetische Datenbanken"*. In: **Kupietz, Marc & Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 179-208.

Drude, Sebastian / Broeder, Daan / Trilsbeek, Paul (2014): *The Language Archive and its solutions for sustainable endangered languages corpora.* Book 2.0, 4, 5-20. doi:10.1386/btwo.4.1-2.5_1.

Harbeck, Matthias / Imeri, Sabine / Sterzer, Wjatscheslaw (2018): *"Feldnotizen und Videomitschnitte. Zum Forschungsdatenmanagement qualitativer Daten am Beispiel der ethnologischen Fächer"*. Erscheint in: o-bib, Schwerpunktheft Forschungsdaten (Heft 2/2018)

von Hodenberg, Christine (2018): *Das andere Achtundsechzig: Gesellschaftsgeschichte einer Revolte.* München: Beck.

Kehrein, Roland / Vorberger, Lars (2018): *"Dialekt- und Variationskorpora"*. In: **Kupietz, Marc / Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 125-150.

Klausmann, Hubert / Tschofen, Bernhard (2011): *"'Sprachalltag'. Ein sprach- und kulturwissenschaftliches Projekt. Zur Alltagssprache in Nord-Baden-Württemberg"*. In: **Wicker, Hubert (ed.)**: *Schwäbisch. Dialekt mit Tradition und Zukunft.* Festschrift zum 10jährigen Bestehen des Fördervereins Schwäbischer Dialekt e.V.. Gomaringen 2011, 91-102.

Leh, Almut (2018): *"Zeitzeugenkonserven. Interviews für nachfolgende Forschergenerationen im Archiv 'Deutsches Gedächtnis'"*, in: Archivar, 71. Jg. (Heft 02, Mai 2018), 153-155.

Medjedovic, Irena (2011): *"Secondary Analysis of Qualitative Interview Data: Objections and Experiences. Results of a German Feasibility Study"*. In: Forum Qualitative Sozialforschung/Forum Qualitative Social Research, 12(3), Art.10,

Möller, Katrin / Schmidt, Thomas (2017): *"The Bonn Longitudinal Study on Ageing (BOLSA) as an interdisciplinary research resource"*. Beitrag zu: *Encounters in Language and Aging Research: Pragmatic Spaces, Longitudinal Studies and Multilingualism.* Third International Conference on Corpora for Language and Aging Research (CLARe 3). Berlin. https://wikis.fu-berlin.de/pages/viewpage.action?pageId=736856191

Pagenstecher, Cord / Apostolopoulos, Nicolas (2013): *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt.* Berlin: Metropol.

Pagenstecher, Cord / Pfänder, Stefan (2017): *"Hidden dialogues. Towards an interactional understanding of Oral History interviews"*. in: **Kasten, Erich / Roller, Katja / Wilbur, Joshua (eds.)**: *Oral History Meets Linguistics*, Fürstenberg/Havel: Kulturstiftung Sibirien, 185-207, http://www.siberian-studies.org/publications/orhili_E.html

Schmidt, Thomas (2018): *"Gesprächskorpora"*. In: **Kupietz, Marc & Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 209-230.

Seifart, Frank / Haig, Geoffrey / Himmelmann, Nikolaus P. / Jung, Dagmar / Margetts, Anna / Trilsbeek, Paul (eds.) (2012): *Potentials of Language Documentation Methods, Analyses and Utilization.* Language Documentation & Conversation, Special Publication No. 3. Honolulu: University of Hawaii Press.

Texte digital annotieren und analysieren mit CATMA 6.0

Horstmann, Jan
jan.horstmann@uni-hamburg.de
Universität Hamburg, Deutschland

Meister, Jan Christoph
jan-c-meister@uni-hamburg.de
Universität Hamburg, Deutschland

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

Schumacher, Mareike
mareike.schumacher@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

In diesem hands-on Workshop werden wir die Möglichkeiten der für Geisteswissenschaftler*innen entwickelten Annotations- und Analyseplattform CATMA 6.0 praktisch erkunden. Es werden keinerlei technische Vorkenntnisse vorausgesetzt. Inhaltlich werden wir uns vor allem den theoretischen und praktischen Aspekten der digitalen Annotation von (literarischen) Texten, als auch der Analyse und Visualisierung dieser Texte und der erstellten Annotationen widmen.

CATMA (Computer Assisted Text Markup and Analysis; www.catma.de) ist ein open-source-Tool, das seit 10 Jahren an der Universität Hamburg entwickelt und derzeit von über 60 Forschungsprojekten weltweit genutzt wird. Die neue Version 6.0 wird im Zuge des DFG-Projektes forTEXT (www.fortext.net) entwickelt und implementiert. Neben erweiterten technischen Möglichkeiten (wie beispielsweise die Möglichkeit der Datenversionierung und die Organisation kollaborativer Arbeit in einer Projektstruktur), bietet die neue Version ein völlig überarbeitetes, intuitiver nutzbares User-Interface, das auf Material Design als dem avanciertesten und am Markt etabliertesten Schema basiert, welches die meisten Nutzer*innen bereits internalisiert haben. Das Interface ermöglicht einen leichten Einstieg in die digitale Textannotation und -analyse, ohne dass umfangreiche technische Kenntnisse vonnöten wären, und ohne dass die Nutzer*innen mit zu vielen (Experten-)Funktionen gleichzeitig konfrontiert würden. Das gesamte Repertoire an Funktionen (wie beispielsweise kollaborative Annotation oder automatische Annotation von Textkorpora) kann dann von erfahreneren Nutzer*innen bei Bedarf genutzt werden.

CATMA unterstützt

- individuelle wie kollaborative Annotation und Analyse – Texte können privat, aber auch im Team erforscht werden;
- explorative, non-deterministische Praktiken der Textannotation – CATMA liegt ein diskursiver, diskussionsorientierter Ansatz zur Textannotation zugrunde, der auf die Forschungspraktik hermeneutischer Disziplinen zugeschnitten ist;
- die nahtlose Verknüpfung von Textannotation und -analyse in einer webbasierten Arbeitsumgebung – Analyse und Interpretation gehen nach dem Prinzip des 'hermeneutischen Zirkels' in CATMA damit Hand in Hand.

Von linguistischen Textanalysetools unterscheidet sich CATMA insbesondere durch seinen „undogmatischen" Ansatz: Das System schreibt mit seiner hermeneutischen Annotation (vgl. Piez 2010) weder definierte Annotationsschemata oder -regeln vor, noch erzwingt es die Verwendung von starren Ja-/Nein- oder Richtig-/Falsch-Taxonomien. Wenn eine Textstelle mehrere Interpretationen zulässt (wie es in literarischen Texten häufig der Fall ist), ist es in CATMA daher möglich, mehrere und sogar widersprechende Annotationen zu vergeben und so der Bedeutungsvielfalt der Texte Rechnung zu tragen. Mit dem sog. Query-Builder lassen sich außerdem Schritt für Schritt Textanalysen durchführen. Die Ergebnisse der Analyse können schließlich in verschiedenen Varianten visualisiert und für die literaturwissenschaftliche Interpretation und Argumentation genutzt werden.

Zudem bietet CATMA auch die Möglichkeit, bereits annotierte Texte zu verarbeiten (z.B. durch den Upload von XML-Dateien) und die in anderen Tools erstellten Annotationen anzuzeigen, mit zu analysieren und damit wissenschaftlich nachzunutzen. Außerdem lassen sich in CATMA auch automatische (z.B. POS für deutschsprachige Texte) und halb-automatische Annotationen generieren.

Manuelles und kollaboratives Annotieren

Die Annotation von Texten gehört seit Jahrhunderten zu den textwissenschaftlichen Kernpraktiken (vgl. Moulin 2010). Genauer lassen sich hier Freitextkommentare, taxonomiebasierte Annotation und Textauszeichnung unterscheiden, wobei die Übergänge häufig fließend sind (vgl. Jacke 2018, § 9). Während CATMA 6.0 auch eine Funktion für Freitextkommentare bietet, ist die taxonomiebasierte Annotation das eigentliche Kerngeschäft des Tools – wobei die Taxonomie prinzipiell undogmatisch erstellt werden kann und die Form von sog. Tagsets annimmt, denen für kollaborative Annotationsprojekte wahlweise eine Annotations-Guideline beigegeben werden kann (vgl. auch Bögel et al).

Im Workshop werden wir den Unterschied von *Document* (der eigentliche Text), *Tagset* (die aus *Tags* – d.h. aus einzelnen Beschreibungsbegriffen – gebildete Taxonomie, mit der Texte annotiert werden) und *Annotation Collection* (die nutzerspezifische Sammlung individueller Annotationen zu einem *Document* oder einem Korpus) kennenlernen. Diese Dreigliederung ist spezifisch für CATMA und bietet eine Reihe von Vorteilen:

- Taxonomien können projektübergreifend und unabhängig von Texten und Annotationen wiederverwendet werden;
- Annotationen können als *Collections* nach unterschiedlichen inhaltlichen (z. B. nach Forschungsaspekten) oder auch organisatorischen

Gesichtspunkten (z. B. nach Projektmitgliedern) gruppiert und wiederverwendet bzw. erweitert werden;
- benutzerspezifische Annotationen werden als sog. *Stand-off Markup* gespeichert und können damit wahlweise angezeigt oder ausgeblendet werden. Der eigentliche Text wird hierbei nicht verändert. Arbeitet eine Gruppe von Annotator*innen mit der gleichen Taxonomie an einem Text, lassen sich Übereinstimmungen und Widersprüche direkt und einfach erkennen (vgl. Gius und Jacke 2017), um auf interessante oder problematische Textstellen aufmerksam zu werden und die 'Arbeit am Text' zugleich kritisch zu reflektieren.

Analyse und Visualisierung

Neben der Annotation sind die Analyse und Visualisierung der Text- und Annotationsdaten das andere wichtige Standbein von CATMA. Hier wird *distant reading* mit *close reading* zusammengebracht, denn die zuvor manuell erstellten qualitativen Annotationen werden nun in ihrer Quantität und Verteilung hinterfragt. Dies geschieht in Zusammenhang mit „klassischen" DH-Textanalysemethoden wie dem Erstellen einer Wortfrequenzliste, der Analyse von Keywords in Context (KWIC und *DoubleTree*) oder der Distribution ausgewählter Wörter (oder eben Annotationen) im Text oder in der Textsammlung.

Neben diesen grundlegenden Funktionen, die alle per Klick ausgeführt werden können, bietet CATMA den sog. *Query Builder*, in dem komplexere Abfragen einfach per Mausklick erzeugt werden können, ohne dass tiefergehende Kenntnisse einer Abfragesprache (sog. *Query Language*) verlangt werden. Im Workshop werden wir uns dabei nicht nur den Analysefunktionen widmen, sondern auch die unterschiedlichen Visualisierungsmöglichkeiten zu den einzelnen Abfragen anschauen und hinterfragen.

Im Analysebereich können außerdem halbautomatische Annotationen erstellt werden, d.h. man annotiert wiederkehrende Wörter oder Wortgruppen auf einmal mit einem bestimmten Tag, statt dies manuell und wiederholt im Annotationsmodul zu tun.

Der Wechsel zwischen der Arbeit im Annotations- und Analyse- und Visualisierungs-Modul ist ein iterativer Prozess, der die klassisch-zirkuläre hermeneutische Interpretationsarbeit in der Literaturwissenschaft widerspiegelt (vgl. Gius, in Vorbereitung).

Ablauf

Im Workshop werden wir uns in einer Mischung aus Präsentations- und Hands-on-Phasen der textanalytischen Arbeit in CATMA 6.0 nähern. Nach einer generellen Einführung in das Tool werden die Teilnehmer*innen anhand eines vorgegebenen Beispieltextes den gesamten Workflow von der individuellen taxonomiebasierten Textannotation, über die Analyse hin zur Visualisierung und Interpretation der Text- und Annotationsdaten kennenlernen und praktisch erproben können.

Lernziele

Die Teilnehmer*innen sollen ausgehend vom digitalen Text in die Lage versetzt werden, Annotationen manuell und automatisch unterstützt zu erstellen und in Annotation Collections zu speichern, Tagsets/Taxonomien zu entwickeln und den Text alleine und in Kombination mit den Annotationen zu analysieren und zu visualisieren. Für Diskussionen und individuelle Rückfragen (theoretischer, praktischer und technischer Art) auf jedem Niveau und in Bezug auf die Projekte der Teilnehmer*innen wird ausreichend Möglichkeit bestehen.

Zeitplan

Im Workshop werden wir den Arbeitsablauf der digitalen Texterforschung praktisch kennenlernen:

- analytische Textexploration (ca. 30 Minuten)
- manuelle und automatische Annotation und Spezifikation von Annotationskategorien (ca. 40 Minuten)
- kombinierte Abfragen von Annotations- und Textdaten (ca. 30 Minuten)
- visuelle Darstellungsmöglichkeiten von Abfrageergebnissen (ca. 20 Minuten)

Beitragende (Kontaktdaten und Forschungsinteressen)

Dr. Jan Horstmann
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Jan Horstmann ist Postdoc und koordiniert das DFG-Projekt forTEXT, in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die klassischeren Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Als Literaturwissenschaftler interessiert er sich vor allem für die neuen Perspektiven und Erkenntnispotentiale, die DH-Methoden auf literarische Artefakte bereithalten können, und forscht in diesem Sinne unter anderem zu Entsagung und Ironie bei Goethe.

Prof. Dr. Jan Christoph Meister
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Jan Christoph Meister ist Professor für Digital Humanities mit dem Schwerpunkt Literaturwissenschaft. Als ursprünglicher Erfinder von CATMA hat er etliche Forschungsprojekte zur Annotation und Visualisierung textueller Daten und der Entwicklung und Verbesserung von DH-Tools geleitet.

Marco Petris, Dipl. Inform.
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Marco Petris ist Informatiker mit starker Affinität zu geisteswissenschaftlichen Fragestellungen. Er ist von Anfang an an der Entwicklung von CATMA beteiligt und beschäftigt sich mit allen Aspekten der DH-Toolentwicklung, des Tool-Designs und der Implementierung.

Mareike Schumacher, M.A.

Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Mareike Schumacher promoviert als digitale Literaturwissenschaftlerin über Orte und narratologische Ortskategorien in literarischen Texten, beschäftigt sich besonders mit den Methoden des distant reading (u.a. Named Entity Recognition oder Stilometrie) und ist im forTEXT-Projekt u.a. für die Dissemination in den (sozialen) Medien zuständig.

Zahl der möglichen Teilnehmer*innen

Bis zu 30 Personen.

Benötigte technische Ausstattung

Teilnehmer*innen bringen ihren eigenen Laptop mit, der mit dem Internet verbunden ist (Achtung: Touch-Devices werden derzeit noch nicht unterstützt). Am Workshop können bis zu 30 Personen teilnehmen. Neben einer stabilen Internetverbindung werden ein Beamer und eine Leinwand benötigt.

Bibliographie

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (2015): *„Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative"*, in: DHCommons Journal 1.

Gius, Evelyn (in Vorbereitung): *„Digitale Hermeneutik: Computergestütztes close reading als literaturwissenschaftliches Forschungsparadigma?"* in: **Fotis Jannidis (Hrsg.)**: *Digitale Literaturwissenschaft*. DFG-Symposium 9.–13.10.2018.

Gius, Evelyn / Jacke, Janina (2017): *„The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis"*, in: International Journal of Humanities and Arts Computing 11 (2), 233–254.

Jacke, Janina (2018): *„Manuelle Annotation"*, in: forTEXT. Literatur digital erforschen. http://fortext.net/routinen/methoden/manuelle-annotation (Zugriff: 24. September 2018).

Moulin, Claudine (2010): *„Am Rande der Blätter. Gebrauchsspuren, Glossen und Annotationen in Handschriften und Büchern aus kulturhistorischer Perspektive"*, in: Autorenbibliotheken, Quarto. Zeitschrift des Schweizerischen Literaturarchivs 30/31, 19–26.

Piez, Wendell (2010): *„Towards Hermeneutic Markup. An Architectural Outline"*, in: Digital Humanities Conference 2010, London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html (Zugriff: 24. September 2018).

Texterkennung mit Ocropy – Vom Bild zum Text

Nasarek, Robert
robert.nasarek@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Müller, Andreas
andreas.mueller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Das OCR-Programm ocropy

Die optische Zeichenerkennung (engl. Optical Character Recognition – OCR) von historischen Texten weißt oftmals niedrige Erkennungsraten auf. Mit einem gekonnten Preprozessing und ocropy (auch ocropus), einem modular aufgebauten Kommandozeilenprogramm auf Basis eines neuronalen long short-term memory Netzes, ist es möglich, deutlich bessere Ergebnisse zu erzielen. (Springmann 2015, S. 3; Vanderkam 2015) Ocropy ist in Python geschrieben und enthält u. a. Module zur Binarisierung (Erzeugung einer Rastergrafik), zur Segmentierung (Dokumentaufspaltung in Zeilen), zur Korrektur fehlerhafter Erkennungstexte, zum Training neuer Zeichen und natürlich zur Erkennung von Dokumenten (siehe Abbildung 1). Ein bedeutender Vorteil dabei ist, dass jedes Modul eine Reihe von nachvollziehbaren Einstellungsmöglichkeiten hat, um auf die individuellen Herausforderungen jedes Dokumentes einzugehen. Zusätzlich besteht die Möglichkeit ocropy auf die Erkennung einer bestimmten Schriftart, bzw. eines Zeichensatzes zu trainieren.

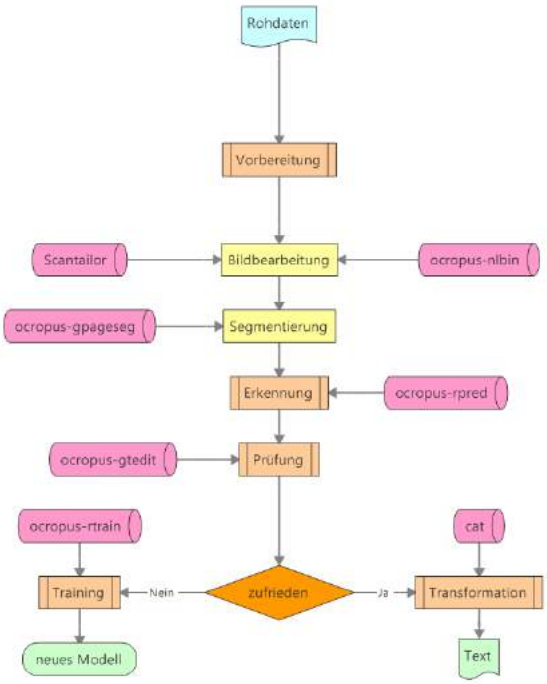


Abbildung 1. Überblick zum Prozessablauf der Texterkennung mit den grundlegenden Software-Modulen

Die Benutzung von ocropy als Kommandozeilenprogramms setzt jedoch den Umgang mit einer Consolen-Umgebung und eine grundlegende Kenntnis von Bash-Kommandos voraus: Für viele potenzielle NutzerInnen stellt dies eine erste Einstiegshürde dar, denn der NutzerInnenanteil von Linuxderivaten beträgt nur 3% (statista 2018), wobei die Gruppe an ShelluserInnen noch kleiner sein dürfte. Im Workshop wird diese Hürde abgebaut, indem alle Schritte „from zero to recognised textfile“ nachvollziehbar und zum Mitmachen aufzeigt wird. Insgesamt werden sechs Themengebiete behandelt, damit die TeilnehmerInnen des Workshops alle benötigten Informationen erhalten, um selbstständig Frakturschriften (oder andere Schriftarten) durch ocropy erkennen zu lassen.

Ubuntu in der VirtualBox

Für bisher ausschließliche NutzerInnen des Betriebssystem Windows oder Mac OS, ist es unverhältnismäßig, allein wegen ocropy Linux als Zweit- oder sogar Hauptsystem zu installieren. Durch die Verwendung einer VirtualBox und des Linux-Derivats Ubuntu kann dieser Schritt umgangen werden. Mit Hilfe einer virtuellen Maschine lässt sich ein Betriebssystem innerhalb eines anderen Betriebssystems emulieren. Das bringt den Vorteil mit sich, keine größeren Änderungen am System vornehmen zu müssen und die Software in einem geschützten virtuellen Rahmen testen zu können. Das Einrichten einer virtuellen Maschine ist daher für die meisten NutzerInnen das Fundament (und vielleicht auch der Einstieg) in Unix-basierte Entwicklerumgebungen. Dabei sind diverse kleinere Einstellungen zu beachten, vom Einschalten der Virtualisierung im BIOS bis hin zur Installation von gemeinsam genutzten Ordnern zwischen Host und Gast. Ubuntu als „Einstiegslinux“ eignet sich hervorragend für die ersten Schritte, da es eine hohe Benutzerfreundlichkeit aufweist und trotzdem alle wichtigen Features mitbringt, die benötigt werden.

Repositorien für brauchbare Digitalisate

OCR-Software erzielt bessere Ergebnisse mit hochauflösenden und fehlerfreien Digitalisaten. Bilddateien sollten mindestens eine Auflösung von 300 DPI besitzen und nicht bereits durch Programme oder Algorithmen bearbeitet worden sein. Google Books referenziert zwar auf viele gescannte Werke, diese liegen aber meist in schlechter Qualität und verlustreich binarisiert vor. Bekannte Repositorien wie das *zentrale Verzeichnis digitalisierter Drucke* oder das *Münchener DigitalisierungsZentrum* bieten exzellente Anlaufstellen zur Beschaffung digitalisierter Drucke; aber auch Sammlungen wie das *Verzeichnis der im deutschen Sprachbereich erschienen Drucke des 16. - 19. Jahrhunderts* der Universität- und Landesbibliothek Sachsen-Anhalt verfügen über frei zugängliche Digitalisate mit einer Auflösung bis zu 600 DPI.

Installation von ocropy

Ocropy ist nicht in den nativen Quellen von den bekanntesten Linux-Derivaten enthalten, sondern muss von Github heruntergeladen und über ein Script installiert werden. Dabei ist die Version der Programmiersprache Python 2.7 zu beachten und die Abhängigkeiten einiger benötigter Module. Im Workshop wird die Installation begleitet und ein bereits auf Drucke des 18. Jahrhundert trainiertes Erkennungsmodul zur Verfügung gestellt.

Preprocessing mit ScanTailor

Eine Texterkennung ist nur so gut wie das Preprocessing des Digitalisates. Bilder, Initiale oder Flecken im Bild stören die Texterkennung und müssen entfernt werden. Darüber hinaus benötigt ocropy binarisierte (schwarz/weiß gerasterte) oder normalisierte Graustufenbilder zur Verarbeitung. Obwohl ocropy mit dem Modul ocropus-nlbin eine eigene Lösung zur Binarisierung von Bilddateien anbietet, hilft dies nicht in Bezug auf Nicht-Text-Elemente, wie Bilder oder schräge Spaltenlinien. Bearbeitungssoftware wie Gimp beinhaltet zwar alle benötigten Funktionen, ist jedoch in Bezug auf die serielle Verwendung bei Textdigitalisaten ineffizent. Im Workshop wird die Software ScanTailor als passgenaues Preprocessing-Tool zur Vorbereitung der Digitalisate favorisiert. ScanTailor ist wie dafür gemacht gescannte Texte in eine einheitliche Form zu bringen und beinhaltet (zum Teil vollständig automatisierte) Funktionen wie

- der Aufsplittung von Spalten oder Seiten
- das Ausrichten der Seite
- des Auswählens des Inhalts

- der Möglichkeit Bereich zu füllen
- der Entzerrung gekrümmter Seiten und
- der Anpassung des Schwellwertes (threshold) bei der Binarisierung.

Außerdem werden Hinweise zu den grundlegenden Eigenschaften eines guten Eingangsbildes gegeben, z. B. in Bezug auf Schwellwert oder DPI-Zahl.

Entwicklung einer Pipeline zur Texterkennung

Die ocropy-Module funktionieren am effizientesten innerhalb einer Pipeline. Ausgehend von der Konvertierung unpassender Dateiformate der Roh-Digitalisate bis hin zur Erstellung einer Korrektur-HTML für die Verbesserung der falsch erkannten Zeichen bietet die Linux-Shell zusammen mit ocropy und dem Programm ImageMagick alle benötigten Werkzeuge. So lassen sich auch große Mengen an Bilddateien stapelweise verarbeiten. In einem Script werden Befehle zur Bildkonvertierung, Zeilenauftrennung, Texterkennung und Textkonvertierung in Reihe geschaltet, um eine stapelhafte Verarbeitung zu ermöglichen. Der Workshop bietet zwei vorgefertigte Scripte zum Gebrauch an und erklärt ihren Ablauf, um eventuelle Anpassungen an die eigenen Bedürfnisse vornehmen zu können.

Training unbekannter Schriftarten

Die eigentliche Stärke von ocropy ist die Möglichkeit Erkennungsmodule für Schriftarten zu trainieren. Die dazu bereitgestellte Ground Truth Data bestimmt maßgeblich die Leistungsfähigkeit der Erkennungsmodule. Dabei stellt sich die Frage, wie eine gute Ground Truth im wörtlichen Sinne auszusehen hat? Wie „schmutzig" dürfen die Daten sein? Sind abgeschnittene Serifen, fehlende Bögen oder i-Punkte ein Problem? Welche Zeichen sollten verwendet werden, um Abbreviationen oder Abkürzungszeichen zu kodieren? Darüber hinaus trainiert ocropy sich nicht permanent besser, sondern baut das neurale Netz zeitweise mit negativen Auswirkungen für die Erkennungsraten um (siehe Abbildung 2). Im Workshop wird ein Script zur Identifikation des besten Trainingsmoduls vorgestellt, um das Beste aus ocropy herauszuholen.

Abbildung 2. Trainingsprozess von ocropus-rtrain mit Ground Truth von Zedlers Universallexikon. training = Ground Truth anhand derer das Modul trainiert wurde, testing = unbekanne Ground Truth zum Test der Performance.

Ablauf

Der Workshop richtet sich vorrangig an Anfänger und leicht fortgeschrittene NutzerInnen im Umgang mit Linux und der Console. Es werden keine Vorkenntnisse in Bash oder Python benötigt und alle im Kurs vorgestellte Software, Scripte und Daten stehen frei zur Verfügung. Der Workshop möchte alle an Interessierten da abholen, wo sie stehen und versucht durch ein schrittweises Vorgehen an die Vorzüge der Consolen-Benutzung und kommandozeilenbasierte Software heranzuführen. Teilnehmer sollten ihr eigenes Notebook mitbringen, auf dem sie auch Administrator-Rechte besitzen. Des Weiteren wird ein Internetzugang benötigt, um fehlende Software oder Abhängigkeiten herunterladen zu können. Größere Softwarepakete (VirtualBox, Ubuntu) werden auch auf USB-Sticks zur Verfügung gestellt, sollten aber nach Möglichkeit vorher selbstständig heruntergeladen werden. Es können je nach Erfahrungsstand der TeilnehmerInnen mit Console und Linux 20 bis 25 Personen betreut werden. Der Workshop dauert drei bis vier Stunden.

Bibliographie

ImageMagick (2018): *Convert, Edit, Or Compose Bitmap Images @ ImageMagick*, URL: https://www.imagemagick.org/ , [zuletzt besucht am 14.10.2018].

MDZ (2018): *Münchner DigitalisierungsZentrum*, Bayerische Staatsbibliothek, München, URL: https://www.digitale-sammlungen.de/ , [zuletzt besucht am 12.10.2018].

ocropy (2018): *Python-based tools for document analysis and OCR*, URL: https://github.com/tmbdev/ocropy , [zuletzt besucht am 14.10.2018].

ScanTailor (2018): *ScanTailor*, http://scantailor.org/ , [zuletzt besucht am 14.10.2018].

Springman, Uwe (2015): *Ocrosis. A high accuracy OCR method to convert early printings into digital text, Center for Information and Language Processing (CIS)*, Ludwig-Maximilians-University, Munich, URL: http://cistern.cis.lmu.de/ocrocis/tutorial.pdfm [zuletzt besucht am 14.10.2018].

statista, Marktanteile der führenden Betriebssysteme in Deutschland von Januar 2009 bis Juli 2018, URL: https://de.statista.com/statistik/daten/studie/158102/umfrage/marktanteile-von-betriebssystemen-in-deutschland-seit-2009/, [zuletzt besucht am 10.10.2018].

Vanderkam, Dan (2015): *Extracting text from an image using Ocropus*, URL: http://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html, [zuletzt besucht am 10.10.2018].

VD (2018): *Digitale Sammlungen des 16. bis 19. Jahrhunderts*, Universitäts- und Landesbibliothek Sachsen-Anhalt, Halle (Saale), URL: http://digitale.bibliothek.uni-halle.de/ , [zuletzt besucht am 14.10.2018].

VirtualBox (2018): *Oracle VM VirtualBox*, URL: https://www.virtualbox.org/ , [zuletzt besucht am 14.10.2018].

Ubuntu (2018): *The leading operating system for PCs, IoT devices, servers and the cloud | Ubuntu*, URL: https://www.ubuntu.com/ , [zuletzt besucht am 14.10.2018].

ZVDD (2018): *Zentrales Verzeichnis Digitalisierter Drucke*, Georg August Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Göttingen, URL: http://www.zvdd.de/ , [zuletzt besucht am 12.10.2018].

Text Mining mit Open Semantic (Desktop) Search – eine digitale Such- und Annotationsumgebung für informationsgetriebene Fragestellungen in den Geisteswissenschaften.

Wettlaufer, Jörg
jwettla@gwdg.de
Georg-August-Universität Göttingen, Deutschland

Ziehe, Stefan
stefan.ziehe@stud.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Mandalka, Markus
info@opensemanticsearch.org
Freier Journalist und Informatiker, Berlin

Einleitung

In den Geisteswissenschaften sind wir bei vielen Fragestellungen auf die Sammlung und das Zusammenstellen von (teilweise sehr großen) Informationsmengen angewiesen, aus denen relevante Details herausgefiltert und anschließend in neue Zusammenhänge gesetzt werden müssen (Wettlaufer 2016). Dieser informationsgetriebene Ansatz geisteswissenschaftlichen Arbeitens ist nicht weit von den Aufgaben entfernt, die Journalisten heute für ihre Arbeit im digitalen Zeitalter lösen müssen (z.B. die Auswertung der sog. „Panama Papers“). In diesem Kontext wurde „Open Semantic Desktop Search“ (OSDS) von Markus Mandalka (Berlin) entwickelt, ein Recherche Tool bzw. eine Enterprise Suchmaschine für die semantische und explorative Suche in großen Dokumentenmengen, das auch Textanalyse und Text Mining mit offenen Standards für Linked Open Data bietet: http://www.opensemanticsearch.org . In diesem Workshop werden einige Funktionen und Anwendungsbeispiele des Softwarepakets in den digitalen Geisteswissenschaften vorgestellt sowie Hinweise zur Installation und Verwendung des Tools gegeben. Ziel des Workshops ist es die TeilnehmerInnen in die Lage zu versetzen, OSDS eigenständig für die Beantwortung wissenschaftlicher Fragestellungen, z.B. aus dem Bereich der digitalen Geschichtswissenschaften, zu verwenden und für ihre jeweiligen Bedürfnisse zu konfigurieren bzw. zu erweitern.

Herausforderungen

Korpora in der informationsgetriebenen geisteswissenschaftlichen Forschung sind oft heterogen und schlecht für Werkzeuge des Natural Language Processing (NLP) aufbereitet. Sie stammen z.B. direkt aus google books oder anderen digitalen Bibliotheken, die manchmal entweder nur eine Bilddatei oder aber einen Textlayer mit einem OCR erkannten, nicht weiter strukturierten Text enthalten. Im Fall von historischen Texten oder Textsammlungen kommen auch unterschiedliche Sprachstufen, Grammatiken etc. hinzu. Für Forschende, die eine geringe Erfahrung mit NLP und den verfügbaren Werkzeugen haben (z.B. Weblicht[1], Standford NLP[2] etc.), übersteigt der Aufwand für die Aufbereitung der Texte für eine korpuslinguistische Analyse oft die zur Verfügung stehende Zeit, zumal die Fragestellungen eher semantischer als sprachhistorischer Art sind. Wie kann man trotzdem diese Texte linguistisch „bewusst“ nutzen und erschließen?

Ein weiterer Aspekt ist das Bedürfnis vieler WissenschaftlerInnen, eine digitale Bibliothek mit fortgeschrittenen Retrieval-Funktionen für das eigene projektspezifische Textkorpus zur Verfügung zu haben. Oft verfügen sie aber selber nicht über die notwendigen Kenntnisse, um eine solche digitale Bibliothek mit einer leistungsfähigen Suchumgebung mit Keyword-in-Context (KWIC) Funktion zu erstellen. Die üblicherweise dafür zur Verfügung stehenden Technologien (Apache SOLR oder ElasticSearch) bedürfen einer gewissen Einarbeitung, einer projektspezifischen Konfiguration und einer Serverumgebung, um die gewünschten Funktionen zur Verfügung zu stellen. Einen Weg aus dieser Situation bieten z.B. Referenzmanager, die heute teilweise Volltextsuche in Dokumenten (zumeist PDF) als Zusatzleistung mit anbieten (Zotero, Mendeley, Citavi). Hier können größere Datenmengen jedoch oft nur über kostenpflichtige Speichererweiterungen realisiert werden und eine Anbindung an eine Referenz bzw. Metadaten ist für eine zu indizierende Datei obligatorisch. Dies kann bei einer großen Sammlung von Texten bzw. einem umfassenden Korpus, für das keine (geeigneten)

Metadaten zur Verfügung stehen, ein großes Problem sein. Eine weitere Erschließung der Texte mittels einer Named Entity Recognition (NER), eine facettierte Suche oder die Visualisierung der gefundenen Zusammenhänge sind dabei bislang ebenfalls nicht möglich.[3]

Open Semantic Desktop Search als integrierte Text-Mining und Retrieval-Umgebung

OSDS ist ein freies Softwarebundle, dass aus Open Source Bestandteilen zusammengestellt wurde und auf dieser Grundlage als Donationware weiterentwickelt wird. Da es sich um ein Softwarepaket handelt, das üblicherweise in Serverumgebungen läuft, wird OSDS lauffertig als virtuelle Maschine (VM) angeboten, die mit Virtual Box von Oracle[4] betrieben werden kann. Die VMs können dabei als Appliance in den folgenden drei Varianten verwendet bzw. eingelesen werden: 1. Open Semantic Desktop Search (OSDS); 2. Open Semantic Search (OSS); 3. Open Semantic Search Server (OSSS).[5]

Paket 1 ist eine VM Appliance, die man mit Virtual Box laden und lokal auf einem Rechner betreiben kann. Die Appliance wird in einer aktuellen Version (Juli/August 2018) in zwei Sprachvarianten angeboten: einmal mit englischen und einmal mit deutschen Keyboard Settings. Die zweite Variante ist ebenfalls eine Appliance, die unter Oracle Virtual Box läuft, aber nur einen Server als „localhost" bereitstellt. Dort fehlt der Gnome Desktop im Debian Linux, auf das alle Distributionen aufsetzen. Die OSDS Version schlägt mit 3GB zu Buche, die Servervariante OSS mit (nur) 1.8 GB. Das dritte Paket (OSSS) ist mit etwa 300 MB am Leichtgewichtigsten, aber erwartet natürlich auch eine (manuelle) Installation und vor allem Konfiguration auf einem Debian oder Ubuntu basierten Server. Vor allem OSDS und OSS benötigen eine moderne Hardware und ausreichend Arbeitsspeicher, um die Indexierung der Texte und die NER Pipeline schnell zu erledigen. Mit 8GB Ram und mind. einem Zweikernprozessor sollten diese Voraussetzungen aber bei den meisten TeilnehmerInnen des Workshops gegeben sein. Im Fall von diesbezüglichen Problemen ist geplant, in Zweiergruppen zu arbeiten.

Kernstück der Enterprise Suchmaschine ist ein Lucene SOLR Indexer, mit dem aufgrund der Einbindung von Apache Tika[6] recht beliebige Dokumente indexiert werden können. Die enthaltenen Informationen werden damit als Keyword im Kontext find- und referenzierbar. In dem Paket ist auch ein sogenanntes Open Semantic ETL (Extract-Transform-Load) Framework integriert, in dem die Texte für die Extraktion, Integration, die Analyse und die Anreicherung vorbereitet werden.[7] Es handelt sich um eine Pipeline, die einem sehr viel Arbeit hinsichtlich der Bereitstellung der Texte für den Indexer abnimmt. OSDS übernimmt automatisiert die Aufbereitung der Texte und kümmert sich nach dem Prinzip eines überwachten Ordners mit einer NLP Pipeline um sämtliche Schritte, die von der Extraktion über die linguistische Analyse bis zur Anreicherung mit weiteren Metadaten notwendig sind. Schließlich stellt OSDS auch einen Webservice (Rest-API) für die maschinelle Kommunikation sowie ein durchdachtes User Interface (Django[8]) zur Verfügung, mit dem die Suchmaschine bedient, konfiguriert und natürlich auch die Daten durchsucht werden können. Die facettierte Suche spielt dabei eine besondere Rolle, da die Facetten mehr oder weniger automatisch aus der linguistischen Analyse der Texte und auf der Grundlage von (konfigurierbaren) Namen Entities (Personen, Orte, Organisationen etc.) gebildet werden. Entsprechend sind auch die Hauptfunktionen des Softwarepakets angelegt: Suchinterface, ein Thesaurus für Named Entities (auch über SPARQL z.B. aus wikidata integrierbar), Extraktion von Entitäten in neu zugefügten Texten, Laden von Ontologien (z.B. in SKOS), eine listenbasierte Suche sowie eine Indexfunktion, die den Aufbau eines neuen Suchindex anstößt. Datenquellen können entweder lokale Ordner oder Ressourcen im Internet (Webseiten, Datenbanken etc.) sein. Diese können in konfigurierbaren Zeitabständen überprüft und neue Informationen dem Index hinzugefügt werden.

Lernziele und Ablauf

Durch den Workshop sollen die TeilnehmerInnen in die Lage versetzt werden, die OSDS Such- und Text Mining Funktionen in einer lokalen Umgebung in Betrieb zu nehmen, zu konfigurieren, Daten in den Suchindex einzulesen und Suchanfragen auszuführen. Dabei soll die Verwendung des NER-Taggers und die Nutzung der Thesaurus-Funktion für die Bereitstellung eigener Suchfacetten erklärt und die Benutzung anhand praktischer Beispiele eingeübt werden. Nach der Erprobung verschiedener Suchanfragen zum Beispielkorpus werden die erweiterten Funktionalitäten von OSDS vorgestellt und getestet. Eine wichtige zusätzliche Funktionalität betrifft die integrierte OCR mit tesseract,[9] die auch Frakturschriften erkennen kann. Desweiteren gibt es eine Schnittstelle der NER zur neo4j Graphdatenbank,[10] die eine Darstellung der Entitäten und ihrer Verknüpfungen als Graphen ermöglicht. Weitere Auswertungsmöglichkeiten eröffnen die Visualisierungsoptionen, die sowohl Ortsnamen auf einer Karte als auch Netzwerke zwischen Personen anzeigen können.

Der Ablauf ist wie folgt vorgesehen:
Block I: 90 Minuten Workshop

- 15 Minuten Begrüßung / Einführung
- 30 Minuten Hands On - Installation von Virtual Box / OSDS auf eigenen Geräten, sofern dies noch nicht vorher durchgeführt wurde. Behebung von Problemen.
- 30 Minuten Lecture zur Funktionalität von OSDS / Parallel: Indexierung des Beispielkorpus
- 15 Minuten Hands On / Ausprobieren des UI / Suchinterface anhand einfacher Abfragen mit boolschen Operatoren, Einsatz der Suchfacetten

30 Minuten Pause
Block II. 90 Minuten Workshop

- 30 Minuten Vorstellung der erweiterten Funktionen von OSDS (OCR, Semantic Tagging, NER, Suche mit Listen, neo4j, Visualisierungen)
- 30 Minuten Hands On: Bearbeitung verschiedener Aufgabenstellungen zu diesen Funktionen (Semantic Tagging, Einbindung von wikdata über SPARQL in die NER)
- 30 Minuten gemeinsame Diskussion der Erfahrungen mit OSDS, Ausblick des Entwicklers auf weitere

Funktionalitäten und zukünftige Entwicklungen sowie Fragen.

Beitragende (Kontaktadressen und Forschungsinteressen)

Der Workshop wird angeboten von Mitarbeitern des Göttingen Centre for Digital Humanities an der Georg-August Universität Göttingen in Zusammenarbeit mit dem Entwickler von OSDS. Das Zentrum hat u.a. einen Forschungsschwerpunkt im Bereich des Text Mining und der digitalen Geschichtswissenschaft. Der Workshop wird ohne besondere Voraussetzungen angeboten (man beachte aber die Anforderungen an die Hardware) und steht TeilnehmerInnen ohne spezielle Programmier- oder Softwarekenntnisse offen. In den Hands-On Teilen wird allerdings individuelle Unterstützung bei der Installation und Einrichtung der notwendigen Virtualisierungsumgebung sowie des Enterprise Suchsystems gegeben.

Jörg Wettlaufer
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen

Die vielfältigen Forschungsinteressen von Jörg Wettlaufer liegen u.a. im Bereich Information Retrieval, Semantic Web Anwendungen in den Geisteswissenschaften und Digitale Geschichtswissenschaft. Er tritt für einen intensiven Dialog zwischen Fachwissenschaften und den Digital Humanities ein und ist seit 2018 zusätzlich im Bereich der Digitalen Lehre in den Geisteswissenschaften tätig.

Stefan Ziehe
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen

Stefan Ziehe ist Masterstudent am Institut für Informatik und zugleich wissenschaftliche Hilfskraft am GCDH und interessiert sich insbesondere für Deep Learning-Verfahren im Bereich der Sentiment Analysis und des NLP.

Markus Mandalka
Warschauerstr. 66
10243 Berlin

Die Expertise von Markus Mandalka liegt in der Schnittstelle zwischen Informatik und Journalismus. Er ist freier Journalist sowie Informatiker und Entwickler des Open Semantic Desktop Search Systems. In diesem Zusammenhang bietet er auch Beratungsservice für Institutionen und Projekte an.

Zahl der möglichen Teilnehmer: 5-25

Benötigte technische Ausstattung:

Es wird außer einem Beamer und WLAN keine besondere technische Ausstattung benötigt. Es sollte sich allerdings um Räumlichkeiten handeln, die eine aktive Betreuung der TeilnehmerInnen in den Hands-On Phasen erlaubt. Die TeilnehmerInnen müssen zudem geeignete Hardware für die praktischen Übungen selber mitbringen (Notebooks mit Prozessoren mit Virtualisierungsbefehlssätzen und 8 GB RAM).

Fußnoten

1. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.
2. https://stanfordnlp.github.io/CoreNLP/.
3. Vgl. zur semantischen Annotation aber INCEPTION: https://inception-project.github.io// und https://recogito.pelagios.org/
4. https://www.virtualbox.org/.
5. https://www.opensemanticsearch.org/de/download.
6. https://tika.apache.org/.
7. https://www.opensemanticsearch.org/etl.
8. https://www.djangoproject.com/.
9. https://github.com/tesseract-ocr/.
10. https://neo4j.com/.

Bibliographie

Gahlke, Lukas / Mai, Florian / Schelten, Alan / Brunsch, Dennis / Scherp, Ansgar (2017): *Using Titles vs. Full-text as Source for Automated Semantic Document Annotation*, in: Computing Research Repository. https://arxiv.org/abs/1705.05311 [zuletzt abgerufen 12.10.2018].

O'Caroll, Ultan (2016): *Open Semantic Desktop Search – good but…*, https://uoccou.wordpress.com/2016/04/22/open-semantic-desktop-search-good-but/ [zuletzt abgerufen 12.10.2018].

Sporleder, Caroline (2010): *Natural Language Processing for Cultural Heritage Domains*, in: Language and Linguistics Compass, 4:9, 750–768.

Wettlaufer, Jörg (2016): *Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern*, in: Zeitschrift für digitale Geisteswissenschaften. text/html Format. DOI: 10.17175/2016_011 [zuletzt abgerufen 12.10.2018]

Wettlaufer, Jörg (2018): *Hands-On „Open Semantic Desktop Search" (OSDS)*, Einstündiger Workshop auf dem Historikertag Münster am 27.09.18. Vgl. http://digigw.hypotheses.org/2475 und http://digihum.de/blog/2018/09/18/hands-on-open-semantic-desktop-search/ [zuletzt abgerufen 12.10.2018].

Usability-Testing für Softwarewerkzeuge in den Digital Humanities am Beispiel von Bildrepositorien

Dewitz, Leyla
leyla.dewitz@tu-dresden.de
Technische Universität Dresden, Deutschland

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

Niebling, Florian
florian.niebling@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Kennen Sie diese Erfahrung? Bei der Suche in einem Bildrepositorium rücken die gesuchten Inhalte erst einmal in den Hintergrund, da sich die Auseinandersetzung mit den Funktionsweisen der Webanwendung als erste, oftmals zeitraubende Hürde erweist.

Usability-Testing ist eine Methode, um die Gebrauchstauglichkeit von bereits fertigen oder sich in der Entwicklung befindlichen Anwendungen zu messen. Methoden der nutzerzentrierten Evaluierung interaktiver Softwarelösungen können die Umsetzung digitaler Arbeitstechniken in den Geisteswissenschaften wesentlich unterstützen (Bulatovic et al. 2016). Dabei spielen *Usability* (dt.: Gebrauchstauglichkeit)[1] und die *User Experience (UX)* (dt.: Nutzererlebnis)[2] eine wesentliche Rolle bei der Entwicklung von Webanwendungen und Softwaretools und deren Zugänglichkeit. Webanwendungen können durch die Evaluation von Nutzergruppen besser auf deren Bedürfnisse angepasst werden und bei der zielgruppengerechten Optimierung zu einem qualitativen Fortschritt in der Anwendbarkeit digitaler Repositorien verhelfen (Burghardt & Wolff 2014).

Schwerpunkte des Workshops sind:

- Teilnehmer lernen grundlegende UX-Techniken kennen und anzuwenden.
- Diese werden anhand des Prototyps erprobt und vertiefend erlernt.
- Usability-Schwachstellen können durch eine simulierte Usability-Testing Situation durch Teilnehmer erkannt und erlernte Methoden in einen Anwendungskontext gestellt werden.

Zunächst werden die zehn Regeln für Usability nach Steve Krug (2016) und die acht goldenen Regeln des Interface Designs nach Ben Schneidermann (2016) vorgestellt und anhand derer eine Einführung in die Thematik gegeben. Daraufhin sollen die Teilnehmer für den Umgang mit Bildrepositorien am Beispiel der 3D-Webanwendung der Nachwuchsforschungsgruppe HistStadt4D sensibilisiert werden. Der Prototyp der Forschungsgruppe soll in einem nächsten Schritt, von Teilnehmern auf Aspekte der Usability hin geprüft werden. Hier werden die Usability-Methoden wie Paper-Prototyping[3] von den Workshop-Teilnehmern angewendet. Ziel ist die Orientierung der Usability an Forschungsfragen spezifischer Nutzergruppen, ihrer Bedürfnisse und Anforderungen an ein System anhand von zuvor gebildeten Personas[4]. Insbesondere soll dadurch eine nutzerzentrierte Optimierung der Funktionalitäten und der Durchsuch- und Auffindbarkeit von Bildmaterialien im Repositorium erreicht werden. Hier werden Feedbacks für die Nachwuchsforschungsgruppe von den Arbeitsgruppen eingeholt, um die Verbesserung der Webanwendung voranzubringen. Dieser Workshop fokussiert dabei vor allem Bildrepositorien; vermittelte Inhalte lassen sich aber ebenso auf andere Repositorien übertragen. Der Workshop richtet sich vorrangig an Kunst- und Architekturhistoriker, Nutzer und Anbieter von Bild- und Fotoarchiven sowie Digital Humanists im Bereich Bild. Während des gesamten Workshops stehen wissenschaftliche Mitarbeitende des Projekts für Fragen und Auskünfte zur Verfügung.

Problemaufriss und Nutzen

Usability-Tests werden nur selten in der Evaluierung von DH-Tools als Messinstrument für die Nutzerzufriedenheit bzw. Gebrauchstauglichkeit von Systemen eingesetzt. Nach Schreibmann & Hanlon (2010) ergab eine Umfrage unter Entwicklern, dass nur etwa 31% Usability-Tests zur Evaluation von Webtools und Software heranziehen. Zudem werden Usability-Methoden im Entwicklungsprozess von Systemen oft erst spät eingebunden (Kirschbaum 2004). Das führt vor allem zu Hürden in der Etablierung digitaler Methoden in den Geisteswissenschaften. Die Integrierung nutzerzentrierter Evaluation würde diese Einstiegshürde für Geisteswissenschaftler mit geringer digitaler Expertise minimieren. Die Usability von Tools ist somit ein Schlüsselfaktor, um die Akzeptanz digitaler Werkzeuge in der Arbeitspraxis von Forschern zu erhöhen. Die gebrauchstaugliche Gestaltung von Arbeitsabläufen kann effektives Arbeiten erleichtern und Bedienungsfehlern vorbeugen. Folglich sind Usability-Tests heutzutage ein Standardelement in den meisten Software-Entwicklungsprozessen. Eine Anwendung dieser Methoden führt darüber hinaus zu einer nachhaltigen Nutzbarkeit von Ressourcen sowie zur Erschließung neuer bzw. fachfremder Communitys (Bulatovic et al. 2016).

Nutzerzentrierte Gestaltung definiert dabei nur ein Vorgehensmodell, in dem die Bedürfnisse von Nutzern am Anfang stehen. Es muss eruiert werden, welche Methoden für die Evaluation des Tools am geeignetsten erscheinen. Die Usability-Forschung liefert hierfür passende Methoden und Lösungen: Benutzertests mit Prototypen, Personas, Beobachtungen, Befragungen, Fokusgruppen, lautes Denken, Expert-Review oder heuristische Evaluationen (Nielsen 1994, Schneidermann 1998, 1996, Sarodnik & Brau 2006, Tullis & Albert 2013).

Projektskizze HistStadt4D im Bereich Digital Humanities

Digitalisate historischer Fotografien und deren Nutzbarkeit für die geschichtswissenschaftliche Forschung und quellenbasierte Vermittlung stellen ebenso wie spatiale Modelle historischer Objekte Kernthemen der Digital Humanities dar. Anhand stadt- und baugeschichtlicher Forschungsfragen und Vermittlungsanliegen zur Historie der Stadt Dresden adressiert das Projekt die Untersuchung und Entwicklung von methodischen und technologischen Ansätzen, um umfangreiche Repositorien historischer Medien und Kontextinformationen räumlich dreidimensional

sowie zeitlich zusammenzuführen, zu strukturieren und zu annotieren sowie diese für Wissenschaftler und Öffentlichkeit mittels eines 4D-Browsers sowie einer ortsabhängigen Augmented-Reality-Darstellung als Informationsbasis, Forschungswerkzeug und zur Vermittlung geschichtlichen Wissens nutzbar zu machen. Prototypische Datenbasis stellen u.a. digitalisierte historische Fotografien der Deutschen Fotothek dar.

Das fünfzehnköpfige, interdisziplinäre Projektteam befasst sich u.a. mit der Usability der Anwendung, einer didaktischen Aufbereitung der Informationen bzw. Visualisierungen sowie der zielgruppenorientierten Entwicklung von Forschungswerkzeugen und Augmented-Reality-Anwendungen. Die Kollaborationsplattform für Wissenschaftler betrachtet wissenschaftliche Standards genauso wie motivierende Gamification-Ansätze, die im Sinn von Citizen Science einen Anreiz für die breite Öffentlichkeit geben, dass virtuelle Modell zu vervollständigen.

Programmplanung – Ablauf des Workshops

Der Ablauf des Workshops gestaltet sich dabei folgendermaßen: Anfangs sollen Grundlagen geklärt und die Relevanz des Einsatzes von Usability-Methoden für die Digital Humanities insbesondere im Kontext digitaler Bildrepositorien vermittelt werden. Im praktischen Teil des Workshops soll das zu testende Tool vorgestellt und in seine Funktionsweise eingeführt werden. Als Testbeispiel dient der in der Nachwuchsforschungsgruppe entwickelte Prototyp – eine Webanwendung, die in einem modernen, HTML5/ webGL-fähigen Browser läuft. Die Workshop-Teilnehmenden werden mittels Personas mögliche Perspektiven von Nutzern einnehmen. Einen Großteil des Interface nimmt ein 3D-Viewport ein, in dem Nutzer die Möglichkeit haben ein 3D-Stadtmodell zusammen mit verorteten Bildern zu erkunden. Das Interface soll im Rahmen der Paper-Prototyping-Methode von den Workshop-Teilnehmern evaluiert werden. Hierdurch möchten wir Kompetenzen der Teilnehmer fördern, indem wir Testsituationen schaffen, in denen Usability-Probleme gefunden und Ansätze für Lösungen proaktiv eingebracht werden können. Der Input der Teilnehmer ermöglicht es uns, Lösungsansätze, die wir in unserer Nachwuchsforschergruppe verfolgen, kritisch prüfen zu lassen und vor allem diejenigen, die mit solchen Quellenbeständen arbeiten, mit ihren Erfahrungen und Wünschen, zu Wort kommen lassen. Den Workshop-Teilnehmern bieten wir einen zielgerichteten Einblick in Theorien und Methoden der Usability, sowie eine qualitativ hochwertige Diskussion und einen tiefgreifenden Erfahrungsaustausch.

Wie bereits während des im letzten Jahr erfolgreich durchgeführten Workshops „Digitale Bildrepositorien – wirkliche Arbeitserleichterung oder zeitraubend“ der Nachwuchsforschungsgruppe HistStadt4D, werden auch diesmal gezielte, auf den Anwendungsbedarf ausgerichtete Aufgaben, die der tatsächlichen kunsthistorischen Tätigkeit entlehnt sind, an die Teilnehmer gestellt. Diese sollen im besten Falle Usability-Schwachstellen des Tools aufzeigen und somit aktiv in den Workshop und in die Lösung des gestellten Problems involviert werden.

Der Ablauf des Workshops gestaltet sich dabei folgendermaßen:

11:00 – 11:15 Uhr Vorstellung der Nachwuchsforschungsgruppe und Einführung in die Thematik

11:15 – 11:45 Uhr Vorstellung von Usability-Methoden

11:45 – 12:00 Uhr Vorstellung des Prototyps (3D-Webanwendung Stadtmodell Dresden)

12:00 – 13:00 Uhr Mittagspause

13:00 – 14:30 Uhr Usability-Testing anhand von Beispielaufgaben

14:30 – 14:45 Uhr Kaffeepause

14:45 – 15:15 Uhr Auswertung mit gemeinsamer Präsentation der Erfahrungen zur Ergebnissicherung

Ausblick

Der Workshop soll zur Sensibilisierung der DH-Community sowie der von Entwicklern der Bildrepositorien beitragen, indem praxistaugliche und allgemein verwendbare Methoden der Usability- und UX-Methoden vermittelt und erste Fähigkeiten erlernt werden.

Fußnoten

1. *„Usability ist das Ausmaß, in dem ein System durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.“* (ISO 9241-11).
2. *„Wahrnehmungen und Reaktionen einer Person, die aus der tatsächlichen und/oder der erwarteten Benutzung eines Produkts, eines Systems oder einer Dienstleistung resultieren. [...] Dies umfasst alle Emotionen […] und Reaktionen, Verhaltensweisen und Leistungen, die sich vor, während und nach der Nutzung ergeben.“* (DIN ISO 9241-210).
3. *Paper-Prototyping* ist eine interaktive Methode im Evaluationsprozess, der Entwicklern hilft, Informationsarchitekturen und Designs nutzerzentriert auszurichten bzw. anzupassen. Nutzer skizzieren auf Papier grobe, für sie visuell sinnvoll positionierte Funktionalitäten und Elemente einer spezifischen Benutzeroberfläche für die zu evaluierende Webanwendung (Martin & Hanington 2012).
4. *Personas* sind fiktive Personen, die konstruiert werden, um bestimmte Personengruppen und deren Bedürfnisse aufzuzeigen um bspw. eine Webanwendung nutzerzentriert ausrichten und Anforderungen definieren zu können. (Martin & Hanington 2012).

Bibliographie

Bulatovic, N / Gnadt, T / Romanello, M / Stiller, J / Thoden, K (2016): *Usability in Digital Humanities - Evaluating User Interfaces, Infrastructural Components and the Use of Mobile Devices During Research Proces*. In 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5–9, 2016, Proceedings. Research and Advanced Technology for Digital Libraries.

DIN EN ISO 9241-11:1998: *Ergonomic requirements for office work with visual display terminals*. Part 11: Guidance on usability.

DIN EN ISO 9241-210:2010: *Ergonomics of human-system interaction*. Part 210: Dialogue principles.

Kirschenbaum, M G (2004): *So the Colors Cover the Wires': Interface, Aesthetics, and Usability*. In A Companion to Digital Humanities, herausgegeben von Susan Schreibman, Ray Siemens und John Unsworth, 523–42. Blackwell Publishing.

Krug, St (2005): *Don't Make Me Think! A Common Sense Approach to Web Usability*, 2nd Edition. New York: New Riders Press.

Martin, B & Hanington, B (2012): *Universal Methods of Design. 100 Ways to Research Complex, Problems, Develop Innovative Ideas, and Design Effective Solutions.* Beverly: Rockport.

Nielsen, J (1994): *Usability Engineering.* San Francisco: Morgan Kaufmann.

Schneiderman, B (2016): *The Eight Golden Ruels of Interface Design*. URL: https://www.cs.umd.edu/users/ben/goldenrules.html (28.09.2018).

Shneiderman, B (1998): *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Boston: Addison-Wesley.

Shneiderman, B (1996): *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*. In Proceedings of the IEEE Symposium on Visual Languages, pages 336-343, Washington. IEEE Computer Society Press.

Sarodnick, F / Brau, H (2006): *Methoden der Usability Evaluation*. Huber Verlag.

Tullis, T / Albert, B (2013): *Measuring the User Experience. Collecting, Analyzing, and Presenting Usability Metrics*. Burlington/New York: Morgan Kaufmann/Elsevier.

Versionskontrolle mit Git + Kollaboratives Arbeiten im Web mit GitHub

Druskat, Stephan
stephan.druskat@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Rockenberger, Annika
annika.rockenberger@nb.no
Nasjonalbiblioteket, Oslo, Norwegen

Einführung

In den Digital Humanities nehmen Softwareentwicklung und Datenverarbeitung eine zentrale Rolle ein, aber auch über diese Felder hinaus ist kollaboratives, verteiltes, digitales Arbeiten inzwischen ein fester Bestandteil des Forschungsprozesses. Um die Qualität ebenso wie die Optimierung des wissenschaftlichen Erkenntnisprozesses sicherstellen und Nachnutzbarkeit der entstehenden Ressourcen gewährleisten zu können, benötigen Forschende Kenntnisse im Bereich grundlegender digitaler Methoden. Derzeit ist eine Einbettung entsprechender methodischer Ausbildung in die Curricula noch nicht überall implementiert, was zum Teil unzureichende Kenntnisse und daraus folgende Unsicherheit in der Anwendung digitaler Methoden bis über die Erfahrungsstufe der Early Career Researchers hinaus zur Folge hat. Zumindest kurzfristig Abhilfe schaffen können entkoppelte Lehrgänge, die Forschenden grundsätzliche Kenntnisse in digitalen Methoden vermitteln. In diesem Bereich haben sich insbesondere die Lehrpläne der Carpentries (https://carpentries.org/) als forschungsbasierte, ergebnis- und lerner*innenorientierte Mittel zum Enablement von Forschungsgemeinschaften hervorgetan. Diese fokussieren auf kleine Lernschritte, formative Lernzielüberprüfungen, mentale Modelle und kurze Feedbackzyklen, sowie praxisorientierte Übungen. Zusätzlich stellt die auf Freiwilligenarbeit und Inklusion fokussierte Community der Carpentries einen Multiplikator dar, der eine höhere Durchsetzung der Disziplinen mit den notwendigen Grundkenntnissen im Bereich computergestützer Methoden möglich macht. Daher verfolgen wir im Rahmen zweier halbtägiger Workshops zwei Ziele:

1. Vermittlung von zur Einhaltung guter wissenschaftlicher Praxis im Sinne von Reproduzierbarkeit von Forschungsergebnissen und Open Science notwendigen Grundkenntnissen in der Versionsverwaltung mit Git und dem kollaborativen Arbeiten im Web mit GitHub.
2. Einführung in gemeinschaftlich entwickelte, offene Lehrmethoden auf Grundlage der Carpentries, hier von Software Carpentry (Wilson 2006).

Die Workshops sind thematisch eng miteinander verwandt und lassen sich daher gut kombiniert besuchen, fokussieren aber jeweils verschiedene abgeschlossene Aspekte des Bereichs Kollaboration via Versionskontrolle, so dass auch der Besuch eines der beiden Teilworkshops möglich und sinnvoll ist.

Teilworkshop „Versionskontrolle mit Git“

Der halbtägige Workshop „Versionskontrolle mit Git“richtet sich an Neulinge im Bereich Versionskontrolle. Vermittelt werden Grundkenntnisse der Arbeit mit dem Versionskontrollsystem Git (Chacon & Straub 2014). Git, ursprünglich entwickelt für die kollaborative Zusammenarbeit am Linux-Kernel, ist nicht beschränkt auf die Versionierung von Software-Quellcode, sondern kann – dank seiner textzeilenbasierten Arbeitsweise – auch für andere Dokumenttypen gewinnbringend eingesetzt werden: Manuskripte, Forschungsdaten in nicht-binären Formaten, etc.

Versionskontrollsysteme (VCS) schützen vor Verlust von Arbeitsergebnissen und erlauben einfaches Zurückspulen auf vorangegangene Stände. Zudem erlauben Sie – bei Einsatz in verteilten Arbeitsgruppen – im Rückblick die Zuweisung bestimmter Änderungen zu Mitarbeitenden und ermöglichen so einfache Klärung von Nachfragen. Schließlich weist das VCS zuverlässig auf konfligierende Änderungen hin, was bei anderen Formen der verteilten Kollaboration nicht immer der Fall ist.

Auch einzelnen Forschenden bietet der Einsatz von VCS, und insbesondere Git, Vorteile, wie etwa Commitbeschreibungen, die den Zugang zu auch zeitlich weit zurückliegenden Änderungen und Projektständen erleichtern.

Git oder ein anderes VCS kommt heutzutage in allen großen Softwareprojekten zur Verwendung und wird von den Entwickelnden auch für kleine Projekte genutzt.

Um ein wirklich fundiertes Grundwissen über Git vermitteln zu können verzichtet dieser Teilworkshop auf den Einsatz von grafischen Benutzeroberflächen für Git. Stattdessen kommt die Kommandozeilenimplementierung zum Einsatz, die für alle wichtigen Betriebssysteme erhältlich ist. Um ein komplikationsfreies Arbeiten aller Teilnehmenden auf der Kommandozeile zu ermöglichen werden daher entsprechende Kenntnisse vor Workshopbeginn abgefragt und bei Bedarf zu Beginn des Workshops zielgerichtet aufgebaut.

Im Anschluss werden VCS, ihre Funktionsweisen und Zielstellungen erläutert und die Einrichtung von Git vorgenommen. Es folgen die Vermittlung der grundlegenden Funktionen: Änderungsverfolgung, Exploration der Änderungshistorie, Ausschluss bestimmter Objekte von der Änderungsverfolgung, Änderungskonflikte und ihre Auflösung und die Arbeit mit entfernten Repositorien.

Darüber hinaus gehende Themen, die von der Versionskontrolle mit Git berührt werden finden ebenfalls Beachtung: Workflows für kollaboratives Arbeiten, Open Science, Lizenzierung, Zitierung und das Bereitstellen eigener Git-Instanzen. Nach Möglichkeit wird abschließend auf GUI-unterstützte Versionierung und Markdown-Formate eingegangen. Diese und vorhergehende Themen können von Teilnehmenden im zweiten Teilworkshop vertieft werden.

Der Teilworkshop beruht auf der entsprechenden Unterrichtseinheit „Version Control with Git“ (Ahmadia et al. 2016) aus dem Curriculum von Software Carpentry. Die Workshopleiter*innen sind zertifizierte Software Carpentry Instructors und es kommen die für das Curriculum entwickelten Methoden zum Einsatz, die es den Teilnehmenden ermöglichen, sich das Gelernte durch angeleitete Übungen und häufige formative Evaluationen aktiv anzueignen.

Teilworkshop „Kollaboratives Arbeiten im Web mit GitHub“

Dieser Teilworkshop führt ein in die Möglichkeiten der Nutzung von GitHub als Plattform für kollaboratives, verteiltes Arbeiten im Web. Komplexer werdende digitalisierte Forschungsprozesse und Projektstrukturen fordern zunehmend kollaboratives Arbeiten. Dabei stellen „Plain Text“-Formate ein problemlos nachnutzbares Mittel der Wahl dar. GitHub ist die derzeit wichtigste Codeplattform und bietet neben Git-Repositorien und eigenen Workflows für das Versionskontrollsystem Git weitere Features, die ein solches Arbeiten ermöglichen und erleichtern. Obgleich die Kernfunktionen von GitHub vornehmlich für die Verwaltug von Softwareprojekten genutzt wird, können diese ebenso gut für andere digitale Objekte, wie Forschungsdaten oder Manuskripte, genutzt werden.

Obwohl GitHub das Git schon im Namen trägt ist es möglich, die kollaborativen Funktionen der Plattform auch ausschließlich über die Weboberfläche von GitHub zu nutzen. Auf diese Weise kann nicht nur Versionskontrolle ausgeübt werden, es können in wenigen Schritten auch Websites erstellt und bereitgestellt werden. All dies, ohne auf die Kommandozeile zurückgreifen zu müssen.

Für den Workshop sind keine Vorkenntnisse über Versionskontrollsysteme, Git oder GitHub notwendig. Teilnehmende des ersten Teilworkshops werden jedoch grundsätzliche Prinzipien der Versionierung wiedererkennen und vertiefen können, beziehungsweise GUI-gestützte Umsetzungen kennenlernen. Die Teilnehmenden lernen, wie man Repositorien einrichtet, Dateien unter Versionskontrolle stellt, direkt mit Anderen an diesen Dateien zusammenarbeitet, beispielsweise durch Issues und Pull Requests. Auch weiterführende Themen wie Konfliktresolution, die Verwendung von Änderungsvorschlägen und Review und das Aufsetzen einer Website sind Themen des Workshops.

Auch dieser Teilworkshop beruht auf den Methoden der Software Carpentry.

Zusammenfassend ermöglichen die beiden Teilworkshops vor allem in Kombination den Teilnehmenden einen fundierten Einstieg sowohl in die Versionskontrolle allgemein, als auch in die praktische Arbeit mit den derzeit am weitesten verbreiteten Instrumenten, Git und GitHub. Die vermittelten Kenntnisse erlauben es, die Vorteile der Arbeit mit diesen von Anfang an auszunutzen und sind Grundlage für den täglichen Einsatz von Versionskontrolle und kollaborativen Prozessen. Beide sind ein zentraler Stützpfeiler für eine auf Best Practices basierende Forschungsarbeit in den digitalen Geisteswissenschaften und eine Voraussetzung von Open Science.

Workshopleiter*innen

Stephan Druskat, Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik, Unter den Linden 6, 10099 Berlin. E-Mail: stephan.druskat@hu-berlin.de.

Stephan Druskat ist Magister der Anglistik, Linguistik und neueren deutschen Literatur und arbeitet seit 2009 als Research Software Engineer (RSE) in Forschungsprojekten der Linguistik und der digitalen Geisteswissenschaften. Er ist Co-Convener der Dhd-AG „Research Software Engineering“, Mitbegründer und Vorstandsmitglied von *de-RSE e.V. – Gesellschaft für Forschungssoftware* und Special Collaborator des britischen *Software Sustainability Institute*. In seiner Forschung beschäftigt er sich vor allem mit Nachhaltigkeit von Forschungssoftware und Softwarezitierung.

Annika Rockenberger, Ph.D., Research Librarian for Digital Humanities, The National Library of Norway, Postboks 2674 Solli, NO-0203 Oslo, Norwegen. E-Mail: Annika.Rockenberger@nb.no.

Annika Rockenberger hat in Berlin u.a. Literaturwissenschaft, Geschichte und Kommunikationswissenschaft studiert. Mit einer Arbeit zur Analytischen Philosophie der Editionsphilologie ist sie an der Universität Oslo promoviert worden. Programmieren hat sie sich selbst beigebracht. Sie ist DH Aktivistin und Community Builder in Norwegen, dem skandinavischen Raum und Europa und derzeit für die Entwicklung einer DH Strategie an der norwegischen Nationalbibliothek zuständig.

Bibliographie

Chacon, Scott/ Straub, Ben (2014): *Pro Git (2nd Edition)*. New York City, NY, USA: Apress. 10.1007/978-1-4842-0076-6 .

Ahmadia, Aron / Allen, James / Appling, Alison / Aubin, Sean / Bachant, Pete / Banaszkiewicz, Piotr / Barmby, Pauline / Batut, Berenice / Bekolay, Trevor / Blischak, John / Bonsma, Madeleine / Borrelli, Jon / Boughton, Andy / Bouquin, Daina / Brauning, Rudi / Brett, Matthew / Brown, Amy / Cabunoc, Abigail / Charlesworth, Jane / Charlton, Billy / Chen, Daniel / Christensen, Garret / Collings, Ruth / Corvellec, Marianne / Davis, Matt / Dolson, Emily / Duchesne, Laurent / Duckles, Jonah / Emonet, Rémi / Estève, Loïc / Farsarakis, Emmanouil / Fauber, Bennet / Fouilloux, Anne / Förstner, Konrad / Geiger, Stuart / Gonzalez, Ivan / Guarinello, Marisa / Hadwin, Jamie / Hannah, Nicholas / Hansen, Michael / Heroux, Martin / Hertweck, Kate / Hinsen, Konrad / Huang, Daisie / Ismiraldi, Yuandra / Jackson, Mike / Jacobs, Christian / Jarecka, Dorota / Johnston, Luke W. / Jones, David / Jędrzejewski-Szmek, Zbigniew / King, W. Trevor / Kluyver, Thomas / Konrad, Bernhard / Kuzak, Mateusz / Labrie, Kathleen / Lapp, Hilmar / Latornell, Doug / Laufersweiler, Mark / LeBauer, David / Lee, Kate / Liffers, Matthias / Loucks, Catrina / Ma, Keith / Marwaha, Kunal / Michonneau, François / Mills, Bill / Mueller, Andreas / Nagraj, VP / Nederbragt, Lex / Nunez-Iglesias, Juan / O'Brien, Brenna / O'Leary, Aaron / Olsson, Catherine / Pawsey, Chris / Pfenninger, Stefan / Pipitone, Jon / Poisot, Timothée / Preney, Paul / Rice, Timothy / Riemer, Kristina / Rio Deiros, David / Robinson, Natalie / Rohl, Andrew / Rokem, Ariel / Sarahan, Michael / Schmeier, Sebastian / Schmider, Hartmut / Silva, Raniere / Smithyman, Brendan / Soranzo, Nicola / Steinbach, Peter / Stevens, Sarah / Timbers, Tiffany / Traphagen, Danielle / Tröndle, Tim / van der Walt, Anelda / Vandervalk, Steve / Weaver, Belinda / Wheelhouse, Mark / White, Ethan / Wilson, Greg / Wu, Steven / Zhang, Qingpeng (2016): *Software Carpentry: Version Control with Git (Version 2016.06)*, in: Huang, Daisie / Gonzalez, Ivan (Hrsg.): https://github.com/swcarpentry/git-novice [letzter Zugriff 13. Oktober 2018]. 10.5281/zenodo.57467 .

Wilson, Greg (2006): *Software Carpentry: Getting Scientists to Write Better Code by Making Them More Productive*, in: *Computing in Science & Engineering* 8: 66-69. 10.1109/MCSE.2006.122 .

Vom gedruckten Werk zu elektronischem Volltext als Forschungsgrundlage Erstellung von Forschungsdaten mit OCR-Verfahren

Boenig, Matthias

boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Herrmann, Elisa

elisa.herrmann@hab.de
Herzog August Bibliothek Wolfenbüttel

Hartmann, Volker

volker.hartmann@kit.edu
Steinbuch Centre for Computing / KIT

EINLEITUNG

In den vergangenen 30 Jahren ist ein beträchtlicher Teil des in Deutschland gedruckten Materials aus der Zeit von 1500 bis ca. 1850 in mehreren, durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Kampagnen in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16.-18. Jahrhunderts (VD16, VD17, VD18) zunächst nachgewiesen und seit 2006 digitalisiert worden. Zusätzlich vorliegender Volltext wird mittlerweile auf breiter disziplinärer Front als Schlüssel zu einer ganzen Reihe von geistes- und kulturwissenschaftlichen Forschungsfragen gesehen und gilt zunehmend als elementare Voraussetzung für die Weiterentwicklung der transdisziplinär arbeitenden Digital Humanities. Deshalb werden bereits an verschiedenen Stellen OCR-Verfahren angewendet; viele dieser Unternehmungen haben allerdings noch sehr starken Projektcharakter. Die informationswissenschaftliche Auseinandersetzung mit OCR kann an der großen Zahl wissenschaftlicher Studien und Wettbewerbe ermessen werden, die Möglichkeiten zur Verbesserung der Textgenauigkeit sind in den letzten Jahrzehnten enorm gestiegen. Der Transfer der auf diesem Wege gewonnenen, oftmals sehr vielversprechenden Erkenntnisse in produktive Anwendungen ist jedoch häufig nicht gegeben: Es fehlt an leicht nachnutzbaren Anwendungen, die eine qualitativ hochwertige Massenvolltextdigitalisierung aller historischen Drucke aus dem Zeitraum des 16. bis 19. Jahrhundert ermöglichen.

Auf dem DFG-Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“ (Deutsche Forschungsgemeinschaft 2014) im März 2014 formulierten Expertinnen und Experten daher folgende Desiderate um die Weiterentwicklung von OCR-Verfahren zu ermöglichen. Es bestehe eine dringende Notwendigkeit für freien Zugang zu historischen Textkorpora und lexikalischen Ressourcen zum Training von vorhandener Software zur Texterkennung bestehe. Ebenso müssen Open-Source-OCR-Engines zur Verbesserung der Textgenauigkeit weiterentwickelt werden, wie auch Anwendungen für die Nachkorrektur der automatisch erstellten Texte. Daneben sollten Workflow, Standards und Verfahren der Langzeitarchivierung mit Blick auf zukünftige Anforderungen an den OCR-Prozess optimiert werden. Als zentrales Ergebnis dieses Workshops stand fest, dass eine koordinierte Fördermaßnahme der DFG notwendig ist. Die „Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)“, kurz OCR-D, begann im September 2015 und versucht seitdem einen Lückenschluss zwischen Forschung und Praxiseinsatz, indem für die Entwicklungsbedarfe Lösungen erarbeitet und der aktuelle Forschungsstand zur OCR mit den Anforderungen aus der Praxis zusammengebracht werden.

ARBEITEN IM PROJEKT OCR-D

Das Vorhaben hat zum Ziel, einerseits Verfahren zu beschreiben und Richtlinien zu erarbeiten, um einen optimalen Workflow sowie eine möglichst weitreichende Standardisierung von OCR-bezogenen Prozessen und Metadaten zu erzielen, andererseits die vollständige Transformation des schriftlichen deutschen Kulturerbes in digitale Forschungsdaten in (xml-strukturierter Volltext) konzeptionell vorzubereiten. Am Ende des Gesamtvorhabens (d.h. unter Einschluss der Modulprojektphase) sollte ein in allen Aspekten konsolidiertes Verfahren zur OCR-Verarbeitung von Digitalisaten des schriftlichen deutschen Kulturerbes stehen und eine Dokumentation, die Antworten auf die damit verbundenen technischen, informationswissenschaftlichen und organisatorischen Probleme und Herausforderungen gibt sowie Rahmenbedingungen formuliert.

Das Projekt ist in zwei Phasen geteilt: In der ersten Phase hat das Koordinierungsgremium von OCR-D Bedarfe für die Weiterentwicklung von OCR-Technologien analysiert und sich intensiv mit den Möglichkeiten und Grenzen der Verfahren zur Text- und Strukturerkennung auseinandergesetzt. Zahlreiche Gespräche mit ExpertInnen aus Forschungseinrichtungen und Bibliotheken sowie Sichtung vorhandener Werkzeuge aber auch Betrachtung vorhandener Textsammlungen sowie aktueller und geplanter Digitalisierungsvorhaben mündeten in der Erkenntnis, dass der Lückenschluss zwischen Wissenschaft und Praxis das primäre Desiderat im Bereich der Textdigitalisierung darstellt. Zudem hat sich im Lauf der ersten Projektphase eine technologische Wende auf dem Gebiet der Zeichenerkennung vollzogen - an die Stelle traditioneller Verfahren der Mustererkennung, die auf einer Segmentierung von Textabschnitten in Zeilen, Wörter und schließlich einzelne Glyphen basieren, die anschließend aufgrund charakteristischer Merkmale (z.B. Steigung an Kanten) erkannt werden, ist eine zeilenorientierte Sequenzklassifizierung auf Basis statistischer Modelle, insbesondere verschiedener Arten neuronaler Netze (sog. *Deep Learning*), getreten. Grund für diesen Technologiewechsel ist die vielfach nachgewiesene Überlegenheit segmentierungsfreier Erkennungsverfahren bezüglich der resultierenden Textgenauigkeit. Diese Überlegenheit gilt insbesondere für schwierige, historische Vorlagen. Dieser Technologiewandel hat sich bisher nicht oder nur äußerst begrenzt auf die Digitalisierungspraxis ausgewirkt. Der Grund dafür liegt vor allem in den bisher bestehenden Hürden beim Einsatz verfügbarer OCR-Lösungen auf Basis neuronaler Netze. Ohne weitreichende projektspezifische Anpassungen ist ein produktiver Einsatz derzeit nicht möglich. Das betrifft unter anderem die Erstellung passender Erkennungsmodelle, die durch das Trainieren eines neuronalen Netzes auf Basis ausgewählter Ground-Truth-Daten generiert werden. Dafür sind zum einen hochqualitativer und umfangreicher Ground Truth aber auch Erfahrungen bzgl. freier Parameter wie z.B. Anzahl der Trainingsschritte, Lernrate, Modelltiefe unabdingbar. Aus OCR-D heraus ist daher ein Datenset mit Trainings- und Ground-Truth-Daten entstanden, welches für Trainings und Qualitätsanalysen im Vorhaben selber genutzt wird aber auch durch andere Forschungsprojekte nachgenutzt werden kann. Neben der Qualität der Zeichenerkennung sind es vor allem Umfang und Korrektheit der strukturellen Annotationen, die die Utilität eines Volltexts für wissenschaftliche Kontexte determinieren. Auch im Bereich der automatischen Layouterkennung (OLR) gab es innerhalb des bisherigen Projektzeitraums vielversprechende Forschungsergebnisse durch den Einsatz innovativer statistischer Verfahren. Der Übertrag in die Praxis in Form nachnutzbarer Software ist hier jedoch noch nicht gegeben. Kommerzielle OCR-Lösungen ignorieren diesen Bereich weitestgehend und bieten nur minimale Strukturinformationen auf Seitenebene (Text, Tabelle, Abbildung etc.) an. Tiefergehende strukturelle Auszeichnungen (Kapitelstruktur, Bildunterschriften, Inhaltsverzeichnisse) werden daher manuell erfasst und in METS/MODS repräsentiert. Eine Verknüpfung zwischen Struktur und Volltext findet, obwohl technisch möglich, in vielen Digitalisierungsvorhaben nicht statt. Für die philologische, editorische oder linguistische Wissenschaftspraxis bedeutet das eine massive Einschränkung die bspw. eine sinnvolle Transformation in hochstrukturierte Formate wie TEI verhindert.

Die Erkenntnisse dieser Bedarfsanalyse mündeten in einem OCR-D-Funktionsmodell, welches den Rahmen für die Modulprojekt-Ausschreibung der DFG im März 2017 bot. Vor diesem Hintergrund wurden acht Modulprojekte bewilligt die seit 2018 an Lösungen zur Bildvorverarbeitung, Layouterkennung, Textoptimierung (inkl. Nachkorrektur), zum Modelltraining und zur Langzeitarchivierung der OCR-Daten arbeitet. Die Entwicklungen schöpfen dabei das Potential innovativer Methoden für den gesamten Bereich der automatischen Texterkennung für die Massenvolltextdigitalisierung von historischen Drucken aus. Sie werden anschließend nahtlos in den OCR-D-Workflow zur optimierten OCR-basierten Texterfassung integriert. Das so entstehende OCR-D-Softwarepaket steht damit Kultureinrichtungen wie Forschenden für die automatische Texterkennung als Open-Source-Software zur Verfügung.

Die meisten Arbeiten werden im Sommer 2019 abgeschlossen sein, aber bereits Anfang des Jahres wird die Alpha-Version einen Einblick in die zu erwartende Gesamtlösung bieten können.

ZIEL DES WORKSHOPS

Der Workshop soll neben der Vorstellung des Projektes und der Software die Gelegenheit bieten selber die Software zu testen und zugleich über Optimierungen und Anforderungen seitens der Wissenschaft an diese Technologien zu diskutieren. Teilnehmende erhalten somit einen exklusiven Einblick in die Entwicklungsarbeit und haben die Möglichkeit proaktiv auf die Arbeiten Einfluss zu nehmen, die Ihren späteren Forschungsalltag begleiten und verbessern soll.

PROGRAMM

Der Workshop gliedert sich in drei Abschnitte:

- Vorstellung des Projekts OCR-D, des Ground-Truth-Datensets und der Guidelines (30min)
- Demonstration der Eigenentwicklung und eines Test-Workflows (120min)

- Diskussion zu Anforderungen und Optimierungen aus Sicht der Digital Humanities (30min)

Der erste Abschnitt stellt die Hintergründe zum Vorhaben vor und geht auf Besonderheiten der Volltextdigitalisierung von historischen Beständen ein. Anschließend wird das Trainings- und Ground-Truth-Datenset präsentiert, das im Rahmen von OCR-D auf- und weiter ausgebaut wird. Besonders die dazu entwickelten Guidelines geben Hinweise für eine spätere Nachnutzung und die Erstellung eigener Ground-Truth-Daten in anderen Projekten. Der Fokus des Workshops liegt auf dem zweiten Abschnitt, in welche der derzeitige Entwicklungsstand präsentiert wird. Die benötigten Test-Dateien werden auf GitHub[1] veröffentlicht. Abgerundet wird der Workshop durch eine Diskussionsrunde zu Anforderungen aus der Wissenschaft heraus an OCR-Techniken und die dafür eingesetzte Software.

VORAUSSETZUNG

Teilnehmende benötigen einen eigenen Laptop mit Internetanbindung und Ubuntu 18.04 als Betriebssystem. Alternativ kann auch Windows/Mac OSX mit der Software VirtualBox verwendet werden. Die VM wird den Teilnehmenden vom OCR-D-Projekt vor Ort zur Verfügung gestellt. Die Anzahl der Teilnehmenden ist auf 20-25 begrenzt. Python- und Linux-Kommandozeilen-Kenntnisse sind wünschenswert

Fußnoten

1. OCR-D Git-Hub: https://github.com/OCR-D/

Bibliographie

Deutsche Forschungsgemeinschaft (2014): Workshop *Verfahren zur Verbesserung von OCR-Ergebnissen*. Protokoll zu den Ergebnissen und Empfehlungen des Workshops. http://www.dfg.de/download/pdf/foerderung/programme/lis/140522_ergebnisprotokoll_ocr_workshop.pdf [Zuletzt abgerufen 07.01.2019]

Wissenschaftliches Bloggen mit de.hypotheses

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

Menke, Ulla
menke@maxweberstiftung.de
Max Weber Stiftung

Computer und Internet haben die Art und Weise, wie Forscher kommunizieren und zusammenarbeiten, grundlegend verändert. Ab Anfang der 90er Jahre konnten Wissenschaftlerinnen und Wissenschaftler über verschiedene Orte und Zeitzonen hinweg kollaborativ an Text, Bild, Audio, Video und Code arbeiten. Während E-Mail, Newsgroups und Online-Chats eine many-to-many-Kommunikation im virtuellen Raum ermöglichten, kamen die wichtigsten aktuellen Entwicklungen in der wissenschaftlichen Online-Kommunikation durch soziale Medien: mit Microblogging, Blogs, Wikis und Social Network Sites (SNS) wie Facebook, Academia.edu, ResearchGate und anderen. Durch sie wurden die Hindernisse für die Veröffentlichung und Kommunikation im Internet deutlich reduziert. Produktionsprozesse, die bisher professionelles Wissen, Ausrüstung und Kapital erforderten, können nun von einfachen Personen mit Computer- und Internetzugang durchgeführt werden. In der Folge wurde das Ökosystem der wissenschaftlichen Kommunikation breiter, schneller, interaktiver, dynamischer, multimodaler und zunehmend vernetzter (König 2015).

Als öffentlich geführte wissenschaftliche Notizbücher eignen sich insbesondere Wissenschaftsblogs zur selbstkritischen Reflektion 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, sich zu vernetzen, Schreiben zu üben und Gedanken im Schreibprozess zu ordnen. Wissenschaftsblogs haben ein hohes Potential für die schnelle Verbreitung und Diskussion aktueller Forschungsinhalte und nutzen die Möglichkeiten des Web 2.0 für eine direkte und interaktive Publikation, bei der multimediale Inhalte wie Bilder, Grafiken, Animationen und Verlinkungen ohne Mehrkosten eingebunden werden können. Wissenschaftsblogs werden zwar zumeist für die eigene Fachcommunity geschrieben, sie sind jedoch offen einsehbar und werden ebenso von Journalisten und von der breiten Öffentlichkeit wahrgenommen.

Wissenschaftsblogs bieten Einblicke in die Werkstatt von Forschenden und zeigen Forschung im Entstehen (Mounier 2013). Gerade in den Digital Humanities sind Blogs und Twitter die wichtigsten Medien für die Diskussion neuer Forschungsansätze und Methoden (Ullyot 2012). Blogs dokumentieren den Forschungsprozess und damit die Phase vor der abschließenden Projektveröffentlichung. Damit ersetzen sie bisherige Praktiken und Formate der Kommunikation und Publikation zumeist nicht – auch wenn sie es theoretisch könnten –, sondern ergänzen diese und stellen in ihrer Ausprägung etwas Neues dar: ein eigenes Format, das Kennzeichen aus der analogen (mündlich wie schriftlichen) und der digitalen Wissenschaftskommunikation als „missing link“ mischt und um neue Merkmale ergänzt. Wenn das Medium die Botschaft ist (Marshall McLuhan), dann zeigen bloggende Forscherinnen und Forscher, wie sie sich Wissenschaft vorstellen: offen, vernetzt, horizontal, direkt, schnell, vielseitig, multimedial… und mit der akzeptierten Möglichkeit, sich zu irren (König 2015). Forschende schreiben in Blogs über einzelne Aspekte ihres Themas, über Publikationen, die sie gelesen haben, über Vorträge und Veranstaltungen, die sie besucht oder über Begegnungen, die sie inspiriert haben. Blogbeiträge handeln von einem konkreten Ereignis oder Gegenstand oder entwickeln

theoretische und methodische Überlegungen. Zumeist zeigt ein Wissenschaftsblog die subjektive Lebenswelt der Forschenden und macht somit ganz generell die Subjektivität der Wissenschaft und des wissenschaftlichen Tuns deutlich.

Mit de.hypotheses.org wurde Anfang 2012 eine Plattform für geistes- und sozialwissenschaftliche Blogs geschaffen, in deren Umfeld seither eine stetig wachsende deutschsprachige Community als Teil eines europäischen Netzwerks entstanden ist. Mittlerweile sind dort über 500 deutschsprachige Blogs aus allen geisteswissenschaftlichen Disziplinen vereint. Die Blogplattform trägt zur Sichtbarkeit und zur Vernetzung der Bloggenden bei und ist eine zentrale Anlaufstelle, bei der die Blogs langzeitarchiviert werden, eine ISSN verliehen bekommen und die Blogbeiträge mit Permalinks ausgestattet sind. Für die Startseite des Portals werden von einer Redaktion und vom Community Management die aktuell besten Beiträge ausgewählt und kuratiert, die darüber eine erhöhte Sichtbarkeit erhalten.

Die bei de.hypotheses vorhandenen unterschiedlichen Blogtypen belegen die große Vielfalt der geisteswissenschaftlichen Blogosphäre. Es gibt Blogs von Forschergruppen und zu Forschungsprojekten, thematische Gemeinschaftsblogs, Blogs zu Quellen und Methoden, Blogs von Instituten und wissenschaftlichen Einrichtungen wie Archive und Bibliotheken, Seminar- und Tagungsblogs, Blogs, die eine Zeitschrift oder eine Publikation begleiten, Blogs für Lehre und Didaktik, Fotoblogs, Blogs zu einer wissenschaftlichen Debatte etc. (König 2013).

Der auf einen halben Tag angelegte Workshop knüpft direkt an den medientheoretischen und –praktischen Teil des Tagungsthemas an und richtet sich an DH-Forschende, die bisher noch nicht bloggen und ein eigenes Wissenschaftsblog anlegen möchten – ob als Einzel- oder als Gemeinschaftsblog, ob begleitend zur Lehre oder zu einem Forschungsprojekt – und dafür ein Konzept entwickeln und grundlegende inhaltliche und technische Überlegungen anstellen und praktisch einüben möchten.

Im Rahmen des Workshops wird zum einen die theoretische und konzeptionelle Seite des wissenschaftlichen Bloggens als eigenes multimediales Medium besprochen, zum anderen ein praktischer Teil angeboten. Zunächst werden einleitend verschiedene aktuelle Praktiken des Wissenschaftsbloggens, der besondere Schreibstil und die Interaktion mit der Leserschaft thematisiert und Elemente für die Strategiebildung für ein eigenes Wissenschaftsblog erläutert (darunter: was bloggen? wie bloggen? wie viel Zeit investieren? für welches Publikum? alleine oder kollaborativ bloggen? wie Themen für das Wissenschaftsblog finden? (Scherz 2013). Es werden best practice Beispiele aus verschiedenen geisteswissenschaftlichen Disziplinen und aus den DH vorgestellt. Gegenstand der Diskussion sind darüber hinaus rechtliche Fragen (vom Einbinden fremder Inhalte wie Bilder und Videos und dem Lizenzieren eigener Inhalte, über die Bestimmungen der DSGVO bis hin zum „Eigenplagiat“ bei Promovierenden usw.) sowie die Frage nach dem „return of investment“ des Bloggens, das durchaus zeitintensiv sein kann und aufgrund der zumeist mangelnden offiziellen Anerkennung überlegt erfolgen sollte. Thematisiert wird außerdem der Umgang mit Kommentaren im Blog sowie die Frage, was Promovierende über ihre Dissertationen bloggen können und was nicht.

In einem Hands-on-Teil – der etwa drei Viertel der Zeit des Workshops einnehmen wird – werden anschließend Schritt für Schritt die einzelnen Aspekte der Blogpraxis vorgestellt und vorgeführt. Die Teilnehmerinnen und Teilnehmer vollziehen die einzelnen Schritte an eigens eingerichteten Wordpress-Schulungsblogs nach, von der Gestaltung und Einrichtung des Blogs, der Formulierung einer guten Überschrift, einer sinnvollen Navigation und Kategorienbildung bis hin zum Einbetten von Videos, und wenden damit das Gelernte sofort an. Sie lernen darüber die Grundlagen moderner CMS-Systeme kennen. Während des Workshops werden parallel zum praktischen Teil Tipps gegeben für die Anfangsphase eines wissenschaftlichen Blogs und für Themen wie Suchmaschinenoptimierung, Menüführung und graphische Gestaltung.

Inhalte und Übungen des Praxisteils sind: Erstellen einer Menüleiste, öffentliche Autorennamen einstellen und Profil ausfüllen, Rechteverwaltung bei mehreren Nutzerinnen und Nutzern, Titel und Untertitel ändern, Design-Theme auswählen (welche sind für welche Form des Bloggens geeignet?), Bild in die Kopfzeile einfügen, Nennung des Urhebers und Lizenz, eigenen Artikel erstellen, Überschrift auswählen, Zitat einfügen, Fußnoten, Verlinkung einfügen, Weiterlesen-Button, Bildrechte, Bilder und Lizenzen einfügen (Grundlagen CC-Lizenzen), Kategorien, Schlagwörter zuweisen, Seite anlegen, Menü anlegen, Meta, Text, Bild und Link, RSS-Feed, Verknüpfung zu Twitter, Videos einbinden, Statistiken lesen, Reichweite vergrößern, rechtliche Bestimmungen der DSGVO, Umgang mit Kommentaren.

Die Anzahl der Teilnehmenden ist auf 25 begrenzt. Der Workshopraum sollte über ein W-Lan und einen Beamer verfügen. Eine weitere technische Ausstattung wird nicht benötigt. Besondere technische Vorkenntnisse sind für die Teilnahme nicht erforderlich. Ein eigenes Laptop oder ein anderes Endgerät muss selbst mitgebracht werden.

Workshopleiterinnen: Dr. Mareike König: Sie ist Projektleiterin der deutschsprachigen Plattform für geisteswissenschaftliche Blogs de.hypotheses und leitet dort die Redaktion. Ihre Forschungsinteressen beziehen sich auf Wissenschaftskommunikation im Web 2.0 und hier speziell auf das Wissenschaftsbloggen als neue Form des wissenschaftlichen Schreibens als Herausforderung für unsere Wissenschaftskultur.

Kontakt: Dr. Mareike König, DHIP, 8, rue du Parc Royal, 75003 Paris, mkoenig@dhi-paris.fr

Ulla Menke: Sie ist Community Managerin der Blogplattform de.hypotheses seit 2016 und arbeitet in der Max Weber Stiftung. Als Community Managerin kümmert sie sich um die rund 350 Wissenschaftsblogs, die es auf der Plattform de.hypotheses derzeit gibt und steht den Bloggenden seit 2016 mit Tipps und Hilfe bei Fragen rund um Technik, SEO, Blognavigation und Bloggestaltung zur Seite.

Kontakt: Ulla Menke, Max Weber Stiftung, Rheinallee 8, Bad Godesberg, menke@maxweberstiftung.de

Bibliographie

König, Mareike (2013): *Die Entdeckung der Vielfalt: Geschichtsblogs auf der internationalen Plattform hypotheses.org*, in: **Haber, Peter / Pfanzelter, Eva (eds.)**: *Historyblogosphere. Bloggen in den Geschichtswissenschaften*, München: Oldenbourg 181–197.

König, Mareike (2015): *Herausforderung für unsere Wissenschaftskultur: Weblogs in den Geisteswissenschaften* in: **Schmale, Wolfgang (ed.):** *Digital Humanities. Praktiken der*

Digitalisierung, der Dissemination und der Selbstreflexivität, Stuttgart: Steiner 57-74.

Scherz, Sabine (2013a): *Warum sollte ich als Wissenschaftler/in bloggen?* in Redaktionsblog, 21.5.2013, https://redaktionsblog.hypotheses.org/1209 .

Scherz, Sabine (2013b): *Mein erster wissenschaftlicher Blogartikel – was schreibe ich bloß?* in Redaktionsblog, 24.5.2013, https://redaktionsblog.hypotheses.org/1214 .

Scherz, Sabine (2013c): *Wie finde ich Themen für mein Wissenschaftsblog?* in Redaktionsblog, 28.5.2013, https://redaktionsblog.hypotheses.org/1217 .

Scherz, Sabine (2013d): *Texte für das Wissenschaftsblog schreiben, wie?* in Redaktionsblog, 5.6.2013, https://redaktionsblog.hypotheses.org/1220 .

Mounier, Pierre (2013): *Die Werkstatt öffnen: Geschichtsschreibung in Blogs und sozialen Medien*, in: **Haber, Peter / Pfanzelter, Eva (eds.):** *Historyblogosphere. Bloggen in den Geschichtswissenschaften*, München: Oldenbourg 51-59.

Ullyot, Michael (2012): *On Blogging in the Digital Humanities*, in: Ullyot, http://ullyot.ucalgaryblogs.ca/2012/02/24/on-blogging-in-the-digital-humanities/.

Panels

Das Wissen in der 3D-Rekonstruktion

Einleitung

In der architekturgeschichtlichen Forschung finden digitale 3D-Rekonstruktionen seit mehr als 30 Jahren als Wissensträger, Forschungswerkzeuge und Darstellungsmittel Verwendung. Dabei hat zum einen die Zahl der erstellten digitalen Rekonstruktionen von historischer Architektur in den vergangenen Jahren kontinuierlich zugenommen, zum anderen weisen diese höchst unterschiedliche technische, grafische und inhaltliche Qualitäten auf. Darüber hinaus werden zumeist weder Entstehungsprozesse noch die Qualität einer zu Grunde liegenden forscherischen Arbeit transparent. Während eine Diversität von Modellen und Werkzeugen nicht zuletzt durch die Pluralität der damit untersuchten Fragestellungen wünschenswert ist, stellt sich die Frage nach Kriterien und Ansätzen, um den Gebrauch dieser Werkzeuge sowie resultierende Ergebnisse zu bewerten und zu validieren. Mit Blick darauf unterliegt der Einsatz von Methoden digitaler Rekonstruktion in der architekturgeschichtlichen Forschung seit jeher einer Ambivalenz. Eindrucksvollen Anwendungsbeispielen und Forschungspotentialen steht eine ganze Reihe überaus berechtigter wissenschaftlich-methodischer Vorbehalte und Desiderata gegenüber.

Entsprechend verfolgt das 2018 von der DFG genehmigte wissenschaftliche Netzwerk "Digitale 3D-Rekonstruktionen als Werkzeuge der architekturgeschichtlichen Forschung" das Ziel, für digitale Rekonstruktionen im Kontext der Architekturgeschichte erstmals eine umfassende Betrachtung aus Perspektive der Einbettung in wissenschaftliche Kontexte vorzunehmen. Die konkrete Fragestellung lautet dabei, wie digitale Rekonstruktionsmethoden als wissenschaftliche Werkzeuge im Kontext architekturgeschichtlicher Forschung validiert werden können.

Eine initiale Themenstellung und damit verbundene Fragenkomplexe wurden unter Beteiligung von ca. 30 Wissenschaftlerinnen und Wissenschaftlern Ende 2014 in einem Workshop der „Arbeitsgemeinschaft Digitale Rekonstruktion des Digital Humanities im deutschsprachigen Raum (DHd) e.V." entwickelt. Die Arbeitsgruppe „Digitale Rekonstruktion" ging aus der 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (25.-28.03.2014, Universität Passau) hervor. Die Arbeitsgruppe versammelt Kolleginnen und Kollegen, die sich dem Thema digitale Rekonstruktion aus dem Blickwinkel der Architektur, Archäologie, Bau- und Kunstgeschichte sowie Computergraphik und Informatik verschrieben haben. Die Arbeitsgruppe bietet eine Plattform für einen Austausch und eine feste Etablierung der digitalen Rekonstruktion im Dienste einer Erfassung, Erforschung und Vermittlung kultureller und geschichtlicher Inhalte innerhalb der Digital Humanities. Vorrangiges Ziel der Arbeitsgruppe ist es, die Akteure im deutschsprachigen Raum zusammenzubringen, um sich den Fragen der Begriffsklärung und der Arbeitsmethodik sowie der Dokumentation und Langzeitarchivierung von digitalen Rekonstruktionsprojekten zu widmen.

Vor diesem Hintergrund soll zur DHd Jahrestagung gemeinsam von Netzwerk und AG ein Panel durchgeführt werden, das die drei Kernpunkte der DHd-Jahrestagung – Theorie, Modellierung, Synthese – unter dem Blickwinkel der 3D-Rekonstruktion aufgreift. Damit knüpft das Panel an die bisher drei bei vergangenen DHd-Jahrestagungen durch die Arbeitsgemeinschaft durchgeführten Panels an, die „Allgemeine Standards, Methodik und Dokumentation" (Kuroczyński et al., 2014) und „aktuelle Herausforderungen" (Kuroczyński et al., 2015) sowie Transformationsprozesse „vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell" (Kuroczyński et al., 2016b) beleuchteten.

Ein wesentliches Kernmerkmal von 3D-Rekonstruktionen ist, dass diese auf die Genese, Transformation und Vermittlung von Wissen abzielen. Im Gegensatz zu Daten muss Wissen sowohl für Menschen lesbar als auch relevant und kognitiv anknüpfbar sein (vgl. Frické, 2018). Entsprechend ist das Ziel des für die DHd-Jahrestagung 2019 geplanten Panels, insbesondere den Umgang mit Wissen vor dem Hintergrund eines Erkenntnis- oder Vermittlungsinteresses zu thematisieren und dabei jeweils kontrastierende und/oder komplementäre Perspektiven darzulegen.

Theoretische und methodische Grundlagen

Epistemische Perspektive (Sander Münster)

Wie verändert die Digitalisierung die Forschung zu Kulturerbe? Was kennzeichnet eine disziplinäre Kultur der 3D-Rekonstruktion? Aus organisationaler Perspektive changiert die Nutzung digitaler Technologien und Ansätze derzeit zwischen einer Einordnung als Teilbereich der Geisteswissenschaften und der "Neudefinition der traditionellen Geisteswissenschaften mit digitalen Mitteln" (Adams and Gunn, 2013). Zur Untersuchung führten wir 15 Interviews mit Digital Humanists in London, die sich mit Objekten und Bildern beschäftigten (Münster and Terras, accepted paper). Welche allgemeinen Forschungsansätze sind dabei ersichtlich? Neben der durch Technologie ermöglichten Beantwortung neuartiger Forschungsfragen und dem Einsatz von Computertechnologien als Medium "für neue Forschungspraktiken ohne zwangsläufige Transformation der Forschungsmethoden" (Long and Schonfeld, 2014) wird ein dritter Typ sichtbar: geisteswissenschaftliche Forschung, welche technologiebezogene Fragen beispielsweise zu User-Engagement, Forschungsethik oder Wissenschaftsphilosophie (u.a. Harari, 2017) beleuchtet und eine umfassende Erklärung technischer Entwicklungen und Phänomene zu ermöglicht. Dabei sind die Digital Humanities als Mode-2-Forschung durch eine transdisziplinäre und anwendungsorientierte Forschung gekennzeichnet (c.f. Nowotny et al., 2003, Hessels and Lente, 2007). Weitere Attribute wie eine interdisziplinäre Teamarbeit, die arbeitsteilige Wissensproduktion und der Einsatz von Maschinen bzw. Software (De Solla Price, 1963) sind vor allem in den Ingenieurwissenschaften, aber weniger in den Geisteswissenschaften anzutreffen. Dies mag erklären, warum Geisteswissenschaftler über eine gegenüber Ingenieuren tendenziell höhere Hürde für den Einstieg in die Digital Humanities berichten. Dieser Befund gilt insbesondere für die 3D-Rekonstruktion, welche durch die Nutzung von 3D-

Werkzeugen als komplexe Expertensysteme eine langwierige Qualifizierung oder aber eine Arbeitsteiligkeit voraussetzt.

Historische Perspektive (Heike Messemer)

Seit Anfang der 1980er Jahre erarbeiten Experten unterschiedlicher Disziplinen wissenschaftliche digitale 3D-Rekonstruktionen von historischer Architektur (Messemer 2016; Messemer 2018). Hierfür greifen sie auf die jeweils zur Verfügung stehenden historischen Quellen zurück. Insofern geben die 3D-Modelle das zum Zeitpunkt ihrer Erstellung zusammengetragene Wissen zum betreffenden Bauwerk wieder, das je nach Quellenlage von (annähernd) lückenlos bis relativ lückenhaft reichen kann. Interpretationen und Hypothesen sind damit ebenso wesentliche Bestandteile von digitalen Rekonstruktionen. Zu fragen ist nun wie das Wissen, die Unsicherheiten, die Interpretationen und Hypothesen im 3D-Modell visuell dargestellt werden können. Die Visualisierung unterschiedlicher Wissensstände wurde und wird bislang sehr heterogen gehandhabt. Eine Rückschau auf vergangene Projekte bietet ein breites Spektrum an Visualisierungsmöglichkeiten, die kritisch reflektiert zu neuen Lösungsansätzen führen können: Bedeuten mehr Details im Modell auch ein Mehr an Wissen zum dargestellten Bauwerk? Welche Rolle spielt die Weiterentwicklung der Technik in Bezug auf Visualisierungsmöglichkeiten? Wie wurden Unsicherheiten im Wissen zum dargestellten Objekt in der Vergangenheit visualisiert? Auf Basis einer analytischen Rückschau können Vorschläge erarbeitet werden, wie zukünftig Unsicherheiten und gesichertes Wissen im Modell visuell gekennzeichnet werden können.

Modellierung

Datenmodellierung (Peggy Große)

In virtuellen Forschungsumgebungen, wie WissKI (= Wissenschaftliche Kommunikationsinfrastruktur, wiss-ki.eu), können die der Rekonstruktion als Grundlage dienenden Quellen standardisiert erfasst und mit den entsprechenden 3D-Modell-Varianten und –Versionen nachvollziehbar verknüpft werden. Die Erfassung und Verknüpfung weiterer Informationen, z.B. zum historischen Kontext eines modellierten Bauwerkes, sind ebenfalls nachvollziehbar möglich. Das für die Forschungsumgebung WissKI verwendete Datenmodell ist eine Applikationsontologie (CHML) auf Grundlage von CIDOC CRM (http://www.cidoc-crm.org). Alle Informationen werden gleichwertig auf Basis eines formal strukturierten Datenmodells abgebildet, sodass die gewünschte Nachhaltigkeit und Nachnutzung komplexer geisteswissenschaftlicher Inhalte für Mensch und Maschine auf lange Sicht interpretierbar bleiben. Eine Herausforderung stellt die Auswahl der Klassen und Eigenschaften dar, um die zu erfassenden Sachverhalte CIDOC CRM konform abzubilden. Hierfür müssen konkrete Inhalte und Objekte auf ihre gemeinsamen übergeordneten Eigenschaften hin ausgewertet werden. Der Beitrag möchte neben dem Datenmodell anhand ausgewählter Beispiele die Klassenzuweisung der Inhalte erläutern und zur Diskussion stellen.

3D-Modellierung (Sander Münster)

Der Prozess der digitalen 3D-Rekonstruktion umfasst nicht nur die Erstellung des virtuellen Modells mithilfe von Softwarewerkzeugen, durchgeführt zumeist von spezialisierten Modelleuren, sondern auch deren anschließende Visualisierung, also die Übertragung des Modells in ein Präsentationsformat. Dieser Prozess wird zumeist eng von geschichtswissenschaftlicher Forschungsarbeit begleitet, bei der zumeist Historiker anhand von Quellen eine fundierte Vorstellung vom Modellierungsobjekt entwickeln (Münster et al., 2017). Entsprechende Herausforderungen für einen Umgang mit Wissen begründen sich in der interdisziplinären Zusammenarbeit und dem dafür nötigen Konsens der beteiligten Akteure über Sprache, Fragestellungen und Methodologie, der Zusammenführung der Wissensbestände zu einem gemeinsamen Ergebnis sowie mit Blick auf die vielfältigen Entscheidungen im Arbeitsprozess schließlich der Attribution von Autorschaften und die wissenschaftliche Qualitätssicherung. Im Panel sollen diese Aspekte sowohl problematisiert als auch mögliche Strategien vorgestellt und diskutiert werden.

Synthese

Infrastrukturelle Konzepte (Piotr Kuroczyński)

Die Entwicklung der I+K Technologien, allen voran hinsichtlich der Computergrafik, der webbasierten Anwendung und Vernetzung der Information, führt zu neuen Wissenszugängen für die objektbasierte Forschung, welche u.a. auf der Tagung *3D Digital Heritage – Exploring Virtual Research Space for Art History* näher beleuchtet wurden (Kuroczyński and Schelbert, 2017). Seit Mitte der 1990er Jahre konnten die ersten interaktiven 3D-Rekonstruktionen und webbasierte Präsentationsformen in der Forschung und Lehre vorgestellt werden (Frischer et al., 2008). Den objektbezogenen Fächern bieten sich heute eine Reihe von vielversprechenden Technologien an, welche neue infrastrukturelle Konzepte sowie neue Erschließung und Zugriff auf das Wissen mit sich bringen. Eine breite Anwendung finden *Game Engines*, die zum einen eine interaktive Darstellung der Ergebnisse, zum anderen eine Verknüpfung zu einer geisteswissenschaftlichen Datenbank ermöglichen (Clarke, 2016). Darüber hinaus werden Lösungen entwickelt, die Simulationen und eine Interaktion mit den Modellen in einer webbasierten Umgebung ermöglichen (Snyder, 2014). Hinsichtlich der Validierung, Interoperabilität und Nachhaltigkeit werden zurzeit verstärkt *virtuelle Forschungsumgebungen* getestet, die eine CIDOC CRM-referenzierte Datenmodellierung und Kontextualisierung der 3D-Modelle als Linked Data forcieren (Kuroczyński et al., 2016a, Bruschke and Wacker, 2016). Vielversprechend scheint darüber hinaus die Aneignung und Erweiterung von *Building Information Modelling* um die geisteswissenschaftlichen Fragestellungen rund um die Objekte. Der Impulsbeitrag möchte die technologisch getriebene Synthese der Information aus einer digitalen 3D-Rekonstruktion vor dem Hintergrund ausgewählter Beispiele beleuchten, um die Potenziale und Herausforderungen neuartiger Forschungsräume zu evaluieren.

Visualisierungs- und Vermittlungsansätze (Florian Niebling)

Die Visualisierung von Forschungsdaten dient nicht ausschließlich der Abbildung, sondern beinhaltet interaktive Explorationsprozesse die einen Erkenntnisgewinn oder ein Vermittlungsziel unterstützen oder häufig sogar erst ermöglichen (Niebling et al., 2017). Daher sind für die Visualisierung nicht nur Aspekte der Wahrnehmung, sondern insbesondere auch Methoden der Interaktion mit – und Darstellung von – Daten relevant, die einen tieferen Einblick in ansonsten nicht ersichtliche Informationen

ermöglichen. In der Darstellung von Ergebnissen einer 3D Rekonstruktion steht das grafische Endergebnis im Vordergrund, häufig sogar als alleiniges Resultat. Eine interaktive Präsentation oder Installation erlaubt es darüber hinaus, die in die Modellierung eingeflossenen Quellen – Fotografien, Pläne, Texte – erfahrbar zu machen, und damit wissenschaftliche Arbeitsschritte, Annahmen, Unsicherheiten oder alternative Betrachtungsweisen zu veranschaulichen. Diskutiert werden verschiedene Methoden der explorativen Darbietung rekonstruierter 3D Modelle, Desktop-basiert, in der Virtuellen Realität (VR) sowie in einer Verknüpfung aus Realität und virtuellen Inhalten in der Erweiterten Realität (AR). Im Zentrum stehen dabei Aspekte der Interaktion mit Forschungsdaten in den verschiedenen Modalitäten.

Bibliographie

Adams, J. L. / Gunn, K. B. (2013): *Keeping Up With...Digital Humanities.* American Library Association [Online], April 5, 2013.

Bruschke, J. / Wacker, M. (2016): *Simplifying the documentation of digital reconstruction processes. Introducing an interactive documentation system*, in: **Münster, S. / Pfarr-Harfst, M. / Kuroczyński, P. / Ioannides, M. (eds.):** *3D Research Challenges in Cultural Heritage II.* Cham: Springer LNCS.

Clarke, J. R. (2016): *3D Model, Linked Database, and Born-Digital E-Book: An Ideal Approach to Archaeological Research and Publication*, in: **Münster, S. / Pfarr-Harfst, M. / Kuroczyński, P. / Ioannides, M. (eds.):** *3D Research Challenges in Cultural Heritage II.* Cham: Springer International Publishing.

De Solla Price, D. (1963): *Little Science - Big Science,* New York, Columbia Univ. Press.

Frické, M. (2018): *Knowledge pyramid.*

Frischer, B. / Abernathy, D. / Guidi, G. / Myers, J. / Thibodeau, C. / Salvemini, A. / Minor, B. *Rome Reborn. ACM SIGGRAPH 2008 new tech demos on - SIGGRAPH '08 2008* Los Angeles, California. ACM Press.

Harari, Y. N. (2017): *Homo Deus: A Brief History of Tomorrow*, Vintage Penguin Random House.

Hessels, L. K. / Lente, H. V. (2007): *Re-thinking new knowledge production: A literature review and a research agenda,* Utrecht, Utrecht University.

Kuroczyński, P. / Grellert, M. / Hauck, O. / Münster , S. / Pfarr-Harfst, M. / Scholz, M. (2015): *Digitale Rekonstruktion und aktuelle Herausforderungen* (Panel). 2. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2015). Graz.

Kuroczyński, P. / Hauck, O. / Dworak, D. (2016a): *3D models on triple paths – New pathways for documenting and visualising virtual reconstructions*, in: **Münster, S. / Pfarr-Harfst, M. / Kuroczyński, P. / Ioannides, M. (eds.):** *3D Research Challenges in Cultural Heritage II.* Cham: Springer LNCS.

Kuroczyński, P. / Hauck, O. / Hoppe, S. / Münster, S. / Pfarr-Harfst , M. (2016b): *Der Modelle Tugend 2.0 – Vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell.*

Kuroczyński, P. / Pfarr-Harfst, M. / Wacker, M. / Münster, S. / Henze, F. (2014): *Pecha Kucha "Virtuelle Rekonstruktion – Allgemeine Standards, Methodik und Dokumentation"* (Panel). 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014). Passau.

Kuroczyński, P. / Schelbert, G. (2017): *3D Digital Heritage – Exploring Virtual Research Space for Art History* [Website zur Konferenz].

Long, M. P. / Schonfeld, R. C. (2014): *Supporting the Changing Research Practices of Art Historians*, Ithaka S+R.

Messemer, H. (2016): *The Beginnings of Digital Visualisation of Historical Architecture in the Academic Field*, in: **Hoppe, S., Breitling, S. (ed.)** *Virtual Palaces, Part II. Lost Palaces and their Afterlife. Virtual Reconstruction between Science and Media*. (= PALATIUMe-Publications 3) München, 21-54.

Messemer, H. (2018): *Entwicklung und Potentiale digitaler 3D-Modelle historischer Architektur – Kontextualisierung und Analyse aus kunsthistorischer Perspektive.* Dissertation, Ludwig-Maximilians-Universität München (unveröffentlicht).

Münster, S. / Jahn, P.-H. / Wacker, M. (2017): *Von Plan- und Bildquellen zum virtuellen Gebäudemodell. Zur Bedeutung der Bildlichkeit für die digitale 3D-Rekonstruktion historischer Architektur* , in: **Ammon, S. / Hinterwaldner, I. (eds.):** *Bildlichkeit im Zeitalter der Modellierung. Operative Artefakte in Entwurfsprozessen der Architektur und des Ingenieurwesens.* München: Wilhelm Fink Verlag.

Münster, S. / Terras, M. accepted paper. *The visual side of digital humanities. A survey on topics, researchers and epistemic cultures in visual digital humanities.* Digital Scholarship in the Humanities.

Niebling, F. / Münster, S. / Friedrichs, K. / Henze, F. / Kröber, C. / Bruschke, J. (2017): *Zugänglichkeit und dauerhafte Nutzbarkeit historischer Bildrepositorien für Forschung und Vermittlung* (Panel), in: **Stolz, M. (ed.):** *4. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2017).* Bern.

Nowotny, H. / Scott, P. / Gibbons, M. (2003): Introduction. *Minerva,* 41, 179–194.

Snyder, L. M. (2014): *VSim: Scholarly Annotations in Real-Time 3D Environments*, DH-CASE II: Collaborative Annotations on Shared Environments: metadata, tools and techniques in the Digital Humanities - DH-CASE '14. Fort Collins, CA, USA: ACM Press.

Deep Learning als Herausforderung für die digitale Literaturwissenschaft

Zusammenfassung

In den Digital Humanities (DH) wird seit einigen Jahren über einen „computational turn" (Berry 2011, Heyer 2014) diskutiert, der die aktuellen Algorithmen und Computertechniken des maschinellen Lernens und des tiefen Lernens stärker berücksichtigt. Beispiele sind das auf word embeddings basierende *word2vec* Modell, das darauf trainiert ist, sprachliche Zusammenhänge von Wörtern zu

rekonstruieren (Mikolov 2013), oder *fastText* und *GloVe* , die ebenfalls zur Textklassifikationen auf der Grundlage von word embeddings erstellt wurden. Diese Technologien haben sich in verschiedensten Bereichen der Verarbeitung von Bilddateien, Online-Datenbanken, der Erkennung natürlicher Sprache, prosodischer Daten und verschlüsselter Textdaten bewährt. Dennoch fehlt es in der Literaturwissenschaft und insbesondere der Textanalyse (Prosa, Lyrik, Drama) bis heute an einer Anwendung dieser neuen Methoden des maschinellen bzw. tiefen Lernens.

Bisher konzentrierten sich die klassischen „Digital Humanities" eher auf die Generierung und Reflexion digitaler Ressourcen wie Textausgaben, Repositorien oder Bilddatenbanken. Dagegen gibt es nur wenige Versuche, Deep Learning in die digitalen Geisteswissenschaften einzubringen. Zumeist wurde Deep Learning in sehr großen Datenbanken von Unternehmen wie Google, YouTube, Bluefin Labors oder Echonest getestet, etwa um Social Media Signale und den Inhalt von Medien in sozialen Netzwerken zu analysieren. Gerade deshalb blieb in diesem Feld die alte Kluft zwischen traditionellen Geisteswissenschaften und Informatik bestehen. Unser Panel will einen Beitrag leisten, um diese Lücke zu schließen.

Wir wollen vor allem die Probleme erörtern, die bei der rechnerischen Analyse literarischer Texte mit Techniken des tiefen Lernens entstehen, z.B.: Können maschinelle Lerntechniken durch Clustering tatsächlich verdeckte Muster in Textdaten erfassen (Graves 2012)? Wie lassen sich auf der Grundlage eines maschinell erlernten Modells Grenzfälle, Kategorisierungsfehler, Ausreißer und ähnliche Besonderheiten erkennen bzw. in den Klassifikationsprozess einbauen? Wie geht man mit dem großen Problem der "black box" um, wie lassen sich die in den "hidden layers" stattfindenden Klassifikationsprozesse nachvollziehen bzw. gar transparent machen? Und welche Tools für die manuelle (z.B. Sonic Visualizer) und automatische Annotation (z.B. PRAAT, ToBI, oder Sphinx) bzw. welche Softwares für die Modellierung (DyNet, TensorFlow, Caffe, MxNet, Keras, ConvNetJS, Gensim, Theano, und Torch) sind empfehlenswert?

Fragestellung und Aufbau des Panels

Wir haben auf unserem Panel Experten für computergestützte Analysen von literarischen Texten (Prosa, Lyrik, Drama) mit Interesse an vertieften Lerntechniken versammelt, die die computationale Analyse anhand von narrativen (Fokalisierung), dramatischen (Aktantenanalyse), poetischen (Metrik, Metaphorik, Reim) oder gattungsübergreifenden (Stilometrie, Topic Modeling) Textmerkmalen bereits umfangreich erprobt haben. Wir wollen uns vor diesem Hintergrund über bewährte Verfahren, praktische Anwendungen und erlernbare Methoden des Deep Learning austauschen, eine Plattform zur Präsentation und Entdeckung laufender Forschungsprojekte bieten, und die Vorteile und potenziellen Mängel der digitalen Mustererkennung auf der Grundlage der Methoden des tiefen Lernens reflektieren.

Folgende Personen haben dabei eine Teilnahme an dem Panel zugesagt:

- Fotis Jannidis ist Professor für Computerphilologie und Neuere Deutsche Literaturgeschichte an der Julius-Maximilians-Universität Würzburg.
- Christof Schöch ist Professor für Digital Humanities an der Universität Trier und Co-Direktor des Trier Center for Digital Humanities.
- Jonas Kuhn ist Professor für Maschinelle Sprachverarbeitung an der Universität Stuttgart und leitet das Centrum für Reflektierte Textanalyse (CRETA)
- Thomas Haider ist wissenschaftlicher Mitarbeiter am Max-Planck-Institut für empirische Ästhetik in Frankfurt am Main.
- Timo Baumann ist Informatiker am Language Technology Institute der Carnegie Mellon University in Pittsburgh, USA.
- Hussein Hussein ist Informatiker an der Freien Universität Berlin.
- Burkhard Meyer-Sickendiek ist Leiter einer von der Volkswagenstiftung geförderten Forschergruppe im Bereich der maschinellen Prosodieerkennung von Hörgedichten.

Wir werden in einem ersten Teil von maximal 30 Minuten einzelne Impulsvorträge präsentieren, und dann in einem zweiten Teil von ebenfalls 30 Minuten Topic Modeling und Embedding als wichtige Themenfelder des tiefen Lernens in den Geisteswissenschaften fokussieren. In einem dritten Teil von ebenfalls 30 Minuten wollen wir dann das Panel für die Diskussionen mit dem Publikum öffnen.

Panel-Vorträge

1. Thomas Haider vom Max Planck Institut Frankfurt wird über Topic Modeling am Beispiel der Latent Dirichlet Allocation (LDA) sprechen. Sein praktisches Beispiel ist die Anwendung der LDA auf einen Korpus neuhochdeutscher Poesie (Textgrid, mit 51k Gedichten, 5M-Token), um auffällige Themen in ihrem zeitlichen Verlauf (1575-1925 n. Chr.) und deren Verteilung ausfindig zu machen, und diese wiederum hinsichtlich der Klassifizierung von Gedichten in Epochen und mögliche Autorschaften auszuwerten. Die meisten Themen sind leicht interpretierbar und zeigen einen klaren historischen Trend. Um dennoch robustere Ergebnisse zu erhalten, plädiert Haider für eine verbesserte Methodik zur Berechnung der Wahrscheinlichkeit eines Themas bei einem bestimmten Zeitfenster. Für ein exploratives Experiment ist die Klassifizierung in Zeitfenster und Autorschaft zwar vielversprechend, für bessere Themenmodelle im Sinne klarer Trend- und Top-Themen wäre jedoch eine bessere Grundlage für die Modellierung diachroner Transformationen in der Poesie notwendig.
2. Fotis Jannidis von der Universität Würzburg wird dann darüber sprechen, wie man Techniken des Word Embeddings zur Modellierung literarischer Texte verwenden kann. Word Embeddings ermöglichen eine Repräsentation von Worten, die diese in eine gut interpretierbare Relation zueinander setzen: räumliche Nähe kann als semantische Ähnlichkeit verstanden werden. Mit Word Embeddings lassen sich zudem historische Entwicklungen von Worten anhand der historischen Veränderungen ihrer nächsten Nachbarn darstellen. Diese semantischen Netzwerke für die Ideengeschichte lassen sich auch im Rahmen literarischer Gattungen wie Prosa, Drama und Lyrik, etwa zum Vergleich verschiedener Themenfelder im Bereich der Lyrik nutzen. Interessant sind auch die Möglichkeiten zur Textrepräsentation: Während gängige bag-of-words-Modelle Texte über gewichtete Wortfrequenzen

repräsentieren, gehen diese Ansätze von n-dimensionalen Embeddings aus und erreichen eine Textrepräsentation durch Summierung oder Durchschnittsbildung u.a.m. Durch die rasante Entwicklung der Embedding-Verfahren – insbesondere im Jahr 2018 - ergeben sich hier ganz neue Möglichkeiten für die digitalen Geisteswissenschaften.

3. Anschließend wird Christof Schöch von jüngeren Anwendungen von Deep Learning in den Literaturwissenschaften berichten, die von verschiedenen Arbeitsgruppen durchgeführt wurden. Diese Anwendungen verwenden teils Word Embeddings für die interne Repräsentation der Texte, zeichnen sich darüber hinaus aber vor allem durch den Einsatz künstlicher neuronaler Netze mit einer vergleichsweise hohen Anzahl von Schichten aus (vgl. Goodfellow et al. 2016). Diese Architekturen sind für einen großen Teil der Merkmalsmodellierung und für die Lösung der jeweils in Frage stehende Klassifikationsaufgabe verantwortlich. Dabei reichen die Anwendungskontexte von Optical Character Recognition für historische Drucke (u.a. Wick et al. 2018) über linguistische Annotationen (vgl. Kestemont und De Gussem 2017) und die Erkennung von Figurenrede in Romanen (vgl. Jannidis et al. 2018) bis hin zur Autorschaftsattribution (vgl. Kestemont et al. 2016) und der Generierung von Gedichten (Hopkins und Kiela 2017). Während die Performanz der auf Deep Learning basierenden Ansätze häufig beeindruckend ist, werfen sie zugleich Fragen der Interpretierbarkeit und damit auch der Relevanz für die fachwissenschaftliche Theoriebildung auf.
4. Jonas Kuhn von der Universität Stuttgart wird das Verhältnis zwischen Deep Learning und theoriegeleiteter Forschung thematisieren. Im Zentrum stehen jüngere Erfolge mit Deep-Learning-Ansätzen, die auf der Grundlage eines sogenannten End-to-end-Training auf großen Mengen von realen Eingabe-/Ausgabepaaren (beispielsweise Maschinelle Übersetzung oder automatische Spracherkennung) basieren. In Hinblick auf eine Optimierung der Vorhersagequalität für diese Aufgaben werden klassische theoretische (Zwischen-)Konstrukte obsolet (für die Übersetzung etwa eine symbolische Repräsentation der syntaktischen Struktur oder der wörtlichen bzw. einer kontextuell implizierten Bedeutung). Aber werden theoretische Analysemodelle (typisch umgesetzt etwa im maschinellen Lernen auf handannotierten Trainingsdaten) damit insgesamt obsolet? Dies hängt davon ab, ob man eine Systemoptimierung für spezifische Aufgaben, oder aber ein wissenschaftliches Erkenntnisinteresse verfolgt. Gleichwohl stellt sich gerade bei Forschungsfragen der Digital Humanities, die oftmals unterschiedliche etablierte Beschreibungsebenen überspannen, die Frage der Motivation für etwaige symbolische Zielkategorien ganz neu.
5. Burkhard Meyer-Sickendiek und Hussein Hussein von der Freien Universität Berlin sowie Timo Baumann von der CMU Pittsburgh werden über character embeddings in der Gedichtanalyse sprechen. Kann man prosodische bzw. rhythmische Muster in moderner Lyrik mittels neuronaler Netztechniken auf der Grundlage von embeddings identifizieren? Für solche Klassifikationen werden feature-basierte Verfahren und Techniken tiefen Lernens vergleichend untersucht (vgl. Escudero et al. 2017). Dem merkmalsbasierten Ansatz, der auf Lyriktheorien im Bereich der freien Versprosodie basiert, wird dabei ein neuronaler Ansatz gegenübergestellt, anschließend wird die Integration von höherem Wissen in das neuronale Netzwerk erläutert. Unser Projekt arbeitet mit character embeddings auf der Grundlage bidirektionaler, rekurrenter neuronaler Netzwerke (RNN, unter Verwendung von Gated Recurrent Unit (GRU)-Zellen (Cho et al. 2014)), die die Sequenz von Zeichen in eine mehrdimensionale Darstellung überführen. Es trainiert also nicht auf Wortebene, kann jedoch Informationen (z.B. Wortteile) über die konstituierenden Zeichenfolgen dekodieren. Dabei werden wir über Techniken des End-to-End-Lernens (Hannun et al., 2014; Graves and Jaitly, 2014) sprechen, die z.B. in der Spracherkennung verwendet werden, die weder Phoneme noch Wörter explizit modelliert, sondern direkt Audio-Features auf Charakter-Streams überträgt.

Diskussion

Mögliche Themengebiete für die Paneldiskussion im dritten und letzten Teil wären insbesondere die Verwendung tiefer Lernverfahren in den digitalen Geisteswissenschaften, etwa mit Blick auf Stilometrie, Computerstilistik, Reim- und Metrikenanalyse, Aktantenanalyse, oder Themenmodellierung. Dabei soll die Publikumsdiskussion all jenen ein Forum bieten, die sich ein tieferes Verständnis und eine praktische Schulung in deep learning sowie eine Plattform für den Austausch von Praktiken, Ergebnissen und Erfahrungen im Umfeld mit einschlägigen Tools erhoffen. Dies kann sich auch auf Kenntnisse aus den Nachbardisziplinen erstrecken, insofern diese über bereits vorhandenes Wissen hinsichtlich der Anwendung „tiefer Lerntechniken“ etwa im Bereich des Data Mining, der Statistik oder der Verarbeitung natürlicher Sprache verfügen. Auf diese Weise erhoffen wir uns eine effektive Fokusverlagerung innerhalb der digitalen Geisteswissenschaften: von der Erstellung und Archivierung digitaler Artefakte und Repositorien hin zu echten Rechenlösungen auf der Grundlage maschinellen Lernens.

Bibliographie

Ananthakrishnan, Sankaranarayanan / Narayanan, Shrikanth S. (2008): *“Automatic prosodic event detection using acoustic, lexical, and syntactic evidence”* IEEE Transactions on Audio, Speech, and Language Processing, 16(1), pp. 216-228.

Berry, David M. (2011): *“The Computing Turn: Thinking About the Digital Humanities”*, in: Culture Machine. 12, 1-22.

Brett, Megan R. (2012): *“Topic Modeling: A Basic Introduction. Journal of Digital Humanities”.*

Cho, Kyunghyun / Merrienboer, Bart van / Gulcehre, Caglar / Bahdanau, Dzmitry / Bougares, Fethi / Schwenk, Holger / Bengio, Yoshua (2014): *“Learning phrase representations using RNN encoder-decoder for statistical machine translation”*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language.

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): *“Stylometry with R: a suite of tools”*. In: Digital Humanities, Conference Abstracts, University of Nebraska, Lincoln, 487-89.

Escudero, David / González, César / Gutiérrez, Yurena / Rodero, Emma (2017): *"Identifying characteristic prosodic patterns through the analysis of the information of Sp_ToBI label sequences"*. Computer Speech & Language 45: 39-57.

Goodfellow, Ian / Bengio, Yoshua / Courville, Aaron (2016): *"Deep Learning"*. Cambridge, Massachusetts: The MIT Press.

Graves, Alex (2012): *"Supervised sequence labelling with recurrent neural networks"* (Vol. 385). Springer.

Graves, Alex / Jaitly, Navdeep (2014): *"Towards end-to-end speech recognition with recurrent neural networks"*. In International Conference on Machine Learning, pages 1764–1772,.

Hannun, Awni / Case, Carl / Casper, Jared / Catanzaro, Bryan / Diamos, Greg / Elsen, Erich / Prenger, Ryan / Satheesh, Sanjeev / Sengupta, Shubho / Coates, Adam (2014): *"Deep speech: Scaling up end-to-end speech recognition"*. arXiv preprint arXiv:1412.5567.

Hasegawa-Johnson, Mark / Chen, Ken / Cole, Jennifer / Borys, Sarah / Kim, Sung-Suk / Cohen, Aaron / Zhang, Tong / Choi, Jeung-Yoon / Kim, Heejin / Yoon, Taejin (2005): *"Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus"*. Speech Commun 46: 418–439, 2005.

Heyer, Gerhard (2014): *"Digital and computational humanities"*, in: http://dhd-wp.hab.de/files/book_of_abstracts.pdf , pp.66f.

Hopkins, Jack / Douwe Kiela (2017): *"Automatically Generating Rhythmic Verse with Neural Networks"*. ACL. http://dx.doi.org/10.18653/v1/P17-1016

Hsu, Chih-Wei / Lin, Chih-Jen (2002): *"A comparison of methods for multi-class support vector machines"*. IEEE Transactions on Neural Networks, 13:415-425.

Jannidis, Fotis / Konle, Leonard / Zehe, Albin / Hotho, Andreas (2018): *"Analysing Direct Speech in German Novels"*. DHd Jahrestagung. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf

Jeon, J. / Liu, Y. (2009): *"Semi-supervised learning for automatic prosodic event detection using co-training algorithm"*. In Proc. of the 47th Annual Meeting of the ACL, pp. 540-548.

Kestemont, Mike / De Gussem, Jeroen (2017): *"Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning"*. Journal of Data Mining & Digital Humanities. http://arxiv.org/abs/1603.01597 .

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): *"Efficient estimation of word representations in vector space"*. arXiv preprint. arXiv:1301.3781.

Moretti, Franco: *Distant reading*. Verso 2013.

Schöch, Christof: *Quantitative Semantik: Word Embedding Models für literaturwissenschaftliche Fragestellungen*. In Digitale Literaturwissenschaft, hrsg. Fotis Jannidis. Stuttgart: Metzler (im Druck).

Wick, Christoph / Reul, Christian / Puppe, Frank: Calamari (2018): *"A High-Performance Tensorflow-Based Deep Learning Package for Optical Character Recognition."* ArXiv:1807.02004 [Cs] http://arxiv.org/abs/1807.02004 , 2018.

Digital Humanities "from Scratch" Herausforderungen der DH-Koordination zwischen Querschnittsaufgaben und "one-(wo)man-show"

[Abstract] Immer häufiger schaffen universitäre, akademische und andere wissenschaftliche Institutionen neue Stellen, um Aktivitäten im Bereich der Digital Humanities zu koordinieren und Infrastrukturen aufzubauen. Beginnt man heute damit ganz von vorn, sei es im Rahmen einer Querschnittsstelle, einer neuen Abteilung oder auch ohne offizielles Mandat, steht man vor anderen Voraussetzungen als vor einem Jahrzehnt. Während einige wenige Standorte inzwischen auf eine langjährig gewachsene DH-Infrastruktur zurückblicken können, stellt die Etablierung von DH-Strukturen "from scratch" im laufenden Betrieb hohe institutionelle, organisatorische, personelle und technische Anforderungen. Übergreifender Austausch, Vernetzung und Bündelung von Initiativen und Aktivitäten jenseits von Institutionen und Verbünden stehen jedoch bislang aus. In diesem moderierten Panel stellen die Beitragenden, die durch ihre Institutionen mit solchen Aufgaben betraut wurden oder diese initiieren, ihre bisherigen Herausforderungen, Erfahrungen und Lösungsansätze zur Diskussion, um aktuelle Desiderate zu identifizieren und sowohl Handlungsoptionen als auch mögliche Strategien aufzuzeigen – und damit einen Impuls für weiterführende Initiativen zu geben.

Einleitung

Wenngleich Zentrumsbildung als ein Merkmal der Institutionalisierung der DH gilt (Sahle 2015), findet DH auch abseits solcher "Zentren" statt. Wie aktuelle Ausschreibungen zeigen, entsteht derzeit Koordinationsbedarf an vielen Standorten, die erst jetzt Strukturen aufbauen können oder wollen. Während auf Projektebene bereits Erfahrungsberichte und Best Practices vorliegen (Pitti 2004), wird die organisatorisch-strukturelle bzw. institutionelle Dimension in Deutschland, Österreich und der Schweiz zwar im bibliothekarischen Kontext (Maier 2016), aber nur selten aus DH-Sicht reflektiert (im anglo-amerikanischen Raum: Posner 2016, Anne et al. 2017).

Zudem basiert DH auf Interdisziplinarität und Kollaboration und resultiert in Koproduktionen, die nicht mehr auf eine Einzelperson zurückführbar sind (Unsworth 1997). In Kombination mit der methodologisch angelegten Überwindung der Disziplinengrenzen entsteht ein Spannungsfeld, das kommunikative Prozesse vor neue Herausforderungen stellt (Edmond 2016). Auch dies wird in den DH kaum erörtert (Griffin / Hayler 2018), obwohl Koordinationsbedarf sowohl in einzelnen Forschungsprojekten als auch in Institutionen selbst besteht.

Konzeption des Panels

Dieses Panel eröffnet ein Diskussionsforum für Akteure, die mit dem Aufbau von DH-Abteilungen, DH-Infrastrukturen oder DH-Querschnittsaufgaben befasst sind. Im Vordergrund stehen Strategien und Methoden für einen reflektierten Umgang mit standortspezifischen Voraussetzungen. Anhand eines Vergleichs der vorgestellten Strukturen, Strategien und Erfahrungen der Teilnehmenden werden grundlegende technische, organisatorische und strukturelle Herausforderungen herausgearbeitet, um den Anstoß für eine breitere Diskussion der Anforderungen und Potenziale aktueller Entwicklungen zu geben.

Das Panel bildet die Heterogenität der Akteure im deutschsprachigen Raum exemplarisch ab. Leitend für die Zusammenstellung ist die Diversität institutioneller Formen hinsichtlich Größe, Ausrichtung und Auftrag sowie als Grundbedingung der konzeptionelle Aufbau einer DH-Struktur "from scratch". Werden im aktuellen Diskurs bislang zuvorderst bereits bestehende DH-Infrastrukturangebote wie diejenigen der zwei großen Verbundprojekte DARIAH-DE und CLARIN, von etablierten geisteswissenschaftlichen Datenzentren oder der im Aufbau begriffenen Nationalen Forschungsdaten-Infrastruktur (NFDI) in den Blick genommen, gilt demgegenüber der Situation von "One-Person-Digital-Humanists" hier besondere Aufmerksamkeit. Das Panel bietet ein Forum für die auf der DHd-Konferenz vertretene Gruppe solcher DH-Koordinator*innen, auch solcher avant la lettre oder ohne explizites Mandat.

Das Panel beginnt mit kurzen Statements der Beitragenden, die ihre Ausgangslage skizzieren und problematisieren ('30). Die nachstehenden Leitfragen werden dann, geleitet vom Moderationsteam, gemeinsam mit Panel und Plenum diskutiert ('45). Abschließend werden im Plenum Rückmeldungen zur Bearbeitung des Themenfeldes eingeholt und Möglichkeiten für anschließende Aktivitäten ausgelotet ('15).

Das Autorenteam strebt die Ergebnissicherung in Form eines Blog-Berichtes an und formuliert gleichzeitig das Angebot, weiterführende Vernetzungs- und Vermittlungsaktivitäten durch Workshops, Arbeitsgruppen oder Publikationen anzustoßen und zu unterstützen.

Leitfragen

Die Verortung verstehen

Die institutionelle Verortung definiert unmittelbar Arbeitsweise und Tätigkeitsfelder einer koordinierenden DH-Stelle. Die Heimatinstitution bestimmt den Zugang zu personellen, finanziellen und infrastrukturellen Ressourcen. Die Positionierung im Organigramm hat wiederum wesentlichen Einfluss auf die Handlungsoptionen innerhalb der Einrichtung und der gesamten Organisation. Der Anspruch der DH, "querliegend" zu operieren, changiert auf institutioneller Ebene zwischen struktureller Isolation und Entgrenzung. Welche Auswirkungen haben Zuordnungen zu Verwaltung, Fachbereich, Lehrstuhl, Bibliothek, Forschungsinstitut, Verbundprojekt und Bezeichnungen wie Abteilung, Koordinationsbüro, Stabsstelle?

Die Kompetenzen abdecken

DH-Kompetenz ist in diversen Spezialisierungen sowohl kontinuierlich als auch temporär aufzubauen, heranzuziehen oder auch zusammenzuführen. Längerfristige Perspektiven für DH-Personal tun sich nur unter besonderen Umständen auf (vgl. Boyles et al. 2018); ebenfalls selten sind Werkverträge mit dedizierten DH-Dienstleistern. Kooperationen mit anderen Institutionen, die über DH-Zentren verfügen, oder mit übergreifenden Infrastrukturprojekten wie CLARIN/DARIAH bieten sich an, existieren aber nicht als standardisierte Vorgänge und decken die Anforderungen nur partiell ab. Welche Expertise sollte eine DH-Koordination mitbringen? Sind Aufteilungen in eine wissenschaftliche sowie eine technische DH-Koordination wünschenswert? Wie können Kompetenzen ausgelagert und wieder eingeholt werden?

Die Aufgaben definieren

DH-Koordination steht anfangs vor der Aufgabe, den Status Quo an ihrer Institution zu kartieren. Welche Aktivitäten, welche Expertise, welche Infrastrukturen bestehen bereits? Welche Projekte und Akteure sind einzubeziehen, wo sind (womöglich fächerspezifische) Lücken und Bedarfe im Hinblick auf Technik, Kompetenzverteilung, Empfehlungen, Best Practices und Policies identifizierbar? Bestehen institutionelle Angebote benachbarter, DH-relevanter Themenfelder wie Forschungsdatenmanagement, Digitalem Publizieren und Open Science? Welche Arbeitsschritte sind wann zu priorisieren, und sind z.B. modulare Konzepte solchen vorzuziehen, die eine institutionelle Gesamtstrategie verfolgen? Wie umfänglich kann und soll Beratung geleistet werden? Welche Instrumente und Formate stehen zur Kompetenzvermittlung zur Verfügung?

Die Fachforschung erreichen

Wissenschaftsgeleitete, forschungsgetriebene Angebote, wie sie DH-Koordinationen und -zentren leisten sollen, basieren auf der fundierten Kenntnis fächerspezifischer Bedarfe, Methoden und Forschungspraktiken. Auf der Grundlage fachwissenschaftlicher Forschungsfragen und im Verbund mit Projekten, Vorhaben und Gremien lassen sich bedarfsgerechte Angebote (Software, Dienste, Beratung) entwickeln und etablieren. Zugleich sind experimentelle Formate wünschenswert, die im Sinne eines "Lab" die Tauglichkeit neuer Angebote evaluieren. Jedoch muss der Kontakt zur Fachforschung aktiv initiiert und etabliert werden: Wie kann die Zielgruppe erreicht, eine Außendarstellung kommuniziert, können Akteure untereinander vernetzt, eine Community aufgebaut, DH als Service etabliert werden? Welche Modelle und Formate bieten sich an: DH-Workshops, Vorlesungsreihen oder fächerübergreifende Arbeitsgruppen?

IT und Bibliothek mitnehmen

Die DH verändern das traditionelle Verhältnis zwischen Fachwissenschaft, IT/Rechenzentrum und Bibliothek.

Deren Aufgabenfelder greifen an mehreren Stellen ineinander und überlagern sich, sind jedoch von unterschiedlichen Aufgaben und Mandaten geprägt. Während die IT die Grundversorgung mit EDV und Netzinfrastruktur stellt, benötigen die DH spezifische Entwicklungsumgebungen und Webapplikationen. Ebenso fordern sie bei Bibliotheken informationswissenschaftliche Expertise, spezielle Digitalisierungsverfahren sowie Datenmodellierung, -archivierung und -kuration an, während bisher bibliografische Erfassung und Literaturbeschaffung das Arbeitsfeld dominierten (Maier 2016). Die neuen Aufgaben übersteigen nicht selten die inhäusigen Kapazitäten und Kompetenzen. Wie können IT und Bibliothek in den Aufbau von DH-Strukturen integriert werden? Was ist bei der Nutzung außerhäusiger Dienste zu berücksichtigen?

Den Wandel begleiten

In einigen Forschungsprojekten gelingt unter engen thematischen, organisatorischen und personellen Rahmenbedingungen mitunter die Integration neuer digitaler Methoden oder Praktiken. Doch lassen sich diese Erfahrungen auf die institutionelle Ebene skalieren, z.B. für Forschungseinrichtungen, SFBs oder Exzellenzcluster? Wenn sich Arbeitsabläufe und Forschungsprozesse an einer Einrichtung grundsätzlich oder dauerhaft ändern sollen, werden ein institutionelles Veränderungsmanagement und eine Nachhaltigkeitsstrategie notwendig. Dieser Aspekt ist im Fachdiskurs der DH bisher nicht präsent. Dennoch fällt den koordinierenden DH-Stellen entweder diese Aufgabe zu oder wird durch das Aufgabenfeld notwendig. “The ultimate function of the digital humanities center at the present time, then, is to be an agent of change.” (Freistat 2012)

Beitragende (alphabetisch)

Swantje Dogunke (Weimar)

Seit dem Jahr 2013 bündeln das Deutsche Literaturarchiv Marbach, die Klassik Stiftung Weimar und die Herzog August Bibliothek Wolfenbüttel ihre Forschungsaktivitäten auf Empfehlung des Wissenschaftsrats in einem Verbund, der vom Bundesministerium für Bildung und Forschung gefördert wird. Ziele des Verbunds sind neben der Erforschung der Bestände der drei Einrichtungen der Aufbau einer gemeinsamen Forschungsinfrastruktur, um die digitalen Sammlungen der Institutionen bestandsübergreifend durchsuchbar zu präsentieren und in projektspezifischen Arbeitsumgebungen Tools und Services zu deren Beforschung bereitzustellen.

Frederik Elwert (Bochum)

Das Centrum für Religionswissenschaftliche Studien (CERES) ist eine Zentrale Wissenschaftliche Einrichtung der Ruhr-Universität Bochum. Seit 2010 ist CERES an Drittmittelprojekten in den Digital Humanities beteiligt. Seit 2016 besteht eine entsprechende Koordinationsstelle, die Projekte begleitet, bei Anträgen berät und Schulungen durchführt. An der Ruhr-Universität gibt es derzeit keine dedizierte DH-Infrastruktur, allerdings bietet die neu eingerichtete AG Forschungsdatenmanagement Anknüpfungspunkte. Dies verspricht auch Verbesserungen für die noch unbefriedigende Nachhaltigkeit der DH-Projekte.

Harald Lordick (Essen)

Das Steinheim-Institut erforscht deutsch-jüdische Geschichte in breitem Spektrum. Rare, weltweit verstreute Quellen, das "Setting" einer kleinen, außeruniversitär und interdisziplinär arbeitenden Einrichtung mit spezifischen Anforderungen sind Antrieb jahrzehntelangen Engagements, computergestützte Verfahren für geisteswissenschaftliches Tun nutzbringend anzuwenden. Zunächst eher auf jeweils aktuelle Bedarfe wissenschaftlichen Arbeitens, Recherche und Publikation bezogen, entstehen zunehmend auch forscherische, digital-analytische Perspektiven. Aus diesen Erfahrungen, mit einer gewachsenen digitalen Strategie und ohne dedizierte DH-Stellen beteiligen sich MitarbeiterInnen aktiv an Infrastrukturbestrebungen wie DARIAH-DE, GND und Community-getriebenen Linked-Data-Initiativen.

Torsten Roeder (Halle)

Als die Leopoldina 2008 zur Nationalen Akademie der Wissenschaft ernannt wurde, entstand das Bedürfnis nach einer Einrichtung, an der die traditionellen und neuen Aufgaben der Akademie historisch und theoretisch reflektiert werden. Vor diesem Hintergrund wurde 2012 das Leopoldina-Studienzentrum gegründet, um die wissenschaftshistorischen, wissenschaftstheoretischen und wissenschaftsphilosophischen Aktivitäten der Akademie zu koordinieren. 2018 wurde eine Referentenstelle mit Schwerpunkt Digital Humanities geschaffen, um die eigenen Forschungsvorhaben und diversen Drittmittelprojekte auf eine digital konsistente Basis zu stellen und ein DH-Gesamtkonzept für das Studienzentrum zu erarbeiten.

Sibylle Söring (Berlin)

Seit 2016 weitet die Freie Universität Berlin ihre Angebote im Kontext infrastruktureller und serviceorientierter Unterstützung (geistes-)wissenschaftlicher Forschung aus. Im Zentrum des Aufbaus einer DH-Infrastruktur stehen die Beratung geisteswissenschaftlicher Forschungsvorhaben bei der Entwicklung und Umsetzung einer digitalen Strategie, der Aufbau von Infrastrukturen für das geisteswissenschaftliche Forschungsdatenmanagement sowie die Nutzbarmachung relevanter Technologien und forschungsnaher Dienste. Ziel ist der Aufbau universitätseigener Strukturen und Lösungen. Dabei werden u.a. die Ergebnisse INF-Teilprojekts des SFB 980 “Episteme in Bewegung” nachgenutzt.

Thorsten Wübbena (Paris)

Die Abteilung Digital Humanities am Deutschen Forum für Kunstgeschichte Paris wurde Ende 2014 eingerichtet und arbeitet an der Schnittstelle zwischen kunstgeschichtlicher Forschung und IT. Neben der Durchführung eigener Forschungsarbeit auf diesem Gebiet versteht sie sich als

Ansprechpartnerin für alle Abteilungen des Hauses bei Fragen rund um Digitalität in der Forschung. Dazu gehören sowohl die Konzeption digitaler Editionen, die Digitalisierung und Pflege geisteswissenschaftlicher Forschungsdaten als auch deren Analyse. Die Verknüpfung mit den internationalen Entwicklungen der Forschungsgemeinschaft geschieht direkt aus den Forschungsvorhaben und Kollaborationen heraus.

Moderation

Fabian Cremer, Referent, Geschäftsstelle der Max Weber Stiftung, Bonn

Anne Klammt, Geschäftsführerin mainzed, Hochschule Mainz

Bibliographie

Anne, Kirk M., et al. (2017): *Building Capacity for Digital Humanities: A Framework for Institutional Planning*, ECAR working group paper. Louisville, CO: ECAR, https://library.educause.edu/~/media/files/library/2017/5/ewg1702.pdf [zuletzt abgerufen 27.09.2018].

Boyles, Christina et al. (2018): *Precarious Labour in the Digital Humanities.* Panel, ADHO DH2018, Mexico DF, https://dh2018.adho.org/precarious-labor-in-the-digital-humanities/ [zuletzt abgerufen 27.09.2018].

Edmond, Jennifer (2016): *Collaboration and Infrastructure*, in: **Schreibman, Susan et al. (eds.):** *A New Companion to Digital Humanities*, 2nd Edition, Malden, MA: Wiley-Blackwell 54-65.

Freistat, Neil (2012): *The function of digital humanities centers at the present time*, in: Gold, Matthew K. (ed.): Debates in the Digital Humanities, http://dhdebates.gc.cuny.edu/debates/text/23 [zuletzt abgerufen 27.09.2018].

Griffin, Gabriele / Hayler, Matt Steven (2018): *Collaboration in Digital Humanities Research – Persisting Silences*, in: Digital Humanities Quarterly 12(1), http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html [zuletzt abgerufen 27.09.2018].

Maier, Petra (2016): *Digital Humanities und Bibliothek als Kooperationspartner,* in: DARIAH-DE Working Papers 19. Göttingen, urn:nbn:de:gbv:7-dariah-2016-5-6.

Pitti, Daniel. V. (2004): *Designing Sustainable Projects and Publications*, in: **Schreibman, Susan et al. (eds.):** *A companion to digital humanities*, Oxford: Blackwell 31, http://www.digitalhumanities.org/companion/ [zuletzt abgerufen 27.09.2018].

Posner, Miriam (2016): *Here and There: Creating DH Community*, in: **Gold, Matthew K. (ed.):** *Debates in the Digital Humanities*, 2016 Edition, http://dhdebates.gc.cuny.edu/debates/text/73 [zuletzt abgerufen 27.09.2018].

Sahle, Patrick (2015): *Digital Humanities? Gibt's doch gar nicht!*, in: **Baum, Constanze / Stäcker, Thomas (eds.):** *Grenzen und Möglichkeiten der Digital Humanities*, https://dx.doi.org/10.17175/sb001_004 .

Unsworth, John (1997): *Creating Digital Resources: the Work of Many Hands*, Vortrag am 14.09.1997, Digital Resources for the Humanities, Oxford, England, http://www.people.virginia.edu/~jmu2m/drh97.html [zuletzt abgerufen 27.09.2018].

Multimodale Anreicherung von Noteneditionen: Edirom und Datenbank

In den vergangenen Jahren hat man in der Musikwissenschaft zunehmend den traditionellen und vor allem im 19. Jahrhundert konturierten Werkbegriff insbesondere für Musik, die vor 1800 entstand, in Frage gestellt. Das Verständnis von Werken als unveränderbare Entitäten, die als Geniestreich des Komponisten eine fixe Gestalt besitzen, wurde aufgegeben zugunsten einer Sichtweise, die ein Werk als Produkt zeitgenössischer Performanz und kultureller Praxis versteht. Die von Aufführung zu Aufführung wechselnde Gestalt eines Werks lässt es eher als ständiges „work in progress" erscheinen (Calella 2014, 2007; Goehr 2007; Over i.Dr.). Projekte wie *A Cosmopolitan Composer in Pre-Revolutionary Europe – Giuseppe Sarti* (www.sarti-edition.de) , das die Modularität von zwei Sarti-Opern (*Fra i due litiganti il terzo gode*, *Giulio Sabino*) editorisch darstellt (vgl. auch Albrecht-Hohmaier/Siegert 2015; Herold/Siegert 2016), oder *Beethovens Werkstatt* (https://beethovens-werkstatt.de) , das Aspekte der Werkgenese digital aufarbeitet, versuchen, diesen Gegebenheiten gerecht zu werden.

Das Anfang des Jahres gestartete, polnisch-deutsche, auf drei Jahre angelegte und von DFG und NCN (Narodowe Centrum Nauki) finanzierte Forschungsprojekt *PASTICCIO. Ways of Arranging Attractive Operas* der Universitäten Warschau und Mainz untersucht das Opernpasticcio. Ähnlich der seit dem 19. Jahrhundert beliebten Potpourris oder Revuen im frühen 20. Jahrhundert griff man im Pasticcio auf musikalisches Material zurück, das sich in vielerlei Hinsicht bereits bewährt hatte, und stellte daraus ein „Werk" zusammen: Arien, die sich einer gewissen Popularität beim Publikum erfreuten, mit denen Sänger Erfolge gefeiert hatten, die aus dem Repertoire eines Sänger-Stars stammten, die einen neuen Stil präsentierten usw. Die zentrale These lautet, dass das Pasticcio als kulturelle Praxis entscheidende europaweite Sozialitäts- und Autoritätsvorstellungen mit Blick auf Komponisten, Sängerinnen und Sänger, Librettistinnen und Librettisten sowie Impresari in der ausgehenden Frühen Neuzeit widerspiegelt. Diese Vorstellungen sind im Spannungsfeld einer aktiven Kenntnis sowie Verwendung des europäischen Musikrepertoires und der frühneuzeitlichen Subjektkonstitution stetigen Veränderungen unterworfen. Um die sich daraus entwickelnden Autorschafts- und Stilkonzepte als transregionale Prozesse und entstehende Normen erforschen zu können, werden digitale Editionen von drei Pasticci und einer Oper mit einer kulturwissenschaftlichen Datenbank verbunden, in der Daten zu Reisewegen, Transfermechanismen, Netzwerken, musikalischen Institutionen, den beteiligten Akteuren, der Materialität der Quellen, musikalischen Vorlagen sowie textlichen und musikalischen Bearbeitungen gesammelt sind.

Der Zuschnitt des Projekts geht von der aktuellen Forschungslage aus, dass Opernpasticci Produkte einer

multiplen Autorschaft sind. Sowohl Sängerinnen und Sänger forderten die Einlage ihrer sogenannten Kofferarien, anhand derer sie ihre individuellen virtuosen Fertigkeiten bestmöglich demonstrieren konnten, als auch Impresari, Agenten und der lokale Adel den Einbau von Arien oder die Vertonung bekannter Libretti vorschlugen (Strohm 2011; Brandt 2002; Holmes 1993; Freeman 1992). Komponisten nutzten das Pasticcio unter anderem, um zu Beginn der Stagione ihr Sängerensemble vorzustellen, um sich mit Opernkompositionen aus Italien auseinanderzusetzen oder um mit wenig Zeitaufwand neue Produktionen fertigzustellen (Strohm 2009a, 2009b). Innerhalb dieses Zuschnitts ist die frühneuzeitliche Mobilität von Musikerinnen und Musikern von eminenter Wichtigkeit, denn gerade durch sie kamen SängerInnen-orientierte Opernproduktionen sowie eine europaweite Zirkulation von Musik zustande (zur Nieden 2015; Burden 2013; Korsmeier 2000; Woyke 1998; LaRue 1995). Die Verbreitung von Musik durch Pasticci unterliegt somit bestimmten ästhetischen wie sozialen Grundvorstellungen, die von einer Orientierung an der Modellkultur Italien bis hin zu lokalen Konjunkturen bestimmter, sich in Musik spiegelnder Soziabilitätsmodelle reichen. Ziel ist es, die Erarbeitung eines ästhetischen Verständnisses des Arbeitens mit präexistentem Material als Praxis auf die Konturierung musikalischer Stile und Autorschaftsfunktionen im 18. Jahrhundert in ihrem Wandel zurückzuspiegeln. Auf diese Weise ist es nicht nur möglich, die Erforschung des Opernpasticcios stärker in die interdisziplinäre Forschung zu Pastiche, Burlesque, Parodie oder Capriccio einzubinden, sondern auch, den bisherigen Akzent auf dem Opernpasticcio als Produkt oder Werk stärker auf seine praxisrelevanten Dimensionen des Vergleichens, Transkribierens oder Kopierens zu legen (zu pasticcio-ähnlichen Arbeitsweisen in anderen Disziplinen vgl. etwa Décout 2017; Kanz 2002).

Um die dargestellten kulturwissenschaftlichen Dimensionen der Pasticcio-Produktion in der Edition von drei Opern-Pasticci und einer Oper abzubilden, fließen neben der Integration von digitalisierten Quellen (Partituren, Manuskripte von Einzelarien, Libretti) kulturwissenschaftliche Daten aller Art ein. Diese umfassen hauptsächlich Informationen zu Werken, ihren Vorlagen, den Ursprungswerken, Sängerkarrieren und -itineraren, beteiligten Akteuren wie Impresari, Widmungsträger und Publikum. Diese Informationen sind einerseits in einer Datenbank recherchierbar und können andererseits in der Edition bei den einzelnen das Pasticcio konstituierenden Bestandteilen abgerufen werden.

Die Aufbereitung der Daten erfolgt anhand des FRBR-Modells (Tillett 2005). Dieses für den Bibliotheksbedarf entwickelte Modell (Functional Requirements for Bibliographic Records) bietet den Vorteil einer quasi neutralen Strukturierung von Daten, die sich um einen Autor und ein Werk sammeln. FRBR strukturiert Daten u. a. in „works“, „expressions“, „manifestations“, „items“, „persons“ und „corporate bodies“. Adaptiert auf das Projekt werden in Bezug auf das „work“ einerseits inhaltliche Dimensionen („work“ und „ work component“), andererseits Abhängigkeitsverhältnisse („work“ und „derivative“, wie etwa ein Arrangement) dargestellt. „Expressions“ zeigt die performative Ebene auf. Hier werden etwa alle Produktionen erfasst. „Manifestations“ fasst die materialen Aspekte zusammen, d. h. die verschiedenen Arten von Quellen (Partitur, Einzelarie, Libretto u. ä.). Jede einzelne Quelle wird als „item“ bezeichnet. „Persons“ und „corporate bodies“ (etwa Institutionen wie Opernhäuser) verstehen sich von selbst.

Das Panel nähert sich der Einbindung der Datenbank in die Edition aus verschiedenen Perspektiven. Einerseits thematisiert es das Erkenntnisinteresse, das mit der Integration der Datenbank verbunden ist (Gesa zur Nieden). Andererseits wird aus Quellen- und Datensicht die Integration beleuchtet (Berthold Over). Denn während die Anreicherung von Editionen durch Quellen bereits seit Jahren in Edirom praktiziert wird, ist eine weitergehende Anreicherung durch Daten Neuland. Die Herausforderungen der Pasticcio-Edition (Martin Albrecht-Hohmaier) führt schließlich zur Zusammenführung von Datenbank und Edition (Kristin Herold), die als eines der Projektziele integraler Bestandteil des Projekts bildet und in ihrer Innovativität Pilotcharakter für künftige Projekte hat. Eine kritische Beurteilung der unterschiedlichen Perspektiven erfolgt durch Christine Siegert, die durch eigene Projekte als ausgewiesene Spezialistin auf dem Gebiet der digitalen Edition und der Instabilität frühneuzeitlicher musikalischer „Werke“ gelten muss. Als Diskutantin gibt sie ein direktes Feedback auf die Impulsreferate und regt durch ihre Statements zur weiteren Diskussion im Plenum an.

Christine Siegert (Beethoven-Haus Bonn)

Diskutantin

Gesa zur Nieden (Johannes Gutenberg-Universität Mainz)

Pasticcio-Forschung und Digital Humanities

Innerhalb der Erforschung des frühneuzeitlichen Opernpasticcios ist die Entwicklung angemessener Darstellungsweisen der komplexen kultur- und musikgeschichtlichen Zusammenhänge ein zentrales Anliegen. Die Mobilität nicht nur der beteiligten Akteure, sondern auch der zahlreichen Arien, ihrer Abschriften und Drucke, die allesamt in die Produktion von Opernpasticci eingingen, ist in Vorträgen kaum anschaulich beschreibbar. Zum einen sind die Transferwege, die als Grundlage für einzelne musikalische Modifikationen anzusehen sind, an sich bereits zu filigran und verzweigt. Zum anderen potenziert sich diese Filigranität noch weiter, sobald gezielte Forschungsfragen wie z.B. ästhetisch-inspirierte überregionale Austauschprozesse untersucht werden.

Im Vortrag sollen die Vorzüge einer digital unterstützten Arbeitsweise in der Pasticcio-Forschung zwischen Präsentation und zentralen Forschungsfragen eruiert werden. Eine Verknüpfung von digitaler Edition (Edirom) und einer Datenbank ermöglicht dabei nicht nur kollaboratives Arbeiten und projektspezifische Visualisierungen samt ihrer operationellen und explorativen Möglichkeiten, sondern ebenso eine kulturhistorisch informierte Bewahrung dieser bisher selten beachteten musikalischen Gattung, die aber als kompositorisches Arbeitsprinzip für das 18. Jahrhundert von zentraler Bedeutung war.

Berthold Over (Johannes Gutenberg-Universität Mainz)

Die Basis der multimodalen Anwendung: Quellen und Daten

Im Projekt *PASTICCIO. Ways of Arranging Attractive Operas* werden drei Pasticci und eine Oper ediert: das von Georg Friedrich Händel arrangierte Pasticcio *Catone* (1732), das von der Operntruppe Mingotti aufgeführte Pasticcio *Catone in Utica* (1744), Johann Adolf Hasses Eigenpasticcio *Siroe* (1763) und dessen Vorlage (1733). Sowohl Quellen, die sich um diese Werke gruppieren (Libretti, Vorlagen für das Arrangement, Einzelarien), als auch Daten zu Akteuren (Sängerinnen/Sänger, Komponisten, Arrangeure, Impresari,

Widmungsträger) und Institutionen (Opernhäuser) werden in einer Datenbank nach FRBR strukturiert gesammelt, um sie in der Edition als Zusatzinformationen abrufen zu können. Ebenso fließen Aufführungsdaten von Werken und Aufenthaltsdaten von Personen in die Datenbank ein. Durch diese Zusatzinformationen können Fragen zur Herkunft von übernommenem musikalischen Material, zu den ursprünglichen Sängerinnen und Sängern, ihren Karrieren und Itineraren sowie zum Quellenmaterial und dessen Zirkulation beantwortet werden.

Martin Albrecht-Hohmaier (Johannes Gutenberg-Universität Mainz)

Die digitalen Editionen

Das Projekt erarbeitet zwei Editionen, in denen jeweils eine Werkgruppe präsentiert wird; zum einen mit zwei Opern-Pasticci, zum anderen mit einer Oper und einem Pasticcio. Diese Bündelung von je zwei Werken dient der beispielhaften Darstellung einer Opern-Praxis im 18. Jahrhundert, wobei die beiden Editionen bewusst sehr divergierende Praktiken darstellen. Die beiden *Catone*-Vertonungen, von unterschiedlichen Autoren und an unterschiedlichen „Institutionen" aufgeführt, haben fast keine Überschneidungen und zeigen zwei sehr voneinander abweichende Versionen, das Libretto zu vertonen bzw. musikalisch zu realisieren; die beiden *Siroe*-Vertonungen hingegen stammen vom gleichen Autor/Komponisten, entstanden mit 30 Jahren Abstand und sind an einigen Stellen deckungsgleich. Im Rahmen des Panels sollen an dieser Stelle einerseits Beispiele für die komplexe Quellenlage der ersteren Werke, andererseits solche für die Zusammenhänge bei einem Self-Pasticcio gezeigt werden – wobei an beiden die Vorteile einer digitalen Edition und die Sinnfälligkeit der Verknüpfung mit der Datenbank zutage treten.

Kristin Herold (ZenMEM Detmold/Paderborn)

Zusammenführung von Datenbank und Edition

Sowohl die Informationssammlung in Form einer Datenbank, als auch die digitale Edition wurden in anderen Projekten bisher als Einzelbestandteile betrachtet; jedoch erst durch die Kombination beider Komponenten kann im Rahmen des Projektes *PASTICCIO. Ways of Arranging Attractive Operas* das volle Potential ausgeschöpft werden, welches sich aus diesem Wechselspiel ergibt. Mit digitalen Methoden versucht das *PASTICCIO* -Projekt erstmals, Informationen über Sängerinnen und Sänger, die einzelnen Bestandteile der Pasticci und ihre Provenienzen und die Informationen über die Aufführungskontexte in Beziehung zu den Musikeditionen zu setzen.

Am Beispiel von Sängerkarrieren wird das spannende Zusammenspiel von Informationen aus Datenbank und Edition illustriert. Durch notwendige Umarbeitungen in Form von direkten Eingriffen in den Notentext (so mussten u. a. die Stimmumfänge von unterschiedlichen Sängerinnen und Sängern durch Transpositionen berücksichtigt werden) bei wechselnden Besetzungen ist beispielsweise die in der Datenbank vorgehaltene Information der Mobilität von Sängerinnen und Sängern auch direkt in der Edition belegbar und macht eine Karriere sichtbar.

Timetable

5 Min. Einführung

10 + 5 Min. Gesa zur Nieden, *Pasticcio-Forschung und Digital Humanities* + Feedback

10 + 5 Min. Berthold Over, *Die Basis der multimodalen Anwendung: Quellen und Daten* + Feedback

10 + 5 Min. Martin Albrecht-Hohmaier, *Die digitalen Editionen* + Feedback

10 + 5 Min. Kristin Herold, *Zusammenführung von Datenbank und Edition* + Feedback

25 Min. Diskussion des Plenums

Bibliographie

Albrecht-Hohmaier, Martin / Siegert, Christine (2015): *Eine codierte Opernedition als Angebot für Wissenschaft, Lehre und Musikpraxis. Überlegungen am Beispiel von Giuseppe Sarti (1729–1802)*, in: **Thomas Bein (ed.):** *Vom Nutzen der Edition. Zur Bedeutung moderner Editorik für die Erforschung von Literatur- und Kulturgeschichte* (= Beihefte zu editio 39). Berlin: de Gruyter 1–17.

Brandt, Stefan (2002): *Gleicher Text, unterschiedliche Realisierungen. Zum Einfluss des sängerischen Personals auf Arienkompositionen bei Porpora und Händel*, in: Basler Jahrbuch für historische Musikpraxis XXVI: 109–127.

Burden, Michael (2013): *Regina Mingotti: Diva and Impresario at the King's Theatre London* (= Royal Musical Association Monographs 22). Farnham.

Calella, Michele (2007): *Zwischen Autorwillen und Produktionssystem. Zur Frage des ‚Werkcharakters' in der Oper des 18. Jahrhunderts*, in: **Ulrich Konrad (ed.):** *Bearbeitungspraxis in der Oper des späten 18. Jahrhunderts. Bericht über die Internationale wissenschaftliche Tagung vom 18. bis 20. Februar 2005 in Würzburg* (= Würzburger musikhistorische Beiträge Bd. 27). Tutzing: Schneider 15–32.

Calella, Michele (2014): *Musikalische Autorschaft. Der Komponist zwischen Mittelalter und Neuzeit* (= Schweizer Beiträge zur Musikforschung 20). Kassel: Bärenreiter.

Décout, Maxime (2017): *Qui a peur de l'imitation?* Paris : Les éditions de minuit.

Freeman, Daniel E. (1992): *An 18th-Century Singer's Commission of ‚Baggage' Arias*, in: Early Music 20/3: 427–433.

Goehr, Lydia (2007): *The Imaginary Museum of Musical Works: an Essay in the Philosophy of Music*. Oxford: Oxford University Press.

Herold, Kristin / Siegert, Christine (2016): *Die Gattung als vernetzte Struktur. Überlegungen zur Oper um 1800*, in: **Kristina Richts / Peter Stadler (eds.):** *„Ei, dem alten Herrn zoll' ich Achtung gern". Festschrift für Joachim Veit zum 60. Geburtstag*. München: Allitera Verlag 671–702.

Holmes, William (1993): *Opera Observed: Views of a Florentine Impresario in the Early Eighteenth Century*, Chicago: University of Chicago Press.

Kanz, Roland (2002): *Die Kunst des Capriccio. Kreativer Eigensinn in Renaissance und Barock*. München: Deutscher Kunstverlag.

Korsmeier, Claudia Maria (2000): *Der Sänger Giovanni Carestini (1700–1760) und „seine" Komponisten* (= Schriften zur Musikwissenschaft aus Münster 13). Eisenach: Verlag der Musikalienhandlung Wagner.

LaRue, C. Stephen (1995): *Handel and His Singers. The Creation of the Royal Academy Operas*, 1720–1728. Oxford: Clarendon Press.

Over, Berthold (i.Dr.): *Zwischen multipler Autorschaft und autonomem Kunstwerk. Musikalische 'Werke' im 17. und 18. Jahrhundert*, in: **Panja Mücke (ed.):** *Geschichte der Musik im Barock*, Bd. 1: Weltliche Vokalmusik (Handbuch der Musik des Barock 1). Laaber: Laaber Verlag.

Siegert, Christine (2016): *Zum Pasticcio-Problem*, in: **Thomas Betzwieser (ed.):** *Opernkonzeptionen zwischen Berlin und Bayreuth. Das musikalische Theater der Markgräfin Wilhelmine. Referate des Symposiums anlässlich der Aufführung von L'Huomo im Markgräflichen Opernhaus in Bayreuth am 2. Oktober 2009* (= Thurnauer Schriften zum Musiktheater 31). Würzburg: Königshausen & Neumann 155–166.

Strohm, Reinhard (2011): *Wer entscheidet? Möglichkeiten der Zusammenarbeit an Pasticcio-Opern*, in: **Daniel Brandenburg / Thomas Seedorf (ed.):** *„Per ben vestir la virtuosa". Die Oper des 18. und frühen 19. Jahrhunderts im Spannungsfeld zwischen Komponisten und Sängern*, (= Forum Musikwissenschaft 6). Schliengen: Edition Argus 62–78.

Strohm, Reinhard (2009a): *Händels Pasticci*, in: **Arnold Jacobshagen / Panja Mücke (ed.):** *Händels Opern*, Teilband 2. Laaber: Laaber Verlag 351–358.

Strohm, Reinhard (2009b): *Der wandernde Gluck und die verwandelte ‚Ipermestra'*, in: **Irene Brandenburg / Tanja Gölz (ed.):** *Gluck der Europäer* (= Gluck-Studien 5). Kassel: Bärenreiter 37–63.

Tillett, Barbara (2005): *What is FRBR? A Conceptual Model for the Bibliographic Universe*, in: Australian Library Journal 54: 24–30.

Woyke, Saskia (1998): *Faustina Bordoni-Hasse. Eine Sängerinnenkarriere im 18. Jahrhundert*, in: Göttinger Händel-Beiträge 7: 218–257.

zur Nieden, Gesa (2015): *Mobile Musicians: Paths of Migration in Early Modern Europe*, in: European History Yearbook 16: 111–129.

Wie es Euch gefällt? Perspektiven wissenschaftsgeleiteter Organisationsformen des Datenmanagements für die Geisteswissenschaften

Ausgangslage

Durch die zunehmende Digitalität in den Geisteswissenschaften wird die Frage nach der Sicherung der langfristigen Verfügbarkeit digitaler Daten und Ergebnisse geisteswissenschaftlicher Forschung immer dringlicher. Inzwischen ist unter Beteiligung der wichtigsten Akteure (Wissenschaftler*innen, Wissenschaftseinrichtungen, Forschungsförderer) der Prozess der Einrichtung einer nationalen Forschungsdateninfrastruktur zur Sicherung der Nachhaltigkeit des Forschungsdatenmanagements im Sinne des Rats für Informationsinfrastrukturen (RfII) (RfII 2016) angegangen worden. Im Mittelpunkt des NFDI-Prozesses stehen Lösungsstrategien für grundlegende Probleme, wie die langfristige Förderung vielfach projektbasierter Infrastrukturen, die Stärkung der Koordination zwischen bestehenden Akteuren und die Sicherstellung der Interoperabilität durch gemeinsame Standards.

Die Geistes- und Kulturwissenschaften bilden momentan im NFDI-Prozess eine eigene Interessengemeinschaft. Im Rahmen verschiedener Workshops wurden Bedarfe und mögliche Organisationsformen der Infrastruktur für diese Community sondiert.[1] Im Mittelpunkt des unter dem Leitbild des Aufbaus *wissenschaftsgeleiteter Forschungsinfrastrukturen* stehenden NFDI-Prozesses steht der Ausgleich zwischen den Bedarfen der Fachdisziplinen/ Fachgesellschaften und bereits existierenden Angeboten bezüglich einer adäquaten Organisationsform (RfII 2017). Diese muss auch kleine Fachdisziplinen einbeziehen und mit entsprechenden Angeboten bedienen sowie interdisziplinäre Aspekte berücksichtigen.[2] Beim dritten NFDI-Workshop im Oktober 2018 stand daher ein „Konzept einer forschungsdaten- & methodenbasierten Infrastruktur für die Geisteswissenschaften" im Fokus.[3] Mit diesem Workshop ist der Konsortialbildungsprozess jedoch noch nicht abgeschlossen. Es werden weitere Möglichkeiten bestehen, die Bedarfe der Community zu signalisieren. Daher erscheint es zeitgemäß, aus Sicht der Praxis unter besonderer Berücksichtigung der digitalen Geisteswissenschaften, gemeinsam über *wissenschaftsgeleitete* Organisationsformen des Datenmanagements zu diskutieren.

Verortung der Thematik

Innerhalb der Geisteswissenschaften haben sich bereits verschiedene Organisationsformen des Forschungsdatenmanagements herausgebildet. Die AG Datenzentren hat in ihrem Grundsatzpapier (AG Datenzentren 2017, S. 20-22) drei idealtypische Formen geisteswissenschaftlicher Datenzentren typisiert: 1) Datenzentren innerhalb von Institutionen, 2) Datenzentren mit regionaler oder fachwissenschaftlicher Ausrichtung und 3) Datenzentren als Teil einer übergreifenden Infrastruktur mit internationaler Perspektive. An Institutionen angesiedelte Forschungsdatenzentren zeichnen sich häufig durch eine starke Nähe zur Wissenschaftspraxis der jeweiligen Hochschule aus und sind auf diese explizit ausgerichtet. Sie bieten häufig entsprechende Lösungsstrategien an oder finden diese in Verbünden. Fachwissenschaftliche Einrichtungen stellen übergreifende Anlaufstellen für Problemstellungen einzelner Fachcommunities dar, können dabei jedoch auch methodisch verwandte Wissenschaften bedienen. Verteilte Organisationsstrukturen mit problemorientierten Entwicklungen und Modellen verfolgen den Ansatz (teil)generische Services für eine breitere wissenschaftliche Öffentlichkeit zu schaffen.

Ziele des Panels

Die AG Datenzentren des DHd möchte im Rahmen der DHd 2019 ein Panel organisieren, in dem Herausforderungen und Lösungswege der Gestaltung einer *wissenschaftsgeleiteten* Organisationsform des geistes- und kulturwissenschaftlichen Datenmanagements aus verschiedenen Perspektiven näher beleuchtet werden. Die Zielstellung den Diskurs weiter zu fördern fügt sich organisch zu den strategischen Zielen

der AG Datenzentren: Mit mittlerweile 25 Institutionen weist die AG eine hohe Diversität an unterschiedlichen Schwerpunkten und Strukturen auf. Seit ihrer Gründung gehört zu den Kernanliegen der AG Datenzentren der Diskurs um die Gestaltung und Herausforderungen fachbezogenen Datenmanagements. So hat die AG Datenzentren 2017 ein Grundsatzpapier veröffentlicht (AG Datenzentren 2017), gemeinsam mit dem DHd-Vorstand eine Stellungnahme zum NFDI-Prozess verfasst (DHd Verband und AG Datenzentren 2018), einen Vorschlag zur Konsortialbildung formuliert (AG Datenzentren 2018) und arbeitet seit einiger Zeit an einem Leistungskatalog der innerhalb der AG Datenzentren vertretenen Akteure (Moeller et al. 2017, Helling et. al. 2018). Durch den Austausch im Rahmen des Panels sollen Impulse für die Agenda für die koordinierte Einrichtung einer nationalen Forschungsdateninfrastruktur gegeben werden, die insbesondere die Bedarfe der digitalen Geistes- und Kulturwissenschaften wiedergibt.

Leitfragen

Im Mittelpunkt der Diskussion sollen die Vor- und Nachteile einer forschungsdaten- und methodenbasierten bzw. einer fachgetriebenen Organisation der Infrastruktur stehen:

- Welche Rollen spielen disziplinäre Grenzen und disziplinspezifische Fragestellungen für die Etablierung einer Forschungsdatenkultur und ihre Organisationsform in den Geistes- und Kulturwissenschaften?
- Welche Auswirkungen hat die jeweilige Organisationsform des Datenmanagements auf multimediale, interdisziplinäre Forschungsprojekte? Welche besonderen Herausforderungen und Bedarfe stellen sich hier?
- Wie können wir sicherstellen, dass auch kleinere Fachdisziplinen beteiligt und ihre Bedarfe adressiert werden?
- Wie erreichen wir, dass wir uns nicht zu weit vom Forschungsalltag der Wissenschaftler*innen entfernen?
- Wo sind spezifische Angebote notwendig, wo reichen generische Dienste?
- Was können wir von anderen Disziplinen lernen, in denen sich schon eine gemeinsame Forschungsdatenkultur etabliert hat?

Mittels kurzer Impulsvorträge werden die Panelisten ihre Sicht auf diesen komplexen Themenbereich bieten. Besonderer Wert soll darauf gelegt werden einen Raum dafür zu bieten, die Meinung der Community der digitalen Geistes- und Kulturwissenschaften einzufangen. Die AG Datenzentren plant im Vorfeld des Panels einen Diskussionsimpuls online zu stellen (DHd-Blog). Die eingegangenen Kommentare werden in die Eingangsmoderation und die Schärfung der Leitfragen für die gemeinsame Diskussion einfließen. Die Ergebnisse des Panels werden wiederum öffentlich gemacht.

Aufbau des Panels

Das Panel wird mit einer einleitenden Moderation und 5 Impulsvorträgen einen Überblick über bestehende *Best Practices*, Bedarfe und Herausforderungen geben. Die Referent*innen skizzieren aus ihren Erfahrungen Problemfelder und setzen Impulse zur Lösung der Fragestellungen, um immer wieder rasch in eine offene Diskussion mit dem Publikum überzugehen.

Ablauf (Gesamtdauer 90 Minuten)

- Teil 1: Einführung in das Thema, strategische Ziele der AG Datenzentren, Vorstellung des Panels (5 Minuten) Ulrike Wuttke & Kai Wörner
  - Kurzfassung Ziel und bisherige Ergebnisse Projekt „Dienstekatalog“ in Hinsicht auf das Panelthema (10 Minuten): Patrick Helling
- Teil 2: 5 Impulsreferate der Panelist*innen unter besonderer Berücksichtigung der oben formulierten Leitfragen jeweils gefolgt von einer kurzen Diskussion (50 Minuten, jeweils 4 Minuten Referat + 6 Minuten Diskussion)
- Teil 3: Moderierte Diskussion zum Themenkomplex unter besonderer Berücksichtigung von drei aus der vorhergehenden Diskussion und den aus der Community im Vorfeld der DHd eingegangenen Rückmeldungen zugespitzten Leitfragen (25 Minuten): Ulrike Wuttke & Kai Wörner

Zusammensetzung des Panels und Impulsreferate

Organisation, Eingangs- und Endmoderation

- Ulrike Wuttke (Fachhochschule Potsdam): Dr. Wuttke ist stellvertretende Sprecherin der AG Datenzentren und Akademische Mitarbeiterin der Fachhochschule Potsdam für das EU-Projekt PARTHENOS. Sie verfügt über Expertise im Bereich Forschungsdatenmanagement unter besonderer Berücksichtigung nationaler und internationaler Infrastrukturen und Community-Anforderungen. Sie wird das Panel aus der spezifischen Sicht der Geistes- und Kulturwissenschaften moderieren.
- Dr. Kai Wörner (Universität Hamburg): Kai Wörner ist Convenor der AG Datenzentren und stellvertretender Leiter des Zentrums für nachhaltiges Forschungsdatenmanagement an der Universität Hamburg. Dort erarbeitet er ein Portfolio an Diensten und Beratungsleistungen zum Forschungsdatenmanagement, die aus dem bereits länger arbeitenden gwin-Projekt aus den Geisteswissenschaften auf alle Fächer der Universität ausgeweitet werden.

Organisation, Präsentation Dienstekatalog der AG Datenzentren

- Patrick Helling (Universität zu Köln): Patrick Helling ist wissenschaftlicher Mitarbeiter am Data Center for the Humanities (DCH) und am Institut für Digital Humanities (IDH) an der Universität zu Köln. Am DCH berät er Forscher*innen bei Fragen des geisteswissenschaftlichen Forschungsdatenmanagements. Im Rahmen der AG Datenzentren ist er mit dem Dienstekatalog betraut und beteiligt sich an AG Aktivitäten bezüglich einer Nationalen Forschungsdateninfrastruktur.

Impulsreferate

- Alexander Czmiel (Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Leiter Informationstechnologie und Digital Humanities) vertritt die Sicht einer außeruniversitären Forschungseinrichtung, die seit über 15 Jahren in den Digital Humanities aktiv ist, digitale geisteswissenschaftliche Forschungsressourcen entwickelt und diese langfristig kuratiert. Das Impulsreferat berichtet über die Erfahrungen, die während dieser Zeit gesammelt wurden und wird versuchen die Bereiche Forschungsdatenmanagement und DH aus der Position einer Wissenschaftsakademie einzuordnen. Dabei wird erläutert, was der Begriff „wissenschaftsgeleitet" für eine Einrichtung, die sowohl die Anbieterperspektive, als auch die Nutzerperspektive in sich vereint, bedeuten kann.
- Dr. Katrin Moeller (Leiterin des Historischen Datenzentrum Sachsen-Anhalt) vertritt die Perspektive eines Datenzentrums mit regionaler bzw. fachwissenschaftlicher Ausrichtung. Präsentiert wird die Notwendigkeit einer fachspezifischen Datenkuration in unmittelbarer Forschungsumgebung mit dem Ziel Wissenschaftler*innen dabei zu unterstützen, Forschungsvorhaben bzw. Fragestellungen in nachnutzbare Forschungsdatenstrukturen zu überführen und am Ende Daten in "Buchqualität" zu veröffentlichen, ohne dabei den Aufwand überbordend werden zu lassen.
- Peter Gietz (DAASI International, DARIAH-DE, DHd-Vorstand) vertritt die Perspektive eines Datenzentrums als Teil einer übergreifenden Infrastruktur mit internationaler Anbindung. Er versucht eine geisteswissenschaftliche NFDI so zu denken, dass sowohl die Interessen der Infrastrukturen DARIAH-DE und CLARIN-D, als auch die der im DHd-Verband organisierten Datenzentren und Fachwissenschaften berücksichtigt und gewahrt sind. Das Impulsreferat thematisiert die verschiedenen Ebenen von Knoten einer geisteswissenschaftlichen Forschungsdateninfrastruktur, die sich sowohl über geografische Zuständigkeiten als auch über Fachspezialisierung ausdifferenzieren, sowie mögliche Geschäftsmodelle. Über einen Brokerdienst könnten Forschungsprojekte oder Einzelwissenschaftler*innen forschungsrelevante generische und individuelle Dienste verschiedener Anbieter über einen einzigen Vertragspartner beziehen.
- Dr. Cosima Wagner (Universitätsbibliothek/ Campusbibliothek der Freien Universität Berlin) vertritt exemplarisch fachwissenschaftliche Anforderungen an das Forschungsdatenmanagement aus der Perspektive eines Regionalstudienfaches (sogenannte "Kleine Fächer"). Basierend auf einem BMBF-Projekt zum Aufbau und zur Erprobung von Strategien zum Forschungsdatenmanagement mit dem Schwerpunkt Ostasien- und Orient- bzw. Altertumswissenschaften skizziert das Impulsreferat FDM-Herausforderungen und Bedarfe von Regionalstudienfächern, die in Forschung und Lehre stark international ausgerichtet sind und bei denen FD in nicht-lateinischen Schriften anfallen. Ausgangspunkt ist die Beobachtung, dass im Zuge der digitalen Transformation des Wissenschaftssystems entstehende neue Infrastrukturen, Standards und Modelle nur in geringem Maße nicht-lateinische Schriften berücksichtigen, was zu einer Verschlechterung der Forschungsinfrastrukturen für diese Fächer führt.
- Dr. Barbara Ebert (Geschäftsstelle des Rats für Informationsinfrastrukturen, RfII) vertritt die wissenschaftspolitische Perspektive. Der RfII hat 2016 die Errichtung einer NFDI empfohlen, um damit einen neuartigen, effizienten Rahmen für das bisher zu kleinteilige und überwiegend befristet finanzierte Gefüge von Forschungsdatendiensten zu schaffen. Vorbereitend sind insbesondere die forschenden Communities bzw. Fachgemeinschaften gefordert, ihre Bedarfe zu definieren und mit aus ihrer Sicht geeigneten Infrastrukturpartnern Dienste-Portfolios mit längerer Perspektivplanung für ganze fachlich-thematische Domänen zu entwickeln. Vorgestellt werden Anforderungen diese Community-Infrastrukturpartnerschaften, die sog. Konsortien, sowie Eindrücke aus den vorbereitenden Diskursen verschiedener Akteure.

Fußnoten

1. Mehr Informationen zu diesen Veranstaltungen sowie weitere Hintergrundinformationen finden sich auf https://forschungsinfrastrukturen.de [letzter Zugriff 15. Oktober 2018].
2. Zusammenfassung der zweiten Veranstaltung der Workshopreihe *Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland*, https://forschungsinfrastrukturen.de/doku.php/zusammenfassung-2018-06-15 [letzter Zugriff 15. Oktober 2018].
3. Programm der dritten Veranstaltung der Workshopreihe *Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland*, https://forschungsinfrastrukturen.de/doku.php/programm-2018-10-04 [letzter Zugriff 15. Oktober 2018].

Bibliographie

DHd-AG Datenzentren (2017): *Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum. Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten*. Hamburg, Online: http://doi.org/10.5281/zenodo.1134760.

DHd AG Datenzentren (2018): *Vorschlag der AG Datenzentren im DHd zur Bildung und Strukturierung eines NFDI-Konsortiums für die Geisteswissenschaften*. Online: http://doi.org/10.5281/zenodo.1442845.

DHd Verband (2018): *Stellungnahme des Verbandes Digital Humanities im deutschsprachigen Raum (DHd) zur Nationalen Forschungsdateninfrastruktur (NFDI)*. Online: https://dig-hum.de/stellungnahme-dhd-nfdi [letzter Zugriff 15. Oktober 2018].

Helling, Patrick / Moeller, Katrin / Mathiak, Brigitte (2018): *Forschungsdatenmanagement in den Geisteswissenschaften - der Dienstekatalog der AG-Datenzentren des Verbands 'Digital Humanities im deutschsprachigen Raum' (DHd)*, in: ABI Technik, Band

38, Heft 3, 251 - 261. DOI: https://doi.org/10.1515/abitech-2018-3006 .

Moeller, Katrin / Ďurčo, Matej / Ebert, Barbara / Lemaire, Marina / Rosenthaler, Lukas / Sahle, Patrick / Wuttke, Ulrike / Wettlaufer, Jörg (2018): *Die Summe geisteswissenschaftlicher Methoden? Fachspezifisches Datenmanagement als Voraussetzung zukunftsorientierten Forschens*, Jahrestagung Digital Humanities im deutschsprachigen Raum (DHd), Book of Abstracts. DOI: http://doi.org/10.18716/KUPS.8085 .

RfII - Rat für Informationsinfrastrukturen (2016): *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*, Göttingen.

RfII - Rat für Informationsinfrastrukturen (2017): *Schritt für Schritt - oder: Was bringt wer mit? Ein Diskussionsimpuls für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)*, Göttingen.

Zeitungen und Zeitschriften als multimodale, digitale Forschungsobjekte: Theorien und Methoden

Einleitung

Zeitschriften und Zeitungen aus den verschiedenen historischen Epochen und Ländern werden seit Jahren digitalisiert und der Öffentlichkeit und Forschung zugänglich gemacht. Allein im europäischen Raum gibt es dazu groß angelegte Projekte, die untereinander z.T. eng vernetzt sind (Europeana, ANNO, The British Newspaper Archive, Delpher, Gallica, Hemeroteca digital, ZDB, ZEFYS). Diese Erzeugung neuer Forschungsdaten bzw. das Zugänglichmachen von bisher auf der ganzen Welt verstreuter kultureller Zeitzeugen wirft für sich allein genommen bereits eine Reihe praktischer und theoretischer Fragen auf. Da die meisten Digitalisierungsprojekte von Bibliotheken initiiert werden, herrschen zumindest in der Art und Weise der Erstellung und Verwaltung gewisse Standards vor (DFG 2013). Die Einigung auf ein Dateiformat, in dem die Zeitungen und Zeitschriften präsentiert werden oder welche Zugriffsrechte bestehen, scheint aber dennoch schwierig zu sein. Wenn sich Forschende ein Zeitungs- oder Zeitschriftenkorpus zusammenstellen, dann gehen sie zumeist eine Kooperation mit der besitzenden Bibliothek ein oder verfügen selbst über die Rechte an der Sammlung (vgl. dazu DH-Projekte wie: Oceanic Exchanges, illustrierte magazine, Die Fackel, Chinese Entertainment Newspapers, Chinese Women's Magazines, WeChangEd, Blue Mountain, MJP, MMP, Revistas Culturales 2.0, Journalliteratur, Literatur im Zeitalter der Illustrierten, u.v.m.). Es stellen sich schon bei der Datenerzeugung eine Reihe von Fragen, deren Klärung zu einer vereinfachten Zugänglichkeit der Forschungsdaten und -objekte führen würde (vgl. Panelbeiträge Neudecker, Kampkaspar/Resch).

Daneben sind Zeitungen und Zeitschriften als Medien per se multimodal (Igl / Menzel 2016; Fricke 2012) und -medial konzipiert. Sie setzen sich aus Bild-Text-Arrangements zusammen und nutzen verschiedene Modalitäten (Interaktion versch. Codes im selben Medium) zur Kommunikation. Da Zeitungen und Zeitschriften in verschiedene Bereiche untergliedert sind, verlangen sie von Leser*innen einen anderen Lektüremodus als das Buch, der Brief, das Plakat, das Comic oder andere Druckwerke. Dabei können sie aber selbst diese anderen Medien enthalten und abbilden. Gerade die Mischung der Formate und Inhalte zeichnet das Medium aus und definiert Sub-Genres (Tages- oder Wochenzeitung, Sport- oder Kulturzeitschrift). Zusätzlich zur schlichten Menge der Inhalte, kommt normalerweise auch ein Layout, das sich von anderen Druckwerken unterscheidet. Bei Zeitungen sind es zumeist Kolonnen, bei Zeitschriften können es auch buchähnliche Aufteilungen sein, die sich zudem im Laufe der Entwicklung vom 17.-21. Jahrhundert verändern (Rißler-Pipka 2017). All diese Merkmale, die hier nur angerissen werden können, bescheren uns beim Wechsel vom analogen Druckwerk zum Digitalisat eine Reihe von Herausforderungen – sowohl theoretischer als auch technischer Art. Von theoretischer Perspektive aus behaupten wir, historische Brüche und Veränderung anhand multimodaler Wahrnehmung diagnostizieren zu können. Das geht einher mit der von Walter Benjamin beschriebenen technischen Reproduzierbarkeit des Kunstwerks (Benjamin 1935-39) und reicht bis zu aktuell kritischen Beobachtungen aus Lateinamerika, die die Rezeption der modernen Presse und Zeitschriften mit der Nutzung digitaler Medien vergleichen (Macciuci 2015: 209). Aus technischer Perspektive stellt sich zum einen das Problem der Texterkennung auf Artikelebene (OCR, NewsEye). Zum anderen gilt es zu überlegen, wie man diese Besonderheiten des Artefakts Zeitung oder Zeitschrift ins digitale Forschungsobjekt überträgt (Drucker 2009: 109; Gooding 2017: 173). Nicht nur die Lesesituation wird eine andere sein, ob das Werk aus Papier im Kaffeehaus gelesen wird oder ob es am Bildschirm digital geblättert oder als XML-Datei dargestellt wird. Auf der anderen Seite gibt es Sorgen um den Verlust der Materialität, der Haptik, der multimodalen Wahrnehmung. Dazu haben sich schon lange vor der Digitalisierung eigene Theorien gebildet (Gumbrecht et al. 1995; Genz / Gévaudan 2016). Niemand möchte ernsthaft ein Äquivalent zum Artefakt finden, aber es sollte auf der anderen Seite die Flexibilität und Interoperabilität bestehen, um sowohl die Anforderungen von FAIR (Wilkinson et al.) als auch diejenigen der Forschung zu erfüllen. Entsprechend der vorgestellten Beiträge sollen gemeinsame Fragen diskutiert werden wie:

- Welche für die Forschung relevanten Elemente werden im digitalen Forschungsobjekt repräsentiert?
- Wie können die Daten und Metadaten für die Analyse genutzt werden?
- Welche Metadatenstandards werden aktuell international verwendet und welche Probleme könnten damit gelöst werden?
- Wie verändert sich die Forschung und ihre Fragestellungen durch die Arbeit am digitalen Forschungsobjekt?
- Kann die Multimodalität des Forschungsobjekts durch digitale Methoden analysiert werden oder wird das Nebeneinander eher verstärkt (Podewski 2018)?

Die Panelvorträge zeigen verschiedene Seiten aus den Bereichen der Digitalisierung und Bereitstellung (1-2) sowie der späteren Analyse (3-4), um so anhand konkreter Beispiele aus der Praxis die Diskussion anzustoßen.

Panelvorträge

Masse vs Klasse oder (wie) lässt sich die Dynamik von Digitalisierung und Digital Humanities organisieren? (Clemens Neudecker)

Die Zeitungsdigitalisierung wurde in Deutschland im internationalen Vergleich lange Zeit vernachlässigt. Vom DFG-Pilotprojekt zur Digitalisierung historischer Zeitungen wurde 2016 ein Master-Plan vorgelegt, dem neben einer DFG-Ausschreibung zur Massendigitalisierung von Zeitungen auch ein Projekt zur Errichtung eines nationalen Zeitungsportals innerhalb der Deutschen Digitalen Bibliothek (DDB) folgte. Das Ziel dieser Initiativen ist es, in den kommenden Jahren eine maßgebliche Erhöhung der digitalisierten Zeitungen in Deutschland und deren zentralen Nachweis für die Forschung zu ermöglichen. In Vorbereitung darauf wurde bereits die Zeitschriftendatenbank (ZDB) überarbeitet und bietet nun forschungsrelevante Mehrwertdienste wie u.a. die Visualisierung von Zeitungsnetzwerken. Aber auch auf europäischer (Europeana Newspapers) und internationaler Ebene finden Bemühungen statt, die zum Ziel haben, Suche und Analyse für digitalisierte Zeitungen über Sprach- und Landesgrenzen hinaus zu vereinfachen. Parallel dazu sind zahlreiche Digital Humanities Projekte entstanden (Oceanic Exchanges, impresso, NewsEye) die mit unterschiedlichen Verfahren und Forschungsdesigns an umfangreichen Zeitungskorpora arbeiten. Die Interoperabilität von Datenmodellen spielt dabei für die Digital Humanities eine ebenso große Rolle wie die Möglichkeit, zusätzliche durch Analyse und Annotation gewonnene Informationen zu vorhandenen Metadaten hinzuzufügen. Andererseits stoßen auch die im Zuge der Digitalisierung entstehenden Metadaten zur Provenienz auf zunehmendes Interesse bei Forschenden. Vor diesem Hintergrund stellt sich die Frage, wie sich die Dynamiken von (Massen-)Digitalisierung und spezifischen Anforderungen der diversen Forschungsvorhaben in Einklang bringen lassen und inwieweit hierbei neue Technologien wie IIIF und das W3C Web Annotation Data Model von Nutzen sein können.

Die Wiener Zeitung als Fallstudie: Partizipative Ansätze in der Praxis (Dario Kampkaspar / Claudia Resch)

Wie sich historische Zeitungen vom analogen Druckwerk in digitale Medien überführen lassen, wird derzeit am Beispiel der ältesten, bestehenden Zeitung der Welt, der Wiener Zeitung (im 18. Jahrhundert: "Wien(n)erisches Diarium") an der Österreichischen Akademie der Wissenschaften erprobt. Das Projektteam nutzt die Plattform "Transkribus", um aus mehreren Hundert historischen (Frakturtext-) Ausgaben, die derzeit im Portal ANNO als Bild-Digitalisate vorliegen, verlässliche Volltexte zu generieren und diese schrittweise online zur Verfügung zu stellen. Am Beispiel des Projekts soll gezeigt werden, wie gemeinsam mit künftigen User*innen im Rahmen der technischen Möglichkeiten eine Benutzeroberfläche entwickelt wird. In diesem Zusammenhang wird das Projektteam seine aktuelle Umfrage unter Forschenden, die mit der Zeitung vertraut sind und häufig mit ihr arbeiten, vorstellen. Denn je konkreter die individuellen Interessen der zukünftigen Zielgruppen am Datenmaterial formuliert sind, desto besser lässt sich einschätzen, worin die Herausforderungen bei der Digitalisierung historischer Zeitungs- und Zeitschriftenbestände in Zukunft liegen (vgl. Gooding 2017: 172).

Kulturzeitschriften als soziokulturelle Produkte (Teresa Herzgsell / Jörg Lehmann)

Im DFG-Projekt "Literarische Modernisierungsprozesse und transnationale Netzwerkbildung im Medium der Kulturzeitschrift" werden Metadaten zu Kulturzeitschriften manuell erhoben und in ein feingliedriges Klassifikationssystem eingeordnet: Name des/der Beiträger*in (nebst Geschlecht, Herkunftsland, etc.), Pseudonyme, Titel, Genre und Sprache des Beitrags. Gegebenenfalls werden auch Widmungen, Erstpublikation, Originalsprache, etc. erfasst. Gegen Ende des Projekts werden fast 60 spanischsprachige Kulturzeitschriften ausgewertet sein. Diese umfangreiche Datensammlung bildet die Grundlage für die Analyse der Akteurs- und Genrenetzwerke (Ehrlicher / Herzgsell 2016). Sie bietet aber auch die Möglichkeit, statistische Auswertungen der erhobenen Merkmale der Zeitschriften sowie der Beiträger*innen durchzuführen.

Um Zeitschriften als soziokulturelle Produkte zu charakterisieren, präsentieren wir in unserem Beitrag Auswertungen relevanter Metadaten. Unser Anliegen ist es, auf einer quantitativen Grundlage sichtbar zu machen, wie die Zeitschriften aufgemacht sind und wie die soziodemographischen Daten zu den Beiträger*innen diese Aufmachung konfigurieren. Wir wollen folgenden Forschungsfragen nachgehen: Inwiefern können die manuell erhobenen Daten – wie etwa die verwendeten Genreklassifikationen und die Titel der Beiträge – Auskunft geben über das Profil der Zeitschriften? Lassen sich die verschiedenen Zeitschriften auf Grundlage dieser Merkmalskombinationen in Gruppen mehr oder weniger ähnlicher Kulturprodukte unterteilen? Kann man auf statistischer Basis Vorlieben bestimmter Beiträger*innen für bestimmte Text- und Bildformen ausmachen, d.h. Autor*in-Genre-Korrespondenzen? Lassen sich Korrelationen zwischen den verwendeten Genres und den biografischen Daten der Beiträger*innen finden?

Forschen mit Metadaten: Über notwendige und zusätzliche Metadaten in Zeitungs- und Zeitschriftenprojekten (Nanette Rißler-Pipka)

Was bezeichnen wir überhaupt als Metadaten von Zeitungen und Zeitschriften und welchen Mehrwert haben sie für die Forschung?

Von Seiten der Literaturwissenschaft wird das Desiderat der Forschung bezüglich fehlender (Meta-)Daten, die es ermöglichen Zeitungen und Zeitschriften als kulturelle Zeitzeugen in Beziehung zu setzen, schon von Frank et al. genau bestimmt, (Frank et al.: 2010, 10). Es wird dort gefragt, wer wo über wen und aus welcher Perspektive geschrieben habe. Teilweise lässt sich dies mit bibliographischen Metadaten oder denjenigen aus Inhaltsverzeichnissen quantitativ ermessen. Teilweise müssen aber auch je nach Forschungsfrage Metadaten in Form von Annotationen hinzu gefügt werden. Der Fokus soll hier auf die Interoperabilität der Metadaten gelegt werden, d.h. wie werden sie angelegt? Dies zielt auf die Frage ab, welche theoretischen Vorüberlegungen für die Entwicklung einheitlicher Metadatenstandards speziell in dem Bereich der Zeitungen- und Zeitschriftenforschung notwendig sind. Neben den händisch angelegten Metadaten kommen in einem digitalen Analyseworkflow meist auch Ergebnisse aus Tools hinzu, die auch als Metadaten des digitalen Forschungsobjekts behandelt werden und im Idealfall gemeinsam mit den vorhandenen Metadaten ausgewertet werden sollten.

Bibliographie

ANNO (AustriaN Newpapers Online): http://anno.onb.ac.at [letzter Zugriff 31. August 2018]

Benjamin, Walter (1935-39): *"Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit"* in: ders.: Gesammelte Schriften, Bd. 1,2 Frankfurt am Main: Suhrkamp, {1980} 471–508.

Blue Mountain Project: http://bluemountain.princeton.edu/exist/apps/bluemountain/index.html [letzter Zugriff 31. August 2018]

Chinese Women's Magazines in the Late Qing and Early Republican Period: https://kjc-sv034.kjc.uni-heidelberg.de/frauenzeitschriften/ [letzter Zugriff 31. August 2018]

Chinese Entertainment Newspapers: http://projects.zo.uni-heidelberg.de/xiaobao/index.php?p=start [letzter Zugriff 31. August 2018]

Das Wien(n)erische Diarium: *Digitaler Datenschatz für die geisteswissenschaftlichen Disziplinen (GD2016/16)* https://www.oeaw.ac.at/de/acdh/projects/wiennerisches-diarium-digital/ [letzter Zugriff 31. August 2018]

Delpher: https://www.delpher.nl [letzter Zugriff 31. August 2018]

Deutsche Forschungsgemeinschaft (2013): *"DFG-Praxisregeln 'Digitalisierung'"*, http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_2013.pdf [letzter Zugriff 31. August 2018]

DFG Ausschreibung *"Digitalisierung historischer Zeitungen des deutschen Sprachgebiets"* vom 22. März 2018: http://www.dfg.de/foerderung/info_wissenschaft/info_wissenschaft_18_08/index.html [letzter Zugriff 31. August 2018]

DFG Projekt *"Errichtung eines nationalen Zeitungsportals auf der Basis der organisatorischen und technischen Infrastruktur der Deutschen Digitalen Bibliothek (DDB) – 'DDB-Zeitungsportal'"*: http://gepris.dfg.de/gepris/projekt/404633104 [letzter Zugriff 31. August 2018]

Die Fackel, Austrian Academy Corpus: http://corpus1.aac.ac.at/fackel [letzter Zugriff 31. August 2018]

Drucker, Johanna (2009): *SpecLab: digital aesthetics and projects in speculative computing*. Chicago: University of Chicago Press.

Ehrlicher, Hanno / Herzgsell, Teresa (2016): *Zeitschriften als Netzwerke und ihre digitale Visualisierung. Grundlegende methodologische Überlegungen und erste Anwendungsbeispiele*. http://www.revistas-culturales.de/de/buchseite/hanno-ehrlicher-teresa-herzgsell-zeitschriften-als-netzwerke-und-ihre-digitale [letzter Zugriff 31. August 2018]

Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung): http://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf [letzter Zugriff 31. August 2018]

ESPrit: http://www.espr-it.eu [letzter Zugriff 31. August 2018]

Europeana Newspapers: http://www.europeana-newspapers.eu [letzter Zugriff 31. August 2018]

Forschergruppe Journalliteratur: https://journalliteratur.blogs.ruhr-uni-bochum.de/ueber-die-forschergruppe/ [letzter Zugriff 31. August 2018]

Frank, Gustav / Podewski, Madleen / Scherer, Stefan (2010): *"Kultur – Zeit – Schrift. Literatur- und Kulturzeitschriften als 'kleine Archive'"*, Internationales Archiv für Sozialgeschichte der deutschen Literatur (IASL). 34, Nr. 2 (Januar): 1-45. https://doi.org/10.1515/iasl.2009.013 .

Fricke, Ellen (2012): *Grammatik multimodal: wie Wörter und Gesten zusammenwirken*. Linguistik, Impulse & Tendenzen 40. Berlin / Boston: De Gruyter.

Gallica (presses et revues): https://gallica.bnf.fr/html/und/presse-et-revues/presse-et-revues [letzter Zugriff 31. August 2018]

Genz, Julia / Gévaudan, Paul (2016): *Medialität, Materialität, Kodierung: Grundzüge einer allgemeinen Theorie der Medien*. Bielefeld: De Gruyter / transcript.

Gooding, Paul (2017): *Historic Newspapers in the Digital Age. "Search All About It!"*. London / New York: Routledge.

Gumbrecht, Hans Ulrich / Elsner, Monika / Pfeiffer, Karl Ludwig (eds.) (1995): *Materialität der Kommunikation*. Frankfurt am Main: Suhrkamp.

Hemeroteca digital (Bibiblioteca Nacional de España): http://www.bne.es/es/Catalogos/HemerotecaDigital/ [letzter Zugriff 31. August 2018]

Igl, Natalia / Menzel, Julia (eds.) (2016): *Illustrierte Zeitschriften um 1900: mediale Eigenlogik, Multimodalität und Metaisierung*. Edition Medienwissenschaft. Bielefeld: Transcript.

illustrierte magazine: http://magazine.illustrierte-presse.de [letzter Zugriff 31. August 2018]

impresso: https://impresso-project.ch/ [letzter Zugriff 31. August 2018]

International Image Interoperability Framework: https://iiif.io/ [letzter Zugriff 31. August 2018]

Literatur im Zeitalter der Illustrierten: *Stationen komplexer Text-Bild-Beziehungen im 19. Jahrhundert*: http://gepris.dfg.de/gepris/projekt/282457319 [letzter Zugriff 31. August 2018]

Macciuci, Raquel (2015): *"Tecnica, soporte, ámbitos de sociabilidad y mecanismos de legitimacíon: sobre la construccíon de espacios de literatura en la prensa períodica"* in: **Schlünder, Susanne / Macciuci, Raquel (eds.)**: *Literatura y técnica: derivas ficcionales y materiales. Libros, escritores,*

textos, frente a la máquina y la ciencia, Actas del VIII Congreso Orbis Tertius. La Plata: Ediciones del lado de acá 205–231.

MJP, Modernist Journals Project, http://modjourn.org [letzter Zugriff 31. August 2018]

MMP, Modernist Magazines Project: http://www.modernistmagazines.com [letzter Zugriff 31. August 2018]

NewsEye: https://www.newseye.eu/ [letzter Zugriff 31. August 2018]

Oceanic Exchanges: http://oceanicexchanges.org/ [letzter Zugriff 31. August 2018]

Podewski, Madleen (2018): *"›Kleine Archive‹ in den Digital Humanities – Überlegungen zum Forschungsobjekt ›Zeitschrift‹"*, in: **Huber, Martin / Krämer, Sybille (eds.)**: *Sonderband 3 der Zeitschrift für digitale Geisteswissenschaften.* https://doi.org/10.17175/sb003_010 [letzter Zugriff 31. August 2018]

Revistas Culturales 2.0: http://www.revistas-culturales.de [letzter Zugriff 31. August 2018]

Rißler-Pipka, Nanette (2017): *"Image and Text in Numbers: Layout Analysis for Hispanic and Spanish Modern Magazines"*, in: **Busch, Hannah / Fischer, Franz / Sahle, Patrick (eds.)**: *Kodikologie und Paläographie im digitalen Zeitalter 4 - Codicology and Palaeography in the Digital Age 4*, Books on Demand, Norderstedt 25–42, http://kups.ub.uni-koeln.de/id/eprint/7780 [letzter Zugriff 31. August 2018]

The British Newspaper Archive: https://www.britishnewspaperarchive.co.uk [letzter Zugriff 31. August 2018]

Web Annotation Data Model: https://www.w3.org/TR/annotation-model/ [letzter Zugriff 31. August 2018]

WeChangEd: http://www.wechanged.ugent.be [letzter Zugriff 31. August 2018]

Wilkinson, Mark u. a. (2016): *"The FAIR Guiding Principles for Scientific Data Management and Stewardship"*, Scientific Data 3 (15. März 2016): 160018, https://doi.org/10.1038/sdata.2016.18 [letzter Zugriff 31. August 2018]

ZDB: Zeitschriftendatenbank: https://zdb-katalog.de/ [letzter Zugriff 31. August 2018]

ZEFYS: http://zefys.staatsbibliothek-berlin.de [letzter Zugriff 31. August 2018]

Vorträge

Aristoteles multimodal – Mit ediarum in den Graphen

Kuczera, Andreas
andreas.kuczera@adwmainz.de
Akademie der Wissenschaften Mainz - Digitale Akademie, Deutschland

Martin, Fechner
fechner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Aristoteles multimodal – Mit ediarum in den Graphen

In diesem Beitrag wird am Beispiel der Forschungsdaten des Akademienvorhabens "Commentaria in Aristotelem Graeca et Byzantina" (CAGB) gezeigt, welchen Mehrwert das Zusammenspiel von Editionsumgebung (Ediarum) und Graphdatenbank (neo4j) bei Analyse von Forschungsdaten bieten kann und warum Schnittstellen Digitaler Editionen für Graphdatenbanken von zentraler Bedeutung sind. [1]

Abb. 1: Neue Homepage des Aristoteles-Projekts (Quelle: Fechner[2]).

Die Forschungssoftware *ediarum* stellt Werkzeuge zur Eingabe, Verwaltung und Präsentation von Digitalen Editionen zur Verfügung (vgl. Abb. 2). Die an der Berlin-Brandenburgischen Akademie der Wissenschaften entwickelte Software[3] erlaubt es historische Texte zu transkribieren, auszuzeichnen und zu kommentieren, sowie die einzelnen Text mit anderen Texten und Registereinträgen zu verknüpfen. Aktuell wird das einheitliche Webframework *ediarum.WEB* entwickelt, mit dessen Hilfe die Erstellung von Webpublikationen Digitaler Editionen aus *ediarum* heraus vereinfacht wird. Grundlage dafür ist ein allgemeiner Ansatz, der die Funktionalitäten von Digitalen Editionen über eine Schnittstelle nachhaltig verfügbar macht. [4] Dafür wird der bisherige Programmcode von schon bestehenden auf ediarum basierenden Digitalen Editionen nachgenutzt, vereinheitlicht und erweitert. In absehbarer Zeit ist eine Veröffentlichung von *ediarum.WEB* auf *github* geplant, damit es wie auch die übrigen Teile von *ediarum* frei genutzt werden kann.

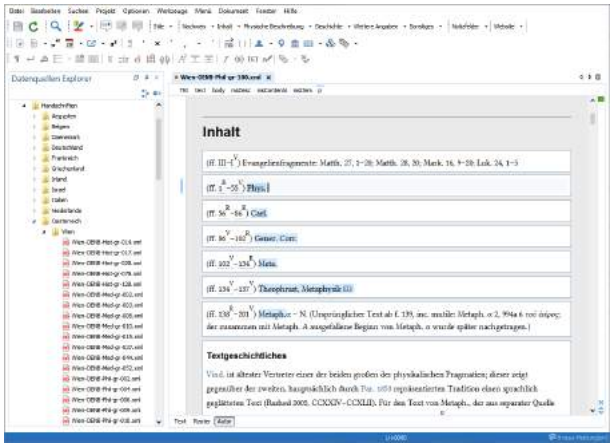


Abb. 2: Handschriftenbeschreibung von "Vind. phil. gr. 100" in Ediarum (Quelle: Fechner).

Durch die Nutzung dieses einheitlichen Webframeworks kommen alle darin enthaltenen und zukünftig entwickelten Features den damit publizierten Digitalen Editionen zugute. Auch schon bestehende Webseiten, wie etwa die des hier besprochenen Akademienvorhabens CAGB, können in Zukunft auf *ediarum.WEB* umgestellt werden. Bereits abgedeckte Funktionalitäten sind unter anderem die Anzeige und Filterung von Texten und Registereinträgen, sowie der Zugriff auf Daten und Metadaten über Schnittstellen. Auch gibt es eine Schnittstelle, die dafür entwickelt wurde , die Objekte und ihre Verknüpfungen von Digitalen Editionen in einem JSON-Format zu exportieren, das sich unter anderem für den Import in Graphdatenbanken eignet.

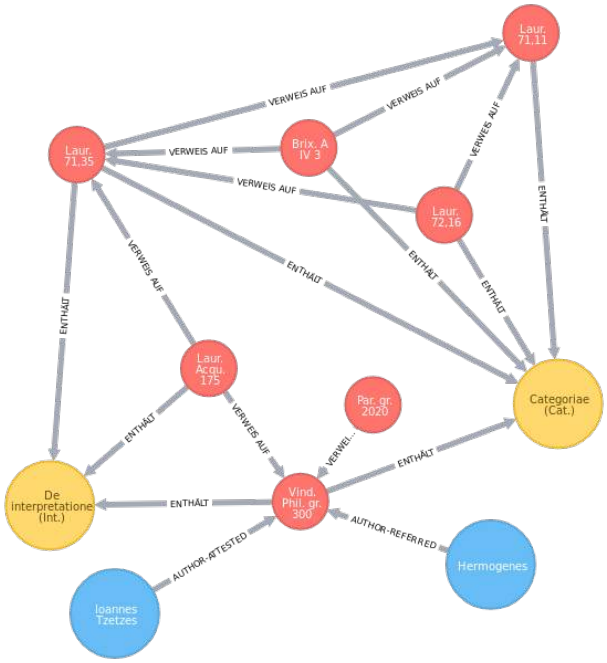


Abb. 3: Die gelben Knoten stellen die Aristoteleswerke dar, die roten Knoten Manuskripte, die sich jeweils auf Aristotelesstellen beziehen oder in deren Handschriftenbeschreibung von der Forschung auf andere Manuskripte verwiesen wird. Johannes Tetzel (die Person im linken blauen Knoten) hat vermutlich das rote Manuskrip geschrieben, auf welches es mit der AUTHOR_ATTESTED-Kante zeigt (Quelle: Kuczera/Fechner).

Graphdatenbanken haben sich in den letzten Jahren zu einem sehr flexiblen Werkzeug für das Management hochvernetzter Forschungsdaten entwickelt. Im Gegensatz zu relationalen Datenbanken, in denen Daten in verknüpften Tabellen abgespeichert werden, gibt es in Graphdatenbanken Knoten und Kanten (Kuczera2017 und Kuczera2018). Sie bieten die Möglichkeit, die Daten sehr flexibel aber gleichzeitig eng an der Forschungsfragestellung zu modellieren und zu analysieren. Damit ergeben sich Abfrageperspektiven, die mit relationalen oder XML-basierten Datenbanksystemen nur schwer zu realisieren sind.Solche hochvernetzten Forschungsdaten entstehen im Akademienvorhaben “Commentaria in Aristotelem Graeca et Byzantina” (vgl. Abb. 3). Mit Hilfe von ediarum entsteht ein detaillierter Handschriftenkatalog, dessen Ziel es ist, relevante Manuskripte mit mittelalterlichen Kommentaren zu Texten des Aristoteles vor allem aus dem 13. und 14. Jahrhundert zu verzeichnen. [5] Die digitale Publikation enthält darüber hinaus auch prosopographische Informationen zu den Autoren, Besitzern und Kommentatoren, sowie zu den in den Handschriften enthaltenen Texten. Allein für die drei Objekttypen "Personen", "Manuskripte" und "Aristoteles Werke" sind in der Datenbank mehrere Beziehungsmöglichkeiten eindeutig definiert. So kann ein Manuskript ein Werk des Aristoteles ganz oder teilweise enthalten, eine Person kann etwa Kopist oder Besitzer einer Handschrift sein, ebenso können aber auch die Handschriftenbeschreibungen Verweise auf damit in Beziehung stehende Handschriften enthalten.

Neben den teilweise sehr umfangreichen Handschriften- und Personenbeschreibungen enthalten insbesondere die Beziehungen zwischen den Datenobjekten für die Aristoteles-Forschung interessante Informationen. Im Rahmen der Handschriften- und Personenbeschreibungen wird durch die Beziehungen zunächst eine Kontextualisierung des Gegenstandes vorgenommen, welche die Einordnung und das Verständnis der jeweiligen Handschrift bzw. Person verbessert. Die Beziehungen selbst spannen letztlich ein Netzwerk auf, welches sich mit Hilfe von graph-basierten Werkzeugen filtern, visualisieren und analysieren lässt. Durch die Möglichkeit an einen solchen Graphen präzise Abfragen zu richten, lassen sich schließlich forschungsrelevante Fragen beantworten, deren Erforschung bisher nur durch zusätzliche mühsame Detailarbeit gelingen konnte und die weit über rein statistische Ansätze hinausgeht Im Gegensatz zur facettierten Filterung, die bei der Filterung auf die direkte Beziehung zwischen zwei Objekten angewiesen ist, können graphbasierte Abfragen auch Abhängigkeiten mit längeren Beziehungspfaden finden und darstellen. Durch die gleichzeitige Visualisierung und damit Kontextualisierung von Handschriften, Personen und Aristoteles Werken in einem gemeinsamen Netzwerk können die bekannten Verbindungen neu bewertet werden. Gerade die Möglichkeit zur Modellierung von multimodalen Netzwerken [6] in Graphdatenbanken bietet flexible Möglichkeiten den Forschungsgegenstand strukturiert zu erfassen und zu analysieren. Im Gegensatz zur historischen Netzwerkanalyse, bei der oft Algorithmen aus anderen Wissenschaftsbereichen wie z.B. der sozialen Netzwerkforschung angewendet werden, können sich Abfragen an eine Graphdatenbank sehr eng an der vom Wissenschaftler selbst gewählten Modellierung orientieren.

p1.label	Texte	Anzahl_von_Handschriften
"Giorgio Valla"	["De sensu (Sens.)", "De memoria (Mem.)", "De insomniis (Insomn.)", "De longitudine vitae (Long.)", "De iuventute (Juv.)", "De respiratione (Respir.)", "De coloribus (Col.)", "Mechanica (Mech.)", "De lineis insecabilibus (Lin.)", "Topica (Top.)", "Organon (Organon)", "De interpretatione (Int.)", "Categoriae (Cat.)"]	13
"Albertus Pius"	["Topica (Top.)", "Categoriae (Cat.)", "De interpretatione (Int.)", "Organon (Organon)", "Topica (Top.)", "Organon (Organon)", "De interpretatione (Int.)", "Categoriae (Cat.)", "De interpretatione (Int.)", "Topica (Top.)", "Organon (Organon)", "Categoriae (Cat.)"]	12
"Gerardus Vossius "	["Rhetorica (Rhet.)", "Topica (Top.)", "De caelo (Cael.)", "De anima (An.)", "Metaphysica (Metaph.)", "Physica (Phys.)", "De caelo (Cael.)", "Physica (Phys.)", "Meteorologica (Mete.)", "De anima (An.)"]	10
"Manuel Kointhos"	["De respiratione (Respir.)", "De vita et morte (VM)", "De memoria (Mem.)", "De insomniis (Insomn.)", "De longitudine vitae (Long.)", "De iuventute (Juv.)", "De caelo (Cael.)", "Meteorologica (Mete.)", "De anima (An.)", "De sensu (Sens.)"]	10
"Georgios Scholarios"	["De anima (An.)", "De sensu (Sens.)", "De longitudine vitae (Long.)", "De memoria (Mem.)", "De vita et morte (VM)", "De iuventute (Juv.)", "De spiritu (Spir.)", "De respiratione (Respir.)", "Magna moralia (MM)", "Ethica Nicomachea (EN)"]	10
"Augerius von Busbeck"	["Ethica Nicomachea (EN)", "Physica (Phys.)", "Metaphysica (Metaph.)", "De interpretatione (Int.)", "De mundo (Mu.)", "De sensu (Sens.)", "De insomniis (Insomn.)", "De memoria (Mem.)", "De mundo (Mu.)", "De sensu (Sens.)"]	10
"Marullo"	["De memoria (Mem.)", "De sensu (Sens.)", "De longitudine vitae (Long.)", "De insomniis	9

Abb. 4: Personen und Texte, auf die sie Zugriff hatten (Ausschnitt) (Quelle: Kuczera/Fechner).

Für die Forschung ist es etwa von Interesse herauszufinden, auf welche Texte eine bestimmte Person Zugriff hatte (Abb. 4) oder welche Kopisten und Besitzer von Handschriften eines bestimmten Textes es in einem bestimmten Zeitraum gab. Leider ist in bisherigen gedruckten Katalogen, sowie auch In der Datenbank diese Information nur implizit vorhanden: Die einzelnen Handschriftenbeschreibungen enthalten Angaben zu ihren Besitzern und den in ihnen enthaltenen Texten. Die Rechercheaufgabe besteht also etwa darin, zu einer bestimmten Person alle Handschriften, zu denen sie Zugang hatte, zu finden und weiterhin alle darin enthaltenen Texte. In einer graphbasierten Datenbank ist lässt sich dies in einem einfachen Query ausdrücken, wodurch die Antwort sogar nicht nur für eine Person, sondern auch gleich für alle ausgeführt werden kann.

p1.label	p2.label	Manuscripts	Anzahl
"Rodolfus Pius "	"Albertus Pius"	["Mut. gr. 199", "Mut. gr. 118", "Mut. gr. 208", "Mut. gr. 210", "Mut. gr. 198", "Mut. gr. 207", "Mut. gr. 197", "Mut. gr. 194", "Mut. gr. 86", "Mut. gr. 195", "Mut. gr. 91", "Mut. gr. 201", "Mut. gr. 180", "Mut. gr. 214", "Mut. gr. 209", "Mut. gr. 235", "Mut. gr. 29", "Mut. gr. 206", "Mut. gr. 184", "Mut. gr. 205", "Mut. gr. 204"]	21
"Albertus Pius"	"Rodolfus Pius "	["Mut. gr. 199", "Mut. gr. 118", "Mut. gr. 208", "Mut. gr. 210", "Mut. gr. 198", "Mut. gr. 207", "Mut. gr. 197", "Mut. gr. 194", "Mut. gr. 86", "Mut. gr. 195", "Mut. gr. 214", "Mut. gr. 91", "Mut. gr. 201", "Mut. gr. 180", "Mut. gr. 209", "Mut. gr. 235", "Mut. gr. 29", "Mut. gr. 206", "Mut. gr. 184", "Mut. gr. 205", "Mut. gr. 204"]	21
"Andreas Darmarios"	"Sambucus, Johannes"	["Vind. Phil. gr. 85", "Vind. Phil. gr. 247", "Vind. Phil. gr. 41", "Vind. Phil. gr. 325", "Vind. Theol. gr. 120"]	5
"Sambucus, Johannes"	"Andreas Darmarios"	["Vind. Phil. gr. 85", "Vind. Theol. gr. 120", "Vind. Phil. gr. 325", "Vind. Phil. gr. 41", "Vind. Phil. gr. 247"]	5
"Albertus Pius"	"Giorgio Valla"	["Mut. gr. 91", "Mut. gr. 189"]	2
"Giorgio Valla"	"Albertus Pius"	["Mut. gr. 91", "Mut. gr. 180"]	2

Abb. 5: Personen und Manuskripte, auf die sie jeweils über ihre persönliche Sammlung Zugriff hatten (Quelle: Kuczera/Fechner).

In der Graphdatenbank sind aber auch darüber hinausgehende Abfragen möglich. So lassen sich die Texte finden, die von den meisten Personen besessen wurden, oder Personenpaare, welche die gleichen Manuskriptsammlungen besaßen (Abb. 5). Die Visualisierungen vonN etzwerken ermöglichen den Forschern eine neue Perspektive auf die komplexen Beziehungen von Personen, Manuskripten und Texten. Einen guten Überblick über die Forschungslage zur Rezeption bestimmter Texte, ergibt etwa die Abfrage nach einzelnen Texten, nach den sie enthaltenen Manuskripten und den dazugehörigen Besitzern. Aus der Analyse wird leicht ersichtlich, bei welchen Personen der Zugriff auf einen oder mehrere Texte nachgewiesen werden kann (Abb. 6).

Die Anreicherung von Texten mit Kontext, etwa über Register und Kommentare ist ein zentraler Aspekt von

Editionen und Digitalen Editionen. Die Daten selbst basieren auf akribischer Forschungsarbeit an den Quellen, die ein großes Hintergrundwissen voraussetzt. Mit dem Export der von Digitalen Editionen behandelten Objekte und Relationen über eine nachnutzbare Schnittstelle wie in ediarum.WEB wird die Möglichkeit eröffnet, die vernetzten Forschungsdaten neuen Analysen zugänglich zu machen. Die graphbasierten Analysen reproduzieren die in den Quellen abgebildeten Kontexte und stellen diese erstmals gemeinsam und vernetzt dar. Sie sind nicht als Gegensatz zur bisherigen Forschungspraxis zu sehen, sondern als Fortentwicklung der bestehenden Methoden. Damit helfen sie bei der Bewertung, Kontextualisierung und Einordnung der einzelnen Quellen und des Gesamtkorpus.

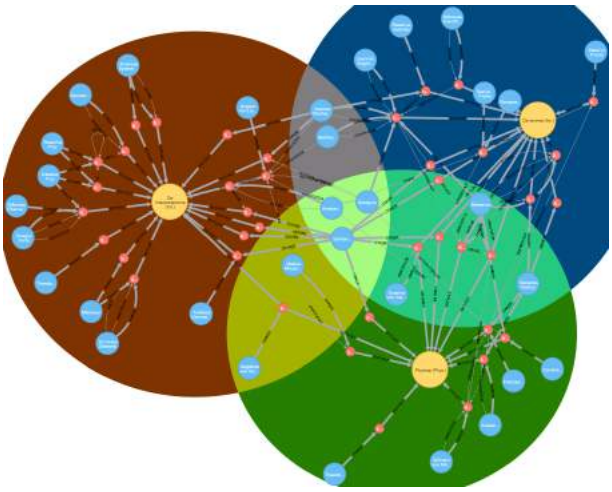


Abb. 6: Welche Personen (blau) haben über welche Manuskripte (rot) Zugriff auf die Werke (gelb) “De interpretatione” (links), “De anima” (oben rechts) und “Physica” (unten rechts). (Quelle: Kuczera/Fechner).

Fußnoten

1. Den Autoren ist kein Projekt bekannt, bei dem eine digitale Editionsumgebung wie z.B. Ediarum über Schnittstellen zu einer Graphdatenbank Datenauswertungen vornimmt, die über die Möglichkeiten der Editionsumgebung hinausgehen. Eine nicht text-zentrierter Anwendungsfall von Graphdatenbanken wird im Dariahpaper "Alte Daten neu verknoten: Die Verwendung einer Graphdatenbank für die Bilddatenbank REALonline" von Ingrid Mattschinegg (u.a.) beschrieben (vgl. https://de.dariah.eu/working-papers). www.creta.uni-stuttgart.de
2. Die Bilder sind allesamt von den Autoren selbst erstellt worden.
3. *ediarum* wird seit 2012 an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) entwickelt und wird mittlerweile in über 15 Projekten in- und außerhalb der BBAW eingesetzt (vgl. Fechner 2017). URL: http://www.bbaw.de/telota/software/ediarum . Der Quellcode für die publizierten Module von ediarum findet sich unter https://github.com/ediarum .
4. Das Konzept hierzu wurde bereits auf der DHd-Tagung 2018 vorgestellt und diskutiert (vgl. Fechner 2018).
5. Das Akademienvorhaben „Commentaria in Aristotelem Graeca et Byzantina“ ist Teil des von Bund und Ländern geförderten Akademienprogramms. URL der digitalen Publikation: https://cagb-db.bbaw.de/ .
6. Multimodale Netzwerke haben mehrere Typen von Knoten. Unimodale Netzwerke haben dagegen nur einen Knotentyp.

Bibliographie

Martin Fechner: *“ediarum: eine digitale Arbeitsumgebung für Editionsvorhaben”*, Folien zum Workshop »ediarum« im Rahmen des Workshops der Akademienunion am 20. Oktober 2017 in Mainz, URL: https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-28169

Martin Fechner: *Eine nachhaltige Präsentationsschicht für digitale Editionen*, in: DHd 2018 - Kritik der digitalen Vernunft. Konferenzabstrakts, hg. v. Georg Vogeler, Köln 2018, URL: http://dhd2018.uni-koeln.de/

Andreas Kuczera: *Graphentechnologien in den Digitalen Geisteswissenschaften*, in: ABI Technik 2017; 37(3): 179–196 (https://doi.org/10.1515/abitech-2017-0042). URL: https://www.degruyter.com/downloadpdf/j/abitech.2017.37.issue-3/abitech-2017-0042/abitech-2017-0042.pdf

Andreas Kuczera: *Graphentechnologien in den digitalen Geisteswissenschaften*. Modellierung – Import – Exploration (Online-Publikation unter https://kuczera.github.io/Graphentechnologien/) . (2018)

Automatic Font Group Recognition in Early Printed Books

Weichselbaumer, Nikolaus
weichsel@uni-mainz.de
JGU Mainz, Deutschland

Seuret, Mathias
mathias.seuret@fau.de
FAU Erlangen, Deutschland

Limbach, Saskia
limbach@uni-mainz.de
JGU Mainz, Deutschland

Christlein, Vincent
vincent.christlein@fau.de
FAU Erlangen, Deutschland

Maier, Andreas
andreas.maier@fau.de
FAU Erlangen, Deutschland

Introduction

Early modern books were printed with a large variety of different fonts. In the first decades after Gutenberg's invention, every printer had to start out by cutting his own punches and casting his own fonts. This diversity was somewhat standardised with the advent of an organised font trade in the 16th century. However, this was a very long process and one that was not completed before the 19th century. Only then did industrialised mass production make fonts as stable and - at least for text fonts - predictable as we know them today. The diversity of fonts is one of the cornerstones of analytical bibliography: with detailed descriptions of the individual fonts we can identify the printer of almost any given incunabula or at least narrow the possible candidates down to a very small group.

For OCR, however, this is a major drawback. Most OCR-models are trained to work with one of three different training sets, based on either just modern antiqua-fonts or on 19th-century standard Fraktur or on all fonts that ever existed. Specialised OCR-models, e. g. for Rotunda or Textura, almost don't exist as they would be very difficult to apply. One reason for this is that metadata for digitised books usually does not include the the font group or even the font of the main text face. Therefore, these models would - at the moment - only be applicable if the font is recognised manually. Given the vast amount of digital copies rendered by the large-scale digitisation projects like those for VD16, VD17 and VD18, this is out of the question.

Our project addresses this problem in two ways. Firstly, we will create a tool that can identify font groups automatically, i.e. fonts which are similar to each other and thus can be used jointly for training an OCR model (Christlein / Weichselbaumer 2016). Secondly, we will create OCR-models for various font groups. In this way, we hope to significantly improve the recognition rates of OCR for early printed books. In this paper, we will present and discuss first results on automatic font group recognition.

Basis

Fortunately, we have outstanding Ground Truth data: the Gesamtkatalog der Wiegendrucke (GW) (Staatsbibliothek zu Berlin 2019a) and its side project, the Typenrepertorium der Wiegendrucke (TW) (Staatsbibliothek zu Berlin 2019b). Both were initiated by Konrad Haebler at the turn of the last century and are still maintained today at the Berlin State Library. The GW provides us with bibliographical data for all known incunabula editions (books printed in the 15th century) as well as some 15,000 digital copies from all over the world. The corresponding records in the TW list over 6,000 fonts used for these books and later editions. This Ground Truth data was painstakingly collected in over a century and was recently - thankfully - converted to a database by the Berlin State Library (Eisermann / Duntze 2014).

Method

In a first step, we have also accumulated a large body of material from our collaboration partners: The University Library of Cologne, the Herzog-August Library in Wolfenbüttel, the University Library of Heidelberg, the University Library of Erlangen, the Berlin State Library, the Göttingen State Library, the Stuttgart State Library and the Bavarian State Library. We will also be able to work soon - for the first time - with digitised copies from the British Library, which is currently scanning its large incunabula collection. All in all, we have over a million images which provide a sound basis for our goals.

For evaluation purposes, random pages were taken out from the labelled data. We have two such subsets: validation and test. The validation data is used for tuning the classification method and evaluating it during its development, while the test data is kept until the end for an unbiased method evaluation on never seen, never used data.

A deep convolutional neural network (CNN) is used for the font recognition. To have both a robust and proven network architecture, we used one inspired by a residual network with 50 layers, also known as ResNet50 (He et al. 2016). A typical ResNet50 has layers with a large amount of neurons (from 64 to 512 in the convolutional layers), which can lead to overfitting the training data. To avoid this pitfall, we restrict the layers to having 96 neurons, with a penultimate fully-connected layer of 384 neurons. The training is done by stochastic gradient descent on batches of 64 samples, with a constant momentum of 0.9 and initial learning rate and weight decay of respectively 0.01 and 0.0005. The learning rate and weight decay are divided by 10 every 300,000 samples. The training is stopped after processing a million samples because the error stagnates.

The CNN has a receptive field of 224x224 pixels, which is insufficient for processing a whole page image at once. To identify the font of a page, we present 25 random crops from this page to the CNN, and average the results of the last linear layer (not the softmax output), then the class with maximum average value is taken. Crops full of text contain between 15 and 500 characters, depending on image resolution and text size. Typically, if a crop is misclassified (e.g., if it did not contain text), it will have little impact on the average result as the CNN is likely to produce results will low confidences.

Evaluation

We used, as training data, 280 pages with a median resolution of 2 megapixels from the 15th century containing text with fonts from four different groups: Antiqua, Bastarda, Rotunda, and Textura.

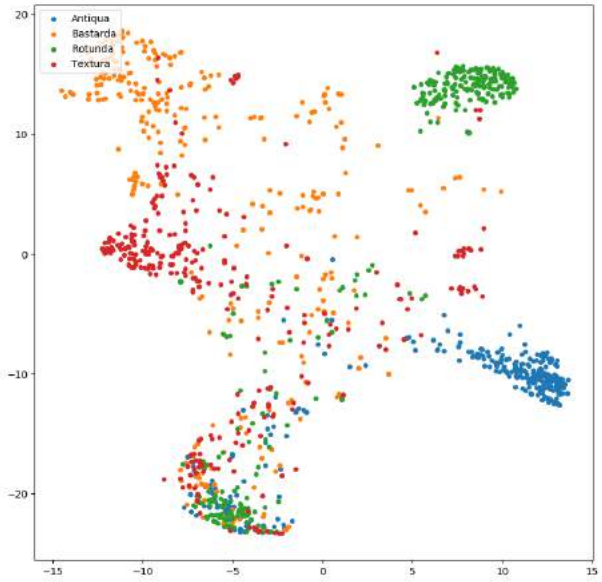
This means the CNN does not have the ability to answer something else. It is however useful to investigate what happens when pages with other fonts are given to the CNN. So as a test we provided 100 images of Fraktur - a font very closely connected to Bastarda. In addition to that, we also provided 30 images of each Greek, Hebrew and Italic - fonts that are rather different to the others. The results obtained on the test data for the four base fonts (15 pages each), as well as on the other pages, are given in the following confusion matrix:

	Antiqua	Bastarda	Rotunda	Textura
Antiqua	14			1
Bastarda		13	1	1
Rotunda			15	
Textura		1		14
Fraktur	0	92	4	4
Greek	5	20	4	1
Hebrew	3	17	6	4
Italic	1	26	2	1

Rows correspond to fonts, and columns to results given by the CNN. We can see that the test pages with the fonts known by the CNN are well classified, with an accuracy of 93%. The other fonts are more spread, but mostly classified as Bastarda.

This is a very significant result. Not only does the network recognize the connection between Bastarda and Fraktur but it also perceives the significant difference between Bastardas and the other three groups. After all, from the very beginning of typography the font group Bastarda differed considerably from other font groups. This is especially true for Texturas and Rotundas which have very uniform characteristics: Textura letters are upright, narrow and stand on crooked feet and Rotunda letters are small, curved and reveal a great contrast between thick and thin strokes. In contrast, Bastardas show much more variety - letters tend to slope forward and have flourished ascenders, yet many of them do not have these characteristics. Therefore it is very plausible that the CNN would categorize 'unknown' fonts as Bastardas.

This matches what can be seen from the data produced by the penultimate fully-connected layer of the CNN. As it has 384 dimensions, a t-Distributed Stochastic Neighbor Embedding (t-SNE) can be applied for visualization purpose. In the figure below, the dots correspond to individual random crops from test images. We can see five main areas. The one at the bottom corresponds to crops with little or no text, and therefore the CNN produces similar values regardless of the type group of the page. The points between this cluster and the center of the graphics might correspond to crops with text content, but not in a quantity large enough for identifying the script. Then, we have three well defined clusters for Antiqua, Rotunda, and Textura. Finally, the part corresponding to Bastarda is well spread and significantly less dense than for the other type groups. Thus, the CNN produces more variability in its penultimate layer for the Bastarda than for the other type groups, and a more important area of the feature space is considered, by the CNN, as corresponding to Bastarda. This could also explain why unseen type groups are frequently classified as belonging to the Bastarda.



Outlook

These results show that it is feasible to recognise font groups automatically. The authors are currently working on improving accuracy further and to expand the scope of the recognition tool from the 15th to the 16th-18th century. At the same time preliminary steps are taken to recognise not only font groups but exact fonts. This feature would not only make it much quicker to date and identify the printer of early modern books based on their fonts but also make this procedure much more accessible to scholars who are not highly specialised in analytical bibliography.

The source code used in this paper is available on github (https://github.com/seuretm/typegroups-classification-projection). Please note that the exact same

results cannot be obtained due to the randomness of initial parameters, and that the data is currently not publicly available.

Bibliographie

Christlein, Vincent / Weichselbaumer, Nikolaus (2016): *Automatische Typenbestimmung in historischen Drucken.* Poster at DHd2016, Abstract online: http://dhd2016.de/boa-2.0.pdf [Access date 9 October 2018].

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 [Access date 9. October 2018].

He et al. (2016): https://ieeexplore.ieee.org/abstract/document/7780459.

Staatsbibliothek zu Berlin (2019a): https://www.gesamtkatalogderwiegendrucke.de [Access date 11 January 2019].

Staatsbibliothek zu Berlin (2019b): https://tw.staatsbibliothek-berlin.de [Access date 11 January 2019].

Automatic recognition of direct speech without quotation marks. A rule-based approach

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Brunner, Annelen
brunner@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Motivation

Many texts incorporate multiple voices: the voice of the narrator and those of the characters or people who are quoted by the narrator. Separating these quotations from the surrounding text is relevant for many applications: In the field of literary studies it is a requirement for studies concerning character representation like sentiment analysis (e.g. Blessing et al. 2016; Schmidt / Burghardt / Dennerlein 2018) or character networks (e.g. Rydberg-Cox 2011; Dimpel 2018). For non-fictional texts, recognizing quotations is relevant for question answering and similar tasks. As those applications rely on processing a lot of textual material, having a way to automatically detect instances of direct speech (DS) is crucial.

As long as a specific pattern of quotation marks is used consistently this is a trivial task. Unfortunately, this is not necessarily the case: First, there is an astounding number of ways to encode quotation marks and incorrect or inconsistent usage is very common. Mistakes like missing closing quotation marks happen easily and can throw off a parser relying solely on those markers. (Brunner 2015: 180-182) The problem is worse for older texts, where the typographic rules are even less standardized and additional errors can occur in the digitization process. Finally, in literature it is not uncommon that authors deliberately choose not to use any markers for stylistic reasons. Those types of texts are especially common in the Digital Humanities and it is useful to have a tool that is not dependent on the use of quotation marks. In addition to that, the tool presented in this talk is rule-based and thus requires no training material.

Related Work

The detection of DS is usually a pre-processing step for another task so it rarely gets much focus in the respective papers. There are many applications for English that could be cited here, but we will deliberately focus on applications for German.

Pouliquen / Steinberger / Best 2007 develop a tool for automatic quotation detection and speaker attribution in newspaper articles for several languages. They look for the proper name of a public character, followed by a verb associated with speech representation, followed by quotation marks. Due to this strict pattern, their recall is relatively low (76%) but they identified 81.7% correct quotations.

The tool GutenTag, developed by Brooke et al. 2015, implements several NLP techniques to use on the texts of Gutenberg corpus. The GutenTag DS recognition relies solely on quotation marks. Before processing the text, the tool checks which types of quotation marks (single or double quotes) are used in it.

In her study about the automatic recognition of speech representation in German literary texts, Brunner 2015 implements two strategies for DS detection: Her rule-based approach uses quotation marks as well as pattern matching to identify frames (proper name - speech verb - colon/comma). This leads to some success in texts with unmarked DS. On a corpus of 13 German narrative texts an f-score of 0.84 for the category *direct* is achieved in a sentence-wise evaluation. Brunner also implements a machine learning approach using random forests and features based on POS-tags and markings for typical speech/thought/writing words. In a ten-fold cross validation, this model achieves an f-score of 0.87 for *direct*. If quotation marks are ignored entirely in training and evaluation, the model still achieves an f-score of 0.81.

Jannidis et al. 2018 implemented a recognizer for DS based on deep learning, specifically trained to work without quotation marks. It is trained on a corpus of 300 German fictional texts in which the quotation marks were removed. This recognizer achieves an accuracy of 0.84 in sentence-wise evaluation and 0.90 in token-wise evaluation on the corpus that is called “Gutenberg” in our evaluation.

To our knowledge, there is at the moment no recognizer for DS in German texts that is rule-based and does not use quotation marks.

Data

We evaluated our algorithm on four different and distinct data sets.

1) Gutenberg[1]: 33 short stories, taken from Project Gutenberg

2) DROC_red[2] (Krug et al. 2018): 85 text samples of German novels (1800-1950). Each sample spans at least one chapter.

3) RW[3]: Text samples (1840-1920) with 200-1000 tokens, manually annotated in the project "Redewiedergabe"

3a) RW_fict: 222 fictional text samples

3b) RW_nonfict: 206 non-fictional text samples

Methods

The algorithm tries to detect whether a given token or a given sentence is part of a DS. Currently, it cannot reliably determine the exact borders of individual DS instances. The technique is purely rule-based and does not rely on machine learning or training data of any sort. The following pre-processing steps were performed: a) tokenization with the OpenNLP Tokenizer[4], b) sentence detection with the OpenNLP SentenceDetector[5], c) tagging with the RFTagger[6], d) parsing with the mate dependency parser[7] and e) named entity recognition[8].

The algorithm tries to solve the problem in the following steps:

1) Segment the text into narrative and non-narrative sections (use paragraphs, if available):

If meaningful paragraphs are present in the text, those paragraphs usually tend to represent either narrative sections or dialogue sections. As it is unlikely that a narrative section contains DS at all, this knowledge can be used to adapt the recognizer rules. However if no such paragraphs are available, the algorithm starts by reconstructing surrogate sections, using the concept of "coherence" between narrative perspective and tempus. It is determined whether the sentence contains first/second or third person pronouns and which tense is used. If consecutive sentences agree on both, they belong to the same segment. If the narrative perspective is in third person and the dominant tense is past, the segment is categorized as 'narrative'. Inside those sections a penalty is introduced, so that more than a single weak indicator has to be found to assume DS.

2) Determine sentences containing DS with high probability: After the sections have been introduced, each sentence is classified as either "directspeech" or "other". The algorithm utilizes a scoring mechanism with manually defined features (e.g. imperative mode, interjections, tempus shift). These features (as well as their weights) were created by inspecting parts of the DROC corpus. They are optimized to identify DS with high precision. The intuition of this pass is that all sentences which are rather "obvious" (at least to the human reader) are now correctly marked as DS.

3) Use the sentences from step 2) as anchors to expand the annotation: The next pass expands the instances within their previously detected coherent sections. It starts at an anchor sentence and adds the adjacent sentences to the "directspeech" annotation if tempus and narrative perspective still agree. The border of a coherent section serves as a definitive stopping point for this process.

The resulting sentences are the final result for an evaluation based on sentence borders.

4) If token-level accuracy is required: Remove sentence parts that are considered frames from the annotation: If the exact span on a token level is required, the sentences are split into sub-sentences (enclosed by commas) and whenever a frame of a DS is detected, this sub-sentence is removed, yielding the final result of the detection.

Evaluation

For evaluation, two performance metrics are applied: 1) Sentence level accuracy → a true positive is achieved by correctly predicting whether DS is contained in the sentence 2) Token level accuracy → each token is evaluated individually.

The results on the different corpora are depicted in the table below. We report micro accuracy values to not favor documents which are much shorter than others.

Corpus	Sentence level accuracy (in %)	Token level accuracy (in %)
DROC_red	80.5	80.4
Gutenberg	85.4	86.8
RW_fict	81.9	81.7
RW_nonfict	60.8	53.4

Both accuracy metrics show very similar results on three corpora. An interesting fact is that the corpus which was used to create the rules (DROC_red) shows the worst results of all three fictional datasets. Gutenberg contains rather schematic narratives which appear to be easier compared to the other data sets. A more fine grained analysis shows that the best document for DROC_red yields a token level accuracy of 99.7% and the worst document an accuracy of 21.5%. The largest gap can be found in the RW_fict dataset with 0% accuracy for the worst and 100% accuracy for the best document. This shows that while in average the results are promising, there are still phenomena that need to be addressed separately.

For the only non-fictional corpus, RW_nonfict, the scores drop by a large margin. This is because the algorithm finds anchors in sections without DS and propagates those incorrectly to the surrounding section. Those incorrect detected anchors are sentence written in first person or rhetorical questions, which are mistaken as DS. This resulted in about 1800 false positives while only 89 sentences of DS were not detected.

Conclusion

We proposed a rule-based approach to detect DS without the help of any quotation markers. Our approach creates coherent sections which segment the documents. Specialized rules detect DS on the sentence level inside those segments. The annotation is then expanded from those anchor sentences. Post-processing removes frame sub-sentences to get an exact span for the utterance. Our evaluation shows that the results appear stable throughout different datasets in the fictional

domain and are comparable to the results achieved in related work. The tool even achieves a higher score compared to Jannidis et al. 2018 on the sentence level. The current algorithm still has issues with non-fictional texts and some types of fictional texts (especially romantic letters and reports written in first person singular) which suggests that it should be extended to detect the type of document in advance in order to classify in a more robust approach across different domains.

Fußnoten

1. http://gutenberg.spiegel.de/
2. https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release
3. www.redewiedergabe.de
4. https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/opennlp/tools/tokenize/Tokenizer.html
5. https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/opennlp/tools/sentdetect/package -summary.html
6. http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/
7. https://code.google.com/archive/p/mate-tools/downloads
8. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/14333/file/Jannidis_Figurenerkennung_Roman.pdf

Bibliographie

Blessing, Andre / Bockwinkel, Peggy / Reiter, Nils / Willand, Marcus (2016): *"Dramenwerkbank - Automatische Sprachverarbeitung zur Analyse von Figurenrede"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 281-284.

Brooke, Julian / Hammond, Adam / Hirst, Graeme (2015): *"GutenTag: An NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus"*, in: North American Chapter of the Association for Computational Linguistics – Human Language Technologies 42-47.

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie* (= Narratologia 47). Berlin: De Gruyter.

Dimpel, Friedrich Michael (2018): *"Narratologische Textauszeichnung in Märe und Novelle"*, in: **Bernhart, Toni / Willand, Marcus / Albrecht, Andrea / Richter, Sandra (eds.)**: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: De Gruyter 121-148.

Jannidis, Fotis / Zehe, Albin / Konle, Leonard / Hotho, Andreas / Krug, Markus (2018): *"Analysing Direct Speech in German Novels"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 114-118.

Krug, Markus / Weimer, Lukas / Reger, Isabella / Macharowsky, Luisa / Feldhaus, Stephan / Puppe, Frank / Jannidis, Fotis (2018): *"Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]"*, in: DARIAH-DE Working Papers Nr. 27 http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-2-9 [letzter Zugriff 27. September 2018].

Pouliquen, Bruno / Steinberger Ralf / Best, Clive (2007): *"Automatic Detection of Quotations in Multilingual News"*, in: International Conference 'Recent Advances in Natural Language Processing' – Proceedings 487-492.

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): 1-11.

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin (2018): *""Kann man denn auch nicht lachend sehr ernsthaft sein?" - Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 244-249.

Automatisierungs-potenziale in der qualitativen Diskursanalyse. Das Prinzip des „Filterns“

Koch, Gertraud
gertraud.koch@uni-hamburg.de
Universität Hamburg, Deutschland

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

Diskursanalytische Verfahren werden in vielen Disziplinen verwendet, so dass die Frage nach Automatisierungspotentialen in diesem Bereich für ganz unterschiedliche Geisteswissenschaften und auch qualitativ arbeitende Sozialwissenschaften relevant ist (vgl. grundlegend Foucault 1971 und 1973; Keller 2011). Die Frage wird aktuell in einem Teilprojekt des Verbundforschungsprojektes hermA (Gaidys et al. 2017) erforscht, aus dem dieser Beitrag hervorgeht.[1] Angesichts steigender digital vorliegender Textmengen, die nicht zuletzt über das Internet verfügbar sind, stellt sich die Frage nach Automatisierungspotentialen immer drängender. Die Perspektive der wissenssoziologischen bzw. -anthropologischen Diskursanalyse, die hier eingenommen wird, kann insofern als exemplarisch angesehen werden, wird aber je nach Schwerpunktsetzung auch Variationen aufweisen.

In der Soziologie und Wissensanthropologie wird die Diskursanalyse als Methode eingesetzt, um zu analysieren, wie sich (neue) gesellschaftliche Phänomene herausbilden und etablieren können, beispielsweise für ein vergleichsweise junges Phänomen wie die Telemedizin. Diese wird zunehmend als Lösung für den Ärztemangel in ländlichen Räumen aber auch zur Betreuung von chronisch kranken Patienten diskutiert und zwar in ganz unterschiedlichen Diskursarenen und durch verschiedene Akteure, vom Bundestag über Ärzte, Krankenkassen bis hin zu Patientenverbänden und den Patienten selbst. Mittels der Diskursanalyse kann herausgearbeitet werden, wie angesichts heterogener Interessenslagen der beteiligten Gruppen die verschiedenen Auffassungen von

Telemedizin sowie ihrer Notwendigkeit verhandelt werden, ob sich ggf. ein allgemeines gesellschaftlich weitgehend akzeptiertes Verständnis von Telemedizin herausbildet, sich schließlich konkrete Arbeitsweisen (Praktiken), institutionelle Zusammenhänge oder Organisationsformen sowie gesetzliche Regelungen verfestigen (materialisieren).

Spezifik der diskursanalytischen Datengrundlage

Für Diskursanalysen wird heterogenes Quellenmaterial verwendet, es können ganz unterschiedliche Textsorten eine Rolle spielen. In der dem Beitrag zugrunde liegenden Forschung zu Akzeptanzproblematiken der Telemedizin werden Webseiten (Homepages, Blogs, Foren, etc.), wissenschaftliche Beiträge und Bundestagsprotokolle analysiert. Potentiell können auch multimodale Daten wie Bilder oder audio-visuelles Material hinzugezogen werden. Zunächst erfolgt jedoch eine Beschränkung auf Textquellen, um die Datenmodellierung mit Methoden der Digital Humanities gezielt verfolgen zu können.

Die Korpuserstellung erfolgt für Diskursanalysen iterativ nach dem in der Grounded Theory (GT) formulierten Prinzip des Theoretischen Sampling, mit dem eine tendenziell unüberschaubare Datenmenge epistemologisch geleitet reduziert wird. Dieses Verfahren zeichnet sich dadurch aus, dass in mehreren zyklischen Prozessen Daten erhoben, annotiert und interpretiert werden. Dabei wird nach jedem Zyklus anhand der vorliegendem Ergebnisse über die weitere Datenerhebung entschieden (vgl. Glaser/ Strauss 1967). „Manuelle" Filterprozesse sind somit eine iterative Abfolge verschiedener interpretativer, methodisch geleiteter Arbeitsschritte: theoretisches Sampling, offenes Kodieren, selektives Kodieren, axiales Kodieren, theoriegeleitete Interpretation (Bryant/Charmaz 2007; Götzö 2014). Die Informationsfülle wird so sukzessive in eine für qualitative Forschungen handhabbare Größenordnung gebracht. Im Laufe des Forschungsprozesses entsteht so in interativen Prozessen der Datenerhbung und -interpretation ein relativ kleiner Datenkorpus, meist aus unterschiedlichen Quellen, die aussagekräftig für die Fragestellung sind (vgl. Strauss/ Corbin 1996). Es stellt sich die Frage, inwieweit diese „manuellen" Filterprozesse von heterogenen Textsorten durch automatisierte Verfahren im Sinne einer höheren Effizienz oder verbesserten analytischen Qualität ergänzt werden können. Die GT zielt nicht auf Repräsentativität der Daten, sondern auf Viabilität also eine hohe Passung in der Erklärungskraft (Glasersfeld 1997) unabhängig davon ob Daten digital oder analog vorliegen.

Filtern als Prinzip methodisch geleiteter Analyse

Im Sinne der iterativen Korpuserstellung wird untersucht, wie sich manuelles und automatisiertes Filtern miteinander verbinden lassen und was die jeweilige Form des Filterns auszeichnet. „Filtern" wird dabei als ein Arbeitsschritt in der wissenschaftlichen Analyse von Daten verstanden, welches nach methodischen Prinzipien umgesetzt und zur Reduktion der im Alltag beobachtbaren Komplexität eingesetzt wird. Bei den automatisierten Ansätzen werden neben Information Retrieval (vgl. Klinke 2017; Manning et al. 2009) auch strukturelle Ansätze des Data und Text Mining für die Reduktion verfügbarer Informationen eingesetzt. Manuell wird entsprechend der Grounded Theory nach dem Prinzip des Theoretical Sampling gearbeitet, um wesentliche, also soziale Wirklichkeit setzende Dokumente bzw. Textabschnitte für die Analyse zu identifzieren. Während es bei automatisierten Ansätzen in der Tendenz darum geht, einen möglichst vollständigen Korpus aller relevanten Dokumente zu generieren, der wiederum eine Basis für weitere automatische Filteransätze bietet, erfolgt die manuelle Korpuserstellung hochgradig selektiv nach Relevanz bzw. Viabilität.

Filtern manuell

Beim manuellen Filtern in der qualitativen Diskursanalyse ist ein hohes Maß an Vorwissen notwendig, welches sich auf mögliche Akteure, Diskursarenen und Kontexte des Themas, hier der Telemedizin bzw. der damit verbundenen Akzeptanzproblematiken, bezieht. Auf der Basis dieses Vorwissens wird der Einstieg in die Frage möglich, wo überhaupt Quellen für die Analyse des Phänomens zu finden sind. Dabei wird heute nicht nur „manuell" gefiltert, sondern die Hilfe von Suchmaschinen im Internet oder Suchabfragen von Stichwörtern in Archivkatalogen in Anspruch genommen. Allerdings sind damit die Informationen weiterhin tendenziell unüberschaubar und auch hinsichtlich ihrer Relevanz höchst heterogen. Ebenso ist offen, inwieweit tatsächlich alle relevanten Akteure und Diskursarenen erfasst worden sind, so dass eine Vielzahl an weiteren manuellen Filterprozessen vorgenommen werden müssen. Dabei wird das Wissen des Forschenden zum Thema stets erweitert, so dass die (stetig wachsende) Expertise der Forschenden in dem Themenfeld eine wesentliche Voraussetzung für eine hohe analytische Qualität der Diskursanalyse darstellt. Die in der Grounded Theory angelegten methodischen Arbeitsschritte profitieren wesentlich von dieser stetig wachsenden Expertise, sind dabei jedoch auch zur Objektivierung der vom Forschenden formulierten Hypothesen, im Sinne von Falsifizierungen oder Bestätigungen, unerlässlich (vgl. Glaser 1978).

Filtern maschinell

Maschinelles Filtern beruht auf strukturellen Analysen von Sprache und bedarf vielfältiger Ressourcen (Ontologien, Wörterbücher, Tools, Korpora). Ein rein automatisiertes Filtern zum Thema Telemedizin ist aufgrund fehlender Ressourcen auf dem aktuellen Wissensstand nicht möglich. Insgesamt darf man davon ausgehen, dass dies meist der Fall ist, wenn neue gesellschaftliche Phänomene auftreten und sich die soziale Wirklichkeit, die Themen und die Sprache wandeln. Es geht insofern darum auszuloten, unter welchen Umständen und wo im Forschungsprozess automatisierte Filterprozesse, zielführend aufgegriffen werden können.

Entsprechend dieser Hypothese wurden bisher unterschiedliche halbautomatisierte Verfahren für die Unterstützung der qualitativen Diskursanalyse mit ihren verschiedenen Arbeitsschritten erprobt: Suchmaschinen, Suchfunktionen, verschiedene Webcrawler, sowie manuelle und automatisierte Annotationen mit proprietären und open-source Tools. Auch die Vorbereitung automatisierter Filterprozesse spielt für die Verfahren eine wichtige Rolle, insbesondere a) die Erstellung von Wortfeldern zur Spezifizierung der Filter-Anwendung, b) die Aufbereitung von Dokumenten für die automatisierte Analyse, c) die Klärung von Arbeitsweisen verschiedener Crawler – etwa iCrawl[2], Apify[3] und IssueCrawler[4] oder das auf Crawlen basierende Webarchiv der Deutschen Nationalbibliothek – und wie diese für eine zielführende Diskursanalyse aufgesetzt werden können, d) die Klärung der Datenlage in bestehenden Korpora wie dem Dokumentations- und Informationssystem des Deutschen Bundestages[5], den alternativ angebotenen Open Data-Beständen[6] sowie dem GermaParlTEI-Korpus[7] und der Aufarbeitung von Crawling-Ergebnissen. Für eine Vorbereitung der Auswertung erfolgt zudem e) die Erprobung und der Vergleich verschiedener Annotationstools von

proprietären Programmen der qualitativen Datenanalyse, welche eine Computerunterstützung für Geistes- und Sozialwissenschaftler*innen zugänglich machen und bereits weit verbreitet sind (vgl. Gasteiger/Schneider 2014; Sattler 2014) und open-source Optionen wie etwa CATMA[8] sowie f) Koreferenzannotationen mittels CoRefAnnotator[9], um Analysen zur Konkordanz und Netzwerkanalysen vorzubereiten. Dabei stehen die spezifischen Potentiale und Probleme für die Korpuserstellung und -auswertung bei diskursanalytischem Datenmaterial im Zentrum.

Filtern als Automatisierungspotential in der Diskursanalyse

Für einen lösungsorientierter Ansatz zur Anreicherung von qualitativen hermeneutischen Verfahren der Diskursanalyse mit den strukturell arbeitenden Ansätzen von Methoden der Digital Humanities hat sich bisher gezeigt, dass es vor allem einfache Verfahren des Information Retrival sind, die unterstützend für die qualitative Forschung nach GT wirken und auch gegenwärtig eingesetzt werden, allerdings vor allem auf der Ebene generischer Tools (Suchfunktionen, automatisiertes Abrufen und Speichern von Dokumenten, Wordfelder). In dem Moment, in dem Anpassungen von Tools notwendig werden, relativieren sich die Vorteile automatisierter Verfahren, insbesondere weil sich die scheinbare Fülle der Informationen im Sinne sogenannter „big data“ bei näherem Ansehen der Daten bisher nicht erfüllt hat, diese aufgrund weniger valider bzw. viabler Textstellen rasch zu „small data“ werden, die wiederum leichter qualitativen Analysen zugänglich sind. Dies hängt wohl zentral mit der Neuheit des Phänomens Telemedizin zusammen, was für Untersuchungsgegenstände der Kulturanthropologie als typisch angesehen werden kann. Gleiches gilt für die Heterogenität unterschiedlicher Textsorten und entsprechend die Herausforderungen automatischer Aufbereitungsschritte. Der gegenwärtige Lerneffekt im Projekt bezieht sich so insbesondere auf die Spezifizierung, wo und wie automatisierte Verfahren in der Diskursanalyse überhaupt sinnvoll eingebettet werden können, sowie auf die Erfahrungen hinsichtlich der Erprobung verschiedener Ansätze des Filterns inklusive der dafür notwendigen Aufbereitungen des generierten Materials und der iterativ ineinander greifenden Schritte der automatischen und manuellen Filterung.

Fußnoten

1. Der Forschungsverbund „Automatisierte Modellierung hermeneutischer Prozesse – Der Einsatz von Annotationen für sozial- und geisteswissenschaftliche Analysen im Gesundheitsbereich (hermA)“ ist ein interdisziplinäres Projekt an der Universität Hamburg, der Technischen Universität Hamburg und der Hochschule für Angewandte Wissenschaften Hamburg, das durch die Landesforschungsförderung Hamburg finanziert wird. Das Teilprojekt „Automatisierungspotenziale hermeneutischer Prozesse in der Diskursethnographie zu Akzeptanzproblematiken der Telemedizin“ ist ein Beitrag aus der Kulturanthropologie. Vgl. https://www.herma.uni-hamburg.de/.
2. http://icrawl.l3s.uni-hannover.de/
3. https://www.apify.com/page-analyzer
4. https://www.issuecrawler.net/
5. http://dipbt.bundestag.de/dip21.web/bt
6. https://www.bundestag.de/service/opendata
7. https://github.com/PolMine/GermaParlTEI
8. http://catma.de/
9. https://github.com/nilsreiter/CorefAnnotator/releases

Bibliographie

Bryant, Antony; Charmaz, Kathy (Hg.): *The SAGE handbook of Grounded Theory*, Los Angeles 2007.

Foucault, Michel: *Archäologie des Wissens*, Frankfurt a.M. 1973.

Foucault, Michel (1971): *Die Ordnung des Diskurses*, Frankfurt a.M. 1991.

Keller, Reiner (2005): *Wissenssoziologische Diskursanalyse. Grundlegung eines Forschungsprogramms*, 3. Auflage Wiesbaden 2011.

Gaidys, Uta / Gius, Evelyn / Jarchow, Margarete / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Zinsmeister, Heike: *Project Descripition. HermA: Automated Modelling of Hermeneutic Processes*, in: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Glaser, Barney G.: *Theoretical Sensitivity. Advances in the Methodology of Grounded Theory*, Mill Valley, Calif. 1978.

Gasteiger, Ludwig / Schneider, Werner: *Diskursanalyse und die Verwendung von CAQDA-Software*, in: **Angermuller, Johannes / Nonhoff, Martin / Herschinger, Eva / Macgilchrist, Felicitas / Reisigl, Martin / Wedl, Juliette / Wrana, Daniel / Ziem, Alexander (Hg.):** *Diskursforschung. Ein interdisziplinäres Handbuch. Band 2: Methoden und Analysepraxis, Perspektiven auf Hochschulreformdiskurse*, Bielefeld 2014, S. 852–872.

Glaser, Barney G. / Strauss, Anselm L. (1967): *Grounded Theory. Strategien qualitativer Forschung*, Bern 2010.

Glasersfeld, Ernst von: *Radikaler Konstruktivismus* Frankfurt 1997.

Götzö, Monika: *Theoriebildung nach Grounded Theory*, in: **Bischoff, Christine / Oehme-Jüngling, Karoline / Leimgruber, Walter (Hg.):** *Methoden der Kulturanthropologie*, Bern 2014, S. 444–458.

Keller, Reiner (2005): *Wissenssoziologische Diskursanalyse. Grundlegung eines Forschungsprogramms*, 3. Auflage Wiesbaden 2011.

Klinke, Harald: *Information Retrieval*, in: **Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (Hg.):** *Digital Humanities. Eine Einführung*, Stuttgart 2017, S. 268–278.

Manning, Christopher D. / Raghavan, Prabhakar / Schütze, Hinrich: *Introduction to Information Retrieval* Cambridge 2009.

Sattler, Simone: *Computergestützte qualitative Datenbearbeitung*, in: **Bischoff, Christine / Oehme-Jüngling, Karoline / Leimgruber, Walter (Hg.):** *Methoden der Kulturanthropologie*, Bern 2014, S. 476–487.

Über die Ungleichheit im Gleichen. Erkennung unterschiedlicher Reproduktionen desselben Objekts in kunsthistorischen Bildbeständen

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland

Es finden sich mittlerweile mannigfaltige Studien, die Objekte historischen Ursprungs in den Fokus rücken und zwischen ihnen bestehende Relationen und Ähnlichkeitsverhältnisse mathematisch abzubilden versuchen (darunter Bergel et al., 2013; Monroy et al., 2014; Hentschel et al., 2016). Eine fundamentale Herausforderung *per se* heterogener historischer Inventare bleibt dabei unberührt: digitale Reproduktionen ein und desselben Objekts, sogenannte *Near-duplicates* oder *Near-replicas*, die als separate Einträge in Aggregatordatenbanken vorgehalten werden und sich bspw. hinsichtlich ihres Farbstichs oder ihrer Helligkeit unterscheiden. Aufgrund nicht-standardisierter, teils unstrukturierter oder selten kontrollierten Vokabularen zugeordneter Metadaten ist es zumeist nicht oder nur mit größtem Aufwand möglich, derartige „Kopien" auf Basis textueller Information sowohl nachhaltig als auch zuverlässig ohne händische Nacharbeit zu verknüpfen.

Unser Ansatz zielt auf dreierlei: erstens die automatische Zusammenführung unterschiedlicher Reproduktionen desselben Objekts; die zweitens das *Retrieval* variierender Reproduktionen erlaubt, um weiterführende Analysen, z. B. quellenkritischer Art, über eben jenes Objekt anstoßen zu können; und drittens die Extraktion textueller Information von Objekten, die ausschließlich visuell, d. h. als digitales Bild, vorliegen und, im Sinne einer *Reverse Image Search*, mit einem bereits annotierten Inventar von Objekten abzugleichen sind. Auf diese Weise wird nicht nur eine effiziente, bildbasierte Suche in Datenbanken insbesondere (aber nicht ausschließlich) kunsthistorischer Objekte ermöglicht, sondern nahezu *Bias*-freie statistische Untersuchungen, wie sie durch den Einfluss häufig reproduzierter Werke bislang nicht gegeben waren.[1]

Methode

Wir stützen uns hauptsächlich auf *Scale Invariant Feature Transform* (*SIFT*; Lowe, 1999; Lowe, 2004), um aus Bildern historischer Objekte lokale Schlüsselpunkte (*Keypoints*) zu extrahieren und mittels Deskriptoren (*Descriptors*) weiterverarbeiten zu können. Schlüsselpunkte bilden einzelne Interessenregionen eines Bildes ab, die statistisch, aber nicht notwendigerweise semantisch relevante Merkmale tragen, und stellen sie in einem 128-dimensionalen Histogramm dar. Im mathematischen Sinne sind sie Extremwerte, die über einen Raum mehrfach skalierter und mit einem Gaußfilter geglätteter Bilder ermittelt werden. Mit üblichen Ähnlichkeits- und Distanzmaßen, z. B. der euklidischen Distanz, ist es so möglich, die Nähe zwischen zwei Schlüsselpunkten zu quantifizieren, und demnach auch, über die Summe der *matchenden* Schlüsselpunkte, die Nähe zwischen zwei Digitalisaten; wobei anzunehmen ist, dass die Anzahl der übereinstimmenden Schlüsselpunkte für variierende Abbildungen desselben Objekts höher ist als für Abbildungen unterschiedlicher Objekte.

Ein Bild wird je nach Größe und Detailgrad mit Hunderten bis Tausenden derartiger Schlüsselpunkte assoziiert. Daraus resultierende computationale Kosten fangen wir durch drei Erweiterungen des Verfahrens ab. Erstens reduzieren wir die Dimensionalität der Deskriptoren mittels *Principal Component Analysis* (*PCA*). Im Gegensatz zu Ke und Sukthankar (2004) greifen wir nicht in den Deskriptionsprozess selbst ein, sondern ermitteln den Eigenraum auf Basis der standardmäßig durch *SIFT* eruierten Deskriptoren. Zweitens verringern wir die Anzahl der Schlüsselpunkte, indem wir zunächst die mit dem höchsten Kontrast auswählen und auf Basis dessen jene filtern, welche die größten Interessenregionen charakterisieren. Damit werden auf flächenmäßig kleinen Arealen zu findende, kontrastreiche Schlüsselpunkte getilgt, die bspw. in textuellen Ergänzungen von Kupferstichen auftreten und für das *Matching* irrelevant bis schädlich sind. Drittens setzen wir mit *Hierarchical Navigable Small World* (*HNSW*; Malkov und Yashunin, 2016) einen *Approximate Nearest Neighbor*-Ansatz mit polylogarithmischer Komplexität ein, der in aktuellen repräsentativen Benchmarks andere Graph-basierte Ansätze in Präzision und Schnelligkeit übertrifft (Aumueller et al., 2018). Eine adaptive Intensitätskorrektur jedes Bildes durch *Contrast Limited Adaptive Histogram Equalization* (*CLAHE*; Zuiderweld, 1994) wird vor der Extraktion der Schlüsselpunkte durchgeführt, um stark über- oder unterbelichtete Reproduktionen anzupassen.

Daten

Ein geeigneter *Gold Standard* wird in drei Schritten etabliert. Zunächst ziehen wir eine Zufallsstichprobe von 3.581 kunsthistorischen Objekten, die in der Datenbank *ArteMIS* des Instituts für Kunstgeschichte der Ludwig-Maximilians-Universität München verzeichnet sind[2] und einen angemessenen Querschnitt verschiedener kunsthistorischer Stile und Epochen erlauben; Holzschnitte sind ebenso inkludiert wie realistische Landschaftsmalerei und Werke des französischen Impressionismus.[3] In einem zweiten Schritt speisen wir Titel und Künstler jener Objekte in die 94 Datenbanken kumulierende *Application Programming Interface* von Prometheus.[4] Da in den jeweiligen Suchergebnissen nicht nur Digitalisate ein und desselben Objekts zu finden sind – Vorzeichnungen und aufgrund ihrer Metadaten ähnliche Reproduktionen sind auch darunter –, schließen wir einen dritten Schritt an, in dem auf unterschiedliche Objekte referenzierende Abbildungen manuell entfernt werden. Es verbleiben 9.934 Reproduktionen. Ein derart selegiertes Digitalisat trägt einen eineindeutigen Identifikator, der sowohl auf das es abbildende

Objekt zeigt als auch auf die an das Digitalisat gekoppelten Metadaten weist.

Um weitere in der bildarchivarischen Praxis existente, aber durch zuvor extrahierte *reale* Digitalisate unzureichend abgedeckte Modifikationen, bspw. größere Änderungen des Kontrasts oder der Sättigung eines Bildes, untersuchen zu können, generieren wir 278.152 zusätzliche, sogenannte *synthetische* Kopien. Jede Ursprungsreproduktion wird dementsprechend dupliziert und 28 mathematischen Transformationen unterzogen; ähnlich zu jenen in Ke et al. (2004), Qamra et al. (2005) und Foo et al. (2007). Unter anderem modelliert werden sich in ihrer Stärke unterscheidende nicht-lineare Verzerrungen, die Wölbungen nahe des Buchrückens von Fotografien kunsthistorischer Publikationen suggerieren.

Evaluation

Drei im *Information Retrieval* gewöhnliche Gütekriterien dienen der Evaluation der angewandten Methoden: Precision, Recall und F_1-Maß. Wir gehen wie folgt vor. Der Satz an Objekten, die mit realen und synthetischen Reproduktionen assoziiert sind, wird unterteilt in 25 zufällig separierte Trainings- und Teststichproben, wobei jeweils 80 Prozent der Objekte der Trainings- und 20 Prozent der Teststichprobe zuzuordnen sind. Auf Basis der 25-fachen Kreuzvalidierung erhalten wir durchschnittliche Werte für Precision, Recall und F_1-Maß, die von einem jeweils gegebenen Schwellenwert abhängen, der die minimale Anzahl der Schlüsselpunkte bezeichnet, die zwischen zwei Digitalisaten übereinstimmen müssen, damit diese als unterschiedliche Reproduktionen desselben Objekts gelten und entsprechend zusammengeführt werden können. Der jeweils für eine Parameterkonstellation optimale Schwellenwert bildet sich aus dem Modus der 25 Einzelschwellenwerte, die mit dem höchsten F_1-Maß verknüpft sind.

Ergebnisse

Wir stellen fest, dass aufgrund des hohen Anteils in den Reproduktionen enthaltener digitaler Artefakte – unter anderem Bildrauschen und Unschärfe –, Konfigurationen mit im Vergleich zu Standardwerten höherem Skalenparameter σ, der die Stärke des in *SIFT* angelegten gaußschen Weichzeichners reguliert, und niedrigeren Schwellenwerten, welche die Aufnahme von kontrastarmen oder auf Kanten situierten Schlüsselpunkten steuern, zu bevorzugen sind. Eine Reduktion der Dimensionalität der Deskriptoren und der Anzahl der Schlüsselpunkte auf jeweils 50 führt zu marginalen Einbußen in Precision und Recall, steigert jedoch maßgeblich die Performanz und mindert den notwendigen Speicherbedarf. Eine so klassifizierte, aus 500 Objekten bestehende Zufallsstichprobe resultiert in Precision = 0,9857, Recall = 0,9820 und F_1-Maß = 0,9839, wenn für *HNSW* moderate Kompromisse zwischen der Geschwindigkeit, die der Aufbau des Index und die eine Suche im Index benötigt, formuliert werden; mindestens 7 näherungsweise übereinstimmende Schlüsselpunkte sind erforderlich, damit Digitalisate als demselben Objekt zugehörig erkannt werden. Größere Einbrüche in Recall, d. h. bis zu 5 Prozentpunkte, sind für stärkere Farbänderungen und nicht-lineare Verzerrungen zu beobachten. Insbesondere drei Gruppen von Objekten erfordern weitere Anpassungen: Digitalisate von Druckgrafiken mit hohen Kontrastunterschieden; Reproduktionen, die Rahmen oder rahmenähnliche Strukturen abbilden; Werke des Impressionismus und nicht-gegenständliche oder diffuse Werke, die unterdurchschnittlich viele Schlüsselpunkte, teilweise nur bis zu 10, produzieren.

Auch ohne zusätzliche Modifikationen zeigt sich, dass die hier präsentierte Methode hinreichend exakte Ergebnisse erwarten lässt und kaum, oder lediglich im Falle hoch spezialisierter Korpora, manuelle Adjustierungen erfordert; selbst wenn stärkere Abweichungen in Kontrast oder Sättigung auftreten. Durch die Integration eines *Approximate Nearest Neighbor*-Ansatzes ist weiterhin gewährleistet, dass das Verfahren auch auf größere historische Bildbestände skaliert.

Fußnoten

1. Dies schließt anderweitige Verzerrungen, bspw. einen *Selection Bias*, natürlich nicht aus.
2. Eine Online-Schnittstelle ist zu erreichen unter http://artemis.lmu.de/ (25.09.2018).
3. Ausgenommen werden Reproduktionen eindeutig dreidimensionaler Objekte, z. B. Plastiken und Skulpturen, da sich diese zusätzlich durch den bei der Aufnahme eingenommenen Blickwinkel unterscheiden können und folglich gesondert zu evaluieren wären.
4. http://www.prometheus-bildarchiv.de/ (25.09.2018).

Bibliographie

Aumueller, Martin / Bernhardsson, Erik / Faitfull, Alec (2018): *ANN Benchmarks*, http://sss.projects.itu.dk/ann-benchmarks/index.html (26.09.2018).

Bergel, Giles / Franklin, Alexandra / Heaney, Michael / Arand-Jelovic, Relja / Zisserman, Andrew / Funke, Donata (2013): *„Content-based Image Recognition on Printed Broadside Ballads. The Bodleian Libraries' Imagematch Tool"*, in: Proceedings of the IFLA World Library and Information Congress.

Foo, Jun Jie / Zobel, Justin / Sinha, Ranjan (2007): *„Clustering Near-duplicate Images in Large Collections"*, in: Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval 21–30.

Hentschel, Christian / Wiradarma, Timur P. / Sack, Harald (2016): *„An Approach to Large Scale Interactive Retrieval of Cultural Heritage"*, in: Proceedings of the 23th IEEE International Conference on Image Processing 3693–3697.

Ke, Yan / Sukthankar, Rahul (2004): *„PCA-SIFT. A More Distinctive Representation for Local Image Descriptors"*, in: Proceedings of the IEEE International Conference on Computer and Pattern Recognition 506–513.

Ke, Yan / Sukthankar, Rahul / Huston, Larry (2004): *„An Efficient Parts-based Near-duplicate and Sub-image Retrieval System"*, in: Proceedings of the >12th ACM International Conference on Multimedia 869–876.

Lowe, David G. (1999): *„Object Recognition from Local Scale-invariant Features"*, in: Proceedings of the 7th IEEE International Conference on Computer Vision 1150–1157.

Lowe, David G. (2004): *„Distinctive Image Features from Scale-invariant Keypoints“*, in: International Journal of Computer Vision 60 (2): 91–110.

Malkov, Yury A. / Yashunin, Dmitry A. (2016): *Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs*.

Monroy, Antonio / Bell, Peter / Ommer, Björn (2014): *„Morphological Analysis for Investigating Artistic Images“*, in: Image and Vision Computing 32 (6): 414–423.

Qamra, Arun / Meng, Yan / Chang, Edward Y. (2005): *„Enhanced Perceptual Distance Functions and Indexing for Image Replica Recognition“*, in: IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3): 379–391.

Zuiderveld, Karel (1994): *„Contrast Limited Adaptive Histogram Equalization“*, in: Academic Press 474–485.

Besucherexperimente im Science Center: Welche Einsichten in die Herstellung von Wissen in Interaktion werden erst durch die Zusammenschau von Audio-, Video- und Eye-Tracking-Daten möglich?

Kesselheim, Wolfgang

wolfgang.kesselheim@ds.uzh.ch
Universität Zürich, Schweiz

Hottiger, Christoph

christoph.hottiger@access.uzh.ch
Universität Zürich, Schweiz

Die interaktive Herstellung von Wissen zu Naturphänomenen im Science Center ist gleich in mehrfacher Hinsicht multimodal. Nicht nur nutzen die Besucherinnen und Besucher zur gemeinsamen Konstruktion von Wissen die unterschiedlichsten körperlichen Ausdrucksressourcen („embodied modes“): Sprache, Gesten, Blick, Körperhaltung und -position etc. Sie bedienen sich auch multimodaler Ressourcen im Ausstellungsraum: Sie manipulieren Objekte, betrachten Grafiken, lesen Anweisungs- und Erklärungstexte und machen sie für ihre interaktive Wissenskonstruktion relevant (vgl. Kesselheim 2017, 2012). Möchte man also die Praktiken rekonstruieren, mit denen sich Besucherinnen und Besucher im Science Center gemeinsam Wissen erarbeiten, ist das im vollen Umfang nur dann möglich, wenn man in der Analyse nachzeichnet, wie alle beteiligten Modi aufeinander bezogen und miteinander verwoben werden.

Diese These möchten wir in unserem Vortrag mit Hilfe einer Reihe von exemplarischen Kurzanalysen belegen, indem wir anhand der analysierten Beispiele aufzeigen, welche konkreten Einsichten in die interaktiven Praktiken der Wissenskonstruktion in Science Centern erst möglich sind, wenn man die Multimodalität des Phänomens ernst nimmt.

Der Vortrag referiert Ergebnisse eines vom Schweizer Nationalfonds geförderten Forschungsprojekts mit dem Titel „Interactive Discoveries: A video and eye-tracking based study on knowledge construction in the science center“. Dieses Projekt geht mit den Methoden der multimodal erweiterten Konversationsanalyse der Frage nach, wie Besucherinnen und Besucher eines modernen, Naturwissenschaft und Technik gewidmeten Mitmachmuseums Naturphänomene ‚entdecken‘, wenn sie im Zuge eines gemeinsamen Gangs durch die Ausstellungsräume die dort aufgestellten Hands-on-Exponate und Experimentierstationen nutzen. Wissenskonstruktion wird, das ist die grundlegende Idee des Projekts, im gemeinsamen Handeln und Sprechen beobachtbar – sie muss nicht, wie dies typischerweise gemacht wird, im Nachhinein und indirekt erhoben werden (etwa per Fragebogen oder Interview).

Der Vortrag folgt einer doppelten Bewegung:

Die erste Reihe von Beispielanalysen beginnt mit reinen Audioaufnahmen. Dann wird gezeigt, wie sich Schritt für Schritt ein besseres Verständnis der musealen Wissenspraktiken ergibt, wenn man die Audiodaten erst durch Video- und schließlich durch Eye-Tracking-Daten ergänzt.

Die zweite Serie von exemplarischen Analysen geht den umgekehrten Weg. Anhand von reinen Eye-Tracking-Daten wird aufgezeigt, inwiefern diese ein unzureichendes, ja verfälschendes Bild des menschlichen Blicks geben, jenes Phänomens also, das sie mess- und beobachtbar machen sollen. Dann wird Schritt für Schritt vorgeführt, wie die Hinzunahme von Video- und Audio-Aufnahmen bessere Einsichten in die Rolle des Blicks für die interaktive Wissenskonstruktion ermöglichen.

Die erste Analysereihe illustriert die folgenden Resultate unserer Forschung:

- Die Arbeit ausschließlich mit Audioaufnahmen (wie sie in der Forschung zu Fachsprache und -kommunikation gang und gäbe waren), bestärkt den Sprach- *bias* der Linguistik: Wissenskonstruktion wird als sukzessive Formulierung von Regelhaftigkeiten, als Erschließungsarbeit an Fachtermini, als konversationelle Aushandlung von unbestreitbaren Sachverhalten verstanden; gleichzeitig erscheint der gemeinsame Science-Center-Besuch als „Gespräch“.
- Die Hinzunahme von Video macht die Körper der Besucherinnen und Besucher sichtbar und ermöglicht es, deren Beitrag zur Wissenskonstruktion in den Blick zu nehmen: die gemeinsame Manipulation von Objekten, die Stabilisierung von erfolgreichem Handeln, die Nutzung von Anleitungs- und Erklärungstexten im Raum usw.; gleichzeitig werden die Praktiken der Besucherinnen und Besucher nicht mehr als „Gespräch“, sondern klar als gemeinsames „Experimentieren“ beschreibbar.
- Durch Eye-Tracking schließlich wird beobachtbar, wie individuelle Wahrnehmungsprozesse in sozial-kommunikative Prozesse der Aushandlung der wahrgenommenen Phänomene münden.

Die zweite Analysereihe illustriert die Ergebnisse unserer Arbeit mit mobilen Eye-Trackern und die Rolle des Blicks in der Besucherinteraktion:

- Anhand von reinen Eye-Trackingdaten problematisiert unser Vortrag die Frage, wie die Grenzen „eines" Blicks zu definieren sind und wie dessen Bedeutung festgestellt werden kann. – Anders als die Leseforschung, die die Blickrichtung plausibel als indirektes Maß für Interpretationsprozesse auffasst (Clifton Jr., Charles et al. 2016), beantwortet Eye-Tracking in Interaktionsstudien nur die Frage wohin geblickt wurde, aber weder was wirklich gesehen, noch *als was* das Gesehene gesehen wurde.
- Wir zeigen, wie die Hinzunahme von Audiodaten (Sprache!) und die Dokumentation der Körper der Besucherinnen und Besucher per Video Hinweise darauf gibt, was die Blicke für die Interaktionsteilnehmer jeweils bedeuten; und wir zeigen, wie die Hinzunahme von Audio und Video untersuchbar macht, wie ‚sozial dargestellter Blick' und gemessener Blick auseinanderklaffen. Blick ist, wie wir zeigen werden, keineswegs nur eine Frage der Augenrichtung (wie in Eye-Tracking-Studien oft vorausgesetzt wird), sondern eine multimodale Gestalt, für die die Ausrichtung von Becken, Schultern und Kopf eine ebenso große Bedeutung haben kann wie die Blickrichtung.

Die feinkörnigen, qualitativen Analysen basieren auf der Methode der linguistischen Konversationsanalyse (s. etwa Sidnell 2010; Gülich et al. 2008) und sind deren strikter Oberflächenorientierung und deren Grundüberzeugung verpflichtet, dass die soziale Wirklichkeit in Interaktion durch das gemeinsame Handeln der Interaktionsteilnehmenden hergestellt wird.

Dies gilt auch für den Zugang zum Phänomen *Wissen*. Aus konversationsanalytischer Perspektive interessiert Wissen nicht als innerer Zustand, als eine individualpsychologische Repräsentation im Kopf eines Besuchers oder einer Besucherin, sondern als ein Phänomen der kommunikativen ‚Oberfläche': eine Gewissheit über das Zutreffen bestimmter Sachverhalte, die in der Interaktion erarbeitet und als gültig ausgehandelt wird (vgl. Bergmann und Quasthoff 2010). Diese Perspektive auf Wissen lenkt den Blick *weg* von der Frage, welche fachlich korrekten Aussagen Besucherinnen und Besucher vor oder nach einem Science-Center-Besuch über ein bestimmtes Phänomen treffen können, und *hin* zu der grundlegenderen Frage, über welche alltäglichen Methoden (die „Ethnomethoden", s. Gülich 2001) die Besucherinnen und Besucher verfügen, um Wissen aus ihrer räumlichen Umwelt zu gewinnen. Dieser Blick weg vom Wissen als Produkt hin zu den Methoden der Wissensgeneration erlaubt es, einen Einblick in den grundlegenden Mechanismus zu gewinnen, dem die Science-Center-Bewegung verpflichtet ist: dem „Discovery Learning" (Eisenberg 2001), dem selbstgesteuerten Lernen am Objekt.

Zu unserer Datengrundlage:

Materialbasis dieses Vortrags ist ein Audio-, Video- und Eye-Tracking-Korpus, das im Rahmen des oben erwähnten Forschungsprojekts erhoben worden ist. Es dokumentiert die ungesteuerte Nutzung von Hands-on-Exponaten und Experimentierstationen im Swiss Science Center Technorama (Winterthur, Schweiz) durch Gruppen von zwei bis maximal vier Besucherinnen und Besuchern.

Bei der Erhebung der Daten wurde auf größtmögliche Natürlichkeit geachtet: Bei den Gefilmten handelt es sich um authentische Besucher, die im Verlauf ihres gemeinsamen Besuchs von den Forschenden angesprochen worden sind. Ihnen wurden keine Vorgaben darüber gemacht, welche Exponate sie anschauen sollten, und ihre Instruktion bestand lediglich darin, mit ihrem Besuch fortzufahren. Obwohl den Versuchspersonen natürlich bewusst ist, dass ihr Verhalten im Rahmen eines Forschungsprojekts dokumentiert wird, legen bisherige Analysen nahe, dass diese Tatsache nur einen geringen Einfluss hat. Das mag daran liegen, dass viele der uns interessierenden Phänomene (Blick, Positionierung des Körpers usw.) sich nur schwer bewusst steuern lassen. Das Korpus dokumentiert das Verhalten von über 200 Besucherinnen und Besuchern. Insgesamt umfasst es mehr als 30 Stunden Aufnahmen, davon ca. die Hälfte mit Eye-Tracking-Brillen.

Für die Analyse wurden die HD-Aufnahmen der beiden Feldkameras, die das Interaktionsgeschehen aus unterschiedlichen Entfernungen dokumentieren, sowie – wenn vorhanden – die Blick-Videos der beiden Eye-Tracker über den Ton synchronisiert und mithilfe des Schnittprogramms FinalCut zu einer Split-Screen-Darstellung zusammengefasst, die dann die Grundlage der Transkription der Interaktion bildete (auf Basis von GAT2, s. Selting et al. 2009). Die Dateien, aus denen jeder Split-Screen-Clip zusammengesetzt ist, wurden zusammen mit den Transkriptionen, Analysebeobachtungen sowie den Metadaten zur Aufnahmesituation und den Aufgenommenen in einer FileMaker-Datenbank verzeichnet, sodass die vielfältigen multimodalen Bestandteile der Daten gezielt durchsucht und aufeinander bezogen werden können.

Welchen Beitrag leistet unser Vortrag zur aktuellen Forschung?

Indem der Vortrag Methoden in den Mittelpunkt stellt, mit denen die Besucherinnen und Besucher in ihrer Wissenskonstruktion Verbalität (etwa die Formulierung von Regelhaftigkeiten) und andere körpergebundene Modi (Gesten, Manipulation von Objekten usw.) verbinden, erweitert er das Untersuchungsspektrum der linguistischen Forschung zur Vermittlung von Wissenschaft (etwa im Paradigma der "Transferwissenschaft", begründet durch Wichter und Antos 1999, oder zur Experten-Laien-Kommunikation, s. etwa Birkner und Ehmer 2013) um eine dezidiert multimodale Perspektive.

Diese multimodale Perspektive auf die Konstruktion von Wissen fehlte bisher weitgehend innerhalb der Fachsprachen- und Fachkommunikationsforschung. Präsent war sie bisher nur in der ethnomethodologischen und konversationsanalytischen Forschung zu Wissenskonstruktion und -vermittlung (s. etwa Hindmarsh und Pilnick 2007 zum ‚verkörperten' Wissen), die in den letzten Jahren einen starken Aufschwung erfahren hat (s. Dausendschön-Gay et al. 2010; Stivers et al. 2011).

Gleichzeitig schließt der Vortrag mit seiner Konzentration auf die interaktive Nutzung der Hands-on-Exponate an eine aktuelle Debatte zur Frage an, wie Objekte (s. Nevile et al. 2014) oder Elemente des gebauten Raums (s. Hausendorf et al. 2016) von Interaktionsteilnehmerinnen und -teilnehmern für ihre Interaktion in Anspruch genommen werden können. Dabei gerät auch dort die Multimodalität in den Mittelpunkt: Wie in unserem Vortrag wird dort nämlich gefragt, wie die Teilnehmerinnen und Teilnehmer unterschiedliche Ausdrucksressourcen zu einer „multimodalen Gestalt" verbinden und wie diese Verbindung analytisch zu erfassen ist.

Bibliographie

Bergmann, Jörg R. / Quasthoff, Uta M. (2010): *"Interaktive Verfahren der Wissensgenerierung. Methodische Problemfelder"*, in: **Dausendschön-Gay, Ulrich / Domke, Christine / Ohlhus, Sören (eds.)**: *Wissen in (Inter-)Aktion. Verfahren der Wissensgenerierung in unterschiedlichen Praxisfeldern.* Berlin / New York: de Gruyter 21–34.

Birkner, Karin / Ehmer, Oliver (eds.) (2013): *Veranschaulichungsverfahren im Gespräch*. Mannheim: Verlag für Gesprächsforschung. http://www.verlag-gespraechsforschung.de/2013/birkner.html [letzter Zugriff am 10.03.2014].

Clifton Jr., Charles / Ferreira, Fernanda / Henderson, John M. / Inhoff, Albrecht W. / Liversedge, Simon P. / Reichle, Erik D. / Schotter, Elizabeth R. (2016): *"Eye movements in reading and information processing: Keith Rayner's 40 year legacy"*, in: Journal of Memory and Language 86: 1–19. DOI: 10.1016/j.jml.2015.07.004.

Dausendschön-Gay, Ulrich / Domke, Christine / Ohlhus, Sören (eds.) (2010): *Wissen in (Inter-)Aktion. Verfahren der Wissensgenerierung in unterschiedlichen Praxisfeldern.* Berlin / New York: de Gruyter.

Eisenberg, M. (2001): *"Discovery Learning, Cognitive Psychology of"*, in: **Smelser, Neil Joseph (ed.)**: *International encyclopedia of the social and behavioral sciences*. Amsterdam: Elsevier 3736–3739.

Gülich, Elisabeth (2001): *"Zum Zusammenhang von alltagsweltlichen und wissenschaftlichen 'Methoden'"*, in: **Brinker, Klaus / Antos, Gerd / Heinemann, Wolfgang / Sager, Sven F. (eds.)**: *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. Berlin: de Gruyter 1086–1093.

Gülich, Elisabeth / Mondada, Lorenza (2008): *Konversationsanalyse. Eine Einführung am Beispiel des Französischen*. Unter Mitarbeit von Ingrid Furchner. Tübingen: Niemeyer.

Hausendorf, Heiko / Schmitt, Reinhold / Kesselheim, Wolfgang (eds.) (2016): *Interaktionsarchitektur, Sozialtopographie und Interaktionsraum*. Tübingen: Narr Francke Attempto.

Hindmarsh, Jon / Pilnick, Alison (2007): *"Knowing bodies at work. Embodiment and ephemeral teamwork in anaesthesia"*, in: Organization Studies 28 (9): 1395–1416.

Kesselheim, Wolfgang (2012): *"Gemeinsam im Museum: Materielle Umwelt und interaktive Ordnung"*, in: **Hausendorf, Heiko / Mondada, Lorenza / Schmitt, Reinhold (eds.)**: *Raum als interaktive Ressource*. Tübingen: Narr: 187–231.

Kesselheim, Wolfgang (2017): *"Die Museumsausstellung - ein Text?"*, in: Germanistik in der Schweiz – Zeitschrift der Schweizerischen Akademischen Gesellschaft für Germanistik 14: 1–29.

Nevile, Maurice / Haddington, Pentti / Heinemann, Trine / Rauniomaa, Mirka (eds.) (2014): *Interacting with objects. Language, materiality, and social activity.* Amsterdam: Benjamins.

Selting, Margret / Auer, Peter / Barth-Weingarten, Dagmar / Bergmann, Jörg R. (2009): *"Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)"*, in: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10: 353–402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf [letzter Zugriff am 28.06.2013].

Sidnell, Jack (2010): *Conversation analysis. An introduction*. Oxford: Wiley-Blackwell.

Stivers, Tanya / Mondada, Lorenza / Steensig, Jakob (eds.) (2011): *The morality of knowledge in conversation*. Cambridge: Cambridge University Press.

Wichter, Sigurd / Antos, Gerd (eds.) (1999): *Wissenstransfer zwischen Experten und Laien. Umriss einer Transferwissenschaft.* Frankfurt am Main: Lang.

Das GeSIG-Inventar: Eine Ressource für die Erforschung und Vermittlung der Wissenschaftssprache der Geisteswissenschaften

Meißner, Cordula
cordula.meissner@uni-leipzig.de
Universität Leipzig, Deutschland

Wallner, Franziska
f.wallner@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

Für die fachübergreifende Wissenschaftssprache geisteswissenschaftlicher Disziplinen liegen aktuell keine lexikografischen Nachschlagewerke bzw. Informationsressourcen vor. Der wichtige Bereich der übergreifend in dieser Fächergruppe gebrauchten sprachlichen Erkenntniswerkzeuge ist damit durch die linguistische Sprachbeschreibung noch nicht erschlossen. Dies stellt zugleich ein Praxisdesiderat dar, denn es fehlt damit in Bezug auf die geisteswissenschaftlichen Disziplinen eine wesentliche Grundlage für die wissenschaftspropädeutische Sprachvermittlung. Angesichts der besonderen Bedeutung der Sprache für die Wissensgewinnung in geisteswissenschaftlichen Disziplinen ist hier jener Wortschatzbereich zentral, der Ausdrucksmittel für wissensmethodologische Inhalte bereitstellt – der Bereich der fachübergreifend gebrauchten Lexik. Dieser Wortschatzbereich war für die geisteswissenschaftlichen Fächer jedoch bislang empirisch nicht erschlossen.

Der Beitrag stellt Ergebnisse eines Projekts[1] vor, in dem der fachübergreifend gebrauchte Wortschatz – das gemeinsame sprachliche Inventar – der Geisteswissenschaften (GeSIG-Inventar) datengeleitet ermittelt wurde. Das GeSIG-Inventar steht nun als Ressource für die Erforschung und Vermittlung der Wissenschaftssprache der Geisteswissenschaften zur Verfügung. Im Folgenden wird zunächst der Wortschatzbereich der fachübergreifenden Lexik in seiner

Bedeutung insbesondere für geisteswissenschaftliche Fächer dargestellt (2). Darauf wird die empirische Ermittlung des GeSIG-Inventars nachgezeichnet (3) und das Inventar vorgestellt (4). Anschließend werden Anwendungsfelder aufgezeigt, für die das GeSIG-Inventar als Ressource genutzt werden kann (5) und ein wissenschaftspropädeutischer Anwendungsfall näher beschrieben (6).

Fachübergreifender Wortschatz geisteswissenschaftlicher Disziplinen

Der Sprache kommt in geisteswissenschaftlichen Disziplinen eine wichtige Rolle zu. Sie ist das wesentliche Werkzeug, mit dem Erkenntnisse fixiert, in dieser Fixierung präzisiert und damit weiterentwickelt werden. Dies gilt im Grunde für alle Wissenschaften, fällt aber bei den Geisteswissenschaften besonders ins Gewicht, da hier selbst die Gegenstände der Forschung größtenteils sprachlich oder symbolisch verfasst sind (vgl. Kretzenbacher 2010: 494) bzw. erst durch eine Überführung in Sprache wissenschaftlich behandelbar werden. Die gewonnenen Begriffe und Formulierungen tragen dabei sowohl semantische Assoziationen als auch historische Verweise und Einordnungen in sich, die wesentlich sind für die geisteswissenschaftliche Arbeit (vgl. Weigel 2013: 57f.). Diese ist damit auch in besonderer Weise an die jeweilige Sprache gebunden, in deren Assoziationsraum diese Formulierungen gefunden wurden. Die Sprache in ihrer einzelsprachlichen Vielfalt ist daher das wichtigste Werkzeug der Geisteswissenschaften, ihre „Mathematik", wie es Hagner (2013) formuliert. In diesem Zusammenhang ist der Blick insbesondere auf jenen Wortschatzbereich zu richten, der Ausdrucksmittel für wissensmethodologische Inhalte bereitstellt. Es handelt sich hier um den Bereich der fachübergreifend gebrauchten Lexik, der unter den Begriffen der ‚allgemeinen' (vgl. etwa Schepping 1976) bzw. ‚alltäglichen' Wissenschaftssprache (vgl. Ehlich 1993 u. a.) gefasst wird. Dieser Bereich zeichnet sich durch eine besondere Nähe zur Gemeinsprache aus: Die ihm angehörenden Ausdrucksmittel existieren zumeist auch gemeinsprachlich, haben aber in der Wissenschaftssprache eine darüberhinausgehende spezifische Bedeutung bzw. Funktion erlangt.

Zur fachübergreifenden Lexik gehören diejenigen nicht-terminologischen sprachlichen Mittel, die pragmatisch-methodische Inhalte ausdrücken und so disziplinenübergreifend Verwendung finden. Die mit der Polysemie bzw. Vagheit der Ausdrücke verbundene inhaltliche Flexibilität in verschiedenen fachlichen und textuellen Kontexten gilt als eine charakteristische Eigenschaft dieser Lexik (vgl. Ehlich 2007: 104f.). Die disziplinenübergreifend verwendeten sprachlichen Mittel stellen Ausdrucksressourcen bereit etwa für Formen des Voraussetzens, des Begründens, des Folgerns, des Übertragens und des Ableitens. Eine Analyse der Lexik der allgemeinen Wissenschaftssprache gestattet somit Einblicke in zentrale Prozesse wissenschaftlichen Handelns. Sie ermöglicht es, die Funktionsweise von Wissenschaftssprache als facettenreiches, differenziertes Erkenntnisinstrument näher zu beleuchten, welches insbesondere für die geisteswissenschaftliche Forschung von grundlegender Bedeutung ist.

Die Ermittlung des GeSIG-Inventars

Mit dem am Herder-Institut der Universität Leipzig angesiedelten Projekt GeSIG (Das gemeinsame sprachliche Inventar der Geisteswissenschaften) wurde das Inventar der allgemeinen Wissenschaftssprache der Geisteswissenschaften auf empirischer Grundlage datengeleitet bestimmt.

Als Datengrundlage wurde ein Korpus geisteswissenschaftlicher Dissertationen aufgebaut. Die Dissertation als Textsorte wurde gewählt, da sie das gesamte Spektrum des in Textform niedergelegten wissenschaftlichen Erkenntnisprozesses in besonderer Breite und Vollständigkeit abbildet.

Zur Operationalisierung des Gebietes der Geisteswissenschaften wurde die Umfangsbestimmung des Wissenschaftsrates (2010) zugrunde gelegt. Diese ist an die Systematik des statistischen Bundesamtes angelehnt und umfasst die dort unterschiedenen Fächergruppen Sprach- und Kulturwissenschaften (ohne Psychologie, Erziehungswissenschaften und Sonderpädagogik) sowie Kunst und Kunstwissenschaften. So operationalisiert umfassen die Geisteswissenschaften 19 Fachbereiche. Eingeschlossen sind Fächer wie Philosophie, Sprach- und Literaturwissenschaften, Geschichtswissenschaften, Regionalstudien, religionsbezogene Wissenschaften, die bekenntnisgebundenen Theologien, die Ethnologien sowie die Kunst-, Theater- und Musikwissenschaften (vgl. Statistisches Bundesamt 2013). Die 19 geisteswissenschaftlichen Fachbereiche bildeten die Grundlage für den Aufbau von 19 entsprechenden Fachbereichskorpora. Dabei wurde für jeden Bereich ein Korpus aus mindestens 10 Dissertationen und mit einer Mindestgröße von 1 Mio. Token zusammengestellt. Insgesamt umfasst die so erhobene Datengrundlage 197 Dissertationen und rund 22,8 Mio. Token. Die Sprachdaten wurden anschließend für die korpuslinguistische Analyse bereinigt und aufbereitet. Sie wurden mit Hilfe des TreeTaggers (Schmid 1995) nach Wortarten annotiert und lemmatisiert. Dabei lag das Stuttgart-Tübingen-Tagset (STTS) zugrunde (Schiller et al. 1999).

Um das Konzept der allgemeinen Wissenschaftssprache zu operationalisieren, wurde das Charakteristikum ihrer disziplinenübergreifenden Verwendung herangezogen. Die sprachlichen Mittel der allgemeinen Wissenschaftssprache der Geisteswissenschaften wurden demnach empirisch bestimmt als Schnittmenge der Wortschätze einzelner geisteswissenschaftlicher Fachbereiche. Der in dieser Schnittmenge enthaltene Wortschatz umfasst jene sprachlichen Mittel, die der Form nach in den 19 geisteswissenschaftlichen Fachbereichen übergreifend gebraucht werden und repräsentiert damit die sprachlichen Mittel der allgemeinen Wissenschaftssprache der Geisteswissenschaften. Abb. 1 illustriert das Vorgehen.

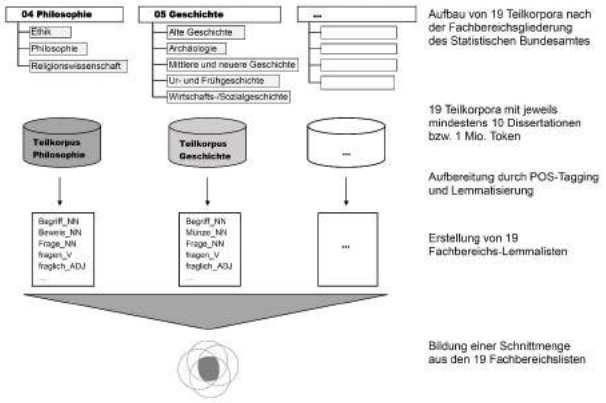


Abb. 1: Ermittlung des GeSIG-Inventars.

Das GeSIG-Inventar

Das so ermittelte GeSIG-Inventar umfasst insgesamt 4.490 Lemmata. 94% davon entfallen auf Inhaltswörter. Nomen bilden mit 1.681 Lemmata (37%) die größte Gruppe. Adjektive nehmen mit 1.171 Lemmata (26%) den zweiten und Verben mit 1.108 Lemmata (25%) den dritten Platz ein. Auf Adverbien entfallen mit 261 Lemmata 6%. Abb. 2 fasst die wortartenbezogene Zusammensetzung des Inventars zusammen.

Wortart	Lemmata
Nomen	1.681
Adjektive	1.171
Verben	1.108
Adverbien	261
Pronomen	105
Präpositionen	68
Konjunktionen	48
Konjunktionaladverbien	39
Partikeln	7
Artikel	2
gesamt	**4.490**

Abb. 2: Das GeSIG-Inventar in seiner Zusammensetzung nach Wortarten.

In Meißner/Wallner (2019) wird eine umfassende Analyse zum GeSIG-Inventar vorgenommen. Das Inventar wird darin einerseits formseitig nach wiederkehrenden Bestandteilen (wortfamiliäre bzw. wortbildungsbasierte Beziehungen) beschrieben. Andererseits erfolgt eine umfassende bedeutungsseitige Analyse der Lemmata. Dies geschieht in onomasiologischer Perspektive nach den die Lemmata verbindenden semantischen Ähnlichkeitsbeziehungen sowie in semasiologischer Perspektive im Hinblick auf die gemeinsprachenlexikografisch erfasste Polysemie der Inventar-Lemmata.

Für die Nachnutzung steht das GeSIG-Inventar in elektronischer Form frei zur Verfügung.[2] Die digitale Liste umfasst Informationen zum Lemma, zum POS-Tag und zur Häufigkeitsklasse des Lemmas in dem der Erhebung zugrunde liegenden Korpus.

Anwendungsfelder

Mit dem GeSIG-Inventar liegt eine Ressource vor, die verschiedene Anwendungsmöglichkeiten für die Erforschung und Vermittlung der Sprache der Geisteswissenschaften eröffnet. Mit Hilfe des Inventars kann dabei jeweils die fachübergreifende Lexik in den untersuchten Texten bzw. Textkorpora identifiziert werden. Dies erlaubt die gezielte Betrachtung der Vorkommen und Verwendungsweisen dieses Wortschatzbestandes etwa in verschiedenen fachdisziplinären, historischen oder erwerbsbezogenen Kontexten. Damit eröffnen sich Nutzungsmöglichkeiten u.a. für folgende Anwendungsfelder:

- **Vergleichende Untersuchungen zur Verschiedenheit der Wissenschaftssprachen der Geisteswissenschaften:** Durch einen Abgleich mit den GeSIG-Lemmata können in Sprachdatensammlungen verschiedener Disziplinen Belege für die Nutzung der fachübergreifenden Ausdrucksmittel identifiziert werden. Diese lassen sich daraufhin vergleichend in ihren theoriebezogenen, fachinhaltlichen Bezügen und Prägungen betrachten.
- **Untersuchungen zur historischen Entwicklung der fachübergreifenden Lexik:** Durch einen Abgleich mit den GeSIG-Lemmata können in Sprachdatensammlungen verschiedener historischer Sprachentwicklungsstufen Belege für die Nutzung der fachübergreifenden Ausdrucksmittel identifiziert werden. Auf dieser Grundlage lässt sich nachzeichnen, wann sich aus einem ursprünglich gemeinsprachlichen Gebrauch der Lexik eine wissenschaftssprachliche Funktion herausgebildet hat und wie die Entfaltung des wissenschaftssprachlichen Funktionenspektrums im historischen Verlauf erfolgt ist.
- **Untersuchung von Lernersprache / der Entwicklung wissenschaftssprachlicher Kompetenz:** Durch einen Abgleich mit den GeSIG-Lemmata können in lernersprachlichen Texten verschiedener Kompetenzstufen Belege für die Nutzung der fachübergreifenden Ausdrucksmittel identifiziert werden. Auf dieser Basis lässt sich nachzeichnen, wie Lernende fachübergreifende Lexik gebrauchen und wo Lernschwierigkeiten liegen (etwa in Bezug auf Korrektheit, stilistische Adäquatheit und Vielfalt).
- **Wissenschaftspropädeutik:** Eine wichtige Aufgabe der Wissenschaftspropädeutik ist es, Lernende für den fachübergreifenden Wortschatz zu sensibilisieren. Das GeSIG-Inventar bietet einen Ausgangspunkt für die wissenschaftspropädeutische Bearbeitung dieses Lernfeldes (vgl. Meißner/Wallner 2018a, 2018b, 2019 Kap. 7).
- **Lexikografie:** Neben diesen Untersuchungsfeldern stellt das GeSIG-Inventar eine Vorarbeit für die lexikografische Aufarbeitung der allgemeinen Wissenschaftssprache der Geisteswissenschaften dar. Als fachübergreifend gebrauchter Lemmabestand bildet es eine empirisch fundierte Basis für eine entsprechende Lemmaselektion (vgl. Meißner/Wallner 2016, Meißner/Wallner 2019 Kap. 5).

Ein wissenschaftspropädeutischer Anwendungsfall

Die fachübergreifende Lexik stellt für Novizen des Wissenschaftsbetriebs (insbesondere mit Deutsch als L2) eine Herausforderung dar und muss daher im studienvorbereitenden bzw. -begleitenden Sprachunterricht gezielt gefördert werden (vgl. etwa Schepping 1976, Ehlich 1995, Steinhoff 2007). Mit dem GeSIG-Inventar liegt für die (Fremd-)Sprachdidaktik erstmals eine empirisch abgesicherte Bestimmung der fachübergreifenden Lexik der Geisteswissenschaften vor. Damit steht für die Wissenschaftspropädeutik eine verlässliche Wortschatzauswahl zur Verfügung.

Mithilfe des elektronisch zugänglichen Inventars ist es darüber hinaus möglich, den fachübergreifenden Wortschatz nach Wortart und Häufigkeit zu sortieren und Teillisten nach diesen Kriterien zu erstellen. Auf dieser Grundlage kann jeweils eine zielgruppenspezifische Auswahl der zu vermittelnden Einheiten getroffen werden. So ließen sich etwa die 50 häufigsten Nomen, Verben und Adjektive als Einstiegswortschatz für einen studienvorbereitenden Sprachkurs auswählen. Daneben können auf Grundlage der Liste Lemmata für spezifische wissenschaftspropädeutische Lernfelder selegiert werden, etwa Nominalisierungen auf *-ung* oder *- ion*.

Die so ausgewählte Lexik kann dann anhand von im Unterricht behandelten Texten thematisiert werden. In Analogie zu dem von Townsend/Kirnan (2015) beschriebenen „Word and Phrases Tool“[3] ließe sich hierbei die elektronische Liste in ein Textanalysewerkzeug implementieren, mit dem die GeSIG-Lemmata bzw. eine Auswahl davon in den behandelten Texten farblich markiert werden können. Abb. 3 illustriert anhand einer Passage aus der Einleitung einer kunstwissenschaftlichen Dissertation, wie das Ergebnis einer solchen Bearbeitung aussehen könnte. In diesem Beispiel sind alle Wortformen, deren Lemma Teil des GeSIG-Inventars ist, hervorgehoben. Mithilfe einer solchen Aufbereitung kann die Bedeutung und Verwendung der fachübergreifenden Lexik am konkreten Textbeispiel zielgruppenspezifisch in wissenschaftspropädeutischen Kursen behandelt werden.

EINLEITUNG

Der rheinland-pfälzische Maler Max Rupp **erarbeitete die formalen Prinzipien seines Werkes in einem Umfeld der Diskussion um** Figuration **und** Abstraktion. **Auf Grund seiner Affinität zur französischen** Malerei **musste diese Thematik ihm in besonderem Maße präsent sein. Die intensive Beschäftigung mit** Villon, Bazaine, de Staël, Magnelli **und anderen Vertretern der abstrakten französischen, speziell Pariser** Kunstszene **und der häufige Kontakt mit Kunst und Künstlern in** Paris **haben sein Werk stark beeinflusst, wie in der vorliegenden Arbeit zu zeigen sein wird. Ebenso wie die Malerei seiner Vorbilder waren auch seine** geometrischen **und** lyrisch-organischen **Bilder in die Diskussion um die** Zeichenhaftigkeit **abstrakter Darstellungen und ihren Bezug zum** Erleben **des** Betrachters **einbezogen.**

Abb. 3: Lemmata des GeSIG-Inventars in einem Ausschnitt der kunstwissenschaftlichen Dissertation 74_KUGE_5.

Fußnoten

1. Das Projekt „Die allgemeine Wissenschaftssprache der Geisteswissenschaften – Projekt zur Bestimmung des Inventars der allgemeinen Wissenschaftssprache der Geisteswissenschaften auf empirischer Grundlage“ wurde von 2015 bis 2017 aus Mitteln des Freistaates Sachsen im Rahmen des Programms „Geisteswissenschaftliche Forschung“ bei der Sächsischen Akademie der Wissenschaften zu Leipzig gefördert. Vgl. zum Projekt http://research.uni-leipzig.de/gesig/.
2. Vgl. unter: http://GeSIG-Inventar.ESV.info (20.12.2018).
3. Vgl. unter: https://www.wordandphrase.info/ (20.12.2018).

Bibliographie

Ehlich, Konrad (1995) : *„Die Lehre der deutschen Wissenschaftssprache: sprachliche Strukturen, didaktische Desiderate“*. In: **Kretzenbacher, Heinz / Weinrich, Harald (Hrsg.)**: *Linguistik der Wissenschaftssprache* [= Akademie der Wissenschaften zu Berlin. Forschungsberichte 10]. Berlin/New York: De Gruyter 325–351.

Ehlich, Konrad(1993) : *„Deutsch als fremde Wissenschaftssprache“*. In: **Wierlacher, Alois (Hrsg.)**: *Grenzen und Grenzerfahrungen* [= Jahrbuch Deutsch als Fremdsprache 19]. München: iudicium 13–42.

Ehlich, Konrad (2007): *Pragmatik und Sprachtheorie* [= Sprache und sprachliches Handeln 1]. Berlin/New York: De Gruyter.

GeSIG-Inventar: http://GeSIG-Inventar.ESV.info (20.12.2018)

Hagner, Michael (2013): *„Die Mathematik der Geisteswissenschaften ist die Vielfalt der Wissenschaftssprachen“*. In: **Goethe-Institut / Deutscher Akademischer Austauschdienst (DAAD) / Institut für Deutsche Sprache (IDS) (Hrsg.)**: *Deutsch in den Wissenschaften. Beiträge zu Status und Perspektiven der Wissenschaftssprache Deutsch*. München: Klett-Langenscheidt 136–141.

Kretzenbacher, Heinz (2010): *„Fach- und Wissenschaftssprachen in den Geistes- und Sozialwissenschaften“*. In: **Krumm, Hans-Jürgen / Fandrych, Christian / Hufeisen, Britta / Riemer, Claudia (Hrsg.)**: *Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch* [= Handbücher zur Sprach- und Kommunikationswissenschaft 35.1]. Berlin/New York: De Gruyter 493–501.

Meißner, Cordula / Wallner, Franziska (2019): *Das gemeinsame sprachliche Inventar der Geisteswissenschaften. Lexikalische Grundlagen für die wissenschaftspropädeutische Sprachvermittlung*. Berlin: Erich Schmidt Verlag.

Meißner, Cordula / Wallner, Franziska (2018a): *Allgemein-wissenschaftssprachlicher Wortschatz in der Sekundarstufe I? Zu Vagheit, Polysemie und pragmatischer Differenziertheit von Verben in Schulbuchtexten*. In: **Hövelbrinks, Britta / Fuchs, Isabell / Maak, Diana / Duan, Tinghui / Lütke, Beate (Hrsg.)**: *Der - Die - DaZ - Forschungsbefunde zu Sprachgebrauch und Spracherwerb von Deutsch als Zweitsprache*. Berlin/New York: de Gruyter 133–147.

Meißner, Cordula / Wallner, Franziska (2018b): *„Zur Rolle des allgemein-wissenschaftssprachlichen Wortschatzes für die Wissenschaftspropädeutik im Übergangsbereich Sekundarstufe II – Hochschule“*. In: InfoDaF 45 (4): 1–22.

Meißner, Cordula / Wallner, Franziska (2016): *„Persuasives Handeln im wissenschaftlichen Diskurs und seine lexikografische Darstellung: das Beispiel der Kollokation Bild zeichnen“*. In: Studia Linguistica 35: 235–252.

Schepping, Heinz (1976): *„Bemerkungen zur Didaktik der Fachsprache im Bereich des Deutschen als Fremdsprache“*. In: **Rall, Dietrich / Schepping, Heinz / Schleyer, Walter (Hrsg.)**: *Didaktik der Fachsprache.* Beiträge zu einer Arbeitstagung an der RWTH Aachen vom 30. September bis 4. Oktober 1974 [= DAAD Forum: Studien, Berichte, Materialien 8]. Bonn-Bad Godesberg: Deutscher Akademischer Austauschdienst 13–34.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*, unter: www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (27.09.2018).

Schmid, Helmut (1995): *Improvements In Part-of-Speech Tagging With An Application To German*, unter: www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (27.09.2018).

Statistisches Bundesamt (2013): *Studierende an Hochschulen – Fächersystematik*. unter: https://www.destatis.de/DE/Methoden/Klassifikationen/BildungKultur/ StudentenPruefungsstatistik.pdf (16.10.2014).

Steinhoff, Torsten (2007): *Wissenschaftliche Textkompetenz. Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten* [= Reihe Germanistische Linguistik 280]. Tübingen: Niemeyer.

Townsend, Dianna / Kiernan, Darl (2015): *„Selecting Academic Vocabulary Words Worth Learning“*. In: The Reading Teacher 69/1: 113–118.

Weigel, Sigrid (2013): *„Erkenntnispotenzial und ideologische Erbschaften – zur deutschen Wissenschaftssprache in den Geisteswissenschaften und ihrer Geschichte“*. In: **Goethe-Institut / Deutscher Akademischer Austauschdienst (DAAD) / Institut für Deutsche Sprache (IDS) (Hrsg.)**: *Deutsch in den Wissenschaften. Beiträge zu Status und Perspektiven der Wissenschaftssprache Deutsch*. München: Klett-Langenscheidt 57–67.

Wissenschaftsrat (2010): *Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften*, unter: www.wissenschaftsrat.de/download/archiv/10039-10.pdf (27.09.2018).

Words and Phrases Tool: https://www.wordandphrase.info/ (20.12.2018)

Schröder, Petra (2012) / Max Rupp (1908 - 2002): *Wege zur Abstraktion*. Dissertation. Köln: Universität zu Köln, unter: http://d-nb.info/1038452473/34 (20.12.2018).

Das Notizbuch als Ideenspeicher und Forschungswerkzeug: Erkenntnisse aus einer digitalen Repräsentation

Scholger, Martina

martina.scholger@uni-graz.at
Universität Graz, Österreich

„I meant my notebooks to be a storehouse of materials for future use and nothing else" (Maugham 1949: xiv) schreibt der englische Dramatiker William Somerset Maugham in seinem Vorwort *A Writer's Notebook* über den Zweck seiner Notizbücher.

Notizbücher werden in unterschiedlichen Disziplinen verwendet: Schriftsteller, Künstler und Wissenschaftler halten ihre unmittelbaren Eindrücke und Ideen in Form von schriftlichen Notizen und flüchtigen Skizzen fest (Radecke 2013: 161). Diese werden zu einem späteren Zeitpunkt zur Ausführung eines Werkes herangezogen oder erweisen sich als Sackgasse und werden wieder verworfen.

Obwohl Notizbücher zum einen eine reiche Quelle an Informationen zur Entstehungsgeschichte von Werken liefern und zum anderen in ihrer Objekthaftigkeit als eigenständiges Meta(kunst)werk betrachtet werden können, gibt es nur wenige Editionsvorhaben – gedruckt oder digital – die sich der Herausforderung Notizbuch stellen. Zu den wenigen Projekten, die sich ausschließlich den Notizen eines Akteurs widmen, zählen die Hybrid-Edition von *Theodor Fontanes Notizbüchern* (Radecke 2013) oder *Paul Klee - Bildnerische Form- und Gestaltungslehre* (Eggelhöfer & Keller 2012). Als Beispiele für die Berücksichtigung von Notizen als Teil von umfassenderen Editionsprojekten seien hier stellvertretend die digitale Faksimile-Ausgabe Nietzschesource (D'Iorio 2009), die *Digital Edition of Fernando Pessoa* (Sepúlveda & Henny-Krahmer 2017), die *Beckett-Edition* (Van Hulle & Nixon 2018) oder das digitale Archiv *eMunch* (Munch Museum 2011-2015) genannt.

Das hier verwendete Korpus entstand als Teil einer Dissertation zur Erschließung und Untersuchung der Notizbücher des österreichischen bildenden Künstlers Hartmut Skerbisch (1945-2009) mit digitalen Methoden. Dabei handelt es sich um 35 Notizbücher, die knapp 40 Jahre künstlerischer Tätigkeit zwischen 1969 und 2008 auf insgesamt 2200 Seiten dokumentieren. Die im Zuge der Forschungsarbeit entstandene digitale Edition steht – derzeit in einer Betaversion – unter https://gams.uni-graz.at/skerbisch im Geisteswissenschaftlichen Asset Management System (GAMS) des Zentrums für Informationsmodellierung der Universität Graz zur freien Verfügung und soll im Laufe des Jahres 2019 finalisiert werden.

GAMS ist ein OAIS-konformes Asset-Management-System zur Verwaltung, Publikation und Langzeitarchivierung digitaler Ressourcen das sich an den FAIR Data Principles orientiert (Stigler & Steiner 2018: 209-211).

Damit wird eine gleichermaßen künstlerische und wissenschaftliche Ressource digital erschlossen und verfügbar gemacht, die bislang für die Öffentlichkeit unzugänglich war.

Skerbisch befasste sich mit konzeptioneller Kunst, Medienkunst und Objektkunst und entzieht sich einer eindeutigen Zuordnung zu einer bestimmten Kunstrichtung. Seine 35 Notizbücher sind jedoch ohne Zweifel *konzeptionell* (Sol LeWitt 1967; Kosuth 1969): Diese nutzte er für die Konzeption und Entwicklung seiner künstlerischen Ideen, seiner Gedankenexperimente, seines allgemeinen Verständnisses für künstlerische Konzepte und die Detailplanung seiner ausgeführten Kunstwerke, sowohl in textueller als auch grafischer Form.

Das Werk von Skerbisch reflektiert neben der grundsätzlichen Faszination für Technik und Wissenschaft zeitlebens Einflüsse aus Literatur, Musik und bildender Kunst. Er stellte die Materialität und die Weiterentwicklung des Raumbegriffs gegenüber der ästhetischen Hülle in den Vordergrund (Holler-Schuster 2015: 170; Scholger 2015: 83-86).

Im Zusammenhang mit einer digitalen Repräsentation der Notizbuchinhalte stellen sich folgende Forschungsfragen:

a) Wie kann eine digitale Aufbereitung der Notizbücher zum Erkenntnisprozess über das Kunstwerk beitragen?

Besonders für die Rezeption und die Vermittlung zeitgenössischer Kunst erweisen sich Notizen als wertvolle Quelle: Auf Arbeiten mit neuen Medien und Materialien, in denen die Ästhetik des Werks nicht im Vordergrund steht, reagiert das Publikum oft verunsichert und verständnislos. Eine editorische Aufbereitung von Notizbüchern vermag jedoch Antworten und Erklärungen zugänglich zu machen, die das Kunstwerk bzw. der Künstler / die Künstlerin zum Zeitpunkt der Aus- oder Aufführung – mitunter auch durchaus bewusst – schuldig bleibt.

b) Wie können Notizbücher dazu beitragen, eine Werksidee, die nie zur Ausführung gekommen ist, oder eine Installation, die verloren gegangen ist, zu rekonstruieren?

Eine weitere Funktion von Notizbüchern gerade im Kontext zeitgenössischer Kunst ist jene der Dokumentation, insbesondere bei temporären und performativen Kunstwerken: Möglich wird damit eine – wenn auch nicht lückenlose – Rekonstruktion nicht zur Ausführung gelangter Werksideen und nicht mehr in ihrer Gesamtheit erhaltener Kunstwerke, deren Einzelteile oftmals für andere Installationen weiterverwendet wurden, oder die sogar nur noch durch Fotos, Berichte und Aussagen von Zeitgenossen dokumentiert sind.

c) Welche Assoziationsprozesse fließen in die Werkkonzeption ein?

Die Notizbücher enthalten eine Reihe an Querverweisen untereinander sowie Referenzen auf Kunstwerke, Personen, Literatur und Musik in Form von direkten oder indirekten Nennungen oder Zitaten. Mit dem Einbringen zusätzlichen Wissens durch Markup wird die Ressource inhaltlich erschlossen und stellt damit ein Werkzeug bereit, das den systematischen Interpretationsvorgang erleichtert und es ermöglicht, die Genese von Ideen, Assoziationen und Konzepten nachzuzeichnen.

Der Beitrag widmet sich den Resultaten dieser Forschungsfragen an die digital edierten Notizbücher von Hartmut Skerbisch und 1) präsentiert auf die Notizbücher angewandte Methoden, 2) zeigt die daraus gewonnenen Erkenntnisse und Ergebnisse und 3) versucht eine generelle Ableitung der verwendeten Methoden und Anwendbarkeit auf vergleichbare Quellen. Schließlich sollen die aus der digitalen Edition gewonnenen Erfahrungen in Bezug auf die digitale Repräsentation thematisiert werden: Welche Methoden erwiesen sich als effektiv, oder aber als wenig erkenntnisbringend in Relation zum geleisteten Ressourcenaufwand.

Methoden

Textkodierung/Textrepräsentation mit TEI

Die angemessene digitale Repräsentation der textuellen Notizbucheinträge, die in etwa zwei Drittel des Korpus ausmachen, erfordert ein geeignetes Textmodell. Dieses Textmodell muss dem der Quelle inhärenten pluralistischen Charakter (Sahle 2013: 45-49) der Textgattung Notizbuch gerecht werden. So gilt es die Notizen inhaltlich, strukturell, materiell und visuell zu erfassen. Diese Anforderungen konnten weitgehend mittels des Standards der Text Encoding Initiative (TEI) abgedeckt werden, insbesondere mit dem Modul 11 zur Repräsentation von Primärquellen (TEI Consortium 2018: „Representation of Primary Sources"), das Empfehlungen zur Verzeichnung von Faksimiles, zur Verknüpfung von Text und Faksimile sowie dem Umgang mit einfachen und erweiterten Elementen für die Transkription zur Kodierung von Interventionen sowohl am Text als auch am Textträger (Streichungen, Hinzufügungen, unklare Stellen, Beschädigungen, Textumstellungen etc.) gibt. Die Eingriffe am Text lassen sich in Anlehnung an die Methode der *critique génétique* weitgehend in TEI erfassen (Burnard et al. 2008; Brüning et al. 2013). Gerade bei Künstlernotizen steht jedoch nicht die Textgenese, sondern vielmehr die Genese einer künstlerischen Idee, eines Konzepts, im Fokus.

Kodierung von Skizzen in TEI und RDF

Neben dem Text sind es besonders Skizzen, die von Skerbisch zur Konzeption und Planung von Kunstwerken als Ausdrucksform eingesetzt wurden. Diese reichen von flüchtigen Skizzen bis hin zu detaillierten Zeichnungen mit Materialangaben und Abmessungen zur Ausführung. Sie nehmen in etwa ein Drittel der Notizbücher ein und bedürfen daher einer besonderen Betrachtung. Aus diesem Grund braucht es neben dem Textmodell ein Modell zur form- und inhaltsbezogenen Beschreibung von Skizzen, das die grafische Komponenten, die Textfunktionen und die Interpretationsebene berücksichtigt (siehe Abb. 1).

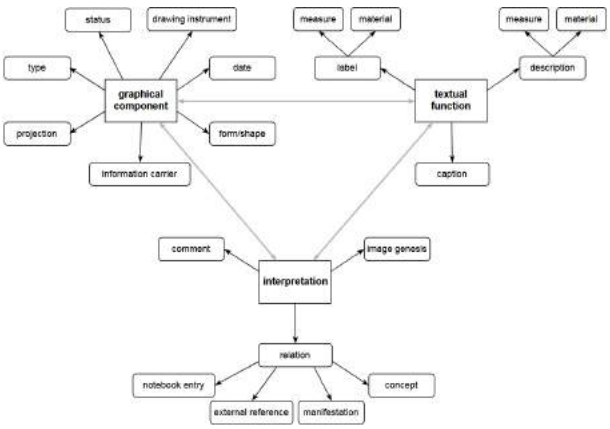


Abbildung 1. Modell zur Beschreibung von grafischen Komponenten

Für die formale Beschreibung der grafischen Komponenten ist es nicht ausreichend, lediglich die Existenz von Skizzen im Text zu verzeichnen oder als illustrativen Zusatz zum Text zu verstehen. Vielmehr müssen die Skizzen sowohl als solitäre Einheit als auch in Verbindung mit dem Text stehend beschrieben werden können. Dazu wurden die Möglichkeiten der TEI und jene der Beschreibung von Ressourcen über das Resource Description Framework (RDF) kombiniert und ein eigenes Modell im Rahmen des Editionsprojektes entwickelt. Es beschreibt die grafischen Komponenten (Typ, Ansicht, Zeichenwerkzeug, Form, Datierung, Informationsträger), die Textfunktionen (Beschriftung, Titel, Beschreibung, Abmessung, Materialbezeichnung) und die Interpretationen (Kommentar, Bildgenese sowie Relationen zu Notizbucheinträgen, externen Referenzen, Manifestationen, Konzepten).

Entitäten über RDF beschreiben

Um die zahlreichen Relationen zwischen den Notizen untereinander aber auch von einzelnen Einträgen zu Entitäten in der real existierenden Welt – wie zu Personen, Literatur, Tonträgern und Kunstwerken – zu verzeichnen braucht es ein Modell, das diese Beziehungen abbildet und operationalisierte Abfragen auf diese Wissensbasis zulässt. Durch die Verknüpfung der Notizbucheinträge mit sachbezogenen Zusatzinformationen in Registern und Thesauri entsteht ein Informationsnetzwerk, das die individuellen Einträge in einen breiteren Kontext einbindet (Vogeler 2015). Diese Aussagen werden als RDF in einem Triple Store gespeichert und über SPARQL-Anfragen die Beziehungen zwischen Einheiten sichtbar gemacht. Neben der Anwendung des Referenzmodells CIDOC-CRM (Le Boeuf et al. 2018) werden, wo möglich, bestehende Authority Files wie die GND (Deutsche Nationalbibliothek 2012-2019), VIAF (OCLC 2012-2019), der Getty Art and Architecture Thesaurus (Getty Research Institute 2019) etc. integriert, um Konzepte formal zu beschreiben und dem Prinzip des *Linked Open Data* zu folgen.

Ergebnisse

Dieser Beitrag kombiniert methodische Ansätze – gewonnen aus der konkreten Arbeit an der digitalen Edition der Notizbücher von Hartmut Skerbisch – und versucht diese zu allgemeinen Aussagen zu formulieren, die für die Anwendung auf ähnliche Quellen herangezogen werden können.

Textgenese auf Dokumentebene

Obwohl die Textkonstitution der Notizbucheinträge meist eine untergeordnete Rolle spielt, hat der Blick auf spezifische Editionsmethoden – wie jene der critique génétique – gezeigt, wie sich bestimmte Textsequenzen durch nachträgliche Interventionen durch den Künstler sinngemäß verändern.

Text- und Ideengenese auf Korpusebene

Textentwicklungen können bei Skerbisch nicht nur über direkte Textinterventionen am Dokument beobachtet werden, sondern auch über mehrere Notizbuchseiten innerhalb eines Heftes und sogar über das gesamte Korpus hinweg. Seine Methode sich einem Thema anzunähern, war auffallend oft eine Vielzahl an Wiederholungen, die er in den Notizbüchern verzeichnete (siehe Abb. 2).

W1	sie	hat	angefangen	ihre	fortlaufenden	Zustände	vorzuträumen
W2	sie	hat	angefangen				
W3	sie	hat	angefangen	ihre		Zustände	vorzuträumen
W4	sie	hat	angefangen	ihre	fortlaufenden	Zustände	vorzuträumen
W5	sie	hat	angefangen				vorzuträumen
W6	sie	hat	angefangen			Ereignisse	vorzuträumen
W7	sie	hat	angefangen	ihre		Entfaltung	vorzuträumen
W8	sie	hat	angefangen	ihre		Vorgänge	vorzuträumen
W9	sie	hat	angefangen	ihre	fortlaufenden	Zustände	vorzuträumen
W10	Sie	hat	angefangen				

Abbildung 2. Wiederholung und Modifikation einer Phrase in mehreren Notizbüchern

Entwicklung einer Idee

Eine zentrale Rolle in der Auswertung und Interpretation der Notizbücher spielt die Entwicklung einer Idee. Skerbischs Werke entstanden über einen langen Zeitraum intensiver Beschäftigung mit den Themen, die er vermitteln wollte. Hier entsteht ein Informationsnetzwerk aus Notizbucheinträgen, künstlerischen Manifestationen, physischen Objekten, externen Einflüssen aus der Literatur und werkübergreifenden künstlerischen Konzepten, das nur über die elektronische Erschließung des Materials zusammengefügt werden kann (siehe Abb. 3).

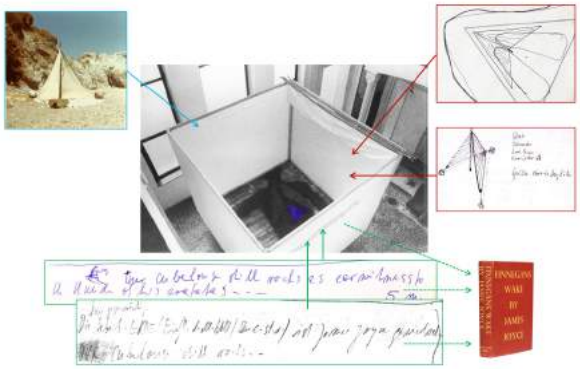


Abbildung 3. Entwicklung einer Idee

Das Editionsmodell eignet sich für Textgattungen mit ähnlichen materiellen, formalen und inhaltlichen Eigenschaften, wie a) fragmentarische Einträge, b) grafische Darstellungen, die als Bedeutungsträger einer gesonderten Betrachtung bedürfen, c) Referenzen auf externe Werke aus Literatur, Kunst und Musik, d) Namens- und Ortsnennungen, e) Verzeichnung von Zitaten, f) Textinterventionen sowie g) Text- und Ideengenese.

Die Zusammenführung unterschiedlicher Methoden (Textrepräsentation mit TEI, Formalisierung der Skizzen mit TEI und RDF, Beschreibung von Entitäten mit RDF) ermöglicht es die Entwicklung spezifischer Ideen nachzuvollziehen. Das Ergebnis ist eine umfangreiche Dokumentation des kreativen Prozesses, von der konzeptionellen Notiz zur künstlerischen Manifestation.

Bibliographie

Allemang, Dean / Hendler, James (2011): *Semantic Web for the Working Ontologist. Effective Modeling in RDFS and OWL*. Oxford: Elsevier.

Burnard, Lou / Jannidis, Fotis / Pierazzo, Elena / Rehbein, Malte (2008-2013): *An Encoding Model for Genetic Editions*. http://www.tei-c.org/Activities/Council/Working/tcw19.html

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

Deutsche Nationalbibliothek (2012–2019): *GND – Gemeinsame Normdatei*, https://www.dnb.de/gnd

D'Iorio, Paolo (2009): *Nietzschesource*. http://www.nietzschesource.org/

Eggelhöfer, Fabienne / Keller Tschirren, Marianne (2012): *Paul Klee – Bildnerische Form- und Gestaltungslehre*. http://www.kleegestaltungslehre.zpk.org

Getty Research Institute (2018): *Art & Architecture Thesaurus (AAT)*. https://www.getty.edu/research/tools/vocabularies/aat/index.html

Holler-Schuster, Günther (2015): *Hartmut Skerbisch's work between media art and land art*, in: **Verein der Freunde von Hartmut Skerbisch (ed.):** *Hartmut Skerbisch. Life and Work. Present as Present*. Wien: Verlag für moderne Kunst 142-171.

Kosuth, Joseph (1969a): *Art after philosophy I*, in: Studio International (915): 134-37.

Le Boeuf, Patrick / Doerr, Martin / Emil Ore, Christian / Stead, Stephen (2018): *Definition of the CIDOC Conceptual Reference Model*. Version 6.2.4. http://www.cidoc-crm.org/sites/default/files/2018-10-26%23CIDOC%20CRM_v6.2.4_esIP.pdf

LeWitt, Sol (1967): *Paragraphs on Conceptual Art*, in: Artforum 5 (10): 80-84.

Maugham, William Somerset (1949): *A writer's notebook*, Melbourn-London-Toronto: William Heinemann.

Munch Museum (2011-2015): *eMunch. Edvard Munchs Tekster. Digitalt Arkiv*. http://www.emunch.no

OCLC - Online Computer Library Center (2010–2016): *VIAF – Virtual International Authority File*, https://viaf.org/.

Radecke, Gabriele (2013): *Notizbuch-Editionen. Zum philologischen Konzept der Genetisch-kritischen und kommentierten Hybrid-Edition von Theodor Fontanes Notizbüchern*, in: editio 27 149-172. 10.1515/editio-2013-010

Sahle, Patrick (2013): *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung*, in: Schriften des Instituts für Dokumentologie und Editorik 9. Norderstedt: Books on Demand.

Scholger, Martina (2015): *Tracing the association processes of the Artist – The artwork as a commentary*, in: **Verein der Freunde von Hartmut Skerbisch (ed.):** *Hartmut Skerbisch. Life and Work. Present as Present*. Wien: Verlag für moderne Kunst 305-3.

Sepúlveda, Pedro / Henny-Krahmer, Ulrike (2017): *Fernando Pessoa – Digitale Edition. Projekte und Publikationen*. http://www.pessoadigital.pt

Stigler, Johannes Hubert / Steiner, Elisabeth (2018): *GAMS – Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten*, in: Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare 71 (1), 207-216. https://doi.org/10.31263/voebm.v71i1.1992

TEI Consortium (2018): *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0.* http://www.tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html/

Van Hulle, Dirk / Nixon, Mark (2019): *Samuel Beckett Digital Manuscript Project.* http://www.beckettarchive.org/

Vogeler, Georg (2015): *Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital editiert?*, in: Grenzen und Möglichkeiten der Digital Humanities, edited by Constanze Baum and Thomas Stäcker. Sonderband der Zeitschrift für digitale Geisteswissenschaften 1. 10.17175/sb001_007

Das Redewiedergabe-Korpus Eine neue Ressource

Brunner, Annelen
brunner@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Engelberg, Stefan
engelberg@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

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

Einführung

In diesem Beitrag[1] wird das Redewiedergabe-Korpus (RW-Korpus) vorgestellt, ein historisches Korpus fiktionaler und nicht-fiktionaler Texte, das eine detaillierte manuelle Annotation mit Redewiedergabeformen enthält. Das Korpus entsteht im Rahmen eines laufenden DFG-Projekts und ist noch nicht endgültig abgeschlossen, jedoch ist für Frühjahr 2019 ein Beta-Release geplant, welches der Forschungsgemeinschaft zur Verfügung gestellt wird. Das endgültige Release soll im Frühjahr 2020 erfolgen. Das RW-Korpus stellt eine neuartige Ressource für die Redewiedergabe-Forschung dar, die in dieser Detailliertheit für das Deutsche bisher nicht verfügbar ist, und kann sowohl für quantitative linguistische und literaturwissenschaftliche

Untersuchungen als auch als Trainingsmaterial für maschinelles Lernen dienen.

Motivation und verwandte Forschung

Redewiedergabe ist sowohl für die Linguistik als auch die Literaturwissenschaft ein interessanter Untersuchungsgegenstand. Die Repräsentation der Figurenstimme in Erzähltexten hat in der Narratologie viel Aufmerksamkeit erfahren und wurde in zahlreichen Kategoriesystemen abgebildet (vgl. z.B. Genette 2010; Martínez / Scheffel 2016). In der Linguistik besteht ein Interesse an sprachlichen Formen der Redewiedergabe, sowie an Redeeinleitungsverben (vgl. z.B. Hauser 2008, Engelberg 2015).

Detaillierte, manuell annotierte Korpora mit diesem Themenschwerpunkt sind bislang vor allem für das Deutsche sehr rar. Ein Vorbild mit detaillierter, literaturwissenschaftlich motivierter Annotation mehrere Redewiedergabetypen für das Englische ist das Korpus von Semino/Short 2004. Das ebenfalls manuell annotierte DROC-Korpus hat seinen Schwerpunkt auf Figurenreferenzen in Romanen, enthält in diesem Kontext allerdings auch Annotationen direkter Wiedergabe mit Sprecherzuordnung (Krug et al. 2018b). Unser Korpus ist eine direkte Weiterentwicklung des aus 13 Erzähltexten bestehenden Korpus aus Brunner 2015, unterscheidet sich jedoch von diesem vor allem in folgenden Aspekten: Es enthält neben fiktionalen auch nicht-fiktionale Texte, die Annotationen sind durch Mehrfachannotation wesentlich verlässlicher und es ist deutlich umfangreicher (für das Beta-Release ca. 350.000 Tokens vs. 57.000 Tokens in Brunner 2015).

Korpuszusammensetzung

Das RW-Korpus umfasst Textmaterial aus dem Zeitraum 1840-1920. Es beruht auf den folgenden drei Textquellen, aus denen jeweils nur die Texte ausgewählt wurden, die in den Untersuchungszeitraum passen:

- Erzähltexte aus der Digitalen Bibliothek, in TEI-Format konvertiert vom Projekt TextGrid (https://textgrid.de/digitale-bibliothek)
- Texte der Zeitschrift „Die Grenzboten", digitalisiert durch die Staats- und Universitätsbibliothek Bremen und in TEI-Format konvertiert durch das Deutsche Textarchiv (http://www.deutschestextarchiv.de/)
- Das Mannheimer Korpus Historischer Zeitungen und Zeitschriften (MKHZ, https://repos.ids-mannheim.de/mkhz-beschreibung.html), bereitgestellt vom Institut für Deutsche Sprache und in TEI-Format konvertiert durch das Deutsche Textarchiv

Bei der Korpuszusammenstellung sollte eine möglichst große Diversität der enthaltenen Texte erzielt werden. Um dies zu erreichen, setzt sich das Korpus aus Textausschnitten (Samples) zusammen. Diese haben mindestens 500 Wörter für fiktionale Texte bzw. 200 Wörter für nicht-fiktionale Texte – mit dieser großzügigeren Grenze war es möglich, auch kurze, abgeschlossene Artikel aufzunehmen, die für Zeitungen/Zeitschriften typisch sind. Die Samples wurden mit folgenden Besonderheiten randomisiert aus dem vorhandenen Textmaterial gezogen: Bei den Texten der Digitalen Bibliothek wurde erzwungen, dass jeder vertretene Autor innerhalb einer Dekade gleichermaßen berücksichtigt wird. Entsprechend wurde beim MKHZ erzwungen, dass alle in einer Dekade vertretenen unterschiedlichen Zeitungen/Zeitschriften gleichermaßen berücksichtigt werden. Damit wurde verhindert, dass Autoren bzw. Zeitungen/Zeitschriften mit wenig Material beim Sampling-Prozess vollkommen herausfallen. Das Beta-Release enthält Texte von etwa 140 unterscheidbaren Autoren und aus 20 unterschiedlichen Zeitungen/Zeitschriften.

Die Quelltexte wurden größtenteils in ihrem Ursprungszustand belassen, mit zwei Ausnahmen: Da für die Zeitschrift „Die Grenzboten" nur automatische OCR-Erkennung durchgeführt wurde, wurden die Samples aus dieser Textquelle manuell nachkorrigiert. In den Texten aus den beiden anderen Quellen wurden häufige Sonderzeichen, wie das Schaft-S, durch moderne Äquivalente ersetzt, jedoch weisen die Texte dennoch in unterschiedlichem Maße altertümliche Schreibungen und z.T. auch Sonderzeichen auf. Insgesamt ist festzuhalten, dass die Textformen im RW-Korpus sehr divers sind, so sind z.B. Texte im Dialekt enthalten, sowie Zeitungsausschnitte, die reine Listen sind. Wir haben uns bewusst dagegen entschieden, solch ungewöhnliches Material herauszufiltern, um eine realistische Repräsentation des Textmaterials aus den untersuchten Dekaden zu erhalten.

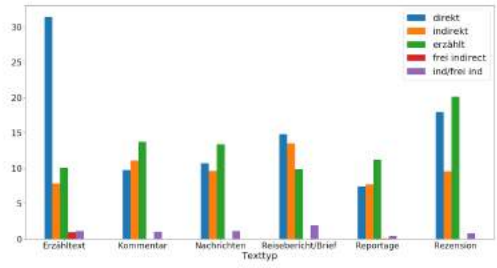
Beim RW-Korpus wurde eine Ausgewogenheit in der zeitlichen Dimension (Textmaterial pro Dekade) sowie zwischen fiktionalen und nicht-fiktionalen Texten angestrebt.

Entgegen ursprünglicher Annahmen stellte es sich als nicht sinnvoll heraus, die Trennung fiktional - nicht-fiktional rein aufgrund der Textquelle zu treffen: Es liegt in der Natur der Textsorte Zeitung/Zeitschrift, dass dort auch fiktionale Texte abgedruckt werden (im Feuilleton, als Fortsetzungsromane u.Ä.). Somit wurde das Kriterium ‚fiktional' für jedes Sample individuell festgelegt. Unsere Definition für ‚Fiktion' ist dabei angelehnt an Gabriel 2007: „Ein erfundener (‚fingierter') einzelner Sachverhalt oder eine Zusammenfügung solcher Sachverhalte zu einer erfundenen Geschichte" (Gabriel 2007: 594). Bei der Identifizierung wurde besonderer Wert auf paratextuelle Merkmale (z.B. Untertitel, Rubriken u.Ä.) gelegt. Von den Samples aus dem MKHZ und den „Grenzboten" wurden auf diese Weise ca. 12% als fiktional eingestuft.

Die folgende Tabelle zeigt die wichtigsten Metadaten des RW-Korpus, welche nach dem Sampling und der Textkorrektur vergeben werden.

Metadatum	Werte	Beschreibung
year	Zahl zwischen 1840 und 1919	Erscheinungsjahr des Textes (bei digBib-Texten: Ersterscheinungsjahr, falls verfügbar)
decade	Zahl in 10er Schritten	Erscheinungsdekade des Textes
source	digBib, grenz, mkhz	Textquelle; bei mkhz wird noch ein Kürzel für die jeweilige Zeitung/Zeitschrift beigefügt
title	String, undefined	Titel des Textes, falls bekannt
author	String, undefined	Autor des Textes, falls bekannt
fictional	yes, no	Ist der Textausschnitt fiktional?
text_type	Erzähltext, Kommentar, Anzeige, Reportage, Nachrichten, Biographie, Rezension, Reisebericht/Brief, unsure	Texttyp; wenn ein Ausschnitt mehrere Texttypen enthält (z.B. Kommentar und Anzeigen), wird nach dem dominanten Typ klassifiziert oder ansonsten ‚unsure' vergeben

Aufgrund der Diversität der in Zeitungen/Zeitschriften vertretenen Texte wurde für jedes Sample eine nähere Klassifikation des Texttyps vorgenommen, so dass auch dessen Einfluss auf die Verteilung der Redewiedergabetypen untersucht werden kann. Die folgende Abbildung gibt einen ersten Eindruck, welch deutliche Abweichungen hier erkennbar sind. Gezeigt werden nur die Texttypen, für beim Korpusstand vom 25.09.2018 mehr als 10 Samples vorlagen. Die Y-Achse zeigt Prozent der Tokens im Text.



Annotationssystem

Wir unterscheiden die vier Typen direkte, indirekte, frei indirekte (‚erlebte') und erzählte Wiedergabe, sowie die drei Medien Rede (*speech*), Gedanke (*thought*) und Schrift (*writing*), so dass sich insgesamt zwölf Annotationsmöglichkeiten ergeben.

Außerdem annotieren wir die Rahmenformel, die eine direkte oder eine indirekte Wiedergabe einleiten kann. In den Rahmenformeln sowie den Instanzen von erzählter Wiedergabe wird das zentrale Wort markiert, das auf die Sprech-/Gedanken-/Schreibhandlung verweist (z.B. *sagte, Gedanke*). Zudem wird für alle Wiedergabetypen der Sprecher markiert, falls vorhanden.

Während die Unterscheidung der drei Medien nur in Ausnahmefällen problematisch ist, bedürfen die vier Typen genauerer Definitionen.

Die direkte Wiedergabe (*direct*) ist eine wörtlich zitierte Äußerung einer Figur. Sie kann von einer Rahmenformel eingeleitet werden und als einziger Wiedergabetyp Anführungszeichen verwenden.

Er fragte: „Wo ist das Mittagessen?"

Die indirekte Wiedergabe (*indirect*) ist eine nicht-wörtliche Wiedergabe einer Äußerung. Sie ist in unserem Annotationssystem formal definiert und besteht aus einer Rahmenformel und einem abhängigen Nebensatz, der häufig im Konjunktiv steht. Dies kann ein Nebensatz mit Verbzweitstellung sein, mit *dass*, *ob* oder *w*-Fragewort oder ein (erweiterter) Infinitivsatz.

Er fragte, wo das Mittagessen sei.

Die freie indirekte Wiedergabe (*free indirect*) – in der Literatur oft ‚erlebte Rede' genannt – definiert sich über die Überlagerung von Figuren- und Erzählerstimme und ist daher eine typische Form fiktionaler Texte. Sie weist keine Rahmenformel und keine sonstigen formalen Markierungen wie Anführungszeichen auf. Finden sich Elemente der Erzählerstimme wie das Tempus Präteritum oder Personalpronomen der dritten Person und gleichzeitig Elemente der Figurenstimme wie Deiktika, Subjektivität, Ausrufe oder figurentypischer Wortschatz, sind dies Indikatoren für freie indirekte Wiedergabe.

Woher sollte er denn jetzt bloß ein Mittagessen bekommen?

Die erzählte Wiedergabe (*reported*) ist die Erwähnung eines Sprech-, Denk oder Schreibakts, aus der man nicht auf den eigentlichen Inhalt schließen kann. Hinweise auf erzählte Wiedergabe geben Wiedergabewörter, die Thematisierung einer Wiedergabehandlung sowie der Inhalt derselben.

Er sprach über das Mittagessen.

Ein Sonderfall sind uneingeleitete Konjunktivsätze, die zur Wiedergabe verwendet werden. Diese werden als Mischform zwischen indirekter und frei indirekter Wiedergabe markiert.

Sie stellte viele Fragen. Wo sei das Mittagessen?

Darüber hinaus gibt es zusätzliche Attribute, die Besonderheiten bei der Wiedergabe markieren und in der folgenden Tabelle dargestellt werden:

level	Verschachtelungstiefe der Wiedergabe
non-fact	nicht-faktische Wiedergaben (z.B. Negationen, zukünftigen Aussagen oder Absichten)
border	Fälle, die an der Grenze von Rede-, Gedanken- oder Schriftwiedergabe liegen, also nicht alle prototypischen Kriterien der jeweiligen Wiedergabeart erfüllen
prag	sprachliche Wendung, die das Muster einer Wiedergabe aufweist, pragmatisch aber einen anderen Zweck erfüllt (z.B. Höflichkeitsfloskeln)
metaph	Metaphern in Form von Wiedergaben (z.B. *Die Klugheit riet mir davon ab.*)

Die detaillierten Annotationsrichtlinien können unter http://redewiedergabe.de/richtlinien/richtlinien.html eingesehen werden.

Annotationsprozess

Die genaue Identifizierung und Klassifizierung der Redewiedergaben auf der Grundlage des detaillierten Annotationssystems ist eine schwierige Aufgabe.

Jedes Sample des RW-Korpus durchläuft darum einen mehrschrittigen Prozess. Zunächst wird es von zwei Annotatoren unabhängig voneinander annotiert. Danach vergleicht ein weiterer Experte die Annotationen und erstellt, falls notwendig, eine Konsens-Annotation, die dann ins finale Korpus aufgenommen wird. Jedes Sample wird also von drei Personen bearbeitet, um größtmögliche Konsistenz zu gewährleisten.

Die Annotatoren arbeiten mit dem eclipse-basierten Annotationswerkzeug ATHEN (entwickelt von Markus Krug im Projekt Kallimachos, www.kallimachos.de), für das im Projekt eine spezielle Oberfläche für die Redewiedergabe-Annotation implementiert wurde (für eine detaillierte Beschreibung vgl. auch Krug et al. 2018a). Das Werkzeug ist frei verfügbar unter der Adresse http://ki.informatik.uni-wuerzburg.de/nappi/release/.

Verfügbarmachung und Ausblick

Das Beta-Release wird in einem standardisierten und dokumentierten Textformat im Langzeitarchiv des Instituts für Deutsche Sprache zur freien Nutzung zur Verfügung gestellt (https://repos.ids-mannheim.de/). Spätestens für das finale Release im Frühjahr 2020 garantieren wir ein TEI-

kompatibles XML-Format. Zudem wird weiteres im Kontext des Redewiedergabe-Projekts entstandenes Material zur Verfügung gestellt, wie nur einfach annotiertes Textmaterial und annotierte Volltexte. Im Jahr 2020 werden auch Werkzeuge fertiggestellt sein, die es erlauben, in Texten verschiedene Formen der Redewiedergabe automatisch zu erkennen.

Nutzungsszenarien für das Korpus sind vielfältig: Aus NLP-Perspektive kann es als Test- und Trainingsmaterial für automatische Redewiedergabeerkenner verwendet werden. Aus linguistischer Perspektive bieten sich Korpusstudien zu sprachlichen Eigenheiten der Redewiedergabe an, wie z.B. die laufenden Studien zu Redewiedergabeeinleitern von Tu. Aus literaturwissenschaftlicher Perspektive erlaubt das Korpus z.B. Untersuchungen zu der Häufigkeit und Form von Wiedergaben in Erzähltexten in ihrer Relation zur Figurencharakterisierung.

Fußnoten

1. Die ersten beiden Autoren haben zu gleichen Teilen an der Erstellung dieses Beitrags mitgewirkt.

Bibliographie

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe.* Ein Beitrag zur quantitativen Narratologie (=Narratologia 47). Berlin: De Gruyter.

Engelberg, Stefan (2015): *„Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons“*, in: **Eichinger, Ludwig M. (eds.)**: *Sprachwissenschaft im Fokus*. Positionsbestimmungen und Perspektiven. Jahrbuch 2014 des IDS. Tübingen: Narr 205-230.

Gabriel, Gottfried (2007): *„Fiktion“*, in: **Fricke, Harald et al. (eds.)**: *Reallexikon der deutschen Literaturwissenschaft*. Bd. 1: A-G. Berlin / New York: De Gruyter 594-598.

Genette, Gérard (2010): *Die Erzählung.* 3., durchges. und korrigierte Aufl. Paderborn: Fink.

Hauser, Stefan (2008): *„Beobachtungen zur Redewiedergabe in der Tagespresse. Eine kontrastive Analyse“*, in: **Lüger, Heinz-Helmut / Lenk, Hartmut E.H. (eds.)**: *Kontrastive Medienlinguistik*. Landau: Verlag Empirische Pädagogik 271-286.

Krug, Markus / Tu, Ngoc Duyen Tanja / Weimer, Lukas / Reger, Isabella / Konle, Leonard / Jannidis, Fotis / Puppe, Frank(2018a): *„Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment“*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 19-21.

Krug, Markus / Weimer, Lukas / Reger, Isabella / Macharowsky, Luisa / Feldhaus, Stephan / Puppe, Frank / Jannidis, Fotis (2018b): *Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]*. DARIAH-DE Working Papers Nr. 27, Göttingen: DARIAH-DE, 2018, URN: urn:nbn:de:gbv:7-dariah-2018-2-9.

Martínez, Matías / Scheffel, Michael (2016): *Einführung in die Erzähltheorie*. 10. Auflage. München: C.H.Beck.

Semino, Elena / Short, Mick (2004): *Corpus stylistics.* Speech, writing and thought presentation in a corpus of English writing. London / New York: Routledge.

Den Menschen als Zeichen lesen. Quantitative Lesarten körperlicher Zeichenhaftigkeit in visuellen Medien

Howanitz, Gernot
gernot.howanitz@uni-passau.de
Universität Passau, Deutschland

Radisch, Erik
erik.radisch@uni-passau.de
Universität Passau, Deutschland

Decker, Jan-Oliver
Jan-Oliver.Decker@Uni-Passau.De
Universität Passau, Deutschland

Rehbein, Malte
malte.rehbein@uni-passau.de
Universität Passau, Deutschland

Trotz der Textlastigkeit der Digital Humanities ist in der letzten Zeit ein gewisser beginnender „visualistic turn“ in der Disziplin zu beobachten, der in den Kulturwissenschaften bereits seit einiger Zeit konstatiert wird. (Sachs-Hombach 1993) Auch der Schwerpunkt dieser Konferenz weist in diese Richtung. Visuelle Medien bieten zweifelsohne eine große Chance für eine Weiterentwicklung der Digital Humanities, trotzdem ist festzustellen, dass sich dieser Turn bisher noch größtenteils auf die Kunstgeschichte konzentriert.[1] Das liegt aufgrund des visuellen Schwerpunktes dieser Disziplin zwar nahe, jedoch widmen sich auch andere Disziplinen, in unserem Falle die Kulturwissenschaften, aus der Perspektive ihrer spezifischen Fragestellungen den visuellen Medien.

Das Anliegen dieses Vortrages ist es, einen Zugang zu präsentieren, der es Kulturwissenschaftler/inne/n ermöglicht, große Bild- und Videokorpora in einem Methodenmix sowohl qualitativer als auch quantitativer Verfahren zu erfassen. Auf der letzten DHd-Konferenz haben wir erste Vorarbeiten in Form eines Ansatzes präsentiert, der aus visuellen Medien den relevanten symbolischen Kontext identifiziert und damit eine automatisierte Informationsextraktion und -analyse ermöglicht (Bermeitinger/Howanitz/Radisch 2018). Ein großes Videokorpus konnte auf diese Weise bearbeitet werden, ohne alle Videos im einzelnen ansehen zu müssen. Das Auftreten von Menschen in diesen Videos oder Bildern wurde dabei lediglich am Rande geschnitten, indem die Möglichkeit in Erwägung gezogen wurde, bestimmte Personen auf Einzelframes zu identifizieren.

Hier möchte unser diesjähriger Vortrag ansetzen und Jurij Lotmans (1984) semiotischen Ansätzen folgend menschliche (Ab-)Bilder zeichenhaft in visuellen Korpora lesen. Als

Beispielkorpus dient uns eine Sammlung von Starpostkarten von Marlene Dietrich aus den 1930er und 1940er Jahren. Ziel ist es, ähnlich wie bei Distant Reading-Ansätzen zu versuchen, versteckte Muster, Gemeinsamkeiten und Abweichungen ohne Betrachtung der Einzelbilder identifizieren zu können. Dabei erweist sich das Erscheinen von Menschen in diesen Bildern als zentral und muss besonders berücksichtigt werden. Um menschliche Abbildungen in visuellen Korpora möglichst umfassend analysieren zu können, schlagen wir eine dreistufige Herangehensweise vor, nämlich die Identifikation einzelner Personen, die Analyse von Körperhaltungen, sowie die Mimikerfassung. Im Folgenden werden alle drei Teilbereiche vorgestellt.

Identifikation

In vielen Fällen der Analyse menschlicher Zeichenhaftigkeit ist es von großem Vorteil, bestimmte Personen identifizieren zu können. So ist es zum Beispiel interessant, die Auswahl an Personen in einem visuellen Korpus durch Kriterien der Relevanz in bezug auf eine konkrete Forschungsfrage einzuschränken. Auch können auf diese Weise verschiedene Protagonisten und in weiterer Folge deren Kookkurrenzen innerhalb des Korpus identifiziert werden, was Rückschlüsse auf Personenkonstellationen zulässt. Handelt es sich bei den analysierten Bildern um Einzelframes aus Videos, erlaubt es dieser Ansatz zudem, resultierende Muster der Personenkonstellationen als Netzwerke zu visualisieren – analog zu der von Franco Moretti etablierten Netzwerkanalyse von Dramen, die in den letzten Jahren zahlreiche Anwendung gefunden hat.

Zum Erkennen von Gesichtern verwenden wir David Sandbergs *Facenet*, das auf OpenFace (Amos/Ludwiczuk/Satyanarayanan 2016) beruht und folgen dabei den Empfehlungen unseres Kollegen Sebastian Gassner (Gassner 2018). Hinzuweisen ist aber hier ebenfalls auf das Distant-Viewing Toolkit von Lauren Tilton und Taylor Arnold, das ebenfalls ermöglicht, bestimmte Akteure in Filmen identifizieren zu können (Tilton/Arnold 2018).

Haltung

Mit der bloßen Anwesenheit bestimmter Personen in visuellen Medien ist aber noch kaum Aussagen über deren Zeichenhaftigkeit möglich. Vielmehr gilt es auch die Körperhaltung der Personen einer eingehenden Analyse zu unterziehen. Gerade hier manifestiert sich die Zeichenhaftigkeit eines Körpers besonders. Der gesamte Körper fungiert als ein Zeichen. Je nach Haltung können völlig unterschiedliche Aussagen getätigt werden, die nur zum Teil einerseits bewusst intendiert sind, andererseits bewusst wahrgenommen werden. Hier erlauben es algorithmische Verfahren, einen Blick auf das Phänomen zeichenhafter Körperlichkeit zu ermöglichen, der etwas weniger durch menschliche Erfahrung beeinflusst ist als dies bei qualitativen Ansätzen der Fall ist.

Körperhaltungen lassen sich auf Grundlage einiger wichtiger Keypoints (Bourdev/Malik 2009) analysieren, die die Position der Hände, Ellenbogen, Schultern, der Hüfte, der Knie und der Füße sowie des Kopfes angeben. Je nachdem, wie diese Positionen in Relation zueinander stehen, ergeben sich verschiedene Haltungen die Rückschlüsse auf symbolische, über den Körper kommunizierte Bedeutungen erlauben. Solche Ansätze wurden erst kürzlich von Leonardo Impetti und Franco Moretti zur Klassifizierung von Aby Warburgs Bilderatlas verfolgt (Impett/Moretti 2017). Ähnliche Ansätze werden auch in unserem Projekt herangezogen. So erlaubt es beispielsweise die Clusterung nach den Keypoints in unserem Korpus bestimmte Körperhaltungen zusammenzufassen und deren Entwicklung diachron über die Zeitspanne des Korpus zu untersuchen. Der Zusammenhang zwischen Körperhaltungen und Genderstereotypen ist vieldiskutiert (für einen körpersemiotischen Zugang vgl. Mühlen-Achs 1998); gerade im Zusammenhang mit Marlene Dietrichs androgyner Selbstinszenierung, die männlich konnotierte Elemente aufgreift, ermöglicht eine quantitative Analyse dieser Körperhaltungen diesbezüglich neue Einblicke. Daneben interessieren uns auch Kopfhaltung und insbesondere die Blickrichtung, die ihrerseits die Grundparameter der zugrunde liegenden kommunikativen Situation offenbart.

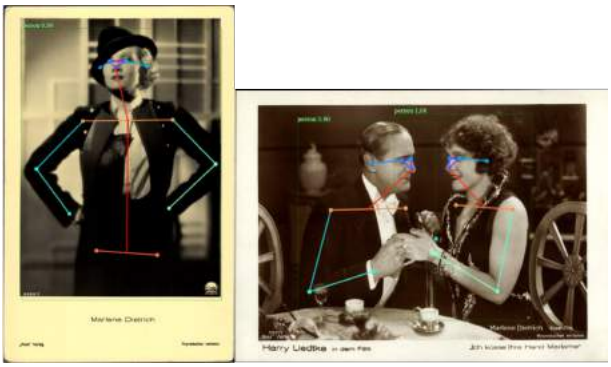


Abb. 1: *Detectron* identifiziert Marlene Dietrichs typische Doppelamphorenhaltung (links), rechts eine Figurenkonstellation.

Der Ansatz der Posen-Erkennung hat noch ein wesentlich höheres Potential für zukünftige Forschung. Hierüber könnten sich auch Bewegungsabläufe in Videos analysieren lassen, da sich bestimmte Bedeutungsebenen erst im Kontext des Zusammenspiels ergeben.

Technisch greifen wir einerseits auf die Keypoint-Ebene von *Detectron* (Girshik et al. 2018) zurück, eines Frameworks für Deep Learning und Objekterkennung, zur Verfügung gestellt von *Facebook Artificial Intelligence Research*, andererseits auf das von Cao et al. (2017) veröffentlichte System zur Posen-Erkennung, das im Vergleich zu *Detectron* weitere Keypoints identifiziert und darüber hinaus zwischen den beiden Körperhälften unterscheiden kann.

Ein Nachteil der hier verwendeten Algorithmen ist, dass sie die Keypoints nur zweidimensional erkennen, also die Position auf dem Bild erfassen, nicht aber deren Lage im Raum. Für eine dreidimensionale Erfassung zumindest des Gesichts wird mit *OpenFace* (Baltrušaitis et al. 2018) deshalb noch ein dritter Algorithmus verwendet, der eine Abschätzung der dreidimensionalen Lage – insbesondere auch der Kameraposition in Relation zum Gesicht – und der Blickrichtung erlaubt. Mithilfe dieser Informationen können grobe Rückschlüsse auf die Gesamtkomposition einer Szene gemacht werden.

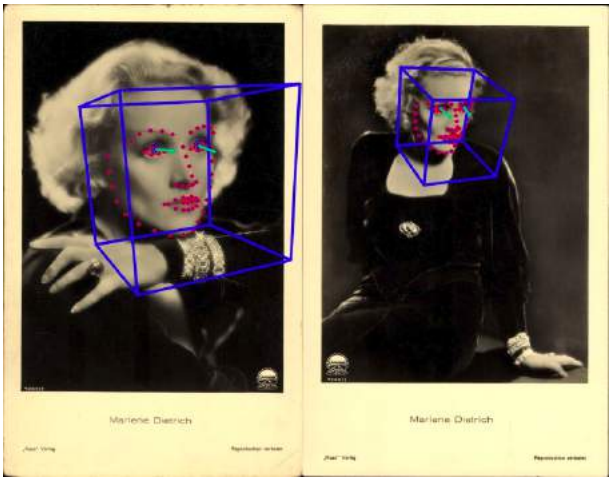


Abb. 2: OpenFace blickt Marlene Dietrich ins Gesicht: Lage im Raum (blau), Keypoints (rot), berechnete Blickrichtung (grün)

Mimik

OpenFace erlaubt es des Weiteren eine weitere Ebene menschlicher Zeichenhaftigkeit in visuellen Medien zu erschließen – das Gesicht. Dabei wird neben der dreidimensionalen Position des Kopfes ebenfalls alle wichtigen Punkte innerhalb des Gesichtes, darunter die Umrisse von Augen, Nase, Lippen, Augenbrauen und Kinn erkannt (Abb. 2). Der Algorithmus ermittelt die Abweichung dieser Gesichtspunkte von einem Standardmodell, welche wiederum mithilfe des Facial Action Coding System (FACS), ein System der Taxonomie menschlicher Gesichtsausdrücke, beschrieben werden können (Ekman/Friesen 1978). FACS besteht aus einer Reihe von Action Units, die basale Gesichtsbewegungen klassifizieren, wie beispielsweise die Hebung der inneren Augenbrauen (AU1), die Senkung der Augenbrauen (AU4), das Zusammenkneifen der Lider (AU7) und das Anziehen der Mundwinkel (AU12).

Diese AUs stellen einen Standard zur Beschreibung von Gesichtsausdrücken dar.

Damit stellt OpenFace ein ideales Werkzeug dar, um Mimiken in visuellen Medien auslesen zu können, mit einer großen potenziellen Breite an Einsatzmöglichkeiten. So lassen sich beispielsweise damit Bilder mit Personen nach Gesichtsausdrücken clustern oder die Häufigkeit bestimmter Action Units im Zeitverlauf zu analysieren, um so Rückschlüsse über veränderte kulturelle Praxen machen zu können. Im Falle von Marlene Dietrich ist beispielsweise beobachtbar, dass AU2 (Hebung der äußeren Augenbrauen) eine Zeit lang häufig auftritt, was auf die Schminkgewohnheiten der damaligen Zeit zurückzuführen ist. Auf diese Weise können auch Modetrends und Stilfragen bis zu einem gewissen Grad quantitativ erfasst werden.

Solche Informationen können aber auch in anderen Kontexten von großen analytischen Wert sein. In einem Korpus von 290 Zigarettenbildern zeigen sich etwa klare Unterschiede zwischen den Geschlechtern. AU6 (Wangen heben) kombiniert mit AU12 (Mundwinkel heben) entspricht etwa der Emotion „Freude“, die bei den Frauen häufiger anzutreffen ist. AU15 (Mundwinkel senken) ist hingegen bei den Männern weitaus verbreiteter (siehe Abb. 3).

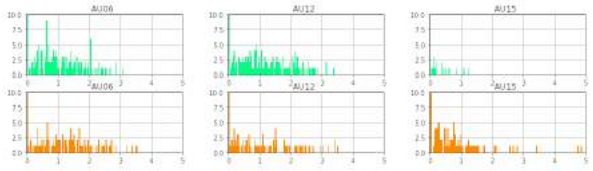


Abb. 3: Histogramme von AU6, AU12 und AU15 bei Frauen (grün) und Männern (orange).

Das Puzzle setzt sich zusammen: Quantitative Lesarten körperlicher Zeichenhaftigkeit

Die oben beschriebene Methodenkombination ist einerseits nötig, um körperliche Zeichenhaftigkeit in visuellen Medien auf den verschiedenen skizzierten Bedeutungsebenen erfassen zu können, erschwert aber andererseits das Arbeiten mit dem Korpus und verlangt deshalb nach einer entsprechenden methodischen Aufbereitung. Das erste Ziel unseres Beitrages ist deshalb, die Methoden durch ein Testkorpus von rund 200 Bildern systematisch zu evaluieren. Anschließend wenden wir die Methoden auf ein Korpus von mehreren tausend deutschen Starpostkarten vorwiegend aus den 1930ern und 1940ern an, die wir in Zusammenarbeit mit dem Filmmuseum Potsdam digitalisiert haben. Als Basis unserer Arbeiten dient ein eigenständiges Framework, das wir für mehrere Forschungsprojekte innerhalb des vom BMBF geförderten Passau Centre for eHumanities (PACE, Fördernummer: 01UG1602) entwickelt haben und die es ermöglichen verschiedene quantitative Aspekte mit qualitativen Fragestellungen kurzzuschließen. Wie bereits erwähnt, ist es ein *desideratum,* das System auch für Bewegtbilder zu öffnen und weiterzuentwickeln, auch wenn der damit verbundene Aufwand nicht trivial ist.

Fußnoten

1. Der Schwerpunkt auf Kunstgeschichte spiegelt sich beispielsweise im DFG-Schwerpunktprogramm “Das digitale Bild” wieder, dass zwar prinzipiell auch für die Kulturwissenschaften geöffnet war, aber in der Ausschreibung einen deutlichen Fokus auf die Kunstgeschichte nahm.

Bibliographie

Amos, Brandon / Ludwiczuk, Bartosz / Satyanarayanan, Mahadev (2016): *"Openface: A general-purpose face recognition library with mobile applications,"* CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016. https://www.cs.cmu.edu/~satya/docdir/CMU-CS-16-118.pdf [letzter Zugriff 29. 9. 2018].

Arnold, Taylor / Tilton, Lauren (2018): *„Distant Viewing Toolkit (DVT) for the Cultural Analysis of Moving Images“*,

https://github.com/distant-viewing/dvt [letzter Zugriff 29. 9. 2018].

Baltrušaitis, Tadas / Zadeh, Amir / Chong Lim, Yao / Morency, Louis-Philippe (2018): *"OpenFace 2.0: Facial Behavior Analysis Toolkit"*, IEEE International Conference on Automatic Face and Gesture Recognition 2018., https://ieeexplore.ieee.org/document/8373812 [letzter Zugriff 29. 9. 2018].

Bermeitinger, Bernhard / Howanitz, Gernot / Radisch, Erik (2018): *"Contextualising Bandera. Eine Distant Watching-Methode"*, DHd 2018.

Bourdev, Lubomir / Malik, Jitendra (2009): *"Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations"*, ICCV 2009. https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/human/poselets_iccv09.pdf [letzter Zugriff 29. 9. 2018].

Ekman, Paul / Friesen, Wallace V. (1978): *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting.

Gassner, Sebastian (2018): *„How to Use David Sandberg's Facenet Implementation"*, https://github.com/sepastian/facenet/blob/master/HOWTO.md [letzter Zugriff 29. 9. 2018].

Girshick, Ross / Radosavovic, Ilija / Gkioxari, Georgia / Dollár, Piotr / He, Kaiming (2018): *Detectron*. https://github.com/facebookresearch/detectron [letzter Zugriff 29. 9. 2018].

Cao, Zhe / Simon, Tomas / Wei, Shih-En / Sheikh, Yaser (2017): *„Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields"*. CVPR 2017. https://arxiv.org/pdf/1611.08050.pdf [letzter Zugriff 29. 9. 2018].

I **mpett, Leonardo / Moretti, Franco (2017)**: *"Totentanz. Operationalizing Aby Warburg's Pathosformeln."* In: Literary Lab Pamphlet 16, November 2017, https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf [letzter Zugriff 29. 9. 2018].

Lotman, Jurij (1984): *"O semiosfere"*. Učen. zap. Tart. gos. un-ta 641 (1984), 5-23. (=Trudy po znakovym sistemam 17).

Mühlen-Achs, Gitta (1998): *Geschlecht bewusst gemacht. Körpersprachliche Inszenierungen*. München: Frauenoffensive.

Sachs-Hombach, Klaus (1993): *Das Bild als kommunikatives Medium. Elemente einer allgemeinen Bildwissenschaft.* Köln: Halem.

Detecting Character References in Literary Novels using a Two Stage Contextual Deep Learning approach

Krug, Markus
markus.krug@informatik.uni-wuerzburg.de
Chair of Applied Computer Science and Artificial Intelligence, Universität Würzburg, Deutschland

Kempf, Sebastian
sebastian.kempf@informatik.uni-wuerzburg.de
Chair of Applied Computer Science and Artificial Intelligence, Universität Würzburg, Deutschland

David, Schmidt
david.schmidt@informatik.uni-wuerzburg.de
Chair of Applied Computer Science and Artificial Intelligence, Universität Würzburg, Deutschland

Lukas, Weimer
lukas.weimer@uni-wuerzburg.de
Chair of Literary Computing, Universität Würzburg, Deutschland

Frank, Puppe
frank.puppe@informatik.uni-wuerzburg.de
Chair of Applied Computer Science and Artificial Intelligence, Universität Würzburg, Deutschland

Motivation

In recent years analyzing constellations of fictional entities in literary fiction has seen a lot of interest (Elson et al. 2010, Agarwal et al. 2012). Those constellations are often visualized as networks of entities apparent in the document. Even though the pipelines used for preprocessing vary drastically they all share some common steps. In order to draw a network, one needs to define nodes and edges. An obvious choice for the nodes are the fictional entities. Those entities appear either as pronouns, nominal references or names. The detection of those character references (CR) was used for a multitude of different applications in automatic digital humanities processing, most notably genre detection (Hettinger et al. 2015) and coreference resolution (e.g. Lee et al. 2013).

Related Work

This section only mentions the most dominant work in terms of Named Entity Recognition techniques as well as those with comparable results in terms of domain. The standard approach is to segment the input document into sentences and apply a sequence classifier on these sentences. The most robust classifier applied to this task was a Conditional Random Field (CRF) (Lafferty et al. 2001). Until the success of deep learning approaches, the classifier published by Stanford (Finkel et al. 2005) and its adaptation to German (Faruqui et al. 2010) were the dominant approaches. Lately the combination of a Bi-LSTM with word-embeddings (e.g. Mikolov et al. 2013) which automatically derived features for a CRF classifier in a deep learning approach surpassed manually created features (Huang et al. 2015). Most recently, Riedl and Padó (2018) included pretraining into the Bi-LSTM-CRF architecture and achieved state of the art results. A maximum entropy classifier with cluster features derived by word2vec and manually crafted features yielded an F1-score of 90% (Jannidis et al.

2017), serving as the only comparable work for German literary data.

However, all of these approaches classify each sentence and every token inside on their own so that subsequent sentences do not benefit from previously detected names. This does not only overcomplicate the detection, it can also introduce inconsistent results (e.g. “Effi” is detected as a name in sentence 10 and 27, but was not detected in sentence 25). Introducing no dependencies between each individual reference seems a wasted opportunity, especially in novels, because the same character reappears multiple times. This paper experimented with network architectures to leverage this shortcoming. The idea of this approach is not new and can even be dated back to the Brills Tagger (Brill 1992) - the classification by two separate classifiers each with its individual perspective on the problem.

Data

The corpus DROC (Krug et al. 2018) provides the data used in this paper. It contains about 393.000 tokens from 90 different samples taken from German novels. Each sample comprises at least one chapter. In total, this corpus contains about 53.000 manually annotated character references. 100-dimensional Word2Vec word embeddings trained on 1700 novels of project Gutenberg[1] were used as secondary input.

Methods

The method used in this paper follows the intuition that, especially in literary fiction, entities appear many times throughout the text. Because each document introduces its own fictional world, each word (meaning the set of all appearances of a token with the same string) has a dominant meaning. However, not every instance of a word can be easily detected. While some might be surrounded by verbs of communication (“sagen”, “antworten”, ...), others might only be surrounded by stop words, which are not beneficial for classification. Therefore, this work introduces two passes through the text. The first pass tries to assign the dominant meaning to a word and is assumed to produce a high recall but a mediocre precision. The purpose of the second pass is to disambiguate individual instances which have been classified as a character reference but could have multiple senses. Furthermore, while the first pass might detect “Effi” and “Briest” as references, there is no information about whether the string “Effi Briest” is a single reference or two distinct references. This is solved in the second pass, which is trained for a sequential prediction and is supposed to detect the exact span of a reference.

The architectures of both neural networks can be depicted as follows:

An instance fed into the first network consists of a list of tuples, each comprising the span of the token, encoded by a word embedding, as well as a left context and a right context. Our previous work determined a context of the previous two and the next two tokens (also encoded by a pre-trained word embedding) as best performing for the determination of character references. The last input vector was derived by a Bi-LSTM character encoding of the target word. The tuples were arranged in order of appearance in the original text and encoded by a Bi-LSTM, feeding an additional tupel at each time step. The Bi-LSTM subsequently generates a condensed representation of those tuples into a vector of 256 units. The intuition is that this vector contains the most informative parts of all contexts for a given target word. The network is trained using log-loss and predicts whether the target word is a reference or not.

The network was trained for 15 epoches on 58 documents (longer training did not necessarily result in a better classification accuracy) and applied to a separate set spanning 14 documents. This second set is then used to train the second network, using Bi-LSTM character embeddings with a subsequent Bi-LSTM. However, the network is only applied to tokens that had been classified as a character reference in the first pass. This follows the intuition that it can now be decided if the current instance is of a different semantic category, which can be detected by analyzing its context. The input of this network is the snippet around the target word with a context size of two. The second task of this network, detecting the exact bounds of a reference, is done by predicting labels in an I-O-B setting. It is noteworthy that words that were not detected in the first pass can not be recovered. The second network is trained with 25 epochs and finally tested on 18 test documents.

Evaluation

We compared the architecture described in Section 4 (denoted 2-stage) with the state of the art architecture (Bi-LSTM-CRF) similar to Riedl and Padó (2018). A Bi-LSTM-CRF using character embeddings in Tensorflow[2] was implemented and applied to the data of DROC. At the current stage our implementation does not make use of pretraining. We used the 90 documents and split them into five folds each comprising 18 different test documents. The remaining 72 documents are used for training. The results are shown in Table 1.

System	Token			Entity		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
2-stage	89.6	74.9	80.7	89.2	69.5	78.1
Bi-LSTM-CRF	91.8	91.6	91.7	90.9	90.8	90.8

Table 1: Results of the two systems applied to DROC. The numbers were derived in a 5-fold scenario and are noted in %. Evaluation is done on token and on entity level using Precision, Recall and F1.

Even though the 2-stage approach seems intuitive at first, it can not compete with the results obtained by the state of the art architecture. Surprisingly, the architecture failed to provide a high recall (this is already apparent after the first pass, where the recall is similar to the state of the art system, however no exact borders can be predicted). A possible explanation for this result is the high amount of about 50% of tokens with only a single appearance in the text. Since only two context tokens to the left and the right are used, the architecture has a shortcoming compared to the Bi-LSTM-CRF, which encodes the entire sentence. The architecture does

especially fail to recognize references that contain the token "von" (such as "Baron von Instetten)". While being competitive in terms of the precision, further work has to be done to increase the recall for this approach.

Conclusion

This paper presents a 2-stage contextual approach to detect character references using deep learning. The results show that while the precision yields competitive results, the recall is still much lower. Possible approaches for this shortcoming might be changing the loss function - currently a false negative and a false positive yield the same penalty - and combining both models. The state of the art model can then be used for words that only appear a single time in the text and the 2-stage approach for words appearing more than once. This could retain the high quality while still generating a consistent labeling by making use of the dependencies between individual appearances of a word.

Fußnoten

1. https://www.gutenberg.org/
2. https://www.tensorflow.org/

Bibliographie

Agarwal, A., Corvalan, A., Jensen, J., & Rambow, O. (2012): *Social network analysis of alice in wonderland. In Proceedings of the NAACL-HLT 2012 Workshop on computational linguistics for literature* (pp. 88-96).

Brill, E. (1992, March): *A simple rule-based part of speech tagger.* In Proceedings of the third conference on Applied natural language processing (pp. 152-155). Association for Computational Linguistics.

Elson, D. K., Dames, N., & McKeown, K. R. (2010, July): *Extracting social networks from literary fiction*. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 138-147). Association for Computational Linguistics.

Faruqui, M., Padó, S., & Sprachverarbeitung, M. (2010, September): *Training and Evaluating a German Named Entity Recognizer with Semantic Generalization*. In KONVENS (pp. 129-133).

Finkel, J. R., Grenager, T., & Manning, C. (2005, June): *Incorporating non-local information into information extraction systems by gibbs sampling*. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.

Huang, Z., Xu, W., & Yu, K. (2015): *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991.

Hettinger, L., Becker, M., Reger, I., Jannidis, F., & Hotho, A. (2015, September): *Genre classification on German novels.* In Database and Expert Systems Applications (DEXA), 2015 26th International Workshop on (pp. 249-253). IEEE.

Jannidis, F., Reger, I., Weimer, L., Krug, M., & Puppe, F. (2017): *Automatische Erkennung von Figuren in deutschsprachigen Romanen.*

Krug Markus, Puppe Frank, Reger Isabella, Weimer Lukas, Macharowsky Luisa, Feldhaus Stephan, Jannidis Fotis (2018, April): *Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]*. DARIAH-DE Working Papers Nr. 27. Göttingen: DARIAH-DE. URN: urn:nbn:de:gbv:7-dariah-2018-2-9

Lafferty, J., McCallum, A., & Pereira, F. C. (2001): *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.*

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013): *Deterministic coreference resolution based on entity-centric, precision-ranked rules*. Computational Linguistics, 39(4), 885-916.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013): *Distributed representations of words and phrases and their compositionality*. In Advances in neural information processing systems (pp. 3111-3119).

Riedl, M., & Padó, S. (2018): *A Named Entity Recognition Shootout for German*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 120-125).

DH is the Study of dead Dudes

Hall, Mark
mark.hall@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

DH is the Study of dead Dudes

Die Digital Humanities (DH) werden oft als Möglichkeit gesehen neue Methoden und neue Datenmengen zu nutzen um neue Ansätze auf Fragen der Geisteswissenschaften zu entwickeln (Barry 2011). Als Teil des DH Forschungsprozesses muss in vielen Fällen ein Datensatz oder eine Person als Fokus der Studie gewählt werden. Dass dabei existierende soziale Trends verstärkt werden, besonders was den Fokus auf den Kanon betrifft ist bekannt und in verschiedenen Hinsichten problematisiert worden (Liu 2012; McPherson 2012; Wernimont 2013; Fiormonte 2016; Gallon 2016; Rhody 2016). Es basiert aber keine dieser Kritiken auf einer expliziten Analyse der im DH Bereich unternommenen Forschung. Dieser Beitrag unternimmt einen ersten Schritt um dies zu korrigieren und präsentiert einen quantitativen, distant-reading Ansatz um der Frage nach Gender, Sprache, und Herkunftsland von in DH untersuchten Personen nachzugehen.

Methode

Die hier präsentierte Analyse basiert auf den Konferenzabstracts der DHd Konferenzen 2016, 2017, und 2018. Aus den Abstracts wurden alle erwähnten Personen manuell identifiziert. Auf eine automatisierte Identifikation wurde verzichtet, da dies nur eine neue Fehlerquelle in die Analyse einführen würde, ohne inhaltlich irgendeinen Unterschied zu machen. Da die Abstracttexte selbst nicht

im Detail gelesen wurden, ist der Ansatz als distant-reading zu verstehen und berücksichtigt auch keine Abstracts, die sich mit diesem Thema beschäftigen, aber keine Personen expliziten erwähnt.

Für die Identifikation der Personen wurden die Inklusions- und Exklusionsregeln in Tabelle 1 angewandt. Das zentrale Prinzip für das Zählen einer Person ist, dass sie explizit ausgewählt worden sein muss. Das heißt, Personen werden nur gezählt, wenn sie oder ihr Werk das primäre Untersuchungssubjekt sind (#1) oder weil sie als exemplarisches Beispiel eines Themas (#2) oder als Sample eines Datensatzes (#3) präsentiert werden. Personen, die sowohl als Studiensubjekte auftreten, aber auch selbst akademische Theorien entwickelt haben, werden nur gezählt wenn ihre Arbeiten das Ziel der Analyse sind, nicht aber wenn sie Teil des methodischen Ansatzes sind (#4). Zum Beispiel, wenn die Schriften Adornos automatisiert analysiert werden, dann wird er gezählt. Wenn aber die Theorien Adornos zur Interpretation anderer Daten genutzt werden, dann wird er nicht gezählt. Letztens werden Personen nicht gezählt, wenn sie in einer vollautomatisch erstellten Liste auftreten, bei der kein manueller Eingriff in der Erstellung auftrat (#5). Zum Beispiel wenn bei einer Häufigkeitsanalyse die 20 häufigsten Namen gelistet werden, dann werden diese nicht gezählt, da die Studie Verzerrungen in der Auswahl von Studiensubjekten betrachtet und nicht Verzerrungen in den Datensätzen.

#	Regel	Aktion
1	Person als primäres Untersuchungssubjekt	Inklusion
2	Person als Beispiel für das untersuchte Thema	Inklusion
3	Person als Sample aus dem untersuchten Datensatz	Inklusion
4	Person erwähnt als Theoriegeber_in	Exklusion
5	Person erwähnt als vollautomatisiertes Ergebnis	Exklusion

Tabelle 1: Inklusions- und Exklusionsregeln für die Erstellung des Personendatensatzes.

Für die derart identifizierten Personen wurden dann Gender, Sprache, und Herkunftsland identifiziert, wobei als Quelle primär Wikipedia genutzt wurde. Bei der Klassifikation von Sprache und Herkunftsland wurde eine Mehrfachzuordnung durchgeführt, um Unterschiede zwischen heutigen und historischen Länder und Sprachgrenzen abzubilden und Mehrsprachigkeit der historischen Personen zu berücksichtigen. Aus den insgesamt 230 erwähnten Personen konnte nur für 4 keine Zuordnung durchgeführt werden. Basierend auf der Zuordnung wurde dann pro Abstract die Zahl der Erwähnungen ermittelt. Pro Abstract wurde eine Person die mehrmals erwähnt wurde nur einmal gezählt. Zugleich wurde eine Person die in verschiedenen Abstracts erwähnt wird, pro Abstract je einmal gezählt.

Die Abstracts wurden mit Publikationsjahr und Katgeorie (Vortrag, Poster, Workshop, …) annotiert, wobei hier nur die Vortrags- und Poster-abstracts analysiert werden.

Ergebnisse

Insgesamt wurden 342 Abstracts analysiert (162 Vorträge, 180 Poster), in denen in 104 explizit Personen erwähnt wurden (60 in Vorträgen, 44 in Postern) (Tabelle 2). Bei den Vorträgen stellt dies einen Anteil zwischen 33% und 50% aller Abstracts da, bei den Postern liegt der Anteil zwischen 19% und 28%. Für die Analyse wurden erstens die Erwähnungen pro Abstract aggregiert und dann die Zahl der Abstracts mit zumindest einer Erwähnung betrachtet. Zweitens wurden die Erwähnungen auch unabhängig von den Abstracts analysiert.

	2016	2017	2018			
	Gesamt	mit Personen	Gesamt	mit Personen	Gesamt	mit Personen
Vortrag	60	20	36	18	66	22
Poster	77	18	36	7	67	19

Tabelle 2: Zusammenfassung des analysierten Datensatzes.

Gender

Tabellen 3 und 4 zeigen die Genderverteilung für die Vortrags- und Posterabstracts. In drei Jahren Vortragsabstracts gibt es keinen einzigen in dem nur weibliche Personen erwähnt werden. Bei den Posterabstracts ist die Situation marginal besser, mit je einem Abstract in 2016 und 2017. Insgesamt wurden über die drei Jahre hinweg nur in 15% aller Vortragsabstracts und in 7% aller Posterabstracts, in denen Personen erwähnt wurden, Frauen namentlich erwähnt.

	2016			2017				2018			
	M	F	M&F	?	M	F	M&F	?	M	F	M&F	?
Vortrag	17	0	3	1	15	0	3	0	19	0	3	0
Poster	16	1	0	1	6	1	0	0	18	0	1	0

Tabelle 3: Anzahl an Abstracts die entweder nur männliche (M), nur weibliche (W), sowohl männliche und weibliche (M&W), oder nur unbekannte (?) Personen erwähnen.

Bei den Einzelerwähnungen (Tabelle 4) ist die Situation ähnlich. Bei den Vortragsabstracts machen Frauen 9% der Erwähnungen aus, bei den Posterabstracts 5%.

	2016			2017			2018		
	M	F	?	M	F	?	M	F	?
Vortrag	33	4	3	55	7	0	63	3	0
Poster	19	1	1	6	1	0	33	1	0

Tabelle 4: Anzahl an Erwähnungen männlicher (M), weiblicher (W), oder unbekannter (?) Personen.

Sprache

Bei den Sprachen der erwähnten Personen stellt sich die Situation anders da. Bei den Vortragsabstracts bewegt sich der Anteil an Abstracts mit nicht-deutschsprachigen Personen zwischen 29% und 58% (Tabelle 5) und der Anteil an erwähnten nicht-deutschsprachig Personen zwischen 46% und 60% (Tabelle 6). Bei den Posterabstracts sind die jeweiligen Zahlen etwas niedriger, mit Anteilen zwischen 20% und 60% (Tabellen 7 und 8).

	Deutsch	Englisch	Französisch	Portugisisch	Spanisch	Andere
2016	14	3	2	2	0	3
2017	14	3	4	0	0	2
2018	13	7	3	1	2	5

Tabelle 5: Anzahl an Vorträgsabstracts die mindestens eine Person mit dieser Sprache erwähnen. Sprachen die nur einmal vorkommen sind in die Kategorie “Andere” aggregiert.

	DE	EN	FR	PT	ES	NL	SE	IT	Andere
2016	19	4	2	3	7	0	0	0	2
2017	31	16	7	0	0	2	2	0	0
2018	24	13	11	0	4	0	0	3	5

Tabelle 6: Anzahl an Personen pro Sprache in den Vortragsabstracts.

	Deutsch	Englisch	Andere
2016	14	2	2
2017	2	3	0
2018	11	3	5

Tabelle 7: Anzahl an Posterabstracts die mindestens eine Person mit dieser Sprache erwähnen.

	DE	EN	FR	IT	RU	Andere
2016	16	2	0	0	0	2
2017	2	3	0	0	0	0
2018	16	3	2	2	2	2

Tabelle 8: Anzahl an Personen pro Sprache in den Posterabstracts.

Herkunftsland

Die Ergebnisse für das Herkunftsland sind ähnlich (Tabellen 9, 10, 11 und 12), wobei durch die Unterteilung der deutschsprachigen Länder (Deutschland, Österreich, Schweiz) die Anteile nicht-deutscher Personen etwas höher sind (37% bis 71%).

	DE	AT	CH	FR	GB	US	ES	IT	Andere
2016	11	3	0	2	0	2	0	0	8
2017	10	1	3	4	3	2	0	0	4
2018	12	2	0	3	6	2	2	2	7

Tabelle 9: Anzahl an Vortragsabstracts die mindestens eine Person mit diesem Herkunftsland. erwähnen.

	DE	AT	CH	FR	GB	US	ES	IT	NL	SE	AR	Andere
2016	15	4	0	2	0	3	7	0	0	0	2	6
2017	27	0	3	7	3	15	0	0	2	2	0	2
2018	19	6	0	11	9	7	4	3	0	0	0	7

Tabelle 10: Anzahl an Personen pro Herkunftsland in den Vortragsabstracts.

	DE	AT	GB	US	Andere
2016	10	3	2	0	2
2017	2	0	0	0	3
2018	10	0	0	2	8

Tabelle 11: Anzahl an Posterabstracts die mindestens eine Person mit diesem Herkunftsland erwähnen.

	DE	AT	GB	RU	US	FR	IT	CH	Andere
2016	12	3	2	0	0	0	0	0	2
2017	2	0	1	0	1	0	0	1	0
2018	15	0	0	5	2	2	2	2	4

Tabelle 12: Anzahl an Personen pro Herkunftsland in den Posterabstracts.

Diskussion

Die Ergebnisse der Länder- und Sprachanalyse zeichnen ein positives Bild des Engagements mit Inhalten und Personen außerhalb Deutschlands, mit über 15 verschiedenen Herkunftsländern. Dies ist positiv, da aufgrund der Datenquelle eine Tendenz zu deutschsprachigen Themen zu erwarten ist.

Zugleich sind die Ergebnisse im Genderbereich katastrophal. Über einen Zeitraum von drei Jahren werden in insgesamt 342 Abstracts in 100 (29%) Männer erwähnt, aber nur in 12 Frauen (3.5%), wobei nur in 2 Abstracts (0.5%) ausschließlich weibliche Personen das Ziel der Studie bzw. die erwähnten Beispiele sind. Zwar stellen Abstracts mit Personenerwähnungen nur einen Teil der DH Forschung da, aber trotzdem zeigt die Genderverteilung eine Schieflage, welche den historischen Blickwinkel, der Frauen aus dem öffentlichen Diskurs verdrängt, verstärkt. Der Zugang zu den Outputs von Frauen ist zwar aufgrund der historischen Sozialstrukturen schwieriger, aber die Wahrscheinlichkeit, dass diese Verteilung das Vorhandensein potentieller weiblicher Studiensubjekte auch nur ansatzweise korrekt abbildet ist unwahrscheinlich. Besonders da Studien wie Fischer und Jäschke (2018) quantitativ zeigen, dass es hier hinreichend weibliche Studiensubjekte gäbe. Es ist wesentlich wahrscheinlicher, dass hier ein kognitiver Bias existiert, dem die DH als Disziplin aktiv entgegen treten muss.

Bibliographie

Berry, David M. (2011): *"The Computational Turn: Thinking about the Digital Humanities"*, in: Culture Machine 12: 1–22.

Fiormonte, Domenico (2016): *"Toward a Cultural Critique of Digital Humanities"*, in **Gold, Matthew K / Klein, L. (ed.)**: *Debates in the Digital Humanities 2016*, University of Minnesota Press 438-458.

Fischer, Frank / Jäschke, Robert (2018): *"Liebe und Tod in der Deutschen Nationalbibliothek"*. in DHd2018: Kritik der digitalen Vernunft 261-266.

Gallon, Kim (2016): *"Making a Case for the Black Digital Humanities"*. in **Gold, Matthew K / Klein, L. (ed.)**: *Debates in Digital Humanities 2016* 42-49.

Liu, Alan (2012): *"Where is cultural criticism in the digital humanities?"* in **Gold, Matthew K (ed.)**: *Debates in the Digital Humanities*, University of Minnesota Press 490-509.

McPherson, Tara (2012): *"Why are the digital humanities so white? Or thinking the histories of race and computation"*. in **Gold, Matthew K (ed.)**: *Debates in the Digital Humanities*, 139-160.

Rhody, Lisa M. (2016): *"Why I dig: Feminist approaches to text analysis"*. in **Gold, Matthew K / Klein, L. (ed.)**: *Debates in Digital Humanities 2016* 536-539.

Wernimont, Jacqueline (2013): *Whence Feminism? Assessing Feminist Interventions in Digital Literary Archives*. in DHQ: Digital Humanities Quarterly, 7(1).

Die Generierung von Wortfeldern und ihre Nutzung als Findeheuristik. Ein Erfahrungsbericht zum Wortfeld „medizinisches Personal“

Adelmann, Benedikt
adelmann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

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

Krüger, Katharina
katharina.krueger@uni-hamburg.de
Universität Hamburg, Deutschland

Vauth, Michael
michael.vauth@tuhh.de
Technische Universität Hamburg, Deutschland

In vielen geistes- und sozialwissenschaftlichen Forschungsprojekten wird mit umfangreichen Textkorpora gearbeitet, die einerseits zu groß sind, um im Sinne eines *Close-Reading*-Ansatzes vollständig annotiert zu werden, in denen es andererseits aber einzelne Textpassagen von besonderer Forschungsrelevanz gibt, bei denen eine solche detaillierte Annotation wünschenswert ist. Das – nach Möglichkeit automatisierte – Auffinden solcher relevanten Textpassagen wird dadurch zu einem notwendigen Teilschritt des Arbeitsprozesses.

Eine simple, aber sehr effektive Möglichkeit der Suche ist die Nutzung von Wortfeldern. In der von Trier in den 1930ern entwickelten Wortfeldtheorie (Trier, 1973) wird das sprachliche Lexikon als – lexikalisch oder konzeptuell – strukturiert betrachtet.[1] Zwischen den Wörtern eines Wortfeldes bestehen Zusammenhänge, die es zu einer weitgehend stabilen semantischen Einheit machen, allerdings sind die Grenzen zwischen benachbarten Wortfeldern meist unscharf. Im digitalen Forschungskontext spielen Wortfelder vor allem eine wichtige Rolle für die Taxonomieerstellung im Semantic Web und werden außerdem in der Lexikonforschung als Werkzeug genutzt (z.B. Bindi et al. 1994, Hamp und Feldweg 1997, Fellbaum 1998). Während diese Zugänge theoretisch fundiert sind, wird der Einsatz von Wortfeldern für die digitale Textanalyse bislang eher *ad hoc* genutzt, reflektierte Ansätze wie Heuser und Le-Khac (2011) sind die Ausnahme. Wir versuchen deshalb die Erstellung von Wortfeldern zu systematisieren und ihren Nutzen als Findeheuristik zu bewerten.

In diesem Beitrag berichten wir über die Erstellung von Wortfeldern unter Verwendung dreier Typen von Verfahren, die von der Erstellung aus bestehenden Ressourcen über die manuelle Generierung bis hin zu stark automatisierten Vorgehensweisen reichen, exemplarisch am Wortfeld „medizinisches Personal“. Dabei können Wortfelder u. a. semantische Netze, standardisierte Vokabulare oder Konzepttaxonomien, unstrukturierte Wortlisten oder Kombinationen all dessen sein. Das Wortfeld bildet eine thematische Schnittstelle unserer Projekte im Forschungsverbund „Automatisierte Modellierung hermeneutischer Prozesse – Der Einsatz von Annotationen für sozial- und geisteswissenschaftliche Analysen im Gesundheitsbereich” (hermA, vgl. Gaidys et al., 2017): Das Verbundprojekt möchte anhand des Bereichs Gesundheit erarbeiten, wie die Automatisierung von Annotationen für hermeneutische Analyse- und Erkenntnisprozesse verbessert werden kann.

Verfahren zur Wortfeldgenerierung

Wortfelder aus bestehenden Ressourcen

Im Sinne der Wortfeldtheorie ist es naheliegend, bestehende lexikalische Ressourcen insbesondere strukturierter Daten zu verwenden, um daraus Wortfelder zu erstellen.

In dieser Hinsicht geeignet scheint der Online-Thesaurus *GermaNet* (Henrich und Hinrichs, 2010), der aktuell 164.814 lexikalische Einheiten umfasst, die in 128.100 sogenannte „Synsets“ semantisch strukturiert sind. Die hierarchisch-semantische Struktur des Wortnetzes ermöglicht es, schnell Subfelder zu identifizieren, die bei sehr spezifischem Erkenntnisinteresse nützlich sind. Wir haben für die Wortfelderstellung „medizinisches Personal“ alle 247 Hyponyme verwendet, die dem Begriff *Heilberufler* zugeordnet sind. Mit diesem Wortfeld wurden in einem Korpus aus 32 dystopischen Gegenwartsromanen 63 Begriffe (779 Erwähnungen) gefunden.

Eine Alternative sind kontrollierte Vokabulare, die mit Begriffen operieren, welche durch eine Redaktion definiert werden. Die Begriffe werden mit Metadaten, etwa Synonymen oder Übersetzungen, angereichert und hierarchisiert. Für medizinisches Personal wurde exemplarisch am größten deutschsprachigen Vokabular gearbeitet, der gemeinsamen Normdatei (GND[2]) der Deutschen Nationalbibliothek (DNB). Wir haben die in der GND enthaltenen Begriffe zum Themenfeld „medizinisches Personal“ recherchiert, indem wir die hierarchische Struktur der Begriffe ausgehend vom allgemeinen Begriff „Arzt“ durchgegangen sind. In der Folge wurden dann alle relevanten Teilbäume aus der XML-Datei des gesamten Vokabulars extrahiert. Der Versuch, im Vokabular als verwandt markierte Begriffe einzubeziehen, war nicht

zielführend, da sich inhaltlich disparate Felder ergaben. Schließlich konnte ein Wortfeld „medizinisches Personal“ mit 127 Begriffen aus der GND generiert werden. Die Suche wurde auf ein Korpus von 195 Bundestagsprotokollen angewendet und erbrachte 5.686 Fundstellen, mindestens ein Begriff des Wortfeldes war in fast jedem der vorher manuell als relevant recherchierten Protokolle vorhanden. Allein 4.508 Treffer entfallen dabei auf den Begriff „Arzt“ oder flektierte Formen davon, zahlreiche Begriffe tauchten im speziellen Korpus nicht auf.

Die Erstellung von Wortfeldern aus bestehenden Ressourcen ist verhältnismäßig unaufwändig. Allerdings muss ein besonderes Augenmerk auf eventuell fehlende, fehlerhafte oder unausgewogene Wörter bzw. Zusammenhänge gelegt werden, um diese Defizite nicht in die generierten Wortfelder zu übernehmen.

Manuell generierte Wortfelder

Eine manuelle *Ad-hoc*-Zusammenstellung von Wörtern als Wortfeld ist für literarische Texte meist nicht geeignet, etwa wenn diese historisches Vokabular enthalten. Um auch Textstellen finden zu können, die zeitgenössische Termini für medizinisches Personal nutzen, haben wir prototypische literarische Texte um 1900 mit dem Annotationstool *CATMA* (Meister et al., 2016) ausgezeichnet. In den sieben annotierten Romanen wurden insgesamt 21 zeitspezifische Bezeichnungen identifiziert, die dem Wortfeld „medizinisches Personal“ zugeordnet wurden.

Ein alternatives manuelles Verfahren ist die Auswertung historischer Lexika, um weitere historische Bezeichnungen zu ermitteln. In anderen Fällen eignen sich Sachregister, etwa aus wissenschaftlichen Publikationen.

Solche manuellen Verfahren bieten sich insbesondere für Texte an, die vom üblichen Sprachgebrauch abweichen. Das *Close Reading* der Texte und die manuelle Annotation relevanter Termini orientieren sich an herkömmlichen geisteswissenschaftlichen Arbeitsweisen. Gleichzeitig können sie Ausgangspunkt für (halb-)automatische Verfahren sein.

Automatische Verfahren

Ausgehend von bereits identifizierten Begriffen haben wir mit *Word Embeddings* gearbeitet. Verfahren der Wahl war *word2vec* (Mikolov et al., 2013) in der Implementation „gensim“ (Řehůřek & Sojka, 2010), das aus einer großen Menge an Sätzen unüberwacht Vektoren vorgegebener Dimension zu jedem vorkommenden Wort erzeugt, sodass in ihrer Bedeutung ähnliche Wörter möglichst ähnliche Vektoren erhalten und umgekehrt. Für die Vektordimension, die Größe des zu berücksichtigenden Kontextes und alle anderen Parameter des Verfahrens verwendeten wir die voreingestellten Standardwerte. Als Trainingsdaten dienten Volltexte von über 2.500 Erzähltexten um 1900, wobei die Aufteilung in Wörter und Sätze nach einer einfachen Heuristik erfolgte. Zur Erstellung von Wortfeldern bestimmten wir anschließend die Kosinus-Ähnlichkeit vorher bekannter Schlüsselwörter wie „Arzt“ oder „Doktor“ zu allen anderen Wörtern im word2vec-Modell. Wir sortierten die Wörter in absteigender Reihenfolge der so bestimmten Vektorähnlichkeit, verwarfen alle mit Ähnlichkeit unter 50 % und erhielten auf diese Weise zu jedem Schlüsselwort je eine Liste im Korpus potenziell semantisch ähnlich verwendeter anderer Wörter.

Die daraus hervorgehenden Listen wurden manuell hinsichtlich der Begriffe durchsucht, die tatsächlich medizinisches Personal bezeichnen. Die gefundenen 131 Begriffe (z. B. Physikus, Wundarzt, Hebmutter) erweiterten das manuell erstellte Wortfeld umfassend. Ohne die halbautomatische Unterstützung wäre es kaum denkbar gewesen, die Vielzahl an Berufsbezeichnungen in den literarischen Texten zu ermitteln.

Zum Einsatz von Wortfeldern

Die vorgestellten Verfahren zur Generierung von Wortfeldern nutzen wir als Findeheuristiken für die Bearbeitung der drei Korpora, die wir für unsere differenten Forschungsfragen untersuchen.

Ausgehend von der Suche mit dem *GermaNet*-Wortfeld wurden Texte in dem Korpus dystopischer Romane mit dem Annotationstool *CATMA*[3] identifiziert, in denen medizinisches Personal besonders häufig genannt wird. In diesen Texten konnten nun wiederum unter Einbezug der Annotationen der Kapitelüberschriften Kapitel identifiziert werden, in denen diese Figuren präsent sind. Erste Stichproben haben gezeigt, dass so Textpassagen gefunden werden, in denen Figuren in medizinischer Hinsicht handelnd auftreten. In literaturwissenschaftlicher Hinsicht hat die wortfeldbasierte Suche damit das Potenzial, bestimmte Motive, Themen und Figurentypen auffindbar zu machen.

Mit dem aus dem GND-Vokabular erstellten Wortfeld haben wir ein Korpus von Bundestags- und Bundesratsprotokollen in *MaxQDA*[4] automatisch annotiert. Dieses Korpus war im Vorfeld auf Grundlage manueller Recherchen als thematisch relevant markiert worden, um Diskurse zu Akzeptanzproblematik von Telemedizin zu analysieren. In dieser Diskursarena spielt medizinisches Personal eine zentrale Rolle. Die Suchergebnisse strukturieren die weiteren Textanalysen, denn sie zeigten relevante Textpassagen zur Rolle des medizinischen Personals sowohl im Diskurs über als auch im Realisieren der Telemedizin auf.

Die manuell zusammengestellten Wortfelder aus literarischen Texten um 1900 sowie einem historischen Lexikon haben wir ebenso wie die automatisch mit *word2vec* generierten und manuell bereinigten Listen auf unser Textkorpus von über 2.500 Prosatexten angewendet. Die Begriffe für „medizinisches Personal“ aus den literarischen Texten wurden dabei 42.569-mal in 1.574 Dokumenten gefunden; die aus dem Lexikon ergaben 16.713 Treffer in 1.418 Dokumenten; die der automatisch generierten Listen 57.968 Treffer in 1.704 Dokumenten. Die identifizierten Textstellen können nun zielgerichtet im Zusammenhang mit der Fragestellung zu Krankheit und Geschlecht um 1900 analysiert werden. Unter anderem lässt sich die von Berufsbezeichnungen häufig implizierte Geschlechtszugehörigkeit betrachten (z. B. „Krankenschwester“, „Hebamme“, „Sanitäter“, „College“ bzw. „Krankenpfleger“ vs. „Krankenpflegerin“).

Fazit: Wortfelder als Findeheuristik

Die vorgestellten Methoden der Wortfeldgenerierung können in manuelle, halbautomatisierte und automatisierte Generierungsstrategien unterteilt werden. Außerdem unterscheiden sie sich hinsichtlich der genutzten Ressourcen: Es gibt Verfahren, in denen die Wörter aus den zu analysierenden Texten – und damit direkt aus dem Forschungsobjekt – stammen, und solche, in denen von konkreten textuellen Kontexten bzw. Forschungsgegenständen unabhängige Wörter genutzt werden.

Durch eine Kombination der unterschiedlichen Methoden können textinhärente und textunabhängige Aspekte – folglich induktive und deduktive Ansätze – berücksichtigt werden, was bessere Ergebnisse ermöglicht. Um die Nutzung von Wortfeldern als Findeheuristiken weiter voranzutreiben, sollte deshalb die geeignete Kombination der Verfahren vor dem Hintergrund der Wortfeldtheorie weiter ausgearbeitet und evaluiert werden.

Fußnoten

1. Für eine Übersicht zur linguistischen Wortfeldtheorie vgl. Vassilyev (1974) und Lehrer (1974, S. 15-45).
2. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html und http://gnd.eurospider.com/s , jeweils zuletzt abgerufen am 27.09.2018.
3. Je nach Größe des Wortfelds können mithilfe der Suchfunktion des Tools auch größere Korpora auf die Begriffe des Wortfelds durchsucht werden. Gefundene Begriffe können annotiert und somit beliebig viele Wortfeldsuchen miteinander kombiniert werden. Beim Annotationsvorgang ist es darüber hinaus möglich, die Ergebnisse der Wortfeldsuche manuell gegebenenfalls auch unter Berücksichtigung des Kontextes zu selektieren.
4. In MaxQDA können einzelne Begriffe gesucht werden.Eine programminterne Lemmatisierung findet statt, ist in ihren Regeln innerhalb des proprietären Programms aber nicht transparent. Für das Wortfeld mussten die 127 Begriffe sowie 88 zugehörige Synonyme der Begriffe eingetragen werden, um ein Wortfeld innerhalb der Programmoberfläche wiederum manuell zu erstellen. Offensichtlich falsche Synonyme wurden dabei aus Gründen der Suchgenauigkeit manuell entfernt, fehlende Synonyme wurden nicht hinzugefügt.

Bibliographie

Bindi, Remo, Calzolari, Nicoletta / Monachini, Monica / Pirelli, Vito / Zampolli, Antonio (1994): *Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition*. In: Literary and Linguistic Computing 9, S. 29–46.

Fellbaum, Christiane (Hg.) (1998): *WordNet: an electronic lexical database*. Cambridge, Mass: MIT Press.

Gaidys, Uta / Gius, Evelyn / Jarchow, Margarete / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Zinsmeister, Heike (2017): *Project Description. HermA: Automated Modelling of Hermeneutic Processes.* In: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Hamp, Birgit / Helmut Feldweg (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, 1997.

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, S. 2228–2235.

Heuser, Ryan / Le-Khac, Long (2011): *Learning to Read Data: Bringing out the Humanistic in the Digital Humanities*. In: Victorian Studies 54, S. 79–86.

Lehrer, Adrienne (1974): *Semantic fields and lexical structure*. Amsterdam: North-Holland Publ. Co. [u.a.].

Meister, Jan Christoph / Petris, Marco / Gius, Evelyn / Jacke, Janina (2016): *CATMA 5.0 [Software für Textannotation und -analyse]*: http://www.catma.de (Zugriff: 24.09.2018).

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): *Efficient Estimation of Word Representations in Vector Space*. ArXiv: https://arxiv.org/abs/1301.3781 (Zugriff: 25.09.2018).

Řehůřek, Radim / Sojka, Petr (2010): *Software Framework for Topic Modelling with Large Corpora*. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, S. 45–50.

Trier, Jost (1973): *Über Wort- und Begriffsfelder*. Darmstadt: Wissenschaftliche Buchgesellschaft. [Zuerst in: Trier, Jost (1931): Der deutsche Wortschatz im Sinnbezirk des Verstandes. Heidelberg, S. 1 - 26 und 310 - 322])

Vassilyev, Leonid M. (1974): *The Theory of Semantic Fields: A Survey*. In: Linguistics 12, S. 79–94.

“Eine digitale Edition kann man nicht sehen” - Gedanken zu Struktur und Persistenz digitaler Editionen

Staecker, Thomas
thomas.staecker@ulb.tu-darmstadt.de
ULB Darmstadt, Deutschland

Einleitung

Digitale Editionen haben nach einer Phase des Ausprobierens und Entwickelns nunmehr eine Reife erreicht, dass sie in vielen Disziplinen nicht mehr als exotischer Sonder-, sondern als Regelfall angesehen werden, was sich in Publikationen wie (Apollon et. al 2014; Driscoll/Pierazzo 2016) oder Förderbedingungen (DFG 2015) spiegelt. Trotz dieser greifbaren Fortschritte stehen digitale Editionen nach

wie vor in der Kritik. Genannt wird immer wieder die fehlende Stabilität und ungelöste Frage der Langzeitarchivierung und - verfügbarkeit. Doch dieses Gefühl des Mangels, so die These des Beitrags, resultiert nicht aus noch nicht geklärten methodischen oder technischen Fragen, sondern beruht auf einer Fehleinschätzung der Natur digitaler Editionen, die in Analogie zum Druckmedium meist nur von ihrer Oberfläche her beurteilt werden. Mit einem Perspektivwechsel, der die Eigentümlichkeiten digitaler Editionen und die zugrundeliegende strukturellen und algorithmischen Komponenten ernst nimmt, ist indes vergleichbare Stabilität möglich, zumindest wenn man sich über das Dokumentenmodell und über die Form seiner technischen Realisierung verständigt.

Markup und Overlap - auf dem Weg zu einem konsolidierten Dokumentenmodell

Die Entwicklung und Nutzung der Markupsprache XML war von Anbeginn an begleitet von Kritik über die Unzulänglichkeit des hierarchischen OHCO Modell (DeRose et al. 1990) für die Repräsentation von Text. Trotz verschiedener Vorschläge konnte bis heute keine abschließende, alle spezifischen Kodierungsprobleme klärende Lösung gefunden werden. Nun hat das der Popularität von XML im Allgemeinen und der in diesem Feld maßgeblichen TEI im Besonderen nicht geschadet. Nach wie vor erfreut sich XML/TEI großer Beliebtheit, auch wenn in jüngerer Zeit der Unterschied von TEI und XML betont wird (Cummings 2017). Das ist umso erstaunlicher, als es an alternativen Ansätzen nicht gemangelt hat (DeRoses 2004; Speerberg-McQueen 2007). Von MECS, GODDAG, TexMECS über LMNL bis zuletzt Text as a Graph (TAGML) entstanden Markup-Konzepte, die für sich in Anspruch nehmen und nahmen, XML und seine Beschränkungen zu überwinden. Gerade mit dem neuesten Konzept des *Text as a Graph* (Kuczera 2016; Dekker/Birnbaum 2017) scheint nach dem Selbstverständnis der Autoren nun endlich der Weg aus der Dauerkrise gewiesen. Doch auch dieses Modell, für das ein erster Serialisierungsentwurf vorliegt (Dekker et al. 2018), wirft erneut Fragen auf. Wenn es auch bestechend scheint und mehr Kodierungsflexibilität verheißt, ist doch fraglich, ob die Graphentheorie tatsächlich das Mittel der Wahl ist, um z.B. einem *autopoietischen* Textbegriff, wie (McGann 2016) ihn postuliert, Herr zu werden. Auch wenn die Graphentheorie als Modell von großem Nutzen sein kann, da sie uns hilft, mehr „clarity of thought about textuality“ (McGann 2016: 90) zu gewinnen, scheint es klug, in der Frage der „Textdefinition“ Vorsicht walten zu lassen und nicht aus dem Auge zu verlieren, dass es bei der Übersetzung des Textes in maschinenlesbare Form nicht nur darum geht, den “intellektuellen” Textbegriff zu modellieren - als “model of” (McCarty 2005) -, sondern auch und vor allem darum, das Textverständnis im Sinne eines “model for” zu erweitern. So vermag man einen “Text” zu konzipieren, von dem McCarty sagt, dass er einen “end maker”, keinen “end user” benötigt und der es erlaubt, so etwas wie eine “digitale Pragmatik” oder “digitale Hermeneutik”, vgl. u.a. (Scheuermann 2016), in die digitale Editorik einzuführen - in Anknüpfung an Konzepte von (Robinson 2003), (Shillingsburg 2006) oder (Gabler 2010).

Was aber von Anfang an bei der Diskussion um Overlap und die Unzulänglichkeiten von XML vernachlässigt wurde, ist, dass die Limitationen des Textmodells nicht unbedingt mit Limitationen der Serialisierung und der digitalen Technik in eins gesetzt werden können. So ist es zwar richtig, dass SGML und in der Folge XML mit dem OHCO Modell im Kopf entwickelt wurden, doch haben bereits (Renear et al.1993) in ihrer Revision darauf abgehoben, dass im pragmatischen Sinne deskriptives Markup auch unabhängig von der OHCO These verwendet werden kann. Eben weil XML vor allem eine Syntax ist, war es letztlich immer wieder möglich, für „konforme“ Lösungen zu sorgen, wie die Vorschläge der TEI zu nicht-hierarchischen Strukturen (TEI Guidelines: Chap. 20) verdeutlichen, aber auch die Beiträge von Renear zum Konzept des “trojanischen Markup” (Renear 2004) und zur XMLisierung von LMNL in CLIX (Renear 2004) oder xLMNL (Piez 2012). Der Grund lag auch darin, dass XML nicht nur in gut etablierte Strukturen der X-Familie eingebettet ist (XSD, XSLT, XQuery etc.), sondern auch allgemeiner die zentrale Document Object Modell (DOM)-Schnittstelle mit allen wichtigen Webelementen wie HTML bzw. XHTML (HTML 5), CSS oder Javascript teilt.

Multimodalität des Dokumentenmodels - nicht nur Markup

Trotz aller Experimente ist heute in der Praxis kaum strittig, dass die TEI und das in ihr entwickelte Dokumentenmodel und mit Abstrichen auch ihre Serialisierung in XML das Mittel der Wahl für digitale Editionen ist. Allerdings bleibt das Modell unvollständig, wenn man nicht auch die anderen Komponenten der Edition in die Überlegungen einbezieht. Wie das analoge Buch über die Rolle zum Kodex und über die Handschrift zum Druck gefunden hat, so muss auch das digitale Buch zu einer stabilen Struktur und Form finden, um nicht nur in der Wissenschaft, sondern auch in Gedächtniseinrichtungen wie Bibliotheken langfristig gesichert, reproduziert und als wissenschaftlich referenzierbares Objekt über Schnittstellen zur Verfügung gestellt und genutzt werden zu können. Dabei sind die Besonderheiten digitaler Dokumente bzw. ihre spezifische Dynamik bzw. Potentialität zu beachten (PDF z.B. erfüllt diese Kriterien nicht). So ist es für das Verständnis der digitalen Edition wichtig, in die Debatte um die Zu- oder Unzulänglichkeit bestimmter Markupsprachen auch das eine Edition verwirklichende Ensemble von Dateien und Funktionen einzubeziehen, deren logisches Zusammenspiel zu bestimmen und nicht nur nach dem, wie die TEI es nennt, „abstract model“ und dessen Serialisierungen zu unterscheiden, sondern auch Regelstrukturen, Präsentationsmodelle und differenzierte Metadatenformate als zum Verständnis notwendige Aspekte zu berücksichtigen. Editionen treten uns typischerweise als eine Kombination von Text mit Markup, Schemadatei, Stylesheets, Transformations- und sonstigen Skripten entgegen. Konkret handelt es sich um eine Reihe von Dateien (oder Datenströmen), wie z.B. .xml,.xslt,.xsd,.css oder .html, die zusammen ein funktionales Ganzes bilden, das als solches nicht nur die Stelle des physischen Dokumentes einnimmt, sondern auch die Grundlage der Langzeitarchivierung

bildet. Dieser ganzheitlich betrachteten digitalen Edition ist eigentümlich, dass sie erst durch eine konkrete, meist nutzergesteuerte algorithmische Verarbeitungsanweisung nach dem klassischen EVA-Prinzip im Viewport oder Empfängersystem „verwirklicht" wird, während ihre Persistenz in den in die Edition hineinkodierten und in ihren Darstellungsfunktionen niedergelegten Möglichkeiten, keineswegs aber in der sichtbaren Oberfläche liegt. Letztere reduziert sich zu einem Ausschnitt, der nur bedingt das gesamte Potential der Edition aufzeigen kann. Wenn diese kombinatorisch vollständig beschreibbaren Möglichkeiten der Präsentation, die zutreffend mit dem Begriff der Schnittstelle verbunden werden (Boot/Zundert 2011; Zundert 2018), den Kernbegriff der digitalen Edition konstituieren, resultieren daraus eine Reihe von praktischen und theoretischen Konsequenzen. Ein erster wichtiger Schritt liegt in der Erkenntnis der Superiorität der Kodierungsgrundlage über die erzeugte angezeigte Oberfläche (Turska/Cummings/Rahtz 2016): "Data is the important Long-term Outcome". Das heißt aber nicht, dass die Oberfläche gleichgültig wäre. Sie darf nur nicht, weil sichtbar, als das einzig wichtige, ja nicht einmal als für die Edition maßgebliche Layer begriffen werden. Die Oberfläche, Visualisierung, die Ausgabe, die Schnittstelle, allgemein das algorithmische Erzeugnis, können über die primäre, immer aber reduzierte und mit Blick auf die kombinatorischen Möglichkeiten ausschnitthafte Darstellungsfunktion hinaus ihrerseits eigenständige Produkte bzw „Interpretationen" sein, vgl. (Zundert 2018). Die Edition selbst sind sie aber nicht, denn eine digitale Edition kann man, streng genommen, nicht sehen. Entsprechend sind zum einen eine Reihe von Text- bzw. Dokumentelemente wie das Layout als eigenständiges bedeutungstragendes oder zumindest -beinflussendes Phänomen als digitale "Textästhethek" und als ein Ergebnis einer Funktion mit vielfältigen Parametern neu zu interpretieren (Stäcker 2019), zum anderen verändern sich Nutzungsszenarien etwa bei der Archivierung und Zitierbarkeit von Editionen, denn wenn die Oberfläche eine von mehreren Möglichkeiten ist, kann sie nicht ohne weitere Vorkehrung Gegenstand des Zitierziels sein. Auf anderer Ebene bedeutet es, dass der/die Autor/Autorin oder, vermutlich genauer, das Autorenteam ein genaues Verständnis auch der technischen Dimension des digitalen Textes haben muss, um seinen nicht nur natürlichen, sondern auch maschinellen „Leser" zu erreichen, oder aber, dass die Autorintention die Schaffung von Möglichkeiten der digitalen Hermeneutik und Analyse einschließen muss. Eine wesentlich Dimension der digitalen Edition ist ferner ihre Verankerung im "Netz". Daraus ergeben sich generell Anforderungen an ihre "Hypertextualität", ihre "Schnittstellen" (Zundert 2018; Stäcker 2019) und Fähigkeit, sich in das "semantic web" zu integrieren (Ciotti/Tomasi 2016). Dazu zählt auch die unmittelbare Integration der genutzten „Forschungsdaten", etwa der digitalen Faksimiles, die im Rahmen der *recensio* gesammelt und gesichtet wurden, so dass auf der Oberfläche ein transkludentes Ensemble entsteht, das auf *hypertextuellen* Strukturen aufbaut.

Es besteht die Hoffnung, dass mit dem Blickwechsel von dem zweidimensionalen sichtbaren Ergebnis auf die unsichtbare Potentialität der Edition sich das eher Proteushafte der Oberfläche der digitalen Edition auflöst und auch für die schon lange gärende Frage nach deren Persistenz und Nachhaltigkeit ein zufriedenstellender Ansatz gerade in ihrer, mit dem Motto der Tagung gesprochen: Multimodalität, gefunden werden kann. Der Beitrag möchte diesen Gedanken anhand von Beispielen weiter ausführen, um einen tragfähigen Begriff von einer persistenten digitale Edition als einem funktionalen und organischen Ensemble von exakt definierbaren Komponenten zu entwickeln.

Bibliographie

Andrews, Tara L. / Zundert, Joris J. van (2018): *What Are You Trying to Say? The Interface as an Integral Element of Argument*, in: **Roman Bleier / Martina Bürgermeister / Helmut W. Klug /Frederike Neuber / Gerlinde Schneider (eds. / hrsg.):** *Digital Scholarly Editions as Interfaces*. Norderstedt 3-33. urn:nbn:de:hbz:38-91064

Apollon, Daniel / Claire Bélisle / Philippe Régnier (eds.) (2014): *Digital critical editions. Topics in the digital humanities*. Urbana: University of Illinois Press.

Boot, Peter/ Zundert, Joris van (2011): *The Digital Edition 2.0 and The Digital Library: Services, not Ressources*, in: Digitale Edition und Forschungsbibliothek. (Beiträge der Fachtagung im Philosophicum der Universität Mainz am 13. und 14. Januar 2011). Harrassowitz 2011: 141–52 (Bibliothek und Wissenschaft, 44) http://hdl.handle.net/20.500.11755/c9e80904-8def-438e-a82b-80d4107b36ed

Ciotti, Fabio / Tomasi, Francesca (2016): *Formal Ontologies, Linked Data, and TEI Semantics*, in: Journal of the Text Encoding Initiative. 10.4000/jtei.1480

Cummings, James (2017): *Slides zum Vortrag gehalten auf der DH2017*: https://slides.com/jamescummings/teimyths . Erscheint in Kürze in DSH.

Dekker, Ronald / David J. Birnbaum (2017): *It's more than just overlap: Text As Graph.* Presented at Balisage: The Markup Conference 2017, Washington, DC, August 1 - 4, 2017, in Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies, vol. 19. https://doi.org/10.4242/BalisageVol19.Dekker01

Dekker, Ronald / Elli Bleeker / Bram Buitendijk, Astrid Kulsdom / David J. Birnbaum (2018): *TAGML: A markup language of many dimensions.* Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018, in Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21. https://doi.org/10.4242/BalisageVol21.HaentjensDekker01

DeRose, Steven J. / Durand, David G. / Mylonas, Elli / Renear, Allen H.: *What Is Text, Really?*, in: Journal of Computing in Higher Education 1, Nr. 2 (Dezember 1990) 3–26. https://doi.org/10.1007/BF02941632 .

DeRose, Steven J. (2004): *Markup overlap: a review and a horse*, in: Proceedings of Extreme Markup Languages. http://xml.coverpages.org/DeRoseEML2004.pdf

Deutsche Forschungsgemeinschaft (2015): *Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft*. Merkblatt der DFG. 2015. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/foerderkriterien_editionen_literaturwissenschaft.pdf

Driscoll, Matthew James/ Elena Pierazzo (eds.) (2016): *Digital scholarly editing: Theories and practices*. Cambridge, UK. https://www.openbookpublishers.com/reader/483 .

Gabler, Hans Walter (2010): *Theorizing the Digital Scholarly Edition*, in: Literature Compass 7, Nr. 2 (Februar 2010) 43–56. https://doi.org/10.1111/j.1741-4113.2009.00675.x .

W3C (2017): *HTML 5.2 Recommendation (14 December 2017):* §9: XML Syntax. https://www.w3.org/TR/html52/xhtml.html#xhtml

Kuczera, Andreas (2016): *Digital Editions beyond XML – Graph-based Digital Editions*, in: Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016). ceur-ws.org/Vol-1632/paper_5.pdf

McCarty, Willard (2005): *Humanities computing*. Basingstoke, Hampshire [u.a.].

McGann, Jerome J. (2014): *A new republic of letters: memory and scholarship in the age of digital reproduction*. Cambridge, Mass. [u.a.]: Harvard Univ. Press.

Piez, Wendell (2012): *Luminescent: parsing LMNL by XSLT upconversion.* Presented at Balisage: The Markup Conference 2012, Montréal, Canada, August 7 - 10, in Proceedings of Balisage: The Markup Conference 2012. Balisage Series on Markup Technologies, vol. 8. https://doi.org/10.4242/BalisageVol8.Piez01

Renear, Allen / Mylonas, Elli / Durand, David (1993): *Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies*. Final version, January 6. http://cds.library.brown.edu/resources/stg/monographs/ohco.html

Robinson, Peter (2003): *WHERE WE ARE WITH ELECTRONIC SCHOLARLY EDITIONS, AND WHERE WE WANT TO BE*, in: Jahrbuch für Computerphilologie. http://computerphilologie.uni-muenchen.de/jg03/robinson.html

Sahle, Patrick (2013): *Digitale Editionsformen : zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. Schriften des Instituts für Dokumentologie und Editorik. Bd. 1-3. Norderstedt.

Scheuermann, Leif (2016): *Die Abgrenzung der digitalen Geisteswissenschaften*, in: Digital Classics Online vol. 2. https://journals.ub.uni-heidelberg.de/index.php/dco/article/viewFile/22746/21865

Shillingsburg, Peter L. (2006): *From Gutenberg to Google: electronic representations of literary texts*, Cambridge.

Speerberg-McQueen, C. M. (2007): *Representation of overlapping structures: Proceedings of the 2007 Extreme Markup Languages conference*. http://conferences.idealliance.org/extreme/html/2007/SperbergMcQueen01/EML2007SperbergMcQueen01.html#id96245

Stäcker, Thomas (2018): *‚Von Alexandria lernen'. Die Forschungsbibliothek als Ort digitaler Philologie*. In Frauen – Bücher – Höfe: Wissen und Sammeln vor 1800 Women – Books – Courts: Knowledge and Collecting before 1800. Essays in honor of Jill Bepler. Hrsg. von Volker Bauer, Elizabeth Harding, Gerhild Scholz Williams und Mara R. Wade. Wiesbaden: Harrassowitz 93–103. http://nbn-resolving.de/urn:nbn:de:tuda-tuprints-75938

Stäcker, Thomas (2019): *Literaturwissenschaft und Bibliothek – eine Beziehung im digitalen Wandel*, in: Digitale Literaturwissenschaft. Metzler/Springer (im Druck, voraussichtl. 2019).

TEI Guidelines: *Chapter 20: Non-hierarchical Structures.* http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html

Turska, Magdalena / Cummings, James / Rahtz, Sebastian (2016): *Challenging the Myth of Presentation in Digital Editions*, in: Journal of the Text Encoding Initiative 9 (2016-2017). http://journals.openedition.org/jtei/1453

Eine Infrastruktur zur Erforschung multimodaler Kommunikation

Uhrig, Peter
peter.uhrig@uos.de
Universität Osnabrück, Deutschland

Eine Infrastruktur zur Erforschung multimodaler Kommunikation

Dieser Vortrag zeigt, wie mit Hilfe des Distributed Little Red Hen Lab eine umfassende Datenbank und Forschungsinfrastruktur geschaffen wurde (und immer noch wird), mit der sich viele Aspekte multimodaler Kommunikation auf Basis der Fernsehaufnahmen des NewsScape-Projekts untersuchen lassen (Steen/Turner 2013).

Datensammlung und Datenbasis

Mit dem UCLA Library Broadcast NewsScape steht Forschern mittlerweile eine Datensammlung von über 400.000 Stunden digitaler Fernsehaufnahmen aus über 10 Jahren zur Verfügung, nicht nur vom US-amerikanischen Fernsehmarkt, aber mit einem starken Fokus auf demselben (vgl. Tabellen 1 und 2). Aufgrund besonderer Einschränkungen des Urheberrechts in den USA darf ein Archiv oder eine Bibliothek „News" aufnehmen und Forschern zur Verfügung stellen. Bei NewsScape wird der Begriff *News* relativ weit ausgelegt, so dass sich auch politische Comedy oder verschiedenste Talkshows in den Aufnahmen finden. Aufgrund gesetzlicher Vorgaben müssen in den USA alle Sendungen mit Untertiteln ausgestrahlt werden, die NewsScape ebenfalls mit aufnimmt, so dass sofort eine relativ brauchbare Verschriftlichung vorliegt, wodurch die Aufnahmen durchsuchbar werden. Dies betrifft auch in den USA ausgestrahlte spanischsprachige Sendungen. Insgesamt nimmt Red Hen Sendungen in mehr als 15 Sprachen auf, unter anderem inzwischen auch in China und Indien, wobei nicht in allen Sprachen bzw. nicht auf allen Sendern Untertitel ausgestrahlt werden. Das Distributed Little Red Hen Lab kooperiert mit dem Open-Source-Projekt CCExtractor um verschiedenste technische Umsetzungen von Untertiteln in Text umwandeln zu können.

Video	
Videodateien	451.974
Laufzeit in Stunden	350.223
Text	
Untertiteldateien	452.208
OCR-Dateien (für Text im Bild)	428.920
TPT-Dateien (heruntergeladende Transkripte)	37.148
Wörter in Untertiteldateien	2,82 Mrd.
Wörter in OCR-Dateien	981,54 Mio.
Wörter in TPT-Dateien	440,38 Mio.
Bilder	
Miniaturbilder	126,08 Mio.

Tabelle 1: Zahlen zur gesamten Sammlung, Mitte November 2017 (übersetzt aus Uhrig 2018)

Sprache	Laufzeit	Wörter	Kommentar
Amerikanisches Englisch	298,004:48:10	2,089,518,746	
Spanisch	15,104:47:23	78,075,367	ca. 60% mexikanisches Spanisch
Französisch	11,425:36:07	8,222,300	verschiedene Varitäten; viele Aufnahmen ohne Untertitel
Internationales Englisch	8,271:55:02	35,646,649	Al Jazeera, France 24, Deutsche Welle, Russia Today, ...
Persisch	5,103:04:54	0	Übertragung ohne Untertitel
Norwegisch	3,241:49:55	7,466,801	Aufnahmen seit 2007, Untertitel seit 2012
Britisches Englisch	2,313:59:54	14,545,895	
Russisch	1,905:47:52	6,511,767	
Deutsch	1,362:15:13	6,381,895	
Schwedisch	1,017:41:15	1,661,240	Aufnahmen seit 2011, Untertitel seit 2015
Portugiesisch	873:31:57	4,897,107	
Dänisch	866:47:26	4,628,942	
Niederländisch/ Flämisch	565:56:41	4,363,813	
Tschechisch	413:47:34	2,956,235	
Polnisch	262:57:42	1,672,483	
Arabisch	148:51:14	0	Übertragung ohne Untertitel

Tabelle 2: Verteilung auf die einzelnen Sprachen (übersetzt aus Uhrig 2018)

Inzwischen zeigt sich jedoch, dass z.T. die Ergebnisse automatischer Spracherkennung im Englischen näher am gesprochenen Wort liegen als die Untertitel, so dass mittelfristig davon auszugehen ist, dass das Vorhandensein von Untertiteln an Relevanz verliert.

Datenverarbeitung

In einem seit 2014 laufenden Projekt wird die Datensammlung auch für die linguistische Forschung aufbereitet, so dass sie nicht nur für traditionell linguistische Fragestellungen sondern auch für multimodale Forschung genutzt werden kann (momentan vorrangig für das Englische).

Textbasierte maschinelle Sprachverarbeitung

Amerikanische Untertitel werden fast ausschließlich in Großbuchstaben ausgestrahlt. Dies stellt eine nicht zu unterschätzende Herausforderung für die weitere Verarbeitung dar, da viele Natural Language Processing (NLP)-Werkzeuge massiv schlechtere Ergebnisse liefern, wenn das Eingabeformat nur aus Großbuchstaben besteht. Das bereits fängt mit der Satzsegmentierung an, für die (gerade im Englischen) die Großschreibung am Satzanfang ein wichtiger Hinweis darauf ist, wo Satzgrenzen zu finden sind. Ein neu entwickelter Satztrenner speziell für die Untertiteldaten und ihre Besonderheiten – Zeilen sind 32 Zeichen lang; Sätze beginnen oft auf einer neuen Zeile – verbesserte die Ergebnisse deutlich, was auch für nachfolgende Verarbeitungsschritte, vor allem das syntaktische Parsing, vorteilhaft ist.

Auch beim Part-of-Speech Tagging zeigte sich, dass reine Großbuchstaben zu schlechten Ergebnissen führen. Das „caseless“ Modell von Stanford CoreNLP (Manning et al. 2014) für das Englische sorgte hier für gute Ergebnisse, die den Ergebnissen für Text mit Groß- und Kleinschreibung kaum nachstehen. Zusätzlich kann man optional ein Modul namens „TrueCase“ nachschalten, das versucht, auf Basis der PoS tags die ursprüngliche Groß- und Kleinschreibung zu erraten.

Für das syntaktische Parsing bietet Stanford ebenfalls ein „caseless“-Modell an, das jedoch auf relativ alter Parsertechnologie aufbaut (klassischer PCFG-Parser mit regelbasiertem Konverter für Dependenzen) und bei unseren Tests auf englischen Daten mit normaler Groß- und Kleinschreibung deutlich schlechter abschneidet das aktuelle Modell (Chen and Manning 2014) namens *dependency neural network* (F-Score labeled attachment: 76,22 vs. 79,56). Es war also notwendig, eine genaue Evaluation durchzuführen, um die bestmögliche Parameterkombination zu ermitteln. Insgesamt wurden 576 Parameterkombinationen evaluiert. Dazu wurde das Korpus ANC MASC mit verschiedenen Parsern und Modellen sowie mit der Originalschreibweise, nur Großschreibung, nur Kleinschreibung und mit den Ergebnissen des TrueCase-Moduls (wo das möglich war) geparst. Es zeigte sich, dass mit TrueCase das Ergebnis des modernen *dependency neural network* Parsers aus Stanford CoreNLP die Ergebnisse relativ nahe an den Ergebnissen mit Originalschreibweise lagen (79,18 vs. 79,56) und damit diese Art von Vorverarbeitung zu deutlich besseren Parsing-Ergebnissen führt als die Verwendung des „caseless“ Modells. Ein Überblick findet sich in Tabelle 3:

	F Original-schreibw.	F bestes caseless	F Klein-schreibung	F Groß-schreibung				
Parser-modell	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled
factored	76.29	80.32	75,90	80,18	72,63	77,68	35,20	48,18
pcfg_cased	76.34	80.16	75,99	79,98	74,77	79,34	29,17	43,68
pcfg_case-less	76.22	80.30	76,20	80,30	76,20	80,30	76,11	80,24
shift-reduce	76.05	79.82	75,74	79,54	72,20	77,16	42,77	54,46
shift-reduce with beam search	76.80	80.90	78,05	81,86	72,08	77,10	42,93	54,92
relational neural network	78.24	82.20	77,76	81,80	76,87	81,34	24,94	40,32
dependen-cy neural network	79.56	83.06	79,18	82,80	77,70	81,68	40,70	51,96

Tabelle 3: Vergleich der Parsermodelle bei unterschiedlicher Groß- und Kleinschreibung

Audio

Zur Vorbereitung der Audioverarbeitung mussten darüber hinaus die Untertitel von allem Text befreit werden, der nicht gesprochen wird. Wesentliche Punkte sind dabei

Sprecherangaben („Reporter:“) und Angaben über den nicht-gesprochenen Ton („[Doorbell rings.]“ oder „[Applause]“). Mittels einer Frequenzliste wurde ein Filter erstellt, der ca. 95 % des nicht-gesprochenen Texts entfernt.

Im nächsten Schritt wurde mit Forced Alignment Software (in diesem Fall *Gentle*) versucht, für jedes Wort in den Untertiteln die genaue Position in der Audiospur des Videos zu ermitteln. Die Software selbst gibt an, etwas über 91 % der Wörter zu alignieren, aber Stichproben zeigen, dass auch bei den alignierten noch falsche Ergebnisse auftreten. Die genaue Größenordnung des Fehlers muss noch ermittelt werden.

Video

Schließlich wurde mittels Computer-Vision-Software die visuelle Ebene auf verschiedene Merkmale hin annotiert, die für die Erforschung multimodaler Kommunikation relevant sind. Der Computer versucht hier, automatisch zu erkennen, ob eine Person auf dem Bild zu sehen ist, ob diese der Sprecher bzw. die Sprecherin ist, ob die Person ihre Hände bewegt und ob bestimmte high-level-Gesten (in diesem Fall als Test sogenannte „timeline gestures“) zu sehen sind (Turchyn et al. 2018). Erste Tests zeigen, dass das System auf OpenCV-Basis eine sehr gute Präzision jenseits der 90% für die Personenerkennung schafft, aber leider bei den Handbewegungen nur eine Präzision im Bereich von ca. 33 % erreicht. Es ist also im Moment immer ein nachgeschalteter manueller Analyseschritt nötig. Aktuell laufen Experimente, die Erkennung mittels OpenPose zu verbessern.

Abfragemöglichkeiten

Alle Daten wurden in *CQPweb* (Hardie 2012), einer korpuslinguistischen Abfrageplattform mit großem Funktionsumfang, gespeichert und können so effizient und komfortabel abgefragt werden. Es wird gezeigt, wie mit einer Abfrage sowohl linguistische als visuelle Parameter abgefragt werden können, so dass man sofort die jeweils passenden Stellen im Video angezeigt bekommt.

Weiterhin werden Abfragemöglichkeiten über ein Geoinformationssystem (GIS) in Verbindung mit linguistischer Analyse (z.B. für die kulturgeographische Forschung) sowie die Suchmaschinen der UCLA demonstriert.

Um die oben erwähnte manuelle Analyse zu beschleunigen, wurde im Rahmen des *Google Summer of Code* 2018 die Version 2 des *Red Hen Rapid Annotator* entwickelt, mit dem komfortabel und schnell große Mengen an Videoschnipseln klassifiziert werden können. Im Vortrag wird dieser ebenfalls kurz demonstriert.

Anwendungen

Abschließend wird ein kurzer Überblick über laufende und abgeschlossene Projekte mit der Infrastruktur gegeben, um zu zeigen, welch vielfältige Fragestellungen bereits heute bearbeitet werden können.

Bibliographie

Chen, Danqi / Manning, Christopher D. (2014): *A Fast and Accurate Dependency Parser using Neural Networks*, in: **Moschitti, Alessandro / Pang, Bo / Daelemans, Walter (eds.):** *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Doha: Association for Computational Linguistics, 740-750. http://aclweb.org/anthology/D14-1082 [Letzter Zugriff 15. Januar 2018]

Hardie, Andrew (2012): *CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool*, in: International Journal of Corpus Linguistics, 17.3, 380-409

Manning, Christopher D. / Surdeanu, Mihai / Bauer, John / Finkel, Jenny Rose / Bethard, Steven / McClosky, David (2014): *The Stanford CoreNLP Natural Language Processing Toolkit*, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14). Baltimore, MD: Association for Computational Linguistics, 55-60. http://aclweb.org/anthology/P14-5010.pdf [Letzter Zugriff 15. Januar 2018]

Steen, Francis F. / Turner, Mark (2013): *Multimodal Construction Grammar*, in: **Borkent, Michael / Dancygier, Barbara / Hinnell, Jennifer (eds.):** *Language and the Creative Mind*. Stanford, CA: CSLI Publications, 255-274

Turchyn, Sergiy / Olza Moreno, Inés / Pagán Cánovas, Cristóbal / Steen, Francis / Turner, Mark / Valenzuela, Javier / Ray, Soumya (2018): *Gesture Annotation with a Visual Search Engine for Multimodal Communication Research*, in: IAAI-18, article 72.

Uhrig, Peter (2018): *NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts*, in: **Anne-Julia Zwierlein / Jochen Petzold / Katharina Böhm / Martin Decker (eds.):** *Anglistentag 2018 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English*. Trier: Wissenschaftlicher Verlag Trier.

Ein neues Format für die Digital Humanities: Shared Tasks. Zur Annotation narrativer Ebenen

Willand, Marcus
marcus.willand@gs.uni-heidelberg.de
Universität Heidelberg; Universität Stuttgart

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

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart

Einleitung

Dieses Paper führt unsere letztjährige Präsentation[1] des Vorhabens fort, den ersten **Shared Task (ST) zur Annotation literarischer Phänomene** zu organisieren und solch ein kompetitives Verfahren als fruchtbares Format für die *Digital Humanities* einzuführen. Dieser ST hat mit dem abgehaltenen Workshop der teilnehmenden Teams einen Meilenstein erreicht.

Bei einem ST bewerben sich Teams mit einem Vorschlag für die Lösung eines durch die Organisatoren ausgeschriebenen Problems, den Task. STs sind kompetitive Verfahren, weil die Lösungsvorschläge vergleichend evaluiert und gemäß einer definierten Metrik in eine Rangfolge gebracht werden. Vor allem in der Sprachverarbeitung (NLP, natural language processing) sind diese Arbeitszusammenhänge weit verbreitet und ein wesentlicher Antrieb für die Fortschritte bei wichtigen Aufgaben, etwa des syntaktischen Parsings.[2] Wir haben dieses kompetitive Verfahren für literaturwissenschaftliche Problemstellungen durch kooperative Aspekte modifiziert und möchten hier sowohl den angepassten Workflow als auch zentrale Einsichten vorstellen, die wir durch den o.g. Workshop generieren konnten.[3] Wir gehen davon aus, dass durch solch adaptierte STs sehr viele andere Problemstellungen der Geisteswissenschaften adressiert werden können, wodurch sich STs als Verfahren für die Digital Humanities natürlicherweise anbieten. Dies ist insbesondere der Fall, wenn computationelle Verfahren auf geisteswissenschaftliche Konzepte treffen und diese in einem intersubjektiven Aushandlungsprozess operationalisiert werden sollen.

Wir haben uns für ein zweiphasiges Verfahren entschieden. Die erste Phase – „SANTA" genannt: Systematic Analysis of Narrative Texts through Annotation – widmet sich der Erstellung von Annotationsrichtlinien für das Phänomen narrativer Ebenen.[4] Die von den acht teilnehmenden Teams eingereichten und auf dem Workshop diskutierten Richtlinien bilden die Grundlage für den Task der geplanten zweiten Phase: die automatisierte Identifikation von Erzählebenen auf Basis von Daten, die nach den Richtlinien annotiert wurden (wird vsl. 2019 ausgeschrieben).[5]

Die acht Teams divergieren hinsichtlich ihrer:

- **Größe**: 1-4 Mitglieder, wobei drei Teams die Richtlinien im Seminarkontext, also mit einer Vielzahl an Studierenden entwickelten (was eine von uns vorgeschlagene Option war)
- **Disziplin**: Literaturwissenschaft, Informatik, Linguistik, Computerlinguistik, Mediävistik, *Digital Humanities*
- **Forschungsziele**: Narratologische Konzeptentwicklung, bzw. -reflexion, Anwendung narratologischer Konzepte für die Einzeltextinterpretation (bzw. ausschließlich für die Texte im von uns vorgegebenen Korpus), linguistische Diskursanalyse, Automatisierung der Annotation
- **Nation**: Deutschland, USA, Schweden, Irland, Kanada
- **Narratologie**: Überwiegend Genette und/oder Ryan, teilweise linguistische oder selbstentworfene Level-Definitionen

Der Workshop selbst war konzeptionell offen angelegt und sollte den Teilnehmer/innen wie auch uns Organisator/innen ermöglichen, den geplanten Ablauf in Reaktion auf die Arbeitsergebnisse zu verändern. Dies war realisierbar, da bis auf wenige Kurzvorträge (z.B. stellte jedes Team zu Beginn in 5 Minuten die zentralen Aspekte seiner Richtlinien vor) hauptsächlich in kooperativen Formaten wie Gruppenarbeiten, Feedback-Runden und Plenumsdiskussionen gearbeitet wurde.

Guidelines: Unterschiede und Gemeinsamkeiten

Am **ersten Workshop-Tag** sollten die Teilnehmer/innen einen differenzierten Einblick in alle acht Annotationsrichtlinien und deren Spezifika bekommen. Dabei wurden **Unterschiede** auf mehreren Abstraktionsebenen identifiziert: Die erste und grundlegendste Einsicht bestand in der Beobachtung, dass die Definitionen narrativer Level mitunter stark differieren und diese Unterschiede durch die divergierenden Forschungsfragen (siehe oben) erklärt werden können: etwa, ob die Level-Annotation im Dienste narratologischer Konzeptentwicklung eingesetzt wird oder zur Erkennung linguistischer Diskursebenen in Texten. Das spezifische Level-Verständnis hat gleichsam Auswirkungen auf den Modus des Definierens. So werden narrative Level teilweise inhaltlich bestimmt über die Elemente der „Story" oder aber – etwas abstrakter – über die Elemente der Erzählung der Story. Zu unterscheiden sind davon Ansätze, die narrative Level über ihre Grenzen bestimmen, teilweise ohne auf die Erzählinhalte zurückzugreifen. Dies ist der Fall, wenn Erzähler- oder Weltwechsel als Definiens eingesetzt werden. Deutlich wurde zudem, dass gerade literaturwissenschaftliche Ansätze nicht immer eindeutig zwischen *identifizierenden* und (bloß) *charakterisierenden* Texteigenschaften narrativer Ebenen – wie etwa Fokalisierung – unterscheiden, wobei die Frage aufkommt, ob letztere einen berechtigten Ort in den Guidelines haben. **Gemeinsamkeiten** der Guidelines wurden ebenfalls diskutiert. Dabei zeigte sich, dass die eingereichten Guidelines zwei recht homogene Gruppen bilden hinsichtlich *Forschungsziel* (Narratologie vs. Automatisierung) und *Konzeptverständnis* (komplex vs. vereinfachend).

Evaluation der Guidelines: Drei Bewertungsdimensionen

Der **zweite Workshop-Tag** widmete sich der Evaluation der Guidelines: Das Ziel der Organisator/innen war es, mit den Teilnehmer/innen gemeinsam die ‚besten' Richtlinien zu finden, wobei zunächst unklar blieb, ob es *einen* oder *mehrere* Gewinner geben sollte (etwa jeweils einen aus den beiden o.g. recht homogenen Gruppen). Die von den Organisator/innen im Vorfeld ausgearbeiteten Evaluationskriterien zur Beurteilung der Stärken und Schwächen der Einreichungen folgen der Idee, den interdisziplinären Aushandlungsprozess anhand von drei Dimensionen zu strukturieren und so neben in der Computerlinguistik etablierten Kriterien weitere relevante geisteswissenschaftliche Aspekte in die Bewertung zu integrieren.

Diese Kriterien wurden zuerst während des Workshops im Plenum reflektiert und anschließend per online-Fragebogen live in die Evaluation überführt. Jede Dimension sollte auf nachvollziehbare Weise potentielle Guideline-

Stärken vergleichbar machen und so für eine ausgewogene Beurteilung durch die Workshopteilnehmer/innen sorgen. Die drei Dimensionen sind – um in der Tagungssprache zu bleiben – *Conceptual Coverage*, *Applicability* und *Usefulness*.

Die erste Dimension beurteilt anhand von vier Fragen die Qualität der Guidelines hinsichtlich ihrer **Abdeckung der zugrundeliegenden (meist narratologischen) Theorie**:

1. Is the narrative level concept explicitly described?
2. Is the narrative level concept based on existing concepts?
3. How comprehensive are the guidelines with respect to aspects of the theory?
4. How adequate is the narrative level concept implemented by this guidelines in respect to narrative levels?

Die zweite Dimension evaluiert die **Anwendbarkeit der Guidelines auf den Text** anhand von zwei Fragen und eines von uns im Vorhinein gemessenen *Inter-annotator agreements*:[6]

1. How easy is it to apply the guidelines for researchers *with* a narratological background?
2. How easy is it to apply the guidelines for researchers *without* a narratological background?

Die dritte Dimension bewertet anhand von vier Fragen, wie der auf Basis einer Richtlinie annotierte Text das T **extverstehen und die weitergehende Textarbeit** befördert:

1. Thought experiment: Assuming that the narrative levels defined in the annotation guidelines can be detected automatically on a huge corpus. How helpful are these narrative levels for an interesting corpus analysis?
2. How helpful are they as an input layer for subsequent corpus or single text analysis steps (that depend on narrative levels)?
3. Do you gain new insights about narrative levels in texts by applying the foreign guidelines, compared to the application of your own guidelines?
4. Does the application of these guidelines influence your interpretation of a text?

Jede der Fragen wurde von den Teams in einer Feedback-Runde erläutert und anhand einer vierstufigen Likert-Skala online beurteilt.

Die dergestalt relativ differenziert abgefragten drei Evaluationsdimensionen lassen sich – stark abstrahiert – auch verstehen als prozedurale Vergewisserungskriterien für gute Guidelines zur Annotation literaturwissenschaftlicher (oder allgemein: geisteswissenschaftlicher) Konzepte auf Texten unterschiedlicher Genres. Sie können prozessorientiert abgebildet werden:

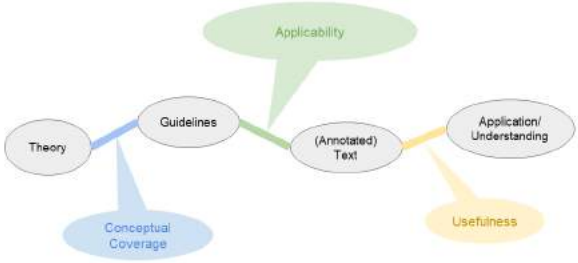


Abb. 1: Prozedurale Darstellung der Evaluationsdimensionen

Evaluation der Evaluation: Ergebnisse des Workshops

Am **dritten Workshop-Tag** wurden die Ergebnisse der Evaluation vorgestellt und diskutiert. Als methodisch ausgesprochen interessantes Resultat zeigte sich, dass die qualitative Plenumsdiskussion der Guidelines zu Einschätzungen führte, die durch die Resultate der quantitativen Evaluation in den Fragebögen abgebildet wurden: Die als theoretisch hochdifferenziert gelobten Guidelines waren just diejenigen, die in der ersten Dimension die meisten Punkte erzielten usw. Eine vierstufige Skala scheint also den gesamten Evaluationsbereich mit den drei komplexen Theorie-, Anwendungs- und Brauchbarkeitsfragen ausreichend differenziert abbilden zu können. Problematisch erschien den Teilnehmer/innen allerdings, dass die Fragen zunehmend schwerer zu beantworten waren. Dies resultierte aus der Schwierigkeit, für einige Fragen potentielle Anwendungsfälle zu antizipieren, in denen bereits annotierte Texte sinnvolle Forschungsfragen ermöglichen. Allerdings wurden im Gegensatz zur subjektiven Wahrnehmung der Teilnehmer/innen diese Fragen der letzten Dimension mit einer zunehmend geringeren Standardabweichung beantwortet. Trotz *gefühlt* größerer Schwierigkeiten mit den Fragen zur Usefulness wurden die Guidelines dort einvernehmlicher evaluiert.

Die drei Dimensionen erwiesen sich damit als praktikables Instrument einer differenzierten Bewertung der Guidelines. Das Experiment "Shared Task für die DH" ist also geglückt. Die drei Bewertungsdimensionen stehen allerdings auch weiterhin für eine der großen methodischen Herausforderungen im *Digital Humanities*-Bereich: die Evaluation von Operationalisierung, Analyse und Interpretation in interdisziplinären Kontexten.

Fußnoten

1. Siehe Gius et al. (2018), aber auch Reiter et al. (2017), bzw. zur Projekt-Dokumentation Gius et al. (2016ff.).
2. Bspw. Daniel Zeman et al. 2017 dokumentieren, wie ein typischer NLP Task funktioniert.
3. Wir danken der VolkswagenStiftung für die Finanzierung dieses Workshops, der vom 17.-19. Sept. 2018 an der Universität Hamburg stattgefunden hat. CRETA (Stuttgart)

danken wir für die Finanzierung der notwendigen Annotationsarbeiten durch unsere HiWis Linda Kessler, Tanja Preuß, Nina Stark, Hanna Winter. Katharina Krüger hat uns bei der Organisation in Hamburg unterstützt und Carla Sökefeld den Workshop protokolliert.

4. Ausführlichere Informationen und Literaturhinweise zu einführender (etwa Jahn 2017), grundlegender (etwa Ryan 1991 und Genette 1980: 227-237) und vertiefender (etwa Mani 2013) narratologischer Literatur finden sich auf der Projekthomepage https://sharedtasksinthedh.github.io/levels

5. Die ausführliche Dokumentation der Abläufe des STs, die Publikation der Guidelines inkl. Reviews wird in zwei Sonderheften der *Cultural Analytics* publiziert (vsl. Q4/2018 und Q3/2019).

6. Alle Guidelines wurden 1) durch die jeweiligen Autor/innen selbst, 2) durch ein zufällig ausgewähltes anderes teilnehmendes Team und 3) durch von uns eingesetzte Hilfskräfte annotiert. Grundlage der Annotation waren acht literarische Prosatexte unterschiedlichen Umfangs, die zwischen 1797 und 1931 publiziert wurden. Als Metrik für das Agreement wurde Gamma verwendet (Mathet et al., 2015).

Bibliographie

Genette, Gérard (1972): *Narrative Discourse. An Essay in Method*. Ithaca 1980. (Franz. Figures III. Paris 1972).

Gius, Evelyn, / Nils Reiter / Marcus Willand (2016ff.): *Shared Tasks in the Digital Humanities. Systematic Analysis of Narrative Texts through Annotation*, Projektwebsite und -dokumentation https://sharedtasksinthedh.github.io/ [letzter Zugriff 28 September 2018].

Gius, Evelyn, / Reiter, Nils / Strötgen, Jannik / Willand, Marcus (2018): *SANTA: Systematische Analyse Narrativer Texte durch Annotation*. DHd2018, Köln.

Mani, Inderjeet (2013): *Computational Narratology*. Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert (Hrsg.). The living handbook of narratology. Hamburg University Press http://www.lhn.uni-hamburg.de/article/computational-narratology [letzter Zugriff 28 September 2018].

Pier, John (2014): *Narrative Levels* (revised version; uploaded 23 April 2014). Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert (Hrsg.). The living handbook of narratology. Hamburg University Press ttp://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014 [letzter Zugriff 28 September 2018].

Reiter, Nils / Gius, Evelyn, / Strötgen, Jannik / Willand, Marcus (2017): *A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts*. DH2017, Montreal.

Ryan, Marie-Laure (1991): *Possible Worlds, Artificial Intelligence, and Narrative Theory.* Bloomington: Indiana University Press.

Mathet, Yann / Widlöcher, Antoine / Métivier, Jean-Philippe (2015): *The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment*. Computational Linguistics, 41(3):437–479.

Zeman, Daniel et al. (2017): *Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL http://www.aclweb.org/anthology/K17-3001 [letzter Zugriff 28 September 2018].

Ein unscharfer Suchalgorithmus für Transkriptionen von arabischen Ortsnamen

Scherl, Magdalena
magdalenascherl@gmail.com
Hochschule Mainz, Deutschland

Unold, Martin
martin.unold@gmail.com
Hochschule Mainz, Deutschland

Homburg, Timo
timo.homburg@hs-mainz.de
Hochschule Mainz, Deutschland

Einleitung

Motivation

Digitale Ortsverzeichnisse (Gazetteers) beinhalten Informationen über Orte sowie deren geographische Lage. Eine der grundlegendsten Aufgaben im Umgang mit solchen Ortsverzeichnissen ist die Suche nach Ortsnamen. Diese Suche kann sehr schwierig sein für Ortsnamen, die in verschiedenen Transliterations- oder Transkriptionsvarianten vorliegen, wie es oft bei arabischen Ortsnamen der Fall ist. In diesen Fällen reicht eine reine Volltextsuche nicht aus. Hier können unscharfe String-Matching-Algorithmen eine bessere Trefferquote für Suchen erreichen.

Zielsetzung

Unser Ziel war es, einen Suchalgorithmus zu entwickeln, der in der Lage ist, arabische Ortsnamen in verschiedenen Transliterationen und Transkriptionen zu identifizieren. Einerseits sollte der Algorithmus fehlertolerant sein, sodass er einen Suchbegriff findet, selbst wenn er etwas anders geschrieben wurde als im Ortsverzeichnis hinterlegt. Andererseits sollte er genau genug sein, um nur tatsächliche Transliterations- und Transkriptionsvarianten einzuschließen. Zum Beispiel sollte die Suche nach "Agaga" den Ort "Ajaja" finden, da es sich um verschiedene Transliterationen des selben arabischen Wortes handelt, aber nicht "Dagaga", da dies ein ganz anderer Ort ist. Um diese beiden Ziele zu erreichen, haben wir einen Algorithmus mit einer modifizierten gewichteten Levenshtein-Distanz (Levenshtein 1965) entwickelt. Eine weitere Eigenschaft unseres Suchalgorithmus ist, dass er für andere Anwendungsfälle als arabische Schrift leicht angepasst werden kann. Wir haben daher auch eine

Version für Keilschriftsprachen implementiert und auf einem sumerischen Wörterbuch getestet.

Forschungsstand

Die gewichtete Levenshtein Distanz wurde bereits für Autokorrektur (Kukich 1992), für die Korrektur von Fehlern bei der Optical Character Recognition (OCR) (Lasko 2001, Mihov 2002) und für die automatische Spracherkennung (Ziolko 2010, Zgank 2012) genutzt. Um die Kosten für die Editieroperationen zu bestimmen, schlägt Weigel (1995) einen iterativen überwachten Lernalgorithmus vor. Lasko (2001) beschreibt die Verwendung einer probabilistischen Substitutionsmatrix und Schulz / Mihov (2002) schlagen die Implementierung eines endlichen Zustandsautomaten vor, um die Performanz des Levenshtein-Algorithmus zu verbessern.

Arabische Schrift

Variationen in der Schreibweise von arabischen Toponymen sind sehr häufig, da es mehrere Transliterationsstandards und verschiedene gebräuchliche Transkriptionsschemata gibt (Brockelmann 1953, Schlott-Kotschote 2004, UNGEGN 2016, Pedersen 2008). Insbesondere die Darstellung jener arabischen Buchstaben, die im lateinischen Alphabet keine direkte Entsprechung haben, variiert hier teilweise beträchtlich. Während einige Standards hierfür diakritische Zeichen verwenden, setzen andere Standards auf die Verwendung von Kombinationen aus zwei Buchstaben. Eine andere Quelle der Variation ist die fehlende Vokalisierung in der arabischen Schrift. Besonders regionale Variationen der Aussprache und Dialektdiversität führen dazu, dass arabische Vokale in der lateinischen Schreibweise unterschiedlich wiedergegeben werden. Zu Abweichungen führen auch unterschiedliche Traditionen der Transkription, die sich entweder eher an der englischen oder an der französischen Aussprache orientieren. Ein weiteres Problem, das zu Variationen führen kann, sind Wortgrenzen und divergierende Ansätze in der Zusammen- und Getrenntschreibung, insbesondere bei der Verwendung des Artikels "al".

Keilschriftsprachen

Die Entwicklung von Software für die Verbesserung der Bearbeitbarkeit von Keilschriftsprachen traf in der Vergangenheit auf ein reges Interesse in der Digital Humanities Community.

Homburg (2016, 2017, 2018) zeigten, dass Fortschritte in der Erstellung einer Natural Language Processing Pipeline und in der Erstellung von State-Of-The-Art semantischen Wörterbüchern für verschiedene Keilschriftsprachen in Entwicklung sind. Homburg (2015) entwickelte eine auf einem Präfixbaum De La Briandais (1959) basierende Eingabemethode für Keilschriftsprachen, die auf der DHd 2015 präsentiert wurde. State Of The Art Eingabemethoden wie Sogou Pinyin[1] für Chinesisch oder Google Japanese Input[2] (Krueger 2000) für Japanisch beinhalten jedoch prädiktive Algorithmen, welche es erlauben die Korrektheit von Texteingaben in ihrem jeweiligen Kontext einzubeziehen und mit Fuzzy Search Algorithmen ebenfalls eine Korrektur von Tippfehlern vorzunehmen. Für die Eingabe von Keilschrift wurden solche Algorithmen bisher noch nicht erprobt, obwohl diese die Eingabe auch durch Einblendung von Zusatzinformationen ernorm vereinfachen kann und mehr relevante Suchergebnisse angezeigt werden können. Für Keilschriftsprachen im Speziellen ist eine Fuzzy Search für die Unterscheidung gerade auch der verschiedenen Dialekte und Transliterationen der Keilschriftarten interessant, da in diesen unter anderem Vokalverschiebungen und Variationen durch verschiedene Transliterationskonventionen auftreten können. Beispiele hierfür sind die Unterscheidungen von diakritischen Zeichen vs. einer numerischen Annotation (ù vs. u2), Transliterationsunterschiede wie die Verwendung von sh vs. sz und sprachliche Entwicklungen über die Zeit hinweg, in denen z.B. endende Konsonanten weggefallen sind (sogenannte Mimation).

Ansatz

Wir verwendeten ein modifiziertes Levenshtein Distanz Maß, welches speziell für die arabische Schrift angepasst wurde. Der Quellcode des Projektes ist unter der GPLv2 Lizenz in unserem Gitlab freigegeben worden.[3] Die Kosten für die Editieroperationen wurden hierbei durch ein überwachtes Lernverfahren ermittelt. Wir verwendeten eine Substitutionsmatrix sowie eine Matrix für Lösch- sowie Einfügeoperationen, um die jeweiligen Kosten der Überführung von einer Transliteration in die nächste zu bestimmen.

	b	d	e	i	r	u	ī	ay
b	0	1	0,97	0,91	1	0,86	0,86	2
d	1	0	0,91	0,98	1	1	0,81	2
e	0,97	0,91	0	0,48	0,89	0,63	0,28	0,19
i	0,91	0,98	0,48	0	1	0,92	0,09	0,56
r	1	1	0,89	1	0	0,69	0,75	2
u	0,86	1	0,63	0,92	0,69	0	1	0,49
ī	0,86	0,81	0,28	0,09	0,75	1	0	0,45
ay	2	2	0,19	0,56	2	0,49	0,45	0

Abbildung 1: Beispielwerte aus der Substitutionsmatrix.

Für die Matrix für Löschungen und Einfügungen haben wir zwei unterschiedlichen Ansätze verfolgt: Im ersten Ansatz (Levenshtein1) wurden die Lösch- sowie Einfügekosten für jeden Buchstaben ohne Betrachtung des Buchstabenkontexts ermittelt. Im zweiten Ansatz

(Levenshtein2) wurden die Lösch- und Einfügekosten in Abhängigkeit des voranstehenden Buchstabens ermittelt.

	b	d	e	i	r	u	ī	ay
b	1	1	0,79	0,98	0,37	1	1	1
d	0,83	1	0,93	0,85	0,7	0,81	1	1
e	0,62	0,55	1	1	0,64	0,53	1	1
i	0,74	0,36	1	0,64	0,56	1	1	1
r	0,91	1	0,74	0,63	1	0,65	1	1
u	0,9	1	1	1	0,6	1	1	1
ī	0,3	0,19	1	1	0,31	1	1	1
ay	2	2	2	2	2	2	2	2

Abbildung 2: Beispielwerte aus der Matrix für Löschungen und Einfügungen (Levenshtein2). Reihen repräsentieren den zu löschenden bzw. den einzufügenden Buchstaben; Spalten repräsentieren den voranstehenden Buchstaben.

Desweiteren wurden spezielle Anpassungen für die arabische Schrift wie z.B. diakritische Zeichen (ī), welche typisch für die gegebenen Transliterationen sind, in das erweiterte Alphabet aufgenommen. Außerdem waren Kombinationen aus zwei Buchstaben zu berücksichtigen, die ein arabisches Phonem repräsentieren (z.B. sh). Da die klassische Levenshtein Distanz nicht aus Buchstabenkombinationen errechnet werden kann, musste der Algorithmus auf diese angepasst werden. In einer vereinfachten Version (Levenshtein1Simple und Levenshtein2Simple) wurden die Buchstabenkombinationen im Vorhinein durch einen Index ersetzt, sodass eine klassische Berechnung über den originären Levenshtein Algorithmus erfolgen konnte. Dieser vereinfachte Ansatz wies eine deutlich höhere Performanz auf.

		b	d	e	r	i
	0	0,99	1,83	2,38	3,11	3,67
b	0,99	0	0,83	1,38	2,12	2,68
u	1,89	0,9	1	1,47	2,08	2,64
d	2,7	1,71	0,9	1,45	2,18	2,74
a	3,25	2,25	1,45	1,38	2,11	2,67
y	4,25	3,25	2,45	1,09	1,82	2,38
r	5,25	4,25	3,45	2,09	1,09	1,65
ī	5,56	4,57	3,76	2,4	1,4	1,18

Abbildung 3: Berechnung der Levenshtein Distanz für ein Beispiel Wortpaar mit Levenshtein2. Buchstabenkombinationen werden durch einen modifizierten Algorithmus berücksichtigt.

		b	d	e	r	i
	0	0,99	1,83	2,38	3,11	3,67
b	0,99	0	0,83	1,38	2,12	2,68
u	1,89	0,9	1	1,47	2,08	2,64
d	2,7	1,71	0,9	1,45	2,18	2,74
ay	4,7	3,71	2,9	1,09	1,82	2,38
r	5,7	4,71	3,9	2,09	1,09	1,65
ī	6,01	5,02	4,21	2,4	1,4	1,18

Abbildung 4: Berechnung der Levenshtein Distanz für ein Beispiel Wortpaar mit Levenshtein2Simple.

Buchstabenkombinationen werden vorab auf einen Index gematcht.

Experimente und Ergebnisse

Die arabische Version des Suchalgorithmus wurde auf zwei Wörterbüchern getestet. Das erste Wörterbuch beinhaltete Toponyme von archäologischen Fundorten in Syrien, im Irak und in der Türkei, welche aus dem TEXTELSEM Repositorium des i3Mainz stammten (tts_arch)[4]. Das zweite Wörterbuch beinhaltete syrische Toponyme aus GeoNames (geo_SY)[5]. Zusätzlich wurde die Übertragbarkeit des Suchalgorithmus auf andere Sprachen auf einem sumerischen Wörterbuch getestet, das aus dem "Semantic Dictionary for Ancient Languages" extrahiert wurde (sum)[6]. Alle Wörterbücher wurden in ein Trainings- sowie ein Testkorpus aufgeteilt. Gemessen wurde die Mean Average Precision (MAP) bei einer Rückgabe der Ergebnisse in Form eines Rankings. Da die durchgeführten Tests so konzipiert waren, dass jeweils nur ein Ergebnis als zutreffend gewertet wurde, genügte für jedes Suchwort die Berechnung eines Präzisionswertes, der anschließend über alle Testsuchwörter gemittelt wurde. Die Ergebnisse unserer Tests sind in Tabelle 1 festgehalten. Sie zeigen, dass unser Algorithmus in der Lage war, Toponyme mit einer Präzision zwischen 90% und 95% abhängig vom Wörterbuch zu finden. Verglichen mit einem ungewichteten Levenshtein Distanzmaß kann unser Ansatz somit eine Verbesserung der Präzision zwischen 9 Prozentpunkten auf dem sumerischen Wörterbuch und 27 Prozentpunkten auf dem TEXTELSEM Wörterbuch erreichen.

Datenset	MAP bei Volltextsuche	MAP bei ungewichteter Levenshtein Distanz	MAP bei eigenem Algorithmus (beste Version)	Algorithmus
tts_arch	0.24	0.63	0.90	Levenshtein2Simple
geoSY_xs	0.01	0.81	0.95	Levenshtein1
Sum	0.01	0.83	0.92	Levenshtein2Simple

Tabelle 1: Testergebnisse. Die Tests zeigen, dass die Levenshtein2Simple Version des Algorithmus im allgemeinen Fall eine bessere Präzision sowie die beste Performanz aufweisen konnte.

Zusammenfassung

Unsere Version der gewichteten Levenshtein Distanz erwies sich als ein vielversprechender Ansatz für die Verbesserung von Suchergebnissen in digitalen Gazetteeren. Zusätzlich konnten wir durch die Anwendung des Algorithmus auf das sumerische Keilschriftwörterbuch die Übertragbarkeit des Algorithmus auf andere Sprachen demonstrieren. Obwohl die vorgeschlagene Adaption des Levenshtein Algorithmus für sumerische Keilschrift erfolgreich war, könnten in anderen Fällen möglicherweise neue Probleme auftreten. Da der Algorithmus bisher nur Kombinationen aus zwei Buchstaben berücksichtigt, würde er nicht für Transliterationen funktionieren, die auch Kombinationen aus mehr als zwei Buchstaben enthalten, beispielsweise für die Transliteration des kyrillischen Alphabets, die Kombination wie ``shtsh'' für den kyrillischen Buchstaben Щ enthält. Für Fälle wie diesen müsste der Ansatz weiterentwickelt werden. Darüber hinaus wäre zu überlegen, inwieweit die Performanz des Algorithmus weiter verbessert werden könnte. Durch die Verwendung eines Burkhard-Keller-Baumes konnte die Performanz immerhin so weit gesteigert werden, dass die Suchzeit auf einem Testkorpus mit über 35.000 Einträgen auf unter eine halbe Sekunde im Durchschnitt reduziert wurde. Für die Verwendung mit größeren Wörterbüchern könnte jedoch eine weitere Verbesserung der Performanz wünschenswert sein. Als Möglichkeit hierfür wäre etwa die Verwendung eines Levenshtein-Automaten nach Schulz / Mihov (2002) zu prüfen, der als besonders effiziente Umsetzung des Levenshtein-Algorithmus gilt.

Fußnoten

1. https://pinyin.sogou.com/
2. https://www.google.co.jp/ime/
3. https://gitlab.rlp.net/mscherl/FuzzySearch
4. http://www.higeomes.org/
5. http://www.geonames.org/
6. https://situx.github.io/SemanticDictionary/

Bibliographie

Brockelmann, C. / Fischer, A. / Heffening, W. / Taeschner, F. (1935): *Die Transliteration der arabischen Schrift in ihrer Anwendung auf die Hauptliteratursprachen der islamischen Welt. Denkschrift dem 19. Internationalen Orientalistenkongreß in Rom.*

De La Briandais, R. (1959): *File searching using variable length keys*. In: Papers presented at the the March 3-5, 1959, western joint computer conference. pp. 295–298. ACM.

Homburg, T. (2017): *Postagging and semantic dictionary creation for hittite cuneiform*. In: DH2017 .

Homburg, T. (2018): *Semantische Extraktion auf antiken Schriften am Beispiel von Keilschriftsprachen mithilfe semantischer Wörterbücher*. In: Dhd2018 .

Homburg, T., Chiarcos, C. (2016): *Word segmentation for akkadian cuneiform*. In: LREC 2016 .

Homburg, T., Chiarcos, C., Richter, T., Wicke, D. (2015): *Learning cuneiform the modern way*, http://gams.uni-graz.at/o:dhd2015.p.55 .

Krueger, M.H., Neeson, K.D. (2000): *Japanese text input method using a limited roman character set*, uS Patent 6,098,086

Kukich, K. (1992): *Techniques for automatically correcting words in text*. ACM Computing Surveys 24,4 .

Lasko, T.A., Hauser, S.E. (2001): *Approximate string matching algorithms for limited-vocabulary ocr output correction*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.1064&rep=rep1&type=pdf .

Levenshtein, V.I. (1966): *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady 10,8 .

Pedersen, T.T. (2008): *Transliteration of arabic*, http://transliteration.eki.ee/pdf/Arabic_2.2.pdf .

Schlott-Kotschote, A. (2004): *Transkription arabischer Schriften. Vorschläge für eine einheitliche Umschrift arabischer Bezeichnungen.*

Schulz, K.U., Mihov, S. (2002): *Fast string correction with levenshtein automata*. International Journal on Document Analysis and Recognition 5 .

UNGEGN Working Group, R.S. (2016): *Arabic. report on the current state of united nations romanization systems for geographical names*. version 4.0, http://www.eki.ee/wgrs/rom1_ar.pdf .

Weigel, A., Baumann, S., Rohrschneider, J. (1995): *Lexical postprocessing by heuristic search and automatic determination of the edit costs*. In: Proceedings of the Third International Conference on Document Analysis and Recognition .

Zgank, Kacic (2012): *Predicting the acoustic confusability between words for a speech recognition system using levenshtein distance*, http://eejournal.ktu.lt/index.php/elt/article/download/2628/1917 .

Ziółko, B., Gałka, J., Skurzok, D., Jadczyk, T. (2010): *Modified weighted levenshtein distance in automatic speech recognition*, http://www.dsp.agh.edu.pl/_media/pl:bziolko_kkzmbm2010final.pdf .

eXist-db und VueJS für dynamische UI-Komponenten

Pohl, Oliver
opohl@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Dogaru, Teodora
teodora.dogaru@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Müller-Laackman, Jonas
jonas.mueller-laackman@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Einführung

Für die Erstellung und Bearbeitung digitaler Editionen im Bereich der Digital Humanities werden vorzugsweise XML-basierte Lösungen verwendet. Eine der dabei am häufigsten benutzten Datenbanktechnologien ist eXist-db[1]. Dabei ist eXist-db nicht nur eine Lösung zum Bearbeiten und Speichern von XML-Daten, sondern bietet auch Möglichkeiten zur Programmierung von Webanwendungen auf Grundlage von XQuery. Im weiteren Programmierer*innen-Umfeld sind X-Technologie-Lösungen sowohl für die Datenhaltung als auch für das Erstellen von Webanwendungen eher eine Nischenlösung[2].

In den letzten Jahren werden verstärkt Frameworks wie Vue.js[3] oder React[4] für die Erstellung von Frontendmodulen in Webanwendungen verwendet.[5] Ein großer Vorteil solcher Frameworks gegenüber herkömmlichen Lösungen ohne Framework oder mit jQuery[6] ist die Möglichkeit, in sich geschlossene, reaktive User-Interface-Komponenten zu erstellen. Reaktiv heißt in diesem Fall, dass wenn sich das Datenmodell innerhalb einer Komponente ändert, ohne weiteren Aufwand sofort auch die Benutzeroberfläche entsprechend angepasst wird. Gleiches gilt umgekehrt, z.B. wenn eine Nutzerin ein Textfeld ausfüllt, auch das Datenmodell der Komponente sofort den entsprechenden Wert übernimmt, und so weitere Operationen im Hintergrund ausführen kann.

Da diese JavaScript-Frameworks jedoch mit JSON statt mit XML arbeiten, soll der hier vorgeschlagene Workshop eine Einführung in die Verwendung von Vue.js in Verbindung mit eXist-db geben. Im Workshop soll konkret gezeigt werden, wie man mit eXist-db abgefragte Daten nach JSON serialisieren kann, um diese dann für die Erstellung einer Vue.js basierten, reaktiven Benutzeroberfläche einzusetzen. Grundsätzlich soll den Teilnehmenden hierdurch das Zusammenspiel zwischen einem beliebigen Backend - JSON-API - JavaScript-Frontend-Framework vermittelt werden.

Workshop

Als praktisches Beispiel für den Workshop werden die Daten und der Quellcode von quoteSalute[7] zur Veranschaulichung benutzt. quoteSalute aggregiert Grußformeln aus verschiedenen digitalen Briefeditionen, bereitet diese auf, sodass NutzerInnen Briefabschlüsse von z.B. Alexander von Humboldt, Friedrich Schleiermacher oder Anna Gräfin von Lehndorff über einen Klick in ihre eigene E-Mail-Korrespondenz einfügen können. Die Daten dafür werden in einer eXist-db-Instanz als TEI-XML vorgehalten und die Benutzeroberfläche über Vue.js realisiert. Dabei kommuniziert die Benutzeroberfläche im Browser mit dem eXist-db-Backend, sodass auf Knopfdruck via XQuery in der Datenbank eine zufällige Grußformel ausgewählt und nach JSON serialisiert wird, sodass letztendlich die Nutzeroberfläche mit der neuen Grußformel angepasst werden kann. Um die passende Grußformel für die eigene E-Mail zu finden, können Nutzer*innenden Korpus nach Höflichkeitsstufe (formell, informell), Sprache (derzeit deutsch, englisch, französisch, italienisch, spanisch, griechisch, latein) oder nach Geschlecht des Absenders und Empfängers filtern.

Sowohl der Datenkorpus, als auch der Quellcode sowie die Dokumentation sind als freie Software verfügbar. [8]

Der Workshop ist für zwei halbe Tage und für ca. 15-30 Teilnehmende ausgelegt. Dabei ist das Workshopprogramm in insgesamt vier Zwei-Stunden-Blöcke aufgeteilt, die inhaltlich aufeinander aufbauen. Im Folgenden werden die im Vorfeld des Workshops zu verteilenden Materialien, die zu erwartenden Voraussetzungen an die Teilnehmenden, sowie die Inhalte der einzelnen Blöcke und ihre Lernziele beschrieben.

Voraussetzungen

Von den Teilnehmenden wird erwartet, dass sie grundlegende Kenntnisse in der JavaScript- und Web-Programmierung (HTML, CSS) haben. Weiterhin wird darum gebeten, einen Computer mitzubringen und vorab eXist-db (kostenfrei) sowie einen Texteditor zum Programmieren wie Visual Studio Code oder Atom zu installieren.

Materialien

Vor dem Workshop stellen wir eine eXist-db-Applikation zusammen, die bereits alle benötigten Daten, Skripte und Datenbankabfragen enthält, um im Workshop sofort in die Arbeit einsteigen zu können, ohne selbst noch alles einrichten zu müssen. Die Teilnehmenden können dann diese App in ihrer eXist-db per Drag & Drop installieren. Eine entsprechende Anleitung wird bereitgestellt.

Sollte die Einrichtung der eXist-Instanz oder die Installation der App wider Erwarten nicht funktionieren, stellen wir auch eine eigene eXist-Instanz für den Workshop mit personalisierten Logins für die Teilnehmenden sowie eine JSON-REST-API bereit, sodass die weiteren Workshopinhalte in jedem Fall praktisch nachvollzogen werden können.

Des Weiteren bereiten wir ein HTML- und VueJS-Code-Skelett vor, sodass auf JavaScript-Seite wenig Zeit auf die Einrichtung der Programmierungsumgebung verwendet werden muss.

Die hier aufgelisteten Daten werden auf Github bereitgestellt und dokumentiert.

Inhalte

Block 1

Zu Anfang sollen die im Workshop verwendeten Tools und Konzepte, insbesondere das Zusammenspiel zwischen exist-db, JSON und Vue.js, vorgestellt werden. Es wird den Teilnehmenden noch etwas Zeit eingeräumt, die nötige Software herunterzuladen, zu installieren und ihre Entwicklungsumgebung einzurichten. Anschließend wird der Aufbau der quoteSalute-eXist-db-App inklusive des mitgelieferten Datenkorpus erklärt und gezeigt, wie innerhalb von eXist-db Daten abgefragt und ins JSON-Format serialisiert werden. Diese Schritte sind bereits voreingerichtet, sodass auch Teilnehmende ohne XQuery oder eXist-db-Kenntnisse den Ablauf nachvollziehen können. Aus praktischer Sicht werden die Teilnehmenden den quoteSalute-Grußformel-Korpus abfragen, eine zufällige Grußformel extrahieren und nach JSON serialisieren.

Block 2

Im zweiten Block sollen die Teilnehmenden anfangs versuchen, die vorgegebenen Beispiele für Datenbankabfragen und JSON-Serialisierungen anzupassen bzw. zu erweitern, um beispielsweise nur Grußformeln, die an Frauen adressiert waren, aus dem Datenbestand herauszufiltern. So soll die Arbeit mit eXist-db und JSON-Serialisierung geübt und gefestigt werden.

Um die aus der eXist-db gelieferten JSON-Daten client-seitig verwenden zu können, muss das JavaScript-Frontend-Framework Vue.js mit der JSON-REST-API kommunizieren. Dafür soll zuerst in die Grundprinzipien von Vue.js (data, templates, model-view-binding, methods, computed) anhand kleiner, aufeinander aufbauenden Code-Beispiele eingeführt werden.

Block 3

In Block 3 sollen die in den ersten beiden Blöcken erarbeiteten Methoden und Abläufe miteinander in Verbindung gebracht werden. Über eine asynchrone Abfrage zum lokalen exist-db-Server (JavaScript-Skelett wird gestellt), soll die Vue.js-App eine neue Grußformel vom Server zu erfragen und diese dann im eigenen Datenmodell und so auch in der Benutzeroberfläche abbilden. Um asynchrone Operationen mit JavaScript nachvollziehen zu können, wird an dieser Stelle zusätzlich das Konzept von Promises vorgestellt und erläutert.

Die Teilnehmenden können die verbleibende Zeit in diesem Block dazu nutzen, ihre Vue.js-App zu programmieren und gegebenenfalls visuell anzupassen.

Block 4

Sollte die Zeit für die Programmierung der Vue.js-App nicht genügen, dient der vierte Block als zusätzlicher Puffer. Für bereits fertige Teilnehmer und Interessierte bieten wir noch ein optionales Programm mit zusätzlichen Aufgaben an. So soll das Zusammenspiel zwischen eXist-db und Vue.js noch an einem weiteren, konkreten Fallbeispiel für die Digital Humanities geübt werden. Die Teilnehmenden erhalten einen Auszug aus einer digitalen Edition und sollen mit den vorgestellten Technologien ein Personenregister erzeugen. Durch diese Transferleistungsübung sollen die erlernten Vue.js-Methoden und Prinzipien gefestigt werden.

Übersicht

Das Grundkonzept des Workshops lässt sich in folgender Tabelle zusammenfassen:

	Technologien & Konzepte	Lernziel
Block 1	Überblick eXist-db, VueJS & JSON	JSON-Serialisierung von XML-Daten mit eXist-db verstehen
Block 2	eXist-db, JSON, XPath, XQuery	eXist-db Datenbankanfragen modifizieren und Ergebnisse nach JSON serialisieren
VueJS	Grundkonzepte von VueJS verstehen und anwenden	
Block 3-4	VueJS, Asynchrone Anfragen, Promises	VueJS-App programmieren und an JSON-REST-API anbinden

Nach dem Workshop sollen die Teilnehmenden in der Lage sein

- nachzuvollziehen, wie man mit eXist-db XML-Daten nach JSON serialisiert,
- die Bedeutung von der Vue.js-Begriffe data, methods, model-view-binding und templates zu verstehen, anwenden und selbst erklären zu können,
- eine einfache, reaktive Vue.js-Applikation zu programmieren,
- diese Vue.js-Applikation an eine JSON-REST-Schnittstelle anzubinden, und

- das grundlegende Prinzip im Zusammenspiel von Backend - JSON-API - JavaScript-Frontend auch auf andere Technologien und Frameworks übertragen zu können.

Technische Ausstattung

Für die Durchführung des Workshops wird ein Beamer (vorzugsweise mit VGA-Anschluss) sowie Internetzugang benötigt. Aufgrund der Länge des Workshops wären mehrere Mehrfachsteckdosen zum Aufladen der Geräte der Teilnehmenden wünschenswert.

Fußnoten

1. http://exist-db.org
2. In ProgrammiererInnen-Umfragen von JetBrains und StackOverflow oder Programmier- und Datenbank-Beliebheits-Indizes sind eXist-db, XQuery und verwandte Technologien nicht einmal genannt.
https://www.jetbrains.com/research/devecosystem-2018/
https://insights.stackoverflow.com/survey/2018#technology
https://www.tiobe.com/tiobe-index/
http://pypl.github.io/DB.html
3. https://vuejs.org/
4. https://reactjs.org/
5. http://2016.stateofjs.com/2016/frontend/
https://2017.stateofjs.com/2017/front-end/results
5. http://2016.stateofjs.com/2016/frontend/
https://2017.stateofjs.com/2017/front-end/results
6. https://jquery.com/
7. http://quotesalute.net
8. https://github.com/telota/quoteSalute

Grundzüge einer visuellen Stilometrie

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Dubray, David
ddubray@uni-potsdam.de
Universität Potsdam, Deutschland

Was kennzeichnet visuellen Stil? Nachdem die digitalen Geisteswissenschaften stark durch textanalytische Verfahren aus der Computerlinguistik und verwandten Gebieten geprägt waren, sind in den letzten Jahren vermehrt Methoden zur Beschreibung visuellen Materials vorgeschlagen worden. Diese sollten insbesondere den Bildwissenschaften neue methodische Zugänge ermöglichen. Die Frage nach einer formalen Beschreibung visuellen Stils etwa hat die kunstgeschichtlichen Forschung seit ihrem Beginn umgetrieben (Wölfflin 1915). In der Form- und Strukturanalyse dominieren jedoch verbale Beschreibungen, eine quantitative Lösung jenseits deskriptiver Ansätze steht aus. Neuere Entwicklungen im Bereich des maschinellen Sehens (*Computer Vision*) lassen nun eine formale Beschreibung visuellen Stils greifbar werden. Diese basiert auf Repräsentationen in den tieferen Schichten sogenannter Convolutional Neural Networks (CNNs).

CNNs sind eine Klasse tiefer neuronaler Netze, die inspiriert von biologischen visuellen Systemen entwickelt wurden, um ingenieurwissenschaftliche Probleme wie z.B. Handschrifterkennung zu lösen (LeCun et al. 1989). Durch nur lokale Konnektivität sind CNNs deutlich sparsamer und effizienter als klassische „ *fully connected*“ neuronale Netze. CNNs lassen sich beschreiben als eine hierarchische Anordnung computationaler Einheiten, die visuelle Information in einem *Feedforward*-Prozess verarbeiten. Jede Schicht der Hierarchie lässt sich interpretieren als eine Menge von Filtern, die bestimmte Merkmale des Eingabebildes extrahieren. Die Filterkoeffizienten werden durch Anpassung an die Daten gelernt. Die Ausgabe einer Schicht besteht aus einer Menge sogenannter Merkmalskarten, welche unterschiedlich gefilterte Versionen des Eingabebildes sind. Filter auf höheren Schichten erhalten als Eingabe im Wesentlichen eine gewichtete Rekombination der Merkmalskarten niedrigerer Schichten. In den unteren Schichten sind die Repräsentationen relativ einfach und entdecken beispielsweise Kanten oder Farben. Repräsentationen mittlerer Schichten können z.B. texturartig sein, während höhere Schichten deutlich komplexer sind und z.B. Objektteile repräsentieren können. Die unterschiedlichen Repräsentationsebenen weisen eine starke Ähnlichkeit mit der hierarchischen Verarbeitung im für Objekterkennung zuständigen ventralen Pfad des menschlichen visuellen Systems auf (Yamins and DiCarlo 2016), weshalb CNNs auch aussichtsreiche Kandidaten für die nähere und quantitativ fundierte Untersuchung ungelöster Probleme der sogenannten *Mid-Level Vision* sind. Auch in diesem Bereich dominierten bis vor kurzem qualitative Beschreibungen wie z.B. gestaltpsychologische Ansätze.

Neural Style Transfer. Wie kodieren CNNs nun Stil? Leon Gatys, Alexander Ecker und Matthias Bethge haben die Methode des Style Transfer entwickelt, in der sie zeigen, dass Stil und Inhalt eines Bildes in CNNs zu einem gewissen Grad unabhängig voneinander repräsentiert werden. Am Beispiel des VGG-Netzwerkes (Simonyan and Ziisserman 2014) demonstrieren Gatys et al., dass stilistische Elemente auf niedrigeren Schichten und Bildinhalte auf höheren Schichten des Netzwerks kodiert werden. Der Stil eines Bildes A lässt sich deshalb prinzipiell auf den Inhalt eines Bildes B übertragen. Neuere Arbeiten haben diesen *Style Transfer* weiter optimiert (Sanakoyeu et al. 2018). Interessanterweise ist Stil in diesen Arbeiten von gegenständlichen und auch abstrakten Gemälden extrahiert worden, obwohl das zugrundeliegende VGG für die Klassifikation von Fotos vortrainiert war. Die gelernten Filter sind offensichtlich hinreichend generisch, um derartigen Transfer zu ermöglichen.

Stil von Illustratoren. In Vorarbeiten haben wir gezeigt, dass sich CNN-Repräsentationen auch zur Beschreibung grafischer Literatur wie Comics, Graphic Novels etc. eignen (Laubrock, Hohenstein and Kümmerer 2018). *Welche Repräsentationen sind nun aber charakteristisch für den Stil von Illustratoren?* Dieser Frage gehen wir in der vorliegenden Untersuchung nach. Wir nutzen dazu das Xception-Netzwerk (Chollet 2016), das deutlich effizienter ist als VGG und bei weniger Parametern

typischerweise eine bessere Klassifikationsleistung erbringt. Als „Signatur" eines Zeichners extrahieren wir das Muster der mittleren Antwortstärke über verschiedene Filter. Diese benutzen wir als Prädiktor für eine Illustrator-Klassifikation. Experimentell variieren wir dabei, aus welchen Ebenen des Netzwerks Filter zu Klassifikation genutzt werden. Die Güte der Klassifikation als Funktion der benutzten Filter dient zur Abschätzung dafür, wie relevant auf einer bestimmten Hierarchieebene repräsentierte Merkmale für den Individualstil sind. Zusätzlich berechnen wir eine Ähnlichkeitsmatrix basierend auf den CNN-Aktivierungen als Grundlage für eine bildbasierte Suche.

Material

Als Material verwenden wir zwei Sammlungen grafischer Literatur: (a) das Graphic Narrative Corpus (GNC; Dunst, Hartel and Laubrock 2009) und (b) Manga109 (Matsui et al. 2017). Das GNC ist eine kuratierte Sammlung über 200 zeitgenössischer Graphic Novels aus den Jahren 1979 bis 2017 mit einem Gesamtumfang von mehr als 50.000 Seiten. Der GNC beinhaltet Werke verschiedener Genres (z.B. Autobiographie, New Journalism, Crime, Superhelden). Manga109 besteht aus 109 Manga-Bänden (mehr als 20.000 Seiten), die zwischen 1970 und 2010 im Handel erhältlich waren und 2017 der Wissenschaft zur Verfügung gestellt wurden. Die Korpora wurden durch zufällige geschichtete Stichprobenziehung in ein Trainings- und ein Testcorpus unterteilt.

Methode

Der CNN-Teil eines auf dem ImageNet-Datensatz (Deng et al. 2009) vortrainierten Xception-Netzwerk wurde benutzt, um Illustratoren in den beiden Korpora zu klassifizieren.

Bildsignatur. Für jedes Bild wurden zunächst Merkmalskarten auf verschiedenen ausgewählten Schichten des Xception-Netzes berechnet. Für die Merkmalskarten pro Filter wurde dann die mittlere Aktivierung (*global average pooling*) berechnet. Ein Vektor der mittleren Aktivierung über eine Menge von Filtern wurde als Signatur des Bildes gespeichert. Dies resultiert in einer recht kompakten Repräsentation mit einem Kompressionsverhältnis im Vergleich zum Ausgangsbild von ca. 1:800 für frühe bzw. ca. 1:100 für späte Schichten und 1:21 bei Verwendung aller Filter.

Klassifikation. Zur Klassifikation trainierten wir ein einfaches *fully-connected* neuronales Netz mit einer verdeckten Schicht von 1024 Einheiten. Diese erhielten als Input die Bildsignatur (*average-pooled feature maps*, s.o.) und als Output die Illustratoren. Das Trainings-Set enthielt 90% der Seiten eines jeden Buches, zufällig bestimmte 10% der Seiten pro Comicband wurden nicht während des Trainings präsentiert, sondern als Test-Set zur Seite gelegt zur Bewertung der Klassifikationsleistung des trainierten Netzes.

Läsionsexperimente. Weil wir uns dafür interessierten, welche Art von Merkmal am charakteristischsten für den Stil eines Illustrators ist, haben wir das CNN auf fortschreitend niedrigeren Ebenen läsioniert und die Klassifikationsleistung mit dem vollen, auf allen Schichten basierenden Modell verglichen.

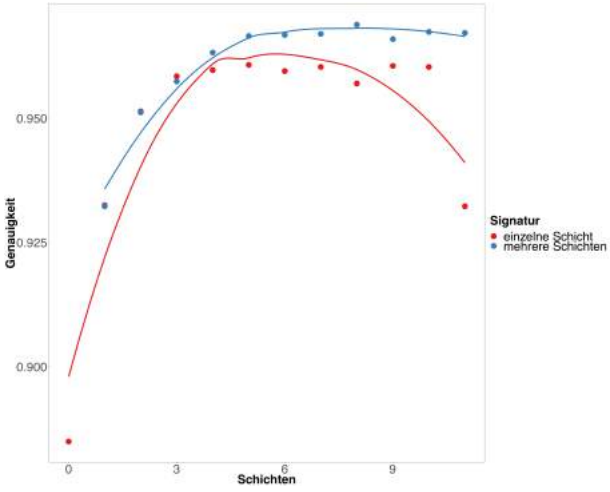
Zusätzlich haben wir Klassifikationen basierend auf dem Output einzelner Schichten berechnet. Insgesamt vergleichen wir also die Klassifikation unter Berücksichtigung einzelner Schichten 0, 1, ..., k vs. mit der bei Berücksichtigung aller Schichten von 0 bis k. Der Merkmalsvektor wurde im letzteren Fall durch einfache Verkettung der Signaturen gebildet.

Ähnlichkeitsmatrix. Die Ähnlichkeiten der Merkmalsvektoren aller Bilder wurden mittels euklidischer Distanz berechnet. Basierend auf dieser Matrix wurde eine Ähnlichkeitssuche implementiert.

Semantische Segmentierung. Mit Hilfe von CNN-Repräsentationen lassen sich auch sehr gut einzelne Bildelemente identifizieren. Zur Detektion von Sprechblasen trainieren wir ein Fully-Convolutional neuronales Netz nach der U-Net-Architektur (Ronneberger et al. 2015) auf 750 annotierte Comicseiten. In dieser Architektur wird neben einem Enkodier- auch ein Dekodierpfad benutzt, in dem die recht abstrakten, semantiknahen Repräsentationen höherer Ebenen mit Kopien der Information niedrigerer Ebenen verrechnet wird, um Ortsinformation zu rekonstruieren. Auch hier haben wir beim Enkodierpfad wieder mit vortrainierten Repräsentationen begonnen und nur ein Feintuning vorgenommen.

Ergebnisse

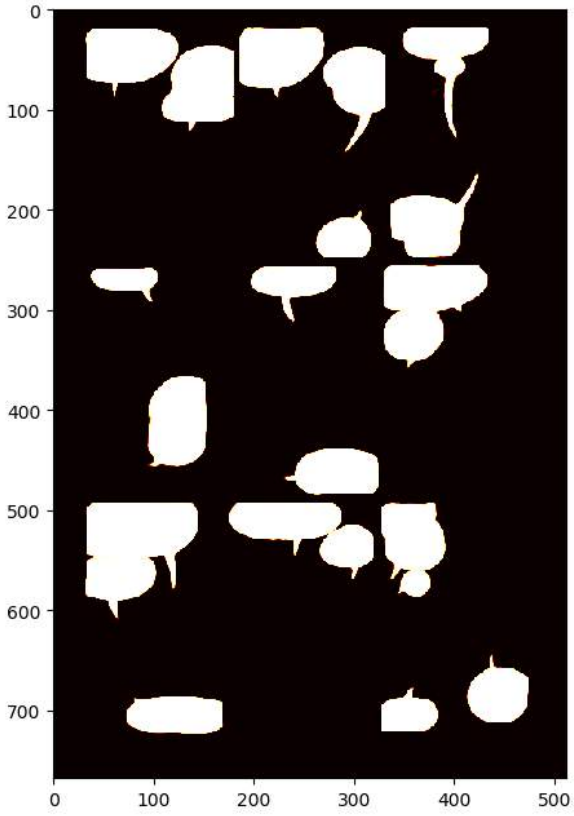
Abbildung 1 zeigt die Genauigkeit der Klassifikation als Funktion der zugrundeliegenden Merkmale. Insgesamt lassen sich die Seiten aufgrund rein visueller Analyse sehr gut ihren Urhebern zuordnen. Man erkennt am Abfall der Kurve für Merkmale aus einzelnen Schichten, dass für die Illustrator-Klassifikation die Repräsentationen mittlerer Ebenen am entscheidendsten sind. Die stilistische Signatur einer Graphic Novel basiert scheinbar eher auf Merkmalen mittlerer Komplexität wie Schraffuren, Texturen oder Schwüngen als auf höher integrierten Merkmalen wie Objektteilen oder spezifischen Motiven.



Genauigkeit der Klassifikation. Die x-Achse gibt an, welche Netzwerkschichten zur Berechnung der Signatur herangezogen wurden (siehe Text für Details). Die Linienfarbe unterscheidet, ob die Klassifikation auf Signaturen einzelner Netzwerkschichten k basiert (rot) oder auf Signaturen der Schichten von 0 bis k, k∈{0, ..., 11} (blau).

Basierend auf den Merkmalsvektoren haben wir eine bildbasierte Ähnlichkeitssuche implementiert. Nach Eingabe eines Suchbildes werden beispielsweise die 10 ähnlichsten Bilder ausgegeben. Die Untersuchung der Klassifikationsfehler ist interessant, sie zeigt beispielsweise, dass unterschiedliche Werke eines Autors zusammen gruppiert werden. Verwechslungen treten eher innerhalb von als zwischen Genres auf. Selbst historische Entwicklungen lassen sich abbilden: In „750 Years in Paris" illustriert Vincent Mahé die Entwicklung eines Häuserblocks in Paris von 1265 bis 2015. Die Bildsuche mit einer „frühen" Seite liefert Bilder aus der frühen Zeit, ebenso liefert die Bildsuche mit einer „späten" Seite Bilder aus einer späteren Epoche.

Bei der semantischen Segmentation von Sprechblasen haben wir ein hervorragendes Ergebnis erzielt. Der F1-Score auf dem Testset betrug 0.935. Auch Elemente wie ein geschwungener Hinweisstrich / Dorn und an den Rändern offene Sprechblasen konnten sehr gut segmentiert werden. Abbildung 2 zeigt ein Beispiel einer Seite, auf der alle Sprechblasen korrekt detektiert und sehr gut segmentiert wurden.



Beispiel für Sprechblasen-Segmentation.

Diskussion

Wir haben verschiedene Sammlungen grafischer Literatur mit CNNs beschrieben und den Beitrag interner CNN-Repräsentationen unterschiedlicher Schichten zur Klassifikation von Zeichenstilen untersucht. Unsere Ergebnisse zeigen, dass der Individualstil eines Zeichners eher durch Merkmale mittlerer als durch solche höherer Komplexität charakterisiert ist. Allgemein haben CNN-basierte Repräsentationen das Potenzial, eine formale Beschreibung stilistischer Merkmale abzubilden. Sie sind deshalb aussichtsreiche Kandidaten für eine quantitative Fundierung bildwissenschaftlicher Form- und Strukturanalyse.

Bibliographie

Chollet, F. (2016): *Xception: Deep learning with depthwise separable convolutions*. In: CoRR: abs/1610.02357.

Chu, W. / Wu, Y. (2018): *"Image style classification based on learnt deep correlation features."* In: IEEE Transactions on Multimedia 20(9): 2491–2502.

Deng, J. / Dong, W. / Socher, R. / Li, L.-J. / Li, K. / Fei-Fei, L. (2009): *„ImageNet: A Large-Scale Hierarchical Image Database."* In: 2009 IEEE Conference on Computer Vision and Pattern Recognition: 248–255.

Dunst, A. / Hartel, R. / Laubrock, J. (2017): *„The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the Digital Humanities."* In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 03: 15–20.

Laubrock, J. / Hohenstein, S. / Kümmerer, M. (2018): *„Attention to comics: Cognitive processing during the reading of graphic literature."* In: Dunst, A., Laubrock, J., and Wildfeuer, J., editors, Empirical Comics Research: Digital, Multimodal, and Cognitive Methods, ch.12: 239–263. Routledge, New York.

Matsui, Y. / Ito, K. / Aramaki, Y. / Fujimoto, A. / Ogawa, T. / Yamasaki, T. / Aizawa, K. (2017): *"Sketch-based manga retrieval using Manga109 dataset."* In: Multimedia Tools and Applications 76(20): 21811–21838.

Gatys, L. A. / Ecker, A. S. / Bethge, M. (2016): *"Image style transfer using convolutional neural networks."* In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016: 2414–2423.

LeCun, Y. / Boser, B. / Denker, J. S. / Henderson, D. / Howard, R. E. / Hubbard, W. / Jackel, L. D. (1989): *"Backpropagation applied to handwritten zip code recognition."* In: Neural Computation 1(4): 541–551.

Ronneberger, O. / Fischer, P. / Brox, T. (2015): *„U-Net: Convolutional Networks for Biomedical Image Segmentation."* In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241.

Sanakoyeu, A. / Kotovenko, D. / Lang, S., / Ommer, B. (2018): *"A Style-Aware Content Loss for Real-time HD Style Transfer."* In: arXiv preprint arXiv:1807.10201.

Wölfflin, H. (1915): *Kunstgeschichtliche Grundbegriffe: das Problem der Stilentwickelung in der neueren Kunst.* München: Bruckmann.

Yamins, D. L. K. / DiCarlo, J. J. (2016): *"Using goal-driven deep learning models to understand sensory cortex"*. In: Nature Neuroscience 19(3):356–365.

Simonyan, K. / Zisserman, A. (2014): *"Very deep convolutional networks for large-scale image recognition."* In: arXiv preprint arXiv:1409.1556

Herausforderungen des Digital Storytelling am Beispiel des VRlabs des Deutschen Museums

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

Geipel, Andrea
a.geipel@deutsches-museum.de
Deutsches Museum, Deutschland

Henkensiefken, Claus
c.henkensiefken@deutsches-museum.de
Deutsches Museum, Deutschland

Digitale Entwicklungen in den Bereichen Virtual Reality (VR) und Augmented Reality (AR) bieten auch für Museen neue Möglichkeiten im Umgang mit und in der Vermittlung des kulturellen Erbes. Gleichzeitig stellen sich Herausforderungen hinsichtlich infrastruktureller Ressourcen und nachhaltiger Vermittlungskonzepte.

Ausgangslage

Das 1903 gegründete Deutsche Museum von Meisterwerken der Naturwissenschaft und Technik ist das weltweit größte Technikmuseum und als Forschungsmuseum und Mitgliedseinrichtung der Leibniz-Gemeinschaft ein international führendes Zentrum für die Erforschung der wissenschaftlich-technischen Kultur. Im Rahmen einer Zukunftsinitiative wird bis 2025 das gesamte Ausstellungsgebäude saniert und erneuert. Ein bedeutender Bestandteil dieser Zukunftsinitiative ist die Maßnahme „Deutsches Museum Digital“. Mit diesem Projekt führt das Deutsche Museum eine der größten Digitalisierungsmaßnahmen an deutschen Kulturinstitutionen durch. Bis 2025 wird ein Bündel umfassender Maßnahmen umgesetzt, um das Deutsche Museum auch als digitale Forschungsressource zu etablieren und es um einen digitalen Erlebnisraum für breite Gesellschaftsschichten zu erweitern [Hohmann 2013].

In den vergangenen Jahren wurden die technischen Grundlagen gelegt, die Infrastruktur aufgebaut, digitale Arbeitsabläufe etabliert und umfangreiche Digitalisierungsmaßnahmen in den Objektsammlungen, der Bibliothek und dem Archiv des Deutschen Museums durchgeführt. Dabei wurde bereits ein hoher Prozentsatz der insgesamt rund 120 000 Objekte des Hauses digital abgelichtet, in einer Datenbank erfasst und mit standardisierten Metadaten angereichert. Die Ergebnisse werden der allgemeinen Öffentlichkeit über das Portal „digital.deutsches-museum.de“ zur Verfügung gestellt.

3D-Erfassung und -Visualisierung

3D-Digitalisierung

Neben der „klassischen“ Digitalisierung von Kulturgut wurden und werden in flankierenden Projekten auch innovativere Methoden der Digitalisierung erprobt. Besonders die Möglichkeiten der 3D-Digitalisierung haben in den letzten Jahren eine hohe Aufmerksamkeit bekommen [CORDIS 2008], so dass es Nahe lag, die Möglichkeiten aktueller Methoden und Techniken der 3D-Digitalisierung zu eruieren. Eine besondere Herausforderung stellte dabei die Art und Beschaffenheit unserer Sammlungsobjekte dar.

Im Projekt kamen sowohl 3D-Laserscan als auch fotogrammetrische Verfahren zum Einsatz. Um einen möglichst differenzierten Einblick in die Möglichkeiten zu bekommen, haben wir bedeutende Objekte zur Digitalisierung herangezogen: Erste Präzisionsventil-Dampfmaschine von Sulzer (1865), die Nachbildung des Normal-Segelapparat von Otto Lilienthal (1894/1962), das Modell des Apollo-15 Mondfahrzeugs (Lunar Roving Vehicles, 1971/2009) und der Benz Patentmotorwagen (1886). Die Objekte unterschieden sich stark in Größe, Material und Beschaffenheit.

Die Sulzer Dampfmaschine ist über 5 Meter hoch und in einer Wand in der Kraftmaschinenhalle des Deutschen Museums verbaut. Zunächst wurde unter Verwendung eines 3D-Laserscanners Aufnahmen aus verschiedenen Positionen gemacht, die hinterher zu einem Modell zusammengefügt wurden. Zusätzlich wurden mit einem experimentellen Aufbau ein photogrammetrischer Ansatz verfolgt und zahlreiche digitale Blitzlicht-Fotografien angefertigt. Insgesamt wurden 45.000 Einzelbilder angefertigt, auf deren Basis auf einem Rechner-Cluster ein 3D-Modell errechnet wurde. Beide Verfahren zeitigten einen sehr hohen technischen und organisatorischen Aufwand. Für die Anfertigung der 3D-Modelle waren sehr umfangreiche händische Bereinigungen und Nacharbeiten notwendig. Durch die hohe Qualität der verwendeten Techniken konnten die Ergebnisse in der Detailgenauigkeit sehr beeindrucken, waren aber ohne massive Reduktion nicht auf normaler Rechnerhardware nutzbar.

3D-Re-Engineering

Als zukunftsweisend haben sich schließlich die Möglichkeiten des (Reverse) 3D-Re-Engineering erwiesen. Auf Basis der 3D-Scans und bereits im Haus vorhandenen Archivmaterial (Baupläne etc.) wurden die genannten Objekte im Detail und originalgetreu als 3D-Modell nachgebaut. Auch dieses Verfahren hat sich personell und technisch ebenfalls als sehr aufwändig erwiesen, allerdings hatte das Ergebnis einen weit höheren Nutzwert.

Die Objekte des Deutschen Museums besitzen ihren musealen Wert in der Regel über ihre Funktion bzw. ihren Einsatzzweck. Die Maschinen konnten nun beispielsweise virtuell in Bewegung gesetzt werden, in ihre Bestandteile zerlegt, mit Zusatzinformationen angereichert oder in ihrem originalen Kontext visualisiert werden. Der Erkenntnisgewinn kann bei entsprechender Aufbereitung am beim virtuellen Modell höher sein als bei der reinen Betrachtung des Objekts im Museumskontext. Daraus ergab sich die Frage, wie

solche Modelle auch dem/-r Museumsbesucher/in erfahrbar gemacht werden konnten.

Durch das Projekt „Museum4punkt0“ wurde dem Deutschen Museum die Möglichkeit gegeben, die bisherigen Ansätze weiter zu verfolgen und wissenschaftlich zu bearbeiten.

Das Verbundprojekt „Museum4punkt0“

„Museum4punkt0 - Digitale Strategien für das Museum der Zukunft“ ist ein Verbundprojekt der Stiftung Preußischer Kulturbesitz und ihrer Staatlichen Museen zu Berlin, der Humboldt Forum Kultur GmbH, des Deutschen Auswandererhauses Bremerhaven, des Deutschen Museums München, der Fastnachtsmuseen Langenstein und Bad Dürrheim mit weiteren Museen der schwäbisch-alemannischen Fastnacht und des Senckenberg Museums für Naturkunde Görlitz. Es ist auf drei Jahre angelegt und wird von der Beauftragten der Bundesregierung für Kultur und Medien gefördert [Museum4punkt0 2018].

Abbildung 1. Projektstruktur des Verbundprojekts Museum4punkt0

In sechs modellhaften Teilprojekten widmet sich das Projekt Fragen rund um den Einsatz digitaler Technologien im Museum. Dabei werden neue Formate und digitale Prototypen für Bildung, Vermittlung, Partizipation und Kommunikation entwickelt und zugleich die Herausforderungen in den Blick genommen, die die Nutzung digitaler Technologien für Personal, Infrastruktur und Arbeitsabläufe nach sich zieht.

Das Teilprojekt des Deutschen Museums

Das Thema „Virtual Reality“ gewinnt in Museen zunehmend an Bedeutung [Freeman Becker Cummins 2016], weshalb sich das Deutsche Museum im Verbundprojekt diesen Schwerpunkt gesetzt hat. Das Teilprojekt des Deutschen Museums „Perspektiven dreidimensionaler Visualisierungen in der musealen Vermittlung“ basiert auf den Ergebnissen der Vorprojekte und widmet sich dem Bereich der 3D-Visualisierung sowie der Virtual Reality oder Augmented Reality in großer Breite. Ziel ist die Entwicklung und prototypische Umsetzung von Methoden und Techniken, die den effektiven und zielgesteuerten Einsatz von 3D-Technologien im musealen Kontext ermöglichen. Der Schwerpunkt liegt auf der digitalen Vermittlung in den Besucherbereichen, Ausstellungen und im Web als eine tragende Säule der Museumsarbeit, aber auch die Anwendungs- und Nutzungsszenarien in den Bereichen Sammeln, Forschen und Bewahren werden mit einbezogen.

Das Entwicklungsprogramm lässt sich grob in zwei Bereichen gliedern. Der erste Bereich umfasst die Entwicklung von Methoden und Techniken, die benötigt werden, um dreidimensionale Inhalte zu erstellen oder zu generieren. Der zweite Bereich eruiert die Einsatzmöglichkeiten und Nutzungsszenarien für dreidimensionale Inhalte. Bei der Umsetzung wird eng mit einschlägigen Partnern aus Forschung und Wirtschaft zusammengearbeitet. Folgende Schwerpunkte sind im Teilprojekt definiert:

- Normen und Standards für AR/VR-Anwendungen im Museum
- Vom analogen Objekt zur digitalen Forschungsressource
- Digital Storytelling in virtuellen Museumswelten
- Online, Modular, Mobil – 3D-Anwendungen im Museum

Ein erster Meilenstein war die Einrichtung eines Virtual-Reality-Labors im Ausstellungsbereich des Deutschen Museum, um die Besucher/innen von Anfang an in den Entwicklungsprozess mit einzubeziehen und die Praxistauglichkeit zu testen. Parallel dazu wurden Strategien für das Digital Storytelling entwickelt, um zunächst die bereits vorhandenen digitalen Inhalte für den/die Besucher/in aufzubereiten und entsprechend dem Medium in Szene zu setzen.

Das Virtual-Reality-Labor des Deutschen Museums

Seit August 2018 können Besucher/innen des Deutschen Museums insgesamt vier dieser 3D-digitalisierten Objekte im VRlab mit Hilfe von VR-Brillen und Controllern interaktiv erkunden [Kommunikationsraum 2017].

Abbildung 2. Interaktionsmöglichkeiten beim Lilienthal-Normalsegelapparat

Die Verwendung bzw. Inszenierung der digitalen Objekte in der virtuellen Realität erfolgte während der Aufbauphase des Labors in engem Austausch zwischen Kuratorium, VR-Spezialisten/innen, Besuchern/-innen und der Museumsdidaktik. Aus den Erwartungen, Erfahrungen und Aspekten der technischen Machbarkeit, die zur Diskussion standen, wurde eine zweigleisige Storytelling-Strategie entwickelt. Neben der Darstellung in einem

virtuellen Ausstellungsraum mit Hintergrundinformationen und Verweisen auf den materiellen Zwilling im realen Ausstellungsraum, werden die Objekte auch in virtuellen Sequenzen erfahrbar. So können z.B. zentrale mechanische Elemente der Sulzer Dampfmaschine in einer Spinnerei bedient und zusammenhängende Prozesse besser verstanden werden.

Digital Storytelling im Spannungsfeld der Nutzergruppen

Die Ausgestaltung bzw. Erweiterung der VR-Szenen geschieht in einem Spannungsfeld der Interessen von Besuchern/-innen, Entwickler/innen und dem Museum, die nicht unbedingt deckungsgleich sind.

Im Vorfeld wurden einige nicht repräsentative Umfragen und Interviews mit Nutzergruppen durchgeführt. In der Gruppe der Besucher/innen – im Deutschen Museum mit einer sehr heterogenen Zusammensetzung – konnte konstatiert werden, dass in der Regel keinerlei Vorkenntnisse im Bereich VR vorhanden sind. Das Erlernen von 3D-Interaktionen in VR stellt eine große Herausforderung dar und benötigt umfassende Betreuung. Abseits von etwaigen Inhalten steht zunächst das pure Erlebnis im VR-Raum im Fokus.

Für VR-Entwickler/innen ist der Umgang mit VR-Technologien eine tägliche Erfahrung. Dadurch sind für diese Nutzergruppe die Erwartungen und Probleme der Besucher/innen in der 3D-Interaktion nur schwer zu antizipieren. Avisierte Lösungen zeichnen sich in der Regel dadurch aus, dass sie schon sehr viele Erfahrungen im VR-Umgang voraussetzen. Umgekehrt werden „einfache" Umsetzungen schnell als uninteressant und langweilig angesehen.

Für das Museum steht vor allem die inhaltliche Dimension im Vordergrund. Inhaltliche und physikalische Korrektheit sowie die Tiefe an – im klassischen Sinne – vermittelter Information werden grundsätzlich als wichtiger eingestuft als Innovation und Interaktionsmöglichkeiten. Ein grundsätzliches Misstrauen gegenüber dem Medium ist durchgängig festzustellen.

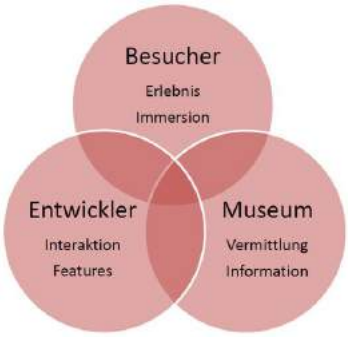


Abbildung 3. Spannungsfeld der Nutzergruppen.

Die Gruppe der Besucher/innen erwartet vor allem ein innovatives Erlebnis, sowohl inhaltlich als auch technisch. Dieses Erlebnis soll sich besonders durch eine hohe Immersion auszeichnen, die der VR-Technologie auch ohne Vorerfahrung zugeschrieben wird. Aus Museumssicht ist zwar auch das Besuchserlebnis wichtig, im Vordergrund steht aber die Vermittlung von wissenschaftlicher Information, und dies auf eine Art und Weise, die besonders nachhaltig ist und sich der analogen Informationsvermittlung überlegen erweist. Schon aufgrund des wesentlich breiteren Vorwissens über die Möglichkeiten des Mediums stehen aus Sicht der Entwickler/innen die Umsetzung innovativer Interaktionsmöglichkeiten im Vordergrund sowie das Angebot besonderer Features. Die Einschätzung, was als „innovativ" in Hinblick auf Features und Interaktion gesehen wird, ist in der Regel technologiegetrieben.

Die zentrale Herausforderung des Digital Storytelling in diesem Projekt besteht darin, in diesem Spannungsfeld einen gemeinsamen Nenner für eine „gelungene" Umsetzung zu finden.

Vorläufiges Fazit

Im laufenden Betrieb sollen bis Ende 2020 verschiedene Ansätze des Digital Storytelling erprobt und ökonomische Arbeitsweisen zur inhaltlichen Aufbereitung und Kontextualisierung von 3D-Inhalten evaluiert werden. Daneben stellt sich auch die Frage, inwiefern die Rückkopplung des virtuellen Objekts an seine materielle Vorlage im Haus technisch ebenso wie didaktisch denkbar ist und in welcher Form die 3D-Digitaliserung und virtuelle Einbettung materieller Objekte neue Sichtweisen sowohl für Museen als auch für Besucher/innen ermöglicht.

Das VRlab versteht sich in diesem Sinne als Experimentierfeld, um im stetigen Austausch zwischen Besucher/innen, Kuratoren/-innen, Software-Entwickler/innen sowie dem didaktischen Fachpersonal, Inhalte und Darstellungsformen weiterzuentwickeln. Mit Hilfe unterschiedlicher Evaluierungsmethoden, wie Interviews, Besucherbefragungen und qualitative Beobachtungen, sollen hierdurch Normen und Standards im Umgang mit VR-Technologien im musealen Umfeld identifiziert sowie darauf aufbauend Handlungsempfehlungen formuliert werden.

Bibliographie

CORDIS (2018): *3D-COFORM.* https://cordis.europa.eu/project/rcn/89256_de.html [letzter Zugriff 14. Oktober 2018]

Freeman, A. / Adams Becker, S. / Cummins, M. / McKelroy, E. / Giesinger, C. / Yuhnke, B. (2016): *NMC Horizon Report 2016*. Museum Edition. Austin: 42-43

Hohmann, Georg (2014): *„Deutsches Museum Digital"* in: Deutscher Museumsbund (Hrsg.): Museumskunde 79: 24-28.

Kommunikationsraum (2018): *Welche Möglichkeiten bieten Virtual Reality und Augmented Reality für das Museum der Zukunft?* https://www.kommunikationsraum.net/ausstellungenmuseen/welche-moeglichkeiten-bieten-virtual-reality-und-augmented-reality-fuer-das-museum-der-zukunft/ [letzter Zugriff 14. Oktober 2018

Museum4punkt0 (2018): *Über uns*. http://www.museum4punkt0.de/ueber-uns/ [letzter Zugriff 14. Oktober 2018]]

HistoGIS: Vom Punkt zur Fläche in Raum und Zeit

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
:sterreichische Akademie der Wissenschaften, Osterreich

Andorfer, Peter
peter.andorfer@oeaw.ac.at
:sterreichische Akademie der Wissenschaften, Osterreich

Geographische Angaben können in historischen Kontexten nicht als simple 2-dimensionale Daten (Längen- und Breitengrade) verstanden werden. Punkte die auf der Karte nur wenige Kilometer voneinander entfernt sind waren vor 500 Jahren wegen geographischer Hürden (Berge, Schluchten, Flüsse etc.) vielleicht ewig weit voneinander entfernt. Ähnliches gilt für politische Grenzen: Vor 30 Jahren waren Orte in Deutschland die heute wenige Autominuten voneinander entfernt liegen durch den eisernen Vorhang voneinander getrennt. Kamzelak (2018) hat es so formuliert: "Orte haben eine historisch-politische Dimension, die bei einer übergreifenden Registererfassung erst sichtbar zu einem Problem wird. [...] Für die Visualisierung von Briefen etwa sind historische Karten ein Desiderat; generell auch Geodaten für Flächen. Und alle mit Geodaten versehenen Einträge müssen mit einem Zeitstempel kombiniert sein, denn beispielsweise die Altstadt von Jerusalem ist eben heute nicht am selben Ort wie vor 2.000 Jahren.".

Trotzdem arbeitet eine Vielzahl an Digital Humanities Projekten auch heute noch mit 2-dimensionalen geographischen Angaben. Unserer eigenen Erfahrung nach liegt das Hauptsächlich an der Verfügbarkeit von Daten und Services. Moderne Punkt-Daten können einfach als geonames, openstreetmap etc. dump heruntergeladen werden bzw. über API Schnittstellen abgefragt werden. Für historische Daten existieren diese Services noch nicht vollumfänglich. Mit Pelagios gibt es ein Ökosystem an Services für Ortsdaten in der antiken Welt (http://commons.pelagios.org/), das im Zuge des "World-Historical Gazetteer" Projektes (http://whgazetteer.org/) auch in jüngere Zeiten ausgedehnt wird. Abgesehen von diesen notwendigen Initiativen und sehr wertvollen Datensätzen fehlen den Digital Humanities immer noch Polygondaten zu politischen Entitäten im Verlauf der Zeit.

HistoGIS hat sich zum Ziel gesetzt genau diese Lücke zu füllen in dem:

- ein Repository historischer Polygondaten mit einheitlichen Metadaten aufgebaut und zum Download angeboten wird. Für dieses Repository werden zunächst schon existierende Polygondaten eingesammelt, aufbereitet und erst in einem zweiten Schritt für historisch wichtige Zeitspannen neue Polygone erstellt. Dabei setzt sich HistoGIS zunächst zum Ziel die Periode zwischen dem ausgehenden 18. Jhdt. und 1918 für das Gebiet der KuK Monarchie und des deutschen Bundes abzudecken.
- und RestAPI Services zur einfachen Anreicherung historischer Daten mit Hilfe des Repositories angeboten werden. Z.B.: Wo war Punkt X/Y zum Zeitpunkt Z[1]? Die API antwortet mit den Metadaten zu den einzelnen überlappenden Polygonen die für den Zeitraum (oder Zeitpunkt) Gültigkeit haben. Nehmen wir an: Ein Projekt hat Reiseberichte in seiner Datenbank. Eine Station war Bolzano 1910. Die Geokoordinaten (46.49067, 11.33982) wurden über Geonames gefunden. Schickt man diese mit dem Jahr an die API bekommt man die Metadaten für Bozen Stadt, An der Etsch, Tirol und Österreich-Ungarn zurück[2].

Eine Fokussierung auf die politischen Verwaltungseinheiten und ihre Grenzen erlaubt es in vielen Bereichen - zumindest für jüngere Zeiten - einen Mix aus modernen Daten/Methoden und historischen zu verwenden ohne historisch falsche Daten zu generieren. So können bei oben angeführten Beispiel Bozen die Google Maps API, Geonames oder Openstreetmap zur Geolokalisation verwendet werden (die Geokoordinaten von Bozen haben sich ja nicht geändert) und die HistoGIS API um die Eingliederung in die Verwaltungshierarchien zu verbessern (Bozen war 1910 Teil der KuK Monarchie). Dieses Vorgehen reduziert nicht nur den Aufwand für die Erstellung/Kuratierung der Daten drastisch, es ermöglicht es auch leichter Punkte denen schon Längen- und Breitengrade zugewiesen wurden politisch/historisch zu verorten.

Datenmodell, Technische Grundlage und Workflow

Die Modellierung historischer Verwaltungsräume hinsichtlich ihrer räumlichen und zeitlichen Ausdehnungen erfolgte bewusst in einer äußerst vereinfachten Art und Weise. Das mittels Python (bzw. GeoDjango) definierte und als Postgresql implementierte Datenmodell besteht in seinem Kern aus den drei Hauptklassen bzw. Tabellen "TempSpatial", "Source" und "TempSpatialRel"

Ein historischer Verwaltungsraum (Temporalized Spatial Entity oder eben "TempSpatial") definiert sich über die im gesamten Datenset einzigartige Kombination der Eigenschaften zeitliche Ausdehnungen ("start_date" und "end_date", Datumsfelder), räumliche Ausdehnung ("geom"; Multipolygon) sowie einem Feld "date_accuracy", welches Auskunft über den Grad der Genauigkeit der angegebenen Datumswerte gibt. Ergänzt wird diese Klasse um die Eigenschaften "name" für, entsprechend den Projektkonventionen einen zeitgenössischen Namen der Entität, "alt_names" für alternative Namen sowie einem Feld "additional_data", welches das Speichern arbiträrer weiterer Daten im JSON Format ermöglicht.

Das Feld "administrativ_unit" zeigt auf eine Hilfsklasse ("SkosConcept") für kontrolliertes Vokabular, welche in weiten Teilen das SKOS Datenmodell implementiert.

Die Quelle jeder Instanz der Klasse TempSpatial bzw. jeder im Datenset erfassten historischen Verwaltungseinheit wird mit Hilfe der Klasse "Source" beschrieben. Darin werden URLs zu verwendeten Daten anderer Projekte gespeichert wie auch eine Beschreibung der (weiteren) Datenerhebung und Kuration im Rahmen des Projektes sowie ein Zitationsvorschlag. Jedes Source Objekt ist außerdem auch mit einem ESRI-Shapefile verbunden welches Projektintern als primäres Datenformat dient. Mehr dazu im Abschnitt Workflow.

Die Modellierung beliebiger Relationen zwischen beliebigen historischen Verwaltungsräumen ist im Datenmodell durch die Klasse TempSpatialRelation grundgelegt. Hier können jeweils zwei TempSpatial Objekte ("instance_a" und "instance_b") für eine Zeitspanne ("start_date" und "end_date") in eine typisierte (Verweis auf die bereits erwähnte Hilfsklasse "SkosConcept") Relation gebracht werden. Hierbei ist jedoch weniger an die in Gazetteern üblichen "part of" Beziehungen gedacht, sondern an Relationen wie beispielsweise "ist Vorgänger von" oder "wurde zusammengelegt mit". Allerdings muss darauf hingewiesen werden, dass im derzeitigen Status des Projektes noch keine derartigen Beziehungen erfasst werden.

Die Art und Weise wie eben erwähnte "part of" Beziehungen erfasst werden sollen, wurde im Projekt ausgiebig diskutierte, wobei hier neben formal- konzeptionellen Argumenten vor allem auch die konkreten Arbeitsvoraussetzungen und -bedingungen im Projekt berücksichtigt werden mussten.

Schlussendlich wurde eine explizite Modellierung hierarchischer Strukturen der erfassten Verwaltungseinheiten verzichtet, sprich im Datenmodell wird nicht ausdrücklich festgehalten, dass z.B. die Verwaltungseinheit A für einen gegebenen Zeitraum, Teil der Verwaltungseinheit B und diese Teil der Verwaltungseinheit D war. Dies erscheint deshalb als zulässig, weil im Projekt von der Prämisse ausgegangen wird, dass, zumindest für den für das Projekt primär interessante (Zeit)Raum, die Fläche der übergeordnete Einheit stets die Summe aller ihr untergeordneten Einheiten bildet. Dies ermöglicht es, dass part-of Beziehungen zwischen TempSpatial Objekten 'on the fly' mit Hilfe von spatial queries und unter Berücksichtigung der jeweiligen Start- und Enddatumswerten berechnet werden können. Und dies wiederum erleichtert einerseits die Arbeit der DatenkuratorInnen im Projekt, da diese keine expliziten Verbindungen pflegen müssen, andererseits ermöglicht dieses Flexibilität die Integration anderer Datensatz mit relativ geringem Aufwand.

An dieser Stelle muss jedoch betont werden, dass die bis dato kuratierte Menge an Daten noch nicht ausreicht um konkrete Aussagen hinsichtlich der Belastbarkeit des hier vorgestellten Datenmodells treffen zu können. Erste Tests diesbezüglich fielen jedoch durchwegs positiv aus, sowohl was die Genauigkeit der spatial queries, vor allem aber auch was deren Performanz betrifft.

Die technischen Komponenten des Projektes sind überschaubar. Als Storage Layer fungiert eine Postgresql Datenbank mit PostGIS erweiterung. Die Interaktion damit erfolgt über einen mittels dem Python basierten Webframework (Geo)Django implementierten Applikation Layer, wobei mit (Geo)Django, respektive django-rest-framework sowohl die HistoGIS-Webapplikation als auch ein entsprechender REST Webservice implementiert wurde bzw. wird.

Die eigentlich Datenkuration erfolgt davon völlig unabhängig und unter Verwendung der open source Software Qgis. Bis dato wurden damit vorwiegend bereits existierende Daten harmonisiert und als ESRI shapefiles gespeichert. Das Schema dieser Dateien entspricht dabei weitgehend dem oben skizzierten Datenmodell. Als 'fertig' erachtete Datensätze (gezippte Shapefiles) werden dann über ein Webformular in HistoGIS als sogenannte "Source" objekte hochgeladen, entpackt, die Features der Shapfiles als TempSpatial Objekte gespeichert und mit dem Source Objekt verknüpft.

Neben der Kuration und Harmonisierung bereits bestehender "Vektordaten" werden im Projekt aber auch selbst Daten erzeugt. Dazu wählt ein Historiker im Team verwertbare (historische) Karten aus, welche idealerweise bereits digitalisiert (gescannt) sind. Die Datakuratorinnen georeferenziert diese Scans (geotiffs) und extrahieren die darin auffindbaren Informationen zu historischen Verwaltungsgrenzen als Vektordaten.

Kartenmaterial bis dato und Ausblick

Bis dato befinden sich knapp 4000 Polygone in der Production Instanz des Systems. Das Anpassen und Einspielen schon vorhandener Polygone wurde mit Daten aus dem Census mosaic Projekt (https://censusmosaic.demog.berkeley.edu/) und dem HGis Archiv (http://www.hgis-germany.de/) begonnen. Damit können große Teile des 19. Jhdts für die Gebiete Österreich-Ungarns und des Deutschen Bundes (inkl. der jeweiligen Nachfolgeentitäten) bereits abgedeckt werden. Die Daten werden nicht nur technisch aufbereitet, sondern auch inhaltlich von einem Verwaltungshistoriker überprüft. HistoGIS implementiert dafür ein Ampelsystem. Vom HistoGIS-Team technisch wie inhaltlich überprüfte Daten werden mit Grün markiert, vom HistoGIS-Team ausgewählte Daten die noch nicht überprüft wurden mit Gelb und von Usern zur Verfügung gestellte, nicht überprüfte Daten mit Rot (momentan befinden sich lediglich gelbe und grüne Daten im System).

Die Aufbereitung der Daten, wie auch die Entwicklung des technischen Systems schreitet erfreulicher Weise schneller voran als geplant. Es wurde deshalb erst kürzlich beschlossen in HistoGIS schon während der Projektlaufzeit auch Daten außerhalb der geplanten räumlich-zeitlichen Grenzen aufzunehmen.

In unserer Präsentation werden wir vor allem das Datenmodell und die RestAPI Schnittstellen des Systems diskutieren und vorstellen.

Fußnoten

1. Eine auf diese API aufbauende Abfragemaske findet sich hier: https://histogis.acdh.oeaw.ac.at/shapes/where-was/
2. Beispielhaft das Polygon für Tirol: https://histogis.acdh.oeaw.ac.at/shapes/shape/detail/3352

Bibliographie

Kamzelak, Roland S. (2018): *"Von der Raupe zum Schmetterling oder Wie fliegen lernen – Editionsphilologie zwischen Infrastruktur und Semantic Web."* In: **Kamzelak, Roland S / Steyer, Timo (eds.)**: *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). text/html Format. DOI: 10.17175/sb002_004 [letzter Zugriff 28. September 2018]

Nüssli, Marc-Antoine / Nüssli, Christos (2017): *"A formal model for historical atlases and historical knowledge"*, http://

www.academia.edu/35853762 [letzter Zugriff 28. September 2018]

Historic Building Information Modeling (hBIM) und Linked Data – Neue Zugänge zum Forschungsgegenstand objektorientierter Fächer

Kuroczyński, Piotr
piotr.kuroczynski@hs-mainz.de
Hochschule Mainz – University of Applied Sciences

Brandt, Julia
julia.brandt@hs-mainz.de
Hochschule Mainz – University of Applied Sciences

Jara, Karolina
karolina.jara@hs-mainz.de
Hochschule Mainz – University of Applied Sciences

Grosse, Peggy
p.grosse@gnm.de
Germanisches Nationalmuseum

Mit der Entwicklung der Computergrafik seit den 1960er, und explizit seit den 1990er Jahren, wurde die virtuelle Rekonstruktion für die objekt- und raumbezogene Forschung entdeckt und eingesetzt (Messemer, 2016). Drei Dekaden nach ihrer Popularisierung stellen wir fest, dass der Einsatz der 3D-Modellierung und Visualisierung vorrangig der Vermittlung in Form altbewährter Bildpublikationen bzw. Filmanimationen erfolgt. Die vielfältigen Möglichkeiten des Computers werden in Folge einer fehlenden digital-orientierten Methodik und Infrastruktur, allen voran der wissenschaftlichen Dokumentation und Publikation der Ergebnisse, nicht ausgeschöpft.

Dabei stellt die 3D-Retrodigitalisierung vorhandener Artefakte infolge der 3D-Laserscanner und Photogrammetrie sowie die quellen-basierte, hypothetische 3D-Rekonstruktion physisch nicht (mehr) vorhandener Objekte einen adäquaten Zugang zum Forschungsobjekt in der Archäologie, Kunst- und Architekturgeschichte sowie der Denkmalpflege dar. Als Repositorien für die wissenschaftlichen 3D-Modelle bieten sich in erster Linie die Universitätsbibliotheken an, sodass über eine metadatenbasierte Kontextualisierung der 3D-Datensätze nachgedacht wird (Blümel, 2013), auch wenn noch keine Institution bereit ist die Ergebnisse aus 3D-Rekonstruktionen als wissenschaftlichen Sammlungsbestand aufzunehmen und langfristig zur Verfügung zu stellen.

Das Potenzial der 3D-Modelle für die objektorientierte Forschung steckt zum einen in der genauen Wiedergabe der geometrischen und materiellen Eigenschaften eines Objekts. Zum anderen geht mit der vorausgehenden, tiefgreifenden Interpretation der Quellen und der kreativen, hypothetischen Nachbildung des Objekts ein umfangreiches Objektverständnis beim Autor der Rekonstruktion einher (Favro, 2012). Der Mehrwert eines digitalen 3D-Modells beginnt für die Wissenschaft mit der nachhaltigen Verknüpfung der wissenschaftlichen Fragestellungen mit den Quellen, ihrer Interpretation und den daraus resultierenden Ergebnissen in mensch- und maschinenlesbarer Form (Kuroczyński, 2017).

In Anlehnung an die Bauindustrie, welche als Antwort auf den einsetzenden digitalen Wandel seit Mitte der 1990er Jahre die Methode des *Building Information Modelling* (BIM) und das Datenaustauschformat *Industry Foundation Classes* (IFC) hervorgebracht hat, müssen sich die Digital Humanities erst noch auf eine digital-orientierte Methodik im Umgang mit den wissenschaftlichen 3D-Modellen einigen. Ohne eine Standardisierung des Datenmodells und des Austauschformats wird die Nachhaltigkeit, Interoperabilität und Nachvollziehbarkeit der digitalen 3D-Objekte, und somit der eigentliche Mehrwert der Modelle, nicht gewährleistet sein.

Als zukunftsweisende Technologie in diesem Kontext setzt sich die Formalisierung des Wissens in strukturierten Datenmodellen (Ontologien), die Vernetzung digitaler Ressourcen (Linked Data) und eine webbasierte interaktive Visualisierung von 3D-Datensätzen infolge der WebGL-Technologie durch. Die computergerechte Art der Formalisierung und Strukturierung des Wissens, wie sie bereits Anfang der 1980er Jahre seitens der Kunstgeschichte postuliert wurde (Heusinger, 1983), ermöglicht die Operationalisierung der Daten und fördert den computergestützten Erkenntnisgewinn und die web-basierte Wissensvernetzung. Im Bereich der Bauforschung und Denkmalpflege seien Projekte, wie *MonArch* (Freitag & Stenzer, 2017), und *SACHER* (Apollonio et al., 2017) zu nennen, die ein umfassendes und kollaboratives Management des (digitalen) Kulturerbes anbieten und innovative Web-3D-Viewer, wie *3DHOP* (http://3dhop.net/), einsetzen. Sie erlauben eine umfassende Dokumentation der Schadenskartierung und der Konservierungsarbeiten, sowie eine Kontextualisierung des Objekts mit weiteren Linked Data-Ressourcen.

Bei der quellen-basierten historischen Rekonstruktion sind allen voran Projekte zu nennen, welche den interpretativen Prozess der 3D-Modellieurng erfassen und die web-basierte Visualisierung nachvollziehbar machen. Projekte, wie *Digital 3D Reconstructions in Virtual Research Environments* (Kuroczyński, Hauck, & Dworak, 2016) oder *DokuVis* (Bruschke & Wacker, 2016), zeigen das Potenzial einer nachhaltigen Erfassung der Prozesse und der Verknüpfung der 3D-Datensätze mit Ereignissen, Quellen und Akteuren als Linked Data auf (Abb. 1).

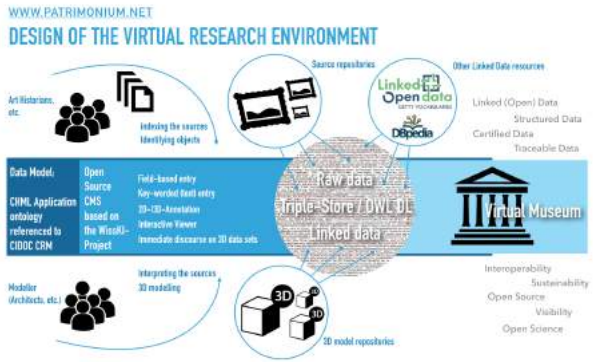


Abb. 1: Konzept einer web-basierten virtuellen Forschungsumgebung für quellen-basierte, digitale 3D-Rekonstruktionen; umgesetzt im Projekt *Digital 3D Reconstructions in Virtual Research Environments* , www.patrimonium.net (Copyright: Piotr Kuroczyński, 2016)

Für die Digital Humanities eröffnen diese Projekte einen neuen Zugang zum Wissen hinter den 3D-Modellen. Eine Grundvoraussetzung hierfür bieten mensch- und computerlesbare Datenmodelle, die eine digital vernetzte Abbildung des Wissens rund um die 3D-Modelle innerhalb einer Applikationsontologie erlauben. Die strukturierten, digitalen Forschungsdaten können in Folge mit der graphen-basierten SPARQL-Abfragesprache für RDF operationalisiert werden, sodass neue Erkenntnisse und Rückschlüsse aus explizitem und implizitem Wissen generiert werden können.

Basierend auf den Anforderungen und Erfahrungen der vorangegangenen Projekte möchte der Beitrag der Fragestellung nach der fehlenden digital-orientierten Methodik und mangelnder Infrastruktur in der wissenschaftlichen, quellen-basierten 3D-Rekonstruktion nicht mehr vorhandener Kunst und Architektur nachgehen. Hierbei wird untersucht, inwieweit bewährte Lösungen aus dem Bauwesen, welche eine fachübergreifende modell-basierte Dokumentation und Kommunikation mittels der IFC-Schnittstelle praktizieren, für die geisteswissenschaftlichen Fragestellungen objektbezogener Disziplinen nutzbar sind. Als Beispielobjekt wird dabei die 1938 zerstörte *Synagoge am Anger* in Breslau/Wrocław, im heutigen Polen, herangezogen. Im Zuge der Auseinandersetzung mit dem soziokulturellen Kontext des Synagogenbaus in der zweiten Hälfte des 19. Jahrhunderts soll die objektbezogene 3D-Modellierung innerhalb einer weitverbreiteten Open BIM-Software untersucht werden. Liegen bereits erste Erkenntnisse aus der BIM-konformen Modellierung von historischen Objekten vor (Murphy, 2012) und häufen sich die Leitfäden in dem Bereich des Bauwesens, insbesondere im *Bauen im Bestand* (Antonopoulou & Bryan, 2017), so fehlt es weiterhin an einer übergeordneten digital-orientierten Methodik, die über die 3D-Modellierung hinaus eine historische Untersuchung und Verankerung der Objekte in einem breiten geisteswissenschaftlichen Kontext ermöglicht (Abb. 2).

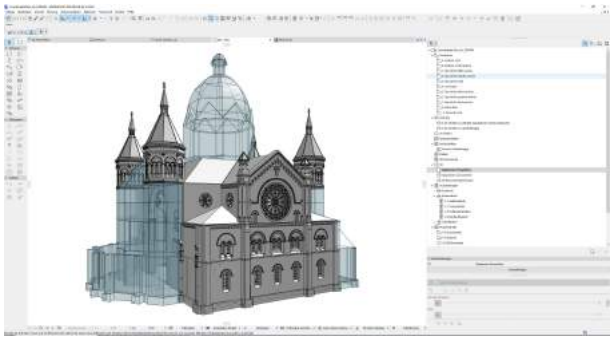


Abb. 2: Semantische Segmentierung *WestPart* und 3D-Modellierung innerhalb der Open BIM-/CAD-Software von ARCHICAD 21 (Copyright: Hochschule Mainz, 2018)

Anhand der *Synagoge am Anger* soll diese Methodik und eine anwendungsorientierte, virtuelle Forschungsumgebung (open source, basierend auf Wiss-KI, http://wiss-ki.eu/) für die Erfassung und Verarbeitung der digitalen Daten, sowie die Publikation der Forschungsergebnisse im Sinne von Open Access untersucht und vorgestellt werden. Die Arbeit behandelt zum einen die Fragen der Quellenerschließung, der Rechteverwaltung und Lizenzen der zu publizierenden Inhalte, zum anderen wird der klassische Arbeitsprozess (Workflow) einer digitalen 3D-Rekonstruktion unter den neuen Anforderungen einer wissenschaftlichen Dokumentation der Arbeitsprozesse im Sinne von seitens der London Charter proklamierten *Paradata* in Frage gestellt (Baker, 2012). Der Ansatz einer Verbindung von objektorientierter, BIM-konformer 3D-Modellierung mit den dehnbaren Abbildungsmöglichkeiten komplexerer geisteswissenschaftlicher Zusammenhänge zwischen den Objekten, Quellen, Ereignissen und Akteuren innerhalb einer flexiblen virtuellen Forschungsumgebung bietet eine neuen *virtuellen Forschungsraum* für die objektorientierten Fächer (Abb. 3).

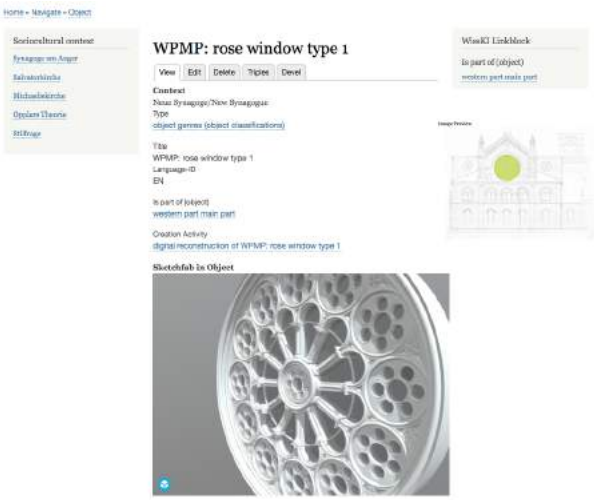


Abb. 3: Frontend vom Objekteintrag *WestPartMainPart: rose window type 1* innerhalb der virtuellen Forschungsumgebung zur digitalen Rekonstruktion der Synagoge von Oppler in Breslau, http://www.vfu-oppler.hs-mainz.de/ (Copyright: Hochschule Mainz, 2018)

Die Attributierung und Auszeichnung der BIM-Objekte und die Möglichkeit der Verknüpfung einzelner Elemente mit externen Daten, z.B. mit den Instanzen innerhalb einer virtuellen Forschungsumgebung, ermöglicht die seitens der Kunstgeschichte lang ersehnten digitalen *Fußnoten des Modells* (Hoppe, 2001). Die webbasierten BIM-Viewer, bspw. *Solibri Model Checker*, bieten heute einen einfachen Zugriff auf das Wissen hinter dem 3D-Datensatz, das weit über die reinen Geometrie- und Materialinformationen hinausreichen kann. Der vorgestellte Begriff eines *Historic Building Information Modelling* für parametrische Objekte (Murphy, 2012) bekommt durch die Erweiterung des Wissenshorizonts infolge der begleitenden Verknüpfung mit den Daten einer virtuellen Forschungsumgebung eine neue Dimension (Abb. 4).

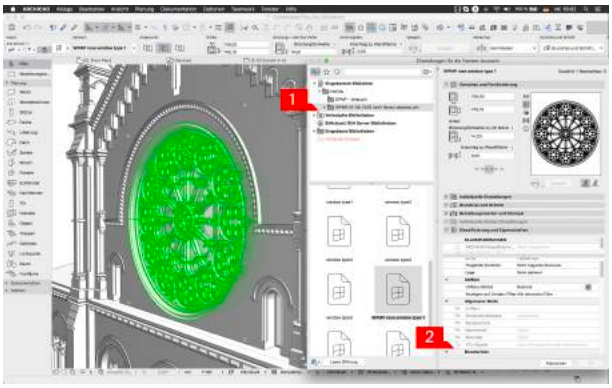


Abb. 4: Semantische Anreicherung und Verknüpfung der Objekte infolge der Attributierung unter den Objekteigenschaften; das Objekt *Rosette* als (1) Bibliotheksobjekt mit der attributierten (2) Verknüpfung zur Objektinstanz in der virtuellen Forschungsumgebung; die URL führt zur Obejktdarstellung gemäß Abb. 3 (Copyright: Hochschule Mainz, 2018)

Die Nutzung des Datenaustauschformats IFC als Träger von geisteswissenschaftlichen Informationen wird in diesem Kontext näher untersucht. Die Weiterentwicklung eines Datenmodells (Applikationsontologie) zur formalen Abbildung der Sachverhalte einer quellen-basierten, hypothetischen 3D-Rekonstruktion wird vorgestellt. Dabei lehnt sich die Datenmodellierung an das CIDOC Conceptual Reference Model (CRM), ISO 21127:2006, an und stellt weiterführende Fragestellungen an die Modellierung und Anbindung weiterer relevanter Link (Open) Data-Ressourcen. Hier stellt sich die Frage, inwieweit heute die Einbeziehung weiterer einschlägiger Ressourcen neue Möglichkeiten der Auswertung komplexer Sachzusammenhänge und implizierten Wissens unterstützen kann. Bekanntermaßen befinden sich die kontrollierten Vokabulare und Thesauri in der Entwicklung und werden mit den Herausforderungen der Übersetzung von Fachbegriffen konfrontiert. Eigene Lösungen zur Verwaltung von Klassifizierungen spielen weiterhin eine wichtige Rolle (Piotrowski, Colavizza, Thiery, & Bruhn, 2014) und werden in die Betrachtung mit einbezogen.

Im Zentrum des Beitrags steht die Dokumentation einer kreativ-interpretativen, quellen-basierten 3D-Rekonstruktion und die Sicherung der Wissenschaftlichkeit, indem die Arbeitsprozesse nachvollziehbar bleiben. Die projektbezogenen Ergebnisse lassen fundierte Erkenntnisse zu, die in einem breiten Blickfeld der technologischen Entwicklung gesetzt werden. Die Entwicklung der VR-/AR-/MR-Technologien führt zu neuen immersiven und interaktiven Erfahrungen, welche wiederum neue Zugänge zu den Forschungsergebnissen und neue Formen der Vermittlung mit sich bringen, die bei der Koordination fachübergreifender Abläufe im Bauwesen bereits eingesetzt werden (Abb. 5).

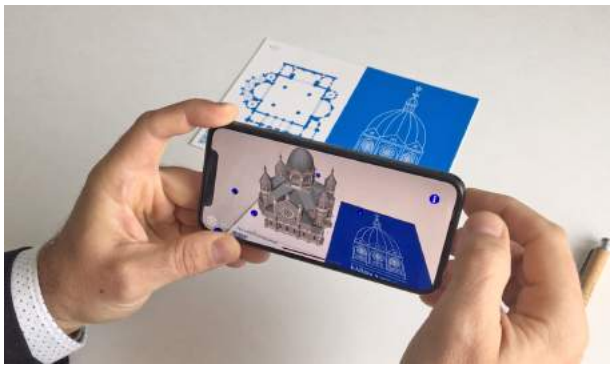


Abb. 5: AR-Anwendung „kARtka z Synagogą“ (https://arvr.hs-mainz.de/pl/) für die öffentlichkeitswirksame Vermittlung der Synagoge mittels Postkarte und Grundriss als Tracker; 2000-fach ausgeteilt während des Gedenkens zur 80. Jährung der Reichspogromnacht am 10.11.2018 in Breslau/Wrocław, Polen (Copyright: Hochschule Mainz, 2018)

Wir befinden uns am Anfang eines technologischen Prozesses, bei dem die Digital Humanities im Bereich der Handhabung und Nutzbarmachung der 3D-Datenstätze von vorausgehenden Branchen lernen können. Wir stecken noch in *Kinderschuhen* der Digitalisierung und sind am Anfang einer langwierigen inhaltlich-methodologischen Etablierung digitaler 3D-Rekonstruktion als einer bewährten Forschungsmethodik. Erst wenn unsere 3D-Modelle semantisch strukturiert, langfristig und offen im Netz zur Verfügung gestellt werden können, können zukunftsweisende Anwendungen wie *3D Wikipedia* (Russell, Martin-Brualla, Butler, Seitz, & Zettlemoyer, 2013) sowie ein daraus resultierender Diskurs im Sinne von Open Science an den Modellen geführt werden. Der vorliegende Beitrag möchte eine Diskussion und Entwicklung in diese Richtung anstoßen.

Bibliographie

Antonopoulou, S., & Bryan, P. (2017): *BIM for Heritage - Developing a Historic Building Information Model. Historic England.*

Apollonio, F. I., Rizzo, F., Bertacchi, S., Dall'Osso, G., Corbelli, A., & Grana, C. (2017): *SACHER: Smart Architecture for Cultural Heritage in Emilia Romagna*. In **C. Grana & L. Baraldi (Eds.)**: *Digital Libraries and Archives (Vol. 733, pp. 142–156)*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68130-6_12

Baker, D. (2012): *Defining Paradata in Heritage Visualization. In Paradata an Transparency in Virtual Heritage (pp. 163–175)*. Ashgate.

Blümel, I. (2013): *Metadatenbasierte Kontextualisierung architektonischer 3D-Modelle*. Berlin.

Bruschke, J., & Wacker, M. (2016): *Simplifying Documentation of Digital Reconstruction Processes*. In **S. Münster, M. Pfarr-Harfst, P. Kuroczyński, & M. Ioannides (Eds.)**: *3D Research Challenges in Cultural Heritage II (Vol. 10025, pp. 256–271)*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47647-6_12

Favro, D. (2012): *Se non e vero, e ben trovato (If Not True, It Is Well Conceived): Digital Immersive Reconstructions of Historical Environments*. Journal of the Society of Architectural Historians, 71(3), 273–277. https://doi.org/10.1525/jsah.2012.71.3.273

Freitag, B., & Stenzer, A. (2017): *MonArch – A Digital Archive for Cultural Heritage. In Das Digitale und die Denkmalpflege: Bestandserfassung - Denkmalvermittlung - Datenarchivierung - Rekonstruktion verlorener Objekte (Vol. 26, pp. 122–129)*. Heidelberg: arthistoricum.net.

Heusinger, L. (1983): *Kunstgeschichte und EDV: 8 Thesen*, 11(4). https://doi.org/10.11588/kb.1983.4.9808

Hoppe, S. (2001): *Die Fußnoten des Modells. CAD-Modelle als interaktive Wissensräume am Beispiel des Altenberger-Dom-Projektes*. In **M. Frings (Ed.)**: *Der Modelle Tugend. CAD und die neuen Räume der Kunstgeschichte* (pp. 87–102). Weimar: VDG.

Kuroczyński, P. (2017): *Virtual Research Environment for digital 3D reconstructions – Standards, thresholds and prospects*. Studies in Digital Heritage, 1(2), 456. https://doi.org/10.14434/sdh.v1i2.23330

Kuroczyński, P., Hauck, O., & Dworak, D. (2016): *3D Models on Triple Paths - New Pathways for Documenting and Visualizing Virtual Reconstructions*. In **S. Münster, M. Pfarr-Harfst, P. Kuroczyński, & M. Ioannides (Eds.)**: *3D Research Challenges in Cultural Heritage II (Vol. 10025, pp. 149–172)*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47647-6_8

Messemer, H. (2016): *The Beginnings of Digital Visualization of Historical Architecture in the Academic Field*. In **S. Hoppe & S. Breitling (Eds.)**: *Virtual Palaces, Part II – Lost Palaces and their Afterlife. Virtual Reconstruction between Science and Media (Vol. 3)*. Heidelberg: arthistoricum.net.

Murphy, M. (2012): *Historic Building Information Modelling (HBIM) for Recording and Documenting Classical Architecture in Dublin 1700 to 1830*. Dublin.

Piotrowski, M., Colavizza, G., Thiery, F., & Bruhn, K.-C. (2014): *The Labeling System: A New Approach to Overcome the Vocabulary Bottleneck*. In DH-CASE '14 DH-CASE II: Collaborative Annotations on Shared Environments: metadata, tools and techniques in the Digital Humanities (Vol. Article No. 1).

Russell, B. C., Martin-Brualla, R., Butler, D. J., Seitz, S. M., & Zettlemoyer, L. (2013): *3D Wikipedia: using online text to automatically label and navigate reconstructed geometry*. ACM Transactions on Graphics, 32(6), 1–10. https://doi.org/10.1145/2508363.2508425

Interaktion im öffentlichen Raum: Von der qualitativen Rekonstruktion ihrer multimodalen Gestalt zur automatischen Detektion mit Hilfe von 3-D-Sensoren

Mukhametov, Sergey

s.mukhametov@gmail.com
Institut für Geoinformatik, WWU Münster, Deutschland

Kesselheim, Wolfgang

wolfgang.kesselheim@ds.uzh.ch
Sprache und Raum Laboratorium, Universität Zürich, Schweiz

Brandenbeger, Christina

c.brandenberger@access.uzh.ch
Sprache und Raum Laboratorium, Universität Zürich, Schweiz

Wenn Menschen miteinander interagieren, ist dies für kompetente Gesellschaftsmitglieder selbst aus einem gewissen Abstand heraus leicht zu erkennen: Die Körper von Interaktionsbeteiligten sind auf spezifische Art und Weise aufeinander ausgerichtet. Sie bilden einen "Interaktionsraum". Die Herstellung, Aufrechterhaltung und Auflösung von Interaktionsräumen sind in Soziologie, Psychologie und Linguistik untersucht worden – das in den letzten Jahren vor allem basierend auf Videoaufnahmen authentischer Interaktionsereignisse. Dabei basieren die Erkenntnisse auf feinkörnigen qualitativen Analysen, also auf einer sehr zeitintensiven Auseinandersetzung mit Einzelfällen.

Ziel unseres Vortrags ist es nun, zu zeigen, wie sich die bisherigen Ergebnisse der qualitativen Forschung zu Interaktionsräumen nutzen lassen, um Interaktionsräume automatisch detektieren zu können. Hierfür erheben wir multimodale Datensätze aus 2-D-Videos, Annotationen und 3-D-Sensordaten. Diese machen es m̈oglich, pr̈azise Informationen zu Bewegungen, Verhalten, und sozialen Interaktionen einer gr̈oßeren Anzahl von Probanden automatisiert und ber̈uhrungsfrei zu erheben und im r̈aumlichen Kontext zu analysieren. Während z.B. Computer-Vision-basierte Methoden zur Erkennung von Konversationsgruppen (z.B. Setti et al 2013) die Positionen einzelner Körperteile oft nur unzureichend erkennen, berechnet unsere Methode die präzisen Positionen und Distanzen von Skelett-Daten im 3-D-Raum.

Mit der Darstellung unserer Methode möchten wir exemplarisch aufzeigen, wie ethnografisch oder konversationsanalytisch arbeitende Studien quantitative Ansätze fruchtbar machen können. Unser Beispielfall ist die Untersuchung der körperlich-räumlichen Formationen,

die für die soziale Praxis des Museumsbesuchs charakteristisch sind. Mit der quantitativen Auswertung von Interaktionsräumen wird es möglich, im Vergleich zu den bisherigen Studien die Datenbasis beträchtlich zu erweitern, von der aus Generalisierungen zu den typischen Mustern und Merkmalen von Interaktionsräumen formuliert werden. Darüber hinaus lassen sich mit Hilfe der quantitativen Auswertung von körperlich-räumlichen Formationen schnell interessante Fälle der Exponatnutzung oder der Interaktion mit anderen Besuchern identifizieren, die dann qualitativ im Detail untersucht werden können.

Während diese technischen Methoden die Untersuchung von Interaktionsräumen erheblich beschleunigen, bergen sie auch Herausforderungen: die Integration der Daten zeitgleich aufnehmender Tiefenbildkameras, die Integration von geometrischen Körperprofilen in den raum-zeitlichen Ablauf, und die raum-zeitliche Analyse mehrerer Bewegungsprofile zur Identifikation von Gruppen und sozialen Interaktionen, und zur Definition von Interaktionsräumen. Zudem weichen die tatsächlichen Formationen interagierender Personen teilweise auch stark von den in der Literatur beschriebenen ab. Die automatische Erfassung macht es daher notwendig, die bestehenden Definitionen zu verschärfen und zu erweitern (aber liefert eben auch die Grundlagen für diese Arbeit an den Definitionen).

Schon in den 1970er Jahren haben sich Forscherinnen und Forscher aus Soziologie und Psychologie für die räumlichen Aspekte der sozialen Interaktion interessiert. So geht Goffman (1963, 1971) aus alltagssoziologischer Perspektive der Frage nach, unter welchen Bedingungen Interaktion in Gang kommen kann, wenn Menschen sich im öffentlichen Raum begegnen. Inspiriert durch Goffman, untersucht der Psychologe Adam Kendon (z.B. 1990 [1973, Ciolek / Kendon 1980) die räumlichen Bedingungen des Entstehens von Interaktion. Anders als Goffman, dessen Analysen auf direkter visueller Beobachtung und Feldnotizen beruhen, arbeitet Kendon erstmals mit Kameras und der detaillierten multimodalen Transkription und Annotation der aufgenommenen Daten. Dabei interessiert sich Kendon nicht nur für den Beginn der Interaktion, sondern zunehmend auch für deren Aufrechterhaltung durch die dynamische Abstimmung der Körperpositionen der Interaktionsteilnehmer im Raum. Zentral sind hier die Konzepte der "F-Formation" und des "o-space", die bis in die aktuelle Forschung hinein folgenreich geblieben sind. Mit "F-Formation" ist eine dynamische Konfiguration der Interaktionsteilnehmer im Raum gemeint, die die räumlichen Voraussetzungen für das gemeinsame Interagieren zur Verfügung stellt. In einer F-Formation bringen die Interaktionspartner ihre "transactional segments" zur Übereinstimmung, also die Raumbereiche, in denen ihr Körper mit der Umwelt in Kontakt treten kann. Durch eine F-Formation entsteht der "o-space": der Raum, der von den Interaktionsteilnehmern umständen wird, in den hinein ihr Sprechen und Gestikulieren gerichtet ist und dem deshalb ihre Aufmerksamkeit gewidmet ist. Alle Teilnehmer haben gleichberechtigten Zugang zu diesem Raum und grenzen ihn gegen die räumliche Umwelt aktiv ab.

Die Linguistik hat erst in den letzten Jahren begonnen, sich für das Zusammenspiel von Raum und Interaktion zu interessieren, parallel in der textlinguistischen Multimodalitätsforschung und in der linguistischen Gesprächsanalyse. Dabei baut die linguistische Gesprächsanalyse auf einer gut etablierten Forschung in der angelsächsischen Soziologie und Anthropologie auf (etwa Goodwin 1986 oder Heath 1986), in deren videobasierter Praxis die Relevanz des Raums schon seit den 1980er Jahren erkennbar geworden ist.

An dieser Tradition knüpft die linguistische Forschung zum "Interaktionsraum" an (s. etwa Mondada 2009) und präzisiert dabei Kendons Untersuchungen zu "F-Formations" und dem "o-space" in wichtigen Punkten. Zum einen differenziert sie den gemeinsamen Interaktionsraum in drei separate Räume, den "Wahrnehmungs-", den "Bewegungs-" und den "Handlungsraum" (Hausendorf 2010). Diese Räume entstehen durch jeweils eigene Koordinationsleistungen: die Koordination der Körperbewegungen im Raum, die der visuellen Wahrnehmung und das aufeinander abgestimmte soziale Handeln. Gleichzeitig arbeitet die linguistische Forschung heraus, wie die konkrete Gestalt des Interaktionsraums auf die Besonderheiten des gebauten Raums reagiert (Hausendorf / Kesselheim 2016) oder auf die spezifischen Notwendigkeiten der gemeinsam ausgeführten Aktivität (etwa die Nutzung bestimmter Objekte, vgl. Nevile et al. 2014). Unsere Methode der automatischen Erkennung von Interaktionsräumen setzt diese Präzisierungen um, indem sie zum einen separat Wahrnehmungs- und Bewegungsräume identifiziert, und zum anderen, indem sie die Rolle von für die Interaktion relevanten Objekten im Raum mitberücksichtigt. Beides ist, wie wir zeigen werden, für das Verständnis des Interaktionsgeschehens in unserem Museumssetting essenziell.

Datengrundlage unseres Vortrags ist ein Teilkorpus von 42 Stunden (1 Woche) Aufnahmen in einem Museum im Norden Deutschlands (Gesamtkorpus: 100 Wochen). Mithilfe von mehreren modifizierten Kinect-v2-Geräten wurden gleichzeitig von mehreren Standpunkten aus 2-D-Videoaufnahmen und 3-D-Tiefenbild-Sensordaten eines Teiles der Dauerausstellung aufgezeichnet. Um eine optimale Erfassung der Besucheraktivitäten zu ermöglichen, kalkulieren wir die Anordnung der Sensoren mit einem speziellen Verfahren, um Okklusion zu vermeiden. Die Nutzung von ToF-Kameras erlaubt es, mit einer zeitlichen Auflösung von 20 ms die 3-D-Koordinaten der Körperteile von Besuchern zu ermitteln, die während der Beobachtungszeit im Blickfeld der Sensoren erschienen sind.

Unsere Methode besteht in dem Tracking möglichst aller menschlicher Aktivitäten in einem mit Sensoren erfassten Innenraum und der nachfolgenden Analyse der entstehenden Tracking-Datensätze. Zunächst werden die Daten mehrerer Sensoren kombiniert. Aus den kombinierten und gefilterten Datensätzen werden die Trajektorien einzelner Besucher zusammengestellt, alle Stellen des Hovering-Verhaltens und Stopp-Positionen detektiert und diese mit den Positionen naheliegender Ausstellungsobjekte in Beziehung gesetzt. Auf der Grundlage der Positionen von Schultern, Becken und Köpfen der betreffenden Personen werden alle “transactional segments” (Kendon, s.o.) berechnet, um Überschneidungen zu finden und diese bestimmten Mustern zuzuordnen. Diese Muster werden dann weiter analysiert, um gemeinsame Interaktionsräume zu berechnen.

So lassen sich sowohl Gruppen von Besuchern definieren, die sich miteinander in Interaktion befinden, als auch körperlich-räumliche Bezugnahmen der Interagierenden auf Exponate im Raum. Dies erlaubt z.B. zu bestimmen, mit welchen Exponaten sich die Besucher auseinandersetzen und ob sie dies alleine oder gemeinsam tun. Aufgrund von einer Sammlung akkumulierter Interaktionsräume werden

schließlich die Konturen von typischen Interaktionsräumen eines Exponats berechnet (typische Distanz zum Exponat, Blickwinkel, Dauer und Fragmentierung der Betrachtung). Daraus ergeben sich wichtige Hinweise für die Ausstellungsgestaltung und Wissensvermittlung als auch neue Möglichkeiten zur Auswertung der Nutzung von Ausstellungen durch die Besucher.

Ein Vorteil der Methode besteht schließlich auch in Bezug auf ethische und rechtliche Fragen, die sich im Zusammenhang mit videobasierten Besucherstudien stellen. Die Analyse der Besucherinteraktion auf Grundlage der 3-D-Skelette und schematischen Visualisierungen der Interaktionsräume, die aus den Sensordaten berechnet worden sind, ist vom Gesichtspunkt des Persönlichkeitsschutzes deutlich weniger heikel als die Arbeit mit Video-Daten, die nur mit hohem technischen Aufwand anonymisiert werden können.

Zur Gliederung unseres Vortrags.

Zunächst werden wir die von der Interaktionsraum-Forschung herausgearbeiteten Merkmale beschreiben, die für die räumlichen Formationen von Interaktionsbeteiligten charakteristisch sind. Dann erläutern wir, wie wir diese Einsichten genutzt haben, um in unserem Korpus Interaktionsräume automatisch zu identifizieren. Dabei arbeiten wir heraus, worin unsere Methode den im Rahmen der Computer Vision entwickelten Methoden zur Interaktionsraumerkennung überlegen ist, sowohl in technischer Hinsicht als auch im Hinblick auf die Differenziertheit der Abbildung qualitativer Forschungsergebnisse in der mathematischen Modellierung der Interaktionsräume.

Unser Beitrag zur Forschung.

Die automatisierte Detektion von Interaktionsräumen ermöglicht es, die Analysebehauptungen der bisherigen Forschung zu Interaktionsräumen auf breiter Datenbasis zu überprüfen und zu konsolidieren. So zeigen unsere Analysen etwa, dass die Kendon'sche Beschreibung von F-Formations zu schematisch ist. Tatsächlich lassen sich in unseren Daten unterschiedliche Ausprägungen von Interaktionsräumen beobachten: beispielsweise solche, zu denen nicht alle Teilnehmer gleichberechtigten Zugang haben, oder solche, in denen Objekte die Position von 'Beteiligten' zugewiesen bekommen. Darüber hinaus kann die Liste der in einem Korpus detektierten Interaktionsräume von qualitativ Forschenden genutzt werden, um in ihrem Material schnell Fälle von 'unproblematischen' Interaktionsräumen zu identifizieren und ausgehend hiervon die Besonderheiten von weniger eindeutigen oder von den Erwartungen abweichenden Interaktionsräumen im Kontrast profilieren zu können.

Bibliographie

Ciolek, T. Matthew / Kendon, Adam (1980): *Environment and the Spatial Arrangement of Conversational Encounters*, in: Sociological Inquiry 50 (3-4), 237–271.

Ge, Weina / Collins, Robert T. / Ruback, R. Barry (2012): *Vision-based analysis of small groups in pedestrian crowds*, in: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 34(5)1003–1016. PMID: 21844622 https://ieeexplore.ieee.org/document/5989835 [letzter Zugriff 02. Januar 2019].

Goffman, Erving (1963): *Behavior in public places. Notes on the social organization of gatherings*. New York, NY: Free Press.

Goffman, Erving (1971): *Relations in public. Microstudies of the public order*. New York: Basic Books.

Goodwin, Charles (1986): *Gesture as a Resource for the Organization of Mutual Orientation*, in: Semiotica 62 (1-2) 29–49.

Hausendorf, Heiko (2010): *Interaktion im Raum. Interaktionstheoretische Bemerkungen zu einem vernachlässigten Aspekt von Anwesenheit*, in: **Deppermann, Arnulf / Linke, Angelika (eds.):** *Sprache intermedial. Stimme und Schrift, Bild und Ton*. Jahrbuch 2009 des Instituts für deutsche Sprache. Berlin: de Gruyter (Jahrbuch des Instituts für deutsche Sprache, 2009) 163–197.

Hausendorf, Heiko / Kesselheim, Wolfgang (2016): *Die Lesbarkeit des Textes und die Benutzbarkeit der Architektur. Text- und interaktionslinguistische Überlegungen zur Raumanalyse*, in: **Hausendorf, Heiko / Schmitt, Reinhold, Kesselheim, Wolfgang (eds.):** *Interaktionsarchitektur, Sozialtopographie und Interaktionsraum*. Tübingen: Narr Francke Attempto (Studien zur deutschen Sprache 72) 55–85.

Heath, Christian (1986): *Body movement and speech in medical interaction*. Cambridge: Cambridge University Press.

Hung, Hayley / Kŕose, Ben (2011): *Detecting F-formations as Dominant Sets*, in: International Conference on Multimodal Interfaces (ICMI) 231–238. homepage.tudelft.nl/3e2t5/HungKrose_ICMI2011.pdf [letzter Zugriff 02. Januar 2019].

Mondada, Lorenza (2009): *Emergent focused interactions in public places. A systematic analysis of the multimodal achievement of a common interactional space*, in: Journal of Pragmatics (41) 1977–1997. DOI: 10.1016/j.pragma.2008.09.019.

Qin, Zhen / Shelton, Christian R. (2012): *Improving Multi-target Tracking via Social Grouping*, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 1972–1978. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6247899 [letzter Zugriff 02. Januar 2019].

Nevile, Maurice / Haddington, Pentti / Heinemann, Trine / Rauniomaa, Mirka (eds.) (2014): *Interacting with objects. Language, materiality, and social activity*. Amsterdam: Benjamins.

Setti, Francesco / Hung, Hayley / Cristani, Marco (2013): *Group detection in still images by F-formation modeling: A comparative study*, in: International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) 1–4. https://ieeexplore.ieee.org/document/6616147 [letzter Zugriff 02. Januar 2019].

Interlinked: Schriftzeugnisse der klassischen Mayakultur im Spannungsfeld zwischen Stand-off- und Inlinemarkup in TEI-XML

Sikora, Uwe
sikora@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Deutschland

Gronemeyer, Sven
sgronemeyer@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland; La Trobe University, Department of Archaeology and History, Australien

Diehr, Franziska
f.diehr@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz

Wagner, Elisabeth
ewagner@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Prager, Christian
cprager@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Brodhun, Maximilian
brodhun@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Deutschland

Diederichs, Katja
katja.diederichs@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Grube, Nikolai
ngrube@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Die computergestützte Erforschung einer nur teilweise erschlossenen Schrift und Sprache wie im Falle der Hieroglyphenschrift der klassischen Mayakultur steht vor zahlreichen Herausforderungen, insbesondere bei der Erfassung der Komplexität von Schrift- und Bildzeugnissen. Gerade historisierende Wissenschaftsdisziplinen sind auf diverse Informationsquellen angewiesen, um ihre Untersuchungsgegenstände nicht bloß in der historischen Vergangenheit sondern auch in der modernen Gegenwartskultur zu vergesellschaften: Informationen zu ursprünglichen Aufstellungsorten von mit Hieroglyphen reliefierten Stelen, Angaben zum aktuellen Aufbewahrungsort dekorierter und beschriebener Keramiken oder jahrzehntealte Zeichnungen monumentaler Tempelinschriften - das Wissen nicht bloß über die historischen Kontexte sondern auch über die wissenschaftliche Arbeit mit und an den antiken Schriftzeugnissen durch Forscher und Sammler bildet den essentiellen Rahmen, um wissenschaftliche Aussagen und Hypothesen zu formulieren, zu überprüfen und zu plausibilisieren.

Die Grundlage dieses Rahmens bilden Modelle, die zum einen eine formalisierte Beschreibung der benötigten Informationen erlauben und zum anderen eben jene Informationen miteinander in Beziehung setzen. Hier stellen Ontologien und domänenspezifische (Daten-)Modelle unerlässliche Hilfsmittel und notwendige Werkzeuge dar, um Wissen über Objekte, die im Fokus des wissenschaftlichen Interesses stehen, einheitlich und vor allem aussagekräftig zu dokumentieren. Vor diesem Hintergrund verfolgt das Projekt 'Textdatenbank und Wörterbuch des Klassischen Maya' (TWKM)[1] das Konzept einer ontologisch-vernetzten Datenbeschreibung: Der antike Text als kulturgeschichtliches Artefakt und somit umfassendes Wissens- und Informationsobjekt wird in einzelne unterschiedliche Informationsbereiche unterteilt, die jeder für sich nach besonderen Anforderungen und eigenen Datenmodellen beschrieben aber aufeinander bezogen werden.

Um den Informationsgehalt der antiken Textressourcen differenziert in maschinenlesbarer Form zu beschreiben, werden verschiedene Informationsbereiche auf unterschiedlichen Ebenen voneinander abgegrenzt: Zunächst werden die Schriftträger anhand eines ontologisch-basierten Metadatenschemas in RDF erfasst: Hier werden Kerninformationen zum Schriftträger (Bezeichnung, Zustand, Material und Technik, Maße etc.), seines archäologischen Kontexts sowie auch darüber hinausgehend historische Ereignisse und Persönlichkeiten der Maya-Kultur dokumentiert (Textdatenbank und Wörterbuch des Klassischen Maya 2017). Die Auszeichnung der textlichen Informationen wird separat in TEI-P5 Dokumenten vorgenommen, die mit den entsprechenden in RDF dokumentierten Ressourcen über persistente URIs (Uniform Resource Identifier) verknüpft werden.

Für die Auszeichnung der etwa 10.000 Maya-Texte dient ein projektspezifisches TEI-P5 Anwendungsprofil. Das TWKM unterscheidet hier zwischen den drei Informationsbereichen (1) Form, (2) Inhalt und (3) linguistische Analyse, die nach jeweils spezifischen Anforderungen separat erschlossen werden. So wird die Erschließung der Form bzw. der Texttopographie, d.h. der strukturellen Anordnung von Schrift- und Bildbereichen auf dem Schriftträger, unabhängig von der linguistischen Analyse und den in diesem Rahmen verwendeten Beschreibungskategorien durchgeführt.

Das TWKM bedient sich hier am Konzept des sog. Stand-off-Markups[2]: die individuierte Auszeichnung von Informationen, die durch Verweise auf andere ausgezeichnete Informationen virtuell[3] in einen gemeinsamen Zusammenhang gebracht werden. Durch diesen Ansatz kann nicht nur die Komplexität der Auszeichnung einzelner Informationsbereiche individuell angehoben bzw. abgesenkt werden. Auch der direkte Einfluss struktureller Anforderungen des XML-Formats auf die Auszeichnung, die in der Praxis häufig zu Problemen und Herausforderungen führen (z.B. die Wohlgeformtheitsregel und die hiermit einhergehende Maßgabe, dass XML-Elemente sich nicht überschneiden dürfen), wird hierdurch minimiert.[4]

Der Inhalt eines Textes lässt sich somit kohärent d.h. in einem logisch-thematischen Zusammenhang beschreiben, obwohl seine topographische formal-strukturelle Anordnung einer gänzlich anderen Logik folgt: z.B. kann ein Text topographisch in geflochtener Form arrangiert und dementsprechend maschinenlesbar beschrieben werden. Die inhaltlich-logische Struktur des Texts wird separat gemäß ihren eigenen Ordnungsprinzipien beschrieben aber mit den topographischen Strukturen in Beziehung gesetzt:

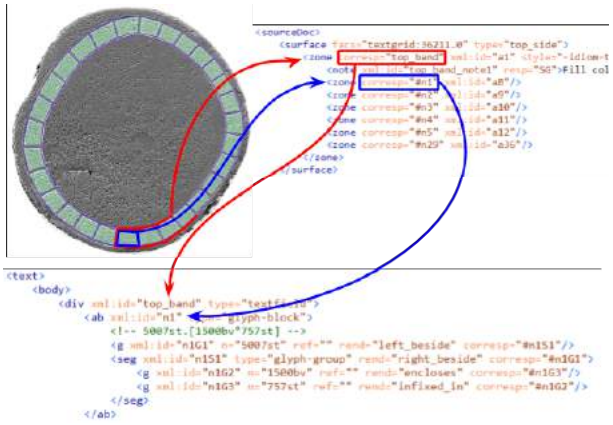


Abbildung 1. Virtueller Zusammenhang zwischen Bildbereich (oben links), Texttopographie (oben rechts) und Inhalt (unten)

Zunächst werden die texttopopgraphischen Eigenschaften des Schriftträgers (tei:sourceDoc) beschreiben. In diesem Kontext werden unterschiedliche Oberflächen (tei:surface) erfasst (z.B. die Flächen einer rechteckigen Stele, die Innen- und Außenseite einer Vase oder Vorder- und Rückseite eines Kodexblattes) und mit digitalen Faksimiles verbunden, die zuvor aus analogen Repräsentationen oder direkt vom Objekt, etwa über 3D-Scanning, erstellt wurden.[5] Diese Oberflächen enthalten einen bis mehrere Textbereiche (tei:zone), die mit Hilfe des Text-Bild-Link-Editors (TBLE) des TextGrid-Laboratorys (Neuroth / Rapp / Söring 2015) mit einzelnen Bereichen des Faksimiles assoziiert wurden. In diesem Rahmen können weitere Informationen, wie bspw. Orientierung oder Ausmaße erhoben werden. Die inhaltliche Erschließung der einzelnen Hieroglyphenblöcke und Schriftzeichen werden wiederum separat in tei:body erfasst.

Die Beschreibung des logo-sylabischen Schriftsystems des klassischen Maya ist aufgrund ihrer Eigenarten und Komplexität eine herausfordernde Aufgabe. Nicht nur aufgrund ihrer vergleichsweise jungen Entzifferungsgeschichte seit den 1950er Jahren gibt es noch eine Reihe von Desiderata bei Lesbarkeit und Verständnis des Schriftsystems. Trotz verschiedener Zeichenkataloge ist die genaue Anzahl von Zeichen noch immer nicht gesichert, damit auch nicht, wie viele Zeichen nicht oder nur unzureichend entziffert sind. Die Inventarisierung war auch bisher eine Herausforderung wegen multipler Klassifizierungsansätze der Schriftzeichen, etwa über die Form (Thompson 1962) oder die Ikonizität (Macri / Looper 2003). Das Projekt hat erstmals eine Systematik anhand einer taxonomischen Beschreibung der Bildung von Graphemvarianten entwickelt und trennt auf diesem Wege auch das bedeutungstragende Zeichen von seiner graphischen Repräsentation. Die wissenschaftliche Deutung des Zeichens, seine linguistische Information, wird dabei über ein System von Kriterien, die durch Aussagenlogik miteinander verbunden sind, qualitativ bewertet. Darauf basierend wird die Plausibilität der jeweiligen Entzifferungshypothese automatisiert eingestuft. Aufgrund der Dynamik der Entzifferungsarbeit ist es demnach nicht möglich, Transliterationen als Basis für die Textauszeichnung zu benutzen (Diehr et al. 2017: 1191-1192).

Deshalb und wegen des Umstands, dass es innerhalb der Mayaforschung keinen Konsens zum Umgang mit den derzeitigen Unicode-Vorschlägen bzw. -Implementierungen zur Mayaschrift gibt (Pallan Gayol / Anderson 2018), verfolgt das TWKM einen eigenen Ansatz zur Schriftbeschreibung gemäß der TEI-P5 Richtlinien.[6] Die Graphe (tei:g) sind in nahezu rechteckigen Blöcken angeordnet (tei:ab[@type='glyph-block']), die ungefähr einem Wort entsprechen. Jedes Zeichen wird mit einer URI-Referenz (@ref) versehen, die auf die konkrete Graphrepräsentation im Zeichenkatalog verweist. Aufgrund der komplexen Graphematik können einzelne Graphe wiederum zu Gruppen im Block zusammengefasst werden (tei:seg[@type='glyph-group']), z.B. bei einer Infigierung. Über weitere Attribute (@corresp und @rend) wird so die Position jedes einzelnen Graphs im Block und in Relation zu den anderen Graphen eindeutig beschrieben.

So kann der hieroglyphische Inhalt eines Texts maschinenlesbar dokumentiert werden, wobei die Entzifferungsgeschichte und hiermit einhergehende Veränderungen und Reinterpretationen einzelner Schriftzeichen auf der Ebene des Zeichenkatalogs abgebildet werden: Sollte sich die Interpretation eines Zeichens beispielsweise hinsichtlich seiner Semantik im Verlauf der fortschreitenden Forschung ändern, so bleibt die hieroglyphische Auszeichnung aller im TWKM erschlossenen Mayatexte unangetastet. Die Veränderung muss lediglich im Zeichenkatalog dokumentiert werden und steht danach als Information für alle Texte zur Verfügung. Gleiches gilt für die Entzifferungsgeschichte eines spezifischen Texts: Sollte die Deutung eines Graphems in einem konkreten Maya-Text revidiert werden, so erfolgt diese Änderung im Zeichenkatalog. Neu klassifizierte Zeichen werden im Zeichenkatalog durch die Relation owl:sameAs[7] mit den betreffenden falsch klassifizierten Zeichen verbunden. Damit ist garantiert, dass vormals falsch klassifizierte Zeichen weiterhin über deren URI auffindbar sind.[8] Die im Korpus referenzierten URIs vormals falsch klassifizierter Zeichen müssen dadurch nicht geändert werden und das kodierte Korpus bedarf keiner nachträglichen Überarbeitung.

Abbildung 2. Übergang der maschinenlesbaren Textauszeichnung zur linguistischen Analyse

Während die Kodierung des Korpus und die Klassifikation der Zeichen in TextGrid vorgenommen werden, findet die linguistische Analyse der Mayatexte in einem separaten Analysetool statt.[9] Über eine eigens entwickelte Schnittstelle liest das Programm sowohl die TEI/ XML-Dokumente als auch die in RDF gespeicherten

Transliterationswerte aus TextGrid aus und bereitet sie für den folgenden mehrstufigen Annotationsprozess von Transliteration, Transkription und morphosyntaktischer Glossierung auf. Wo bisher nur vereinzelt Studien vorliegen (vgl. Wichmann 2006), besteht nun das Ziel, eine umfassende Grammatik für das Klassische Maya zu entwickeln. Durch ein Verfahren der mehrstufigen Annotation bei gleichzeitigem Anlegen paralleler Analysepfade, die sich jederzeit auf die entsprechende Belegstelle zurückverfolgen lassen, sind ideale Voraussetzungen zur Durchführung grammatikalischer Bestimmungen und Untersuchungen geschaffen.

Im Fokus des TWKM steht die multiperspektivische Erforschung der Sprache und Schrift des Klassischen Maya. Die über die ganze Welt verstreuten und im Original größtenteils nicht zugänglichen Schriftzeugnisse der antiken Mayakultur werden mittels miteinander verknüpfter Ansätze digital erschlossen. In diesem Rahmen finden zahlreiche Technologien Anwendung, um das derzeitige Wissen über die Maya-Texte nach wissenschaftlichen Standards zu dokumentieren und sukzessive auszuweiten. Durch die kombinierte Verwendung von Ontologien zur Beschreibung und Verknüpfung einzelner Ressourcen auf Metadatenebene einerseits und TEI-P5 XML zur maschinenlesbaren Beschreibung der der Textressourcen andererseits ergibt sich ein engmaschiges Netz aus Informationen zu einem längst vergangenen Kapitel der Menschheitsgeschichte. Ein Netz, das unterschiedliche Zugänge für die wissenschaftliche Forschung sowie für die interessierte Öffentlichkeit bereithält, um tiefe Einblicke in einen vor dem Vergessen bewahrten Teil des kulturellen Erbes zu gewähren - frei, transparent und nachnutzbar.[10]

Fußnoten

1. http://mayawoerterbuch.de
2. Siehe hierzu die Definition in http://uahost.uantwerpen.be/lse/index.php/lexicon/markup-standoff/ . Die Methode wurde erstmalig beschrieben von Thompson / McKelvie (1997).
3. “Virtuell” meint, dass der Zusammenhang nicht durch die hierarchische Struktur der Daten vorgegeben, sondern durch den Verweis strukturell entkoppelter Daten aufeinander hergestellt wird. Es ist dementsprechend ein Zusammenhang, der erst durch die Verarbeitung der verknüpften Daten in einem Informationssystem kultiviert wird, d.h. erst durch die Anwendung von Informationsprozessen wird aus den verknüpften Daten eine zusammenhängende Information erzeugt.
4. Die TEI-Community führt eine anhaltende Diskussion über Vor- und Nachteile des Stand-Off Markups und dessen fortlaufender Entwicklung (vgl. Bański 2010 und Spadini / Turska / Broughton 2015). Dabei sind Flexibilität, Interoperabilität und Nachhaltigkeit der erzeugten Dokumente jene zentralen Faktoren, die in den Diskussionen immer wieder miteinander abzuwägen sind: Durch die Anwendung von Prozessierungsmechanismen, die benötigt werden, um die Daten der unterschiedlichen Dokumente zusammenzuführen, ergeben sich Probleme für die Nachhaltigkeit und Nachnutzung (vgl. Rehm et al. 2010). Dem gegenüber stehen die vielseitigen Möglichkeiten und Potenziale der Datenanreicherung und -verarbeitung: Separate Ressourcen können unabhängig voneinander bearbeitet und gleichzeitig flexibel ineinander verschränkt werden. Dadurch entstehen semantisch-reichhaltige Dokumente mit hoher Informationsdichte. Diese Vorteile zeigen sich insbesondere im Umgang mit (überlappenden) Hierarchien und Annotationen (z.B. Ide / Romary 2007 und Dipper 2005).
5. Das Projekt bemüht sich um die Nachnutzung von digitalen Faksimiles, die aber für viele Bereiche noch nicht oder in nicht ausreichender Qualität vorliegen. Über Kooperationen werden daher Archive digitalisiert, so etwa die etwa 40.000 Objekte umfassende Fotothek von Prof. Karl Herbert Mayer, Graz, von denen bereits über 5.700 Digitalisate publiziert werden konnten (https://classicmayan.kor.de.dariah.eu/). Für die Arbeiten zum 3D-Scanning siehe https://blog.sketchfab.com/from-the-rainforest-to-virtual-light-scanning-maya-hieroglyphs/ .
6. Diese Herausforderungen sind auch bei anderen antiken, nicht-alphabetischen Schriftsystemen gegeben (vgl. Rossi / De Santis 2019). Zu diesem Zweck wurde zur Vereinheitlichung von Auszeichnungen 2015 die interdisziplinäre Arbeitsgruppe EnCoWS (Encoding Complex Writing Systems) ins Leben gerufen.
7. Zur Definition von owl:sameAs siehe: https://www.w3.org/TR/owl-ref/#sameAs-def .
8. Durch diese Methode werden unter anderem auch Untersuchungen zur Klassifikationsgeschichte der Schriftzeichen ermöglicht.
9. Eine erste Beschreibung des sich in der Entwicklung befindenden Tools ‘ALMAH’ findet sich im Jahresbericht 2017 des Projekts (Grube et al. 2018: 5-7).
10. Die erzeugten Daten werden sukzessive auf dem zukünftigen Projektportal https://www.classicmayan.org/ zugänglich gemacht werden. Des Weiteren werden die Korpusdaten auch frei zugänglich im TextGrid Repository veröffentlicht werden. Die im Projekt entstandenen Schemata sind im öffentlichen Bereich unseres Git-Repositoriums einsehbar und können unter einer CC BY-4.0 Lizenz genutzt werden: https://projects.gwdg.de/projects/documentations/repository .

Bibliographie

Bańksi, Piotr (2010): *Why TEI stand-off annotation doesn't quite work and why you might want to use it nevertheless*, in: Proceedings of Balisage 2010. Series on Markup Technologies 5 https://doi.org/10.4242/BalisageVol5.Banski01 [letzter Zugriff: 05.01.2019].

Diehr, Franziska / Gronemeyer, Sven / Prager, Christian / Brodhun, Maximilian / Wagner, Elisabeth / Diederichs, Katja / Grube, Nikolai (2017): *Modellierung eines digitalen Zeichenkatalogs für die Hieroglyphen des Klassischen Maya*, in: **Eibl, Maximilian / Gaedke, Martin (eds.):** *Proceedings der INFORMATIK 2017*, Bonn: Gesellschaft für Informatik, 1185–1196 https://doi.org/10.18420/in2017_120 [letzer Zugriff 1.10.2018].

Dipper, Stefanie (2005): *XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation*, in: Proceedings of Berliner XML Tage 2005 39–50.

Grube, Nikolai / Prager, Christian / Diederichs, Katja / Gronemeyer, Sven / Grothe, Antje / Tamignaux, Céline / Wagner, Elisabeth / Brodhun, Maximilian / Diehr, Franziska (2018): *Arbeitsbericht 2017*, in: Textdatenbank und Wörterbuch des Klassischen Maya, Arbeitsstelle der Nordrhein-Westfälischen Akademie der

Wissenschaften und der Künste an der Rheinischen Friedrich-Wilhelms-Universität Bonn http://dx.doi.org/10.20376/IDIOM-23665556.18.pr005.de [letzter Zugriff: 05.01.2019].

Ide, Nancy / Romary, Laurent (2007): *Towards International Standards for Language Resources*, in: **Dybkjaer, Laila / Hemsen, Holmer / Minker, Wolfgang (eds.):** *Evaluation of Text and Speech Systems*, Springer 263–284 https://doi.org/10.1007/978-1-4020-5817-2_9 [letzter Zugriff: 05.01.2019].

Macri, Martha J. / Looper, Matthew G. (2003): *The New Catalog of Maya Hieroglyphs: The Classic Period Inscriptions*, in: The Civilization of the American Indian Series 247. Norman, OK: University of Oklahoma Press.

Neuroth Heike / Rapp, Andrea / Söring, Sibylle (2015): *TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*, Göttingen: Universitätsverlag Göttingen https://doi.org/10.3249/webdoc-3947 [letzer Zugriff 1.10.2018].

Pallan Gayol, Carlos / Anderson, Deborah (2018): *Achieving Machine-Readable Mayan Text via Unicode: Blending "Old World" Script-encoding with Novel Digital Approaches*, in: **Ortega, Élika / Worthey, Glen / Galina, Isabel / Priani, Ernesto (eds.):** *Book of Abstracts Digital Humanities 2018*, Puentes-Bridges 256–261.

Rehm, Georg / Schonefeld, Oliver / Trippel, Thorsten / Witt, Andreas (2010): *Sustainability of linguistic resources revisited*, in: Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML. Balisage Series on Markup Technologies 6 https://doi.org/10.4242/BalisageVol6.Witt01 [letzter Zugriff: 05.01.2019].

Rossi, Irene / De Santis, Annamaria (2019): *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, Berlin: De Gruyter.

Spadini, Elena / Turska, Magdalena / Broughton, Misha (2015): *TEI Standoff Markup - A work in progress*, in: TEI Members Meeting 2015 urn:nbn:nl:ui:17-f4d0afe1-5c62-4999-8271-7e8cadcd4805 [letzter Zugriff: 05.01.2019].

Textdatenbank und Wörterbuch des Klassischen Maya (2017): Ontology of the Sign Catalogue for Classic Mayan https://classicmayan.org/documentations/idiomschema.html [letzter Zugriff 1.10.2018].

Thompson, J. Eric S. (1962): *A Catalog of Maya Hieroglyphs*, in: The Civilization of the American Indian Series 62. Norman, OK.: University of Oklahoma Press.

Thompson, Henry S. / McKelvie, David (1997): *Hyperlink semantics for standoff markup of read-only documents*, in: Proceedings of SGML Europe.

Wichmann, Søren (2006): *Mayan Historical Linguistics and Epigraphy: A New Synthesis*, in: Annual Review of Anthropology 35: 279-294.

Intervalle, Konflikte, Zyklen. Modellierung von Makrogenese in der Faustedition

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

Brüning, Gerrit
bruening@faustedition.de
Freies Deutsches Hochstift / Frankfurter Goethe-Museum

Pravida, Dietmar
dpravida@goethehaus-frankfurt.de
Freies Deutsches Hochstift / Frankfurter Goethe-Museum

Wissenbach, Moritz
wissenbach@faustedition.de
Julius-Maximilians-Universität Würzburg, Deutschland

Die Faustedition versammelt 573 Textzeugen mit ca. 730 datierbaren Objekten zu Goethes „Faust" in einem digitalen Archiv. Für die Benutzer sollen sich die repräsentierten historischen Objekte nicht isoliert voneinander, sondern in einem sinnvollen Zusammenhang darstellen. Eine solcher Zusammenhang kann dadurch entstehen, dass die Objekte genetisch geordnet werden. Genetische Einordnung bedeutet zuallererst, Objekte zeitlich zu situieren. Dabei geht es teils um die Ermittlung von Kalenderdaten, teils um die Bestimmung des relativen zeitlichen Verhältnisses mehrerer Objekte. Die Ermittlung chronologischer Systeme ist eine Grundfrage in vielen Disziplinen. Je nach Sachlage können dabei etwa statistische Verfahren (z.B. Bayliss 2015 zur Archäologie) oder evolutionäre Modelle (Trovato 2014: 189–200 zur mediävistischen Philologie) eingesetzt werden; die Verwendung disziplinfremder Ansätze kann aber auch zu Problemen führen (Pereltsvaig/Lewis 2015 zur Glottochronologie). Beim derzeitigen Stand scheint es am aussichtsreichsten, das in einzelnen Disziplinen oder bei einzelnen Projekten geübte Vorgehen zu formalisieren, um potentiell generalisierbare Verfahren zu entwickeln.

Besonders schwierig ist es, nicht bloß einzelne Objekte zu datieren, sondern eine große Menge in eine chronologische Ordnung zu bringen, wenn die Objekte genetisch voneinander abhängig sind, nur wenige absolute Daten zur Verfügung stehen und sonst nur lokale Anhaltspunkte für relative Chronologien gegeben sind (klassisches Beispiel: die antike Chronographie; Grafton 1993, Burgess/Witakowski 1999). Dies ist auch bei umfangreichen genetischen Handschriftendossiers neuzeitlicher Autoren und Werken mit komplexer Entstehungsgeschichte der Fall. Hier können einzelne Teilentwürfe in relativer zeitlicher Beziehung stehen, vereinzelt sind Datierungen verfügbar; doch die Rekonstruktion der *Makrogenese*, d.h. der chronologischen

Ordnung des Gesamtbestands (zum Begriff Van Hulle 2018: 47–48), kann sich als außerordentlich komplex erweisen.

Und so liegen die Dinge auch bei „Faust“. Nur wenige der makrogenetischen Objekte sind genau datierbar; stattdessen gibt es eine große Menge relativer, aber strikt lokaler Chronologien für jeweils nur einige Objekte. Den bislang einzigen Versuch, Einzelaussagen zu aggregieren und alle Objekte in zeitlich-stemmatische Beziehung zueinander zu setzen, macht Fischer-Lamberg für zwei Akte des “Faust II”. Ihre Stemmata (Fischer-Lamberg 1955: 150–166) markieren die praktische Grenze dessen, was an Einzelinformationen mit menschlichen Mitteln aggregiert werden kann. Die Rekonstruktion einer (theoretisch beliebig großen) Makrogenese verlangt nach maschineller Verarbeitung und visueller Aufbereitung vorhandener Einzelinformationen.

Datengrundlage

Um dies zu ermöglichen, wurde der Informationsgehalt der verfügbaren einschlägigen Aussagen zur Datierung[1] in XML erfasst:

- bibliographische Quelle
- datierter Textzeuge
- absolute Datierung
- relative Datierung
- (ungefähre) Gleichzeitigkeit von Textzeugen

Eine *Relative Datierung* setzt Zeugen[2] in eine zeitliche Reihenfolge. Es wird ausgesagt, dass ein Zeuge vor einem anderen entstanden ist.

```
<relation name="temp-pre">
  <source uri="faust://bibliography/fischer-lamberg1955">S. 160</source>
  <item uri="faust://document/wa/2_III_H.5"/>
  <item uri="faust://document/wa/2_III_H.8"/>
</relation>
```

Abbildung 1. Kodierte Aussage: Laut Fischer-Lamberg 1955: 160 (source), ist 2 III H.5 vor 2 III H.8 (temp-pre) (vereinfacht).

Eine *Absolute Datierung* ordnet einen oder mehrere Textzeugen konkreten Datumsangaben zu. Tagesgenaue Datierungen bilden die Ausnahme; häufig sind Aussagen, dass ein Zeuge nicht vor oder nicht nach einem Zeitpunkt entstanden sei. Typisch sind unscharfe Angaben wie 1800/1801 oder Frühsommer. Solche Angaben werden nach dokumentierten Regeln auf Aussagen der Form nicht vor und nicht nach mit jeweils normierten tagesgenauen Angaben abgebildet.

```
<date notBefore="1825-02-26" notAfter="1825-04-05">
  <source uri="faust://bibliography/bohnenkamp1994">S. 573</source>
  <item uri="faust://document/wa/2_III_H.5"/>
</date>
```

Abbildung 2. Laut Bohnenkamp 1994: 208, ist 2 III H.5 auf den Zeitraum vom 26. Februar zum 5. April 1825 zu datieren (vereinfacht).

Modellierung als Graph

Die verschiedenen Aussagen werden in einem gerichteten Graphen modelliert[3]: Die Zeugen bilden Knoten, die relativen Datierungen Kanten des Graphen. Zur Integration der absoluten Datierungen werden die Datumsangaben (Tage) ebenfalls als Knoten in den Graphen integriert: Aus einer Datierung wie *2 III H.5 wurde nicht vor dem 26. Februar 1825 und nicht nach dem 5. April 1825 geschrieben* wird so ein Teilgraph *1825-05-25 → 2 III H.5 → 1825-04-06*, in dem die Knoten für Ereignisse (Daten oder Zeugen) und die Kanten für *zeitlich vor* stehen. Ergänzt wird der Graph durch »Zeitstrahlkanten«, die von jedem Datum zum nächstfolgenden führen.

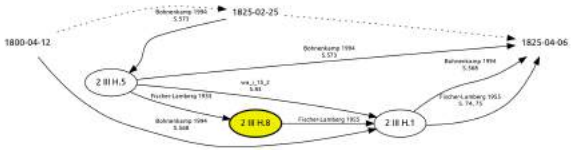


Abbildung 3. Aus Einzelaussagen gebildeter Graph zu 2 III H.8, unmittelbar benachbarten Zeugen und Datierungen (vereinfacht).

Aus diesem Graphen lassen sich Informationen ableiten, die sich erst aus dem Zusammenspiel der Einzelaussagen ergeben: Betrachtet man einen Zeugen *z*, so sind alle von *z* aus entlang gerichteter Kanten erreichbaren Zeugen (hier: 2 III H.1) nach *z* entstanden, und die von *z* aus erreichbaren Daten bilden Grenzen des *terminus ante quem* (des Zeitpunkts, vor welchem ein Zeuge entstand). Der Graph bietet damit die Grundlage für eine ungefähre Datierung auch derjenigen Zeugen, für die keine direkte absolute Datierung vorliegt.

Ist die Gesamtheit der Aussagen nicht widerspruchsfrei, so ergeben sich Zyklen im Graphen. Dies induziert einen Teilgraphen, in dem (vgl. Abb. 4 mit relativen Datierungen einiger Handschriften) jeder Knoten von jedem anderen erreichbar ist (der Teilgraph ist *stark zusammenhängend*).

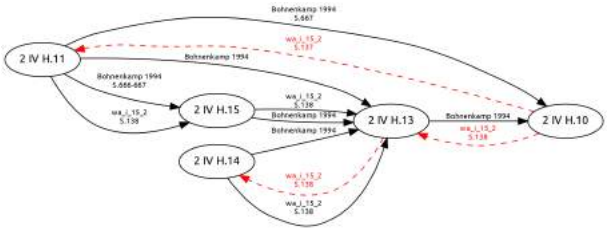


Abbildung 4. Relative Datierungen für einige Handschriften des 4. Akts. Erst die Entfernung aller rotgestrichelten Kanten führt zu einem zyklenfreien Graphen.

Um den Graphen aus Abb. 4 zyklenfrei zu machen, ist die Entfernung von wenigstens drei Kanten notwendig (gestrichelt). Der komplette Makrogenesegraph enthält eine stark zusammenhängende Komponente mit 477 Dokumenten und 2136 Kanten – zu umfangreich, um die Konflikte manuell zu eliminieren.

Konfliktbehandlung

Eine möglichst kleine Menge von Kanten zu entfernen, um einen zyklenfreien Graphen zu erhalten, ist ein als *Minimum Feedback Arc Set* bekanntes, NP-vollständiges (Karp 1972) Problem der Graphentheorie; um eine optimale Lösung zu finden, ist die größte der stark zusammenhängenden Komponenten zu groß (und die aus graphentheoretischer Sicht optimale Lösung muss auch nicht die philologisch korrekte sein, es wäre nur diejenige, die die wenigsten Datierungsaussagen verwirft). Verwendet wird stattdessen eine Heuristik, z.Z. das Verfahren von Eades (1993, Implementierung von igraph[4]), wobei zuviel entfernte Kanten nach Möglichkeit wieder hinzugefügt werden.

Entfernt man alle Konfliktkanten, so erhält man einen zyklenfreien gerichteten Graphen (DAG), der die Basis für die automatisierte Weiterverarbeitung ist. Dessen Knoten können in eine *topologische Ordnung* gebracht werden, d.h. eine Reihenfolge, die mit allen Kanten des Graphen konsistent ist (Manber 1989: 199).

Um die aus einer Vielzahl teils widersprüchlicher Aussagen mechanisch gezogenen Schlüsse nachvollziehbar und verbesserbar zu machen, werden die Daten in einer Reihe verlinkter Darstellungen mit GraphViz (Gansner/North 2000) visualisiert. Die Grundlage bildet der Gesamtgraph mit allen Aussagen, in dem algorithmisch entfernte Aussagen (Konflikte) rotgestrichelt visualisiert werden.

Zu jedem Zeugen gibt es einen Teilgraphen, der seine Nachbarn, die nächsten erreichbaren absoluten Datierungen und alle Aussagen dazwischen visualisiert. Darunter werden die Aussagen tabellarisch aufgelistet. Zu jeder (entfernten) Konfliktkante zeigt eine separate Visualisierung einen Pfad in der Gegenrichtung, mit dem die Kante in Konflikt stand, so dass der Konflikt erkennbar wird und zu den beteiligten Zeugen weiternavigiert werden kann.[5] Daneben gibt es aufbereitete Darstellungen für jede Szene und jede Quelle.

Anhand der Visualisierungen können die vorliegenden Reihenfolgeentscheidungen nachvollzogen, aber auch Datierungskontroversen und -lücken identifiziert undin den Quelldateien behoben werden:

- Aussagen können als *zu ignorieren* markiert werden, um sie bei der Bildung des Graphen auszuschließen.
- Quellen können nach ihrer Zuverlässigkeit bewertet werden, um zu beeinflussen, wie leicht oder schwer entsprechende Kanten als Konfliktkanten entfernt werden.
- Eigene Erkenntnisse können mit entsprechend hohem Gewicht einbezogen werden.

Die visualisierten Ergebnisse erfüllen so einen mehrfachen Zweck: Sie dienen zur systematischen Erschließung der einschlägigen Forschung, zur Klärung der genetischen Verhältnisse für den gesamten Faust und als Hilfsmittel zur Überprüfung, Vervollständigung und Verbesserung der Datenlage, d.h. zur Erweiterung des Forschungsstand. Darüber hinaus wird die ermittelte Reihenfolge in der Faustediton verwendet, um die Zeilensynopse sowie das Balkendiagramm zu sortieren.

Ausblick

Neben der manuellen Nachbearbeitung der Daten kommen zur Verbesserung des Verfahrens alternative Heuristiken für das Minimum-Feedback-Arc-Problem in Frage (z.B. Even et al. 1998).

Bei der vorgestellten Methode werden in der relativen Chronologie die Entstehungsintervalle nur in eine disjunkte vor-nach-Beziehung gesetzt. Möglich wäre es, hier zu prüfen, ob eine weitere Ausdifferenzierung der Beziehungen zwischen Entstehungsintervallen realisierbar ist, so dass Beziehungen der Form „wurde begonnen vor Fertigstellung von“ ausgedrückt werden können. Komplexere Relationen als eine einfache zeitliche Abfolge werden bereits von Allen (1983) in einem Graphmodell beschrieben, allerdings ist hier die globale Konfliktfreiheit ein Problem.

Eine Alternative zu der oben beschriebenen Abbildung unscharfer Aussagen auf scharf begrenzte Intervalle ist etwa die Modellierung über Fuzzy-Mengen (vgl. z.B. Barro et al. 1994). Dies erfordert jedoch auch die Neudefinition der Relationen (Schockaert/De Cock 2008) und der darauf aufbauenden Verfahren etwa zur Konfliktauflösung.

Der vorgestellte Ansatz basiert auf bereits vorhandenen, mit traditionellen philologischen Mitteln gewonnenen Datierungsaussagen. In Wissenbach/Pravida/Middell (2012) wird ein Verfahren vorgestellt, mit der kodierte, textinhärente Eigenschaften von Fassungen für die regelbasierte Bildung genetischer Hypothesen genutzt werden, um auf diesem Weg generelle Hypothesen zur Arbeitsweise des Autors zu verifizieren. Nachdem die genetische Analyse des Korpus nun weiter vorangeschritten ist, kann dieser Ansatz auf einer breiteren Datenbasis evaluiert werden.

Fußnoten

1. Zu den ausgewerteten Quellen siehe faustedition.net/macrogenesis/sources .
2. Das Datenmodell der Faustedition sieht eine konzeptuelle Unterscheidung zwischen Zeugen und Inskriptionen (Niederschriften) vor: Die eigentlichen Objekte der Datierung sind Inskriptionen. Ein Zeuge kann eine oder mehrere unterschiedlich datierte Inskriptionen enthalten. Im folgenden wird der Einfachheit halber nur von Zeugen gesprochen.
3. Mit der Graphbibliothek NetworkX (Hagberg/Schult/Swart 2008).
4. http://igraph.org/
5. Beispiel: faustedition.net/macrogenesis/H_P93--1825-01-01 .

Bibliographie

Allen, James F. (1983): *Maintaining knowledge about temporal intervals*, in: Communication of ACM 21(11): 832–843 doi: 10.1145/182.358434.

Barro, Senén / Marín, Roque / Mira, José / Patón, Alfonso R. (1994): *A model and a language for the fuzzy representation and handling of time*, in: Fuzzy Sets and Systems 61: 153–175. doi: 10.1016/0165-0114(94)90231-3.

Bayliss, Alex (2015): *Quality in Bayesian chronological models in archaeology*, in: World Archaeology 47(4): 677-700. doi: 10.1080/00438243.2015.1067640

Bohnenkamp, Anne (1994): *... das Hauptgeschäft nicht außer Augen lassend. Die Paralipomena zu Goethes Faust*. Frankfurt am Main / Leipzig: Insel.

Burgess, Richard W. / Witakowski, Witold (1999): *Studies in Eusebian and Post-Eusebian Chronography* (= Historia. Einzelschriften; 135). Stuttgart: Steiner.

Eades, Peter / Lin, Xue-Min / Tamassia, Roberto (1993): *A fast and effective heuristic for the feedback arc set problem*, in: Information Processing Letters 47(6): 319–323. doi: 10.1016/0020-0190(93)90079-O.

Even, Guy / Naor, Joseph / Schieber, Baruch / Sudan, Madhu (1998): *Approximating Minimum Feedback Sets and Multicuts in Directed Graphs*, in: Algorithmica 20(2): 151–174. doi: 10.1007/PL00009191.

Fischer-Lamberg, Renate (1955): *Untersuchungen zur Chronologie von Faust II 2 und 3. Diss. phil. (masch.)*, Humboldt-Universität Berlin.

Gansner, Emden R. / North, Stephen C. (2000): *An open graph visualization system and its applications to software engineering*, in: Software: Practice and Experience 30(11): 1203–1233. doi: 10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N.

Grafton, Anthony (1993): *Joseph Scaliger. A Study in the History of Classical Scholarship. Vol. II: Historical Chronology* (= Oxford-Warburg Studies). Oxford: Clarendon.

Hagberg, Aric A. / Schult, Daniel A. / Swart, Pieter J. (2008): *Exploring Network Structure, Dynamics, and Function using NetworkX*, in: **Varoquaux, Gael / Vaught, Travis / Millman, Jarrod (eds):** Proceedings of the 7th Python in Science Conference (SciPy2008) in Pasadena, CA https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495 [letzter Zugriff 14. Oktober 2018].

Karp, Richard M. (1972): *Reducibility Among Combinatorial Problems*, in: **Miller, Raymond Edward / Thatcher, James W. (eds.):** *Complexity of Computer Computations*. New York: Plenum 85–103. doi: 10.1007/978-3-540-68279-0_8.

Manber, Udi (1989): *Introduction to Algorithms: A Creative Approach*. Reading, MA: Addison-Wesley.

Pereltsvaig, Asya / Lewis, Martin (2015): *The Indo-European Controversy. Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

Schockaert, Steven / De Cock, Martine (2008): *Temporal Reasoning about Fuzzy Intervals*, in: Artificial Intelligence 172(8): 1158–1193. doi: 10.1016/j.artint.2008.01.001.

Trovato, Paolo (2014): *Everything You Always Wanted to Know About Lachmann's Method. A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*. Padova: libreriauniversitaria.it

Van Hulle, Dirk (2018): *Editing the Wake's Genesis: Digital Genetic Criticism*, in: **Sartor, Genevieve (ed.):** *James Joyce and Genetic Criticism. Genesic Fields* (= European Joyce Studies; 28). Leiden, Boston: Brill Rodopi 37–54.

Wissenbach, Moritz / Pravida, Dietmar / Middell, Gregor (2012): *Reasoning about Genesis or The Mechanical Philologist*, in: **Meister, Jan Christoph (ed.):** *Digital Humanities 2012. Conference Abstracts*. Hamburg: Hamburg University Press 418–422 http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/reasoning-about-genesis-or-the-mechanical-philologist.1.html [letzter Zugriff 14. Oktober 2018].

Interview-Sammlungen - Digitale Erschließung und Analyse

Pagenstecher, Cord

cord.pagenstecher@cedis.fu-berlin.de
Freie Universität Berlin, Deutschland

Seit der „Geburt des Zeitzeugen“ (Sabrow/Frei 2012) sind in Deutschland und Europa Hunderte von Oral History-Sammlungen entstanden. Zu den thematischen Schwerpunkten dieser „Era of the Witness“ (Wieviorka 2006) zählten neben der Zeit des Nationalsozialismus auch die Erfahrungen von DDR-Bürger/innen sowie die Geschlechter-, Migrations- und Minderheitengeschichte (Klingenböck 2009, Leh 2015). Vor allem seit den 1980er Jahren entstanden neben groß angelegten Dokumentationsprojekten mit jeweils Hunderten von Interviews auch zahlreiche kleine Sammlungen im Bereich von Geschichtswerkstätten, Museen und Gedenkstätten.

Audio-visuell aufgezeichnete lebensgeschichtliche Interviews sind zu einer wichtigen Quelle der Geschichtswissenschaft und ihrer Nachbardisziplinen geworden (Andresen/Apel/Heinsohn 2015). Allerdings ist der Status Quo der Sicherung, Erschließung und Bereitstellung von Oral History-Sammlungen an den verschiedenen Einrichtungen noch unzureichend.

Andererseits sind mit der raschen Entwicklung der Video- und Webtechnologie seit der Jahrtausendwende große digitale Oral History-Archive entstanden, die neue Analysemöglichkeiten bieten (Apostolopoulos/Barricelli/Koch 2016). Einige der am besten erschlossenen Sammlungen stehen am Center für Digitale Systeme der Freien Universität Berlin bereit: Das „Visual History Archive“ der USC Shoah Foundation umfasst über 53.000 Interviews, von denen CeDiS 950 Interviews transkribiert hat (www.vha.fu-berlin.de , www.zeugendershoah.de). Die 590 Interviews von „Zwangsarbeit 1939-1945“ wurden in einem spezialisierten Backend mit Workflow-Management wissenschaftlich erschlossen und in einem mehrsprachigen Online-Archiv mit timecodierten Transkripten, facettierter Suche, interaktiver Karte und Notizfunktion bereitgestellt (www.zwangsarbeit-archiv.de , Apostolopoulos/Pagenstecher 2013, Pagenstecher 2018b). Das über 90 Interviews umfassende Projekt „Erinnerungen an die Okkupation in Griechenland“ setzt den gesamten Prozess von der Interviewführung bis zur Online-Bereitstellung um (www.occupation-memories.org , Droumpouki 2016). Auch die 150 Video-Interviews der britischen Sammlung „Refugee Voices“ und die 4.500 Interviews des „Fortunoff Archives“ der Yale University stehen bereit.

In Vorbereitung sind Sammlungen zur deutsch-chilenischen sowie zur DDR-Geschichte sowie ein sammlungsübergreifender Katalog von Zeitzeugeninterviews. In Zusammenarbeit mit Sammlungsinhabern wie dem Archiv

„Deutsches Gedächtnis“ an der FernUniversität Hagen und linguistischen Experten wie dem Bayerischen Archiv für Sprachsignale an der LMU München werden darüber hinaus zukunftsträchtige Wege der digitalen Archivierung, Aufbereitung und Analyse von Oral History-Interviews gesucht.

Herausforderungen der digitalen Erschließung

Digitale Interview-Archive wie „Zwangsarbeit 1939-1945“ stellen einen ersten Schritt der Oral History in Richtung Digital Humanities dar, aber noch nicht mehr: Die Archive konnten nur mit hohem manuellen Aufwand erstellt werden. Sie sind Einzelprojekte mit unterschiedlichen Erschließungssystemen, was eine sammlungsübergreifende Recherche erschwert. Für digitale Analysen sind die Daten noch unzureichend aufbereitet. Zum Schutz von Urheber- und Persönlichkeitsrechten unterliegen die Bestände unterschiedlichen Zugangsregelungen.

Damit sind einige der Herausforderungen angesprochen, die mit der digitalen Kuratierung von Oral History-Interviews verbunden sind: Spracherkennung, Alignment, Strukturierung, Interoperabilität, Forschungsethik. Die Digital Humanities können hier Lösungswege oder Anregungen anbieten. Gleichzeitig werfen die audiovisuellen, biografisch-narrativen Daten hier besondere technologische, methodische und ethische Fragen auf.

Die audiovisuellen Medien bilden den Kern der Oral History; für Recherche, Analyse und Publikation sind aber textgebundene, timecodierte Transkriptionen von zentraler Bedeutung. Die automatische Spracherkennung hat in den letzten Jahren erhebliche Fortschritte gemacht, liefert aber für die oft dialektal und in mäßiger Aufnahmequalität vorliegenden Zeitzeugen-Interviews heute noch keine Transkripte in lesefähiger Qualität. Immerhin kann sie aber die Volltextsuche in nicht transkribierten Interviews unterstützen (Stanislav/Švec/Ircing 2016).

Um die Transkripte mit den mehrstündigen Audio- oder Videoaufnahmen zu koppeln, müssen Timecodes in die Texte eingefügt werden. Erst diese Segmentierung erlaubt eine Volltextsuche in der Audiodatei und eine synchrone Untertiteldarstellung. Verschiedene Programme unterstützen eine manuelle Transkription und Segmentierung, die aber zeitaufwändig ist. Die in der Linguistik genutzten Werkzeuge wie ELAN sind den meisten Oral Historians zu komplex, so dass in der Transkriptionspraxis meistens unstrukturierte Textdokumente erstellt werden. Erst jüngst sind automatische Alignment-Werkzeuge wie WebMAUS (https://clarin.phonetik.uni-muenchen.de/BASWebServices) so leistungsfähig geworden, dass auch mehrstündige Oral History-Interviews damit bearbeitet werden können. Ein nutzerfreundliches Transkriptionsportal für die Oral History ist in Vorbereitung (https://www.phonetik.uni-muenchen.de/apps/oh-portal/).

Bisher werden Interviews nach unterschiedlichen Richtlinien und Methoden transkribiert und indexiert. Anzustreben ist dagegen eine sammlungsübergreifend standardisierte, maschinenlesbare Auszeichnung der Interviews und ihrer timecodierten Transkripte, die oft auch weitere Auszeichnungselemente, z. B. Sprecherwechsel oder Ortsnamen, enthalten. Damit über die reine Textsuche hinaus auch strukturnutzende Suchverfahren möglich sind, müssen die Transkriptionen mit diesen verschiedenen Annotations-Ebenen strukturiert abgebildet werden. Dafür empfiehlt sich ein auf den TEI-Guidelines der Text Encoding Initiative (http://www.tei-c.org/release/doc/tei-p5-doc/de/html/TS.html) basierendes Schema, das derzeit für die an der FU Berlin erschlossenen Interviewsammlungen erarbeitet wird. Aus linguistischer Sicht wurde dafür der ISO-Standard 24624:2016 entwickelt (Schmidt 2011).

Eine interoperable Erschließung wird erschwert durch die unterschiedlichen Communities, in denen Oral History-Bestände beheimatet sind. Je nach Anbindung der Interview-Sammlungen an ein Archiv, eine Bibliothek, ein Museum oder ein sprachwissenschaftliches Zentrum kommen unterschiedliche Metadatenstandards zur Verwendung. In der Archivwelt ist EAD verbreitet, die angelsächsischen Bibliotheken nutzen MARC21, CLARIN setzt auf das CMDI-Framework. Übergreifende Crosswalks und Discovery-Systeme fehlen, was eine sammlungsübergreifende Recherche erschwert.

Schließlich sind die Persönlichkeitsrechte der Interviewten besonders zu beachten. Angesichts der kollaborativen Produktion sensibler Daten im Interviewprozess hat die Oral History-Community schon früh über forschungsethische Verantwortung diskutiert (Leh 2000). Für die digitale Bereitstellung und Analyse von Interviews ist daher große Sensibilität und ein abgestuftes Rechtemanagement erforderlich, oft auch eine Anonymisierung der Interviews. Dies stellt – zusammen mit Fragen der Langzeitarchivierung und persistenten Auffindbarkeit von Audio- und Videodateien – eine weitere Herausforderung dar.

Forschungsfragen und Analysemöglichkeiten

Während sich die Linguistik verstärkt für die umfangreichen Datenkorpora der Oral History interessiert (Kasten/Roller/Wilbur 2017, Armaselu/Danescu/Klein 2018), bleiben viele Historiker/innen skeptisch gegenüber quantifizierenden, womöglich dekontextualisierenden Analysemethoden. Hier dominieren hermeneutische und textbasierte Zugänge in der Auswertung weniger, oft selbst geführter Interviews anhand der Transkriptionen. Nun aber erleichtern die digitalen Interview-Archive die Sekundäranalyse vorhandener Interviews unmittelbar anhand der Ton- und Videoaufnahmen. Damit entstehen neue Forschungsfragen, einerseits in Bezug auf Multimodalität und Interaktion in der Gesprächssituation, andererseits auf komparative und korpuslinguistische Auswertungsmöglichkeiten.

Eine Fallstudie untersuchte Erzählmuster und Ausdrucksweisen, Intonation und Mimik in zwei Interviews aus dem „Visual History Archive“ und dem Online-Archiv „Zwangsarbeit 1939-1945“. In den beiden Aufnahmen von 1998 und 2006 berichtet die gleiche Zeitzeugin über ihr Leben: Anita Lasker-Wallfisch, britische Cellistin, Holocaust-Überlebende und Breslauer Jüdin. Im Vergleich zeigt das spätere Interview einen klareren Anspruch auf epistemische Autorität sowie eine gewachsene narrative Erfahrung und performative Leistung (Pagenstecher 2018a). Dabei fällt besonders die nonverbale Interaktion auf: Gerade in kritischen Momenten schmiedet die Erzählerin eine visuell-argumentative Allianz mit dem Interviewer.

Diese narrativen Muster blieben unbemerkt bei einer konventionellen Interview-Analyse anhand des Transkripts. Digitale Methoden helfen auch, die im Zentrum jedes Interviews stehende Arbeitsallianz zwischen Erzähler/in und Interviewer/in besser zu verstehen. Dieser mehr oder weniger versteckte Dialog ist zwar aus Sicht der praktischen Interviewführung vielfach empfehlend beschrieben, gelegentlich auch selbstkritisch reflektiert, aber noch kaum vergleichend analysiert worden (Pagenstecher/ Pfänder 2017).

Quantitative Ansätze können helfen, spezifische Erinnerungs- und Erzählmuster in großen Interviewsammlungen zu erkennen. Als Pilotstudie wird hier die Nutzung von Begriffen wie „Sklaverei", „Sklavenarbeit" oder „versklavt" in den Interviews untersucht. Da die Daten für korpuslinguistische Tools noch nicht ausreichend exportierbar sind, werden hier nur die Analysemöglichkeiten des Online-Archivs „Zwangsarbeit 1939-1945" genutzt. Seit den Nürnberger Prozessen wurde die nationalsozialistische Zwangsarbeit immer wieder in den historischen Kontext der Sklaverei gestellt. Im Laufe der Zeit bekam der Begriff „Sklavenarbeit" in den verschiedenen Öffentlichkeiten der betroffenen Länder sehr unterschiedliche Konnotationen (vgl. Pagenstecher 2010).

Wie aber sprechen ehemaligen Zwangsarbeiter/innen 2005/2006, also kurz nach der Entschädigungsdebatte, geführten Interviews über ihre Erfahrung der „Sklavenarbeit"? Eine Volltextsuche nach dem Wortteil „*sklav*" liefert Treffer in 140 von 477 Interviews. Der insgesamt in 29% aller Interviews auftauchende Sklaven-Begriff wird besonders häufig in italienischen (78%) und englischen (66%) Interviews verwendet. Dem entspricht die Verteilung auf Erfahrungsgruppen: Italienische Militärinternierte (60%) und jüdische Überlebende (42%) nutzen den Terminus häufiger, Religiös Verfolgte (14%) sowie Sinti und Roma (16%) seltener als der Durchschnitt.

Entscheidend ist dabei offensichtlich weniger die gruppenspezifische Erfahrung als der landesspezifische Erinnerungsdiskurs. So sprechen jüdische Überlebende in Israel oder Osteuropa seltener über Sklaventum als die im englischen Sprachraum Interviewten. Dass in angelsächsischen Ländern mehr über Sklaven gesprochen wird, liegt vermutlich an der dem Interviewprojekt vorausgehenden, längeren öffentlichen Debatte über die Zwangsarbeiter-Entschädigung, die vor allem dort stark vom Terminus Sklavenarbeit geprägt war. Typisch dafür war etwa die Schlagzeile „Nazi slaves take case to US" (BBC 1999). Offenbar reagieren die Interviewten auf diese Diskurse, teilweise auch direkt auf sprachliche Vorgaben der Interviewer/innen.

Die Art der Zwangsarbeit spielt dagegen eine geringere Rolle; allerdings wird Sklaventum bei Erfahrungen im gemeinhin besonders schweren Einsatzbereich Bau/ Steine/Erden (40%) häufiger erwähnt. Hier ist also eine genauere Untersuchung der Verwendungskontexte erforderlich. Zu prüfen ist beispielsweise, ob mit der „Sklaven"-Referenz eher eine damalige (Arbeits-, Gewalt- oder Diskriminierungs-)Erfahrung wiedergegeben oder eher aus heutiger (biografischer oder politischer) Sicht über eigene Erfahrungen reflektiert wird. Für solche Fragestellungen reicht der von der Korpus-Software angebotene Kontext von einigen Wörtern links und rechts der Fundstelle nicht aus. Hier ist ein Close Reading oder Viewing des Interviews erforderlich.

Resümee

Anhand der digitalen Interview-Sammlungen an der Freien Universität Berlin skizzierte dieser Beitrag die Potentiale und Herausforderungen des Zusammenwirkens von Digital Humanities und Oral History in der Kuratierung und Analyse.

Digitale Technologien ermöglichen die softwaregestützte Sicherung, Erschließung und Bereitstellung von Zeitzeugen-Interviews sowie ihre sammlungsübergreifende Recherche und quellennahe Analyse. In Zukunft können auch quantitative Analysen genutzt werden, um individuelle und kollektive Muster des Erfahren, Erinnerns und Erzählens zu entdecken.

Gewiss verliert das digital aufbereitete Zeugnis im Zeitalter der technischen Reproduzierbarkeit ein Stück seiner Aura. Seiner fundierten Analyse und sorgsamen, kontextualisierenden Interpretation sollte dies freilich keinen Abbruch tun. Die Digital Humanities eröffnen der Oral History jedenfalls faszinierende neue Forschungsperspektiven.

Bibliographie

Andresen, Knud / Apel, Linde / Heinsohn, Kirsten (eds.) (2015): *Es gilt das gesprochene Wort. Oral History und Zeitgeschichte heute*, Göttingen.

Apostolopoulos, Nicolas / Pagenstecher, Cord (eds.) (2013): *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt*, Berlin.

Apostolopoulos, Nicolas / Barricelli, Michele / Koch, Gertrud (eds.) (2016): *Preserving Survivors' Memories. Digital Testimony Collections about Nazi Persecution: History, Education and Media*, Berlin: Stiftung „Erinnerung, Verantwortung und Zukunft" (EVZ), URL: http://www.stiftung-evz.de/fileadmin/user_upload/EVZ_Uploads/Handlungsfelder/Auseinandersetzung_mit_der_Geschichte_01/Bildungsarbeit-mit-Zeugnissen/Testimonies_Band3_Web.pdf [zuletzt abgerufen: 10. Januar 2019]

Armaselu, Florentina / Danescu, Elena / Klein, Francois (2018): *Oral History and Linguistic Analysis. A Study in Digital and Contemporary European History*, in: CLARIN Annual Conference 2018 Proceedings: 11-15, URL: https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf [zuletzt abgerufen: 10. Januar 2019]

BBC (1999): *Nazi slaves take case to US*, 12.10.1999, URL: http://news.bbc.co.uk/2/hi/europe/472104.stm [zuletzt abgerufen: 10. Januar 2019]

Droumpouki, Anna Maria (2016): *Erinnerungen an die Okkupation in Griechenland. Entstehung, Entwicklung und gesellschaftliche Bedeutung eines deutsch-griechischen Dokumentationsprojekts*, in: BIOS. Zeitschrift für Biographieforschung und Oral History, 29/1: 141-151. https://doi.org/10.3224/bios.v29i1.09 [zuletzt abgerufen: 10. Januar 2019]

Kasten, Erich / Roller, Katja / Wilbur, Joshua (eds.) (2017): *Oral History Meets Linguistics*, Fürstenberg: SEC, 185-207, URL: http://www.siberian-studies.org/publications/orhili_E.html [zuletzt abgerufen: 10. Januar 2019]

Klingenböck, Gerda (2009): *Stimmen aus der Vergangenheit. Interviews von Überlebenden des*

Nationalsozialismus in systematischen Sammlungen von 1945 bis heute, in: **Daniel Baranowski (eds.):** *„Ich bin die Stimme der sechs Millionen". Das Videoarchiv im Ort der Information*, Berlin: Stiftung Denkmal für die ermordeten Juden Europas, 27-40

Leh, Almut (2000): *Forschungsethische Probleme in der Zeitzeugenforschung*, in: BIOS. Zeitschrift für Biographieforschung und Oral History, 13: 64-76

Leh, Almut (2015): *Vierzig Jahre Oral History in Deutschland. Betrag zu einer Gegenwartsdiagnose von Zeitzeugenarchiven am Beispiel des Archivs ‚Deutsches Gedächtnis'*, in: Westfälische Forschungen. Zeitschrift des LWL-Instituts für westfälische Regionalgeschichte, 65: 255-268

Pagenstecher, Cord (2010): *'We were treated like slaves.' Remembering forced labor for Nazi Germany*, in: **Gesa Mackenthun, Raphael Hörmann (eds.)**, *Human Bondage in the Cultural Contact Zone. Transdisciplinary Perspectives on Slavery and Its Discourses*, Münster: Waxmann 275-291.

Pagenstecher, Cord / Pfänder, Stefan (2017): *Hidden dialogues. Towards an interactional understanding of Oral History interviews*, in: **Kasten, Erich / Roller, Katja / Wilbur, Joshua (eds.):** *Oral History Meets Linguistics*, Fürstenberg: SEC: 185-207

Pagenstecher, Cord (2018a): *Testimonies in digital environments: comparing and (de-)contextualising interviews with Holocaust survivor Anita Lasker-Wallfisch*, in: Oral History Journal, 46 (2): 109-118

Pagenstecher, Cord (2018b): *Curating and Analyzing Oral History Collections*, in: CLARIN Annual Conference 2018 Proceedings , ed. by Inguna Skadin and Maria Eskevich: 34-38, URL: https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf [zuletzt abgerufen: 10. Januar 2019]

Sabrow, Martin / Frei, Norbert (eds.) (2012): *Die Geburt des Zeitzeugen nach 1945*, Göttingen: Wallstein

Schmidt, Thomas (2011): *A TEI-based Approach to Standardising Spoken Language Transcription*, in: Journal of the Text Encoding Initiative, 1, DOI:10.4000/jtei.142.

Stanislav, Petr / Švec, Jan / Ircing, Pavel (2016): *An Engine for Online Video Search in Large Archives of the Holocaust Testimonies*, in: Interspeech 2016: Show & Tell Contribution, https://www.isca-speech.org/archive/Interspeech_2016/pdfs/2016.PDF [zuletzt abgerufen: 10. Januar 2019]

Wieviorka, Annette (2006): *The Era of the Witness*, New York.

Jadescheibe oder Kreis – Reflexion über manuelle und automatisierte Erkennung von Schriftzeichen der vorspanischen Mayakultur

Prager, Christian
cprager@uni-bonn.de
Universität Bonn, Abteilung für Altamerikanistik, Deutschland

Mara, Hubert
hubert.mara@iwr.uni-heidelberg.de
Universität Heidelberg, Interdisziplinäres Zentrum für wissenschaftliches Rechnen, Heidelberg

Bogacz, Bartosz
bartosz.bogacz@iwr.uni-heidelberg.de
Universität Heidelberg, Interdisziplinäres Zentrum für wissenschaftliches Rechnen, Heidelberg

Feldmann, Felix
felix.feldmann@iwr.uni-heidelberg.de
Universität Heidelberg, Interdisziplinäres Zentrum für wissenschaftliches Rechnen, Heidelberg

Einleitung

In unserem Vortrag fassen wir Ergebnisse einer laufenden, interdisziplinären Kooperation im Bereich digitaler Epigraphie zwischen dem Akademieprojekt "Textdatenbank und Wörterbuch des Klassischen Maya" (TWKM, Universität Bonn, Prager) und einem auf Schriftforschung spezialisierten Team am Interdisciplinary Center for Scientific Computing (IWR Heidelberg, Mara, Bogacz und Feldmann) zusammen (vgl. Bogacz, Feldmann, Prager, Mara 2018). Im Mittelpunkt der Kooperation zwischen Schriftforschung und angewandter Informatik steht die vollautomatische Erkennung von Zeichen der Hieroglyphenschrift der Klassischen Mayakultur in 3D (Zeitraum: 250 - 900 n.Chr., Region: südliches Mexiko, Guatemala, Belize und Honduras). Am IWR in Heidelberg wird dieses Verfahren seit mehreren Jahren erfolgreich bei der automatischen Erkennung und Transliteration von Keilschriftzeichen auf Tontafeln angewendet, die mittels eines hochauflösenden 3D-Scanners dokumentiert wurden (Bogacz, Gertz und Mara 2015; Bogacz, Klingmann, Mara 2017) (Abbildung 1) .

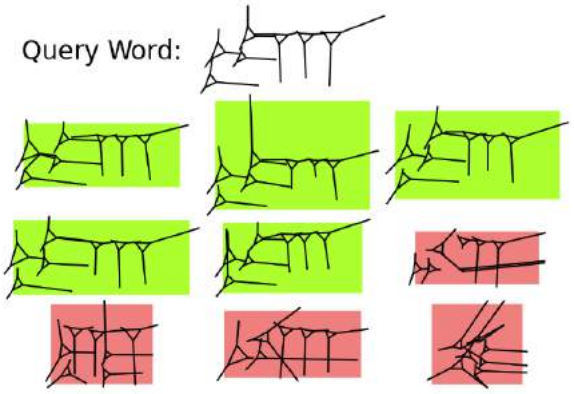


Abbildung 1. Ergebnisauswahl einer automatisierten Abfrage des akkadischen Keilschriftzeichens /ta/ in einem Textkorpus. Korrekte Erkennungen sind grün markiert (Bogacz, Gertz und Mara 2015, Abb. 4).

Im Zuge der Arbeit für die Textdatenbank und das digitale Wörterbuch des Klassischen Maya hat das Projekt Mayainschriften in 3D dokumentiert und zusammen mit den Projektpartnern am IWR Möglichkeiten getestet, ob die aus rund 1000 Elementen bestehende Hieroglyphenschrift der Maya ebenfalls dazu geeignet ist, automatisiert erkannt zu werden. Der wesentlichste Unterschied zu aktuellen Vorgehensweisen in der digitalen Epigraphie, denen typischerweise die manuelle Kodierungen und Verlinkung der Texte (z.B. mit Hilfe von XML TEI, RDF), sowie darauf aufbauende korpuslinguistische Analysen zu Grunde liegen (vgl. Diehr et al. 2018 zum Klassischen Maya; Chiarcos et al. 2018 am Beispiel des Sumerischen Textkorpus), ist der weitgehend vollautomatische Ansatz auf Bilddaten (Abbildung 2). In der konkreten Anwendung handelt es sich um Krümmungsvisualisierungen, die aus 3D-Messdaten berechnet werden. In diesem Sinne entsteht eine Optical Character Recognition (OCR) für Schrift in 3D. Ein Ziel der Kooperation ist es mit Hilfe dieses automatisierten Verfahrens einen auf der Zeichenmorphologie basierenden Katalog der Mayaschriftzeichen einschließlich ihrer Varianten automatisiert zu erstellen.

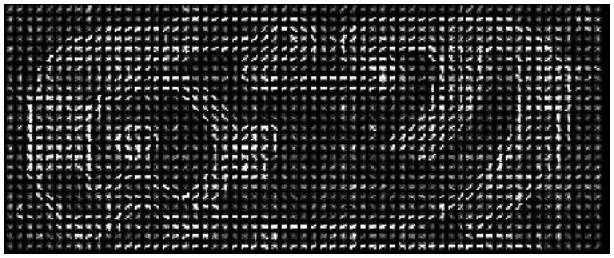


Abbildung 2. Beispiel für automatische Zeichenerkennung der Hieroglyphe /chi/ aus der Dresdner Mayahandschrift mit Hilfe eines *Histogram of Oriented Gradients* (HOG) (Feldmann, Bogacz, Prager, Mara 2017, Abb. 3).

Wir präsentieren Methoden, Herausforderungen und Ergebnisse und zeigen am Material auf, wo wir Fortschritte und Durchbrüche, aber auch (aktuell) Grenzen bei der (voll)automatisierten Erkennung von Maya-Schriftzeichen festgestellt haben. Fokussierend auf Differenzen zwischen manueller und automatischer Zeichenbestimmung ziehen wir Schlussfolgerungen für die Erforschung der Maya-Schrift und die angewandte Informatik.

Geburtsstunden und -wehen der digitalen Epigraphie des Klassischen Maya

Die digitale Epigraphie des Klassischen Maya erlebte ihre Geburtsstunde zu Beginn des Kalten Krieges, als sowjetische Forscher elektronische Rechenmaschinen einsetzten um das Rätsel der Mayaschrift zu lösen. Wenige Jahre zuvor veröffentlichte ein damals führender deutscher Experte für die Maya-Hieroglyphen, die Entschlüsselung der Mayaschrift sei nach über fünfzig Jahren vergeblicher Arbeit ein unlösbares Problem (Schellhas 1945). Nur fünfzehn Jahre später verkündeten sowjetische Mathematiker und Archäologen, dass ihnen gemeinsam mit Hilfe eines Elektronenmaschinenrechners innerhalb von vierzig Stunden die vollständige Entzifferung und Übersetzung der Maya-Handschriften gelungen sei (Sobolev 1961; O'Kane 1962). Für die maschinelle Verarbeitung übertrug man die Texte von zwei erhaltenen Maya-Handschriften mit Hilfe von Nummernschlüsseln in ein maschinenlesbares Format und speicherte sie über Lochkarten in der Rechenmaschine. In Minutenschnelle prozessierte der Rechner das Datenmaterial und erzeugte Häufigkeits-, Vorkommens- und Kookkurrenzanalysen der Schriftzeichen, Zeichenkombinationen, Wörter und ganzen Wortfolgen. Die Ergebnisse der lexikometrischen Erhebung wurde mit Häufigkeiten von Silben, Silbenkombinationen, Wörtern und Sätzen aus yukatekisch-sprachigen und in lateinischer Schrift aufgezeichneten Texten und Wörterbüchern aus dem 16. und 17. Jahrhundert verglichen und korreliert. Die Grundannahmen des Forscherteams waren jedoch falsch, der Versuch das Problem der Mayaschrift maschinell zu lösen, galt schon kurz nach seiner Veröffentlichung als gescheitert (Schlenther 1964).

Digitaler Dornröschenschlaf oder wie das Rätsel doch gelöst wurde

Weitere Versuche die Mayaschrift digital zu erforschen folgten Mitte 1960 (Rendón 1965), dann erst wieder in den 1980er Jahren (Ringle und Smith-Stark 1996). Beide Projekte beschränkten sich darauf Konkordanzen einzelner Hieroglyphen zu kompilieren. Keines der Projekte hatte jedoch einen nennenswerten Impakt für die Forschung: die Anfang der 1960er Jahre begonnene digitale Epigraphik des Klassischen Maya fiel in einen Dornröschenschlaf, aus der sie erst wieder durch das Forschungsprojekt Textdatenbank und Wörterbuch des Klassischen Maya erweckt wurde (Prager et al. 2016). Die Entzifferung der Mayaschrift gelang in den vergangenen 50 Jahren gänzlich ohne komputationelle Hilfsmittel. Wegweisend dazu waren Arbeiten des Ägyptologen Juri Knorozov (1956). Aufgrund der Zeichenzahl schloss er daraus, dass es sich bei der Mayaschrift um ein dem Altäyptischen vergleichbares Schriftsystem handelte und hatte dadurch den logo-syllabischen Charakter des Mayaschriftsystems erkannt. Die Entzifferung der Mayaschrift ist bis heute im Prozess, wir kennen etwa von 60% der rund 1000 verschiedenen Schriftzeichen den Lautwert. Obschon die Ergebnisse des sowjetischen Teams Anfang der 1960er Jahre die Forschung nicht nachhaltig beeinflusste, war es aus heutiger Sicht das erste Projekt, das nicht nur interdisziplinär zusammenarbeitete, sondern EDV zur Lösung eines epigraphisch-linguistischen Problems herangezogen hatte.

Herausforderungen beim Verständnis des Maya-Schriftsystems: Graphematische und graphetische Herausforderungen für die Erkennung von Schriftzeichen

Bedeutend für den Durchbruch bei der Entzifferung war die Entdeckung des Prinzips der Zeichensubstitution, die auch den großen Variantenreichtum in der Mayaschrift erklärt. Abbildung 3 zeigt drei morphologisch unterschiedliche Graphe des Zeichens /pa/, die lediglich das gemeinsame, diagnostische Merkmal einer schraffierten Fläche aufweisen.

Abbildung 3. Sogenannte Standard-, Kopf- und Körpervariante des Zeichens für die Silbe /pa/ in der Mayaschrift (Marc Zender, 1999).

Heute kennen wir eine Bandbreite an Schreib- und Gestaltungsprinzipien, womit nicht nur das einzelne Graphem, sondern auch Wörter des Klassischen Maya variantenreich realisiert wurden. Die Schreiber strebten ein Höchstmaß an visueller Prachtentfaltung und formaler Variation an. Eintönigkeit, Konformität und Wiederholung sollten vermieden werden, kalligraphische Spielarten bestimmten das Werk des Schreibers und stellen heute eine immense Herausforderung für die automatische Erkennung dar.

Automatisiertes Suchen von Mayaschriftzeichen: erste Ergebnisse

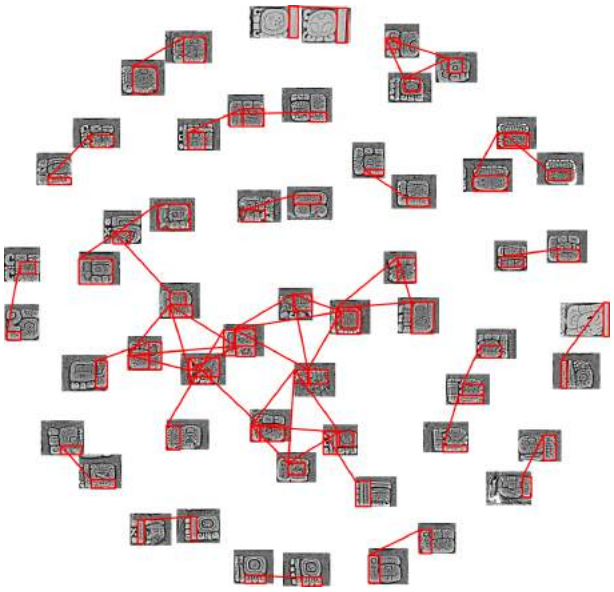


Abbildung 4. Ähnlichkeitsnetzwerk von Zeichen der Mayaschrift, die der Algorithmus vollautomatisch erkannt und verlinkt hat (Bartosz Bogacz et al. 2018).

Mit dem heutigen Stand der Technik in Form von Rechenleistung und Methoden der Bildverarbeitung, Mustererkennung und Maschinellem Lernen konnten wir zeigen, dass die manuellen Entzifferungen der 1980er Jahre in Algorithmen abgebildet werden können. Dabei ist anzumerken, dass z.B. die berechneten Ähnlichkeitsmaße mit Arbeitsplatzrechnern innerhalb weniger Minuten bestimmt werden können. Dies ermöglicht das Testen verschiedener Hypothesen über Zeichenähnlichkeiten in Form von Änderungen an Parametern und Kombinationen von unterschiedlichen Algorithmen. Hierbei wird der bereits in der digitalen Erforschung von Keilschrifttafeln angewendete *Multiscale Integral Invariant Filter* (MSII) eingesetzt um Schriftzeichen in 3D-Objekten zu isolieren. Die 3D-Objekte werden in diesem Verfahren in 2D umgesetzt und mit Hilfe von Projektionsprofilen segmentiert, um ein Gitter aus Spalten und Zeilen zu erzeugen. Anschließend werden die Hieroglyphenblocks selbst nach dem Zufallsprinzip segmentiert, wobei Hintergrund und Vordergrund aufgrund der Oberflächenkrümmung der ursprünglichen 3D-Oberfläche getrennt werden. Die abgerufenen Zeichen werden zunächst nach ihrer Größe zu einem Satz gängiger Größen zusammengefasst. Für jede Glyphe wird ein auf dem Histogramm der Gradienten (HOG) basierender Merkmalsvektor berechnet und für ein hierarchisches

Clustering verwendet (Abb. 3). Ein bemerkenswertes Ergebnis ist das Erkennen von Linienelementen komplexer Zeichen, die auch in gestauchter Form vorkommen können. Damit verhält sich die vollautomatische *Machine Learning Pipeline* sehr ähnlich zur Diagnostik, wie sie von Experten angewendet wird. Mit den jetzigen Parametern werden gedoppelte und gestauchte Elemente der Mayaschrift korrekt identifiziert. Dies wird in einer Visualisierung der Zeichen in einem Graphen bzw. Ähnlichkeitsnetzwerk besonders deutlich. Das System errechnet die Grenzen der Hieroglyphenblöcke (Abbildung 4) und kann zwischen Bild- und Textinformationen unterscheiden.

Das Datenmaterial für unsere Experimente stammt mit einer Tafel aus Cancuen und einer anderen Tafel aus dem Fundort La Corona zudem aus unterschiedlichen Epochen und Regionen. Durch die Verwendung von hoch-aufgelösten 3D-Messdaten handelt es sich hierbei um digitale Primärquellen im Unterschied zu (retro-)digitalisierten Handzeichnungen, die eine Interpretation beinhalten, und somit eher als interpretierte Sekundärquellen zu verstehen sind.

Im erzeugten Ähnlichkeitsnetzwerk der Primärquellen ist klar zu erkennen, dass Übereinstimmungen bei konstanten Elementen wie Zahlzeichen verbunden werden, während komplexere Zeichen in geschlossenen Clustern gezeigt werden, die dem Inschriftenträger entsprechen. Bei erodierten Zeichen im gleichen Cluster wurden korrekte Vorschläge gemacht (Abbildung 4, links). Deutlich werden Verbindungen aufgezeigt zu denen bereits Vermutungen über Übereinstimmungen vorliegen (Abbildung 4, rechts). Bei sauber gearbeiteten Schreibungen ohne Abweichungen ist die Identifizierung immer perfekt und basiert auf den diagnostischen Elementen eines Zeichens. Dies entspricht der intuitiven Analyse durch einen Experten.

Abbildung 5. Korrekte Rekonstruktion des Silbenzeichens /na/, das im oberen Fall fast gänzlich verloren ist.

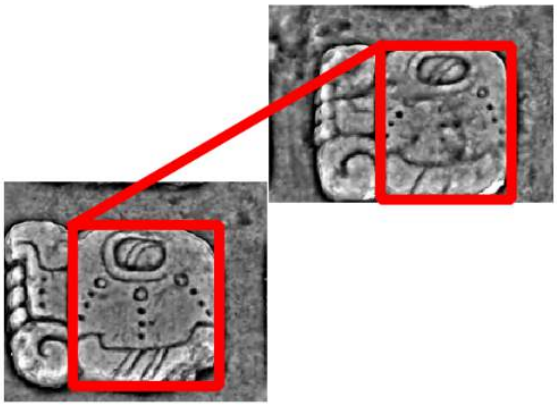


Abbildung 6. korrekte Übereinstimmung identischer Zeichen. Grafik: Bartosz Bogacz, 2018.

Schlussfolgerung

Aus der Sicht der angewandten Informatik ist die Erkennung von Mayaschriftzeichen eine interessante Herausforderung, da die Grundwahrheit nicht bekannt ist. Weil die digitale Beschreibung und Verarbeitung der Zeichen kontinuierlich verbessert wird, kann das Wissen über die Zeichen immer weiter an die Grundwahrheit angenähert werden. Damit entsteht ein Spannungsfeld in dem sich die Entwicklung von Algorithmen und die Entwicklung von geisteswissenschaftlichen Fragestellungen gegenseitig beeinflussen. Dabei kommt es immer wieder zu analytischen hermeneutischen Fragen über die Anwendbarkeit der neuesten Entwicklungen in der Informatik. Insbesondere die – in anderen Anwendungsgebieten – sehr erfolgreichen *Convolutional Neural Networks* bzw. *Deep Learning* scheinen auf Grund der relativ geringen Datenmengen zur Zeit nicht direkt anwendbar. Basierend auf den aktuellsten positiven Ergebnissen sind wir zuversichtlich, dass hier neue Methoden für die digitale Epigraphie in Arbeit sind und Synergien zwischen Gedächtnisleistung, klassischem *Machine Learning* und *Deep Learning* zu weiteren Verbesserungen führen werden.

Bibliographie

Bogacz, Bartosz / Gertz, Michael / Mara, Hubert (2015): *"Character Retrieval of Vectorized Cuneiform Script"*, in: 13th IAPR International Conference on Document Analysis and Recognition (ICDAR2015) https://www.researchgate.net/publication/281781820_Character_Retrieval_of_Vectorized_Cuneiform_Script [letzter Zugriff 28. September 2018].

Bogacz, Bartosz / Feldmann, Felix / Prager, Christian / Mara, Hubert (2018): *"Visualizing Networks of Maya Glyphs by Clustering Subglyphs"*, in: Sablatnig, Robert et al. (eds): Eurographics Workshop on Graphics and Cultural Heritage. Geneva: The Eurographics Association 105-111 http://doi.org/10.2312/gch.20181346 [letzter Zugriff 12. Januar 2019].

Bogacz, Bartosz / Klingmann, Maximilian / Mara, Hubert (2017): *"Automatic Transliteration of Cuneiform from Parallel Lines with Sparse Data"*, in: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR2017) https://www.researchgate.net/publication/321491564_Automating_Transliteration_of_Cuneiform_from_Parallel_Lines_with_Sparse_Data [letzter Zugriff 28. September 2018].

Chiarcos, Christian / Pagé-Perron, Émile / Khait, Ilya / Schenk Niko / Reckling, Lucas (2018): *"Towards a Linked Open Data Edition of Sumerian Corpora"*, in: Calzolari, Nicoletta et al.: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. Paris: European Language Resources Association 2437-2444. http://www.lrec-conf.org/proceedings/lrec2018/pdf/862.pdf [letzter Zugriff 28. September 2018].

Diehr, Franziska / Brodhun, Maximilian / Gronemeyer, Sven / Diederichs, Katja / Prager, Christian / Wagner, Elisabeth / Grube, Nikolai (2018): *"Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya"*, in: Wartena, Christian et al. (eds): Knowledge Organization for Digital Humanities: Proceedings of the 15th Conference on Knowledge Organization WissOrg'17 of the German Chapter of the International Society for Knowledge Organization (ISKO). Berlin: Freie Universität Berlin 37–43 doi: https://doi.org/10.17169/FUDOCS_document_000000028863 [letzter Zugriff 28. September 2018].

Feldmann, Felix / Bogacz, Bartosz / Prager, Christian / Mara, Hubert (2017): *"Histogram of Oriented Gradients for Maya Glyph Retrieval"*, in: Schreck, Tobias et al. (eds): Eurographics Workshop on Graphics and Cultural Heritage. Geneva: The Eurographics Association 115–118 http://dx.doi.org/10.2312/gch.20171301 [letzter Zugriff 28. September 2018].

Knorozov, Yuri (1956): *"New Data on the Maya Written Language"*, in: Journal de la Société des Américanistes 45: 209–217 https://www.persee.fr/doc/jsa_0037-9174_1956_num_45_1_961 [letzter Zugriff 28. September 2018].

O'Kane, Lawrence (1962): *"Computers Solve Mayan Writings; Soviet Mathematicians Use Devices for Translation Original Writing System Glossaries Developed Samples of Translations Expert Reserves Judgment"*, in: The New York Times, 15 April http://query.nytimes.com/gst/abstract.html?res=9800EEDA143DE532A25756C1A9629C946391D6CF [letzter Zugriff 28. September 2018].

Rendón, Juan / Spescha, Amalia (1965): *"Nueva clasificación plástica de los glifos mayas"*, in: Estudios de Cultura Maya 5: 189-252 http://dx.doi.org/10.19130/iifl.ecm.1965.5.668 [letzter Zugriff 28. September 2018].

Ringle, William M. / Thomas C. Smith-Stark (1996): *A Concordance to the Inscriptions of Palenque, Chiapas, Mexico* (= Middle American Research Institute Publication 62) New Orleans, LA: Middle American Research Institute, Tulane University.

Schellhas, Paul (1945): *"Die Entzifferung der Mayahieroglyphen: ein unlösbares Problem?"*, in: Ethnos 10(1): 44–53 https://doi.org/10.1080/00141844.1945.9980637 [letzter Zugriff 28. September 2018].

Schlenther, Ursula (1964): *"Kritische Bemerkungen zur kybernetischen Entzifferung der Maya-Hieroglyphen (mit 10 Abbildungen und 3 Tabellen"*, in: Ethnographisch-archäologische Zeitschrift 5(5): 111-139.

Sobolev, Sergei L'vovich (1961): *"Die vollständige Entzifferung der Maya-Handschriften durch mathematische Methoden"*, in: Wissenschaftliche Zeitschrift der Humboldt Universität 10(4–5): XVII–XXI.

Zender, Marc (1999): *Diacritical Marks and Underspelling in the Classic Maya Script: Implications for Decipherment*. M.A. Thesis. Department of Archaeology, University of Calgary http://dx.doi.org/10.5072/PRISM/19313 [letzter Zugriff 28. September 2018].

Kann Nonstandard standardisiert werden? Ein Annotations-Standardisierungsversuch nicht nur von PoS-Tags am Beispiel des Spezialforschungsbereichs „Deutsch in Österreich“

Seltmann, Melanie E.-H.
melanie.seltmann@univie.ac.at
Universität Wien, Österreich

Einleitung

Unter einer Annotation von Sprachdaten wird eine Markierung, Kategorisierung und Interpretation derselben verstanden. Sie dienen – auch in anderen Wissenschaftsdisziplinen – in der Regel der wissenschaftlichen Auseinandersetzung und dem Forschungsprozess mit einem Datum und halten ebenso dessen Ergebnis fest (vgl. Breuer/Seltmann 2018: 145). Für die technische Realisierung können dabei sehr unterschiedliche Umsetzungen eingesetzt werden, abhängig sowohl von den Daten, die in einem Projekt untersucht werden, als auch den persönlichen Vorlieben der (entscheidenden) Mitarbeiter.

Der Spezialforschungsbereich „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (SFB DiÖ)

Im Vortrag soll anhand des (Aufbaus des) Annotationssystems des Spezialforschungsbereichs „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (FWF F60) (kurz: SFB DiÖ) und dessen Teilprojekts 11 „Kollaborative Online-Forschungsplattform“ die Annotation von Nonstandardvarietäten hinterfragt werden. Der SFB DiÖ beschäftigt sich mit der Vielfalt und dem Wandel der deutschen Sprache in Österreich. Er erforscht den Gebrauch und die subjektive Wahrnehmung von deutscher Sprache in Österreich und untersucht die Einflüsse von Kontaktsprachen auf sie. Der SFB ist an vier wissenschaftlichen Institutionen in Österreich angesiedelt: an den Universitäten Wien, Graz und Salzburg sowie an der Österreichischen Akademie der Wissenschaften. Dabei sind unterschiedliche Forschungsbereiche und -institute beteiligt: von der Germanistik über die Slavistik bis hin zur Translationswissenschaft und der Schallforschung. Die drei thematischen Taskcluster befassen sich mit den inhaltlichen Säulen Variation, Kontakt und Perzeption und werden von zwei Taskclustern für die Organisation und Koordination sowie die Kollaborative Online-Forschungsplattform ergänzt und unterstützt (s. Abb. 1, vgl. Budin et al. 2018).

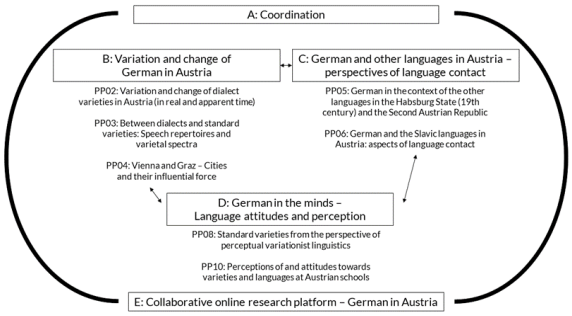


Abbildung 1. SFB-DiÖ Taskcluster und Teilprojekte

Durch die verschiedenen beteiligten Institute bedarf es schon innerhalb des SFB einer großen Flexibilität und Variabilität im Annotationssystem, da in den verschiedenen Teilprojekten nicht nur unterschiedliche Fragestellungen und linguistische Systemebenen untersucht werden, sondern dies auch aus unterschiedlichen disziplinären Perspektiven und in unterschiedlicher Granularität. Hinzu kommt, dass nicht nur sehr viele, sondern auch sehr heterogene empirische Daten (Methodenpluralismus) erhoben werden, welche einheitlich und methodenübergreifend annotiert werden sollen (vgl. Abb.2). Die Anzahl der dadurch entstehenden Annotationen sowie deren Vokabular ist sehr umfangreich. Da eines der Ziele der Forschungsplattform aber ist, die annotierten Daten einem möglichst breiten Publikum zur Verfügung zu stellen, muss das Annotationssystem auch „nach außen“ (intersubjektiv) nachvollziehbar sein und sich, wo möglich, an bereits bestehende Standards halten (z. B. bereits vorhandene Tagsets). Um diesem multidimensionalen Anspruch gerecht zu werden, wurde und wird ein eigenes Annotationssystem entwickelt, was auf ggf. vorhandene Standards (in seinem Vokabular) zurückgreift. Die Besonderheit liegt dabei in einer Zweiteilung von technischer Speicherung und Repräsentation.

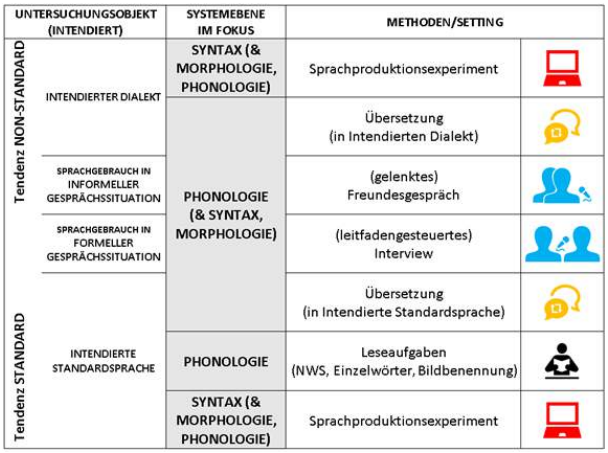


Abbildung 2. Erhebungssettings des SFB DiÖ

Das Annotationssystem des SFB DiÖ

Ziel der Annotation ist es, ein einheitliches Gesamtbild über alle Systemebenen der gesprochenen Sprache hinweg zu verfolgen. Die Annotation wird dabei phänomenbezogen auf einer Annotationsebene je betrachtetem Phänomen vorgenommen. Neben den Systemebenen wie Phonetik/Phonologie, Morphologie/Lexik und Syntax kommen auch qualitativ-inhaltliche sowie ggf. gesprächsanalytische Annotationen hinzu. Um diesem multidimensionalen Anspruch gerecht zu werden, wurde und wird ein eigenes Annotationssystem entwickelt, was auf ggf. vorhandene Standards (in seinem Vokabular) zurückgreift. Die Besonderheit liegt dabei in einer Zweiteilung von technischer Speicherung und Repräsentation. Die Speicherung funktioniert linear innerhalb einer relationalen Datenbank auf mehreren Tagebenen, die zeitgleich im Mehrbenutzerzugriff bearbeitet werden können. Die Repräsentation wird jedoch hierarchisch geliefert. Dies hat den Vorteil, das Annotierende durch die komplexe Annotation geleitet werden und weniger Obacht auf die richtige Reihenfolge der Tags legen müssen, da sich diese aus der Hierarchie automatisch ergibt.

Das vorgestellte Annotationssystem ist insofern hierarchisch, dass es eine child-parent-Struktur für ein spezifisches phänomenbezogenes Tagset fordert. Dies bedeutet, dass einerseits immer Kategorien zu den zu taggenden Features angegeben werden müssen, andererseits ihr Aufbau sinnvoll beschrieben werden muss. Dadurch wird die Kohärenz und Nachvollziehbarkeit des Tagsets erhöht. Auch in der praktischen Anwendung wird hierdurch die Annotation vereinfacht, indem durch die Vorgabe der Struktur selbige ebenso beim Annotationsprozess abgerufen wird: Durch diesen Vorgang wird es erst möglich, dass z.B. eine Tagging-Eingabemaske aus den unzähligen vorhanden Tags

jene herausfiltern kann, die für Annotierende in diesem Moment relevant sind.

Wenn diese Hierarchie jedoch auch in die Speicherung Eingang finden würde, wären Probleme vorprogrammiert. Die Annotationen werden daher linear gespeichert, um eine größtmögliche datenstrukturelle Flexibilität zu ermöglichen. Dies ist insbesondere dann nötig, wenn das Annotationssystem iterativ zum Forschungsprozess erweiterbar sein soll, wie es im SFB DiÖ der Fall ist.

Vergleich von Annotationssystemen

Im Vortrag wird jedoch nicht nur das Annotationssystem des SFB Deutsch in Österreich vorgestellt, sondern auch mit anderen Annotationssystemen verglichen. Hierfür wird insbesondere der Bereich des Part-of-Speech-Taggings herausgegriffen. Unter Part-of-Speech-Tagging oder PoS-Tagging wird die automatische Zuweisung von Wortarten verstanden. Es wird jedem Token automatisch durch einen Tagger eine Wortart zugewiesen (vgl. Lemnitzer/Zinsmeister 2010: 72).

Vergleichssysteme sind etwa das Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) oder das European Dialect Syntax (EDiSyn)-Tagset (Barbiers/Wyngaerd o.J.). Trotz der häufigen Verwendung des STTS ist es in verschiedenen Punkten relativ problematisch. Ziel ist einmal nur die Wortartenbestimmung selbst, zudem im Falle des häufig verwendeten großen Tagsets die Repräsentation von Morphologie und Derivation (vgl. Telljohann et al. 2013: 2). Das Tagset ist in Anbetracht verschiedener Klassifizierungskategorien aufgebaut, die Kategorisierung erfolgte nach morphologischen, syntaktischen sowie semantischen Kriterien (vgl. Schiller et al. 1999: 4). Es besteht aus elf Hauptwortarten, welche unterschiedlich tief klassifiziert sind (vgl. Schiller et al. 1999: 5). Diese für einige Forschungsfragen sehr hilfreichen Subklassifizierungen stellen jedoch auch ein Problem dar, da keine Einheitlichkeit im Aufbau des Tagvokabulars gegeben ist und die Auswahl nicht ohne Vorkenntnisse zu diesem Aufbau ersichtlich ist. Durch den uneinheitlichen Aufbau in Bezug auf Analysetiefe sowie der Struktur der Teilelemente im Tag ist das Set einerseits nur für eingeschränkte Zwecke verwendbar und – wenn dieses Problem ausgeglichen werden soll – nur schwierig nachvollziehbar um die dazu nötigen, neuen Teilelemente eines Tags ergänzbar.

Dies wird insbesondere beim PoS-Tagging von Nonstandardvarietäten ersichtlich, das eine Herausforderung darstellt, da viele varietätenspezifische Ausprägungen von Wortarten bzw. Features von Wortarten und Besonderheiten dieser nicht abgedeckt werden. Eine klarere Strukturierung und weitaus mehr Möglichkeiten bietet hier das European Dialect Syntax (EdiSyn)-Tagset, welches zumindest in Ansätzen die Wortarten als Kategorien und deren Ausprägungen (grammatische Features) als Features der Kategorien modelliert, d.h. eine Trennung von Tags auf verschiedenen Ebenen durchführt. Die Features selbst werden wiederum nicht benannten Kategorien (grammatischen Kategorien wie Kasus, Genus etc.) zumindest zur Gruppierung zugeordnet. Der große Vorteil daran ist, dass Wortarten und Features unabhängig voneinander erweitert werden können und dass Features potentiell auch mehreren Wortarten zugeschrieben werden können – was nicht nur für den sprachtypologischen Vergleich verschiedener Sprachen sinnvoll erscheint, sondern auch, da innerhalb ein und derselben Sprache dieselben Features für unterschiedliche Wortarten auftreten. Diesen Ansatz erweitert das im SFB genutzten Annotationssystems, setzt jedoch für Kategorien und Features eigene Generationen an und lässt damit eine m:n-Verbindung zu.

Standardisierungsversuch von Annotationen

Darauf basierend wird herausgearbeitet, bis zu welchem Grad eine Standardisierung von Annotationssystemen, insbesondere auch sprachen- und varietätenübergreifend arbeitenden Systemen, möglich, handhabbar und sinnvoll ist. Schließlich soll untersucht werden, inwiefern mit Abweichungen von einer Normierung umgegangen werden kann, inwiefern Mehrsprachigkeit (auch im Sinne der „inneren Mehrsprachigkeit“, vgl. Wandruszka (1979)) die Standardisierungsmöglichkeiten beeinträchtigt, bzw. inwiefern Möglichkeiten bestehen, ein Annotationssystem zu modifizieren und zu erweitern, um trotz der großen und nicht trivialen Anforderungen einer Standardisierung standhalten zu können. Hierbei ist vor allem zu hinterfragen, ob eine Standardisierung des Vokabulars möglich und nötig ist oder ob vielmehr die Struktur und Beschreibung der zu standardisierende Aspekt der Annotation ist.

Forschungsfragen

Ziel des Vortrags ist es, sich den folgenden Forschungsfragen zu widmen:

- Welche Anforderungen entstehen an ein Annotationssystem für variationslinguistische Daten und insbesondere Nonstandardvarietäten und wie können diese erfüllt werden? Welche Anforderungen entstehen durch Methoden- und Theorienpluralismus?
- Inwiefern ist ein solches Framework standardisierbar bzw. welche Aspekte davon?
- Welche Vor- und Nachteile birgt ein standardisiertes Annotationssystem?

Bibliographie

Barbiers, Sjef / Wyngaerd, Guido Vanden (o.J.): *Tagging protocol*. http://www.meertens.knaw.nl/pdf/variatielinguistiek/dialectsyntax/Tagging-protocol.pdf [zuletzt abgerufen 13. Oktober 2018].

Breuer, Ludwig M. / Seltmann, Melanie E.-H. (2018): *Sprachdaten(banken) – Aufbereitung und Visualisierung am Beispiel von SyHD und DiÖ* in: **Börner, Ingo / Straub, Wolfgang / Zolles, Christian (eds):** *Germanistik digital.* Digital Humanities in der Sprach- und Literaturwissenschaft. Wien: Facultas, 135-152.

Budin, Gerhard / Elspaß, Stephan / Lenz, Alexandra N. / Newerkla, Stefan M. / Ziegler, Arne 2018): *Der Spezialforschungsbereich ‚Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption‘* in: Zeitschrift für

germanistische Linguistik 46(2), 300-308. DOI: 10.1515/zgl-2018-0017.

Lemnitzer, Lothar / Zinsmeister, Heike (2010): *Korpuslinguistik*. Tübingen: Gunter Narr Verlag.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS* (Kleines und großes Tagset) http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [zuletzt abgerufen 13. Oktober 2018].

Telljohann, Heike / Versley, Yannick / Beck, Kathrin / Hinrichs, Erhard / Zastrow, Thomas (2013): *STTS als Part-of-Speech-Tagset in Tübinger Baumbanken* in: **Zinsmeister, Heike / Heid, Ulrich / Beck, Kathrin (eds.):** *Das Stuttgart-Tübingen Wortarten-Tagset – Stand und Perspektiven.* Journal for Language Technology and Computational Linguistics 28, 1/2013. 1-15.

Wandruszka, Mario (1979): *Die Mehrsprachigkeit des Menschen*. München: Piper.

Klassifikation von Titelfiguren in deutschsprachigen Dramen und Evaluation am Beispiel von Lessings „Emilia Galotti“

Krautter, Benjamin
Benjamin.Krautter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pagel, Janis
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einführung:

In seiner Studie zu Gotthold Ephraim Lessings bürgerlichem Trauerspiel *Emilia Galotti* (1772) formuliert Gisbert Ter-Nedden: „Der erste Akt war der Selbstdarstellung des Prinzen und der höfischen Partei gewidmet; der zweite gehört den Galottis, und zwar zunächst Odoardo als dem Gegenspieler des Prinzen: nach dem Machthaber betritt der Tugendheld die Bühne“ (Ter-Nedden 1986: 189). Die Titelfigur Emilia geht in seiner Darstellung im Kollektiv der Galottis unter. Prinz und Vater Odoardo werden hingegen als die beiden Gegenspieler hervorgehoben. Sind sie also gemäß Ter-Nedden Protagonist und Antagonist, die Helden und Hauptfiguren des Stücks?

Zur Einteilung und Abstufung des Dramenpersonals mithilfe quantitativ erfassbarer Kriterien führt Manfred Pfister in den späten 1970er Jahren die Terminologie der „quantitative[n] Dominanzrelationen“ (Pfister 2001 [1977]: 226) ein. Er benennt hierfür zwei Kriterien, die Haupt- von Nebenfiguren unterscheiden sollen: die „Dauer der Bühnenpräsenz einer Figur“ (ebd.) und den Anteil der Figurenrede am Haupttext. Laut Pfister fehle es allerdings an einer differenzierten Handlungsgrammatik, die auch funktionale Relationen – etwa die aktiven Handlungsschritte von Figuren – operationalisieren könne (vgl. ebd.: 227). Er wirbt letztlich für einen nicht weiter explizierten multidimensionalen Ansatz, der die quantitative Einteilung des Personals zuverlässiger und feingliedriger machen soll.

An diese Idee der quantitativen und zugleich multidimensionalen Einteilung dramatischer Figuren versuchen wir im Folgenden mittels digitaler Analysetechniken anzuschließen (vgl. Fischer u.a. 2018). Dazu fassen wir das Problem der Figureneinteilung als Klassifikationsaufgabe. Zielsetzung ist es, titelgebende Dramenfiguren mit maschinellen Lernverfahren automatisch auszuzeichnen. Dadurch sind wir in der Lage, die genauen Einflussfaktoren zu prüfen und die Ergebnisse transparent zu evaluieren. Nach unserem Dafürhalten sind es zumindest drei Gründe, die titelgebende Dramenfiguren zu einer geeigneten Zielkategorie der Klassifikation machen. Den möglichen alternativen Konzepten mangelt es erstens an einer konsensfähigen Definition und Differenzierung. Das gilt insbesondere für die Begriffe ‚HeldIn‘ und ‚ProtagonistIn‘, die teilweise synonym verwendet werden (vgl. etwa Plett 2002: 21f., Jannidis 2004: 90, 104f.). Überdies ist gerade das Heldenkonzept stark von literaturgeschichtlichen Entwicklungen geprägt und somit historisch variabel (vgl. etwa Alt 1994: 167f., Platz-Waury 2007 [1997]: 591, Martus 2011: 15). Die intersubjektive Annotation der Figurenkategorien bereitet zweitens Schwierigkeiten, vor allem dann, will man ProtagonistInnen oder HeldInnen mit Blick auf ihre Bedeutung für die Handlung bzw. den zentralen Konflikt des Dramas bestimmen.[1] Zieht man dagegen die gegebenen Titelfiguren der Texte heran, entfällt dieses Annotationsproblem. Drittens lässt sich *a priori* annehmen, dass titelgebende Figuren Eigenschaften verkörpern und im Text repräsentieren, die sie in vielen Fällen auch zu Hauptfiguren der Handlung machen.[2]

Stand der Forschung:

In der aktuellen Forschung folgen mehrere Studien der von Wladimir Propp (1986 [1928]) am russischen Volksmärchen eingeführten Figurentypologisierung, um literarische Figuren auf formaler oder automatischer Basis (sub-)klassifizieren zu können (etwa Declerck / Koleva / Krieger 2012 oder Finlayson 2017). Moretti (2011 und 2013) nutzt indessen Netzwerkdarstellungen von Shakespeares *Hamlet* für eine Neukontextualisierung der dramatischen Figureninteraktion. Mit seiner Formalisierung argumentiert er gegen die Dichotomie zwischen ProtagonistInnen und Nebenfiguren und für netzwerkanalytische Zentralitätsmaße.[3] Frank Fischer u.a. (2018) konzentrieren sich auf die dramengeschichtliche Einordnung von Protagonisten in deutschsprachigen Stücken (vgl. auch Algee-Hewitt 2017). Sie nutzen einen multidimensionalen Ansatz, der Netzwerkmetriken mit zählbasierten Maßen kombiniert. Im Gegensatz dazu versuchen Jannidis u.a. (2016) Hauptfiguren in deutschsprachigen Romanen automatisch zu klassifizieren. Als Goldstandard verwenden sie annotierte Zusammenfassungen der Romane.

Versuchsaufbau:

Das untersuchte Korpus umfasst 38 Dramen mit mindestens einer Titelfigur, deren Veröffentlichung sich von der Mitte des 18. bis ins frühe 20. Jahrhundert erstreckt.[4] *Tabelle* 1 gibt einen Überblick über das Verhältnis titelgebender und nicht-titelgebender Figuren. Demnach werden lediglich drei Prozent aller Figuren durch die Titel der Stücke repräsentiert. Bei der Auswahl haben wir darauf geachtet, sowohl ein breites Spektrum literarischer Epochen als auch verschiedene dramatische Strukturen zu berücksichtigen. August Wilhelm Schlegels *Ion* kommt etwa mit sechs Figuren aus, Grabbes *Napoleon oder Die hundert Tage* sieht 183 Figuren vor. Schillers *Die Verschwörung des Fiesko zu Genua* umfasst insgesamt 75 Szenen, Hofmannsthals *Elektra* lediglich fünf.

#Dramen	#Titelfigur (%)	#Nicht-Titelfigur (%)	#Figuren Total
38	42 (3)	1166 (97)	1208

Tabelle 1. Zahl der Titelfiguren und nicht-Titelfiguren.

Für das maschinelle Klassifikationsverfahren nutzen wir mit *Random Forest* (Ho 1995, Breiman 2001) einen Algorithmus, der Entscheidungsbäume zu einem Ensemble fügt und die Parameter mittels mathematischer Regression berechnet.[5] Da die einzelnen erlernten Entscheidungsbäume eingesehen werden können, ist *Random Forest* auch dafür geeignet, die entscheidungstragenden Merkmale näher zu untersuchen. Vorausgehende Experimente mit *Support Vector Machines* (SVM) erzielten zudem schlechtere Ergebnisse. *Random Forest* scheint besser geeignet zu sein, um viele und potentiell korrelierende Features zu verarbeiten.

Um dem Vorsatz eines multidimensionalen Modells gerecht zu werden, kombinieren wir als Features zählbasierte Metriken (*Tokens*: Anzahl der gesprochenen Tokens normalisiert nach Textlänge), Netzwerkrelationen (*Degree*, *weighted Degree*, *betweenness Centrality*, *closeness Centrality*, *eigenvector Centrality*), die Bühnenpräsenz der Figuren (*Active*: Figur spricht innerhalb einer Szene; *Passive*: Figur wird innerhalb einer Szene namentlich genannt), *Topic Modeling* und Metadaten (etwa Zahl der Dramenfiguren, Epochenzugehörigkeit).[6]

	Precision (TF)	Recall (TF)	F1 (TF)	Precision (C)	Recall (C)	F1 (C)	MCC
Majority BL	-	0	-	0.97	1.00	0.98	0.00
Tokens BL	0.34	1	0.51	1.00	0.93	0.96	0.56
alle Features	0.45	1	0.62	1.00	0.96	0.98	0.66

Tabelle 2. Klassifikationsergebnisse von vier Modellen für Titelfiguren (TF) und nicht-Titelfiguren (C): Majority Baseline, Tokens Baseline und alle Features.

Diskussion der Klassifikationsergebnisse:

Die *F1-Scores* der Klassifikation in *Tabelle 2* zeigen, dass das gelernte Modell (alle Features) in der Lage ist, sinnvolle Vorhersagen zu treffen, die über die Heuristiken der Majority und Tokens Baselines hinausgehen. Die verwendeten Features scheinen also tatsächlich Annäherungen an die Eigenschaften titelgebender Figuren wiederzugeben. Die Ergebnisse verdeutlichen aber auch, dass das Modell übergeneralisiert. Zwar wird jede titelgebende Figur als solche erkannt (*Recall* [alle Features] 1.00), das System tendiert jedoch dazu, auch nicht-titelgebende Figuren als Titelfigur zu identifizieren (*Precision* [alle Features] 0.45). Das deutet darauf hin, dass die exponierte Stellung der Figuren im Titel des Dramas zwar durchaus im Text Realisierung findet, andere (Haupt-)Figuren die spezifischen Eigenschaften aber dennoch teilen können. Nur in elf der 38 Dramen werden alle Figuren richtig ausgezeichnet.[7] *Abbildung 1* und *Abbildung 2* zeigen die *Feature Importance* und die Verteilung der Features. Im Entscheidungsprozess des Algorithmus sind die *Tokens* das wichtigste Feature. Die Leistungsfähigkeit der Tokens Baseline ließ das bereits erahnen. Aber auch einige *Topics*, Netzwerkmaße und die Bühnenpräsenz tragen zu den Klassifikationsergebnissen bei. Das erlernte *Topic Model* erreicht beispielsweise auch ohne die Zugabe weiterer Features Klassifikationsergebnisse, die im Bereich der Tokens Baseline liegen (vgl. Reiter u.a. 2018: 3). Um einen besseren direkten Vergleich der Modelle zu ermöglichen, geben wir auch Werte für den *Matthews Correlation Coefficient* (MCC) an. Der MCC ist weder auf eine einzelne Klasse fokussiert noch wird er von starken Klassenunterschieden beeinflusst. Auch hier zeigt sich, dass das Modell mit allen Features den beiden Baselines überlegen ist.

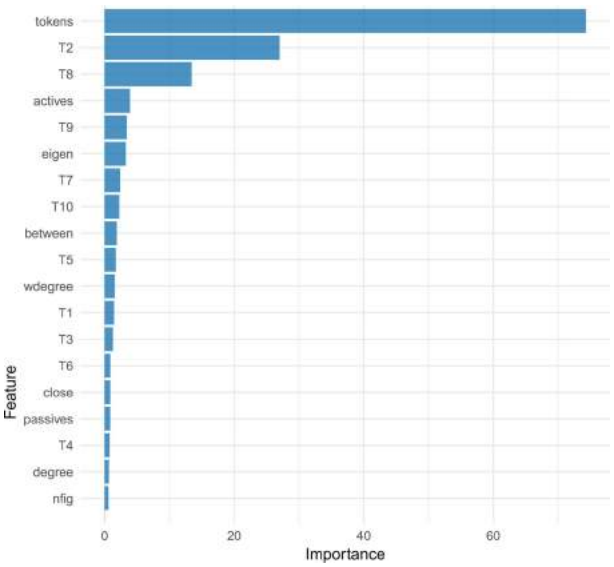


Abbildung 1. Die *Feature Importance* vergleicht die Performanz des Modells, wenn eines der Features entfällt. Die Abnahme an Performanz entspricht der relativen Wichtigkeit des jeweiligen Features für die Klassifikation (Importance). T1–T10 stehen für die zehn erlernten *Topics*.

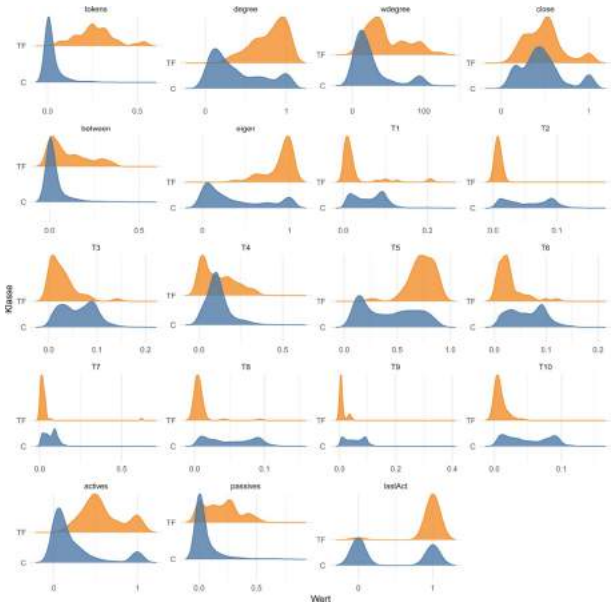


Abbildung 2. Featureverteilung der Klassifikation: Die Skalen auf der x-Achse geben den Wertebereich des jeweiligen Features wieder, die y-Achse zeigt die Verteilung der Feature-Werte auf die jeweiligen Klassen (TF: Titelfigur, C: nicht-Titelfigur).

Emilia Galotti als passiv präsente Titelfigur:

Anhand Lessings *Emilia Galotti* versuchen wir, die aufgeführten Klassifikationsergebnisse genauer nachzuvollziehen. Wir streben hierbei keine Interpretation des Stücks an, sondern wollen auf Grundlage unserer quantitativen Analysen erste Beobachtungen schildern, die sich hauptsächlich auf die Struktur des Dramas beziehen. Zwar beinhaltet die Klassifikation auch *Topic Modeling* Ergebnisse, die erlernten *Topics* sind aber nur begrenzt interpretierbar: sie geben kaum Rückschlüsse auf die Semantiken der Figurenreden. Die Featureverteilung in *Abbildung 2* verdeutlicht, dass lediglich *Topic* 5 ein positives Unterscheidungsmerkmal für die Auszeichnung als Titelfigur ist. Unter den zwanzig wahrscheinlichsten Begriffen von *Topic* 5 finden sich jedoch ausnahmslos Funktionswörter.[8] Die Begriffe sind zudem oftmals Teil weiterer *Topics*, die dann zwar auch semantische Ausdrücke beinhalten, sich aber auf Anredeformeln, Angaben zur sozialen Stellung der Figur oder Berufsbezeichnungen beschränken.[9]

Gleich vier Titelfiguren benennt die Klassifikation für Lessings bürgerliches Trauerspiel *Emilia Galotti*. Neben der tatsächlich titelgebenden Emilia werden ihr Vater Odoardo Galotti, der Prinz Hettore Gonzaga und dessen Kammerherr Marinelli als Titelfiguren ausgezeichnet. Betrachtet man ausschließlich die beiden von Pfister angeführten Kriterien quantitativer Dominanzrelationen, würde man Emilia wohl kaum als zentrale Figur des Dramas wahrnehmen. Mit nur 2363 Tokens spricht Emilia nicht nur weniger als der Prinz und Marinelli, die mit jeweils knapp über 5500 Tokens die umfassendsten Redeanteile am Haupttext haben, sie spricht auch weniger als ihr Vater Odoardo und die Gräfin Orsina (*Abbildung 3*). Des Weiteren ist sie lediglich in sieben der 43 Dramenszenen aktiv präsent (*Abbildung 4*).

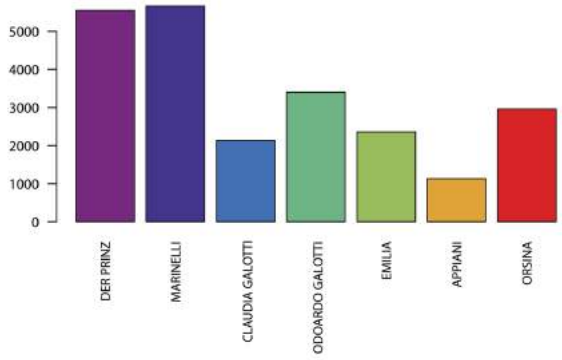


Abbildung 3. Redeanteile in Lessings *Emilia Galotti* gemessen in Tokens. Aufgeführt sind die sieben Figuren mit dem größten Redeanteil am Haupttext.

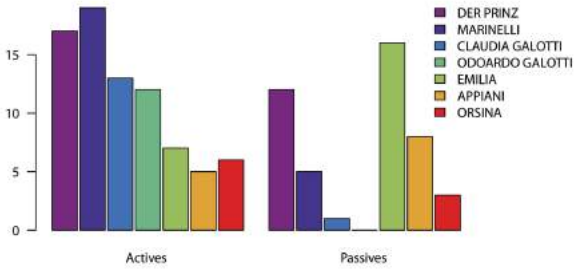


Abbildung 4. Aktive und passive Präsenz in *Emilia Galotti* gemessen an der Zahl der Szenen. Passiv präsent ist eine Figur nur dann, wenn sie in dieser Szene nicht selbst aktiv ist.

Warum also wird Emilia trotzdem als Titelfigur erkannt? Die Featureanalyse in *Abbildung 5* zeigt, dass neben den *Tokens* vor allem die *Topics* 8 und 2 sowie die *betweenness Centrality* den größten Einfluss auf die Klassifikation haben. Dass die Zahl der gesprochenen Wörter positiv bewertet wird überrascht, fällt sie doch so deutlich geringer aus als bei Marinelli oder dem Prinzen. Diese Figuren sprechen zwar mehr als Emilia, verglichen mit anderen Dramen aber wohl nicht übermäßig viel mehr. Der hier als gering wahrgenommene Umfang der Figurenrede reicht somit für den Klassifikator aus, um Emilia zumindest in die Riege potentieller Titelfiguren zu heben – auch, da in Lessings Drama insgesamt recht wenige Figuren auftreten.

Für den Leser ist es dagegen wohl eher Emilias durchgängige passive Präsenz in den Dialogen und Monologen anderer Figuren, die sie als Titelfigur kennzeichnet. Exemplarisch dafür steht bereits der erste Akt. Ausgelöst durch eine Bittschrift verliert sich der Prinz in unruhigen Gedanken an Emilia Galotti: „Ich kann doch nicht mehr arbeiten. – Ich war so ruhig, bild‘ ich mir ein, so ruhig – Auf einmal muß eine arme Bruneschi, Emilia heißen: – weg ist meine Ruhe, und alles! –“ (Lessing 2000 [1772]: 293). Mit wechselnden Gesprächspartnern – Maler Conti, Marinelli und Camillo Rota – wird Emilia immer wieder zum Mittelpunkt der folgenden Dialoge. Das gilt insbesondere für die sechste Szene, als

Marinelli die anstehende Vermählung Emilias mit dem Grafen Appiani preisgibt, woraufhin der eifersüchtige Prinz seinem Kammerherren Marinelli die völlige Handlungsfreiheit in dieser Angelegenheit zugesteht (vgl. ebd., 300–305). Diese passive Präsenz Emilias lässt sich über weite Teile des Dramas nachvollziehen. In 16 Szenen wird in der Figurenrede mit ihrem Namen auf sie referiert, obwohl sie selbst zu diesem Zeitpunkt nicht aktiv am Bühnengeschehen beteiligt ist. Verglichen mit anderen Dramenfiguren ist dieser Wert sehr hoch. Marinellis Name wird beispielsweise nur in fünf Szenen aufgerufen, auf Emilias Mutter Claudia wird ohne ihre Anwesenheit auf der Bühne gar nicht namentlich referiert, wie *Abbildung 4* veranschaulicht.

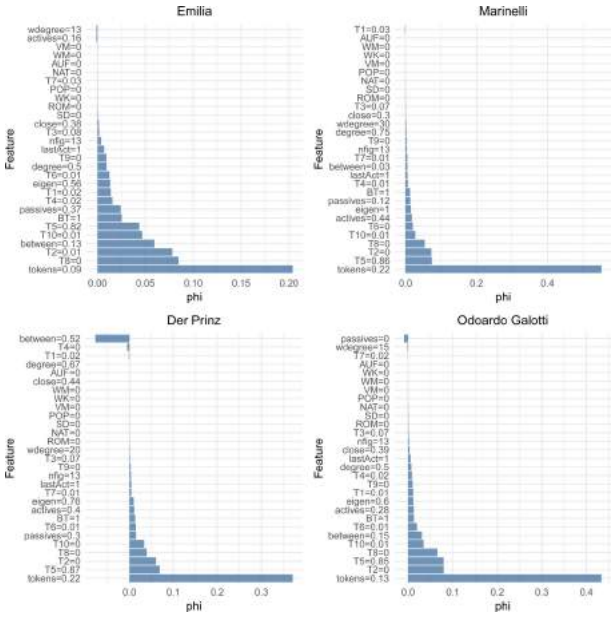


Abbildung 5. *Feature Importance* der vier als titelgebend klassifizierten Figuren in *Emilia Galotti*.

Fazit und Ausblick:

Wir konnten zeigen, dass unser multidimensionales Modell sinnvolle Ergebnisse für die Klassifikation titelgebender Figuren liefert (MCC 0.66). Titelfiguren werden sehr zuverlässig erkannt (*Recall* 1.00), das Modell neigt jedoch zur Übergeneralisierung. Die Beobachtungen anhand Lessings *Emilia Galotti* lassen aber annehmen, dass die Übergeneralisierung genau die Figuren umfasst, die man als handlungstragende Hauptfiguren des Stücks bezeichnen könnte (vgl. Fick 2000: 337.) Diesen Einzelbefund wollen wir in Zukunft durch die Auswertung und Analyse zusätzlicher Dramen systematisieren. Eine weitere Aufgabe wird es sein, das Erlernen der *Topics* besser nachzuvollziehen und auf die semantische Interpretierbarkeit hin zu optimieren, ohne dabei die Leistungsfähigkeit für die Klassifikation einzuschränken.

Anhang der untersuchten Dramen:

Arnim, L. A. von: Marino Caboga
Brentano, C.: Ponce de Leon
Büchner, G.: Dantons Tod
Büchner, G.: Leonce und Lena
Büchner, G.: Woyzeck
Goethe, J. W.: Götz von Berlichingen mit der eisernen Hand
Goethe, J. W.: Iphigenie auf Tauris
Goethe, J. W.: Torquato Tasso
Gottsched, J. Ch.: Der sterbende Cato
Grabbe, Ch. D.: Don Juan und Faust
Grabbe, Ch. D.: Hannibal
Grabbe, Ch. D.: Herzog Theodor von Gothland
Grabbe, Ch. D.: Napoleon oder Die hundert Tage
Gutzkow, K.: Richard Savage, Sohn einer Mutter
Gutzkow, K.: Uriel Acosta
Hofmannsthal, H. von: Elektra
Hofmannsthal, H. von: Ödipus und die Sphinx
Kotzebue, A. von: Die beiden Klingsberg
Laube, H.: Monaldeschi
Laube, H.: Struensee
Lessing, G. E.: Emilia Galotti
Lessing, G. E.: Miss Sara Sampson
Pfeil, J. G. B.: Lucie Woodvil
Romantik Iffland, A. W.: Figaro in Deutschland
Schiller, F.: Die Jungfrau von Orléans
Schiller, F.: Die Piccolomini
Schiller, F.: Die Verschwörung des Fiesko zu Genua
Schiller, F.: Maria Stuart
Schiller, F.: Wallensteins Tod
Schiller, F.: Wilhelm Tell
Schlaf, J.: Meister Oelze
Schlegel, A. W.: Alarkos
Schlegel, A. W.: Ion
Schlegel, J. E.: Canut
Schnitzler, A.: Anatol
Schnitzler, A.: Professor Bernhardi
Tieck, L.: Der gestiefelte Kater
Tieck, L.: Prinz Zerbino
Tieck, L.: Ritter Blaubart
Uhland, L.: Ludwig der Bayer
Wieland, Ch. M.: Klementina von Porretta
Wieland, Ch. M.: Lady Johanna Gray

Fußnoten

1. Bei einem entsprechenden Versuch erzielten wir Übereinstimmungen zwischen 0.43 und 0.83 (Cohen's ϰ).
2. Titelfiguren könnten dann ersatzweise als (Pseudo-)Goldstandard für die Klassifikation von Hauptfiguren genutzt werden, um zu testen, ob Eigenschaften der Titelfiguren auf andere Hauptfiguren übertragbar sind.
3. Probleme von Morettis Vorgehen diskutiert Peer Trilcke (2013: 226–232).
4. Wir nutzen dafür das German Drama Corpus: https://github.com/dracor-org/gerdracor.
5. Die Implementierung erfolgt über die Pakete randomForest und Caret für R: https://cran.r-project.org/

web/packages/randomForest/index.html/ und https://cran.r-project.org/web/packages/caret/. Der vollständige Code zu den Experimenten ist zu finden unter https://quadrama.github.io/blog/2018/12/12/detect-protagonists.de/.
6. Eine genauere Beschreibung aller Features findet sich in Krautter, Benjamin / Pagel, Janis / Reiter, Nils und Willand, Marcus (2018): „Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen", in: *Litlab Pamphlet* 7.
7. Vollständig korrekt ausgezeichnet wurden die folgenden Dramen: *Die beiden Klingsberg*, *Marino Caboga*, *Die Piccolomini*, *Wallensteins Tod*, *Die Verschwörung des Fiesko zu Genua*, *Ludwig der Bayer*, *Die Jungfrau von Orleans*, *Herzog Theodor von Gothland*, *Dantons Tod*, *Hannibal*, *Napoleon oder Die hundert Tage*.
8. Die 20 wahrscheinlichsten Begriffe von *Topic* 5 sind: „ich", „und", „die", „nicht", „der", „ist", „zu", „–", „das", „Ich", „in", „so", „den", „dem", „es", „ein", „mich", „sie", „Sie", „er".
9. Etwa: „Herr", „König", „Königin", „Vater", „Narr", „Graf", „Majestät", „Schuldmeister", „Mutter", „Tochter", „Sohn", „Doktor".

Bibliographie

Algee-Hewitt, Mark (2018): *Distributed Character: Quantitative Models of the English Stage, 1500-1920*, in: Digital Humanities 2017: Conference Abstracts: 119–121.

Alt, Peter-André (1994): *Tragödie der Aufklärung*. *Eine Einführung*. Basel / Tübingen: Francke.

Breiman, Leo (2001): *Random Forests*, in: Machine Learning 24(2): 5–32.

Declerck, Thierry / Koleva, Nikolina / Krieger, Hans-Ulrich (2012): *Ontology-Based Incremental Annotation of Characters in Folktales*, in: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities: 30–34 http://www.aclweb.org/anthology/W12-1006 [zuletzt abgerufen 13. Oktober 2018].

Fick, Monika (2000): *Emilia Galotti*, in: Lessing Handbuch. Leben–Werk–Wirkung. Stuttgart: Metzler 316–343.

Finlayson, Mark A. (2017): *‚ProppLearner': Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory*, in: Digital Scholarship in the Humanities 32(2): 284–300 https://doi.org/10.1093/llc/fqv067 [letzter Zugriff 13. Oktober 2018].

Fischer, Frank / Trilcke, Peer / Milling, Carsten / Skorinkin, Daniil (2018): *To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)*, in: Digitial Humanities 2018: Conference Abstracts: 193–201.

Ho, Tin Kam (1995): *Random Decision Forests*, in: Proceedings of the 3rd International Conference on Document Analysis and Recognition: 278–282.

Jannidis, Fotis (2004): *Figur und Person. Beitrag zu einer historischen Narratologie*. Berlin: de Gruyter.

Jannidis, Fotis / Reger, Isabella / Krug, Markus / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): *Comparison of Methods for the Identification of Main Characters in German Novels*, in: Digital Humanities 2016: Conference Abstracts: 578–582.

Krautter, Benjamin / Pagel, Janis / Reiter, Nils / Willand, Marcus (2018): *Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen*, in: Litlab Pamphlet 7.

Lessing, Gotthold Ephraim (2000 [1772]): *Emilia Galotti. Ein Trauerspiel in fünf Aufzügen*, in: **Bohnen, Klaus (ed.):** *Lessing Werke und Briefe.* Bd. 7. Werke 1770–1773. Frankfurt a.M.: Deutscher Klassiker-Verlag 291–371.

Martus, Steffen (2011): *Transformationen des Heroismus. Zum politischen Wissen der Tragödie im 18. Jahrhundert am Beispiel von J. E. Schlegels Canut*, in: **Burkhard, Torsten / Hundt, Markus / Martus, Steffen / Ort, Claus-Michael (eds.):** *Politik – Ethik – Poetik. Diskurse und Medien frühneuzeitlichen Wissens*. Berlin: Akademie Verlag 15–42.

Moretti, Franco (2011): *Network Theory, Plot Analysis*, in: Literary Lab Pamphlet 2. https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [zuletzt abgerufen 13. Oktober 2018].

Moretti, Franco (2013): *‚Operationalizing': or, the Function of Measurement in Modern Literary Theory*, in: Literary Lab Pamphlet 6. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf [zuletzt abgerufen 13. Oktober 2018].

Pfister, Manfred (2001 [1977]): *Das Drama*. *Theorie und Analyse*. München: Fink.

Platz-Waury, Elke (2007 [1997]): *Art. Figurenkonstellation*, in: **Weimar, Klaus / Fricke, Harald / Grubmüller, Klaus / Müller, Jan-Dirk (eds.):** *Reallexikon der deutschen Literaturwissenschaft*. Bd. 1. Berlin / New York: De Gruyter 591–593.

Plett, Bettina (2002): *Problematische Naturen? Held und Heroismus im realistischen Erzählen*. München / Paderborn / Wien / Zürich: Schöningh.

Reiter, Nils / Krautter, Benjamin / Pagel, Janis / Willand, Marcus (2018): *Detecting Protagonists in German Plays around 1800 as a Classification Task*, in: EADH 2018: Conference Abstracts. http://dx.doi.org/10.18419/opus-10162 [zuletzt abgerufen 22.12.2018].

Ter-Nedden, Gisbert (1986): *Lessings Trauerspiele. Der Ursprung des modernen Dramas aus dem Geist der Kritik*. Stuttgart: Metzler.

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.

Korpuserstellung als literaturwissenschaftliche Aufgabe

Gius, Evelyn

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

Katharina, Krüger

katharina.krueger@uni-hamburg.de
Universität Hamburg, Deutschland

Carla, Sökefeld
carla.soekefeld@studium.uni-hamburg.de
Universität Hamburg, Deutschland

Korpuserstellung in der (digitalen) Literaturwissenschaft

Die Praxis der Zusammenstellung von Primärtexten zu einem Korpus ist gewissermaßen literaturwissenschaftliches Alltagsgeschäft, trotzdem wird sie selten problematisiert. Die beiden Standardfälle der nicht-digitalen Korpusanalyse erscheinen auch bezüglich der Korpuszusammenstellung unproblematisch: (1) Die Forschungsfrage erfordert eine bestimmte Textbasis (etwa bei einer Untersuchung zu Krankheitsdarstellungen bei Thomas Mann), (2) Die Korpuserstellung basiert auf der Kanonizität der Texte. Zumindest im zweiten Fall ist das geeignete Vorgehen allerdings nicht selbstverständlich, denn es müsste begründet werden, nach welchen Kriterien Werke tatsächlich exemplarisch und repräsentativ sind. Dies wird jedoch kaum thematisiert (vgl. Gius, in Vorbereitung).

Bei der konkreten Erstellung eines *digitalen* Korpus ist wiederum aufgrund der Menge der verfügbaren Texte häufig nur eine balancierte Sammlung nach vordefinierten Kriterien realisierbar, die allerdings die Gefahr von unerwünschten Korrelationen birgt (Schöch 2017).[1] Trotz dieser Problematik wird die methodologische Bedeutung der Korpuszusammenstellung auch in den *Digital Humanities* kaum wahrgenommen . So ergibt eine Stichwortsuche nach „Korpus“ bzw. „corpus“ in allen Tagungsbänden der *Digital Humanities-* und *Digital Humanities im deutschsprachigen Raum*-Konferenzen von 2014-2017 jeweils hunderte Beiträge. Allerdings gehen nur acht der 49 Abstracts mit literaturwissenschaftlicher Thematik explizit auf die Korpus *erstellung* ein, obwohl in den übrigen durchaus auch von repräsentativen Korpora die Rede ist. Lediglich fünf[2] dieser Beiträge beschäftigen sich dezidiert mit der Problematik eines literaturwissenschaftlichen Korpus.

Ein Grund für die geringe Auseinandersetzung mit dem Thema kann sein, dass die Praxeologie der Zusammenstellung literarischer digitaler Korpora verhältnismäßig neu ist und sich bislang keine literaturwissenschaftlichen Routinen etablieren konnten, die das Zusammenstellen des Untersuchungsobjektes betreffen.[3] Ein anderer Erklärungsansatz ist, dass schon in der traditionellen Literaturwissenschaft der “literaturwissenschaftliche Objektumgang vor allem durch implizite Normen strukturiert” ist (Schruhl 2018) und entsprechend auch eine Übersetzung dieses nicht-digitalen Zugangs in einen digitalen Zugang – also eine Art Operationalisierung der Korpuserstellung – nicht möglich ist.

Die digitale literaturwissenschaftliche Korpuserstellung stellt also ein ungelöstes methodologisches Grundproblem dar.

Wir erläutern im Folgenden die Problematik der literaturwissenschaftlichen Korpuserstellung exemplarisch, um eine Diskussion darüber anzustoßen sowie mögliche Lösungsansätze vorzustellen.

Erstellung eines Korpus zur Untersuchung von genderspezifischer Darstellung von Krankheit

Im Rahmen unseres Forschungsprojektes zu genderspezifischer Darstellung von Krankheit in literarischen Texten im Forschungsverbund hermA (“Automatische Modellierung hermeneutischer Prozesse”)[4] geht es um die Frage nach der genderspezifischen Darstellung von Krankheit bei literarischen Figuren um 1900. In einem korpusbasierten Ansatz werden Methoden für die Analyse der Darstellung von Krankheit entwickelt. Da wir einen konstruktivistischen Ansatz in Bezug auf Genderkonzepte verfolgen und außerdem grundsätzlich davon ausgehen, dass sich semantische Aspekte in literarischen Texten nur bedingt an der Textoberfläche materialisieren,[5]mussten wir Strategien entwickeln, um ein möglichst großes Ausgangskorpus zu etablieren. Die Grundlage zur Erstellung dieses Korpus war das Kolimo-Korpus nach Herrmann und Lauer (2017). Es enthält über 42.000 Texte mit einem Fokus auf der Zeitspanne von 1880 bis 1930 aus dem Deutschen Textarchiv, dem TextGrid Repository und dem Projekt Gutenberg-DE (vgl. Tabelle 2).

Da die Entwicklung von Strategien zur Bildung themenspezifischer Subkorpora ebenfalls Ziel des Forschungsvorhabens ist, gab es für die Textauswahl zunächst keine inhaltlichen Einschränkungen. Kriterien zur Textauswahl waren deshalb nur das Datum der Erstveröffentlichung, Textsprache, das Genre und Textlänge; Grundlage für ihre Ermittlung bildeten die Kolimo-Metadaten.

Kriterium	Relevanter Bereich	Zweifelsfälle	Festlegung durch:
Erscheinungsdatum	1870-1920	Zeitspannen, widersprüchliche Angaben, fehlende Angaben	Metadaten + Recherche zu Ersterscheinungsdaten bzw. Lebensdaten
Sprache	Deutschsprachig (Standardsprache), keine Übersetzungen		Automatisierte Überprüfung: >90% deutschsprachig
Gattung/Genre	Prosa und dazugehörige Genres	Kinder- und Jugendbuch (im Kernkorpus) Autobiographie, Biographie, Fabel, Märchen, Memoire, Tagebuch (nicht im Kernkorpus) Ohne spezifisches Genre	Metadaten Kategorisierung der Angaben in den Metadaten Recherche
Textlänge	> 1.000 Wörter	Kapitelweise vorliegende Texte u.ä.	automatische Wortzählung

Tabelle 1: Übersicht der Kriterien für die Korpuserstellung

Ins Korpus aufgenommen wurden Texte mit Erstveröffentlichung zwischen 1870 und 1920. Die Auswahl wurde anhand der vorliegenden Metadaten getroffen und bei Bedarf weiter ergänzt. Je nach Repositorium unterschied sich die zusätzlich nötige Recherchearbeit stark: elf DTA-Texte, 7.602 Gutenberg-Texte, und 27.300 TextGrid-Texte hatten keine Datumsangabe.

Da wir uns auf Phänomene konzentrieren, die in der deutschsprachigen zeitgenössischen Literatur verhandelt wurden, lag der Fokus auf standarddeutschen Texten. Übersetzungen wurden in ein Sonderkorpus aufgenommen, zu weniger als 90% deutschsprachige Texte aussortiert.
Außerdem wurden ausschließlich Prosatexte berücksichtigt, wobei für die (Sub-)Genres Autobiographie, Biographie, Fabel,

Märchen, Memoire und Tagebuch eigene Sonderkorpora erstellt wurden. Texte der Kategorie Kinder- beziehungsweise Jugendbuch wurden hingegen in das Kernkorpus miteinbezogen.
Um Kürzestprosa auszuschließen, die sich meist deutlich von der narrativen Struktur anderer Formen unterscheidet, wurden nur Texte mit über 1.000 Wörtern aufgenommen.

Das nach diesen Kriterien manuell erstellte Kernkorpus umfasst nur noch etwa 2.700 Werke.

Korpus	Texte (DTA/Textgrid/Gutenberg)
Ausgangskorpus Kolimo	42.710 (1.317 / 27.412 / 13.981)
Kernkorpus	2.726 (50 / 1.137 / 1.539)
Sonderkorpus Märchen	2.601
Sonderkorpus Übersetzungen	465
Weitere Sonderkorpora	69

Tabelle 2: Übersicht Korpusgrößen

Korpusbereinigung

Die dargestellten Schwierigkeiten bei der Entscheidung über die Aufnahme eines Textes in das Kernkorpus basierten auf unvollständigen, widersprüchlichen oder nicht ohne weiteres aufeinander abbildbaren Metadaten und konnten anhand der dargelegten Setzungen vergleichsweise einfach gelöst werden. Weitaus schwieriger gestaltete sich hingegen die Identifikation von Dubletten und die Konzipierung einer geeigneten Problembehandlung – ein leider typisches Problem bei der Aggregation eines Korpus aus verschiedenen Quellen.

So zeigte sich bei einer ersten manuellen Durchsicht der Liste und der stichprobenartigen Überprüfung der Volltexte, dass Dubletten nicht eindeutig anhand von Metadaten identifizierbar sind. Die naheliegende Lösung, ein Volltextvergleich, müsste jedoch bei 2.700 Texten über sieben Millionen mögliche Paare vergleichen und würde auch bei der Nutzung von Cluster Computing Jahrzehnte dauern. Deshalb mussten Heuristiken entwickelt werden, um das Verfahren abzukürzen.

Entsprechend haben wir einen Workflow entworfen, um mit möglichst hoher Treffgenauigkeit Dubletten zu identifizieren. Ziel war, alle echten Dubletten zu finden und das Korpus entsprechend zu bereinigen, ohne Texte vorschnell zu streichen.

Dafür wurden folgende, größtenteils automatische Prüfungen durchgeführt:[6]

1. Identifikation eindeutiger Dubletten:
   - Werke, denen derselbe Volltext zugeordnet wurde
   - Werke mit einer Edit-Distanz (Levensthein 1966) ≤ 2 bei Autor und Titel
2. Identifikation Dublettenkandidaten:
   - Werke mit einer Edit-Distanz ≤ 2 bei Autor (und mehr beim Titel)
3. Volltextvergleich Dublettenkandidaten:
   - gemessene einseitige Edit-Distanz (in beiden Richtungen):
     - ≤15%: Dublette
     - ≥ 80%: keine Dublette
   - alle anderen Fälle: manuelle Prüfung (vgl. Tabelle 3)
4. Entfernung verbleibender Texte mit
   - ≤ 1.000 Wörter und/oder
   - ≥ 10% nichtdeutscher Textanteil

Bei den verbleibenden Textpaaren gehen wir von echten Dubletten aus. Für diese erfolgt eine Repositorien-Priorisierung nach der Qualität der Repositorien (1. DTA, 2. TextGrid, 3. Gutenberg).

Problem	Bsp.	Lösungsansatz
Text enthält zusätzlichen Paratext	"Hymnen" (Ferdinand von Saar) mit Vorwort des Herausgebers	Variante ohne Paratext
Eindeutig stark veränderte Ausgabe	"Die Pilger der Wildnis" (Johannes Scherr) als "für die Jugend bearbeitete Ausgabe"	"Ursprungsvariante" wählen
Mehrbändiges Werk (≠ Dubletten)	"Auch Einer" (Friedrich Theodor von Vischer, 1879): Band 1 und 2 mit zwei nahezu gleichlautenden Dateinamen	zusammenführen
Teile eines Sammelbandes (≠ Dubletten)	"Der Schmied seines Glückes" von Gottfried Keller, erschienen in "Die Leute von Seldwyla"	aufteilen

Tabelle 3: Dubletten: Lösungsansätze bei manueller Überprüfung

Ansätze für die literaturwissenschaftliche Korpuserstellung?

Die Digitalisierung fördert die literaturwissenschaftliche Korpuserstellung in neuem Umfang und macht dadurch die Textzusammenstellung als literaturwissenschaftliches Problem offensichtlich, das methodologisch kaum beleuchtet ist. Oft dominiert die Frage, welche Texte überhaupt ins Korpus können, also aus welchen bereits digitalisierten oder noch digitalisierbaren Texten ausgewählt werden kann, die literaturwissenschaftlichen Überlegungen zur Textauswahl.

Da die Menge digitalisiert vorliegender Texte stetig steigt, haben Korpora außerdem immer häufiger eine Größe, die von einzelnen Forscher/innen nicht mehr überschaubar ist. Deshalb muss sich die literaturwissenschaftliche Begründung der Relevanz der Texte in einem Korpus einer gewissen Größe im Zweifelsfall auf Aspekte, die als Informationen in den Metadaten der Texte vorliegen – wie Zeit, Gattung/Genre oder Autor/innen – beschränken.

Sowohl die Menge der zur Verfügung stehenden Texte als auch die Qualität der Primär- und Metadaten sind noch steigerungsfähig. Hier ist die Forschungsgemeinschaft genauso gefordert wie Bibliotheken und Archive.

Im Sinne einer wissenschaftlichen Qualitätssicherung ist es daher umso wichtiger, im Hinblick auf das *zur Verfügung* stehende Datenmaterial Qualitätskriterien für die Zusammenstellung der Korpora zu entwickeln und umzusetzen, wie das vorgestellte Verfahren zeigen soll. So können aus Texten mit schlechten oder fehlenden Metadaten vergleichsweise homogene Korpora erstellt und anschließend mit weiteren Informationen angereichert werden.

Da das literaturwissenschaftliche Wissen über ein Korpus mit steigender Korpusgröße notwendigerweise ab- und damit die Gefahr unerwünschter Korrelationen zunimmt, sollte das Korpus außerdem mit Informationen angereichert werden, die für die Interpretation von Analyseergebnissen genutzt werden können (Epochenzugehörigkeit, thematische Zuordnung etc). Dadurch können entdeckte Korrelationen auf einer größeren Basis an Texteigenschaften – also: besser – interpretiert werden. Da das manuelle Ergänzen

von Informationen sehr arbeitsaufwändig ist, sollten dafür (halb-)automatische Verfahren entwickelt bzw. genutzt werden.[7]

Die Anforderungen an ein Korpus variieren je nach Kontext und Forschungsfrage des Projekts, deshalb müssen allgemeine Qualitätskriterien für die Korpuserstellung eine gewisse Offenheit aufweisen. Grundsätzlich gilt aber: Für die Korpuserstellung muss frühzeitig eine Strategie zur Priorisierung von Problemen entwickelt – und dokumentiert! – werden. Dabei geht es darum, sich wie in diesem Beitrag skizziert eine Übersicht über konzeptuelle und technische Aspekte zu verschaffen und anschließend einfach zu lösende Probleme und konzeptuell wichtige Entscheidungen in eine geeignete Reihenfolge zu bringen – den Workflow für die Korpuserstellung.

Fußnoten

1. Zu den verschiedenen Typen von Datensammlungen bzw. Korpora vgl. ebenfalls Schöch (2017).
2. Farrar (2016): Corpus of Revenge Tragedy; Herrmann und Lauer (2016a/b): Kafka/Referenzkorpus; Herrmann und Lauer (2017): Kolimo; Pernes et al. (2017): historisch-literarisches Metaphernkorpus.
3. Trilcke & Fischer (2018) sprechen hier davon, dass das "epistemische Ding" der Literaturwissenschaft sich vom (einzelnen) Primärtext hin zu durch Weiterverarbeitung entstandene Zwischenformate und Korpora entwickelt.
4. Vgl. https://www.herma.uni-hamburg.de und Gaidys et al. (2017).
5. So wird häufig eine für den Texte zentrale Krankheit nur andeutungsweise erzählt, etwa in Schnitzlers Novelle *Sterben* (1894), in der die Krankheit des Protagonisten über den Großteil des Textes ausschließlich über die Symptom- und Behandlungsbeschreibungen als Tuberkulose erkennbar ist.
6. Für die Mitarbeit bei der Planung und die Implementierung der Verfahren danken wir Benedikt Adelmann.
7. Zum Beispiel durch die Einbindung von fachlichen Wissensbasen wie etwa den in der Deutschen Nationalbibliothek verfügbaren bibliographischen Metadaten.

Bibliographie

Farrar, Danielle Marie (2016): *The Corpus of Revenge Tragedy (CoRT): Toward Interdisciplining Early Modern Digital Humanities and Genre Analysis*. In: DH 2016 - Conference Abstracts. S. 789–790.

Gaidys, Uta/Gius, Evelyn/Jarchow, Margarete/Koch, Gertraud/Menzel, Wolfgang/Orth, Dominik/Zinsmeister, Heike (2017): *Project Description. HermA: Automated Modelling of Hermeneutic Processes*. In: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Gius, Evelyn (in Vorbereitung): *Digitale Hermeneutik: Computergestütztes close reading als literaturwissenschaftliches Forschungsparadigma?* In: **Jannidis, Fotis (Hg.)**: *Digitale Literaturwissenschaft*. Metzler.

Herrmann, Berenike/Lauer, Gerhard (2017): *Das „Was-bisher-geschah" von KOLIMO. Ein Update zum Korpus der literarischen Moderne*. In: DHd 2017 Digitale Nachhaltigkeit Konferenzabstracts. S. 107–111.

Herrmann, Berenike/Lauer, Gerhard (2016a): *Aufbau und Annotation des Kafka/Referenzkorpus*. In: DHd 2016 Modellierung - Vernetzung - Visualisierung Konferenzabstracts. S. 158–159.

Herrmann, Berenike/Lauer, Gerhard (2016b): *KARREK: Building and Annotating a Kafka/Reference Corpus*. In: DH 2016 - Conference Abstracts. S. 552–553.

Levenshtein, Vladimir (1966): *Binary codes capable of correcting deletions, insertions, and reversals*. In: Soviet Physics Doklady. Vol. 10, No. 8, S. 707–710.

Pernes, Stefan/Keller, Lennart/Peterek, Christoph (2017): *Aufbau eines historisch-literarischen Metaphernkorpus für das Deutsche*. In: DHd 2017 Digitale Nachhaltigkeit Konferenzabstracts. S. 92–94.

Trilcke, Peer/Fischer, Frank (2018): *Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen*. In: Zeitschrift für digitale Geisteswissenschaften, Sonderband 3 "Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden".

Schöch, Christof (2017): *Aufbau von Datensammlungen*. In: **Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.)**: Digital Humanities: eine Einführung. Stuttgart: J.B. Metzler Verlag. S. 223–233.

Schruhl, Friederike (2018): *Objektumgangsnormen in der Literaturwissenschaft*. In: Zeitschrift für digitale Geisteswissenschaften, Sonderband 3 "Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden".

Makroanalytische Untersuchung von Heftromanen

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

Konle, Leonard
leonardkonle@gmail.com
Universität Würzburg, Deutschland

Leinen, Peter
P.Leinen@dnb.de
Deutsche Nationalbibliothek, Frankfurt a.M.

Einleitung

Heftromane, früher als 'Romane der Unterschicht' (Nusser 1981) abgewertet, sind aufgrund eines weniger wertungsfreudigen Umgangs mit Populärliteratur (Hügel 2007, Kelleter 2012) in den letzten 10-15 Jahren wieder Gegenstand der Literaturforschung geworden (z.B. Nast

2017, Stockinger 2018). 'Heftromane' wurden immer definiert durch das eigene Publikationsformat (zumeist rd. 64 Seiten), eigene Formen der Distribution über den Zeitschriftenmarkt und nicht über den Buchhandel, und auch die Soziographie der Heftromanleser weicht deutlich von der der sonstigen Literatur ab. Im Folgenden berichten wir über erste Ergebnisse einer Auswertung von 9.000 deutschsprachigen Heftromanen aus den Jahren 2009-2017. Möglich wurde die Forschung durch eine Kooperation zwischen der Würzburger Arbeitsgruppe zur literarischen Textanalyse und der Deutschen Nationalbibliothek (DNB), die die Daten vorhält. Ziel dieser ersten, noch weitgehenden explorativen Studie, sind die Antworten auf zwei Fragen: Wie unterscheiden sich die Gattungen der Heftromane untereinander und wie unterscheiden sich die Heftromane von Hochliteratur? Im ersten Schritt gehen wir der Frage nach, wie gut sich die Gattungen klassifizieren lassen und welche Texteigenschaften dabei eine Rolle spielen. Im zweiten Schritt werden die Gattungen inhaltlich erfasst. Zuletzt geht es um die angeblich einfachere Sprache der Heftromane.

Die Analyse stützt sich auf die digitalen Texte, die an die Deutsche Nationalbibliothek abgeliefert wurden. Die Deutsche Nationalbibliothek (DNB) sammelt, archiviert, verzeichnet im gesetzlichen Auftrag die ab 1913 in Deutschland veröffentlichten Medienwerke sowie die im Ausland veröffentlichten deutschsprachigen Medienwerke, Übersetzungen deutschsprachiger Medienwerke in andere Sprachen und fremdsprachige Medienwerke über Deutschland und stellt diese der Öffentlichkeit zur Verfügung. Seit der Gesetzesnovelle von 2006 gehört auch das Sammeln von Medienwerken, die online publiziert werden, ausdrücklich zu den Aufgaben der DNB.

Der Bestand der DNB umfasst derzeit etwa 5 Millionen digitale Objekte, darunter ca. 900.000 E-Books, ca. 1,5 Millionen E-Journal Ausgaben und ca. 2 Millionen E-Paper Ausgaben. Neben dem umfangreichen physischen Bestand steht den Nutzerinnen und Nutzern der DNB damit ein wachsender Fundus von "born digital" Objekten zur Verfügung.

Die Anforderungen an die Informationsversorgung haben sich durch den digitalen Wandel insgesamt stark verändert. Die Einführung neuer Forschungsmethoden wie z.B. automatisierte Daten- und Textanalysen großer digitaler Bestände gehen mit der Notwendigkeit veränderter Formen der Bereitstellung von Beständen einher.

Die DNB sieht in der Kooperation mit den DH-Communities eine strategisch wichtige Fortsetzung[1] ihrer Dienstleistungs- und Benutzungs-Angebote. Ein Aspekt dabei ist die Unterstützung ausgewählter Kooperationspartner durch die Bereitstellung auch umfangreicher Korpora vorrangig aus born digital Objekten wie E-Books sowie einer leistungsfähigen Infrastruktur für die Durchführung automatischer Analysen. Die letzte Urheberrechtsreform eröffnet den registrierten Nutzerinnen und Nutzern der DNB diese Möglichkeit in den Räumen der DNB und auf deren Infrastruktur.

In der Vorverarbeitung wurden, soweit das aufgrund der selbst innerhalb eines Verlags sehr heterogenen Ausgangslage automatisch möglich war, Werbung, Leseproben usw. entfernt. Problematische Texte wurden nicht für die Analyse verwendet. Die unausgewogene Verteilung in Abbildung 1 kommt durch die Neupublikation älterer Hefte zustande. Die Verteilung über die Gattungen (Abbildung 2) ist sehr unausgewogen. Abb. 3 zeigt, dass manche der Gattungen von einzelnen Serien dominiert werden. Außerdem haben wir ein Vergleichskorpus 'Hochliteratur' mit 500 Romanen von Autoren erstellt, die einen literarischen Preis gewonnen haben oder dafür vorgeschlagen wurden.

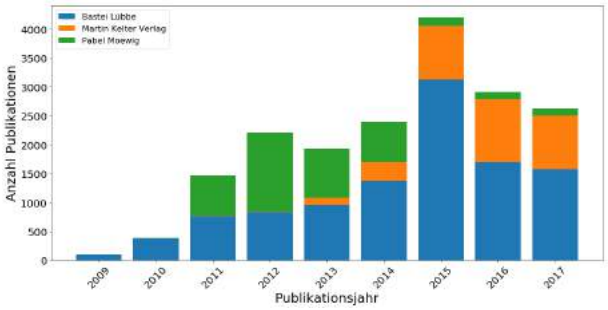


Abbildung 1. Anzahl der Publikationen nach Verlag und Erscheinungsjahr

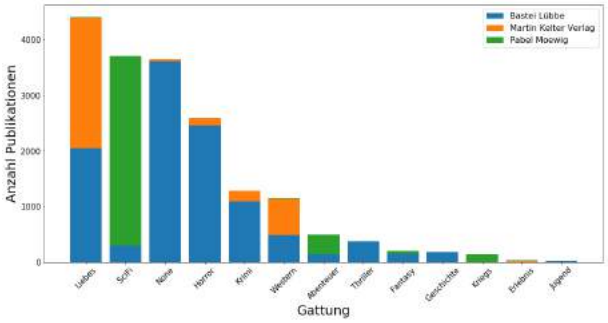


Abbildung 2. Anzahl der Publikationen nach Gattung und Verlag

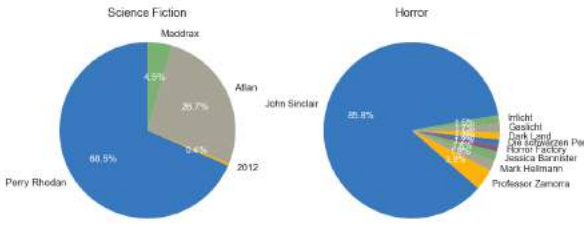


Abbildung 3. Dominanz großer Serien

Methoden

Gattungen erkennen. Um einen Eindruck der Kohärenz der Gattungen aus inhaltlicher Perspektive zu erhalten, wird eine Document-Term-Matrix mit den 8000 häufigsten Substantiven verwendet und eine Dimensionsreduktion mit umap (McInnes 2018) durchgeführt, um eine zweidimensionale Darstellung zu ermöglichen. Um die Leistungsfähigkeit der Substantive für die Klassifikation der Texte nach Gattungen zu prüfen, wurde überwachtes Lernen mit durch Logistic Regression durchgeführt. Außerdem

werden die Texte aufgrund eines stilistischen Maßes gruppiert: Kosinus Delta (Evert et. al 2017) der 2000 häufigsten Worte, deren Leistungsfähigkeit wird ebenfalls durch eine Klassifikation getestet.

Themen und Topoi. Wir verwenden zwei Verfahren, um Themen, Settings, Gegenstände, Figuren, rekurrente Formulierungen usw. der Gattungen zu identifizieren: Topic Modeling und Zeta. Ein Model (100 Topics) (Blei, Jordan, Ng 2002) wird über das gesamte Korpus mit Mallet 2.0.8 (McCallum 2002) erstellt. Zeta wird verwendet, um zu ermitteln, welche Worte in einer Gruppe von Texten im Vergleich mit einer zweiten bevorzugt werden (Burrows 2007, Craig, Kinney 2009). Für unsere Untersuchung wurden jeweils für jede Gattung 200 Texte dieser Gattung und 200 Texte aus allen anderen Gattungen zufällig gezogen und Eder's Zeta mit Stylo berechnet (Eder, Kestemont, Rybicki 2016); Parameter nach Empfehlungen in (Schöch et al. 2018).

Gattungskomplexität und Kontrast zu Hochliteratur . Zur Prüfung der Hypothese, dass Schemaliteratur sprachlich weniger komplex sei als Hochliteratur, wurde Vokabular und Syntax untersucht: Die lexikalische Komplexität wurde durch ein standardisiertes type-token-ratio (sttr, Fenstergröße 10.000) (Kubát, Milička 2013) sowie die Wortlänge ermittelt . Die syntaktische Komplexität wird zum einen durch die durchschnittliche Satzlänge und zum anderen durch die Variabilität von POS-Tags untersucht.

Ergebnisse

Gattungen erkennen. In früheren Arbeiten mit Texten des 19. Jahrhunderts konnten wir stilometrisch (most frequent words) einige Gattungen sehr gut (z.B. Abenteuerroman), andere nur schlecht unterscheiden (z.B. Bildungsroman), während die thematischen Unterschiede (topic models) schlechtere Klassifikationsergebnisse erbrachten (Hettinger et al. 2016). Die Heftroman-Gattungen bilden sowohl inhaltlich (Abbildung 4) als auch stilistisch (Abbildung 5) recht deutlich abgegrenzte Einheiten. Diesen visuellen Eindruck bestätigen die Ergebnisse der Klassifikation mit den 8000 häufigsten Substantiven mit einem 0.90 F1 (macro) und 0.94 F1 (micro). Eine Klassifikation auf Basis der Distanzmatrix erreicht leicht schlechtere Werte von 0.78 F1 (macro) und 0.91 F1 (micro).

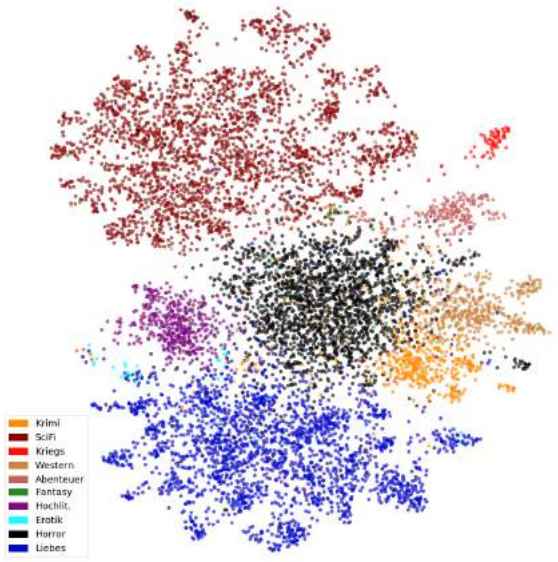


Abbildung 4. Clustering der Texte auf Basis der 8000 most frequent nouns

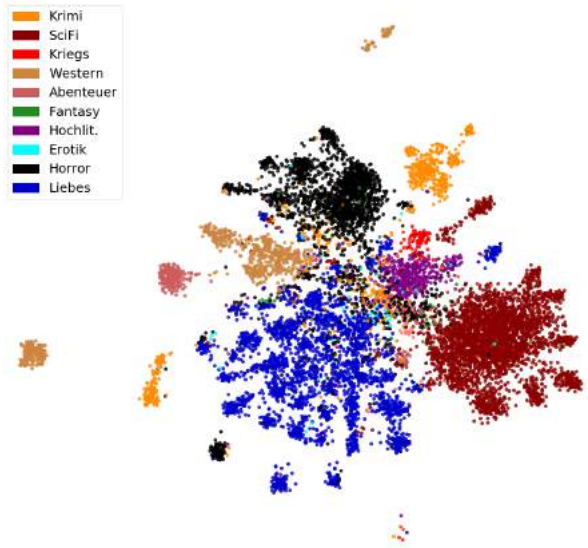


Abbildung 5. Transformation der Distanzmatrix (cosine delta, 2000 mfw)

Themen und Topoi. Das Topic Model eignet sich nur begrenzt zur inhaltlichen Erschließung des Korpus, da die ermittelten Topics zum größten Teil nicht interpretierbar sind. Einige wenige funktionieren sehr gut, z.B. in Abbildung 6 die wichtigsten Topics für die Gattung 'Western'. Sie enthalten zentrale Entitäten (Pferde, Wagen), vor allem aber verdeutlichen sie, dass Kampfhandlungen wichtig für die Gattung sind.

Abbildung 6. Häufigste Topics der Gattung "Western"

Allerdings sind zahlreiche Topics zwar statistisch diskriminativ, aber aus literaturwissenschaftlicher Perspektive undankbar. So ist es offensichtlich, dass das Topic in Abb. 7 Kommunikationswörter als typisch für den Liebesroman aufführt. Aber die Worte des Topic in Abb. 8, fast ebenso diskriminativ für Fantasy und SciFi, überschneiden sich damit in vielen Punkten.

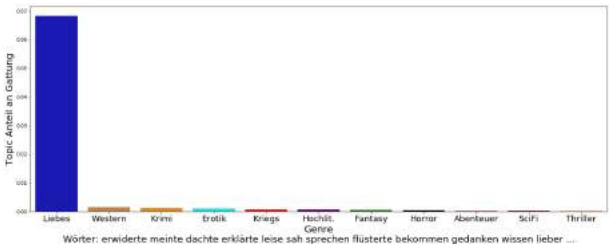


Abbildung 7. Kommunikationswörter des Gattung Liebesroman

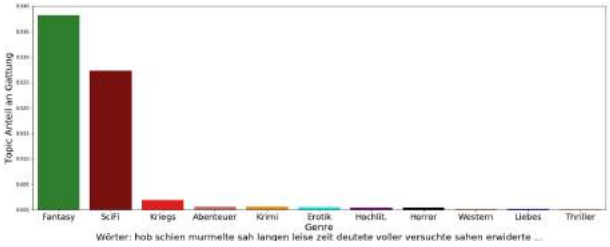


Abbildung 8. Distinktives Topic für Fantasy und SciFi

Wie das Clustering der Texte aufgrund der Topics in Abbildung 9 zeigt, können diese als Proxies für die Verteilungen von Worten in Romanen dienen, aber sie erschließen die Texte - anders als das in vielen vergleichbaren Untersuchungen der Fall ist - nur in wenigen Fällen inhaltlich.

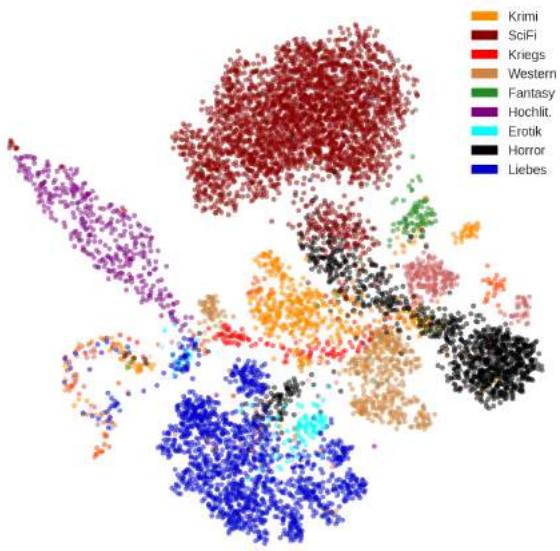


Abbildung 9. Transformation der Verteilung der Topics pro Dokumenten mit umap

Die Wörter, die aufgrund von Zeta, von der Gattung bevorzugt werden, leisten dagegen die inhaltliche Erschließung der Gattungen ausgesprochen gut (siehe die Beispiele in Abbildung 10 und 11): Vor allem Figuren (z.B. Seeleute, Fürsten, Oberarzt, usw.) sowie Objekte und Settings (z.B. Bordwand, BH, Bergwald) entsprechen den Erwartungen, lassen aber zugleich einen Einblick in die Spezifika der Genres zu, z.B. dass es sich im Falle der Science Fiction um eine langandauernde 'space opera' handelt oder dass der Handlungsort der Liebesromane häufig ein Oberschichtmilieu ist. Nur im Fall der Hochliteratur ist die Divergenz der sprachlichen Register / der Begriffe auffällig groß. Dieser Effekt schlägt sich in Abbildung 12 nieder, es ist zu beobachten, dass die Kohärenz, hier interpretiert als Maß für die Geschlossenheit einer Gruppe von Worten, der Zeta Wörter für Hochliteratur in etwa der einer Zufallsstichprobe entspricht. Wie gut die Zeta-Wörter die Gattungen repräsentieren, wird daran deutlich, dass eine Klassifikation (svm, 150 Zetawörter pro Gattung, 600 Texte aus jeder Gattung mit Oversampling, wo notwendig) einen F1 (micro) score von 0.90 erreicht.

Abenteuer	Adels	Arzt	Erotik	Horror	Hochlit
aye	schlosspark	infusion	orgasmus	geisterjäger	klo
spanier	fürsten	zillertal	t-shirt	rover	it
achtern	gräfin	fee	slip	silberkugeln	me
bordwand	baron	sprechstunde	nippel	silberkugel	hitler
ausguck	fürstin	innsbruck	reißverschluss	beretta	texte
außenbords	schlossbewohner	röntgen	po	geweihten	for
kahn	durchlaucht	grünwald	brustwarze	dämons	weltkrieg
gesegelt	hinzusetzte	spatz	schamlippen	zombie	on
degen	fürst	oberarzt	klasse	vampiren	andauernd
holzbein	teenagern	facharzt	strähnen	untoten	wischt
pullen	privaträume	hausarzt	warmer	blutsauger	juden
backbordseite	schlosshof	gebärmutter	penis	friedhofs	russland
mast	fürstenpaar	operationssaal	duschen	abbé	cola
seemann	boxer	pflegerin	nässe	augenhöhlen	wörter
wikinger	cousin	assistenzarzt	angefühlt	dämonischen	präsidenten
backbord	standesgemäß	bergdoktor	lecken	luzifer	what
großmast	bediensteten	notarztwagen	unterleib	dämonenpeitsche	christus

Abbildung 10. Genretypische Wörter (Zeta)

Heimat	Kriegs	Krimi	Liebes	SciFi	Western
madl	leutnant	streifenwagen	dienerschaft	galaxis	saloon
bissel	mg	ford	gnädigen	raumschiff	hufschlag
brotzeit	munition	field	diwan	planet	texas
tonis	russen	officer	gnädiges	universums	winchester
bös	russischen	schalldämpfer	anerbieten	schleuse	cowboys
obstler	deutscher	handschellen	unbeschreiblichen	weltraum	cowboy
förster	einschläge	dienstwaffe	vornehmer	hangar	camp
ausschaut	flugzeuge	detective	gottlob	schutzschirm	well
gell	oberst	brooklyn	destille	raumschiffs	reitern
leut	funker	notebook	teetisch	raumfahrer	county
trenker	meldet	inspektoren	liebenswürdigkeit	jahrtausenden	kansas
bergwald	flanke	ermittlung	mancherlei	terra	creek
bursch	lastwagen	mafia	reizendes	lichtjahre	banditen
bergführer	feindes	plaza	umzukleiden	humanoiden	mountains
feschen	pistole	ganoven	umkleiden	geortet	karabiner
gerad	ne	bewußte	namenlos	unsterblichen	arizona
bergtour	flieger	datenbanken	frohen	projektion	indianern

Abbildung 11. Genretypische Wörter (Zeta)

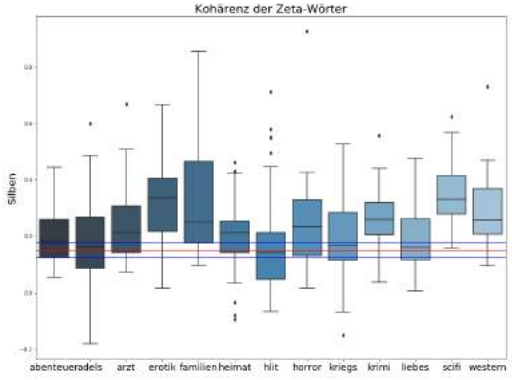


Abbildung 12. Kohärenz der Zetawörter mittels word embedding. Rote/ blaue Linien: Wert für zufällig gewählte Worte. Rot: Mittelwert, Blau +/- eine Standardabweichung

Gattungskomplexität und Kontrast zu Hochliteratur . Die Annahme, dass Heftromane weniger komplex seien als Hochliteratur ist schon Teil des Namens: Schemaliteratur. Zugleich ist es offensichtlich, dass literarische Komplexität unterschiedliche Aspekte betreffen kann. Die hier untersuchten Aspekte sind teilweise von der ideologiekritischen Trivialliteraturforschung bereits an kleinen Samples untersucht worden, die im Fall der Heftromane zu dem Ergebnis kommt, dass der Wortschatz kleiner sei, die durchschnittliche Satzlänge kürzer und die Komplexität der Sätze geringer (Nusser 1982: 88f.) Zwar wurde die ideologiekritische Forschung zur populären Literatur in den letzten Jahren kritisiert, die quantitativen Forschungen sind jedoch nicht durch neue ersetzt worden.

Wir nehmen das Verhältnis von Types zu Tokens in einem Text als Maß für die Variabilität der Sprache und als Größe des Wortschatzes. Wie die Boxplots in Abbildung 13 zeigen, unterscheidet sich Hochliteratur (‘hlit’) in diesem Punkt keineswegs grundlegend vom Heftroman. t-Test.

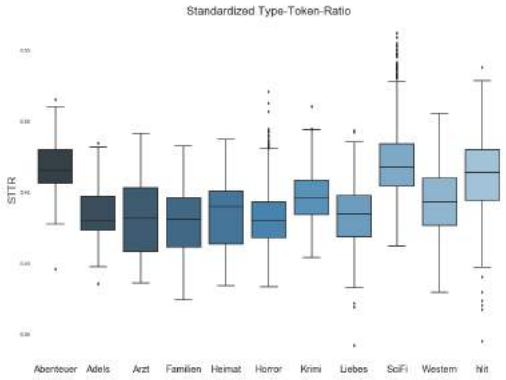


Abbildung 13. Standardized Type-Token-Ratio

Auch bei der durchschnittlichen Länge der Worte, häufig verwendet für Maße der Leseschwierigkeit von Texten, ist die Varianz innerhalb der Heftromane größer der Unterschied zwischen den Heftromanen und der Hochliteratur (siehe Abbildung 14).

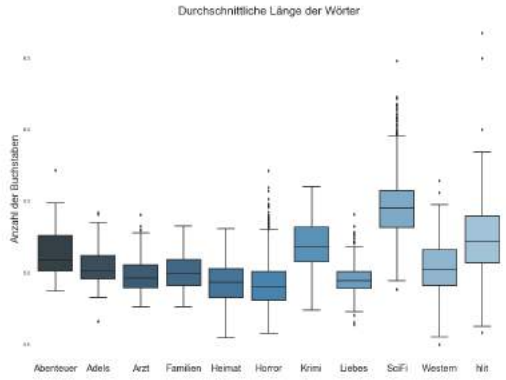


Abbildung 14. Durchschnittliche Wortlänge in Buchstaben

Die Messung der Satzlänge bestätigt die Untersuchungen aus den 1970er Jahren (Abbildung 15): Die Sätze sind im Durchschnitt über alle Gattungen hinweg kürzer, auch die Varianz der Satzlänge ist im Bereich der Hochliteratur deutlich größer. Allerdings widerspricht die Messung der Part-of-Speech Trigramme der Annahme, dass die Satzbaupläne ebenfalls schematischer sind (Abbildung 16).

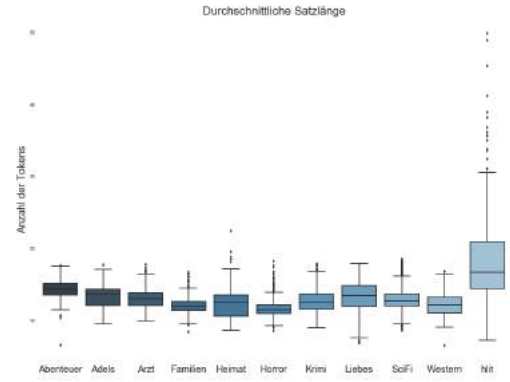


Abbildung 15. Durchschnittliche Satzlänge in Token

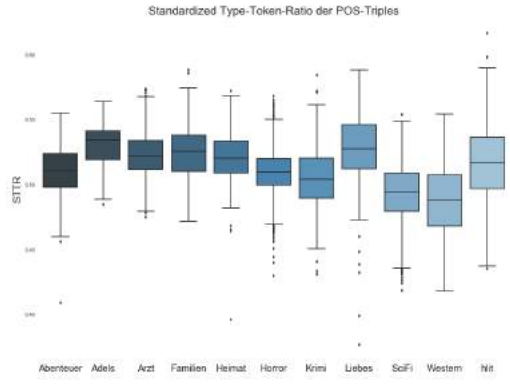


Abbildung 16. Schematisierung von Satzbauplänen anhand von POS-Triples

Diskussion

Die Gattungen der Heftromane sind - so der vorläufige Befund - deutlich umrissen und lassen sich durch die Zetawörter auch inhaltlich gut erschließen. Zwei Thesen zum Verhältnis von Hochliteratur zum Heftroman können dagegen als widerlegt gelten: Das Gebiet der Heftromane ist keineswegs besonders "homogen" (Domagalski 1980), vielmehr ist die Binnenvarianz sehr deutlich. Zum anderen ist die Sprache der Heftromane nicht eindeutig schlichter (Nusser 1982). Auch hier ist die Varianz innerhalb der Gattungen auffällig; insbesondere die Science-Fiction Romane weichen deutlich ab. Diese Arbeit markiert erst den Anfang unserer Untersuchungen. In den nächsten Schritten werden Figurenkonstellation, Analyse von Erzähler- und Figurenrede sowie Sentiment untersucht werden.

Fußnoten

1. Strategische Prioritäten 2017 – 2020, urn:nbn:de:101-2017021403

Bibliographie

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003): *"Latent dirichlet allocation"*. Journal of machine Learning research, 3 (Jan), 993-1022.

Burrows, J.: *»All the Way Through. Testing for Authorship in Different Frequency Strata«*, in: Literary and Linguistic Computing 22.1 (2007), S.27–47.

Craig, H., Kinney, A. (2009): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge.

Eder, M., Rybicki, J., and Kestemont, M. (2016): *Stylometry with R: a package for computational text analysis*. R Journal, 8(1): 107-121.

Foltin, H. F. (1965): *Die minderwertige Prosaliteratur. Einteilung und Bezeichnung*. In: DVjs 39 H. 2, 288–323.

Domagalski, P. (1980): *Trivialliteratur. Geschichte. Produktion, Rezeption.* Freiburg im Breisgau.

Hügel, H.-O. (2007): *Lob des Mainstreams. Zu Theorie und Geschichte von Unterhaltung und Populärer Kultur*. Köln: Halem.

Evert, S. Proisl, T., Reger, I., Pielström, S., Schöch, C. Vitt, T. (2017): *Understanding and explaining Delta measures for authorship attribution*. In: Digital Scholarship Humanities 32, 2,1, p. ii4-ii16.

Hettinger, L., Jannidis, F., Reger, I., Hotho, A. (2016): *Classification of Literary Subgenres*. Abstracts DHd-Tagung 2016, Leipzig 2016.

Kelleter, F. (2012): *Populäre Serialität. Eine Einführung*. In: ders.: (Hg.): Populäre Serialität: Narration – Evolution – Distinktion. Zum seriellen Erzählen seit dem 19. Jahrhundert. Bielefeld: transcript, S. 11-46.

Kubát, M. & Milička, J. (2013): *Vocabulary richness measure in genres*. Journal of Quantitative Linguistics, 20(4), 339–349.

McCallum, A.: *"MALLET: A Machine Learning for Language Toolkit."* http://mallet.cs.umass.edu . 2002.

McInnes, L. Healy, J. (2018): *"UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction"*, ArXiv e-prints 1802.03426

Nast, M. (2017): *"Perry Rhodan" lesen. Zur Serialität der Lektürepraktiken einer Heftromanserie*. Bielefeld: transcript.

Nusser, P.: *Romane für die Unterschicht. Stuttgart 1982.*

Nutz, W.; Schlögell, V. (1991): *Die Heftroman-Leserinnen und -Leser in Deutschland. Beiträge zur Erfassung popularkulturaler Phänomene*. In: Communications 16 (2). DOI: 10.1515/comm.1991.16.2.133.

Schöch, C. (2018): *Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie*. In **Bernhart, T., et al. (eds.)**: *Quantitative Ansätze in der Literatur- und Geisteswissenschaften*. Berlin: de Gruyter. 77-94.

Schöch, C. Schlör, D., Zehe, A., Gebhard, H, Becker, M., Hotho, A.: *Burrows' Zeta: Exploring and Evaluating Variants and Parameters*. Abstracts DH 2018.

Stiftung Lesen. Lesen in Deutschland 2008. https://www.stiftunglesen.de/download.php?type=documentpdf&id=11

Stockinger, C. (2018): *Das Groschenheft*. In: **Carlos Spoerhase und Steffen Martus (Hg.)**: *Gelesene Literatur.*

Populäre Lektüre im Zeichen des Medienwandels. München: text + kritik. (Im Druck)

Metadaten im Zeitalter von Google Dataset Search

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Data Center for the Humanities, Universität zu Köln

Rau, Felix
frau@uni-koeln.de
Institut für Linguistik, Universität zu Köln

Perspektiven auf Kuratieren, Suchen, Finden und Präsentieren aus der Praxis

Relevanz des Themas

Metadaten sind Teil einer jeden datenbezogenen Forschungspraxis und ein essentieller Bestandteil einer jeden Forschungsdatenmanagementstrategie. Auch wenn Metadatenproduktion und -kuration oft einen von den Forscher*innen ungeliebten Aspekt der Projektarbeit darstellt, ist die durch Metadaten sichergestellte Find- und Zitierbarkeit von Daten für die Forschungsgemeinschaft und insbesondere auch für die Geldgeber von höchstem Interesse. Einzelne Datenzentren und nationale und internationale Forschungsinfrastrukturen haben seit Jahren in die Verbesserung der Findmechanismen investiert. Mit dem Eintritt von großen kommerziellen Anbietern wie Google und Elsevier in die Suche wissenschaftlicher Daten(sätze) kündigen sich in den letzten Monaten massive Änderungen in der Landschaft an. Als einzelnes fach- oder datentypspezifisches Repositorium müssen wir in dieser Landschaft den Erwartungen und Anforderungen von Depositor, Geldgebern, und potentiellen Nachnutzern gerecht werden, während wir mit unseren Geld- und Personalressourcen verantwortungsvoll umgehen.

Problemstellung

Forschungsdatenzentren und Forschungsinfrastrukturen wenden einen bedeutenden Teil ihrer personellen und finanziellen Ressourcen für die Modellierung und Kuration von Daten und Metadaten auf. Bereitstellung und Erhalt von oft komplexen Suchfunktionalitäten und die Schaffung von Anschlussfähigkeit der Metadatenschemata mit externen wissenschaftlichen und fachspezifischen Suchportalen ist arbeitsaufwändig und wartungsintensiv.

Viele Forschungsdatenzentren haben in den letzten Jahren komplexe Metadatenschemata entwickelt, die die Vielschichtigkeit des Gegenstandes detailliert abbilden. Dabei gibt es durchaus eine Tendenz zur fachspezifischen Übermodellierung von Metadaten in hochkomplexen Schemata. Diese repositoriums- und gegenstandsspezifischen Schemata bilden sowohl den grundlegenden Inhalt der Webseiten des einzelnen Repositoriums als auch das Ausgangsmaterial um externe wissenschaftliche Datensuchportale zu bedienen.

Da diese komplexen Datensätze der einzelnen Portale in den gängigen Findmechanismen des Webs – insbesondere in der noch stets dominanten Google Web Search – aus verschiedenen Gründen schlecht erfasst werden, wurden in den letzten Jahrzehnt Metasuchportale entwickelt um eine repositoriumsübergreifende Findbarkeit sicherzustellen. Diese Metaportale sind normalerweise an eine Domäne gebunden. Die Domäne kann der Nationalstaat sein, wie beim niederländischen Portal NARCIS[1], sie kann an eine Forschungsinfrastruktur gebunden sein, wie im Falle des CLARIN VLO[2], sie kann fachspezifisch sein, wie im Falle des OLAC-Portals für Spracharchive[3], oder der Service kann an einen anderen Infrastrukturservice gekoppelt sein, wie die Koppelung der DataCite-Suche[4] an die DOI-Registrierung.

Es ist unumstritten, dass eine einfache Auffindbarkeit und Zugänglichkeit von Datensätzen ein zentraler Baustein der Nachnutzung von Forschungsdaten ist und deshalb die Möglichkeit gefunden zu werden höher zu bewerten ist, als die Genauigkeit der zu grunde liegenden Metadaten, und die verschiedenen Metasuchportale versprechen dies zu leisten. Die Existenz dieser Metaportale hat dazu geführt, dass einzelne Repositorien die Investition in maßgeschneiderte Webinterfaces verzichtet und sich Datenzentren auf Kuration und Bereitstellung der Metadaten über (Harvesting-)Schnittstellen konzentrieren. Dies ist auch eine Reaktion auf die Tatsache, dass die in wissenschaftlichen Projekten erstellten Portale oft nicht die Qualität und Nutzerfreundlichkeit erreichen, die Nutzer*innen aus kommerziellen Webangeboten gewöhnt sind, und anspruchsvolle Webseiten, bedingt durch die Kurzlebigkeit modernen Webtechnologien sehr wartungs- und kostenintensiv sind.

Um die Diversität und Komplexität der einzelnen Metadatenschemata handhabbar zu machen, reduzieren die Metaportale die Angaben auf kompakte und flache Strukturen, die oft aus nicht mehr als Listen von Key-Value-Paaren bestehen. Das DARIAH-DE Repository hat sich für den geradezu radikalen Schritt entschieden für das Anlegen einer Kollektion im Deposit-Interface "Publikator" gerade mal drei Pflichtfelder zu definieren (Mache & Klaffki 2018). Diese sind *Titel*, *Urheber*in* und *Rechteverwaltung* (Linzenzbestimmung).

Die komplexen Metadatenschemata in Kombination mit der Verarbeitung der Metadaten in den Metaportale hat aber auch den Effekt, dass Angaben, die nicht mit Blick auf die vereinfachte Repräsentation in den Metaportalen angelegt wurden, ohne Kontext nicht mehr verständlich sind. So kann ein Feld "Description", das sich auf ein Objekt wie Sprache oder Sprecher in einem komplexen Metadatenschemas bezieht, in einem Metaportal so dargestellt werden, als würde es sich auf den gesamten Datensatz beziehen.

Zustand

Das *Kölner Zentrum Analyse und Archivierung von AV - Daten* (KA^3) wird seit 2015 als Verbundvorhaben vom BMBF gefördert. Partner sind das *Max-Planck-Institut für Psycholinguistik* in Nimwegen (Niederlande), das *Fraunhofer-*

Institut IAIS in Sankt Augustin, das *Archiv "Deutsches Gedächtnis"* der FernUniversität Hagen sowie drei Akteure an der Universität zu Köln - das *Regionalen Rechenzentrum*, das *Institut für Linguistik* und das *Data Center for the Humanities*. Ein zentraler Bestandteil des Vorhabens ist die Etablierung eines Forschungsdatenrepositoriums am Kölner Standort. Hier werden vor allem Daten aus der linguistischen Sprachdokumentation kuratiert und archiviert und eine Ausweitung auf annotierte oder transkribierte audiovisuelle Daten aus methodisch verwandten Fachdisziplinen wie z.B. den ethnologischen Fächern oder der Oral History vorbereitet. Ein festgeschriebenes Projektziel ist die vollständige Integration des Repositoriums und Archivs *Language Archive Cologne* in die CLARIN-Forschungsinfrastruktur. Die Nähe zum Forschungsalltag der linguistischen Sprachdokumentation und die Anbindung an CLARIN bringen Bedingungen und Traditionen mit, die zu berücksichtigen sind.

Die Mehrheit der Metadaten in diesem Fachbereich liegen im IMDI-Format[5] vor. Das Format wird aktiv nur noch in wenigen Spracharchiven weltweit eingesetzt, hat aber nachhaltig die Konzeptualisierung von Metadaten in der Fachgemeinschaft geprägt. Der ausladende Charakter und das akademische Eingabetool, das primär genutzt wurde,[6] wurden in dem Wunsch geschaffen, die Bandbreite der Forschungstätigkeit zu inkludieren. Hier verschwimmen die Grenzen zwischen Daten und Metadaten.

Die CLARIN-Forschungsinfrastruktur macht klare Vorgaben bezüglich der Vorhaltung und Publikation von Metadaten durch die angeschlossenen Zentren. Metadaten müssen im CMDI-Metadatenformat verfügbar und über eine OAI-PMH Schnittstelle harvestbar sein. Die in der CMDI Component Metadata Registry definierten Profile können frei im Rahmen der existierenden Best Practices modelliert werden (CMDI 2018). Durch ein semantisches Mapping über die CLARIN Concept Registry soll eine semantische Interoperabilität von Metadatenkategorien gewährleistet werden.

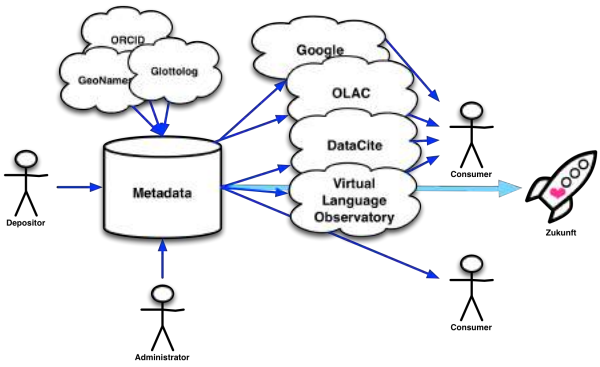
Metadaten in der CLARIN-Infrastrukur werden durch das Metasuchportal *Virtual Language Observatory* (VLO) geharvestet und dargestellt. Durch eine festgelegte Anzahl von Metadatenkategorien, die in jedem spezifischen Metadatenprofil durch Vergabe der entsprechenden Konzepten kenntlich gemacht werden müssen, werden die durchsuchbaren und facettierbaren Kategorien festgelegt. Mit der Zuordnung von bestehenden Metadatenfeldern zu definierten Kategorien (facet mapping) bestimmt der Metadatenprovider also letztendlich das Erscheinungsbild seiner Daten im VLO.

Zu den relevanten internationalen Infrastrukturen und Findmechanismen, die nicht in CLARIN organisiert sind, zählt vor allem der Katalog und die Metasuchmaschine der Open Language Archives Community (OLAC). OLAC hat ein eigenes sprachressourcen-orientiertes Metadatenformat und funktioniert ebenfalls nach dem Harvester-Prinzip über die Bereitstellung einer OAI-PMH-Schnittstelle durch den Metadatenprovider.

Ein wichtiger Gravitationspunkt für eine nicht-fachspezifische übergreifende internationale Adressierbarkeit und Discoverability von Forschungsdatensätzen sind die Dienste des DataCite Konsortiums in Verbindung mit der Registrierung von Digitial Object Identifier (DOIs) für jeden Datensatz. Mit jeder DOI wird ein einfacher, aber erweiterbarer, Metadatensatz registriert, der durch die DataCite Metadatensuche auffindbar wird. DOI und DataCite ist es gelungen, dass die Vergabe mit positiven und progressiven Effekten assoziiert wird, wie Anerkennung akademischer Leistungen und Wissenschaftlichkeit des Inhalts, sodass erfolgreich eine Nachfrage aus der Forschung generiert wird. Das Metadatenschema wird damit als Referenz für eine nicht-fachspezifische Beschreibung von Forschungsdatensätzen weiter an Bedeutung gewinnen. Metadatensätze werden hier aktiv durch den Metadatenprovider über eine Schnittstelle bei DataCite registriert.

Die übergreifenden Metadatenportale müssen sich mit der Nutzerfreundlichkeit und Performanz von kommerziellen Webdiensten messen. Hierbei liegt auf der Hand, dass die Qualität der Suche maßgeblich von der Qualität der Metadaten und vor allem die Verlässlichkeit und Passung der Metadatenkategorien abhängt, über die sie selbst nur bedingt Kontrolle haben.

Es ist also angebracht, dass sich Forschungsdatenarchive und Forscher bei der Konzeptualisierung von Metadaten und bei der Kuratierung nicht ausschließlich auf die Abbildung des zugrunde liegenden Forschungsgegenstands versteifen, sondern vielmehr Metadaten von ihren intendierten Disseminationskanälen her zu denken. In Zukunft wird es immer wichtiger sein, wie Daten in den jeweiligen Meta-Portalen aussehen werden.



Erwartungen und Veränderungen

Der neueste Ankömmling im Bereich der wissenschaftlichen Datensatz-Suche ist Google mit Google Dataset Search[7]. Dieser Service befindet sich noch im Beta-Stadium und ändert sich täglich. Die grundlegende Ausrichtung und die zugrundeliegenden Technologien sind aber bereits dokumentiert[8] und mehr oder weniger stabil. Google basiert die Datensatz-Suche auf Technologien und Strategien der Websuche auf. Nachdem Google seinen Support für OAI-PMH-Schnittstellen 2008 (Mueller 2008) eingestellt hatte, ist es nicht unerwartet, dass sie, anstatt Metadaten, über spezielle Schnittstellen abzufragen oder eine Registrierung zu erfordern, durch das Crawlen der Webseiten von Repositorien und Portalen erfasst werden. Dabei werden strukturierte Daten, die mit Hilfe der Linked-Data-Technologien JSON-LD[9] oder RDFa[10] und den Ontologien *schema.org*[11] oder *W3C Data Catalog Vocabulary*[12] ausgezeichnet wurden, als Grundlage der Darstellung und wahrscheinlich auch des Suchindexes

genommen. Durch diese Technologieentscheidungen wird eine flache Metadatenstruktur, bestehend aus einfachen Key-Value-Paaren und eine allgemeine, fachagnostische Datenfeldsemantik, bedingt. Der Fokus auf HTML-Webseiten und allgemeine Webtechnologien für strukturierte Daten als Hauptschnittstelle für Metadaten bedeutet eine grundlegende Abkehr von den gängigen Praktiken in der wissenschaftlichen Technologielandschaft.

Ausblick und mögliche Antworten

Die aktuellen Veränderungen in der Datensatzsuche bedeuten auf der einen Seite, dass Findbarkeit immer weiter aus der Kontrolle der einzelnen Repositorien in die Hand von Metaportalen und Drittanbietern wie Google und Elsevier[13] geht. Auf der anderen Seite bedeutet der durch Google angedeutete Technologiewechsel eine Stärkung der Webinterfaces der einzelnen Repositorien. Die Webseite ist damit wieder primäre Schnittstelle. Repositorien werden die Semantik ihrer Metadaten stärker von den intendierten Disseminationskanälen her denken müssen und einzeln und in Verbünden Strategien für die nachhaltige Findbarkeit entwickeln.

Fußnoten

1. https://www.narcis.nl/ , 11.01.2019.
2. https://vlo.clarin.eu/ , 11.01.2019.
3. http://search.language-archives.org , 11.01.2019.
4. https://search.datacite.org/ , 11.01.2019.
5. https://www.mpi.nl/ISLE/ , 11.01.2019.
6. https://tla.mpi.nl/tools/tla-tools/arbil/ , 11.01.2019.
7. https://toolbox.google.com/datasetsearch , 11.01.2019.
8. https://developers.google.com/search/docs/data-types/dataset , 11.01.2019.
9. https://www.w3.org/TR/json-ld/ , 11.01.2019.
10. https://www.w3.org/TR/rdfa-primer/ , 11.01.2019.
11. https://schema.org/ , 11.01.2019.
12. https://www.w3.org/TR/vocab-dcat/ , 11.01.2019.
13. https://datasearch.elsevier.com/ , 11.01.2019.

Bibliographie

Blumtritt, J. / Rau, F. (2016): *User-Experience von Spracharchiven: Eine Neubewertung der Interaktion von Archiv und Nutzern*, in: Digital Humanities im deutschsprachigen Raum (DHd 2016), Leipzig. http://dhd2016.de/boa.pdf#page=215 .

CMDI and Metadata Curation task forces of the Standing Committee on CLARIN Technical Centres (2018): *CMDI Best Practices*, Version 1.1.1, https://github.com/clarin-eric/cmdi-best-practices/releases .

Mache, B. / Klaffki, L. (2018): *Das DARIAH-DE Repository. Elementarer Teil einer modularen Infrastruktur für geistes- und kulturwissenschaftliche Forschungsdaten*, in: O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 5(3) 2018, 92-103. https://doi.org/10.5282/o-bib/2018H3S92-103 .

Mueller, John (2008): *Retiring support for OAI-PMH in Sitemaps*. https://webmasters.googleblog.com/2008/04/retiring-support-for-oai-pmh-in.html .

Schaffner, J. (2009): *The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies*, in: Report produced by OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2009/2009-06.pdf .

Methoden auf der Testbank: Ein interdisziplinäres, multimodales Lehrkonzept zur Beantwortung einer fachhistorischen Fragestellung

Moeller, Katrin
katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Müller, Andreas
anderas.mueller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Purschwitz, Anne
anne.purschwitz@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Digital Humanities in die Lehre und Ausbildung stärker zu integrieren, ist eine vielfach geäußerte Forderung im Rahmen von DH und geisteswissenschaftlichen Fachverbänden (Sahle 2017). Während in den ersten beiden Jahrzehnten der Digitalisierung vor allem der Wandel von der analogen zur digitalen Erschließung und Präsentation, Open Access, Blended Learning und digitales Publizieren im Mittelpunkt der Forschung stand, haben sich Ansätze und Themen zur universitären Vermittlung von DH-Technologien in der jüngeren Vergangenheit stark verändert. Mit den Digital Humanities ist eine Community entstanden, die innovative neue Methoden, vor allem aber eine Vielzahl von Tools und Werkzeugen bereitstellt. Sie entspringen nicht einem einzelnen fachspezifischen Kontext, sondern setzen auf interdisziplinäre Konzepte und vor allem vertiefte informatische Kenntnisse. Mittlerweile lässt sich in der Fachlandschaft eine Etablierung neuer Formen von Studiengängen und Curricula beobachten, die solche spezifischen DH-Anwendungen vermitteln und so zur Ausbildung der dringend benötigten DH-Spezialisten beitragen. Grundsätzlich deckt diese Form der parallelen Ausbildung von DH-Spezialisten zu den eigentlichen Fachwissenschaften daher einen sehr wichtigen Bedarf ab (Sahle 2013).

Dennoch wirft diese Entwicklung auch Schwierigkeiten auf, da sie die Entkoppelung von an der DH orientierten Wissenschaftlern und eigentlicher geisteswissenschaftlichen Fachwissenschaft zusätzlich verschärft (Hohls 2017). Da

viele DH-Anwendungen heute noch keinem Fachkanon oder Standards unterliegen, bleibt die Einarbeitung in solche Methoden und Tools ein arbeitsintensiver Prozess interdisziplinärer Verständigung, der meist individuell geleistet werden muss. Dies führt häufig zu Schwierigkeiten im Vermittlungsprozess. In der jüngeren Diskussion wird dies gern mit dem Bild des DH-affinen „Hackers" charakterisiert, der in der Fachwissenschaft dem interessierten „Laien" gegenübertritt. Die Spannungen zwischen beiden Gruppen (Enthusiasten und Skeptiker) haben sich in den letzten Jahren verschärft. Auf dem letzten Historikertag in Münster (September 2018) war die Skepsis gegen neue Methoden des digitalen Arbeitens in mehreren Sessions genauso fassbar, wie letztlich die Forderung danach, dass die DH auch fachlich für die einzelnen Fachverbände konkrete Antworten liefern müsse und nicht nur eine grundlegende Erschließungsfunktion besitzen dürfe. Innerhalb der einzelnen Fachverbände muss geklärt werden, inwieweit digitale Geisteswissenschaft Bestandteil der klassisch fachbezogenen Ausbildung werden kann und soll (Fickers 2014, S. 27, Schulz 2018, S. 79f.). Das gilt nicht nur für die spezifischen Forschungsrichtungen der Hilfs- und Grundwissenschaften und der Quellenkritik, sondern letztlich für alle Teilbereiche der fachbezogenen Forschung (Schlotheuber/Bösch 2015). Während in den Digital Humanities eher die hilfswissenschaftlichen Traditionen der Erschließung, Annotation, Editorik, Messung und Visualisierung wichtige Dimensionen repräsentieren, bleibt für die Geschichtswissenschaft die digitale Quellenkritik, Datenmodellierung, Analyse und vor allem Methoden- und Algorithmenkritik wesentliche Aufgabe (Rehbein 2015, Schulz 2018). Mittlerweile stehen ausgereifte kommerzielle und OpenSource-Programme bereit, um mittels qualitativer und/oder quantitativer Datenanalyse ganz verschiedene methodische Verfahren anzuwenden. Allerdings ist der zeitliche Umfang des Geschichtsstudiums durch die Einführung der Bachelor- und Masterstudiengänge heute streng limitiert. Viele Studierende verlassen nach dem Bachelorstudium die Universitäten und nehmen eine Arbeit auf. Innerfachlich ist daher der Anteil von historischer Fachkompetenz und digitalem methodischen Wissen bei der Ausbildung des Nachwuchses zu gewichten und es müssen zwingend Vorschläge diskutiert werden, welche grundsätzlichen Bausteine Anschlussfähigkeit zu den Digital Humanties herstellen können und wie diese in den Studienkanon integriert werden können.

Welche Grundkompetenzen der Geschichtswissenschaft sind für die Basis der Geisteswissenschaft also ausschlaggebend? Für das Fach selbst dürfte die Verbindung von fachlicher Fragestellung, ihre Übertragung in spezifische Methoden und die Modellierung der damit in Zusammenhang stehenden Daten solche Grundkompetenzen beschreiben.

Im Rahmen des Workshops "Digitale Lehrmethoden und digitale Methoden in der Geschichtswissenschaft. Neue Ansätze für die Lehre" hat die AG Digitale Geschichtswissenschaft im Verband der Historiker und Historikerinnen Deutschlands im Jahr 2018 eine Workshopreihe begonnen, um genau solche Spannungsfelder auszuloten und Angebote für die digitale Vermittlung zwischen DH und Fachcommunities zu leisten (König 2018). Zwar gibt es Summerschools und einzelne Workshops für die Vermittlung von digitalen Methoden, diese bieten aber häufig nur schwerpunktartige Einführungen zu bestimmten Tools und Techniken. Einen breiten Überblick über die vielfältige Landschaft digitaler Methoden, ihre Einbindung in Forschung und Lehre, bzw. ihre Anwendungsmöglichkeiten und -grenzen können diese Formen der Weiterbildung hingegen nicht leisten. Systematisches Wissen steht für die Lehre auf diese Weise bisher nur selten zur Verfügung, wie auch Anreiz- und Gelegenheitsstrukturen für eine tiefergehende digitale Schulung fehlen. Das Digitale wird so schnell zur Last anstatt zur Lust, vor allem wenn hinter den Angeboten auch ausgefeilte didaktische Konzepte und Ideen stecken sollen. Hier braucht es kreative Ideen, um im wissenschaftlichen Alltag Dozierende und Studierende zu erreichen und auf diese Weise, die bereits vorhandenen digitalen Infrastrukturen mit Leben zu füllen.

Ein solches digitales Angebot möchten wir mit unserem Vortrag vorstellen. Im Rahmen des Workshops „Methoden auf der Testbank. Drei Zugänge zur Hexenforschung im Vergleich" haben wir anhand von zentralen Thesen zur Hexenforschung einen Korpus von Texten und Daten zusammengestellt, der sich für drei verschiedene methodische Analysen entlang der gleichen historischen Fragestellung nach der Entwicklung, Abgrenzung und Ausdifferenzierung des Zauberei- und des Hexensabbatkonzeptes im 16. Jahrhundert nutzen lässt. Der Workshop richtet sich an Dozierende, die digitalen Methoden eher abwartend gegenüberstehen. Dabei geht es gezielt um die Verbindung fachlicher, methodischer und digitaler Problemstellungen. Aufgegriffen wird daher eine wichtige These der Hexenforschung (Voltmer 2012, Behringer 1997, Dillinger 2007), die bereits eine lange fachliche Diskussion besitzt und für die durch die Lehrkonzeption auf digitalem Weg Einsichten und Erkenntnisse formuliert werden können. Entwickelt wurde ein Materialkorpus der folgende Materialien bereithält und bis zur Tagung auch frei zur Verfügung gestellt werden soll:

1.) Lehrkonzept und didaktische Stoffentwicklung zur Thematik der Hexenverfolgung sowie die fachliche Entwicklung und Begründung der Fragestellung. Dieser Baustein des Angebots formuliert nicht nur die Fragestellung, sondern bietet zudem eine Einbindung zentraler fachwissenschaftlicher Texte und Thesen zum Thema, um hier verschiedene Aspekte auch für eine projektorientierte Seminargestaltung zu ermöglichen. So werden etwa Vorschläge unterbreitet, welche Formen der Analyse Studierende anhand eines selbstgewählten Projektes mithilfe der methodischen Ansätze wählen können.

2.) Textkorpora: Digitalisiert und transkribiert wurden Urgichten und Verhörprotokolle von insgesamt 52 Personen (ca. 500 Digitalisate), die in der Stadt Rostock (Mecklenburg) während des 16. Jahrhunderts aufgrund von Zauberei oder Hexerei angeklagt wurden. Der Textkorpus erlaubt aufgrund des damit eingefangenen - auch sprachwissenschaftlich oder rechtsgeschichtlich sehr interessanten - Beobachtungszeitraums weitergehende Fragestellungen und Analysen. Aufgrund der hervorragenden Quellenüberlieferung bot der Bestand bereits mehrfach Ansatzpunkte zur auszugsweisen Edition (Koppmann 1900, Krause 1915) wie auch grundlegender Erforschung (Ehlers 1986, Moeller 2007, Müller 2017). Langfristig lassen sich hier sogar Untersuchungen zur Editionspraxis verschiedener Zeiten anstellen. Zugleich werden motivgeschichtliche Analysen möglich, da der Bestand etwa in die volkskundlichen Sammlungen und Korpora des 19./20. Jahrhunderts eingegangen ist.

Diese Texte sind in einem zweimaligen Korrekturprozess und entsprechend der DTA- Transkriptionsrichtlinien

transkribiert worden. Bis zur Tagung sollen sie nach Möglichkeit in das Deutsche Textarchiv oder ein anderes Repositorium überführt werden.

3.) Über die Transkription hinaus erfolgte eine Modellierung der Fragestellung mithilfe eines selbst zu entwickelnden Kategoriensystems, das sich auch zur Implementation (Annotationen) in die Textdateien eignet (hier können praktische Übungen zusätzlich angeboten werden). Zum anderen aber auch in Form von Datenstrukturen zur Analyse mit MAXQDA, einem Statistikprogramm (hier spezifisch SPSS, möglich ist aber auch die Nutzung von OpenSource Software wie R) sowie einem graph- bzw. netzwerkbasierten Datenmodell (hier Gephi) Auswertungen erlaubt. Wir haben uns auf die Anwendungen von Programmen konzentriert, die Lehrenden und Lernenden einen ersten, niedrigschwelligen Zugang zur Anwendung von Methoden im Vergleich ermöglichen und vor allem für das Selbststudium auch genügend ausreichend dokumentiert sind. Über einen didaktischen Aufbau von qualitativer Datenanalyse (orientiert an Kuckartz 2014 und Mayring 2015), statistischer Auswertung und netzwerkorientierter Untersuchung können Studierende hier in der Datenmodellierung vom Konkreten zum Allgemeinen, der Entwicklung von Kategoriensystemen und der Verwendung von Standards geschult werden. Denn grundlegende Probleme in der Lehre sind meist das Verständnis von Datenstrukturen, die Formen der Modellierung, die Entscheidung für eine Methode und die Interpretation von Erkenntnissen aus den erzielten Datenanalysen.

4.) Die Datenmodellierung wird genau dokumentiert und dient Studierenden und Lehrenden damit ebenso zur Institutionalisierung von zentralen Schritten des Forschungsdatenmanagements. Anhand von eher sozialwissenschaftlichen Dokumentationspraktiken werden Richtlinien formuliert, die nicht nur das Verständnis der Daten in ihrer Überführung von der „Quelle zur Tabelle“ (Manfred Hettling) fördern, sondern ebenso ein Beispiel für die „gute wissenschaftliche Praxis“ der Dokumentation von Forschungsleistungen bieten. Die Daten werden in langfristig speicherbaren, spezifischen Datenrepositorien abgelegt. Auf einer gemeinsamen Plattform der AG Digitale Geschichte und des Historischen Datenzentrums Sachsen-Anhalts werden die einzelnen Komponenten zu einer didaktischen Einheit nachnutzbar versammelt.

5.) Das Datenset bzw. die Lehrkonzeption ermöglicht es den Studierenden, einen Überblick über die Verwendung von drei verschiedenen methodischen Ansätze zu erlangen, die Methoden kritisch zu vergleichen und Fähigkeiten zur Operationalisierung und Implementierung von Datenmodellierungen zur Beantwortung von Fragestellungen zu erwerben. Überdies wird eine individuelle Aktivierung der Studierenden möglich, die für die didaktische Vermittlung von Fachinhalten heute eine zentrale Rolle in der Lehre spielt (Helmke 2015). Mit Hilfe von selbsterstellten und bereitgestellten Daten können geschichtswissenschaftliche Analysen durchgeführt und die Vor- und Nachteile einzelner Methoden diskutiert werden. Vor allem aber erweist sich, ob Methodenvielfalt eher zur Bestätigung von Thesen oder zu neuen Perspektiven führt. Das Werkzeug bietet damit nicht nur Möglichkeiten der Quellen-, sondern eben auch der Methodenkritik.

Insgesamt möchten wir damit einen Baustein vorstellen, wie sich digitale Lehre und Methoden der DH unmittelbar mit fachlichen Fragestellungen und Gegenständen der Geschichtswissenschaft verzahnen und bearbeiten lassen und sich der immer wieder beklagte Graben zwischen Digital Humanties und Fachwissenschaften einebnen lässt. Gleichzeitig möchten wir Einschätzungen zum Zeitaufwand, Lehraufwand und zur Ressourcenplanung geben. Die Lehreinheit soll eher unerfahrenen Nutzerinnen und Nutzern der Methoden einen Eindruck vermitteln, was digitale Werkzeuge leisten können und wie sie sich in der alltäglichen, fachspezifischen Lehre einbetten lassen, obwohl sie natürlich klassische Werkzeuge einer allgemeinen digitalen Forschung repräsentieren.

Bibliographie

Wolfgang Behringer: *Hexenverfolgung in Bayern. Volksmagie, Glaubenseifer und Staatsräson in der Frühen Neuzeit*, München 1997.

Johannes Dillinger: *Hexen und Magie. Eine historische Einführung*, Frankfurt/Main 2007.

Ingrid Ehlers: *Über den Glauben an Hexen und Zauberer und ihre Verfolgung im Rostock des 16. Jahrhunderts*. Beiträge zur Geschichte der Stadt Rostock (Neue Folge) 6, 1986, S. 21-40.

Andreas Fickers: *Der ultimative Klick? Digital Humanities, Online-Archive und die Arbeit des Historikers im digitalen Zeitalter*, in: Forum für Politik, Gesellschaft und Kultur in Luxemburg 337, 2014, S. 25-29, http://hdl.handle.net/10993/21285.

Andreas Helmke: *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*, Seelze-Velber 2015.

Rüdiger Hohls: *Digital Humanities vs. Digital History: Differenzen und Gemeinsamkeiten*, in: Videoaufzeichnungen der Ringvorlesung "Digital Humanities: Die digitale Transformation der Geisteswissenschaften, Berlin 2017, URL: http://www.bbaw.de/mediathek/archiv-2017/24-10-2017-digital-humanities.

Mareike König: *Workshopreihe 2018: Digitale Lehrmethoden und digitale Methoden in der Geschichtswissenschaft. Neue Ansätze für die Lehre #digigw18*, 2018, https://digigw.hypotheses.org/1660.

Karl Koppmann: *Aus Hexenprozessen (aus Rostocker Niedergerichtsakten von 1576-1621)*, in: Korrespondenzblatt des Vereins für niederdeutsche Sprachforschung 21, 1900, S. 20-29.

Ludwig Krause: *Die Blocksbergfeste der Hexen und Zauberer nach den Rostocker Kriminalakten des 16. Jahrhunderts*, in: Niedersachsen 15, 1915, S. 238-240.

Udo Kuckartz: *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*, Weinheim und Basel 2014.

Philipp, Mayring: *Qualitative Inhaltsanalyse. Grundlagen und Techniken*, Weinheim und Basel 2015.

Katrin Moeller: *Dass Willkür über Recht ginge. Hexenverfolgung in Mecklenburg im 16. und 17. Jahrhundert*, Bielefeld 2007.

Andreas Müller: *Die Magie der Inhaltsanalyse. Entwurf einer Inhaltsanalyse für den Vergleich von Hexenprozessakten aus Rostock 1584 und Hainburg*, Masterarbeit Universität Wien 2017, https://www.historicum.net/fileadmin/sxw/Themen/Hexenforschung/Themen_Texte/Magister/hexen_mag_mueller.pdf.

Malte Rehbein: *Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht*,

in: H-Soz-Kult, 27.11.2015, www.hsozkult.de/debate/id/diskussionen-2905 .

Patrick Sahle: *DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities*, Göttingen: GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2013 (DARIAH-DE working papers 1).

Patrick Sahle: *Forschung & Karriere. Zur anhaltenden Formierung, Professionalisierung und Professoralisierung der Digital Humanities*, Vortrag Berlin, in: Ringvorlesung des Interdisziplinären Forschungsverbundes Digital Humanities in Berlin, 12.12.2017.

Eva Schlotheuber / Frank Bösch: *Quellenkritik im digitalen Zeitalter. Die Historischen Grundwissenschaften als zentrale Kompetenz der Geschichtswissenschaft und benachbarter Fächer*, auf: VHD-Blog (4 S.; PDF-Version: http://www.historikerverband.de/fileadmin/user_upload/vhd_journal_2015-04_beileger.pdf), 30. Oktober 2015.

Julian Schulz: *Auf dem Weg zu einem DH-Curriculum. Digital Humanities in den Geschichts- und Kunstwissenschaften an der LMU München*, München 2018, https://doi.org/10.5282/ubm/epub.42419 .

Rita Voltmer / Walter Rummel: *Hexen und Hexenverfolgung in der frühen Neuzeit*, Darmstadt 2012.

Multimodale Stilometrie: Herausforderungen und Potenzial kombinatorischer Bild- und Textanalysen am Beispiel Comics

Dunst, Alexander
alexander.dunst@gmail.com
Universität Paderborn, Deutschland

Hartel, Rita
rst@upb.de
Universität Paderborn, Deutschland

Einleitung

Stilometrische Untersuchungen blicken auf eine lange Tradition in der Literaturwissenschaft zurück (Holmes). Im Gegensatz dazu befinden sich quantitative Untersuchungen des Stils visueller Kunst und multimodaler Medien in einem frühen Experimentierstadium, das von explorativen Untersuchungen, Methodenentwicklung und -Adaption geprägt ist. Auch hier sind Fortschritte erkennbar, etwa in der digitalen Kunstgeschichte, der Filmwissenschaft und in der Comicforschung (Manovich, Douglas & Zepel; Qui, Taeb & Hughes; Baxter, Khitrova & Tsivian; Cutting et al.; Dunst & Hartel 2018a). Dabei konzentriert sich die stilistische Klassifikation bisher entweder auf visuelle oder sprachliche Kanäle. Beispielhaft zu nennen sind die Analyse der historischen Entwicklung von Filmfarben bei Barbara Flückiger oder Ben Schmidts thematische Untersuchungen populärer TV-Serien (Flückiger; Schmidt). Während dieser monomodale Fokus bei visueller Kunst der oftmaligen Dominanz der Bildebene geschuldet ist, so sind dafür bei multimodalen Medien wie Film, Fernsehen, Computerspielen oder Comics andere Gründe ausschlaggebend. Mit dem Topic Modeling oder der Textstilometrie stehen Methoden zur Verfügung, die auf sprachlichen Daten basieren und in der digitalen Literaturwissenschaft laufend weiterentwickelt werden. Visuelle Stilometrie, obwohl weit weniger ausgereift, kann auf die erwähnten Arbeiten aus der Kunstgeschichte und empirischen Filmwissenschaft zurückgreifen.

Auch technische Hürden tragen zu monomodalen Analysen bei: je nach Medium ist die digitale Erschließung und Analyse von Informationskanälen mit erheblichen Schwierigkeiten verbunden, etwa die Spracherkennung von Filmdialogen, die automatische Erkennung von handschriftlichen Texten in Comics oder die computergestützte Verarbeitung großer Mengen an Bildmaterial. Die Kombination unterschiedlicher Informationskanäle in visuellen Medien führt jedoch unweigerlich zur Frage, inwieweit Stil durch die Analyse einzelner Modi erfasst werden kann. Auch thematische Untersuchungen setzen sich dem Vorwurf aus, komplexe Medien auf zu schmaler Datenbasis zu interpretieren, wenn Filmanalysen alleine auf Untertiteln oder Drehbüchern basieren. Im Gegenzug verbindet sich mit der Einbeziehung mehrerer Informationskanäle in stilometretische Untersuchungen die Hoffnung, multimodale Medien vollständiger beschreiben und einzelne Autoren, Genres oder Epochen genauer voneinander unterscheiden zu können. Im Folgenden werden erste Untersuchungen vorgestellt, die auf Basis eines Corpus an englischsprachigen Comicbüchern – so genannten „Graphic Novels“ – visuelle und Textstilometrie kombinieren (Dunst, Hartel & Laubrock).

Vorarbeiten und Herausforderungen

Wie bereits dokumentiert, haben wir auf Basis kunsthistorischer und filmwissenschaftlicher Vorarbeiten eine visuelle Stilometrie für Comicbücher entwickelt, die zwischen einzelnen Genres, Autoren und Publikationsformen unterscheiden und diese auf repräsentativer Datenbasis stilistisch beschreiben kann (Dunst & Hartel 2018b). Die relativ geringe Datenbasis, die mit der Analyse eines kulturellen Nischenproduktes einhergeht, bedeutete jedoch, dass nicht in allen Fällen signifikante Ergebnisse erzielt werden konnten. Insbesondere die Entwicklung einzelner Gattungen innerhalb eines Mediums lässt sich im historischen Verlauf nicht signifikant belegen. Auch nationale Traditionen konnten nicht immer stilistisch voneinander unterschieden werden, selbst wo sich diese für den Betrachter deutlich voneinander unterscheiden. Abbildung 1 zeigt, dass die von uns verwendeten visuellen Maße den japanischen Manga-Autor Osamu Tezuka stilistisch nicht von anglo-amerikanischen Werken abgrenzen konnten (Dunst & Hartel 2018a).

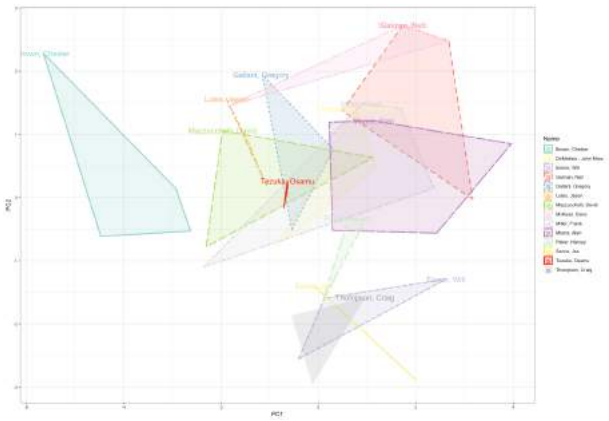


Abbildung 1. Visueller Stil bei Comics-Autoren

In beiden Fällen – also sowohl bei Autor- als auch bei Genreunterscheidungen – liegt es nahe, dass die kombinatorische Analyse von visueller und Textstilistik die Wahrscheinlichkeit erfolgreicher Unterscheidungen erhöhen würde. Allerdings stellen diese zusätzlichen Dimensionen eine multimodale Stilometrie vor methodische Herausforderungen. Wird eine große Zahl an Maßen für die Klassifikation herangezogen, so erschwert dies die qualitative Interpretation der Ergebnisse. Zwar lässt sich mit Hilfe einer Principal Component Analysis (PCA) darstellen, ob sich Genres oder Autoren signifikant unterscheiden. Je mehr Dimensionen in die PCA einfließen, desto schwerer fällt allerdings die Aussage, auf welchen Maßen diese Unterschiede fußen. Hinzu kommen wie erwähnt technische Hürden: abgesehen von den Angeboten bekannter Softwaregiganten führen automatische Spracherkennungssysteme noch zu relativ schlechten Resultaten. Ähnlich liegt der Fall bei Comics, deren Texte auf Handschriften basieren. Erst seit kurzem können diese Texte mit Hilfe automatischer Erkennungssysteme, die auf „Deep Learning“ basieren, zugänglich gemacht werden. Dennoch liegen Fehlerraten weit über jenen, die die Basis der meisten literaturwissenschaftlichen Analysen bilden. Der nächste Abschnitt stellt eine Methode vor, die dennoch die Verwendung einfacher Textmaße für die multimodale Stilometrie ermöglicht.

Datenbasis & Methode

Die Analysen basieren auf dem ersten repräsentativen Corpus englischsprachiger Comicbücher, von uns „Graphic Narrative Corpus“ (GNC) genannt (Dunst, Hartel & Laubrock). Wie in früheren Arbeiten beschrieben (Hartel & Dunst), nutzen wir die Bag Error Rate (BER) für eine Abschätzung, ob die Qualität der erkannten Texte ausreichend gut ist, um diese für die Textanalysen heranzuziehen. Hierzu haben wir als Gold Standard rund 10% der Seiten einer Graphic Novel manuell annotiert. Wir betrachten in unseren Analysen die Multi-Menge (oft als „Bag“ bezeichnet) aller Wörter, also die ungeordnete Menge aller Wörter, wobei Wörter – im Gegensatz zur herkömmlichen Menge – in dieser Multimenge auch mehrfach vorkommen. Wir berechnen also für jedes in einem der beiden Texte enthaltenen Wörter die Differenzen der Häufigkeiten $freq^T(w)$ für die Texte T=GS (Gold-Standard) und T=ET (erkannter Text), summieren diese auf, und normalisieren diese, in dem wir die Summe durch Gesamtanzahl aller Wörter im erkannten Text teilen:

$$BER := (\Sigma_{w \in W} |freq^{GS}(w) - freq^{ET}(w)|) / (\Sigma_{w \in W} freq^{ET}(w))$$

Frühere Analysen haben gezeigt, dass wir den Text als geeignet für die Analyse erachten können, wenn für eine Graphic Novel die BER kleiner als 40 ist. Auf den erkannten Texten betrachten wir die Textähnlichkeit basierend auf einer euklidischen Vektordistanz der Dokument-Vektoren der Term-Dokument-Matrix, die jeweils die für jedes Dokument D die relative Vorkommenshäufigkeit tf(D,t) der 2000 häufigsten Wörter t enthält. Bzgl. der visuellen Maße betrachten wir die mittlere Helligkeit jeder Seite, die Entropie und die Anzahl der Flächen als Maß für die visuelle Unruhe eines Bildes, sowie den Color Layout Descriptor und den Edge Histogram Descriptor des Standards MPEG7 (Martínez, Koenen & Pereira). Diese haben sich in früheren Arbeiten als vielversprechende Maße herausgestellt (Dunst & Hartel 2018a). Wann immer eine Dimensionsreduktion notwendig ist, führen wir diese mithilfe einer PCA durch. Um z.B. die textuellen und visuellen Maße kombiniert zu betrachten, haben wir – um einem Ungleichgewicht der 2000 Dimensionen für die 2000 häufigsten Wörter im Vergleich zu den ca. 40 visuellen Maßen entgegenzuwirken - zunächst via PCA die textuellen Dimensionen auf 40 reduziert. Anschließend haben wir die Dimensionen der textuellen PCA und die Dimensionen der visuellen Maße mit Hilfe einer weiteren PCA kombiniert. Zur Untersuchung signifikanter Zusammenhänge nutzen wir die ANOVA (ANalysis Of VAriance), die untersucht, ob die Varianz zwischen den Kategorien größer ist als die Varianz innerhalb der Kategorien.

Ergebnisse & Diskussion

Abbildung 2 stellt die Ergebnisse der multimodalen Stilometrie im Vergleich mit rein visuellen oder Textmaßen dar. Dabei zeigt sich, dass die Kombination von Bild- und Textanalyse nicht immer zum Erfolg führt. Zwar ergeben sich aus der Analyse beider Informationskanäle statistisch deutlichere Signifikanzen bei der Autor-Identifikation und Genreunterscheidung. Der Effekt ist allerdings gering. Im Fall der Klassifikation nach Originalsprache des Werks sowie unterschiedlicher Publikationsformen – etwa als Einzelband oder als fortgesetzte Serie– führt die Hinzunahme der Textanalyse derzeit nicht zu signifikanten Ergebnissen. Bei der Analyse unterschiedlicher Formen von Autorschaft (Einzelautor*innen, Zusammenarbeit von einer Autor*in und einer Illustrator*in und größeren Autor*innen-Teams) liegt das Resultat der Textanalyse weit über der Signifikanzgrenze. Die deutlich signifikanten Ergebnisse der visuellen Stilometrie setzen sich allerdings auch in der Kombination beider Kanäle durch. Insgesamt lässt sich sagen, dass trotz der gleichen Anzahl der verwendeten visuellen und Textmaße erstere derzeit aussagekräftiger erscheinen. Weiter erscheint es sinnvoll, die Analyse beider Informationskanäle immer auch einzeln zu betrachten und diese nicht immer sofort zu kombinieren.

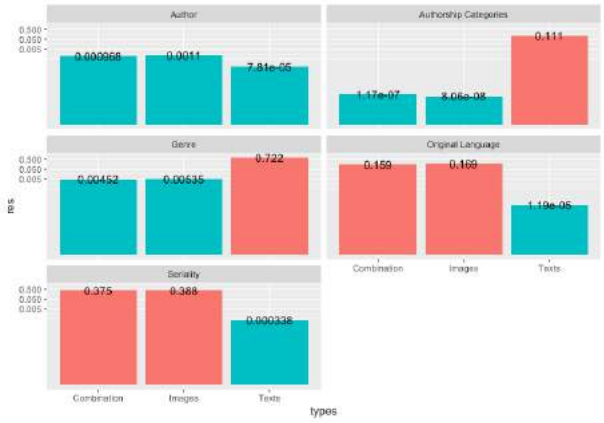


Abbildung 2. Zusammenfassung der stilometrischen Untersuchungen; Signifikanz ab $p < 0,05$

Wie bereits kurz erwähnt, zeigt sich für eine Klassifikation der entscheidende Einfluss der verwendeten Textmaße. Abbildung 3 stellt alle von uns untersuchten Werke in Streudiagrammen dar und unterscheiden diese farblich nach Originalsprache. Im Fall japanischer Manga handelt es sich hier außerdem um eine eigenständige Nationaltradition. Obwohl uns japanische und französischsprachige Werke in englischer Übersetzung vorliegen, führt nur die Textanalyse zu einer klaren Unterscheidung. Dies steht im klaren Gegensatz zu den Ergebnissen in Abbildung 1. Zwei potenzielle Ursachen können für dieses, auf den ersten Blick kontraintuitive, Ergebnis angeführt werden. Erstens erscheint es möglich, dass diese Unterscheidung eine Folge des Übersetzungsprozesses sind. Wahrscheinlicher ist, dass sich typische Merkmale der Texte von Manga auch in Übersetzung erhalten – in diesem Fall die Frequenz einzelner Wörter.

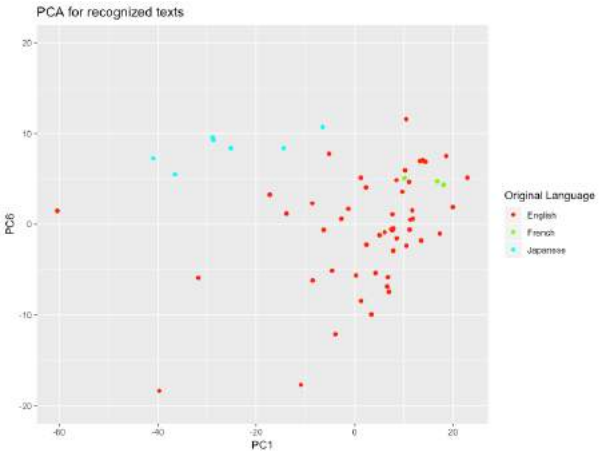


Abbildung 3. Streudiagramm basierend auf Textmaßen

Zusammenfassung & Ausblick

Wir haben erste Untersuchungen vorgestellt, die am Beispiel von Comicbüchern visuelle und Textmaße für eine multimodale Stilometrie kombinieren. Dabei handelt es sich, insbesondere in letzterem Fall, um sehr einfache Maße, die dem derzeitigen Stand der automatischen Texterkennung für Comicschriften geschuldet sind, und insgesamt um erste Pilotversuche. Wie sich zeigte, führt die Kombination der Text- und Bildebene in der Analyse bisher nicht immer zu besseren Ergebnissen. Allerdings ist dies trotz der geringen Anzahl der untersuchten Werke sowohl bei der Gattungsunterscheidung als auch bei der Autoridentifikation der Fall, für die in der Literaturwissenschaft seit längerem ähnliche Maße herangezogen werden. Insgesamt erscheint es sinnvoll, vor einer Kombination die Analyse der visuellen und textlichen Informationskanäle immer auch einzeln zu betrachten. Einen alternativen Zugang zu dem hier gewählten bietet die stilistische Klassifikation mit Hilfe neuronaler Netzwerke (für Comics: Laubrock & Dubray). Obwohl hier potenziell bessere Ergebnisse erzielt werden können, präferieren wir aus mehreren Gründen einen niederschwelligen Ansatz: trotz der Zuhilfenahme der PCA in den hier abgebildeten Darstellungen versprechen wir uns von der Verwendung einzelner Maße eine bessere qualitative Interpretation. Zweitens ist dieser Zugang weniger datenhungrig und daher der geringen Anzahl an Werken in unserem Corpus angemessen. In einem nächsten Schritt wollen wir die Ergebnisse der Texterkennung verbessern. Dies wird es ermöglichen, zusätzliche Textmaße und Werke für unsere Analyse heranzuziehen und unsere Ergebnisse zu verbessern.

Bibliographie

Baxter, Mike / Khitrova, Daria / Tsivian, Yuri (2016): *A Numerate Film Theory? Cinemetrics looks at Griffith, Griffith Looks at Cinemetrics*, in: Mise au Point 8, https://journals.openedition.org/map/2108 [letzter Zugriff 2. Januar 2019].

Cutting, James / Brunick, Kaitlin / DeLong, Jordan / Iricinischi, Catalina / Candan, Ayse (2011): *Quicker, faster, darker: Changes in Hollywood Film over 75 Years*, in: i-Perception 2: 569-576

Dunst, A. / Hartel, R. (2018a): *Automated Genre and Author Distinction in Comics*, DH 2018: Book of Abstracts 184-188.

Dunst, Alexander / Hartel, Rita (2018b): *The Quantitative Study of Comics: Towards a Visual Stylometry of Graphic Narrative*, in: **Dunst, Alexander / Laubrock, Jochen / Wildfeuer, Janina (Eds.):** *Empirical Comics Research: Digital, Cognitive, and Multimodal Methods*, New York: Routledge 43-61.

Dunst, Alexander / Hartel, Rita / Laubrock, Jochen (2017): *The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities*, in: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition 15-20.

Flückiger, Barbar (2017): *Analysis of Film Colors in a Digital Humanities Perspective*, in: Frames 1, http://framescinemajournal.com/article/analysis-of-film-colors-in-a-digital-humanities-perspective [letzter Zugriff 2. Januar 2019].

Hartel, Rita / Dunst, Alexander (2019): *How Good is Good Enough? Establishing Quality Thresholds for the Automatic Text Analysis of Retro-Digitized Comics*, in: Proceedings of the Multimedia Modeling Conference (Springer Lecture Notes in Computer Science 11296), https://easychair.org/publications/preprint_open/Mdf2 [letzter Zugriff 2. Januar 2019].

Holmes, David (1998): *The Evolution of Stylometry in Humanities Scholarship*, in: Literary and Linguistic Computing 13: 111-17.

Laubrock, Jochen / Dubray, David (2018): *Computational Analysis and Visual Stilometry of Comics using Convolutional Neural Networks*, in: DH 2018: Book of Abstracts 228-231.

Manovich, Lev / Douglas, Jeremy / Zepel, Tara (2011): *How to Compare One Million Images*, http://manovich.net/index.php/projects/how-to-compare [letzter Zugriff 2. Januar 2019].

Martínez, J. / Koenen, R. /Pereira, F. (2002): *MPEG-7: the generic multimedia content description standard, part 1*, in: IEEE Multimedia 9: 78-87.

Qi, Hanchao / Taeb, Armeen / Hughes, Shannon (2013): *Visual stylometry using background selection and wavelet-HMT-based Fisher information distances for attribution and dating of impressionist paintings*, in: Signal Processing 93: 541-53.

Schmidt, Ben (2014): *Search for Structures in the Simpsons and everywhere else*, http://benschmidt.org/2014/09/11/simpsons-2 [letzter Zugriff 2. Januar 2019].

Multimodale Versuche der Alignierung historischer Texte

Wagner, Andreas

wagner@rg.mpg.de
Max-Planck-Institut für europäische Rechtsgeschichte, Deutschland

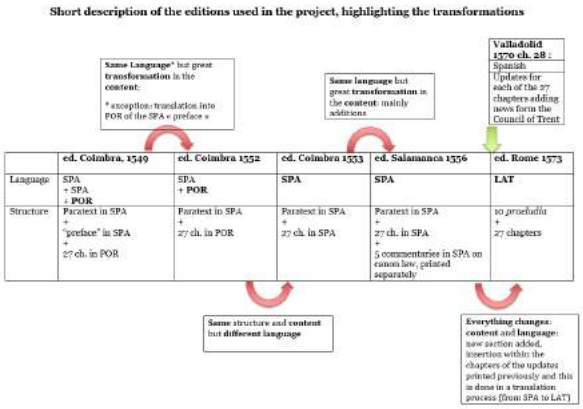
Bragagnolo, Manuela

bragagnolo@rg.mpg.de
Max-Planck-Institut für europäische Rechtsgeschichte, Deutschland

Anhand der Aufgabe einer sprachenübergreifenden Kollationierung berichtet dieser Beitrag von "multimodalen" Analysen digitaler Texte: von einer statistischen über lexikalische bis zu wissensmodellierenden Perspektiven auf den Datensatz. Wir greifen auf diese Ansätze zurück, um verschiedene Überarbeitungsstufen und Übersetzungen eines Textes zu alignieren, und wir diskutieren, warum die Aufgabe noch immer keine in der Praxis zufriedenstellende Lösung gefunden hat. So hilft der Beitrag, eine offene Forschungsfrage der Digital Humanities genauer zu bestimmen.

Das Projekt "Das Beichthandbuch des Martín de Azpilcueta und das Phänomen der Epitomierung" untersucht anhand der Entwicklung eines besonderen Texts und seiner Entwicklung den Wandel normativen Wissens in der Rückkopplung mit diversen Praxiszusammenhängen: Der spanische Kirchenrechtler Martín de Azpilcueta (1492-1586) publizierte 1549 sein "Manual de Confesores y Penitentes" mit Regeln für Verfahren und Beurteilung von Beichten. Die ursprüngliche Publikation erschien auf Portugiesisch, allein zu Azpilcuetas Lebzeiten folgten noch über 60 weitere Editionen, in denen der Autor selbst Übersetzungen und Anpassungen vornahm, etwa um auf Beschlüsse des Konzils von Trient einzugehen, oder um das Werk anderen Adressatenkreisen zu erschließen (vgl. Bragagnolo 2018).

Unser Korpus umfasst zunächst 5 zwischen 1549 und 1573 gedruckte Editionen. Zwei auf portugiesisch: (A) Coimbra 1549, 8°, 720 Seiten umfassend, (B) Coimbra 1552, 8°, 1.000 S.; zwei auf spanisch: (C) Coimbra 1553, 4°, 588 S. und (D) Salamanca 1556, 4°, 813 S.; und auf Latein (E) Rom 1573, 4°, 1.136 S. Wir gehen von drei verschiedenen Transformationsmodi aus: Änderungen des Inhalts innerhalb einer Sprache (A → B, C → D); Übersetzung in eine andere Sprache ohne größere Änderungen des Inhalts (B → C); Übersetzung unter gleichzeitiger Änderung des Inhalts (D → E).



Ein erster Beitrag digitaler Methoden zur Analyse dieser Entwicklungen besteht in der systematischen Alignierung von Texten der verschiedenen Versionen über Modifikationen und Übersetzungen hinweg. Wir diskutieren im Folgenden verschiedene Ansätze der automatischen Alignierung von sogenannten Bitexten und wie diese Ansätze sich in der Konfrontation mit den Besonderheiten des Projekts (historisches Vokabular, Orthographie und Grammatik, publizistische oder typographische Eigenheiten in den Texten, inhaltliche Überarbeitungen in den Übersetzungen usw.) bewähren. Ein wichtiger Gesichtspunkt sind dabei immer auch die Art, der Umfang und die Auswirkungen der nötigen manuellen/intellektuellen Vor- und Nachbereitungen.

Für die Evaluation der verschiedenen Ansätze alignieren wir einen Teil der im Projekt als TEI XML transkribierten Texte in der LERA Umgebung[1] manuell. Da die Texte zum Teil umfangreiche Überarbeitungen enthalten, wird zu sehen sein, ob automatische Methoden der Evaluation (wie Papineni et al. 2002 oder Lin/Och 2004) Verwendung finden können, oder ob doch auf eine manuelle Evaluation zurückgegriffen werden muss (ähnlich Darriba Bilbao et al. 2005).

I. Statistische Modi

Im ersten Teil diskutieren wir Algorithmen, die ausblenden, dass es sich bei unseren Daten um symbolische bzw. sprachliche Ausdrücke handelt. Sie werden gleichsam jeweils als "rohe" Datenmengen verstanden, die auf statistische Weisen vermessen werden können und es werden Übereinstimmungen in den Mustern oder in

den Intervall-Längen zwischen spezifischen Datenpunkten gesucht.[2] Obwohl in allen Fällen bestimmte Besonderheiten des historischen Forschungsgegenstands zu Komplikationen führen, ist die Leistungsfähigkeit dieser Ansätze nicht zu unterschätzen. Denn ihre Unzulänglichkeiten sind weitgehend mit jenen besonderen Zusammenhängen historischer Texte verschränkt, in denen ohnehin manuell nach- oder vorgearbeitet werden muss, und es ist nicht von vornherein auszuschließen, dass es sich lohnen könnte, mit manuellem Aufwand die Texte besser vorzubereiten, um dann mit diesen Ansätzen sehr gute Ergebnisse erzielen zu können.

1. Für die meisten Methoden der computergestützten Übersetzung (*Machine Translation*) stellt der *Satz* die grundlegende Einheit der Übersetzung dar und es haben sich eine Reihe von Ansätzen etabliert, die zur Erkennung von Satzkorrelationen in Bitexten allein auf die bloße Satzlänge als eines der besten Maße für die Wahrscheinlichkeit abstellen, mit der ein Satz im zu untersuchenden Dokument die Übersetzung eines Referenz-Satzes aus dem Quell-Dokument ist.[3] Die Unterschiede in den typischen Satzlängen zwischen zwei Sprachen schlagen sich offenbar in allen Sätzen eines Dokuments in ähnlicher Weise nieder, so dass sich in zwei Dokumenten die Verhältnisse der Satzlängen zueinander stark ähneln. Fälle, in denen allzu kurze Sätze beim Übersetzen verbunden, oder sehr lange Sätze aufgeteilt werden, werden mit geringerer Genauigkeit erkannt; wie häufig dieser Fall aber vorkommt, hängt von den involvierten Sprachen und Übersetzern ab.
2. Ein zweiter Ansatz aus dem *Machine Translation*-Umfeld sind geometrische Ansätze (vgl. Melamed 1999). Sie basieren auf der Annahme, dass es ausreichend sein müsste, sehr grob markierte "Kandidaten" für Satzkorrespondenzen in die richtige Reihenfolge zu bringen. Mit anderen Worten liegt der Fokus nicht auf der eigentlichen Übereinstimmung, sondern auf der Position im Text: Die Ausgangsannahme ist, dass die zu vergleichenden Texte synchron fortschreiten und der erste Satz im einen Text den ersten Satz im anderen übersetzt, der zweite den zweiten usw. Diese Annahme kann in einem durch den Fortschritt in beiden Texten aufgespannten Koordinatensystem als ansteigende Diagonale repräsentiert werden. In einer geometrischen Betrachtung wird dann versucht, durch Umsortierung der vorgefundenen Sätze, die Punkte an diese Diagonale anzunähern.
3. Da unsere Texte in eng verwandten Sprachen vorliegen – Portugiesisch, Spanisch und Latein –, erscheint es lohnenswert, auch mit Ansätzen, die Übereinstimmungen auf der Ebene von Wortstämmen oder -fragmenten untersuchen, einen Versuch zu unternehmen (vgl. Darriba Bilbao et al. 2005). Wir untersuchen also Ähnlichkeiten in den Vektorräumen für die vorkommenden 3- und 4-Gramme.[4]

II. Lexikalische Modi

Eine zweite Menge von Methoden der *Machine Translation* verarbeitet die Daten nur in sprachlogisch aufbereiteter Form, hebt insbesondere auf die übereinstimmende Bedeutung der sprachlichen Ausdrücke ab und setzt viele "klassische DH"-Ansätze ein (vgl. Ma 2006). Diese Ansätze setzen Arbeitsschritte wie Tokenisierung und Stemmatisierung oder Lemmatisierung voraus und in unseren Experimenten evaluieren wir verschiedene weitere, optionale Schritte, um zunächst zu einer treffenden Charakterisierung *eines* Textes zu gelangen. Dies wird für beide Sprachversionen vorgenommen, bevor dann diese "konzentrierten" oder "gefilterten" Charakterisierungen endlich auf der Basis eines Wörterbuchs *miteinander* verglichen werden.[5]

Die von uns evaluierten optionalen Schritte zur Etablierung einer Charakteristik von Sätzen sind (a) "Filter" wie Stopwörter und TF/IDF-Topwerte und (b) "Booster" wie stärker gewichtete Zahlen, Zahlwörter und Named Entities. Offenkundig hängt allerdings das Ergebnis der Vergleiche in dieser zweiten Perspektive mindestens ebensosehr von der Qualität der Wörterbücher wie von der Leistung und der Auswahl der vorgeschalteten "Charakterisierungs"-Algorithmen ab. Daher legen wir ein besonderes Augenmerk auf das relative Gewicht der Qualität des Wörterbuchs und ihrer manuellen Verbesserung auf der einen, des Aufwands und Gewinns beim Einsatzes von Filtern und Boostern auf der anderen Seite.

III. Wissensbasierte Modi

Abschließend stellen wir mit der Graphanalyse eine Perspektive vor, die in aktuellen Diskussionen zur sprachübergreifenden Plagiatserkennung diskutiert wird und eine Modellierung des im Text beschriebenen Wissens unternimmt (vgl. Franco-Salvador/Rosso/Montes-y-Gómez 2016). Anstelle eines Wörterbuchs zur Überbrückung des Sprachunterschieds wird hier ein semantisches Netz – in unserem Beispiel BabelNet (Navigli/Ponzetto 2012) – verwendet, um die Wörter der Texte mit "sprachunabhängigen" Konzepten zu verbinden, die untereinander in taxonomischen, synonymen, kontradiktorischen u.a. Beziehungen stehen. Dabei wird durch die Texte jeweils ein Ausschnitt eines umfassenderen Begriffsgraphen instanziiert, um anschließend die resultierenden Teilgraphen miteinander zu vergleichen. Dies erlaubt die Disambiguierung der verwendeten Wörter und eine differenziertere Vergleichsbasis durch die Einbeziehung des semantischen Kontexts der verglichenen Textpassagen. Die Konstruktion und die Vergleiche der zahlreichen Teilgraphen sind offenkundig rechenintensivere Aufgaben, und im Übrigen setzt der Ansatz ebenfalls Schritte wie Tokenisierung und Lemmatisierung voraus, so dass der mögliche Gewinn in der Vergleichsgenauigkeit hier durch einen höheren Aufwand erkauft wird, der zu einem kaum verminderten Aufwand der Textaufbereitung (z.B. der Normalisierung) hinzu kommt.

Diskussion

Wir diskutieren im Rahmen des Beitrags insbesondere, welche Komplikationen sich in der Arbeit in der Folge von Besonderheiten unseres Fragezusammenhangs und Materials gezeigt haben, und wie diese sich auf die unterschiedlichen Ansätze jeweils auswirken. Als wichtige Faktoren konnten wir historische Orthographie und Abkürzungen, uneindeutige und inkonsistente Interpunktion sowie die Kodierung von

Layout-Besonderheiten wie Fußnoten identifizieren und so versuchen wir zu bestimmen, welchen Gewinn entsprechende manuelle Vorarbeiten wie Satzsegmentierung, absatzweise Alignierung, Verbesserung des Wörterbuchs, Normalisierung von Schreibungen und Typographie erzielen können.

Fußnoten

1. Paul Molitor, Jörg Ritter et al.: LERA - Locate, Explore, Retrace and Apprehend complex text variants , eine im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekts SaDA - Semi-automatische Differenzanalyse von komplexen Textvarianten erstellte Arbeitsumgebung.
2. Im linguistischen Kontext entspricht die Mustersuche des ersten Ansatzes etwa einer n-Gramm-Analyse, und wenn die herausgehobenen Datenpunkte des zweiten Ansatzes Repräsentationen von Interpunktionszeichen sind, entspricht dieser einer Analyse der Satzlängen. Obwohl die Auswahl der verwendeten Maße so durchaus durch sprach- und texttheoretische Überlegungen inspiriert und angeleitet ist, sind die Maße selbst von diesen Motiven doch im Grunde vollkommen unabhängig und könnten in gleicher Weise mit ganz anders gearteten Datenreihen angewandt werden. (In Sankoff/Kruskal 1983 etwa werden Anwendungen der Sequenz-Alignierung in ganz anderen Feldern beschrieben.)
3. Die frühesten Versuche in dieser Richtung wurden wohl im IBM Machine Translation Lab unternommen; vgl. Brown et al. 1990. Klassisch wurde der Algorithmus und der Aufsatz von Gale/Church 1993; aktueller, mit weiteren Methoden kombiniert und auf sog. “low ressourced languages” zielend etwa bei Varga et al 2005.
4. Als zusätzliche Dimension haben wir dem Vektorraum die Position des jeweiligen Satzes im Text sowie die Behandlung der benachbarten Sätze hinzugefügt, so dass von zwei Satzpaaren mit gleichen n-Gramm-Häufigkeiten dasjenige den Vorzug erhalten kann, dessen Sätze näher beieinander liegen oder das eine mit den benachbarten Sätzen vergleichbare Verschiebung darstellt.
5. Ansätze wie Kay/Röscheisen 1993, Fung/Church 1994 oder auch Varga et al. 2005 können ein solches Wörterbuch auf der Basis allein der vorliegenden Texte erstellen.

Bibliographie

Manuela Bragagnolo: *Les voyages du droit du Portugal à Rome. Le ‘Manual de confessores’ de Martín de Azpilcueta (1492-1586) et ses traductions*, (The Travels of Law from Portugal to Rome. Martín de Azpilcueta’s ‘Manual de confessores’ (1492-1586) and its Translations), Max Planck Institute for European Legal History Research Paper Series No. 2018-13 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3287684)

Peter F. Brown / John Cocke / Stephen A. Della Pietra / Vincent J. Della Pietra / Fredrick Jelinek / John D. Lafferty / Robert L. Mercer / Paul S. Roossin: *A statistical approach to machine translation*, in: Computational Linguistics 16 (1990): 79-85, https://dl.acm.org/citation.cfm?id=92860.

V.M. Darriba Bilbao / J.G. Pereira Lopes / T. Ildefonso: *Measuring the impact of cognates in parallel text alignment*, in: Proceedings of the Portuguese Conference on Artificial Intelligence (2005): 338-343. DOI: 10.1109/EPIA.2005.341306.

Ábel Elekes / Adrian Englhardt / Martin Schäler / Klemens Böhm: *Toward meaningful notions of similarity in NLP embedded models*, in: International Journal on Digital Libraries (2018), DOI: 10.1007/s00799-018-0237-y.

Samuel Fernando / Mark Stevenson: *A semantic similarity approach to paraphrase detection*, in: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics (2008), 45-52, https://pdfs.semanticscholar.org/d020/eb83f03a9f9c97e728355c4a9010fa65d8ef.pdf.

Marc Franco-Salvador / Paolo Rosso / Manuel Montes-y-Gómez: *A systematic study of knowledge graph analysis for cross-language-plagiarism detection*, in: Information Processing and Management 52 (2016), 550-570. DOI: 10.1016/j.ipm.2015.12.004.

Pascale Fung / Kenneth W. Church: *K-vec: A new approach for aligning parallel texts*, in: Proceedings of the 15th Conference on Computational Linguistics, Vol. 2 (1994), 1096-1102, DOI: 10.3115/991250.991328.

William A. Gale / Kenneth W. Church: *A program for aligning sentences in bilingual corpora*, in: Computational linguistics 19/1 (1993): 75-102, https://dl.acm.org/citation.cfm?id=972455.

Martin Kay / Martin Röscheisen: *Text-translation Alignment*, in: Computational Linguistics 19/1 (1993), 121-142.

Tom Kenter / Maarten de Rijke: *Short Text Similarity with Word Embeddings*, in: CKIM ’15 Proceedings (2015) DOI: 10.1145/2806416.2806475.

Chin-Yew Lin / Franz Josef Och: *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics*, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (2004). DOI: 10.3115/1218955.1219032.

Xiaoyi Ma: *Champollion: A robust parallel text sentence aligner*, in: 5th International Conference on Language Resources and Evaluation (LREC) 2006, 489-492.

Helena de Medeiros Caseli / Maria das Graças Volpe Nunes: *Evaluation of sentence alignment methods for brazilian portuguese and english parallel texts*, in Brazilian Symposium on Artificial Intelligence (SBIA) (2004), 184-193, DOI: 10.1007/978-3-540-28645-5_19.

I. Dan Melamed: *A Portable Algorithm for Mapping Bitext Correspondence*, in: ACL ’98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (1997), 305-312, DOI: 10.3115/976909.979656.

I. Dan Melamed: *Bitext Maps and Alignment via Pattern Recognition*, in: Computational Linguistics 25/1 (1999), 107-130, https://dl.acm.org/citation.cfm?id=973218.

Robert C. Moore: *Fast and accurate sentence alignment of bilingual corpora*, in: S.D. Richardson (ed.): AMTA 2002. Machine Translation: From Research to Real Users, LNCS 2499 (2002), pp. 135-144. DOI: 10.1007/3-540-45820-4_14.

Roberto Navigli / Simone Paolo Ponzetto: *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, in: Artificial Intelligence 193 (2012), 217-250, DOI: 10.1016/j.artint.2012.07.001

Kishore Papineni / Salim Roukos / Todd Ward / Wei-Jing Zhu: *BLEU: A Method for Automatic Evaluation of Machine Translation*, in: Proceedings of the 40th Annual Meeting on

Association for Computational Linguistics (2002): 311-318. DOI: 10.3115/1073083.1073135.

Christian Paul / Achim Rettinger / Aditya Mogadala / Craig A. Knoblock / Pedro Szekely: *Efficient graph-based document similarity*, in: **H. Sacks et al. (eds.):** *ESWC '16 European Semantic Web Conference / LNCS 9678 Lecture Notes in Computer Science* (2016), 334-349, DOI: 10.1007/978-3-319-34129-3_21.

Alexandr Rosen: *In search of the best method for sentence alignment in parallel texts*, in: **R. Garabık (ed.):** Computer treatment of Slavic and East European languages. Third international seminar (2005), 174-185, http://utkl.ff.cuni.cz/~rosen/public/slovko05.pdf.

David Sankoff / Joseph Kruskal: *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. Addison-Wesley (1983).

André Santos / José João Almeida / Nuno Carvalho: *Structural Alignment of plain text books*, in: LREC '12 Proceedings of the Eigth International Conference on Language Resources and Evaluation (2012), 2069-2074, http://www.lrec-conf.org/proceedings/lrec2012/pdf/967_Paper.pdf.

Danica Seničić: *Automatic alignment of bilingual sentences*. The case of English and Serbian. M.A. thesis, Louvain, 2016, https://dial.uclouvain.be/memoire/ucl/en/object/thesis%3A11186.

Daniel Stein: *Machine translation: Past, present and future*, in: **Georg Rehm / Felix Sasaki / Daniel Stein / Andreas Witt (eds.):** *Language technologies for a multilingual Europe*, TC3 III. Language Science Press (2018), pp. 5-17. DOI: 10.5281/zenodo.1291924.

Joseph P. Turian / Luke Shen / I. Dan Melamed: *Evaluation of Machine Translation and its Evaluation*, in: Proceedings of Machine Translation Summit Proceedings of Machine Translation Summit IX (2003), 386-393, https://nlp.cs.nyu.edu/pubs/papers/turian-summit03eval.pdf.

Dániel Varga / Péter Halácsy / András Kornai / Viktor Nagy / László Németh / Viktor Trón: *Parallel Corpora for medium density languages* [Hunalign], in: RANLP '05 Proceedings of Recent Advances in Natural Language Processing (2005), 247-258, http://kornai.com/Papers/ranlp05parallel.pdf.

Krzysztof Wołk / Krzysztof Marasek: *A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation*, in: Advances in Intelligent Systems and Computing 275 (2014), 107-114, DOI: 10.1007/978-3-319-05951-8_22.

Netzwerkanalyse narrativer Informationsvermittlung in Dramen

Vauth, Michael
michael.vauth@tuhh.de
Technische Universität Hamburg, Deutschland

Einleitung

In diesem Beitrag wird ein Verfahren vorgestellt, das Netzwerkvisualisierungen dramatischer Texte für eine spezifische Form der kommunikativen Interaktion zwischen Figuren fokussiert.

Es wird gezeigt, inwiefern gewichtete, gerichtete und dynamische Figurennetzwerke narrative Informationsvermittlung in der Figurenrede visualisieren können und auf diesem Weg dramennarratologische Analysen bzw. Annotationen ausgewertet werden.

Im Gegensatz zu literaturwissenschaftlichen Netzwerkanalysen, die um die automatisierte Analyse des „kompositorische[n] Grundgerüst[s]" (Trilcke 2013: 224) von großen Dramenkorpora (Piper et al 2017; Trilcke et al. 2015) bemüht sind, steht in diesem Beitrag also die Visualisierung von manuellen Annotationen im Vordergrund.

Darüber hinaus werden mit Rückgriff auf die ermittelten Netzwerkdaten Deutungspotenziale exemplarisch an Kleists *Die Familie Schroffenstein* (DFS) diskutiert.[1] Das Erkenntnisinteresse zielt also auf zwei Aspekte: (1) Inwiefern lassen sich narrative Redebeiträge, die ein zentrales Element der inneren und äußeren Informationsvermittlung im Drama (Pfister 2001: 20-22) sind, durch Annotationen netzwerkgraphisch visualisieren? (2) Inwiefern stellt die literaturwissenschaftliche Netzwerkanalyse in diesem Kontext einen Mehrwert dar?

Annotation narrativer Figurenrede

Ausgangspunkt der vorgestellten Netzwerke ist eine Typologie narrativer Figurenrede bzw. von Binnenerzählungen, die zur Annotation der Dramen Heinrich von Kleists genutzt wurden.[2] Dabei wurden über 800 Vorkommnisse narrativer Figurenrede in den Dramen manuell annotiert. In einem ersten Schritt unterscheidet die Typologie zwischen narrativen Äußerungen, mit denen Figuren über ihre eigene Wirklichkeit erzählen, und narrativer Figurenrede, bei der das nicht der Fall ist. Der erste Phänomentyp, die horizontalen Binnenerzählungen, können mit dem narratologischen Kategorieninventar zur Beschreibung anachronen Erzählens (Lahn & Meister 2008: 138-141) genauer beschrieben und annotiert werden:[3]

Binnenerzählungen		
Horizontal	Analepsen	488
	Simullepsen	123
	Prolepsen	33
Vertikal	Pseudoanalepsen	33
	Pseudosimullepsen	6
	Pseudoprolepse	29

Tabelle 1: Vorkommen narrativer Figurenrede in Kleists Dramen

Diese manuellen Annotationen sind die Grundlage dafür, dass unterschiedliche Formen der Informationsvermittlung netzwerkgraphisch visualisiert werden können.

Netzwerkerstellung

Unter Rückgriff auf die Annotationen der narrativen Figurenrede und die TEI-Annotationen von Sprecherfiguren und Szenen- sowie Aktwechseln im TextGrid-Korpus wurden chronologisierte Sender-Adressaten-Kanten erstellt. Dazu wurden die vier Sprecherfiguren, die auf eine narrative Äußerung folgen oder ihr vorangehen, als Adressaten berücksichtigt, sofern keine Akt- oder Szenenwechsel zwischen narrativer Figurenrede und potenziellem Adressaten liegt und es sich um unterschiedliche Figuren handelt. Die Anzahl der erzeugten Kanten ist somit deutlich höher als die Anzahl der narrativen Äußerungen (siehe exemplarisch in Tabelle 2 die Kanten 4 und 5 sowie 8-10).

Id	Source (Sprecher)	Target (Adressat)	Label (Akt)	Timeset (Beginn & Ende)	Weight
1	Rupert	Eustache	1	"<[1.0,48.0]>"	1
2	Rupert	Jeronimus	1	"<[2.0,48.0]>"	1
3	Rupert	Ottokar	1	"<[3.0,48.0]>"	1
4	Jeronimus	Ottokar	1	"<[4.0,48.0]>"	1
5	Jeronimus	Ottokar	1	"<[5.0,48.0]>"	2
6	Ottokar	Jeronimus	1	"<[6.0,48.0]>"	1
7	Jeronimus	Kirchenvogt	1	"<[7.0,48.0]>"	1
8	Kirchenvogt	Jeronimus	1	"<[8.0,48.0]>"	1
9	Kirchenvogt	Jeronimus	1	"<[9.0,48.0]>"	2
10	Kirchenvogt	Jeronimus	1	"<[10.0,48.0]>"	3

Tabelle 2: Auszug aus der Kantenliste zu DFS

Das Kantengewicht (*Weight*) zwischen einer Sender- und einer Adressatenfigur steigt mit jeder narrativen Äußerung, die eine Senderfigur im ‚Beisein' derselben Adressatenfigur äußert (siehe z.B. Kante 4 und 5).

Die Figurennetzwerke, die auf dieser Grundlage erstellt werden, illustrieren, welche Figuren sich zu welchem Zeitpunkt des Dramenverlaufs narrativ äußern, welche Figuren narrative Informationen bekommen und wie häufig Figuren an narrativem Informationsaustausch beteiligt sind.

Visualisierungs- und Analysebeispiele

Abbildung 1 zeigt dieses Potential der Netzwerkvisualisierung exemplarisch für *Die Familie Schroffenstein*.[4]

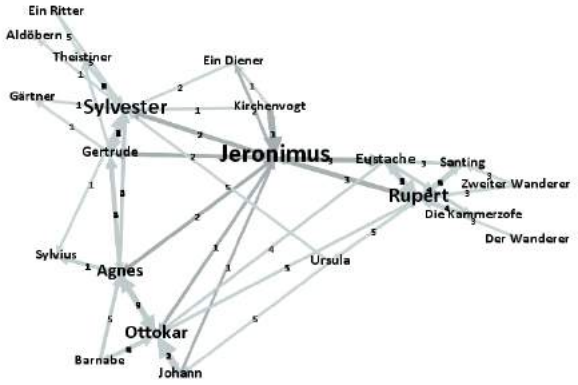


Abbildung 1. Narrative Informationsvermittlung in DFS

Die Größe der Knotenbeschriftung repräsentiert hier die betweenness centrality[5] der Figuren und damit den Stellenwert bzw. Einfluss der Figur auf die narrative Informationsvermittlung innerhalb des Dramas.

Label	betweenness centrality	weighted indegree	weighted outdegree	indegree	outdegree	degree
Jeronimus	125,45	99	82	25	28	53
Gertrude	98,74	46	70	15	15	30
Sylvester	85,37	107	31	29	14	43
Rupert	74,47	145	17	36	10	46
Agnes	45,90	59	85	18	23	41
Eustache	17,25	37	73	14	19	33
Ottokar	5,11	180	51	33	14	47
Ursula	3,31	1	4	1	4	5
Santing	2,94	10	49	6	12	18
Johann	2,35	4	106	3	15	18
Kirchenvogt	0,00	1	80	1	14	15
Ein Diener	0,00	2	2	2	2	4
Sylvius	0,00	7	3	4	2	6
Gärtner	0,00	2	0	2	0	2
Aldöbern	0,00	1	0	1	0	1
Theistiner	0,00	5	15	4	5	9
Der Wanderer	0,00	0	1	0	1	1
Zweiter Wanderer	0,00	0	4	0	3	3
Die Kammerzofe	0,00	2	13	2	6	8
Barnabe	0,00	1	21	1	8	9
Ein Ritter	0,00	0	2	0	2	2

Tabelle 3: Betweenness centrality, (weighted) in- und outdegree in DFS

Schon anhand dieses Beispiels und der netzwerkmetrischen Daten in Tabelle 3 lassen sich einige Vorzüge einer netzwerkgraphischen Annotationsauswertung zeigen:

- Die Reichweite einer Figurenerzählung wird im Hinblick auf den Adressatenkreis ermittelt.
- Die narrativen Funktionen der Figuren werden durch ihre Netzwerkeigenschaften erkennbar:
  - *Brückenfiguren*: hohe betweenness centrality; Verbindung getrennter Netzwerkbereiche (Trilcke 2013: 217): z.B. Jeronimus (siehe hier und nachfolgend Tabelle 2).
  - *Botenfiguren*: Geringe betweenness centrality; mind. eine narrative Äußerung:

    - *Botenfiguren i.e.S.*, die *nach* dem ersten Akt erzählend in Erscheinung treten: z.B. die Wanderer, der Ritter und Barnabe.
    - *Expositionsfiguren*, die *im* ersten Akt/Dramenteil erzählend in Erscheinung treten: z.B. der Kirchenvogt.[6]
  - *Zielfiguren*: hohe betweenness centrality; hoher gewichteter Eingangsgrad; geringer gewichteter Ausgangsgrad: z.B. Sylvester und Rupert, die häufig die Adressaten, aber selten die Sprecher narrativer Figurenrede sind. (Die Handlung des hier gewählten Beispieltexts legt die These nahe, dass diese Figuren entscheidungsmächtige Figuren sind und daher zahlreiche Informationen bekommen.)
  - *Figuren der Informationskontrolle*: hohe betweenness centrality; sehr hoher Ausgangsgrad; Netzwerke mit geringer Kantendichte: z.B. Hermann in Kleists *Hermannsschlacht.*[7]
- Es lassen sich Figurenpaare und Netzwerkbereiche identifizieren, zwischen denen es keinen oder nur vermittelten Informationsaustausch gibt. Hier sind natürlich Dyaden besonders interessant, bei denen die beiden Figuren eine hohe betweenness centrality aufweisen: z.B. Rupert und Sylvester.
- Die Informationsstrukturen geben Aufschluss über den allgemeinen Grad der Informiertheit der Figuren: z.B. die Kantendichte.[8]

Zudem können unterschiedliche Formen der narrativen Figurenrede netzwerkgraphisch miteinander verglichen werden. Die Abbildungen 3 und 4 zeigen dies exemplarisch. In Abbildung 3 werden Figurenerzählungen visualisiert, in denen sich Figuren in Übereinstimmung mit der fiktionalen Wirklichkeit äußern. Abbildung 4 zeigt narrative Äußerungen, bei denen das Gegenteil der Fall ist. Es handelt sich also um narrative Falschaussagen. Der Vergleich ist in diesem Fall aufgrund der vorhandenen Parallelen *und* Unterschiede aufschlussreich. Bei beiden Netzwerken behält Jeronimus die zentrale Position im Netzwerk. Hier schlägt sich nieder, dass er für die Verbreitung von wirklichkeitsgemäßen Informationen ebenso verantwortlich ist, wie für die Verbreitung von falschen Verdächtigungen, Lügen und Vorurteilen. Agnes' Netzwerkposition verändert sich hingegen stark (Abb. 4). Sie ist innerhalb ihrer Familie und in der kommunikativen Interaktion mit Ottokar die zentrale Figur bei der Weitergabe falscher Informationen.

Informationsvermittlung im Dramenverlauf: Dynamische Netzwerke

Wie Agarwal et al. (2012: 94) gezeigt haben, hat die Erstellung von dynamischen Netzwerken den Vorteil, die Veränderlichkeit der netzwerkmetrischen Eigenschaften einer Figur, einer Figurengruppe oder eines gesamten Netzwerks im Verlauf eines Roman- oder Dramengeschehens berücksichtigen zu können. Das bestätigen die netzwerkmetrischen Auswertungen der *Familie Schroffenstein* in Abbildung 2. Hier wird der gewichtete Ausgangsgrad für vier ausgewählte Figuren aktweise dokumentiert. So tritt der Kirchenvogt narrativ nur im ersten Akt in Erscheinung, was seine Funktion als Expositionsfigur unterstreicht. Auch Barnabes Funktion als Vermittlerin von Anagnorisis-Informationen zum Dramenende spiegelt sich wider. Jeronimus Bedeutung relativiert sich, weil ersichtlich wird, dass er – aufgrund seiner Ermordung am Ende des dritten Akts – nur in den ersten drei Akten als informationsvermittelnde Figur auftritt. Seine hohen Werte in Akt zwei und drei zeigen jedoch, dass er für den Handlungsverlauf entscheidende Informationen äußert. Bei Rupert bestätigt sich seine geringe narrative Aktivität (Tabelle 3) als ein relativ konstantes Verhalten. Der niedrige gewichtete Ausgangsgrad für das gesamte Drama ist also nicht darauf zurückzuführen, dass er nur in wenigen Szenen (erzählerisch) in Erscheinung tritt.

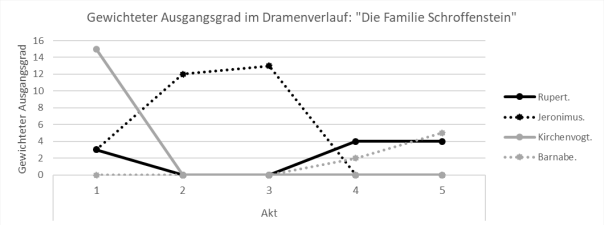


Abbildung 2. Gewichteter Ausgangsgrad im Dramenverlauf in DFS

Schluss

Solange die automatische Annotation narrativer Figurenrede nicht möglich ist, setzt das vorgestellte Verfahren einen relativ großen Annotationsaufwand voraus. Es ermöglicht somit keinen umfassenden Vergleich von Dramen, was unter anderem zur Einordnung der vorgestellten quantitativen Netzwerkanalysen wünschenswert wäre.

In diesem Beitrag wurde jedoch exemplarisch gezeigt, inwiefern netzwerkgraphische Visualisierungen für die Auswertung narratologischer Annotationen einen analytischen Mehrwert haben können. Die formalen Annotationen können und sollen durch inhaltsbezogene Annotationen angereichert werden. Auf dieser Grundlage könnte netzwerkgraphisch der Informationsaustausch über bestimmte Themen oder Figuren visualisiert werden.

Anhang

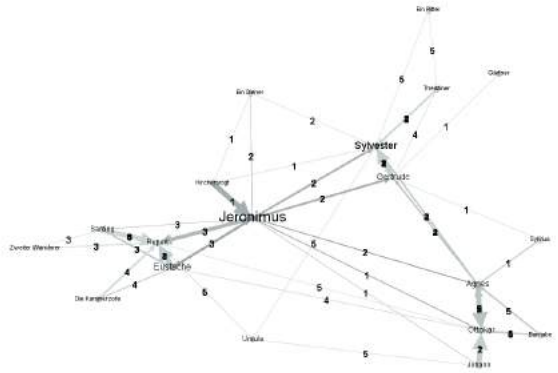


Abbildung 3. Narrative Informationsvermittlung (Horizontale Binnenerzählungen/Wirklichkeitserzählungen) in DFS

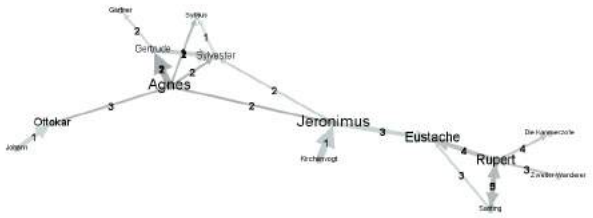


Abbildung 4. Narrative Informationsvermittlung (Pseudoanalepsen/ Falschaussagen) in DFS

Label	**betweenness centrality**	**weighted indegree**	**weighted outdegree**	indegree	outdegree
Hermann	482,7	66	47	36	27
Thuiskomar	107,25	5	10	5	6
Thusnelda	69,62	28	25	13	9
Ventidius	61,95	7	27	5	14
Varus	57,8	10	9	6	8
Dagobert	27	3	3	2	3
Zweite Hauptmann	21	2	1	2	1
Gertrud	20	7	6	4	3
Marbod	16,2	11	5	7	4
Wolf	11,96	5	4	5	4
Aristan	8,5	1	5	1	4
Rinold	6	1	5	1	4
Erste Cherusker	6	1	5	1	4
Der zweite Cherusker	5,2	1	2	1	2
Gueltar	5	1	2	1	2
Der Mann	4	1	3	1	3
Das Volk	2	2	1	2	1
Erste Älteste	0,5	1	1	1	1
Scäpio	0,33	1	5	1	4

Tabelle 4: Figuren mit höchster betweenness centrality in Kleist *Hermannsschlacht*

Fußnoten

1. Textgrundlage der Annotationen war Kleists Werkausgabe von Siegfried Streller, die durch das TextGrid Repositorium digital zur Verfügung steht: https://textgrid.de/en/digitale-bibliothek.
2. Dazu wurde das Annotationstool CATMA (Meister et al. 2016) verwendet. CATMA bietet die Möglichkeit, mit selbstdefinierten literaturwissenschaftlichen Analysetaxonomien zu annotieren und ist damit für narratologische Forschungsprozesse besonders geeignet.
3. Grundlegend für die Annotation ist ein Narrativitätskonzept, das berücksichtigt, dass Texte aller Gattungen narrative Elemente enthalten können, wie es u.a. Wolf (2002) beschreibt. Zu der narratologischen Terminologie vgl. Lahn/Meister 2016: 147-149.
4. Alle Visualisierungen und Netzwerkanalysen wurden mit dem Tool GEPHI (Bastian et al. 2008) erstellt.
5. Mit der betweenness centrality wird gemessen, für wieviele Knotenpaare ein Knoten den kürzesten Netzwerkpfad darstellt. Eine hohe betweenness centrality in Netzwerken, die Informationsflüsse abbilden, indiziert also einen großen Einfluss der Figur auf die Informationsvermittlung im Netzwerk, da sie als Brückenfigur fungiert.
6. Vgl. zum Unterschied Pfister 2001: 280f.
7. Kantendichte der *Hermannsschlacht*: 0,051; Kantendichte *Die Familie Schroffenstein*: 0,388. Siehe zur *Hermannsschlacht* Tabelle 4 im Anhang, in der sich Hermanns propagandistische „Überzeugungsarbeit" (Müller-Salget 2009: 78) widerspiegelt.
8. Hier ist zu berücksichtigen, dass die Kantendichte natürlich auch durch andere Faktoren beeinflusst wird (Trilcke 2013: 225).

Bibliographie

Agarwal, A. / A. Corvalan / J. Jensen / O. Rambow (2012): *Social Network Analysis of Alice in Wonderland*, Proceedings of the Workshop on Computational Linguistics for Literature: 88–96.

Bastian, M. / S. Heymann / M. Jacomy (2008): *Gephi: An open source sofware for exploring and manipulating networks.* AVI 2008 – Proceedings of the Working Conference on Advanced Visual Interfaces.

Lahn, Silke/ J. C. Meister (2008): *Einführung in die Erzähltextanalyse*. Stuttgart: Verlag J.B. Metzler.

Meister, J. C. / M. Petris / E. Gius / J. Jacke (2016): *CATMA 5.0. software for text annotation and analysis*.

Moretti, F. (2011): *Network Theory, Plot Analysis.* Stanford Literary Lab Pamphlets 2.

Müller-Salget, Klaus (2009): *Die Herrmannsschlacht*, in: **Ingo Breuer (Hg.):** *Kleist-Handbuch. Leben – Werk – Wirkung*. Stuttgart: Verlag J.B. Metzler. S. 76-79.

Piper, Andrew / Mark Algee-Hewitt / Koustuv Sinha / Derek Ruths / Hardik Vala (2017): *Studying Literary Characters and Character Networks*. Digital Humanities 2017, Conference Abstracts.

Pfister, Manfred (2001): *Das Drama*. München. Wilhelm Fink Verlag.

Trilcke, P. / F. Fischer / D. Kampkaspar (2015): *Digital Network Analysis of Dramatic Texts*. Digital Humanities 2015: Book of Abstracts.

Trilcke, Peer (2013): *Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft*. Empirie in der Literaturwissenschaft. Hrsg. von Christoph Rauen, Katja Mellmann und Philip Ajouri. Münster: 201–247.

Wolf, Werner (2002): *Das Problem der Narrativität in Literatur, bildender Kunst und Musik: Ein Beitrag zu einer intermedialen Erzähltheorie*. Erzähltheorie transgenerisch, intermedial, interdisziplinär. Hrsg. von Vera Nünning und Ansgar Nünning. Trier: 23–104.

Nomisma.org: Numismatik und das Semantic Web

Wigg-Wolf, David
david.wigg-wolf@dainst.de
Römisch-Germanische Kommission des Deutschen Archäologischen Instituts

Tolle, Karsten
tolle@dbis.cs.uni-frankfurt.de
Johann Wolfgang Goethe-Universität Frankfurt am Main

Kissinger, Timo
Timo.Kissinger@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Einführung

In diesem Beitrag wird eine domänenspezifische Anwendung von Linked Open Data und Ontologien vorgestellt, die als Paradigma für den Umgang mit digitalen Corpora in der Archäologie dienen kann. Zunächst wird die Entwicklung des Projektes Nomisma.org sowie dessen Anwendung für die Verlinkung von digitalen Datenbeständen erläutert. In einem zweiten Teil wird ein Pilotprojekt vorgestellt, das darauf abzielt, textbasierte Münzpublikationen als RDF zur Verfügung zu stellen und somit prüft, inwieweit das Vokabular und die Ontologie von Nomisma.org eingesetzt werden können.

Das 2010 von der American Numismatic Society, New York, initiierte Projekt Nomisma.org[1] definiert und stellt stabile digitale Repräsentationen numismatischer Konzepte nach den Prinzipien von Linked Open Data zur Verfügung. Sie werden als http-URIs veröffentlicht[2], die Zugang zu weiterverwertbaren Informationen zu den Konzepten liefern, dazu noch Links zu weiteren Linked Open Data-Ressourcen (Getty Vocabularies, wikidata, viaf, GND, u.v.m.). Ferner wurde eine numismatische Ontologie[3] entwickelt, die den Bedürfnissen und Arbeitsweisen der Numismatik gerecht ist und eine unkomplizierte Modellierung ermöglicht. Das kanonische Format von Nomisma.org ist RDF/XML, aber auch weitere Formate wie JSON-LD (für Geodaten geoJASON-LD), Turtle, KML und HTML5+RDFa 1.1. werden bedient.

Virtuelle Sammlungen

Zunächst lag der Schwerpunkt der Arbeit von Nomisma.org bei der römischen Numismatik und bis 2016 erfolgte die Onlinestellung der Linked Open Data-Ressourcen *Coinage of the Roman Republic Online* (CRRO)[4] und *Online Coins of the Roman Empire* (OCRE)[5] als virtuelle Münzsammlungen für die Prägungen der römischen Republik und der Kaiserzeit. Die Abfrage erfolgt über RDF Dumps der Projektpartner, die zentral gehalten werden; die dazugehörigen Bilder werden on-the-fly von den Servern der Partner geholt, wodurch nicht nur das Volumen der zentralen Datenhaltung gering gehalten und die Abfragegeschwindigkeit erhöht, sondern auch eventuelle rechtliche Probleme bei der Lizenzierung der Bilder vermieden werden.

Mittlerweile befassen sich erste Projekte auch mit dem disparaten und weitaus komplexeren Stoff der griechischen Welt,[6] z.B. *PELLA* für die Prägungen der makedonischen Dynastie der Argeaden, *Corpus Nummorum Thracorum* für Thrakien, oder *Seleucid Coins Online* für die Prägungen der Seleukiden. Des Weiteren haben Arbeitsgruppen bereits angefangen, die notwendigen Konzepte für die keltische und islamische Numismatik sowie für das Mittelalter zu definieren.

Fast 200.000 antike Münzen von insgesamt 39 Institutionen werden heute in auf Nomisma.org basierenden Online-Ressourcen veröffentlicht.[7] Aktiv an der Entwicklungen beteiligt sind u.a. das Institute for Studies of the Ancient World, New York, das Deutsche Archäologische Institut sowie vier der weltweit bedeutendsten Münzsammlungen: Die American Numismatic Society, das British Museum, die Bibliothèque nationale de France und das Münzkabinett der Staatlichen Museen zu Berlin. Auch eine Reihe kleinerer Sammlungen sind dabei ihre Bestände in Portalen wie OCRE und CRRO online zu stellen. Darunter schon acht der 34 deutschen universitären Sammlungen, die Mitglieder im *NUMiD*-Verbund (*Netzwerk universitärer Münzsammlungen in Deutschland*)[8] sind.

Die verlinkte Antike

Nomisma.org und die darauf bauenden Projekte sind in die Linked Open Data-Welt der Antike fest integriert. Mit Nomisma.org verlinkte, bzw. gemappte Münzen werden bei Ressourcen wie *Pleiades*[9] und *Pelagios*[10] angezeigt und umgekehrt können Einträge in den virtuellen Sammlungen mit kontextvermittelnden Daten aus anderen Quellen angereichert werden. Beispielsweise: Wird Literatur in der Form von iDAI.bibliographie / ZENON[11] URIs zitiert, können umgekehrt in Zenon Verweise auf Münzen beim entsprechenden Titel angezeigt werden. Im Abschlussbericht des Work Package 15 des EU-FP7 Projektes ARIADNE wurde Nomisma.org mehrfach als führendes Beispiel für die fachspezifische Anwendung von Linked Open Data zitiert (Geser 2016).

Nomisma.org hat numismatische Daten auf eine Weise digital zur Verfügung gestellt, die vor 10 Jahren kaum vorstellbar gewesen wäre und macht einen bedeutenden Bestand an Material der Forschung leichter zugänglich. Durch die Verlinkung mit anderen LOD-Ressourcen sind diese Daten auch deutlich sichtbarer und stehen damit

für Forschungsvorhaben in anderen Disziplinen vermehrt zur Verfügung. So kann Nomisma.org beispielsweise einen wichtigen Beitrag zur Aufhebung der immer noch weit verbreiteten Isolation der Numismatik in den Altertumswissenschaften leisten, eine Isolation, die in der (bisher oft fehlenden) Zusammenarbeit mit der Archäologie besonders deutlich erkennbar ist.

Fundmünzen und Archäologie

Konzentrierten sich Projekte wie OCRE und PELLA zunächst auf die Präsentation von Münzen aus öffentlichen Sammlungen, richtet sich der Blick mittlerweile zunehmend auf Münzen im archäologischen Kontext. Nomisma.org kompatible Standards für die Aufnahme und Veröffentlichung von Münzen aus Ausgrabungen wurden bei einem Treffen an der University of Oxford im September 2018 vereinbart, um Re-use und Interoperabilität zu ermöglichen.[12]

Nationale Projekte wie *NUMIS*[13] für die Niederlande oder das *Inventar der Fundmünzen der Schweiz*[14] veröffentlichen umfangreiche Bestände an Fundmünzen im Web, die aber nur über das jeweilige nationale Portal zugänglich sind. Eine länderübergreifende Abfrage ist nicht möglich. Eine Ausnahme bietet das Projekt *Antike Fundmünzen in Europa*[15], das die Datenbanken der Römisch-Germanischen Kommission in Frankfurt und des Unternehmens *Finds of Roman Coins in Poland* auf das Vokabular und die Ontologie von Nomisma.org mappt und mittels eines D2R-Servers als RDF online stellt. Ein SPARQL-Endpoint ermöglicht die gemeinsame Abfrage der beiden Datenbestände und erleichtert damit die Analyse der Fundmünzen aus einem von der Nordsee bis zur Ukraine reichenden Raum.

Text zu RDF: Erster Versuch einer Digitalisierung

Bisher lag der Schwerpunkt der Verlinkung mit bzw. Mapping auf Nomisma.org auf in Datenbanken gehaltene Daten. In diesem Beitrag soll abschließend ein Pilotprojekt vorgestellt werden, das sich mit den Möglichkeiten der Anwendung von Nomisma.org auf textbasierte Datenbestände beschäftigt.

Grundlage

Das Projekt „Fundmünzen der römischen Zeit in Deutschland“ (FMRD) brachte 48 Bände mit weit über 300.000 erfassten Fundmünzen heraus. Im Rahmen eines Praxisprojektes im Studiengang „Digitale Methodik in den Geistes- und Kulturwissenschaften“[16] wurde als Beispieldatensatz aus FMRD der Fundkomplex Domgrabung/ Liebfrauen-Areal in Trier mit 1.157 Münzen digitalisiert (M. Radnoti-Alföldi 2006: 119–206).

Das Verfahren

Die Daten zu den Münzen wurden aus dem PDF in CSV und XML extrahiert und in ein RDF-Dokument umgewandelt, um die Vernetzung und Visualisierung zu ermöglichen (Abb. 1).

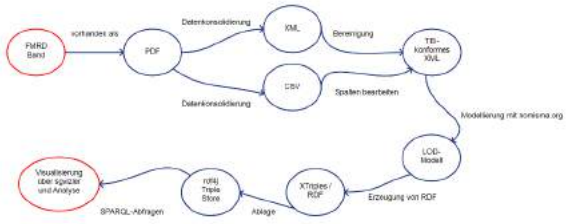


Abbildung 1. Verlaufsplan.

Für die Konvertierung aus dem PDF und die Bearbeitung wurde *Adobe Acrobat Pro DC*[17], der *Oxygen*[18] Editor und *Libre Office*[19] verwendet. Die extrahierten Daten wurden anhand der CSV-Extraktion in ein einheitliches Spaltensystem gebracht. Über das XML-Dokument wurden anhand von regulären Ausdrücken Konvertierungsfehler etc. ausgebessert. Am Ende stand eine einzelne TEI-konforme Tabelle. Der nächste Schritt bestand darin, ein Linked-Open-Data-Modell anhand der Ontologie von Nomisma.org (Abb. 2) zu entwickeln.

```
<rdf:Description rdf:about="https://lod.academy/fmrd/id/*5.">
        <nmo:hasAuthority>Galba</nmo:hasAuthority>
        <nmo:hasDenomination>S</nmo:hasDenomination>
        <nmo:hasDate>68/69</nmo:hasDate>
        <dcterms:date>1. Jh. n. Chr.</dcterms:date>
        <nmo:hasMint>?</nmo:hasMint>
        <nmo:hasReferenceWork>RIC ?</nmo:hasReferenceWork>
        <nmo:hasPeculiarity>NaN-Wert</nmo:hasPeculiarity>
        <dcterms:description>Domgrabung A 5. Vs.: Kopf n.r. m. Lbkr. Fast völlig
            abgegriffen.</dcterms:description>
        <dcterms:bibliographicCitation>339,1</dcterms:bibliographicCitation>
    </rdf:Description>
```

Abbildung 2. RDF-Modell.

Um dieses Modell automatisiert auf alle 1.157 Münzen der Trierer Domgrabung übertragen zu können, wurde der Webdienst *XTriples* verwendet[20], der es erlaubt aus XML ein RDF-Dokument zu erzeugen.

Abbildung 3. Webservice XTriples.

Das RDF-Dokument wurde auf einen Triple Store abgelegt, der über ein RDF4J-Framework verfügt und SPARQL-Abfragen erlaubt. Ein weiterer Webdienst, *sgvizler*,[21] ermöglicht es im

Anschluss die Daten zu visualisieren und zu analysieren (Abb. 4).

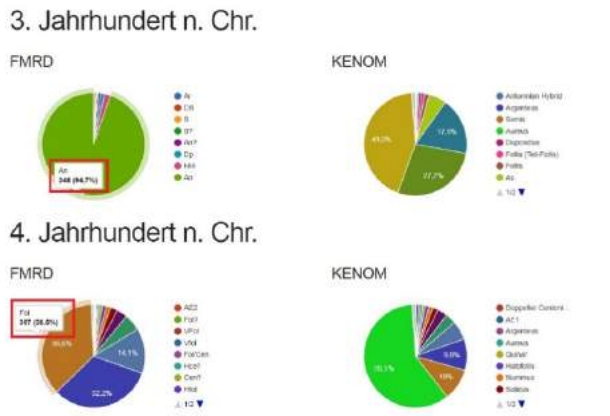


Abbildung 4. Visualisierung über sgvizler zu den Fundmünzen der Trierer Domgrabung.

Ausblick

Das oben beschriebene Verfahren wurde dann erweitert, um eine Pipeline aufzubauen, mit der alle Bände der FMRD-Reihe digitalisiert werden können. Auf manuelle Schritte und proprietäre Software soll dabei möglichst verzichtet werden.

Tabellenextraktion

In einem ersten Schritt wurden aktuelle Open-Source-Angebote verglichen, die es ermöglichen, aus der PDF-Struktur Tabellen in CSV- bzw. XML-Dateien zu extrahieren. Die Daten liegen jedoch so inhomogen vor, dass bisher erschienene Programme und Skripte/Bibliotheken schnell an ihre Grenzen stoßen. Als Beispiel soll hier das browserbasierte Tool *Tabula*[22] dienen, welches Tabellen im PDF automatisch erkennt und anschließend als CSV ausgeben kann. Im Verlauf der Erprobung stellte sich *Tabula* als ungeeignet heraus, da immer noch zu viele manuelle Eingriffe vonnöten sind. So erkennt es manche Tabellenteile nicht (Abb. 5) und die Spalteninhalte im erzeugten CSV werden nicht richtig zugeordnet.

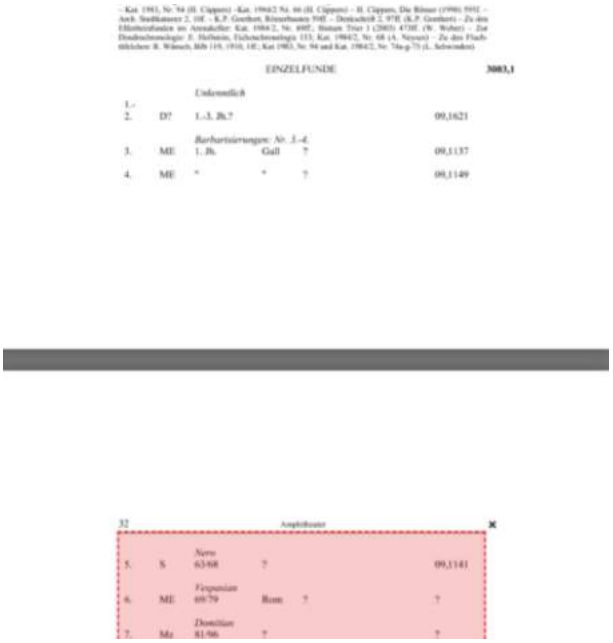


Abbildung 5. Tabula: Fundmünzen Nummer 1 bis 4 werden nicht erkannt.

pdftohtml

Eine andere Herangehensweise erzeugt als Zwischenschritt XML, bevor dieses in CSV umgewandelt werden kann. Für die Erzeugung des XML wird das Tool *pdftohtml* verwendet.[23] Das Ergebnis lässt sich über den *pdf2xml-viewer*[24] überprüfen (Abb. 6).

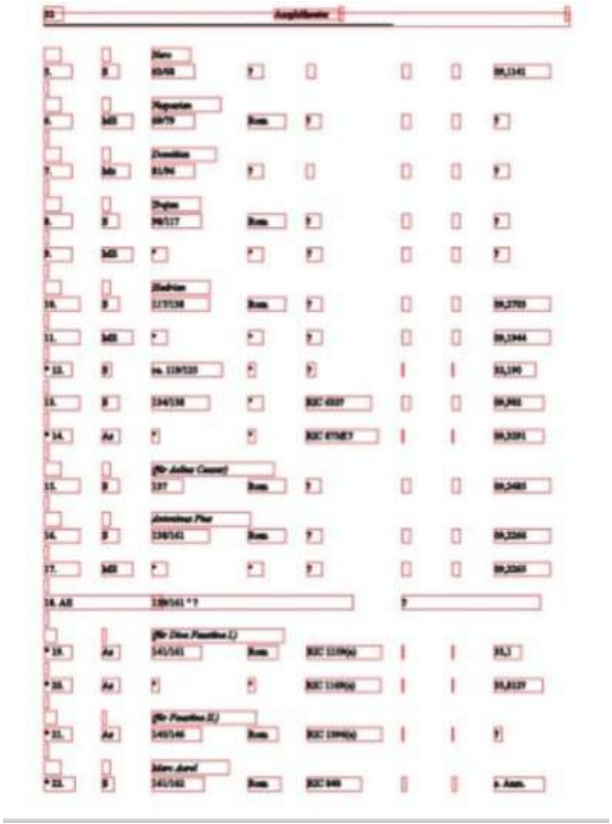


Abbildung 6. OCR-Ergebnis.

Die Normalisierung der Spalten ist hier leichter, da auch leere Tabellenspalten erkannt werden. Jedoch wird auch hier der Spalteninhalt nicht überall korrekt erkannt (Münze 18). Die XML-Struktur erlaubt es über die Tags zu navigieren und diese mit X-Technologien zu manipulieren. Doch ist der Anteil der nicht korrekt erkannten Spalteninhalte immer noch zu hoch, um die Daten wie im Druckband abzubilden.

Reguläre Ausdrücke

Eine andere Möglichkeit an die Tabellendaten zu gelangen sind reguläre Ausdrücke. Mithilfe eines Pythonskriptes wurden diese separiert und als CSV ausgegeben. Dafür wurden beispielhaft verschiedene Münzkomplexe herangezogen und die regulären Ausdrücke um die pro Komplex neu auftretenden Sonderfälle modifiziert. Dieses Verfahren erzeugte gute Ergebnisse für die behandelten Komplexe. So waren nur noch kleinere manuelle Nachbesserungen nötig. Im Vergleich erzeugte dieses Verfahren das beste Ergebnis.

Fazit

Die Digitalisierung von „Fundmünzen der römischen Zeit in Deutschland“ erweist sich als anspruchsvoll. Der Ansatz, die Daten über reguläre Ausdrücke zu gewinnen und zu extrahieren, erscheint bisher am vielversprechendsten. Die dabei auftretenden Abweichungen zwischen den Münzfundkomplexen lassen sich durch auf die FMRD-Struktur zugeschnittene Ausdrücke ausbessern. So erhoffen wir uns, die bisher über Nomisma.org veröffentlichten Datenbestände durch wichtige archäologische Kontexte zu erweitern und zu bereichern.

Fußnoten

1. http://nomisma.org/; Hinweis: Alle im Beitrag erwähnten URLs wurden zuletzt am 14.10.2018 überprüft.
2. z.B. http://nomisma.org/id/sestertius
3. http://nomisma.org/ontology
4. http://numismatics.org/crro/
5. http://numismatics.org/ocre/
6. https://www.greekcoinage.org/portals.html
7. http://nomisma.org/datasets
8. http://www.numid-verbund.de/
9. https://pleiades.stoa.org/
10. http://peripleo.pelagios.org/
11. http://peripleo.pelagios.org/
12. https://www.greekcoinage.org/coins-in-context.html
13. https://www.dnb.nl/en/about-dnb/nationale-numismatische-collectie/numis/numis-database/index.jsp
14. https://www.fundmuenzen.ch/
15. http://afe.fundmuenzen.eu/
16. https://www.digitale-methodik.uni-mainz.de/
17. https://acrobat.adobe.com/de/de/acrobat/acrobat-pro.html
18. https://www.oxygenxml.com/
19. https://de.libreoffice.org/
20. http://xtriples.spatialhumanities.de/index.html
21. http://mgskjaeveland.github.io/sgvizler/
22. https://tabula.technology/
23. https://poppler.freedesktop.org/
24. https://github.com/WZBSocialScienceCenter/pdf2xml-viewer

Bibliographie

G. Geser (2016): *ARIADNE WP15 Study: Towards a Web of Archaeological Linked Open Data* (www.ariadne-infrastructure.eu/.../ARIADNE_archaeological_LOD_study_10-2016.pdf, 31.10.2016).

E. Gruber: *Numishare Blogspot*: http://numishare.blogspot.com/ .

E. Gruber / G. Bransbourg / S. Heath / A. Meadows (2014): *Linking Roman Coins: Current Work at the American Numismatic Society*, in: **G. Earle et al. (eds):** *Archaeology in the Digital Era*: Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, 26-29 March 2012 (Amsterdam) 249–258.

E. Gruber / S. Heath / A. Meadows / D. Pett / D. Wigg-Wolf (2014): *Semantic Web Technologies Applied to Numismatic Collections*, in: **G. Earle et al. (eds):** *Archaeology in the Digital Era*: Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, 26-29 March 2012 (Amsterdam) 264–274.

Radnoti-Alföldi, M. (2006): Die Fundmünzen der römischen Zeit in Deutschland IV 3/2. Stadt und Reg.-Bez. Trier. Die Sog. Römerbauten (Mainz).

K. Tolle / D. Wigg-Wolf (2015): *Uncertainty Handling for Ancient Coinage*, in: **F. Giligny et al. (eds):** *CAA2014. 21st Century Archaeology. Concepts, Methods and Tools*. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology (Oxford) 171–178.

K. Tolle / D. Wigg-Wolf (2016): *How to Move from Relational to 5 Star Linked Open Data – A Numismatic Example*, in: **S. Campana et al. (eds):** *CAA2015. Keep the Revolution Going*. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology (Oxford 2016) 275–281.

D. Wigg-Wolf / F. Duyrat (2017): *La révolution des Linked Open Data en numismatique: Les exemples de nomisma.org et Online Greek Coinage*. Archéologies numériques 1.1, 2017.

Potentielle Privatsphäreverletzungen aufdecken und automatisiert sichtbar machen

Bäumer, Frederik Simon
fbaeumer@mail.upb.de
Universität Paderborn, Deutschland

Buff, Bianca
bbuff@mail.upb.de
Universität Paderborn, Deutschland

Geierhos, Michaela
geierhos@mail.upb.de
Universität Paderborn, Deutschland

Das moderne Web basiert auf Interaktion, Diskussion und Austausch von Informationen. Durch die fortschreitende semantische Anreicherung wird das Web auch zu einer riesigen Informationsquelle für datengesteuerte Anwendungen, wie sie auch in Digital Humanities Verwendung finden. Dies stellt unter Umständen ein Risiko für einzelne Benutzer[1] dar. Da Daten immer effektiver mit bestehenden Ressourcen verknüpft werden, können selbst ungewollt (implizit) offenbarte Einzelinformationen schädliche Folgen für einzelne Nutzer haben. Obwohl *Serviceprovider* im Web die Pflicht und auch das Eigeninteresse haben, die Sicherheit und Privatsphäre von Benutzerdaten zu gewährleisten, gibt es Fälle, in denen Benutzerdaten missbraucht und kompromittiert oder öffentlich verfügbare Informationen gegen dessen ursprünglichen Verfasser verwendet werden (Gross, et al., 2005). Die bestehenden Datenschutzrichtlinien, Betreiberhinweise und (teil-)automatisierte Schutzmechanismen, welche die Privatsphäre von Personen schützen sollen, sind aber oftmals unzureichend. Es ist demnach im Interesse der Kommunizierenden, nur diejenigen Informationen in Textbeiträgen zu platzieren, die einen gewissen selbstbestimmten Grad an Anonymität wahren. Darüber hinaus ist es für alle die, die mit Daten arbeiten wollen (bspw. in der Forschung) im Interesse, private Daten filtern zu können. In diesem Beitrag stellen wir unsere bisherigen Arbeiten an *Text Broom* vor, einem ersten Prototypen, welcher potentielle Privatsphäreverletzungen in Form expliziter als auch inhärenter Angaben in online verfügbaren Fließtexten erkennen sowie sichtbarmachen kann und so eine Hilfe für Benutzer als auch für die gefahrlose Weiterverwendung von Daten darstellen kann.

Privatsphäreverletzungen: Erkennung als Herausforderung

Wie sich unwissentliche Informationspreisgaben in sprachlichen Ausdrücken manifestieren, wurde bisher unzureichend untersucht. Frühere Arbeiten zeigten jedoch auf, dass sprachliche Formulierungen oft mehr Informationen enthalten, als es zunächst den Anschein erweckt. Um diese zu erkennen, wurden vordefinierte Muster verwendet, die nur begrenzt dem Gestaltungsfreiraum natürlicher Sprache gerecht werden und nur offensichtliche (explizite) Informationspreisgabe feststellen (Bäumer, et al., 2017). In diesem Kontext existieren Vorarbeiten, wie die von Sweeney (1996), Dias (2016) sowie von Kleinberg und Mozes (2017). Dabei besonders erwähnenswert ist das Tool NETANOS (*Named Entity-based Text ANonymization for Open Science*) von Kleinberg und Mozes (2017), das benannte Entitäten in Fließtexten erkennen und hervorheben kann. Hierbei handelt es sich stets um benannte Entitäten (z. B. Personennamen), deren wörtliche Nennung zwar eine Gefahr für die Privatsphäre der Betroffenen darstellen kann, deren isolierte Erkennung jedoch trivial im Vergleich zur Behandlung der Ausdruckskomplexität von Privatsphäreverstößen in Fließtexten ist. Denn immer noch fehlt es an Wissen über die genaue sprachliche Manifestierung und an computerlinguistischen Verfahren, die darauf zurückgreifen können. Dies ist allerdings zwingend erforderlich, um entsprechende privatsphäregefährdende Textbestandteile zu identifizieren und mit einer Erläuterung möglicher Risiken zu versehen.

Arztbewertungen als Untersuchungsgegenstand

Service Provider nehmen im Web unterschiedliche Gestalt an, jedoch steht zumeist eine zentrale Dienstleistung im Mittelpunkt, wie es beispielsweise der Erwerb von Produkten bei Online-Shops oder entsprechende Meinungsäußerungen auf Bewertungsportalen sind. Ein medial vielbeachtetes Beispiel in diesem Zusammenhang sind sogenannte *Physician Review Websites* (PRWs), die es den Nutzern ermöglichen, medizinische Dienstleistungen und damit auch die bislang als sensibel geltende Arzt-Patienten-Beziehung anonym zu bewerten. Um eine authentische Bewertung zu erstellen, ergänzen Bewertende vielfach private Informationen, z. B. über Orte, Krankheiten oder Medikamente und sind so potentiell für Dritte identifizierbar (z. B. Ärzte, Freunde). Hier stehen nicht nur explizit genannte Informationen („Ich bin Diabetiker“) im Fokus, sondern auch inhärente Angaben („Ich

bin Vater“ → männlich) sowie Metainformationen (Datum der Bewertung, Alter, Krankenkasse, Ort der Praxis). Oftmals ergibt sich eine Gefahr für die Privatsphäre nicht aus einer einzelnen Information, sondern aus der Summe expliziter und inhärenter Angaben. Das Problem wird verschärft, wenn Patient und Bewertender nicht *in persona* auftreten und somit eine potentielle Privatsphäreverletzung an einer dritten Person (z. B. Kinder, Eltern) vorliegt (Geierhos & Bäumer, 2015). Arztbewertungen eignen sich somit auf Grund der hohen Gefahr unabsichtlich preisgegebener Informationen und auch auf Grund ihrer langjährigen und öffentlichen Zugänglichkeit. So steht uns ein Korpus zur Verfügung, welches ca. 900.000 deutschsprachige Patientenberichte inklusive Metainformationen enthält und die Zeitspanne von 2007 bis 2016 abdeckt (Bäumer et al., 2015).

Annotation privater Angaben in der medizinischen Domäne

Privatsphäreverletzungen können sich vielfältig im nutzergenerierten Texten manifestieren und sind zusätzlich auf Grund stark schwankender Textqualität schwer automatisiert zu erkennen. Deshalb werden neben präzisen, aber gering toleranten linguistischen Mustern auch Methoden des Maschinellen Lernens eingesetzt. Hier bedarf es umfangreicher Annotationen in Trainingsdaten, um private Angaben in unbekannten Texten automatisiert zu erkennen. Um eine große Anzahl an Texten annotieren zu können, nutzen wir das Annotationstool Prodigy, welches insbesondere binäre Annotationsentscheidungen mittels *Active Learning* merklich beschleunigen kann. Die Herausforderungen, die dennoch mit der Annotation einhergehen, werden im Folgenden an dem vermeintlich trivialen Beispiel der Annotation von Krankheiten im oben genannten Korpus aufgezeigt.

Bereits die Frage, was genau annotiert werden soll, ist nicht trivial: Oberbegriffe wie „Krankheit“ und „Symptom“ werden nur ergänzend mit einer Spezifizierung annotiert. Konkreter heißt dies, dass das Wort „Erkrankung“ nicht annotiert wird, jedoch das Kompositum bzw. die Nominalphrase „Herzerkrankung“ und „psychische Erkrankung“ schon. Personenbezeichnungen wie „Magersüchtige“, „Diabetiker“ und „Neurodermitiker“ und Adjektive wie „magersüchtig“ und „laktoseintolerant“ werden annotiert, da sie implizieren, dass eine Person an der jeweiligen Krankheit leidet. Die Annotation von Adjektiven stellt dabei eine Herausforderung dar. In den folgenden Beispielen enthält das attributive Adjektiv einen essentiellen Bestandteil der Semantik der Phrase: „kardiovaskuläre Erkrankung“, „bulimischer Anfall“. Ohne die Attribuierung würde das Substantiv „Erkrankung“ bzw. „Anfall“ nicht als „Krankheit“ annotiert werden. Durch das Adjektiv wird die Phrase jedoch (annähernd) gleichbedeutend mit den Ausdrücken „Herz-Kreislauf-Erkrankung“ und „Bulimie“. Da letztgenannte als „Krankheit“ annotiert werden würden, werden gleichermaßen auch alle Phrasen, die ein unspezifisches Substantiv (z. B. „Erkrankung“) sowie ein attributives Adjektiv (z. B. „kardiovaskuläre“), das den wesentlichen Teil der Semantik trägt, enthalten, als „Krankheit“ markiert. Im Gegensatz dazu werden attributive Adjektive, die eine Krankheit als solche nur näher beschreiben, nicht mit annotiert. Ein Beispiel dafür ist der „wässrige Durchfall“: In diesem Fall wird lediglich das Substantiv als „Krankheit“ annotiert und das Adjektiv nicht beachtet. In Bezug auf das Adjektiv „chronisch“, welches bei Krankheitsbildern oder Symptomen häufig attributiv verwendet wird, gilt, dass dieses Wort nicht mit annotiert wird, da es nicht maßgeblich zu der Semantik der Krankheit beiträgt. In dem Satz „Ich habe chronischen Durchfall und chronische Schmerzen.“ wird folglich nur „Durchfall“ als Krankheit annotiert.

Dies ist nur ein Beispiel für die Komplexität der Thematik und der Herausforderungen bei der Ressourcenerstellung. Ein weiteres Beispiel sind Ambiguitäten, wie sie u. a. bei „Die Ärztin sollte nicht wie meine Mutter sein“ und „Meine Mutter hat auch MC“ vorkommen. Während der erste Satz unkritisch ist, wird im zweiten Satz die Privatsphäre einer dritten Person gefährdet. Die Interpretation des Wortes „Mutter“ gelingt nur im Kontext und zeigt, dass eine rein wortbasierte Vorgehensweise nicht zielführend ist. Aus diesem Grund wird unser Tool Text Broom zwar wie dargestellt auf Basis domänenspezifischer Texte trainiert, nutzt aber eine umfangreiche NLP-Pipeline zur kontextspezifischen Interpretation.

Text Broom

Wie dargestellt, adaptiert unser Prototyp *Text Broom* die Idee von Kleinberg und Mozes (2017), die sich auf benannte Entitäten konzentrieren. Allerdings geht *Text Broom* darüber hinaus, indem es potentielle Privatsphäreverletzungen mit Hilfe einer Textverarbeitungspipeline (*Multi-Stage*-Ansatz) erkennt, die unterschiedliche Granularitätsstufen bietet. Somit verarbeitet die *Text-Broom*-Pipeline ein viel breiteres Spektrum an linguistischen Informationen, aufgeteilt in vier Phasen. Stufe I enthält eine Vorverarbeitung, die grundlegende Sprachverarbeitung wie *Part-Of-Speech* (POS) *Tagging* verwendet. Stufe II kombiniert *Semantic Role Labeling*, linguistische Muster und Eigennamenerkennung. Dies sind nicht-domänenspezifische Komponenten, die eine breite thematische Abdeckung ermöglichen. Im Gegensatz dazu enthält Stufe III eine domänenspezifische Informationsextraktion, eine Komponente zur Phrasenklassifizierung und die finale Bewertungskomponente, die alle bis zu diesem Zeitpunkt gesammelten Informationen zusammenfasst und auswertet. Die letzte Stufe IV enthält Komponenten zur Visualisierung und Erläuterung. Nutzer erhalten somit eine vielschichtige Sicht auf ihre Texte, in denen Privatsphäreverstöße explizit hervorgehoben werden.

Fußnoten

1. Aus Gründen der leichteren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Entsprechende Begriffe gelten im Sinne der Gleichbehandlung für beide Geschlechter.

Bibliographie

Kleinberg, Bennet / Mozes, Maximilian (2017): *Web-based text anonymization with Node.js: Introducing NETANOS*

(Named entity-based Text Anonymization for Open Science). The Journal of Open Source Software.

Dias, Francesco (2016): *Multilingual Automated Text Anonymization*. Inst. Superior Técnico of Lisboa, Lissabon, Portugal.

Bäumer, Frederik S. / Geierhos, Michaela / Schulze, Sabine (2015): *A System for Uncovering Latent Connectivity of Health Care Providers in Online Reviews*. In **Dregvaite, Giedre / Damaševičius, Robertas (Hsg.)**: *Communications in Computer and Information Science*, Band 538. ICIST 2015, Litauen, Oktober 15-16, 2015. Proceedings (S. 3–15). Cham, Schweiz: Springer International.

Bäumer, Frederik / Grote, Nicolai / Kersting, Joschka / Geierhos, Michaela (2017): *Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews*. In **Damaševičius, Robertas & Mikašytė, Vilma (Hsg.)**: *Communications in Computer and Information Science*, Band 756. ICIST 2017, Litauen, 12.-14. Oktober 2017, Proceedings (S. 77–89). Cham, Schweiz: Springer International.

Sweeney, Latanya (1996): *Replacing personally-identifying information in medical records, the Scrub system*. Proceedings of the AMIA annual fall symposium. American Medical Informatics Association. S. 333-337.

Geierhos, Michaela / Bäumer, Frederik S. (2015): *Erfahrungsberichte aus zweiter Hand: Erkenntnisse über die Autorschaft von Arztbewertungen in Online-Portalen*. In: DHd 2015: Book of Abstracts, ZIM-ACDH, Graz, Österreich, 2015, S. 69-72.

Gross, Ralph / Acquisti, Alessandro (2005): *Information Revelation and Privacy in Online Social Networks (The Facebook case)*. In Proceedings of the ACM Workshop on Privacy in the Electronic Society, S. 71-80, Alexandria, USA, 2005.

Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor

Fischer, Frank
frafis@gmail.com
Higher School of Economics, Moskau, Russland

Ingo, Börner
ingo.boerner@univie.ac.at
Universität Wien

Mathias, Göbel
goebel@sub.uni-goettingen.de
WU Wien

Angelika, Hechtl
angelika.hechtl@wu.ac.at
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Christopher, Kittel
contact@christopherkittel.eu
Karl-Franzens-Universität Graz

Carsten, Milling
cmil@hashtable.de
Berlin

Peer, Trilcke
trilcke@uni-potsdam.de
Universität Potsdam

Einleitung

Obwohl sich infrastrukturell einiges getan hat, sieht ein typischer Operationsmodus der digitalen Literaturwissenschaft immer noch so aus, dass eine bestimmte Forschungsmethode auf ein oft nur ephemeres Korpus angewandt wird. Im besten Fall ist das Ergebnis *irgendwie* reproduzierbar, im schlechtesten Fall gar nicht. Im besten Fall gibt es ein offen zugängliches Korpus in einem Standardformat wie TEI, einer anderen Markup-Sprache oder zumindest als txt-Datei. Im schlechtesten Fall ist das Korpus gar nicht zugänglich, d. h., die Forschungsergebnisse müssen einfach hingenommen werden.

Doch seit kurzem gibt es Anzeichen, dass sich dies ändert. Einige Digital-Humanities-Projekte stellen Schnittstellen zu stabilen Korpora zur Verfügung, über die man mannigfaltige Zugriffsmöglichkeiten bekommt und reproduzierbar arbeiten kann. Eines dieser Projekte ist DraCor, eine offene Plattform zur Dramenforschung, die in diesem Vortrag vorgestellt werden soll (zugänglich unter https://dracor.org/ bzw. über die Repos und verschiedene Schnittstellen). DraCor transformiert vorliegende Textsammlungen zu ›Programmable Corpora‹ – ein neuer Begriff, den wir mit diesem Vortrag ins Spiel bringen möchten.

Die Bausteine

Vanillekorpora

Ähnlich wie die COST Action zu europäischen Romanen (Schöch et al. 2018), versucht das DraCor-Projekt als Basis für eine digitale Komparatistik einen Stamm an multilingualen Dramenkorpora aufzubauen, die in basalem TEI kodiert sind. Ein selbst betriebenes russischsprachiges (https://dracor.org/rus) und ein deutschsprachiges Korpus (https://dracor.org/ger) dienen dabei als Einstieg. Diese Korpora sind, ähnlich wie die Sammlung »Théâtre classique« von Paul Fièvre, im weitesten Sinne als Vanillekorpora angelegt, die über das notwendige Markup hinaus zunächst kaum

weitere spezielle Auszeichnungen enthalten, allerdings frei zur Verfügung stehen und damit fork- und erweiterbar sind. Zur Demonstration, dass auch andere, reicher kodierte Korpora dazugebunden und sofort alle bereits bestehenden Extraktions- und Visualisierungsmethoden der Plattform angeboten werden können, wurden das Shakespeare Folger Corpus sowie das schwedische Dramawebben-Korpus geforkt und angedockt (https://dracor.org/shake bzw. https://dracor.org/swe). Dramenkorpora in weiteren Sprachen sollen folgen; einzige Voraussetzung dabei ist jeweils, dass diese in TEI vorliegen.

Die Vorteile von frei auf GitHub gehosteten Korpora liegen auf der Hand. Unabhängig von den letztlich durch die Plattform zur Verfügung gestellten Schnittstellen können die Korpora alternativ durch Klonen oder andere Downloadmethoden, etwa über den SVN-Wrapper von GitHub, direkt bezogen und individuell weiterverarbeitet werden. Ein offen zugängliches GitHub-Repositorium heißt auch, dass Pull Requests zur Fehlerkorrektur und Forks für Erweiterungen möglich und erwünscht sind.

XML-Datenbank (eXist-db) und Frontend

DraCor als Plattform setzt auf die eXist-Datenbank, um die TEI-Dateien zu verarbeiten und Funktionen zur Beforschung der Korpora zur Verfügung zu stellen. Das Frontend wurde mit ReactJS gebaut, ist responsiv und einfach erweiterbar. Der Schwerpunkt liegt aber nicht auf der GUI, sondern auf der API (vgl. generell zur Unterscheidung zwischen beiden Schnittstellenansätzen Bleier/Klug 2018).

API und Entwicklungsumgebung

Um dem Ideal und der Möglichkeit nahe zu kommen, auf einfache Weise »alle Methoden auf alle Texte« anwenden zu können (Frank/Ivanovic 2018), braucht es mehr als offene Korpora. Der zitierte Text von Frank/Ivanovic macht sich hinsichtlich dessen für SPARQL-Endpunkte stark; auch DraCor bietet einen solchen an, besitzt darüber hinaus aber eine reiche API, die über Swagger dokumentiert und erläutert wird (https://dracor.org/documentation/api/). In einem Teilbereich der Korpusphilologie, den Digital Scholarly Editions, hat die Diskussion um eine proaktivere Nutzung von APIs bereits begonnen (zur Vorgeschichte vgl. wiederum Bleier/Klug 2018), als Beispiel hierfür diene die Folger Digital Texts API (https://www.folgerdigitaltexts.org/api), über die man sich spezifische Querys zusammenbauen kann. Der Vorteil einer moderneren Lösung wie Swagger besteht darin, dass API-Querys live und direkt ausgeführt und die Outputs genauer kontrolliert und gesteuert werden können.

Ein einfaches Use-Case-Szenario sieht dann so aus, dass man etwa im RStudio mit zwei, drei Zeilen Code einen Blick in ein Korpus werfen kann, etwa über die zeitliche Entwicklung der Anzahl der Charaktere im russischen Drama zwischen 1740 und 1940, die in der Metadatentabelle festgehalten sind (https://dracor.org/api/corpora/rus/metadata). Diese Datei, beziehbar im JSON- oder CSV-Format, wird in eine Data.Table eingelesen, woraufhin die Werte zweier Spalten (Erscheinungsjahre und Number of Speakers) einfach über ggplot visualisiert werden können (Abb. 1).

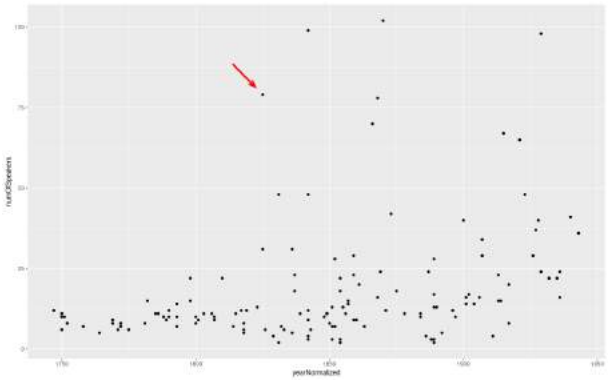


Abbildung 1: Anzahl der Charaktere pro Drama in chronologischer Ordnung (Quelle: RusDraCor).

Anhand dieses sehr simplen Beispiels zeigt sich dann recht deutlich, dass sich mit Puschkins an Shakespeare angelehntem historischen Drama »Boris Godunow« (1825), in dem Sprechakte von 79 Charakteren vorkommen, eine strukturelle Diversifizierung der russischen Dramenlandschaft Bahn bricht.

Die Möglichkeiten beschränken sich aber nicht darauf, vorgefertigte API-Funktionen zu benutzen. Neue Forschungsideen zeitigen immer auch neue Bedarfe an einfach bezieh- und reproduzierbaren Daten und Metriken; die API kann dementsprechend erweitert werden. Dies wird dadurch erleichtert, dass über Apache Ant die gesamte Entwicklungsumgebung auf dem eigenen System nachgebaut werden kann.

Durch bereits implementierte Funktionen können neben Struktur- und Metadaten etwa auch Volltexte ohne Markup bezogen werden (auch Untermengen von Volltexten wie Regieanweisungen), etwa wenn Methoden wie die Stilometrie oder das Topic Modeling der Endzweck sind, also Methoden, die nach dem »bag of words«-Prinzip arbeiten, für das kein Markup vonnöten ist.

Insgesamt wird durch den Aufbau und die Dokumentation offener APIs die bisher oft aufwendige Reproduzierbarkeit von Forschungsergebnissen erheblich erleichtert.

Shiny App

Ein Beispiel für die vielseitigen Nutzungsmöglichkeiten der DraCor-API ist die Shiny App, die Ivan Pozdniakov aufgesetzt hat (https://shiny.dracor.org/). Shiny ist ein auf R basierendes Framework, das es ermöglicht, interaktive Visualisierungen im Browser darzustellen. Die DraCor-Shiny-App tut genau dies und setzt dabei vollkommen auf die DraCor-API für den Datenbezug. So kann zu Lehr- und Forschungszwecken, aber auch zur einfacheren Datenkorrektur, auf Visualisierungen des aktuellen Datenbestandes zugegriffen werden.

Didaxe

Das Markup oder andere Formalisierungen literarischer Texte sind nicht selbsterklärend. Zwar gibt es einige Standards, aber die jeweilige Operationalisierungslösung hängt von der Forschungsfrage ab. Allein das Extrahieren

von Figurennetzwerkdaten ist auf viele Arten und Weisen möglich, was dazu führt, dass etwa alle von verschiedenen Forschungsgruppen extrahierten Netzwerke aus Shakespeares »Hamlet« zu leicht verschiedenen Ergebnissen kommen. Selbst für Dramen ist dies also schon ein nicht-trivialer Akt, von Romanen dann ganz zu schweigen (beispielhaft seien Grayson et al. 2016 genannt, die verschiedene Extraktionsmethoden für Romane durchtesten und die Ergebnisse vergleichen). Um diese Erkenntnis schon in der Lehre zu fördern, wurde das Tool »Easy Linavis« (https://ezlinavis.dracor.org/) entwickelt und in die DraCor-Toolchain integriert. Per Hand können Netzwerkdaten aus Texten extrahiert und dabei das Bewusstsein für die Kontingenz dieses Vorgangs geschärft werden, eine wichtige Vorstufe zur Operationalisierung.

Neben einem Ansatz zur Gamifizierung des TEI-Korrekturvorgangs (Göbel/Meiners 2016) haben wir für Lehrzwecke auch ein Dramenquartett entwickelt, um spielerisch das Verständnis von Netzwerkwerten zu trainieren (Fischer at al. 2018).

Die aufgezählten, um die Plattform herumgruppierten didaktischen Mittel sind integraler Bestandteil des ganzen Projekts, da sie auf dessen Daten und Operationalisierungen aufsetzen. Wichtig dabei war die Erkenntnis, dass Daten mehrere Gestalten annehmen und für Forschung und Lehre gleichermaßen von Bedeutung sein können.

Linked Open Data (LOD)

Im TEI-Code sind PND- bzw. Wikidata-Identifier sowohl für Autor*innen als auch für die Werke hinterlegt. Auf diese Weise lassen sich verschiedene Realien, die außerhalb der eigenen Korpusarbeit liegen, hinzufügen. Eine automatisch erstellte Autor*innengalerie hat dabei noch eher illustrativen Charakter (de la Iglesia/Fischer 2016).

Darüber hinaus kann man aber zum Beispiel feststellen, ob es nicht einen unbewussten regionalen Bias im Korpus gibt. Dafür lässt man sich über die Wikidata-Identifier die Verteilung der Geburts- und Sterbeorte der Autor*innen auf einer Karte anzeigen. So konnte dann für das deutschsprachige Korpus GerDraCor ausgeschlossen werden, dass es einen solchen Bias gibt, da sich die Orte relativ gleichmäßig über die (historisch) deutschsprachigen Gebiete verteilen (Göbel/Fischer 2015).

Ebenso lässt sich über die Wikidata-ID der Stücke herausfinden, wo diese uraufgeführt worden sind (Beispiel-Query: http://tinyurl.com/y9vga68j), d. h., Aspekte der Aufführungsgeschichte lassen sich zuschalten, obwohl diese gar nicht im Fokus des Kernprojekts liegen. Programmable Corpora verbinden sich also auch mit der Welt um sie herum, was sie u. a. von den nach innen gerichteten Workbenches der Korpuslinguistik unterscheidet.

Infrastruktur statt Rapid Prototyping

Projekte wie DraCor versuchen nichts anderes als den digitalen Literaturwissenschaften eine verlässliche und ausbaufähige Infrastruktur zu geben, damit sie sich stärker auf eigentliche Forschungsfragen konzentrieren und reproduzierbare Ergebnisse hervorbringen können.

Eine wichtige Folgerung für uns war, dass wir die Weiterentwicklung unserer seit vier Jahren entwickelten all-in-one Python-Skriptsammlung *dramavis* aufgeben und uns lieber der Arbeit an der API widmen. *Dramavis* (Kittel/Fischer 2014–2018 sowie Fischer et al. 2017) folgte dem in den Digital Humanities nicht untypischen Rapid Prototyping mit direkter Verarbeitung literarischer XML-Daten (Trilcke/Fischer 2018) und einer mittlerweile stark gewachsenen Codebasis, die alles auf einmal kann, deren Maintenance aber immer schwieriger geworden ist und oft genug von den eigentlichen Forschungsfragen weggeführt hat.

Fazit

In Anlehnung an das Projekt »ProgrammableWeb« – das eine Datenbank von offenen APIs unterhält und dessen Slogan lautet: »APIs, Mashups and the Web as Platform« (zugänglich unter https://www.programmableweb.com/) – schlagen wir für infrastrukturell-forschungsorientierte, offene, erweiterbare und LOD-freundliche Korpora den Begriff ›Programmable Corpora‹ vor.

Programmable Corpora erleichtern es, Forschungsfragen auf viele Arten und Weisen um Korpora herum programmieren zu können. Es steht zu erwarten, dass sich infrastrukturelle Anstrengungen dieser Art für die gesamte Community auszahlen mit Effekten, wie sie John Womersley in seiner Präsentation auf der ICRI2018 in Wien aufgezählt hat: a) dramatically increase scientific reach; b) address research questions of long duration requiring pooled effort; c) promote collaboration, interdisciplinarity, interaction.

Der Anschlussmöglichkeiten sind viele, egal ob man gar nicht programmieren möchte, sondern nur eine GEXF-Datei für Gephi benötigt, ob ein Korpus über seine Verbindungen zur Linked Open Data Cloud beforscht oder einfach aus R oder Python heraus bestimmte Daten bezogen werden sollen, ohne dass man sich mit dem Korpus und dessen Maintenance und Reproduzierbarkeit selbst kümmern muss (all dies bleibt natürlich aber eine Option). Programmable Corpora erleichtern die Entscheidung, auf welcher Ebene der eigene Forschungsprozess einsetzt.

Bibliographie

Bleier, Roman / Klug, Helmut W. (2018): *Discussing Interfaces in Digital Scholarly Editing*. In: Digital Scholarly Editions as Interfaces. BoD, Norderstedt, S. V–XV. URL: https://kups.ub.uni-koeln.de/9094/

de la Iglesia, Martin / Fischer, Frank (2016): *The Facebook of German Playwrights*. URL: https://dlina.github.io/The-Facebook-of-German-Playwrights/

Fischer, Frank / Dazord, Gilles / Göbel, Mathias / Kittel, Christopher / Trilcke, Peer (2017): *Le drame comme réseau de relations. Une application de l'analyse automatisée pour l'histoire littéraire du théâtre*. In: Revue d'historiographie du théâtre. № 4. URL: https://hal.archives-ouvertes.fr/hal-01811799

Fischer, Frank / Kittel, Christopher / Milling, Carsten / Schultz, Anika / Trilcke, Peer / Wolf, Jana (2018): *Dramenquartett – Eine didaktische Intervention.* In: Konferenzabstracts zur DHd2018, Universität zu Köln. S. 397 f. DOI: https://doi.org/10.6084/m9.figshare.5926363.v1

Göbel, Mathias / Fischer, Frank (2015): *The Birth and Death of German Playwrights*. URL: https://dlina.github.io/The-Birth-and-Death-of-German-Playwrights/

Göbel, Mathias / Meiners, Hanna-Lena (2016): *Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus*. In: Konferenzabstracts zur DHd2016, Bern/CH. S. 140–143. URL: http://www.dhd2016.de/abstracts/vortr%C3%A4ge-007.html

Grayson, Siobhán / Wade, Karen / Meaney, Gerardine / Greene, Derek (2016): *The Sense and Sensibility of Different Sliding Windows in Constructing Co-Occurrence Networks from Literature*. In: 2nd IFIP International Workshop on Computational History and Data-Driven Humanities. Trinity College Dublin 2016. PDF: http://derekgreene.com/papers/grayson16chdh.pdf

Kittel, Christopher / Fischer, Frank (2014–2018): *dramavis. Python-Skriptsammlung*. Repo: https://github.com/lehkost/dramavis

Schöch, Christoph et al. (2018): *Distant Reading for European Literary History. A COST Action [Poster]*. In: DH2018: Book of Abstracts / Libro de resúmenes. Mexico: Red de Humanidades Digitales A. C. URL:

Trilcke, Peer / Fischer, Frank (2018): *Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen*. In: Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden. Hrsg. von Martin Huber und Sybille Krämer (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3). DOI: http://dx.doi.org/10.17175/sb003_003

Prototypen als Proto-Theorie? – Plädoyer einer digitalen Theoriebildung

Kleymann, Rabea

r.kleymann90@gmail.com
Universität Hamburg, Deutschland

In der Diskussion um den Status der Digital Humanities (DH) als wissenschaftliche Disziplin wird häufig die Frage einer Theoriebildung ins Feld geführt. Konfrontiert mit dem Vorwurf der Theorielosigkeit (vgl. Bauer 2011) sowie der Forderung nach „mehr Theorie“ (Lauer 2013, 112) lässt sich eine ausschweifende Theoriedebatte beobachten (vgl. Gold 2012). Zum einen werden die DH aufgrund der Heterogenität ihrer Gegenstände sowie ihres Aufgabenfeldes jenseits der Theoriebildung verortet (vgl. Winko/Köppe 2013, 328).[1] Zum anderen wird die Theoriebildung in eine zeitliche Relation gebracht (vgl. Hall 2012). So wird mit der Proklamation eines ‚Endes der Theorie‘ eine posttheoretische Ära eingeläutet (vgl. Scheinfeldt 2012), zugleich bekräftigt sich das Verständnis eines vortheoretischen Status der DH (vgl. Boellstorff 2014, 105). Was nun ‚vor‘ oder ‚nach‘ der Theorie liegt bestimmen beide Positionen auf ähnliche Weise, indem sie ihren Fokus auf die Aktivitäten (vgl. Flanders/Jannidis 2015, 3) oder „doing research“ (vgl. Kath et al. 2015, 31; Cecire 2011) richten.

Im Zuge dieser Hinwendung zur digitalen Praxis gewinnt die Entwicklung von prototypischer Software zunehmend an Bedeutung (vgl. Ramsey/Rockwell 2012). Übernommen aus der informatischen Softwareentwicklung wird unter einem Softwareprototyp eine „provisorische“ und „experimentelle“ (Ruecker et al.) Software verstanden, die eine zukünftige Systemkomplexität (Houde/Hill 1997; Mogensen 1992) auf bestimmte Teilkomponenten reduziert und diese unter Annahme eines deskriptiven oder prospektiven Potenzials ausstellt. Softwareprototypen werden als methodisch und theoretisch modellierte „digitale Artefakte“ (Rockwell/Ramsey 2012) verstanden, die neben den bereits in der Informatik etablierten Funktionen der Exploration, Erklärung und Verständigung zunehmend auch in anderen diskursiven Kontexten als Argument (Ruecker/Galey 2010), Provokation (Boer/Donovan 2012), Spekulation (Hinrichs/Forlini 2017), Vermittler von Performanz und Kritik (Drucker 2012) begriffen werden. Ruecker/Galey (2010, 1) sprechen von „Prototypen als Theorie“.

Softwareprototypen scheinen, so nun der Ausgangspunkt des Vortrages, nicht nur an epistemischen Status zu gewinnen, sondern auch den Bedarf der noch ausstehenden oder bereits obsoleten Theoriebildung zu adressieren. Der vorliegende Vortrag möchte darlegen, inwiefern Softwareprototypen als ‚Proto-Theorie‘ der DH begriffen werden können. Dazu stellt der Vortrag nicht nur die theoriebildenden Effekte eines Softwareprototyps vor, sondern argumentiert, dass der Softwareprototyp eine spezifische Form der Theoriegestaltung darstellt. Unter ‚Proto-Theorie‘ wird eine durch die Form des Softwareprototyps gestaltete Theoriearbeit begriffen. Der Vortrag versteht sich als Beitrag zu den Versuchen, die „Geisteswissenschaften als Ort avancierter Theoriebildung“ (Grizelj/Jahraus 2011, 9) neu zu vermessen sowie Spielräume und Grenzen einer auf Softwareprototypen aufbauenden Theoriebildung auszuloten. Zunächst wird ein geisteswissenschaftliches Verständnis der Konzepte *Prototyp* und *Theorie* vorgestellt. Anschließend wird die Theoriemündigkeit des Softwareprototyps eruiert.

Zur Anschauung von Identität und Differenz – Das Konzept des Prototyps

Während die Rede von einem Softwareprototypen im DH-Forschungsdiskurs etabliert ist, steht eine Auseinandersetzung mit dem Konzept des Prototyps über ein informatisches Verständnis hinaus noch aus.[2] Aus dem Griechischen von *prototypon* ‚Urbild‘ abgeleitet, setzt sich das Kompositum aus „ *proto* ‚vorderster, erster, bedeutsamster“ (Kluge 2011, 728) und „ *typos* Schlag, Stoß (von gr. typtein ‚schlagen, prägen) ‚prägende Form‘, Umriss, Gestalt, Muster nach“ (Ritter 1998, 1587) zusammen. Im taxonomischen Referenzrahmen der Klassifizierung und Kategorisierung verortet, dient der Prototyp im Allgemeinen als Beschreibungsbegriff der Validierung von „Gleichheit und Verschiedenheit“ (Mainberger 2003, 42). Es können zwei Bestimmungen des Prototyps unterschieden werden: Unter der kognitionspsychologischen Perspektive (vgl. Rosch 1975) wird der Terminus des Prototyps in der Standardverwendung als abstrakte Entität eingeführt, welche die typischen Eigenschaften einer extensionalen Kategorie als bestes Exemplar repräsentiert (vgl. Kleiber 1993, 117).[2] Die Zugehörigkeit zu einer Kategorie wird in Bezug auf den Prototyp qua analytischer Quantifizierung von Merkmalen ermittelt. Notwendige Bedingung der Zugehörigkeit zu einer Kategorie ist die Übereinstimmung von mindestens einem Merkmal.

Davon zu unterscheiden ist eine morphologische Bestimmung des Prototyps, wie sie unter anderen Ludwig

Wittgensteins Familienähnlichkeit darstellt.[4] Am Beispiel des Begriffs „Spiel" führt Wittgenstein aus, dass nicht alle „Glieder einer Familie" (Wittgenstein 1945, 278) durch ein einzelnes gemeinsam definierendes Merkmal vereint sind, sondern durch „ein kompliziertes Netz von Ähnlichkeiten, die einander übergreifen und kreuzen" (278) verknüpft sind. Die Zugehörigkeit zu einer Kategorie stellt sich über Ähnlichkeitsbeziehungen her.

Gestaltung von Zugängen – Theorien in der Literaturwissenschaft

Der Begriff der Theorie „griechisch: *theoria* geistiges Anschauen bzw. wissenschaftliches Betrachten" bezeichnet „explizite, elaborierte, geordnete und logisch konsistente Kategoriensysteme, die der Beschreibung, Erforschung und Erklärung der Sachverhalte ihres jeweiligen Objektbereichs dienen" (Nünning/Nünning 2010, 6). So adressiert die Theorie eine Mangelerfahrung, die sich in der Differenz zwischen Theorie und Welt expliziert: „Theorien sind ein Effekt der Unzugänglichkeit von Welt dergestalt, dass sie die Zugänglichkeit als Theorie substituieren und somit kompensieren" (Jahraus 2011, 28). Bezogen auf die Literaturwissenschaft konstituieren Theorien nicht nur einen spezifischen Zugang zum Gegenstand Literatur, indem sie „Bedingungen der Produktion und Rezeption […] sowie […] Beschaffenheit und […] Funktionen" reflektieren (Winko/Köppe 2013, 7), sondern sie erzeugen überhaupt erst den Gegenstand als epistemisches Objekt ihrer Betrachtung.[5]

Darüber hinaus stellen Theorien eine spezifische Textgattung mit eigener Literarizität dar (vgl. Culler 2011, 3). Da Theorien vorrangig in Form von schriftlich verfassten Texten erscheinen und daher einer sprachlichen Logik folgen, unterliegen sie einer spezifischen Textualität, die konstitutiv, aber nicht reduzibel für die Gestaltung eines theoretischen Gehalts ist (vgl. Saar 2013, 47). Zu den „Minimalbedingungen" (Winko/Köppe 2013, 8) einer Literaturtheorie gehören dabei erstens, der Abstraktionsgrad der theoretischen Aussage, der „auf mehrere, eventuell sogar alle Einzelphänomene eines bestimmten Typs zutr[ifft] bzw. ein Modell der Phänomene bereitzustell[t]" (Winko/Köppe 2013, 8). Zweitens können Literaturtheorien resümierende, erklärende, koordinierende und prognostische Funktionen zugesprochen werden. Bezogen auf die Gestaltung einer Literaturtheorie ist drittens der systematische und formale Aufbau kennzeichnend. In Folge der Systematisierung kann viertens die Explizitheit des Begriffsapparats angeführt werden (vgl. Winko/Köppe 2013, 8).[6] In der literaturwissenschaftlichen Praxis ermöglicht der Anschluss an einer Theorie eine intersubjektive Nachvollziehbarkeit (vgl. Nünning/Nünning 2010, 2). So dienen Theorien dazu, Evidenz und Plausibilität unter Annahme eines Wahrheit- und Wissenschaftlichkeitsanspruches herzustellen und zu sichern. Mit Blick auf eine „interpretative community" (Bode 2011, 80) und den allgemeinen wissenschaftlichen Diskurs schaffen Theorien einen spezifischen Kommunikationsrahmen.

Zur Theoriemündigkeit des Softwareprototyps in den DH

Vor dem Hintergrund der skizzierten Prototypenkonzepte sowie den Anforderungen an (literaturwissenschaftliche) Theorien stellt sich nun die Frage, inwiefern Softwareprototypen diesem Anspruch gerecht werden können. Zunächst geht der Entwicklung einer prototypischen Software ein Prozess der Datenmodellierung und der Herstellung einer Ontologie (vgl. Jannidis/Flanders 2015, 7) voraus. In den Klassifikationsverfahren zur Gewinnung von Klassen, Eigenschaften und Relationen (vgl. [Gruber 2009] zit. Jannidis/Flanders 2015, 9) sedimentiert sich bereits ein theoretisches Wissen, wie zum Beispiel die Vorstellung vom Text oder von Bedeutung. Über diese theoretische präfigurierende Modellierung hinaus expliziert sich im Softwareprototyp jedoch die für Theorie konstitutive Differenzerfahrung von Beobachtung und Beobachteten. Der Softwareprototyp stellt einen Zugang dar, der interaktiv vermittelt wird. Als reduziertes und komprimiertes Modell[7] eines Sets von meist methodischen Operationen erfüllt der Softwareprototyp oftmals die „synthetische und integrierende Leistung einer […] Schau" (Wirth 2013, 138). Einzelne Sachverhalte bzw. Aktivitäten werden verknüpft, die sodann prototypisch, das heißt prägend und musterhaft, als Teil eines größeren kohärenten Erklärungszusammenhangs erscheinen. Genauer gesagt, repräsentiert der Softwareprototyp die Theorie als funktionales Gefüge und übernimmt mithin eine exemplarische Funktion.[8] Er steht „nicht bloß stellvertretend und uneigentlich für Anderes, sondern prototypisch und materialiter für etwas ein, von dem er selbst ein nur zu demonstrativen [oder explikativen, RK] Zwecken abgetrennter Teil zu sein [verspricht]" (Schaub 2011, 12). So finden sich zum Beispiel in den Abstracts der DHd Konferenz 2018 nicht nur zahlreiche Vorstellungen von Softwareprototypen,[9] sondern auch Formulierungen zur explikativen und prognostischen Leistung sowie zur Funktion der Validierung.[10] Oszillierend zwischen der Vorstellung vom Softwareprototyp als eine Figuration von Evidenz[11] und als verzeitlichte und verräumlichte Repräsentationsform der Theorie, die aufgrund ihres provisorischen Status zum spekulativen Spiel (vgl. Hacking 1983, 352) einlädt, zeigt sich die prekäre epistemische Lage. Denn der Softwareprototyp löst die scharfe „Distinktion zwischen der Welt der Dinge und der Welt der Zeichen" (Rautzenberg/Strätling 2013, 11) auf. Vielmehr gehen die Grenzen fließend ineinander über, sodass zugleich die „differenztheoretische Grundlegung von Theorie" (Jahraus 2011, 36) zugunsten einer interaktiven Erfahrung infrage gestellt wird.[12]

Neben den Nachweis der theoriegenerierenden Effekte stellt sich die Frage nach der Form, in der das theoretische Wissen im Softwareprototypen erscheint. Die visuelle Ordnung des Softwareprototyps in Form einer grafischen Benutzeroberfläche ist nicht der textuellen Ästhetik des Linearen verpflichtet. An die Stelle der (textuellen) Systematizität tritt eine topologische und temporäre Entfaltung der Theorie. Hinsichtlich des Kriteriums der Explizitheit kann eine epistemische Verschiebung zur Präsenz hin angenommen werden. Im Sinne eines „Präsenzphänomen[s]" (Gumbrecht 2012, 216)[13] scheint der Softwareprototyp nicht nur theoretisches Wissen zu repräsentieren, sondern die Visualität und Interaktion generiert einen Überschuss, in dem sich neues Wissen zeitigt. Dabei überlagert sich empirisch gesichertes und hypothetisch spekulatives Wissen.

Begreift man nun die Pluralität von Softwareprototypen in der DH-Forschung aus einer übergeordneten Perspektive können die einzelnen Softwareprototypen im Sinne Wittgensteins als ‚Glieder einer Familie' betrachtet werden. So entsteht kein systematisches Theoriegebilde, sondern vielmehr die Dynamik eines offenen und unabgeschlossenen Netzes.

Schlussfolgerungen

Die Diskussion um den epistemischen Status des Softwareprototyps hinterfragt nicht nur das

Selbstverständnis der DH als (Hilfs)Wissenschaft und als Forschungsinfrastruktur[14], sondern verortet die Arbeit an der Theoriebildung wieder in den Geisteswissenschaften. Wie sieht eine zeitgenössische Literaturtheorie aus, welche die einzelnen Sachverhalte in den Zusammenhang der Digitalität stellt? Die ‚Proto-Theorie' erweist sich als eine mögliche Antwort. Aufgaben der DH-Community sind daher, die Legitimation von unterschiedlichen Theorieformen zu erproben, die Differenzen und Barrieren zwischen textuellen und technischen Gestaltungsformen zu diskutieren und so das hybride Theorienetz zu verdichten.

Fußnoten

1. Vgl. Winko/Köppe (2013, 328): „Der Überblick über ihre heterogenen Arbeitsfelder macht bereits deutlich, dass die Computerphilologie nicht an der Entwicklung einer umfassenden Literaturtheorie interessiert ist, sondern andere Ziele verfolgt." Die eigentliche Theoriearbeit wird dabei verstanden als außerhalb der DH, zum Beispiel in den Medienwissenschaften oder Software Studies (vgl. Jannidis et. al 2017, 19) sowie der Mathematik und Informatik (Flanders/Jannidis 2015, 2).
2. Eine Ausnahme stellt der DHd-Beitrag von Henny-Krahmer et. al. (2018): *Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane* dar. Hier wird das Protypenmodell als Gattungskonzept diskutiert.
3. Auf das problematische Verhältnis von Begriff und Prototyp kann an dieser Stelle nur hingewiesen werden. Weitere Ausführungen finden sich bei Kleiber (1993) sowie Gansel (2017).
4. Zu nennen sind auch die Verfahrensweise der vergleichenden Naturlehre, wie sie von Buffon, Herder und Goethe präsentiert werden.
5. Signifikant für die Praxis der Literaturwissenschaft ist dabei vor allem das Nebeneinander von unterschiedlichen theoretischen Ansätzen (Wellbery 1985, 7; Jahrhaus/ Neuhaus 1995).
6. Methoden unterscheiden sich von Theorien dadurch, dass sie eher die Art und Weise eines Vorgehens zielführend beschreiben. Theorien können folglich eine sowie ein Set von allgemeinen (Lesen, Hypothesen generieren) sowie spezifischen (deduktiven, dialektischen) Methoden bereitstellen. Exemplarisch für die hermeneutische Theorie ist der hermeneutische Zirkel oder das systemtheoretische Re-entry für die Systemtheorie.
7. Stachowiak (1992, 219) definiert das Modell als „pragmatische Entität eines (mindestens) fünfstelligen Prädikats […]: X ist Modell des Originals Y für den Verwender k in der Zeitspanne t bezüglich der Intention Z." Allgemeine Ausführungen zum Modell vgl. auch Epple (2016).
8. Entgegen der These von Rockwell/Ramsey (2012) zur „Thing Theory" geht es nicht um die konkrete Materialisation der Theorie als eines objektiven Gegenstandes, der die textuelle Form der Theorie substituiert.
9. Vgl. DHd 2018, Köln Abstracts: Es finden sich 17 Erwähnung von Softwareprototypen, verstanden als prototypische Software, auf den Seiten 32, 53, 65, 98, 154, 164,175, 194-195, 205, 220, 273, 282, 388, 422, 435, 459, 466-468.
10. Beispiele der Erwähnungen von Softwareprototypen mit Hinweis auf den Leistungsumfang: „zur Diskussion gestellt" (S. 32); „Anforderungen und Interessen aus der Fachcommunity zur Weiterentwicklung" (S. 53); „Prozess zu beschleunigen bzw. zuverlässiger zu machen" (S. 65); „iterativ […] gewonnene[] Erkenntnisse" (S. 195); „die einzelnen Funktionalitäten und Konfigurationsmöglichkeiten ausgetestet und verbessert" (S. 205), „Interaktions-Prototypen mit Nutzern zu testen, um belastbarere Aussagen über seine Validität zu erhalten." (S. 282) etc.
11. Eine Evidenzgenerierung erfolgt bei literaturwissenschaftlichen Softwareprototypen zum Beispiel durch Visualisierung oder Text-Mining. Softwareprototypen erscheinen vor diesem Hintergrund als dezidierte Beweisverfahren der DH, die evidentielle Gewissheit und Präsenz vermitteln.
12. Im Unterschied zur textuellen Theorieform wird das Subjekt, das sich durch die Theorie einen Zugang verschaffen will, scheinbar schon durch die interaktive Ausrichtung des Softwareprototyps ‚mit-konstituiert'. Hier schließt sich die Frage nach der Subjektkonstituierung in textuellen und technischen Theorieformen an.
13. Gumbrecht (2012) in Rekurs auf Jean-Luc Nancy einen Begriff der „Präsenz, der sich auf unser räumliches Verhältnis zu den Dingen der Welt" aufbaut.
14. Unter den Aspekt der Forschungsinfrastruktur fallen Erwägungen über interdisziplinäre Teamstrukturen, Projektkonstellationen sowie Forschungsförderungen und -finanzierungen. Zur Diskussion steht unter anderen die Frage, warum Ressourcen für die Herstellung von Softwareprototypen eingesetzt werden.

Bibliographie

Bauer, Jean (2011): *"Who are you Calling Untheoretical?"* in: Journal of Digital Humanities 1,1. http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/ [letzter Zugriff am 29.08.2018].

Bode Christoph (2011): *"Theorietheorie als Praxis. Überlegungen zu einer Figur der Unhintergehbarkeit oder: Über eine Theorie-Praxis-Asymmetrie"*. in: **Grizelj, Mario / Jahraus, Oliver (eds.)**: *Theorietheorie. Wider die Theoriemündigkeit in den Geisteswissenschaften*. München: Fink, 79-94.

Boellstorff, Tom (2014): *"Die Konstruktion von Big Data in der Theorie"* in: **Reichert, Ramón (ed.)**: *Big data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie.* Bielefeld: transcript 105-131. http://dx.doi.org/10.14361/transcript.9783839425923 [letzter Zugriff am 24.08.2018].

Boer, Laurens / Donovan Jared (2012): *"Provotypes for Participatory Innovation"* in: *DIS 2012*, 388-397. https://www.researchgate.net/publication/254462007_Provotypes_for_participatory_innovation [letzter Zugriff am 29.08.2018].

Cecire, Natalia (2011): *"Introduction: Theory and the Virtues of Digital Humanites"* in: *Journal of Digital Humanities* 1,1. http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/ [letzter Zugriff am 29.08.2018].

Culler, Jonathan D. (2011): *Literary theory. A very short introduction*. Oxford: Oxford Univ. Press.

Drucker, Johanna (2012): *"Humanistic Theory and Digital Scholarship"* in: **Gold, Matthew (ed.)**: *Debates in Digital*

Humanities. Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/ [letzter Zugriff am 28.08.2018].

Epple, Moritz (2016): *"Analogien, Interpretationen, Bilder, Systeme und Modelle: Bemerkungen zur Geschichte abstrakter Repräsentationen in den Naturwissenschaften seit dem 19. Jahrhundert"* in: **Axer, Eva / Geulen, Eva / Heimes, Alexandra (eds.)**: *Forum Interdisziplinäre Begriffsgeschichte* 5, 1, 11-31. http://www.zfl-berlin.org/publikationen-detail/items/fib-5-jg-2016-1.html.html [letzter Zugriff am 29.08.2018].

Galey, Alan / Ruecker, Stan (2010): *"How a Prototype Argues"* in: *Literary and Linguistic Computing* 25,3. doi:10.1093/llc/fqq021 [letzter Zugriff am 28.08.2018].

Gansel, Christina (2017): *"Prototypensemantik und Stereotypensemantik"* in: **Staffeldt, Sven / Hagemann, Jörg (eds.)**: *Semantiktheorien: Lexikalische Analysen Im Vergleich.* Tübingen: Stauffenburg, 77-96.

Gold, Matthew (ed.) (2012): *Debates in Digital Humanities.* Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/ [letzter Zugriff am 28.08.2018].

Grizelj, Mario / Jahraus, Oliver (eds.) (2011): *Theorietheorie. Wider die Theoriemündigkeit in den Geisteswissenschaften*. München: Fink.

Gumbrecht, Hans Ulrich (2012): *Präsenz.* Berlin: Suhrkamp.

Hacking, Ian ([1983] 1996): *Einführung in die Philosophie der Naturwissenschaften.* Stuttgart: Reclam.

Hall, Gary (2012): *"Has Critical Theory Run Out of Time for Data-Driven Scholarship"* in: **Gold, Matthew (ed.)**: *Debates in Digital Humanities.* Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/ [letzter Zugriff am 28.08.2018].

Hinrichs, Uta / Forlini, Stefania (2017): *"In Defense of Sandcastles: Research Thinking trough Visualization in DH"* in: *Digital Humanities Conference 2017* https://dh2017.adho.org/abstracts/133/133.pdf [letzter Zugriff am 28.08.2018].

Houde, Stephanie / Hill, Charles (1997): *"What do prototypes prototype?"* in: **Helander M. / Landauer, T.K. / Prabhu P. (eds.)**: *Handbook of human-computer interaction*. Amsterdam: Elsevier, 367-381.

Jahraus, Oliver / Neuhaus, Stefan (eds.) (2002): *Kafkas "Urteil" und die Literaturtheorie. Zehn Modellanalysen.* Stuttgart: Reclam.

Jahraus, Oliver (2011): *"Theorietheorie"* in: **Grizelj, Mario / Jahraus, Oliver (eds.)**: *Theorietheorie. Wider die Theoriemüdigkeit in den Geisteswissenschaften*. München: Fink 17-39.

Jannidis, Fotis / Flanders, Julia (eds.). (2015): *Knowledge Organization and Data Modeling in the Humanities.* https://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf . [letzter Zugriff am 28.08.2018].

Jannidis, Fotis / Kohle, Hubertus / Rehbein Malte (eds.) (2017): *Digital Humanities. Eine Einführung.* Stuttgart: Metzler.

Kath, Roxana / Schaal, Gary / Dumm, Sebastian (2015): *"New Visual Hermeneutics"* in: *ZDL* 43,1, 27-51.

Kleiber, Georges (1993): *Prototypensemantik. Eine Einführung.* Tübingen: Narr.

Küpper, Joachim / Engell, Lorenz (eds.) (2013): *The beauty of theory. Zur Ästhetik und Affektökonomie von Theorien.* München 2013.

Köppe, Tilmann / Winko, Simone (2013): *Neuere Literaturtheorien. Eine Einführung.* Stuttgart: Metzler.

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, 99-116.

Mainberger, Sabine (2003): *Die Kunst des Aufzählens.* Berlin: de Gruyter.

Mogensen, Preben (1992): *"Towards a Provotyping Approach in Systems Development"* in: *Scandinavian Journal of Information Systems* 4, 31-53.

Nünning, Vera / Nünning, Ansgar (2010): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen*. Stuttgart: Metzler.

Ramsay, Stephen / Rockwell, Geoffrey (2012): *"Developing Things: Notes toward an Epistemology of Building in the Digital Humanities"* in: **Gold, Matthew (ed.)**: *Debates in Digital Humanities.* Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/ [letzter Zugriff am 28.08.2018].

Ritter, Joachim / Gründer Karlfried (eds.) (1998): *Historisches Wörterbuch der Philosophie.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Rosch, Eleanor (1975): *"Cognitive Reference Points"* in: *Cognitive Psychology* 7, 532-547. https://doi.org/10.1016/0010-0285(75)90021-3 [letzter Zugriff am 28.08.2018].

Ruecker, Stan (2015): *"A Brief Taxonomy of Prototypes for the Digital Humanities"* in: *Scholarly and Research Communication* 6, 2. https://doi.org/10.22230/src.2015v6n2a222 [letzter Zugriff am 28.08.2018].

Ruecker, Stan / Scaletsky, Celso / Meyer, Guilherme / Michura, Piotr (unbekannt): Eintrag "Prototype" in: *MLA Commons. Digital Pedagogy in the Humanities. Concepts, Models, and Experiments* https://digitalpedagogy.mla.hcommons.org/keywords/prototype/ [letzter Zugriff am 28.08.2018].

Saar, Martin (2013): *"Kritik und Affekt. Nietzsches Textpolitik"* in: **Küpper, Joachim / Engell, Lorenz (eds.)**: *The beauty of theory. Zur Ästhetik und Affektökonomie von Theorien*. München: Fink, 47-57.

Schaub, Mirjam (2010): *Das Singuläre und das Exemplarische. Zu Logik und Praxis in Philosophie und Ästhetik.* Zürich: diaphanes.

Scheinfeldt, Tom (2012): *"Why Digital Humanities Is "Nice"* in: **Gold, Matthew (ed.)**: *Debates in Digital Humanities.* Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/ [letzter Zugriff am 28.08.2018].

Seebold, Elmar (ed.) (2011): *Kluge. Etymologisches Wörterbuch der deutschen Sprache.* Berlin: de Gruyter.

Stachowiak, Herbert (1992): *"Modell"* in: **Seiffert, Helmut / Radnitzky, Gerard (eds.)**: *Handlexikon zur Wissenschaftstheorie*. München: Ehrenwirth, 219-222.

Wellbery, David (ed.) (1985): *Positionen der Literaturwissenschaft. Acht Modellanalysen am Beispiel von Kleists ‚Das Erdbeben von Chili'*. München: Beck.

Wirth, Uwe (2013): *"Der will bloß spielen! Der Dilettant und die schöne Theorie"* in: **Küpper, Joachim / Engell, Lorenz (eds.)**: *The beauty of theory. Zur Ästhetik und Affektökonomie von Theorien*. München: Fink, 137-148.

Wittgenstein, Ludwig ([1945], 1984): *Philosophische Untersuchungen*. Bd. 1. Werkausgabe. Frankfurt: Suhrkamp.

Scalable Viewing in den Filmwissenschaften

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Universität Leipzig

Pause, Johannes
johannes.pause@uni.lu
Universität Luxemburg

Walkowski, Niels-Oliver
walkowski@nowalkowski.de
unabhängiger Wissenschaftler

Einleitung

Der vorliegende Beitrag greift das Konferenzmotto „multimedial, multimodal“ der DHd 2019 durch die Einführung des methodischen Konzepts des *Scalable Viewings* im Kontext filmwissenschaftlicher Forschung auf. Angelehnt an Designprinzipien aus dem Feld der Datenvisualisierung (Keim et al. 2008) und ersten Beiträgen zu einer Idee des Scalable Readings (Weitin 2017) zielt der Begriff Scalable Viewing darauf, die dynamische, multiperspektivische Repräsentation von Forschungsdaten als methodischen Kernbestandteil einer computergestützten Filmanalyse zu etablieren und zugleich zu systematisieren. Diese Systematisierung basiert auf zwei Forschungsprojekten, die sich mit Aspekten von Farbigkeit, ästhetischen Reizen, Figuren, Montage sowie Dialogen und Filmskripten auseinandergesetzt haben und die zur Illustration des Scalable Viewing-Konzepts im Rahmen des Vortrags vorgestellt werden.

In den letzten Jahren wurde eine Reihe von Tools zur Visualisierung von Bewegtbilddaten präsentiert (vgl. Schoeffmann et al. 2015). Die gestiegene Bedeutung dieses Themas zeigt sich in Teilen auch in der Ausrichtung eines eigenständigen Tracks (https://trecvid.nist.gov/) im Rahmen der *Text Retrieval Conference* (TREC). Während die meisten Tools und Visualisierungen aus dem Bereich der Informatik stammen, gibt es zunehmend auch Visualisierungen, die speziell das Feld der Digital Humanities adressieren. So hat Lev Manovich im Rahmen des *Visualizing Vertov*-Projekts (Manovich 2013) ein exploratives Visualisierungsinterface für die Analyse von Vertov-Filmen kreiert, das weit darüber hinaus zur Anwendung gekommen ist, so zum Beispiel bei der Untersuchung der *Jean Desmet Collection* (Olesen u. a. 2016) oder bei einer Auseinandersetzung mit 55 Filmen aus den *Walt Disney Animation Studios* (Ferguson 2017). *ScriptsThreads* (Hoyt 2014) ist ein weiteres Werkzeug, das die Figurenkonstellation in Filmen automatisiert extrahiert und visualisiert. Wiederum andere Visualisierungswerkzeuge fokussieren auf die Darstellung der Entwicklung von Stimmungen (*sentiment analysis*) oder dominanter Farbstrukturen (Burghardt et al. 2016; Hohman et al. 2017, Pause & Walkowski 2018). Besonders stark rezipiert wurde in diesem Zusammenhang das zum Begriff für eine Reihe von Visualisierungsansätzen gewordene Projekt *MovieBarcodes* (http://moviebarcode.tumblr.com/; vgl. Abbildung 1), das sich konzeptuell mindestens bis 2001 zurückverfolgen lässt (Barbieri, 2001).

Abbildung 1. Star Wars: Episode VIII - The Last Jedi (2017), Quelle: http://moviebarcode.tumblr.com/image/173580407805

Scalable Viewing: Skalierbarkeit von Abstraktion und Visualisierung

MovieBarcodes wie auch ein Großteil der anderen in diesem Kontext relevanten Visualisierungen basieren in der Regel auf einem Distant Reading-Ansatz, da sie die visuelle Ebene der Farbverwendung oder anderer Features innerhalb eines Films auf einen Blick darzustellen versuchen. Mit dem Begriff des ‘Distant Reading’ beschreibt Franco Moretti (Moretti 2013) die rein quantitative Analyse großer literarischer Korpora. Distant Reading soll eine abstrakte Repräsentation literarischer Werke ermöglichen, die auf der Grundlage bestimmter Vorentscheidungen in Daten und Diagramme übersetzt und so großmaßstäblich vergleichbar gemacht werden. Die Vogelperspektive etwa auf die Entwicklung einer ganzen literarischen Gattung über mehrere Jahrzehnte hinweg eröffnet dabei unweigerlich Einsichten, die über die qualitativ-hermeneutische Analyse und das *Close Reading* eines literarischen Texts nicht generierbar wären. Freilich nimmt eine derart distanzierte Perspektive Verluste in Kauf, indem sie dasjenige, was an den literarischen Werken kommensurabel ist, isoliert. Auf dieser Grundlage artikulieren sich dann die meisten der gegenüber digitalen Methoden geäußerten Vorbehalte, zum Beispiel jene des Reduktionismus, Empirismus oder Szientismus, die unmittelbar mit wissenschaftspolitischen Grundsatzentscheidungen zusammenfallen. Sie greifen die vermeintliche Abbildfunktion von Visualisierungen an, indem sie ihre Selektivität in den Vordergrund rücken. Solche Vorbehalte erweisen sich jedoch als nicht länger haltbar, sobald Visualisierungen im Rahmen einer vielgestaltigen Praxis konzipiert werden, die im besten Falle durch die Visualisierung selbst initiiert wird und die Statik des Abbildes auflöst.

In Erweiterung des von Martin Müller (https://scalablereading.northwestern.edu/) eingeführten und von Thomas Weitin (2017) am deutschen Novellenschatz erprobten Konzepts des *Scalable Readings* wollen wir für eine

computergestützte Auseinandersetzung mit Bewegtbildern, die in der Regel eine aisthetische Dimension aufweisen, den Begriff des *Scalable Viewings* vorschlagen. Während sich Moretti in seinem Buch vom Close Reading methodisch in einer Form abgrenzen zu wollen scheint, die nachgerade Unvermittelbarkeit suggeriert, stellt Scalable Viewing die Verbindung zu den traditionelleren Methoden der Geisteswissenschaften im Sinne eines mixed methods-Ansatzes (Kuckartz 2014) wieder her:

> Wer die Oxford-Klassiker-Ausgabe der Odyssee liest (...), ist im Grunde bereits ein ‚distant reader' der Gesänge, die Homer zugeschrieben worden sind. Und gerade der distant reader im herkömmlichen Verständnis, der mit digitalen Analysen Daten und Visualisierungen erzeugt, muss diese verstehen und interpretieren. „Scalable Reading" bedeutet indes nicht nur, dass sich close und distant reading methodisch durchdringen, es steht für ein integriertes Verständnis aller Akte des Lesens … (Weitin 2015: Abschnitt 2)

In diesem Sinne lässt sich die an Müller und Weitin angelehnte Idee einer freien Skalierung als methodische Durchdringung von Distant and Close Viewing fassen. Im Rahmen einer Analyse von Bewegtbildern lässt sich dieser Ansatz mit bestehenden Gestaltungsrichtlinien zur Informationsvisualisierung ("Overview first, zoom and filter, then details-on-demand"; Shneiderman 1996) und dem Ansatz der *Visual Movie Analytics* (Kurzhals et al. 2016) verbinden, wie am Beispiel des nachfolgend beschriebenen *Scalable MovieBarcode*-Tools gezeigt werden kann.

Wie eingangs bereits erläutert wurde, handelt es sich bei MovieBarcodes um eine typische Distant Reading-Visualisierung, werden doch Filme in komprimierter Form anhand der Farbverwendung und deren linearer Abfolge dargestellt. Um im Sinne des vorgeschlagenen Scalable Viewing-Konzepts weitere Detailebenen für die Analyse verfügbar zu machen, wurde das Konzept der *Scalable MovieBarcodes* entwickelt (Burghardt et al. 2018). Dabei kann ein Film in beliebigen Zoom-Stufen dargestellt werden, von einer Überblicksdarstellung im Sinne klassischer MovieBarcodes bis hin zur Detailansicht auf Ebene einzelner Keyframes (vgl. Abb. 2). Darüber hinaus wurde die bei MovieBarcodes rein auf die Farbinformation beschränkte Visualisierung um eine textuelle Ebene ergänzt, indem Dialoge und Figurennamen mit den jeweiligen Keyframes aligniert wurden (für technische Details vgl. Burghardt et al. 2018). Die Dialoge werden je nach Sprecher in unterschiedlichen Farbcodes als kleine Rechtecke unterhalb der Keyframes visualisiert und können bei Bedarf auch als Untertitel angezeigt werden (vgl. Abb. 3). Gleichzeitig erlaubt die textuelle Erschließung des Films die gezielte Suche nach Keyframes mit bestimmten Figuren und Schlüsselwörtern, wodurch eine Skalierung der MovieBarcodes anhand bestimmter Filterkriterien ermöglicht wird (vgl. Abb. 4).

Abbildung 2. Darstellung unterschiedlicher Zoomstufen der Scalable MovieBarcodes-Visualisierung am Beispiel von "Pretty Woman" (Garry Marshall, 1990). Oben: Überblick über alle Einstellungen; unten: Detailansicht einzelner Keyframes.

Abbildung 3. Detailansicht einer Einstellung bei "Pretty Woman" (Garry Marshall, 1990) mit Anzeige der transkribierten Dialoge.

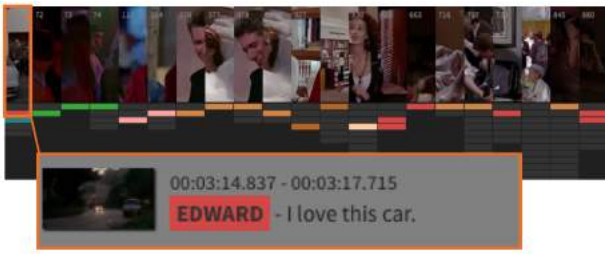


Abbildung 4. Darstellung eine Sub-MovieBarcodes, der Einstellungen mit dem Suchbegriff "love" enthält.

Erweiterung des Scalable Viewing-Konzepts um die Dimensionen Korpusgröße und Medialisierung

Das Beispiel der Scalable MovieBarcodes demonstriert, entsprechend der zuvor geäußerten These, wie durch Interaktivität und Zoomfähigkeit der Visualisierung den Fallstricken einer auf eine Abbildfunktion reduzierten Visualisierung als reines Distant Viewing entgegengearbeitet werden kann. An die Stelle des Abbildes tritt die freie Skalierung zwischen Distant und Close Viewing. Was das Beispiel darüber hinaus implizit auch noch deutlich machen kann, ist die Inkongruenz zwischen der Distant und Close Viewing-Dimension einerseits und der Frage nach der Größe des untersuchten Korpus andererseits. Scalable MovieBarcodes zeigen, dass ein Distant Viewing eines Einzelwerks genauso wie ein Close Viewing von dreißig

Sequenzen (shots) innerhalb desselben Tools denkbar sind. Die Produktivität einer solchen Distant Viewing-Analyse einzelner Filme konnte zuletzt auch anhand der detaillierten Untersuchung dreier Zombiefilme (Pause & Walkowski 2018) unter Beweis gestellt werden. Verhindert werden soll durch diese Differenzierung vor allem eine voreilige Gleichsetzung von Distant Viewing-Methoden mit "Big Data". Aus diesem Grund ist der Scalable Viewing-Ansatz um die Skalierbarkeit der Positionierung zwischen Einzelwerk- und Korpus-Perspektive zu ergänzen. So erlauben es computergestützte Methoden, beide Perspektiven innerhalb eines einzigen Forschungsprozesses positiv aufeinander zu beziehen.

Schließlich konstituiert das gemeinsame Arrangement aus Skripttext, extrahierten Frames und diagrammatischen Elementen eine dritte, eigenständige Dimension des Scalable Viewing-Ansatzes. Bei Weitin wird diese Dimension angedeutet, wenn von "einer weiteren 'scale' medialer Formen und analytischer Aufbereitungen" die Rede ist (Weitin 2015). Genau genommen sind hier zwei miteinander verwandte Möglichkeitsräume angesprochen: Zum einen geht es um den Versuch, möglichst viele Modalitäten des Forschungsgegenstandes selbst zu medialisieren, d.h. in Repräsentationen wie farbige Rechtecke, gestauchte, 1-Pixel breite Frames und ähnliches zu verarbeiten und dadurch analytisch zugriffsfähig zu machen. Zum anderen kann es aber auch darum gehen, ein einziges Merkmal des Forschungsgegenstands in vielen, möglichst unterschiedlichen Modalitäten aufzubereiten.

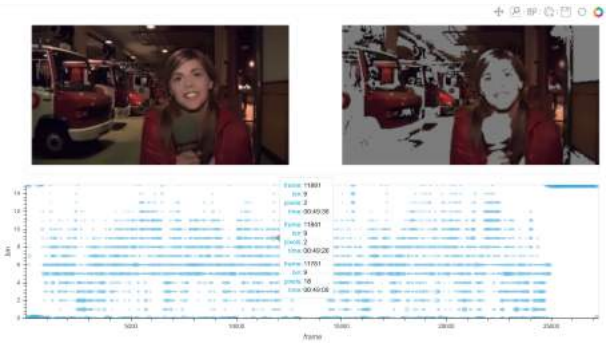


Abbildung 5. Die Repräsentation eines spezifischen Farbeffekts als numerische, diagrammatische und aisthetische Größe in einem interaktiven Filmanalyse-Dashboard.

Ein Beispiel hierfür ist das im zweiten Fallbeispiel präsentierte *HoloViews*-Dashboard (http://holoviews.org/). Das Dashboard wurde im Rahmen einer Analyse von 24 Politthrillern erstellt und kombiniert unterschiedliche Formen der Repräsentation von Farbeffekten in verschiedenen miteinander in Beziehung stehenden Widgets. Abbildung 5 zeigt, wie ein solcher Farbeffekt gleichzeitig: (a) als Punkt auf einem, den gesamten Film umspannenden Scatterplot, (b) als numerische Größe des relativen Pixelanteils am Frame sowie (c) als ein durch Histogramm-Backprojection isolierter Bereich auf dem Filmstill dargestellt werden kann. Die Bedeutung des Farbeffekts ist somit Folge einer multimodalen Kontextualisierung.

Werden nun alle drei diskutierten Dimensionen zusammengedacht, so ergibt sich ein dreidimensionaler Möglichkeitsraum, innerhalb dessen unterschiedliche Scalable Viewing-Ansätze systematisiert werden können (Abbildung 6).

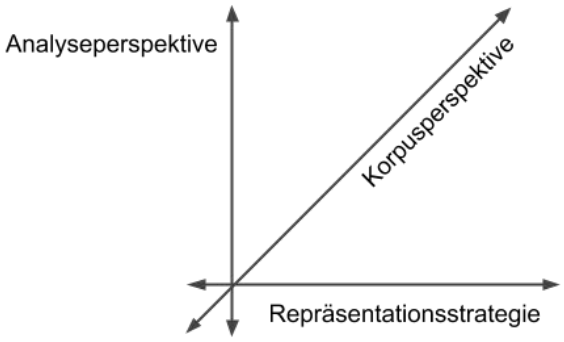


Abbildung 6. Die drei Achsen des Scalable Viewing.

Die Pointe eines Ansatzes, der Scalable Viewing als eigenständiges methodisches Konzept vorstellt, liegt mithin in der Möglichkeit, progressiv innerhalb ein und desselben Forschungsprozesses zu skalieren, anstatt sich für einen Distant oder einen Close Viewing-Ansatz, für eine Fallbeispielanalyse oder einen Big Data-Ansatz oder für die effizienteste Form der Darstellung bzw. das effizienteste Tool zu entscheiden. Dabei bezieht das vorgeschlagene Konzept die schon im Scalable Reading-Ansatz angedachte, integrative Perspektive nicht allein auf die epistemischen Objekte der Filmwissenschaft, sie sieht die Leistung computergestützter Analyseverfahren gerade darin, dass unterschiedliche Formen des Zugriffs auf das Material durch diese selbst immer von Neuem angeregt werden. Die Skalierung dient nicht der Kompensation der einen digitalen Methodik, vielmehr ist die Pluralisierung von Perspektiven auf das Untersuchungsmaterial selbst das Resultat eines durch computerisierte Verfahren erweiterten, wissenschaftlichen Möglichkeitsraums, der im Rahmen eines konzeptuellen Scalable Viewing-Rahmenwerks methodologisierbar wird. Gerade den Digital Humanities sollte es ein Anliegen sein, nicht immer schon festgelegten Methoden zu folgen, sondern den eher experimentellen Weg einer epistemischen "Bastelei" einzuschlagen, in welcher Zugriff wie Untersuchungsobjekt erst im Prozess emergieren.

Bibliographie

Barbieri, M. / Mekenkamp, G. / Ceccarelli, M., / Nesvadba, J. (2001): *The color browser: a content driven linear video browsing tool*, in: Multimedia and Expo Conference. IEEE 627–630.

Burghardt, M. / Kao, M. / Walkowski, N.-O. (2018): *Scalable MovieBarcodes – An Exploratory Interface for the Analysis of Movies*. 3rd IEEE VIS Workshop on Visualization for the Digital Humanities, Berlin.

Burghardt, M. / Kao, M. / Wolff, C. (2016): *Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis*, in: Book of Abstracts of the International Digital Humanities Conference (DH). Krakow.

Ferguson, K. L. (2019): *Digital Surrealism: Visualizing Walt Disney Animation Studios*, in: Digital Humanities Quarterly. 11(1) http://www.digitalhumanities.org/dhq/vol/11/1/000276/000276.html.

Hohman, F. / Soni, S. / Stewart, I. / Stasko, J. (2017): *A Viz of Ice and Fire: Exploring Entertainment Video Using Color and Dialogue*, in: Proceedings of the 2nd Workshop on Visualization for the Digital Humanities (VIS4DH), Phoenix, Arizona.

Hoyt, E. / Ponot, K. / Roy, C. (2014): *Visualizing and Analyzing the Hollywood Screenplay with ScripThreads*, in: Digital Humanities Quarterly. 8(4) http://www.digitalhumanities.org/dhq/vol/8/4/000190/000190.html-

Keim D. / Andrienko G. / Fekete JD. / Görg C. / Kohlhammer J. / Melançon G. (2008): *Visual Analytics: Definition, Process, and Challenges*, in: **Kerren A. / Stasko J.T. / Fekete JD. / North C. (Hrsg.):** *Information Visualization. Lecture Notes in Computer Science.* 4950: 154-175.

Kuckartz, U. (2014): *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: Springer VS

Kurzhals, K. / John, M. / Heimerl, F. / Kuznecov, P. / Weiskopf, D. (2016): *Visual Movie Analytics*, in: IEEE Transactions on Multimedia. 18(11): 2149-2160.

Manovich, L. (2013): *Visualizing Vertov*, in: Russian Journal of Communication. 5(1): 44-55 doi: 10.1080/19409419.2013.775546.

Moretti, F. (2013): *Distant Reading*. London, New York: Verso.

Gosvig Olesen, C./ Masson, E. / Van Gorp, J. / Fossati, G. / Noordegraaf, J. (2016): *Data-Driven Research for Film History: Exploring the Jean Desmet Collection*, in: The Moving Image. 6(1): 82–105.

Pause, J. / Walkowski, N.-O. (2018): *Everything is Illuminated. Zur Numerischen Analyse von Farbigkeit in Filmen*, in: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel doi:10.17175/2018_003

Schoeffmann, K. / Hudelist, M. A. / Huber, J. (2015): *Video interaction tools: A survey of recent work*, in: ACM Computing Surveys (CSUR) 48(1): 1–34.

Shneiderman, B. (1996): *The eyes have it: A task by data type taxonomy for information visualizations*, in: Proceedings of the IEEE Symposium on Visual Languages. 336-343.

Weitin, T. (2017): *Scalable Reading*, in: Zeitschrift für Literaturwissenschaft und Linguistik. 1-6 doi: 10.1007/s41244-017-0048-4.

Skalierbare Exploration. Prototypenstudie zur Visualisierung einer Autorenbibliothek am Beispiel der ›Handbibliothek Theodor Fontanes‹

Busch, Anna
annabusch@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Bludau, Mark-Jan
mark-jan.bludau@fh-potsdam.de
Urban Complexity Lab, Fachhochschule Potsdam, Deutschland

Brüggemann, Viktoria
viktoria.brueggemann@fh-potsdam.de
Urban Complexity Lab, Fachhochschule Potsdam, Deutschland

Dörk, Marian
doerk@fh-potsdam.de
Urban Complexity Lab, Fachhochschule Potsdam, Deutschland

Genzel, Kristina
kgenzel@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Möller, Klaus-Peter
klaus-peter.moeller@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Seifert, Sabine
sabine.seifert@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Trilcke, Peer
trilcke@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Stand der Forschung und Problemaufriss

Im Zuge des anhaltenden Aufschwungs der Schreibprozessforschung geraten seit einiger Zeit Quellentypen in den Blick der literatur- und kulturwissenschaftlichen Forschung, die Prozesse der auktorialen Selbstorganisation, der Notation oder der Lektüre dokumentieren. Dazu gehören etwa Notizbücher (Hoffmann 2008, Efimova 2018), Zettelkästen (Gfrereis, Strittmatter 2013; Krajewski 2011; Schmidt 2016) oder auch Autor*innenbibliotheken. Während dabei für Notizbücher und Zettelkästen avancierte, auf Transkriptionen und TEI-Editionen basierende Techniken der digitalen Präsentation entwickelt werden (vgl. etwa Radecke 2018 für Notizbücher, Schmidt 2016 für Luhmanns Zettelkasten), beschränkt sich die digitale Präsentation von Autor*innenbibliotheken derzeit auf die Bereitstellung entweder von elektronischen Findmitteln, die bibliothekarische Metadaten zur Verfügung stellen (vgl. etwa die Bibliothek Paul Celans: https://www.dla-marbach.de/bibliothek/spezialsammlungen/bestandsliste/bibliothek-paul-celan/) oder von einfachen Digitalisaten, die in einem Viewer und/oder als PDF-Download angeboten werden (vgl. etwa

die Grimm-Bibliothek https://www.digi-hub.de/viewer/browse/gelehrtenbibliotheken.grimmbibliothek/-/1/-/-/) . Wenngleich diese Präsentationsformen dem bibliothekarischen Charakter der Bücher einer Autor*innenbibliothek durchaus gerecht werden, gelingt es ihnen nicht, den autographischen Charakter dieser Bücher zu erfassen, deren Besonderheit in den Lese- und Gebrauchsspuren liegt, die die Autorin oder der Autor (bzw. später die Erb*innen, Nachlassverwalter*innen, besitzenden Institutionen etc.) in ihnen hinterlassen haben. Zugleich erweist sich die Handhabung der Bücher einer Handbibliothek nach den Standards der Edition von Autographen als unverhältnismäßig, ist das Ziel ihrer Präsentation doch eben nicht ein edierter Text; vielmehr soll durch sie der Nachvollzug kreativer Lektüre- und Gebrauchspraktiken ermöglicht werden.

Dabei lassen sich für eine solche Präsentation, in Bezugnahme auf die Bedürfnisse der Forschung, drei Ziele formulieren: 1) die Bereitstellung möglichst der Gesamtheit der Bände einer Autor*innenbibliothek; 2) die Implementierung von Suchanfragen-basierten Findmitteln; 3) die Bereitstellung von Anwendungen, mit denen sich das Profil der Sammlung und die in ihr sich abzeichnenden Kreativitäts- und Lektüremuster entdecken, erkennen und erforschen lassen.

Für Ziel 1) stehen gängige Techniken zur Präsentation von digitalisierten Büchern (zumeist auf METS/MODS basierend) bereits zur Verfügung (DFG-Viewer, Visual Library o.Ä.); auch für Ziel 2) werden heute mit bibliothekarischen oder archivarischen Discovery-Tools oder mit einer auf den X-Technologien basierenden Umgebung bereits Lösungen angeboten. Für Ziel 3), das im Fokus unseres Projektes steht, liegen für den Spezialfall Autor*innenbibliotheken hingegen bisher weder Lösungsansätze noch Konzeptualisierungen vor, obwohl in den letzten Jahren Forderungen nach der Ergänzung von such-basierten Interfaces durch ›glückliche‹ Zufälle ermöglichende Konzepte laut wurden (Dörk et al. 2011, Thudt et al. 2012, Whitelaw 2015). Auch wenn die Etablierung sogenannter Explore-Modi zur Erfüllung dieser Ziele schon teilweise beigetragen hat, fehlt es in der digitalen Verfügbarmachung noch an flexibler Navigation entlang der Relationen zwischen den Objekten (Kreiseler et al. 2017).

Projekt, Vorgehen und Korpus

Im Folgenden stellen wir einen im Verfahren des forschungsbasierten ›Rapid Prototyping‹ entwickelten Entwurf einer explorier- und skalierbaren Gesamtrepräsentation einer Autorenbibliothek vor. Dabei verbindet unsere Prototypenstudie die in der Regel als eine philologische und archiv- bzw. bibliothekswissenschaftliche Problematik adressierte Präsentation von Autor*innenbibliotheken mit der gestaltungsorientierten Forschung zur Visualisierung kultureller Sammlungen (Glinka et al. 2017; Dörk et al. 2017; Windhager et al. 2018). Im Fokus des Forschungsprojekts – das im Mai 2018 gestartet ist und bis März 2019 durchgeführt wird – steht ein in enger Koordination zwischen digitaler Archivwissenschaft, Literaturwissenschaft und informatischer Visualisierungsforschung entwickelter, webbasierter Software-Prototyp, der eine Idee realisiert, wie sich Autor*innenbibliotheken digital repräsentieren ließen. Das Projekt nimmt dabei zwei Impulse auf: Zum einen ist es dem ›Distant Reading‹ (Moretti 2013) verpflichtet, das nach Möglichkeiten einer Mustererkennung auf der Gesamtheit oder auf Teilen einer Autorenbibliothek sucht; zum anderen sollen die Repräsentationen skalierbar sein (Weitin 2017), was Fragen zu Übergängen zwischen verschiedenen Granularitäten aufwirft.

Grundlage für die modellhafte Erschließung ist die ca. 155 Bände umfassende Handbibliothek Theodor Fontanes, die im Theodor-Fontane-Archiv der Universität Potsdam bewahrt und ergänzt wird. Die Bedeutung dieser überlieferten Autorenbibliothek ergibt sich in erster Linie aus ihrer Provenienz. Sie ist Teil der ›Schriftstellerwerkstatt‹, unersetzbar wegen der zahlreichen von Fontane verfassten Marginalien und wertvoll durch die verschiedenen Widmungsexemplare (Rasch 2005). Die Bände, die Fontane zur Abfassung von Essays und Rezensionen herangezogen hat, weisen beispielsweise besonders viele Bewertungen aus Fontanes Feder auf.

Neben der vollständigen Verscannung des Bestandes nach archivarischen Standards und der Verzeichnung der Einzelbände nach bibliothekarischen Kriterien erfolgt eine Datenerfassung sowohl auf Seiten- wie auch auf Korpusebene. Die Daten zu Fontanes Handbibliothek sind gemäß relationalem Datenmodell in Tabellen verzeichnet, in denen die Verknüpfungen der Gegenstände abgebildet werden. Der Zugang zu den Daten erfolgt multiperspektivisch: auf Gesamtkorpus-, auf Objekt-, auf Seiten- und Einzelphänomenebene. Die Offenlegung der Zugänge im Rahmen einer Visualisierung zeigt beispielhaft die thematische und personelle Clusterbildung innerhalb der Sammlung oder die Verbindungen verschiedener Benutzungsspuren.

Visualisierungskonzept

Das im Projekt entwickelte Visualisierungskonzept legt einen besonderen Fokus auf die kontinuierliche, auf mehreren Granularitätsebenen zoom- und filterbare Navigation, welche die Erkundung einzelner Objekte ebenso zulässt wie deren Vergleich. Der interaktive Prototyp, der im März 2019 veröffentlicht wird, bietet drei grundlegende Ebenen, auf denen man sich durch den Bestand bewegt: Autor*innen, Bücher und Seiten.

Abbildung 1. Startseite/Buch-Ebene (Screenshot des Prototypen)

Ausgangspunkt für die Exploration der Visualisierung ist die Buch-Ebene, die eine Übersicht aller Bücher der Handbibliothek, geordnet nach Autor*innen, darstellt (Abb.

1). Dies folgt einem bei Visualisierungsinterfaces typischen Prinzip, den ersten Zugang über einen Überblick zu schaffen (Shneiderman 1996, Whitelaw 2015).

Jedes Buch wird durch einen vertikalen Balken dargestellt, in dem eine Seite wiederum durch ein abgegrenztes Segment repräsentiert wird, sodass sich eine Leseordnung der einzelnen Bücher von oben nach unten ergibt. Die Seitensegmente sind farblich entsprechend ihrer Lese- und Gebrauchsspuren kodiert. Während Seiten ohne Spuren weiß dargestellt werden, unterteilen sich die farbigen Lesespuren in die Kategorien: 1) Provenienzangaben (Grautöne), 2) Markierungen (Rottöne), 3) Marginalien (Blautöne) und 4) zusätzliches Material (Gelb). Mouseover über ein Segment zeigt eine Vorschau des jeweiligen Seiten-Scans und den Buchtitel an (Abb. 2).

Abbildung 2. Seiten-Ebene mit hervorgehobenen Marginalien und Mouseover über ein Element

Ablesbar sind auf der Buch-Ebene, die einem individuellen Strichcode der Bücher ähnelt, zum einen der Umfang eines Buches (Gesamtlänge des Balkens), aber auch die Verteilung der Lesespuren in diesem (Farbkodierung). Die Filterleiste über der Visualisierung dient hierbei als Legende für die Farbkodierung und bietet die Möglichkeit zur Fokussierung auf bestimmte Lesespur-Typen. Die Auswahl eines Lesespur-Typs löst die Entfaltung der entsprechenden Unterkategorien in der Filterleiste aus.

Mit Hilfe eines Suchfeldes können spezifische Textstellen, die einer Suchanfrage entsprechen, hervorgehoben werden. Durch die Selektion eines Buches werden die anderen Bücher zusammengestaucht und eine Detailansicht des ausgewählten Buches wird entfaltet, die zusätzliche Meta-Informationen zum Werk bietet.

Über die Scrollfunktion des Browsers können die Granularitätsebenen der Visualisierung erreicht werden. Dem Funktionsprinzip des Semantic Zoom folgend (Perlin & Fox 1993) führt Scrollen nach oben zu einer höheren Abstraktion und nach unten zu einem höheren Detailgrad – es erlaubt also einen Wechsel zwischen den drei Ebenen. Das Scrollen ermöglicht dabei kontinuierliche, sinnhafte Übergänge zwischen den Ansichten und bietet die Möglichkeit, in eigener Geschwindigkeit vor- und zurückzugehen, mit dem Ziel, die Ansichtswechsel nachvollziehbarer zu gestalten.

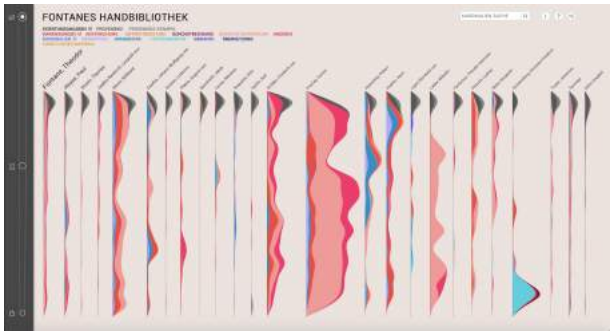


Abbildung 3. Autor*innen-Ebene. Alle Filter-Kategorien sind entfaltet. Hier deutlich zu sehen ist z.B. die bei allen Büchern gleich stark ausgeprägten Provenienzangaben in grau zu Beginn der Bücher.

Im Gegensatz zur mittleren Buch-Ebene sind auf der höheren Ebene (Abb. 3) alle Bücher eines Autors zusammengefasst, indem die Gesamtverteilung der Lesespuren in Form eines Flächen-Diagramms dargestellt wird. Hierdurch ist ein Vergleich von Fontanes Lesespuren, verteilt über die Werke unterschiedlicher Autor*innen, möglich, aber es lassen sich durch die höhere Abstraktion auch umfassende Muster nachvollziehen. Navigiert man von der mittleren Buch-Ebene in die andere Richtung auf die untere Ebene, wird die Visualisierung ins Detail entfaltet (Abb. 2), sodass für eine*n ausgewählte*n Autor*in einzelne Seiten gezielt ausgewählt werden können und Marginalien detailliert auch in der Transkription sichtbar werden.

Alle Selektionen, Filterungen und die ausgewählte Granularitätsebene werden in der URL kodiert, wodurch sowohl die Nutzung der Verlaufsfunktionen des Browsers als auch das Speichern unter Favoriten oder das Teilen und Referenzieren von Ansichten per Link möglich wird.

Reflexion der Ergebnisse

Die im Rahmen des Prototypen entwickelte, neuartige visuelle Annäherung an Autor*innenbibliotheken und die in ihnen bezeugten Lektürespuren verbindet gestaltungsorientierte Ansätze zur Visualisierung kultureller Sammlungen mit philologisch-, archiv- und bibliothekswissenschaftlichen Forschungsfragen.

Anhand visueller Filter können Untermengen identifiziert sowie Kategorien gebildet werden, die Mustererkennungen in der Sammlung ermöglichen. Nutzer*innen sollen in die Lage versetzt werden, Begriffs- und Themenräume innerhalb dezidierter Kategorien und über deren Grenzen hinweg zu erfassen. Gefragt wurde nach der Integration von Suchfunktionen, Skalierung und Sichtbarmachung in Interfaces und wie durch attraktive Einstiege in einen Bestand weitergehende Explorationsmöglichkeiten eröffnet werden können.

Die Entdeckung neuer Forschungsfragen während des Prototypingprozesses und die damit einhergehende Nachjustierung in der Erschließung beleuchtet die Wechselwirkungen zwischen visueller Forschung, Metadatenmanagement und Philologie. Deutlich wurde, dass die Sammlung als Konstrukt zu verstehen ist, das erst die (multimodale) Rückschau generiert. Dabei entsteht ›die‹ Sammlung aus der Verschränkung von Perspektive und Objekt. Die sich daraus ergebende Konsequenz, multiple

Sichten auf ›das‹ und auf ›die‹ Objekt(e) anzubieten, schlägt sich in den verschiedenen Granularitätsebenen des Protoypen nieder und hat eine Ablösung der statischen Verzeichnung eines OPAC-Katalogs durch eine beobachtungsabhängige Visualisierung der verfügbaren Quelldaten zur Folge. Diese Sichtbarmachung fußt auf einem dynamischen Modell der Sammlungserfassung geleitet durch eine digitale Repräsentationsmodellierung.

Bibliographie

Dörk, Marian / Carpendale, Sheelagh / Williamson, Carey (2011): *The Information Flaneur: A Fresh Look at Information Seeking*, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1215-1224.

Dörk, Marian / Pietsch, Christopher / Credico, Gabriel (2017): *One view is not enough: High-level visualizations of a large cultural collection* , in: Information Design Journal, 23:1, 39-47. http://mariandoerk.de/papers/idj2017.pdf

Efimova, Svetlana (2018): *Das Schriftsteller-Notizbuch als Denkmedium in der russischen und deutschen Literatur*. Paderborn: Wilhelm Fink.

Gfrereis, Heike / Strittmatter, Ellen (eds.) (2013): *Zettelkästen. Maschinen der Phantasie*. Ausstellungskatalog. Deutsche Schillergesellschaft. Marbach a.N.

Glinka, Katrin / Pietsch, Christopher / Dörk, Marian (2017): *Past Visions and Reconciling Views: Visualizing Time, Texture and Themes in Cultural Collections* , in: Digital Humanities Quarterly 11.2. http://www.digitalhumanities.org/dhq/vol/11/2/000290/000290.html

Haber, Peter (2010): *Autorenbibliotheken im digitalen Zeitalter*, in: Quatro. Zeitschrift des Schweizerischen Literaturarchivs 30/31, 39-43.

Höppner, Stefan / Jessen, Caroline / Münkner, Jörn (eds.) (2018): *Autorschaft und Bibliothek. Sammlungsstrategien und Schreibverfahren*. Mit einem Vorwort von Reinhard Laube. Göttingen: Wallstein.

Hoffmann, Christoph (2008): *Wie lesen? Das Notizbuch als Bühne der Forschung*, in: **Birgit Griesecke (ed.):** *Werkstätten des Möglichen 1930-1936: L. Fleck, E. Husserl, R. Musil, L. Wittgenstein*, Würzburg: 45-57.

Knoche, Michael (ed.) (2015): *Autorenbibliotheken. Erschließung, Rekonstruktion, Wissensordnung*. Wiesbaden: Harrassowitz.

Krajewski, Markus (2011): *Paper Machines. About cards & catalogs, 1548-1929*. Cambridge: MIT Press.

Kreiseler, Sarah / Brüggemann, Viktoria / Dörk, Marian (2017): *Tracing exploratory modes in digital collections of museum web sites using reverse information architecture*, in: First Monday 22.4.

Moretti, Franco (2013): *Distant Reading*. London: Verso.

Perlin, Ken / Fox, David (1993): *Pad: an alternative approach to the computer interface*, in: SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques, 57-64.

Rasch, Wolfgang (2005): *Zeitungstiger, Bücherfresser. Die Bibliothek Theodor Fontanes als Fragment und Aufgabe betrachtet*, in: **Ute Schneider (ed.):** *Imprimatur. Ein Jahrbuch für Bücherfreunde*. N.F. [Bd.] XIX. Wiesbaden: Harrassowitz 103-144.

Rohmann, Ivonne (2015): *Aspekte der Erschließung und Rekonstruktion nachgelassener Privatbibliotheken am Beispiel der Büchersammlungen Herders, Wielands, Schillers und Goethes*, in: **Michael Knoche (ed.):** *Autorenbibliotheken. Erschließung, Rekonstruktion, Wissensordnung*. Wiesbaden: Harrassowitz 17-59.

Schmidt, Johannes (2016): *Niklas Luhmann´s Card Index: Thinking Tool, Communication Partner, Publication Machine*, in: **Alberto Cevolini (ed.):** *Forgetting Machines. Knowledge Management Evolution in Early Modern Europe*. Leiden: Brill 289-311.

Shneiderman, Ben (1996): *The eyes have it: A task by data type taxonomy for information visualizations*, in: Visual Languages, 1996. Proceedings., IEEE Symposium. IEEE, 336-343.

Thud, Alice / Hinrichs, Uta / Carpendale, Sheelagh (2012): *The bohemian bookshelf*, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1461-1470.

Weitin, Thomas (2017): *Scalable Reading*, in: Zeitschrift für Literaturwissenschaft und Linguistik 47.1: 1-6. https://doi.org/10.1007/s41244-017-0048-4

Wieland, Magnus (2015): *Materialität des Lesens. Zur Topographie von Annotationsspuren in Autorenbibliotheken*, in: **Knoche, Michael (ed.):** *Autorenbibliotheken. Erschließung, Rekonstruktion, Wissensordnung*. Wiesbaden: Harrassowitz 147-173.

Windhager, Florian / Federico, Paolo / Schreder, Günther / Glinka, Katrin / Dörk, Marian / Miksch, Silvia/ Mayr, Eva (2018): *Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges.* in: TVCG: IEEE Transactions on Visualization and Computer Graphics. http://mariandoerk.de/papers/tvcg2018.pdf

Whitelaw, Mitchell (2015): *Generous Interfaces for Digital Cultural Collections,* in: Digital Humanities Quarterly, 9.1. http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Social Media, YouTube und Co: Multimediale, multimodale und multicodierte Dissemination von Forschungsmethoden in forTEXT

Schumacher, Mareike
mareike.schumacher@uni-hamburg.de
Universität Hamburg, Deutschland

Horstmann, Jan
jan.horstmann@uni-hamburg.de
Universität Hamburg, Deutschland

Abstract

Die Bedeutung von Social Media in den digitalen Geisteswissenschaften wächst. Nicht nur als Gegenstand der Analyse (z.B. in Gao et al. 2018 oder Reid 2011) sind Social Media für die Digital Humanities von Interesse, sondern auch zunehmend für die Dissemination von Forschungsergebnissen (vgl. Ross 2012). Vor allem in Blogs und Twitter wurde großes Potential für Diskussionen und die Verbreitung von Ergebnissen erkannt (vgl. Puschmann/Bastos 2015, Terras 2012). Auch in der Rezeptionsforschung der Wissenschaftskommunikation zeigt sich, dass Webmedien besonders relevant sind (vgl. Brossard 2013, 14096–14101) und dass diese darum in besonderem Maße zur „scientific literacy" beitragen könnten (vgl. Schäfer 2017, 283). Generell bietet (informelle) Wissenschaftskommunikation über Webmedien noch viel ungenutztes Potential (vgl. Schäfer 2017, 279–280, Neuberger 2014, Voigt 2012). Unser Beitrag zeigt, wie eine multimediale und multimodale webbasierte Strategie die Dissemination von Digital-Humanities-Methoden unterstützen und die Wissenschaftskommunikation des Forschungsfeldes stärken kann. Die quantitative Analyse der Erfolge dieser Strategie lässt Rückschlüsse darauf zu, welche Methode wem wie vermittelt werden sollte und bildet daher eine wichtige Basis für die Konzeption von Forschungsprojekten und der universitären Lehre.

Konzeptioneller Rahmen – Multimedialität, Multimodalität und Codierungssysteme in forTEXT

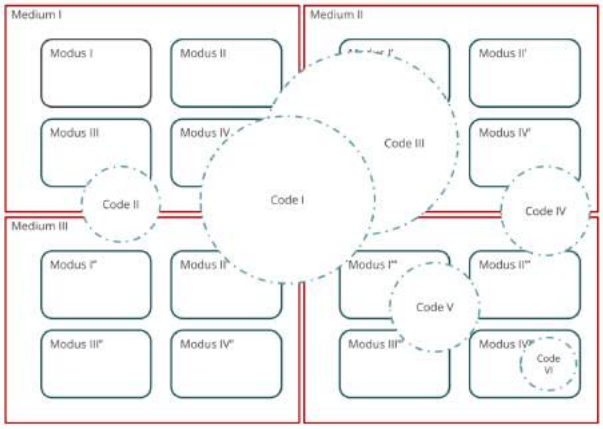
forTEXT ist ein Vermittlungsprojekt für digitale Methoden der Textanalyse, das sich vor allem an Forschende richtet, die bisher noch nicht mit digitalen Methoden arbeiten (siehe https://fortext.net). Neben der 'analogen' Dissemination in Workshops und universitärer Lehre wird auch ein Schwerpunkt auf die online-Vermittlung gelegt, da an wissenschaftlichen Themen Interessierte diese Kanäle häufig als Informationsquelle nutzen (vgl. Brossard 2013, 14098). Die hier vorgestellte webbasierte Strategie als Teil des Disseminationskonzeptes in forTEXT soll darüber hinaus zur DH-Wissenschaftskommunikation beitragen und so die Sichtbarkeit des Forschungsgebiets erhöhen (zur Bedeutung der Geisteswissenschaften in der Wissenschaftskommunikation vgl. Scheu/Volpers 2017). Für die forTEXT-Disseminationsstrategie sind Multimedialität, Multimodalität und multiple Codierungen zentrale Aspekte, die wir wie folgt definieren:

Multimedialität: Aufbereitung und/oder Nutzung unterschiedlicher medialer Kanäle. „Medium" verstehen wir wie Roesler/Stiegler (2005, 150–152) als Vermittlungssystem innerhalb eines Kommunikationsprozesses, bei dem auch das Medium selbst Teil der Vermittlung ist.

Multimodalität: Aufbereitung und/oder Nutzung unterschiedlicher Kommunikationsmodi. Dabei verstehen wir „Modus" als Bezeichnung für eine semiotische Einheit wie z.B. Design oder Sprache (vgl. Bucher 2007, 53).

Multiple kulturelle Codierung: Wir übernehmen hier ein semiotisches Verständnis von „Code " als Bezeichnung für ein System relevanter Informationseinheiten (vgl. Eco 1985, 58f.). Kulturelle Codes funktionieren als Bedeutungsnetz aus Referenzen auf ein kollektives Wissenskorpus (vgl. Barthes 1976, 25). Um den Begriff klar vom informationstechnologischen (Binär-)Code zu trennen, sprechen wir von Codierung oder Codierungssystem.

Medien, Modi und Codes wirken auf unterschiedlichen Ebenen des Vermittlungsprozesses. Dabei sind Medien und Modi stark miteinander verbunden. Modi können aber als Varianten in andere Medien übertragen werden. Codes sind inhaltliche Elemente, weshalb sie für die Dissemination von Forschungsergebnissen zentral sind. Sie können sich auf einen Modus in einem Medium beziehen oder modi- und medienübergreifend sein:



Arbeitspraxis – die webbasierte Disseminationsstrategie

Die forTEXT-Webseite als Basis medialer Vermittlung von Digital-Humanities-Inhalten

Das zentrale Vermittlungsmedium in forTEXT ist die Projektwebseite. Hier werden in Textbeiträgen sowohl Bilder als auch Videos eingebettet. Die forTEXT-Webseite bildet die Basis für die multimediale Web-Strategie, da hier grundlegende Modi und kulturelle Codierungen umgesetzt wurden, die in den sozialen Medien erweitert werden. Die primär genutzten Modi und ihre kulturellen Codierungen sind:

Design: Gedecktes Farbschema und serifenlose Schrift stehen für Schlichtheit und Sachlichkeit. Nur im Logo gibt es verspielte Elemente, die an eine Handschrift erinnern und die Verbindung von Tradition und Modernität vermitteln.

Sprache: Die Beiträge erfüllen die Ansprüche wissenschaftlichen Schreibens. Die Wissenschaftlichkeit wird durch die technische Funktionalität zum Zitieren unterstützt.

Stimme: Grundsätzlich ist die Webseite mehrstimmig angelegt, da hier verschiedene Autor*innen schreiben. Alle nutzen einen sachlichen Tonfall und die implizite Leserin wird stets mit formellem „Sie" angesprochen.

Bildlichkeit: Die eingebetteten Bilder sind zumeist digitale Repräsentationen der eingesetzten Tools und scheinen als solche zunächst gegenstandsneutral. Allerdings sind die Bilder häufig auch Visualisierungen der in forTEXT durchgeführten Fallstudien, d.h. sie zeigen nicht nur grafische, sondern auch textliche Elemente und verweisen auf die Modellierung eines Forschungsgegenstandes, die bei der Erstellung der Grafik stattgefunden haben muss.

Die forTEXT-Webseite richtet sich in erster Linie an drei Zielgruppen, die sich für die forTEXT-Disseminationsstrategie als besonders relevant erwiesen haben:

- Studierende – Lernende der DH-Methodik
- Nachwuchswissenschaftler*innen – Umsetzende der DH-Methodik
- Digitale Geisteswissenschaftler*innen – Lehrende der DH-Methodik

Social Media in forTEXT

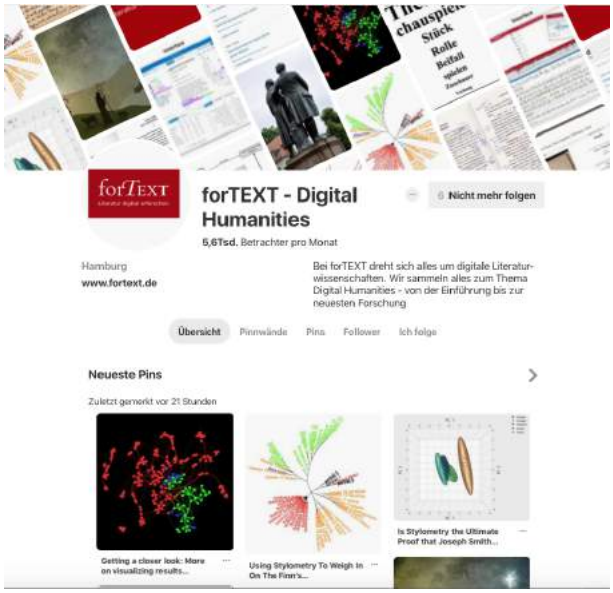
Ausgehend von den Inhalten der Webseite, deren Modi und den entsprechenden Codierungen werden drei soziale Medien zur Vermittlung genutzt. Anders als die Webseite sollen die Social-Media-Kanäle jeweils primär eine Zielgruppe erreichen: YouTube vor allem Zielgruppe 1, Pinterest Zielgruppe 2 und Twitter Zielgruppe 3.

YouTube

Auf YouTube erstellen wir eigene Inhalte, die die Artikel der Webseite aufgreifen, weiterführen und ergänzen. Es gibt derzeit zwei Inhaltstypen; Fallstudien und Tutorials. Beide können als Open-Educational-Ressources genutzt werden. In methodischen Fallstudien wird zum Beispiel mittels NER verglichen, welche Bedeutung die Hauptfiguren in Goethes *Werther* und in Plenzdorfs *neuem Werther* haben. Wir erklären, wie die NER-Machine-Learning-Prozesse funktionieren und verlinken sowohl zur Webseite als auch zu forTEXT-Tutorials. In den forTEXT-NER-Tutorials wird in drei kleinen Einheiten die Installation, Anwendung und das Training eines eigenen NER-Modells gelehrt.

Design und sprachliche Elemente der forTEXT-YouTube-Videos richten sich nach den Vorgaben der Webseite. Abbildungen der eingesetzten Tools werden ergänzt von piktografischen Animationen. Diese sind zwar schlicht, befördern jedoch Unvoreingenommenheit und Autodidaktik. Dem Vorurteil einer geringeren Technikaffinität weiblicher Menschen begegnen wir mit einem weiblichen Voice-over (vgl. Schelhowe 2000). Dadurch werden Schwellenängste abgebaut und der Eindruck vermittelt, dass Nutzer*innen und digitale Tutorin sich gemeinsam autodidaktisch an die Methoden heranwagen.

Pinterest



Die Anpassbarkeit in Hinblick auf Design, Stimme und Bildlichkeit der Kommunikationsmodi ist bei Pinterest am geringsten. Hier werden überwiegend fremde Artikel „gepinnt", die lediglich mit einer kurzen Beschreibung angereichert werden. Auch führt die Besonderheit von Pinterest als Chimäre zwischen sozialem Medium und Suchmaschine dazu, dass die sprachlichen Elemente nicht nur für die menschliche Wahrnehmung, sondern insbesondere technologisch eine Rolle spielen. Primär werden hier DH-Forschende angesprochen, die Pinnwände für Tools (z.B. Stanford-NER, Carto, Gephi, CATMA), einzelne Methoden (z.B. Netzwerkanalyse, Stilometrie, Topic Modeling), Diskussionen und viele andere Themen der Digital Humanities finden.

Twitter

Bei Twitter sind Layout und Typografie der Tweets nicht veränderbar. Jedoch wird bei jedem Tweet das forTEXT-Logo angezeigt. Im Gegensatz zu Pinterest ist auf Twitter die Ausgestaltung der Stimme bedeutsam. Hier steht die Kommunikation mit der eigenen Forschungscommunity im Vordergrund. Die Beschränktheit der Tweets auf 280 Zeichen führt dazu, dass eher Fachbegriffe als Umschreibungen genutzt werden. Hashtags führen zu Themen, die für die Community bedeutsam sind und folgen einem kulturellen Sprachcode. Hier nimmt forTEXT an kollegialen Insider-Gesprächen Teil und betont die forschungsrelevante Seite des Projektes.

Quantitative Analyse

Die webbasierte forTEXT-Disseminationsstrategie wird regelmäßig quantitativ ausgewertet. Zusätzlich zur kontinuierlichen Steigerung von Aufmerksamkeit für das Projekt (quantitativ messbar durch Impressionen, Interaktionen, Betrachtungszeiten), werden auch Analysen durchgeführt, die eher konzeptionelle Aspekte der webbasierten Dissemination von Forschungsmethoden in den Fokus rücken. Auf Basis dieser Analysen entwickeln wir eigene Relevanzmetriken, die neben quantitativen auch qualitative Aspekte berücksichtigen, wie bspw. demografische Daten, die anzeigen, welche Zielgruppen über welche Medien, welche Modi und welche Codes tatsächlich erreicht werden können, aber auch Kommentare, Feedback und Interaktionen mit anderen Nutzer*innen.

Zum jetzigen Zeitpunkt läuft die forTEXT-Social-Media-Arbeit seit drei Monaten. Auf allen medialen Kanälen zeigt sich bereits eine steigende Aufmerksamkeit, auch wenn die Zahlenwerte nach Medium stark differieren. Twitter erzielt mit durchschnittlich 20.000 Impressionen im Monat quantitativ die größte Reichweite. Auch die Interaktionsrate ist mit bis zu 7,4% relativ hoch – die Zielgruppe 3 kann hier sehr gut erreicht werden. Mit Pinterest konnten in den ersten drei Monaten durchschnittlich 1.737 monatliche Impressionen erreicht werden, wobei die einzelnen Monate mit 780 Betrachtern im ersten Monat und 9.300 Betrachtern im dritten Monat stark schwanken. Hier zeigt sich, dass die mit maschinellem Lernen verknüpfte Suchmaschine Pinterest länger braucht, um Inhalte und Interessierte zusammen zu bringen. Neue Inhalte müssen regelmäßig und vergleichsweise hochfrequent (derzeit fünf tägliche Pins) verlinkt werden, damit die Pinterest-Algorithmen ein Profil einzuordnen lernen und anderen Nutzer*innen empfehlen. Eine Einsicht aus der Analyse der Pinterest-Daten ist, dass hier insbesondere Nutzerinnen erreicht werden können. Die Vermittlung von forTEXT-Inhalten über YouTube läuft derzeit erst etwa einen Monat, sodass die Zahlenwerte (140 Impressionen im September) noch relativ gering sind. Qualitative Rückmeldung zeigt aber, dass die Videos bisher vor allem im Rahmen der DH-Lehre auf Interesse stoßen.

Bereits zu diesem frühen Zeitpunkt zeigt die Fallstudie des forTEXT-Projektes, welche Aspekte einer multimedialen, multimodalen und multipel codierten Vermittlungsstrategie sich als produktiv erweisen. Die Vorannahme, dass auf Twitter vor allem die eigene Community erreichbar ist, hat sich bestätigt. Hingegen deutet der Gender-Gap auf Pinterest an, dass hier weniger Zielgruppe 2, sondern eher Zielgruppe 1 erreicht werden kann, da vor allem die Zielgruppe der Studierenden geisteswissenschaftlicher Fächer meist überwiegend weiblich ist. Aus Feedback zu den forTEXT-YouTube-Videos konnten wir erfahren, dass diese derzeit vor allem für Lehrende von Interesse sind. Neben den vor allem im Marketing üblichen Relevanzkriterien von Impressionen, Engagement und Interaktion (die auch für die wissenschaftliche Impactmessung fruchtbar gemacht werden können, vgl. Herb/Beucke 2013) ist für die Vermittlung von DH-Methoden daher die tatsächlich erreichte Zielgruppe und deren Nutzungsmotivation relevant. So kann forTEXT am Ende nicht nur selbst Social Media produktiv nutzen, sondern auch aufzeigen, welche Medien für welche Ziele der Vermittlung von DH-Methoden besonders bedeutend sind.

Bibliographie

Barthes, Roland (1976): *S/Z*. Frankfurt am Main: Suhrkamp.

Brossard, Dominique (2013): *New media landscape and the science information consumer*, in: PNAS 110 (3), 14096–14101. https://www.pnas.org/content/pnas/110/Supplement_3/14096.full.pdf , [Zugriff 21.12.2018].

Bucher, Hans-Jürgen (2007): *Textdesign und Multimodalität. Zur Semantik und Pragmatik medialer Gestaltungsformen*, in: **Roth, Kersten Sven / Spitzmüller, Jürgen (eds.):** *Textdesign und Textwirkung in der massenmedialen Kommunikation*. Konstanz: UVK, 49–76.

Eco, Umberto (1985): *Einführung in die Semiotik*. München: Fink.

Herb, Ulrich / Beucke, Daniel (2013): *Die Zukunft der Impact-Messung. Social Media, Nutzung und Zitate im World Wide Web*, in: Wissenschaftsmanagement. Zeitschrift für Innovation 19 (4), 22–25. https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/23789/1/Die_Zukunft_der_Impact_Messung_fuer_Reps_fertig.pdf [Zugriff: 21.12.2018].

Gao, Jin / Nyhan, Julianne / Duke-Williams, Oliver / Mahony, Simon (2018): *Visualising The Digital Humanities Community: A Comparison Study Between Citation Network And Social Network*, in: Digital Humanities 2018. Book of Abstracts. Puentes-Bridges.

Neuberger, Christian (2014): *Social Media in der Wissenschaftsöffentlichkeit. Forschungsstand und Empfehlungen*, in: **Weingart, Peter / Schulz, Patricia (eds.):** *Wissen – Nachricht – Sensation. Zur Kommunikation zwischen Wissenschaft, Medien und Öffentlichkeit*. Weilerswist: Velbrück, 315–368.

Puschmann, Cornelius / Bastos, Marco (2015): *How Digital Are the Digital Humanities? An Analysis of Two Scholarly Blogging Platforms*, in: PLOS ONE 10 (2): e0115035. https://doi.org/10.1371/journal.pone.0115035 [Zugriff 2.10.2018].

Reid, Alexander (2011): *Social Media Assemblages in Digital Humanities: from Backchannel to Buzz*, in: **Wankel, Charles (ed.):** *Teaching Arts and Science with the New Social Media*. West Yorkshire: Emerald Publishing, 321–338.

Roesler, Alexander / Stiegler, Bernd (eds. 2005): *Grundbegriffe der Medientheorie*. Paderborn: UTB.

Ross, Claire (2012): *Social media for digital humanities and community engagement*, in: **Warwick, Claire / Terras, Melissa / Nyhann, Julianne (eds.):** *Digital Humanities in Practice*. London: Facet Publishing.

Schäfer, Mike S. (2017): *Wissenschaftskommunikation online*, in: *Bonfadelli, Heinz et al. (eds.): Forschungsfeld Wissenschaftskommunikation*. Wiesbaden: Springer. https://link.springer.com/content/pdf/10.1007%2F978-3-658-12898-2.pdf [Zugriff 21.12.2018].

Schelhowe, Heidi (2000): *Informatik*, in: **Braun, Christina von / Stephan, Inge (eds.):** *Gender-Studien. Eine Einführung*. Stuttgart, Weimar: Metzler, 207–216.

Scheu, Andreas M. / Volpers, Anna Maria (2017): *Sozial- und Geisteswissenschaften im öffentlichen Diskurs*, in: **Bonfadelli, Heinz et al. (eds.):** *Forschungsfeld Wissenschaftskommunikation*. Wiesbaden: Springer. https://link.springer.com/content/

pdf/10.1007%2F978-3-658-12898-2.pdf [Zugriff 21.12.2018].

Terras, Melissa (2012): *The Impact of Social Media on the Dissemination of Research: Results of an Experiment*, in: Journal of Digital Humanities 1 (3). http://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras/ [Zugriff 2.10.2018].

Voigt, Kristin (2012): *Informelle Wissenschaftskommunikation und Social Media*. Berlin: Frank & Timme.

Standardisierte Medien. Ein Paradigmenwechsel in den Geisteswissenschaften

Althof, Daniel
daniel.althof@bbaw.de
BBAW, Deutschland

Medien

Medien eröffnen nach Luhmann ein spezifisches Spektrum von Differenzen. (Seel 1998) Text, Bild, Ton sind Medien, weil in ihnen spezifische Unterschiede (wie Worte, Stufungen von Helligkeit, Farben, Tönen) gemacht werden können. In der Bereitstellung liegt die Leistung dieses Mediums. (Seel 1998: 244) Diese Differenzen lassen sich zu Zeichen formen, die *Bedeutung tragen.* Die Geisteswissenschaften (GW) erforschen diese Zeichen. Medien formen so unseren Zugang zur Wirklichkeit. Mehr noch: Wirklichkeit gibt es nicht jenseits von Medien. *Durch Medien ereignet sich Gegebensein.* (Seel 1998: 248) GW untersuchen diesen Zeichen-vermittelten Zugang zur Wirklichkeit.

Im Diskurs um Medien kommt dem Computer eine Sonderrolle zu. Er ist nicht nur ein Instrument zur Berechnung. Er ist auch nicht nur ein Apparat zur Übertragung von Informationen. Er ist ein Supermedium. Denn er enthält nicht andere Medien, sondern transformiert alles in sich und generiert alles dynamisch aus sich.[1] Text, Bild, Ton sind vom Computer selbst generierte Differenzsysteme, *die auf ein und dasselbe digitale Verfahren zurückgehen*. Indem die heterogenen Medien (wie Text, Bild, Ton) durch den Computer transformiert werden, verschmelzen der kommunikative Aspekt des Mediums, ein Mittler von Bedeutung zu sein, d.h. Bedeutung zu *tragen*, mit dem technischen Aspekt des Mediums, ein Mittel (zur Übertragung) zu sein und so Bedeutung zu *übertragen*.

Die Standardisierung

Der Vortrag möchte dieser Transformation nachgehen und die Implikationen für die Arbeit in den GW herausarbeiten. Dabei nutzt der Vortrag die Standardisierung in den DH als ein Vehikel für die Überlegungen, gleich wohl nicht über Standardisierung referieren. Die Standardisierung ist ein Angelpunkt für die Gewährleistung derjenigen Funktionalitäten, die eine digitale Reproduktion von Inhalten ermöglichen. Das Spektrum reicht von rein technischen Standards für Datenspeicherung oder Datenaustausch bis zu »inhaltlichen« Standards für die Strukturierung von Daten. Entscheidend jedoch ist der Umstand, dass diese Ebenen nicht getrennt werden können, sondern sich überlagern und gegenseitig durchdringen. Die Standards ermöglichen also jene Verschmelzung des technischen und des kommunikativen Aspekts, die den Computer auszeichnet.[2] Sie sind technische Grundlage und zugleich inhaltliche Basis für das Tragen und Übertragen von Bedeutung. Technische, syntaktische und semantische Ebene wirken unaufhebbar ineinander. Semantik wird technisch realisiert und Technik auf Semantik hin entworfen.

Und diese Verschmelzung verändert den Umgang mit den Zeichen, ja sogar die Natur des Zeichens selbst, um dessen Verstehen es in den GW grundsätzlich geht. Anders formuliert: Der Vortrag möchte einen Paradigmenwechsel konstatieren, der darin besteht, dass das Zeichen und die Erforschung der Zeichen in den GW eine tiefgreifende Transformation dadurch erfahren, dass sie aus den analogen Medien in ein digitales Supermedium transferiert werden.

Der Paradigmenwechsel

Was die Digitalisierung für die DH leistet, ist damit vergleichbar mit der Mathematisierung und Formalisierung der Naturwissenschaften (NW) im 19. Jahrhundert. Diese Umwälzung in den NW lässt sich Ideengeschichtlich nachverfolgen in ihrem Siegeszug, der mit der Umstellung vom scholastischen, deduktiven Verfahren auf das experimentelle, induktive Verfahren begann. Die Ablösung des aristotelischen Weltbildes, in dem metaphysische Erwägungen über das Sein und Seiendes im Vordergrund standen, durch ein funktionalistisches und kausales Weltbild[3] ist der Katalysator für eine Formalisierung der Erkenntnis, die nicht durch die Auslegung von Zeichen (und Begriffen) gewonnen wird, sondern auf Freilegung quantifizierbarer Entitäten beruht, deren Verhältnisse sich *berechnen* lassen.

Was also für die NW die Rückführung der Sachverhalte auf Zahlen war, ist für die GW die Rückführung der Zeichen auf Bits und Bytes, realisiert in multiplen Standards. In der NW ist das Ziel die umfassende *Berechenbarkeit*. In den DH steht dagegen die *Abbildbarkeit* im Vordergrund. Die Formalisierung ändert den Modus des Erkennens in den NW vom Verstehen zum Berechnen. Die Digitalisierung ändert den Modus des Erschließens vom Zeichen-Lesen zur Zeichen-Verarbeitung. Hier gilt es im Vortrag die Strukturen genau zu untersuchen und zu vergleichen.

Die Rolle des Computers wird sodann von entscheidender Bedeutung sein. Denn aufgrund seiner Funktion als Supermedium führt er die spezifischen Unterschiede aller Medien zurück auf die logisch minimale Differenz von 0 und 1. Dies wird mit der Absicht zu untersuchen sein, das Ersetzen der Medien-spezifischen Materialität durch frei konstruierbare Interaktivität im Mediendiskurs einzuordnen. Das formale Zeichensystem des Computers ist vor diesem Hintergrund als ein Hybrid zu sehen. Als Zeichen herkömmlicher Natur steht es für ein anderes und gibt etwas zu verstehen. Es sagt etwas. Es vermittelt etwas – passiv. Es

ist ein Medium einer Botschaft. Aber als formales Zeichen wird es aktiv, bildet nicht nur ab, sondern implementiert vermittelst der involvierten Standards selbst eine Logik, die das vertretene Zeichen nochmals mit Gehalt anreichert. Die mit den Standards technisch, syntaktisch und semantisch in das Zeichen selbst eingeschleuste Logik bestimmt das Zeichen über sein Gegebensein hinaus unaufhebbar. Das Zeichen bildet nicht nur die Relation zu allen übrigen Zeichen ab, sondern ebenso die durch seine technische Reproduzierbarkeit eingeschleppte Verfasstheit.[4] So kann also gesagt werden, dass der Computer als Super-Medium ein spezifisches Spektrum von Differenzen im Rahmen der implementierten Standards eröffnet. Der Computer stellt Gegebensein her und holt dabei das Zeichen selbst aus der ideell-geistigen Assoziation in die real-technische Verlinkung. Nichts Geringeres als die vollständige Abbildung ist das Ziel. Zudem transformiert der Computer das Zeichen aus der Oberfläche in die Tiefen-Schichtungen der multiplen Standards.[5] Die Syntax-Semantik-Distinktion der Zeichen wird grundiert mit einer technischen Dimension.

War anfangs festgehalten, dass der Computer zwei wichtige Aspekte eines Mediums vereint, also sowohl Bedeutung zu tragen als auch Bedeutung zu übertragen, so soll am Ende des Vortrages einsichtig gemacht worden sein, dass das digitale Zeichen (als Forschungsgegenstand der GW) über die Standardisierung selbst diese beiden Aspekte vereint. An dieser Stelle soll auch klar werden, dass die Digitalisierung in der Tat trotz aller Unterschiede zur Formalisierung entscheidende Gemeinsamkeiten mit dieser hat. Denn es geht bei der umfassenden Abbildbarkeit von Zeichen zugleich auch um eine automatisierbare (und so berechenbare) Verarbeitung, die keine (äußerliche) Manipulation von Zeichen ist, sondern (immanente) Weiterverarbeitung an Hand einer in den Zeichen selbst liegenden Logik. Das Zeichen wird wie die Zahl zur Sache selbst.[6] Das Zeichen wird in formale Strukturen überführt und tendenziell damit identisch. Insofern es die GW selbst sind, die nun als digitale die Zeichen reproduzieren, annotieren und manipulieren, ereignet sich nun nicht mehr Gegebensein, auf das sich die GW beziehen, sondern das Gegebensein und die Zeichen selbst werden von den GW aktiv gestaltet. Und das ändert das Verständnis und die Richtung geisteswissenschaftlicher Forschung nachhaltig.

Fußnoten

1. Aus technischer Sicht Coy 1994: 30; Aus philosophischer Perspektive Seel 1998: 258f und Berry 2011: 9f oder sehr ausführlich Robben 2006.
2. Hier kommen zwei Aspekte zusammen, die sich in der Medientheorie lange Zeit als Oppositionen entwickelt haben. Auf der einen Seite das Verständnis des Computers als Werkzeug. Auf der anderen Seite das Verständnis des Computers als Medium.
3. Zentrale Figuren sind hier F. Bacon und G. Galilei. Vgl. Wandschneider 2004: 55ff.
4. Vgl. P. F. Stefan: „Die Bedingungen des Sag- und Zeigbaren gehen unmittelbar konstitutiv in das Denkbare ein. Darstellung und Herstellung bilden einen Funktionszusammenhang." (Stefan 2000). B. Doltzer widmet diesem Zusammenhang eine ganze Monographie, in der er den wechselseitigen Zusammenhang von „Medien und Wissen, Wissen und Technik" offenlegt, „der sich zeigt, wenn man beide Seiten, Diskurs *und* Medium, als verkörpertes Wissen begreift." (Doltzer 2006: 9) Die hier formulierte These geht über die von S. Krämer formulierte hinaus, die auf einen „Überschuß an Sinn" aufmerksam macht, der jedem Medium qua der „medialen Materialität" unwillkürlich inneliegt (Krämer 1998: 78f).
5. Zur Thematik von Oberfläche und Tiefe vgl. Burkhardt 2015: 73ff.
6. Die Zahl verweist auf kein anderes, sondern steht für sich selbst. Die Zahl ist Teil eines Symbolsystems, das seine Bedeutung qua Regeln definiert. Formal-logische Systeme generell, die aus einem Vokabular (aus Zeichen) und Transformationsregeln bestehen, die die Herstellung weiterer gültiger Ausdrücke beschreiben, heißen generell Kalkül. Das Dezimalsystem und die darauf definierten Rechenregeln ist z.B. ein solches Kalkül. Ein Kalkül verzichtet auf die Bedeutung dessen, was er repräsentiert. Die Zeichen haben nur eine Kalkül-interne Bedeutung, die durch ihre Transformationsregeln bestimmt sind (vgl. Krämer 1988: 60).

Bibliographie

Berry, D. M. (2011): *Philosophy of Software. Code and Mediation in the Digital Age*, Basingstoke, Hampshire: Palgrave Macmillan UK.

Burkhardt, M. (2015): *Digitale Datenbanken. Einen Medientheorie im Zeitalter von Big Data*, Bielefeld: transcript Verlag.

Coy, W. (ed.) (1994): *Aus der Vorgeschichte des Mediums Computer*, München: Wilhelm Fink Verlag.

Doltzer, B. J. (2006): *Diskurs und Medium. Zur Archäologie der Computerkultur*, München: Wilhelm Fink Verlag.

Krämer, S. (1988): *Symbolische Maschinen. Die Idee der Formalisierung in geschichtlichem Abriß*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Krämer, S. (1998): *Das Medium als Spur und als Apparat*. In: **Krämer, S. (ed.)**: *Medien Computer Realität*. Frankfurt/M.: Suhrkamp.

Robben, B. (2006): *Der Computer als Medium. Eine transdisziplinäre Theorie*, Bielefeld: transcript Verlag.

Seel, M. (1998): *Medien der Realität und Realität der Medien*. In: **Krämer, S. (ed.)**: *Medien Computer Realität*. Frankfurt/M.: Suhrkamp.

Stefan, P. F. (2000): *Denken am Modell – Gestaltung im Kontext bildender Wissenschaft*. In: **Brüdeck, B. E. (ed.)**: *Der digitale Wahn*. Frankfurt/M.: Suhrkamp.

State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines

Reul, Christian
christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Springmann, Uwe
uwe@springmann.net
Universität Würzburg, Deutschland

Wick, Christoph
christoph.wick@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Introduction

During the last few years, great progress has been made on OCR methods which can mainly be attributed to the introduction of a line-based recognition approach using recurrent neural networks (Breuel et al. 2013). Since this breakthrough, impressive recognition accuracies beyond 98% have been achieved on a variety of materials, ranging from the earliest printed books (Springmann et al. 2016; Springmann and Lüdeling 2017) to modern prints (Breuel 2017; Wick et al. 2018). Early prints show a high variability in terms of printing types and therefore usually require book-specific training in order to reach desirable character error rates (CER) below 1-2%. On the contrary, modern typography is much more regular and mixed models, i.e. models trained on a variety of fonts and typesets from different sources, comfortably achieve CERs well below 1% without any book-specific training. Apart from the aforementioned introduction of new recognition techniques and network structures, several methodical improvements like pretraining (transfer learning) and majority or confidence voting have been introduced and successfully evaluated, especially for the application on early printed books (Reul et al. 2018).

Printings from the 19th century represent a middle ground between the two periods introduced above, considering both the variability of typesets and the state of preservation of the scans. Mixed models have achieved encouraging results without the need for book-specific training but the expectable recognition accuracy still is substantially lower than for prints from the 21st century (Breuel et al. 2013). Just as for modern prints, there is a great need for highly performant mixed models for 19th Fraktur scripts since there are masses of scanned data available online, consisting of a variety of materials including novels, newspapers, journals, and even dictionaries.

In this paper, we describe the training procedure leading to our own strong mixed models and compare the evaluation results to those achieved by other main OCR engines and their respective models on a variety of Fraktur scripts. In particular, we report results from OCRopus, Tesseract, and ABBYY Finereader each with their own standard Fraktur model as well as OCRopus and Calamari with a mixed model trained on a Fraktur corpus of the 19th century.

Related Work

Only few evaluation results are available on 19th century Fraktur OCR data. A rare exception is the evaluation of the Fraktur model of OCRopus trained on around 20,000 mostly synthetically generated text lines (Breuel et al. 2013). Evaluation on two books of different scan qualities yielded impressive CERs of 0.15% and 1.37% respectively. There exist other evaluations on more recent (Breuel et al. 2013) or older texts (Springmann and Lüdeling 2017) yielding better and worse results, respectively. An evaluation of OCR data on a wider range of Fraktur texts of different quality is missing.

Methods

In this section we briefly describe the OCR engines ABBYY Finereader, OCRopus, Tesseract, and Calamari, our training and evaluation data as well as the transcription guidelines.

OCR Engines

For contemporary material the proprietary ABBYY OCR engine (https://www.ABBYY.com) clearly defines the state of the art for layout analysis and OCR covering close to 200 recognition languages including Fraktur printed in the 18-20th centuries with an "Old German" dictionary which we used for our experiments.

The open source engine OCRopus was the first one to implement the pioneering line-based approach introduced by Breuel et al. (Breuel et al. 2013) using bidirectional LSTM networks. Apart from the superior recognition capabilities compared to glyph-based approaches, this method has the advantage of allowing the user to train new models very comfortably by just providing image/text pairs on line level.

Calamari (https://github.com/Calamari-OCR) , also available under an open source license, implements a deep CNN-LSTM network structure instead of the shallow LSTM used by OCRopus. It yields superior recognition capabilities compared to OCRopus and Tesseract (Wick et al. 2018). Because of its Tensorflow backend it is possible to utilize GPUs in order to support very fast training and recognition. In addition, it supports the training of voting ensembles and pretraining, i.e. it uses an already existing model as a starting point instead of training from scratch.

Until recently, the open source OCR engine Tesseract (https://github.com/tesseract-ocr) used individual glyphs rather than entire text lines for training and recognition. However, version 4.0 alpha also added a new OCR engine based on LSTM neural networks and a wide variety of trained mixed models. Like ABBYY and contrary to OCRopus and Calamari, Tesseract supports the use of dictionaries and language modelling.

Training Data

To achieve high quality results on early prints it is usually necessary to perform a book-specific training. For our 19th century mixed model we try to avoid this by training on a wide variety of sources over four subsequent training

steps (see Table 1). First, we use corpora with texts from different centuries for pretraining to achieve a certain overall robustness. Next, the training continues by incorporating synthetic data generated from freely available Fraktur fonts. The training concludes with the addition of real Fraktur data from the 19th century. After training on the entire data set, we perform a final refinement step in which we only use a subset of at most 50 lines per book in order to prevent the model from overfitting to the books with a high number GT lines available (10,000+ compared to less than 50 for some books). The described data are mostly available online in the GT4HistOCR corpus (Springmann et al. 2018).

Table 1. Corpora used for training our mixed models. Apart from the data available in the GT4HistOCR corpus we also incorporated lines from the Archiscribe project (https://archiscribe.jbaiter.de) and the GitHub repository of Jesper Zedlitz (JZE, https://github.com/jze/ocropus-model_fraktur).

Data	Cent.	# Books	# Lines	Lang	Step
ENHG	15	9	24,766	ger	Pretraining
Kallimachos	15,16	9	20,929	ger, lat	Pretraining
EML	15-17	12	10,288	lat	Pretraining
RIDGES	15-19	20	13,248	ger	Pretraining
UW3	20	-	96,481	eng	Pretraining
Synth.	-	66 fonts	99,214	ger	Synth. Data
DTA19	19	39	243,942	ger	Real Data
Archiscribe	19	103	3,430	ger	Real Data
JZE	19	8	1,636	ger	Real Data
DTA19	19	39	1,950	ger	Refinement
Archiscribe	19	103	3,429	ger	Refinement
JZE	19	8	355	ger	Refinement

Evaluation Data

For evaluation, we used four corpora from the 19th century (Table 2, top), which were completely different from the training data, and consisted of 20 different evaluation sets (Table 2, bottom).

Table 2. "Novels" (N) is a corpus consisting of novels currently collected and captured by the Chair for Literary Computing and German Literary History of the University of Würzburg. The "OCR-Testset" (O, https://github.com/cisocrgroup/Resources/tree/master/ocrtestset) consists of a novel and a journal. "Daheim" comprises four volumes of a German journal and "Sanders" (S) is a German dictionary provided by the Berlin-Brandenburg Academy of Sciences and Humanities.

Data	Period	# Lines	# Books
Novels	1781-1873	3,483	13
OCR-TS	1809-1841	465	2
Daheim	1865-1875	583	4 vol.
Sanders	1865	630	1

ID	(Short) Title	# Lines
N-1781	Eleonore	305
N-1803	Liebe-Hütten	184
N-1810	Der Held des Nordens	264
N-1818	Reinhold	253
N-1826	Frauenwürde	268
N-1836	Die Ruinen im Schwarzwalde	318
N-1848	Levin	269
N-1851	Georg Volker	264
N-1859	Der beseelte Schatten	260
N-1865	Gefahrvolle Wege	333
N-1869	Der Arzt der Seele	250
N-1870	Die Bank des Verderbens	273
N-1873	Natürliche Magie	242
O-1809	Wahlverwandtschaften	223
O-1841	Grenzboten	242
D-1865	Daheim volume 1865	134
D-1875	Daheim volume 1875	144
D-1882	Daheim volume 1882	142
D-1892	Daheim volume 1892	163
S-1865	Sanders Dictionary	630

Figure 1 shows some example lines.

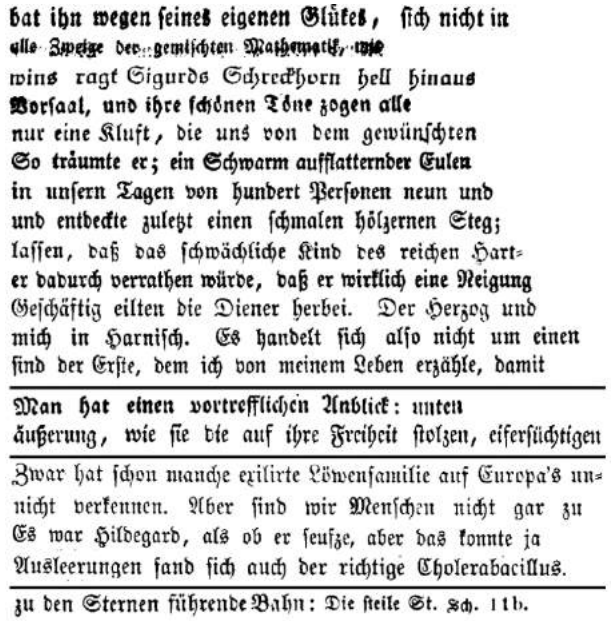


Figure 1. Example line images of the 20 evaluation works in the order given in Table 3. For practical reasons, all lines have been vertically normalized and some lines have been shortened.

Transcription Guidelines and Resulting Codec

Before starting the training, we had to make several decisions regarding the codec, i.e. the set of characters known to the final model. We kept the long s, resolved all ligatures with the exception of ß (sz), regularized Umlauts like å, o͛, u͛, quotation marks, different length hyphens, the r rotunda (#) and mapped the capital letters I and J to J. Applying these rules resulted in a codec consisting of 93 characters:

- special characters:
  #!"\&'()[]*,-./:;=?§⸗
- digits:
  0123456789
- lower case letters:
  abcdefghijklmnopqrsſßtuvwxyz
- upper case letters:
  ABCDEFGHJKLMNOPQRSTUVWXYZ
- characters with diacritica:
  ÄÖÜäöüàèé

Evaluation

Table 3 summarizes the results of applying the four OCR-Engines to the 20 data sets from Table 2. For all evaluations the experiments were performed on well segmented line images provided by ABBYY.

Table 3. CERs in percent of different OCR engines and their respective mixed models: Tesseracts "frk_best" model (Tess), OCRopus with its standard Fraktur model (FRK) and the mixed model trained by us (OCRo), and Calamari with and without voting.

Data	Tess	FRK	OCRo	Abbyy	Calamari	
	single	single	single	default	single	voted
N-1781	6.61	4.08	2.48	2.79	0.81	**0.56**
N-1803	17.17	18.21	11.30	26.54	6.38	**4.75**
N-1810	5.26	5.30	1.92	3.22	0.45	**0.21**
N-1818	7.90	7.73	3.85	9.30	1.85	**0.96**
N-1826	2.77	1.00	0.40	1.04	0.08	**0.01**
N-1836	6.88	4.68	2.01	2.70	0.70	**0.56**
N-1848	1.58	1.17	0.33	0.57	0.08	**0.02**
N-1851	1.93	0.63	0.24	0.70	0.09	**0.04**
N-1855	4.58	4.42	1.38	3.83	0.80	**0.58**
N-1859	2.19	1.42	0.31	0.38	0.17	**0.08**
N-1865	2.44	1.31	0.62	1.23	0.19	**0.13**
N-1870	2.09	1.97	0.43	0.47	0.26	**0.10**
N-1873	2.53	1.14	0.32	0.34	0.22	**0.14**
N-all	4.39	3.42	1.58	3.13	0.71	**0.47**
O-1809	3.04	2.22	1.13	1.62	0.26	**0.20**
O-1841	2.09	1.06	0.60	0.79	0.13	**0.07**
O-all	2.40	1.44	0.77	1.06	0.17	**0.11**
D-1865	2.10	1.85	0.71	**0.16**	0.26	0.17
D-1875	1.50	0.85	0.17	**0.04**	0.09	0.09
D-1882	1.53	1.17	0.43	**0.09**	0.20	0.12
D-1892	0.90	0.45	0.23	**0.01**	0.02	**0.01**
D-all	1.48	1.05	0.38	**0.07**	0.17	0.09
S-1865	5.12	10.02	5.91	5.47	2.74	**2.14**
NOD	3.68	2.80	1.29	2.38	0.55	**0.37**
All	3.87	3.76	1.90	2.80	0.84	**0.61**

Discussion

A striking result is the great variation among the CERs, e.g. by a factor of more than 2,500 from 26.54% to 0.01% for ABBYY and more than 400 from 4.75% to 0.01% for Calamari voted, which probably depends on the quality of the scans as well as the similarity of each font to the training data. Furthermore, training a model on real Fraktur data outperforms a model trained on mostly synthetic data generated for Fraktur (e.g. FRK vs. OCRo). The self-trained Calamari models achieve the best results, outperforming ABBYY by 70% without voting and even by 78% with voting averaged over all 20 datasets yielding an average CER below 1%.

For all approaches, the most frequent error either consists in the insertion (Tesseract) or the deletion of whitespaces (all others) leading to merged or splitted words. This represents a common problem with historical prints, as the inter word distances vary heavily. The error distribution varies considerably for the different engines. For example, in the case of ABBYY the three most frequent errors make up to less than 5% of all errors, whereas OCRopus (close to 9%) and Calamari (over 15%) show a considerably more top-heavy distribution.

Conclusion and Future Work

Our evaluations showed that open source engines can outperform the commercial state-of-the-art system ABBYY by up to 78% if properly trained. The resulting models as well as the data required to adjust the model's codec are publicly available (https://github.com/chreul/19th-century-fraktur-OCR) . Further improvements can be expected by providing more ground truth for training the mixed model and by using even deeper neural networks than the Calamari default. While ABBYY already has strong postprocessing techniques available, this represents an opportunity to improve the results achieved by Calamari and OCRopus even further, in particular the inclusion of dictionaries and language models.

Bibliographie

Breuel, Thomas M. / Ul-Hasan, Adnan / Al-Azawi, Mayce / Shafait, Faisal (2013): *"High-performance OCR for printed English and Fraktur using LSTM networks"* in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE.

Breuel, Thomas M. (2017): *"High performance text recognition using a hybrid convolutional-LSTM implementation"* in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. IEEE.

Reul, Christian / Springmann, Uwe / Wick, Christoph / Puppe, Frank (2018): *"Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning"* in ArXiv preprints: https://arxiv.org/abs/1802.10038 (accepted for JLCL Volume 33 (2018), Issue 1: Special Issue on Automatic Text and Layout Recognition).

Springmann, Uwe / Fink, Florian / Schulz, Klaus-U. (2016): *"Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents"* in ArXiv preprints: https://arxiv.org/abs/1606.05157 .

Springmann, Uwe / Lüdeling, Anke (2017): *"OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus"* in Digital Humanities Quarterly 11, 2: http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html .

Springmann, Uwe / Reul, Christian / Dipper, Stephanie / Baiter, Johannes (2018): *"Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin"* in ArXiv preprints: https://arxiv.org/abs/1809.05501 (submitted to JLCL Volume 33 (2018), Issue 1: Special Issue on Automatic Text and Layout Recognition).

Wick, Christoph / Reul, Christian / Puppe, Frank (2018): *"Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition"* in ArXiv preprints: https://arxiv.org/abs/1807.02004 (submitted to Digital Humanities Quarterly).

Tanz annotieren - Zur Entstehung, den Möglichkeiten und den Perspektiven digitaler Methoden in der Tanzwissenschaft

Rittershaus, David
david.rittershaus@hs-mainz.de
Hochschule Mainz - University of Applied Sciences, Deutschland

Der Choreograph William Forsythe und die digitalen Technologien

In seiner gut dreißigjährigen Schaffensphase in Frankfurt am Main initiierte und unterstützte der Choreograph William Forsythe mehrere Großprojekte, die sich intensiv mit den Möglichkeiten zur digitalen Aufzeichnung, Archivierung, Vermittlung und Publikation von Tanz auseinandersetzten. Es sind im Laufe dieser Projekte Softwareanwendungen entstanden, die in ihrer heutigen Weiterentwicklung die Verwendung in der Tanzpraxis erlauben. Angesichts der insgesamt wachsenden Sammlungen tanzbezogener Daten stellt sich daher die Frage, ob und in welcher Form digitale Methoden auch in der Tanzwissenschaft zum Tragen kommen können. Im Folgenden soll anhand der Initiativen William Forsythes die Entwicklung nachgezeichnet werden, um anschließend grundlegende Fragen bezüglich der Möglichkeiten digitaler Methoden für die Tanzwissenschaft zu diskutieren.

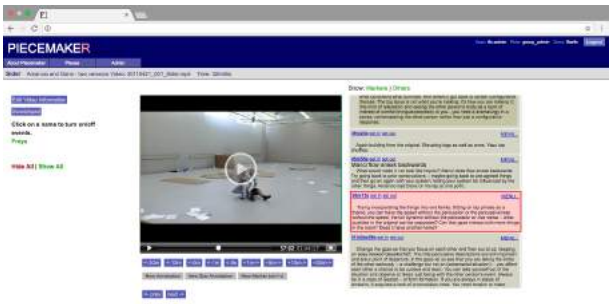
In den insgesamt zwanzig Jahren (1984-2004) unter Forsythes Leitung wurde das Frankfurter Ballett zu einem der bedeutendsten Tanzensembles der Welt. Obwohl Forsythe sich dem Ballett verbunden sah, entwickelte er immer mehr eine eigene Bewegungssprache (Siegmund 2011). Forsythe suchte daher auch nach neuen Möglichkeiten der Aufzeichnung und Vermittlung für seinen Tanz, der sich gar nicht oder nur schwer mit „herkömmlicher" Tanznotation (Jeschke, 2010) erfassen lässt. Er wandte sich dafür bereits in den 1990er Jahren digitalen Technologien zu. So entstand in dieser Zeit in Zusammenarbeit mit dem Zentrum für Kunst und Medien in Karlsruhe (ZKM) die CD-ROM *Improvisation Technologies: a tool for the analytical dance eye*. Sie beinhaltet knapp 65 Videos, in denen Forsythe Bewegungen vorführt und erläutert. Die Videos werden von Grafiken überlagert, die räumliche Relationen und Bezüge in und um den Körper herum veranschaulichen und sich als Annotationen beschreiben lassen (DeLahunta & Jenett 2017: 68).

Die Möglichkeiten digitaler Technologien nutzte Forsythe für seine Zwecke mit seinem neuen Ensemble *The Forsythe*

Company von 2005 an verstärkt. Gemeinsam mit einem Team der Ohio State University wurden die inneren organisatorischen Strukturen der Choreographie *One Flat Thing, reproduced* herausgearbeitet, visualisiert und unter dem Titel *Synchronous Objects* publiziert[1]. Annotationen dienen hier, wie DeLahunta und Jenett darlegen, zu Zwecken der Repräsentation: "[…] not only to draw attention to two key choreographic structuring components, the cueing and alignment systems, but also as a part of instructional videos." (DeLahunta & Jenett 2017: 70).

Piecemaker - Software zur Prozessdokumentation im Tanz

Beinahe gleichzeitig entwickelte das Ensemblemitglied David Kern eine webbasierte Anwendung, die er *Piecemaker* nannte, und die von 2008 bis 2014 *The Forsythe Company* zur digitalen Aufzeichnung von Proben und Aufführungen diente. Die Webanwendung ermöglichte es zeitgleich Videos und schriftliche Anmerkungen zu erfassen und verband beides automatisch miteinander, sodass die Anmerkungen als Videoannotationen gespeichert wurden. In der Zeit der Nutzung von *Piecemaker* durch die Forsythe Company ist ein ausgesprochen umfangreiches und einzigartiges digitales Archiv entstanden, das von einigen Stücken die Entstehung beinahe vollständig dokumentiert.



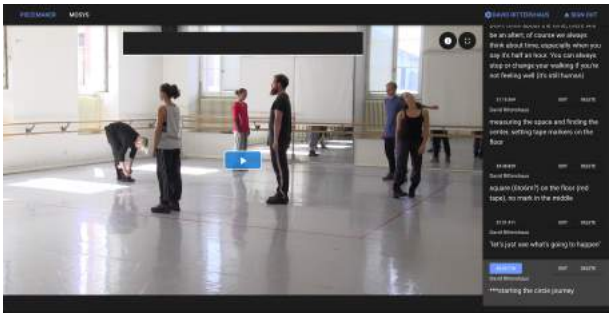
Version 1 der *Piecemaker*-Webanwendung, wie sie von *The Forsythe Company* verwendet wurde.

Mit dem Projekt *Motion Bank*[2] gelang es Forsythe ein weiteres großes Vorhaben zur Erforschung digitaler Tanzaufzeichnung zu initiieren. *Motion Bank* übernahm die *Piecemaker* Anwendung und entwickelte sie weiter, um Choreographien einiger bedeutender Künstler*innen des Zeitgenössischen Tanzes aufzuzeichnen und zu annotieren (DeLahunta & Jenett 2017: 73). Die Ergebnisse wurden in Form sogenannter *Online Scores* im Web veröffentlicht[3] und gewähren einen Einblick in die Konzepte und die Praxis der jeweiligen Choreograph*innen. Nach dem Ende der Förderung durch die Kulturstiftung des Bundes gelang es Florian Jenett *Motion Bank* 2016 als Forschungsprojekt an der Hochschule Mainz zu etablieren[4], wo er es gemeinsam mit Scott DeLahunta leitet. Im Rahmen der Forschungsarbeit konnten seither die *Piecemaker*-Anwendung und die Publikationsplattform zum *Motion Bank Web System* weiterentwickelt und dessen Einsatz mit verschiedenen Partnern in der Tanzpraxis erprobt werden. Dabei wird der Ansatz verfolgt den Tanzschaffenden eine niederschwellige und freie Software an die Hand zu geben, mit der sie ihre eigene Praxis aufzeichnen und annotieren können.

Tanz annotieren

Das Besondere an *Piecemaker* ist, dass die Annotationen zwar auch anhand eines Videos erstellt werden können, vor allem aber im Tanzstudio mit dem Blick auf das Geschehen im Raum aufgezeichnet werden. Dieses Verfahren zur Annotation von Tanz nennen wir Live-Annotation und es unterscheidet sich von der Post-Annotation, bei der Annotationen nachträglich bzw. anhand zeitbasierter Medien hinzugefügt werden. Die über sechs Jahre hinweg entstandenen Annotationen im Tanzstudio der Forsythe Company, wurden größtenteils von der Dramaturgin des Ensembles, Freya Vass-Rhee, als Live-Annotationen erstellt. Sie hielt darin Anweisungen des Choreographen, Diskussionen des Ensembles, Szenenabläufe, Licht- und Musikeinsätze usw. fest. Die Annotationen sind dabei keineswegs „in einer algorithmischen Strenge, Konsequenz und Ausnahmslosigkeit […] formalisiert" (Rapp 2017, S. 256), die eine maschinelle Auswertung ohne Weiteres zulassen würde. Eine „Ontologie" liegt ihnen nicht zugrunde. Dennoch bringen sie Metadaten mit sich, die es ermöglichen, die Daten in verschiedenen Benutzeroberflächen strukturiert zu veranschaulichen. Die Annotationen selbst enthalten für Tanzforschende Informationen, die sich aus dem Video alleine nicht erschließen lassen, sondern die Anwesenheit bei der Probe oder der Aufführung voraussetzen oder Innenansichten der beteiligten Künstler*innen widerspiegeln. Ein großes Potenzial der Annotation zeitbasierter Medien im Tanz ist die Versammlung unterschiedlicher Perspektiven wie Innen- und Außensicht, Nähe zum Geschehen (Live-Annotation) und Distanz (Post-Annotation).

Zu der aktuellen Überarbeitung und der Integration von *Piecemaker* in das *Motion Bank Web System* gehört auch die Änderung der Datenstruktur, die nun auf dem Web Annotation Data Model des W3C[5] aufbaut und mit Linked Data kompatibel ist. So ist es theoretisch in Zukunft möglich verschiedene Datensätze zu verlinken oder mit anderen digitalen Tanzarchiven, wie bspw. dem digitalen Pina Bausch Archiv (Thull 2014), zu verbinden. Dafür müssten jedoch überhaupt erst Top-Level-Ontologies geschaffen werden. Angesichts der Heterogenität der Ansätze und Konzepte im Zeitgenössischen Tanz, die sich oftmals dezidiert versuchen einer Kategorisierung zu entziehen und nicht mehr auf ein standardisiertes Bewegungsrepertoire zurückgreifen, ergeben sich für die Schaffung von „Ontologien" auf unterschiedlichen Ebenen einige Problemstellungen, die auch auf grundsätzliche Fragen digitaler Wissensrepräsentation verweisen.



Ansicht aus der derzeitigen Weiterentwicklung von *Piecemaker,* inzwischen Teil des *Motion Bank Systems.* Live-annotierte Probe der tanzmainz Compagnie am Staatstheater Mainz mit Choreograph Taneli Törmä.

Tanzwissenschaft als kritische Wissenschaft

Wenn es für die Geisteswissenschaften ein "systembedingtes Primat des Individuellen vor dem Allgemeinen" (Jannidis 2017: 107) gibt, gilt das für die Erforschung des Zeitgenössischen Tanzes allemal: Ihn selbst charakterisieren „Diffusionen heterogener Tanzstile und choreographischer Verfahren", ein fortlaufender Wandel von Formen in einem Prozess ständiger „Überarbeitung und Um-Schreibung" (Traub 2001: 181-184). Die Tanz- und Theaterwissenschaft hat es mit Singulärem zu tun, will sie es damit aufnehmen, so der Theaterwissenschaftler Nikolaus Müller-Schöll, muss sie „[…] zu allererst die mit ihrem Namen gesetzten Voraussetzungen – das Theater wie die Wissenschaft – radikal in Frage stellen." (Müller-Schöll 2016: 150). Die Tanzwissenschaftlerin Gabriele Klein sieht gerade in dem Finden einer Sprache für dynamische Vorgänge eine Herausforderung für Wissenschaftler*innen, die scheitern muss. Daher sei Tanzwissenschaft immer auch Wissenschaftskritik „[…] insofern, als sie sich gegen ein Wissen wendet, das dynamische Vorgänge über statische Konzepte zu fassen versucht." (Klein 2007: 33). Tanzwissen, über das Tänzer aufgrund körperlicher Erfahrung verfügen, kennzeichnet Klein als „spezifisches narratives Wissen" (Klein 2007: 32). Vor diesem Hintergrund bleiben Fragen bezüglich der quantitativen bzw. maschinellen Auswertung tanzbezogener Daten offen, beispielsweise inwiefern damit verbundene Verfahren der Kategorisierung und Formalisierung gerade die Spezifik eliminieren, die den Tanz auszeichnet, oder inwiefern die Komplexität zeitgenössischer Tanzpraxis reduziert wird, die von der Tanzwissenschaft sonst narrativ abgebildet werden kann. Als kritische Wissenschaft, die immer auch ihre eigenen Grundbedingungen reflektiert, muss die Tanzwissenschaft, will sie mit digitalen Methoden arbeiten, sich mit den offenen Fragen digitaler Wissensrepräsentation auseinandersetzen:

> „[...] there is a lot of work to be done on unpicking the normative values which underlie the kinds of schema and classification embedded within the algorithmic data structures of linked data. Indeed, the fragmentation of knowledge into chunks that can be composed and recomposed at will points to the nature of a computational episteme which privileges knowledge divided into non-narrative shards of information […]. Additionally, the political economy of linked data, and linked open data, raises important questions for the way in which humanities knowledge is converted into data lakes that become a kind of oil for a postmodern capitalism." (Berry & Fagerjord 2017: 77)

Fazit

Der Einsatz digitaler Methoden und ihre Erforschung stehen in der Tanzwissenschaft noch am Anfang. Während mit der maschinellen und quantitativen Auswertung tanzbezogener Daten noch viele offene Fragen verbunden sind, zeichnen sich neue Möglichkeiten qualitativer Auswertung durch die diagrammatischen Text-Bild-Verbünde (Krämer 2016) der Benutzeroberflächen einer Annotationssoftware wie *Piecemaker* und des *Motion Bank Systems* ab. Sie beruhen auch auf quantitativen Verfahren, bei denen bisher vorranging aus den Metadaten Benutzeransichten mit Visualisierungen zur Betrachtung der Daten abgeleitet werden und somit digital unterstützte qualitative Arbeitsweisen eröffnen. Zukünftig soll das *Motion Bank System* die Möglichkeit bieten individuelle, kontextspezifisch Vokabularien anzulegen, um diese bei der Annotation mit *Piecemaker* anwenden zu können. Während sich die Diskussionen rund um die Erstellung solcher Vokabularien im Praxisfeld des Zeitgenössischen Tanzes und der Tanzausbildung bereits als fruchtbar erweisen[6], ist noch offen, ob und inwiefern sie sich für tanzwissenschaftliche quantitative Methoden eignen oder ob sie in erster Linie dazu dienen können, umfangreiche Inhalte in der Software – im Zuge einer vorrangig qualitativen Auswertung – besser zu filtern, zu strukturieren und zu sortieren.

Fußnoten

1. https://synchronousobjects.osu.edu/ , zuletzt aufgerufen am 23.09.2018.
2. http://www.motionbank.org/ , zuletzt aufgerufen am 23.09.2018.
3. http://scores.motionbank.org/ , zuletzt aufgerufen am 23.09.2018.
4. https://medium.com/motion-bank/motion-bank-at-hochschule-mainz-c89ef4a61643 , zuletzt aufgerufen am 23.09.2018.
5. https://www.w3.org/TR/annotation-model/ , zuletzt aufgerufen am 23.09.2018.
6. https://medium.com/motion-bank/developing-vocabularies-for-dance-education-e4c4584950a8 , zuletzt aufgerufen am 08.01.2019.

Bibliographie

Berry, David M. / Fagerjord, Anders (2017): *„Knowledge Representation and Archives."* In: **Berry, David M. / Fagerjord, Anders (eds.)**: *Digital Humanities. Knowledge and Critique in a Digital Age*. Cambridge, Malden: Polity 60-79.

DeLahunta, Scott / Jenett, Florian (2017): *„Making digital choreographic object interrelate. A focus on coding practices."*

in: **Beyes, Timon / Leeker, Martina / Schipper, Imanuel (eds.)**: *Performing the digital: performativity and performance studies in digital cultures*. Bielefeld: transcript Verlag 63-79.

Jannidis, Fotis (2017): *„Grundlagen der Datenmodellierung“* in: **Jannidis, Fotis /Kohle, Hubertus / Rehbein, Malte (Hrsg.)**: *Digital Humanities. Eine Einführung.* Stuttgart: Metzler 99-108.

Jeschke, Claudia (2010): *„Tanz-Notate: Bilder. Texte. Wissen.“* in: **Brandstetter, Gabriele / Hofmann, Franck / Maar, Kirsten (Hrsg.)**: *Notationen und choreographisches Denken.* Freiburg: Rombach 47-65.

Klein, Gabriele (2007): *„Tanz in der Wissensgesellschaft.“* in: **Gehm, Sabine / Husemann, Pirkko / von Wilcke, Katharina (Hrsg.)**: *Wissen in Bewegung. Perspektiven der künstlerischen und wissenschaftlichen Forschung im Tanz.* Bielefeld: transcript 25-36.

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis. Grundlinien einer Diagrammatologie.* Frankfurt am Main: Suhrkamp.

Müller-Schöll, Nikolaus (2016): *„Das Problem und Potential des Singulären. Theaterforschung als kritische Wissenschaft.“* in: **Cairo, Milena / Hannemann, Moritz / Haß, Ulrike / Schäfer, Judith (Hrsg.)**: *Episteme des Theaters. Aktuelle Kontexte von Wissenschaft, Kunst und Öffentlichkeit.* Bielefeld: transcript 139-150.

Rapp, Andrea (2017): *„Manuelle und automatische Annotationen“* in: **Jannidis, Fotis /Kohle, Hubertus / Rehbein, Malte (Hrsg.)**: *Digital Humanities. Eine Einführung*. Stuttgart: Metzler 253-267.

Siegmund, Gerald (2011): *„Of monsters and puppets. William Forsythe's work after the ‚Robert Scott Complex'“* in: **Spier, Steven (ed.)**: *William Forsythe and the Practice of Choreography*. London, New York: Routledge 20-37.

Thull, Bernhard (2014): *„Das digitale Pina Bausch Archiv“* in: **Pina Bausch Stiftung (Hrsg.)**: *Tanz erben*. Bielefeld: transcript 59-73.

Traub, Susanne (2001): *„Zeitgenössischer Tanz“* in: **Dahms, Sibylle / Jeschke, Claudia / Woitas, Monika. (Hrsg.)**: *Tanz*. Kassel: Bärenreiter Metzler 181-188.

Technologienutzung im Kontext Digitaler Editionen – eine Landschaftsvermessung

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Universität zu Köln, Deutschland

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Deutschland

Harzenetter, Lukas
lukas.harzenetter@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Barzen, Johanna
johanna.barzen@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Breitenbücher, Uwe
uwe.breitenbuecher@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Leymann, Frank
frank.leymann@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Geisteswissenschaftliche Forschung ist schon seit langem eine auch digitale Praxis. Dies spiegelt sich u. a. in den Methoden der Ergebnissicherung wider: Präsentationssysteme, interaktive Visualisierungen, Recherche-Datenbanken und nicht zuletzt Digitale Editionen haben sich neben der klassischen Publikation längst als digitale Instrumente der Ergebnissicherung etabliert. Während für die Persistenz statischer Daten wie z. B. digitaler Dokumente bereits gut etablierte Strategien und Standards wie TEI-XML (TEI 2018) oder OASIS-DITA (OASIS 2018) existieren, stellt Forschungssoftware im Hinblick auf die langfristige Ergebnissicherung noch immer eine besondere Herausforderung dar.

Dies äußert sich in erster Linie in einem Mangel an konkreten Nachhaltigkeitsstrategien und erwächst u. a. aus einem als „Software-Aging“ (Parnas 1994) bekannten Problem, dem Forschungsanwendungen als „Lebende Systeme“ (Sahle/Kronenwett 2013) grundsätzlich unterworfen sind, da sie, wie jede Software, nicht unabhängig von ihrer Laufzeitumgebung bzw. ihrem digitalen Ökosystem gedacht werden können. Die kontinuierliche Evolution dieser Umgebungen sorgt dafür, dass Softwaresysteme, die nicht stetig an diese veränderten Umweltbedingungen angepasst werden, mit der Zeit veralten und letztendlich unbenutzbar werden.

In dem von der DFG geförderten Kooperationsprojekt „SustainLife – Erhalt lebender, digitaler Systeme für die Geisteswissenschaften“, das in einer Zusammenarbeit zwischen dem Data Center for the Humanities (DCH, siehe http://dch.phil-fak.uni-koeln.de) der Universität zu Koln und dem Institut fur Architektur von Anwendungssystemen (IAAS, siehe http://www.iaas.uni-stuttgart.de) der Universitat Stuttgart durchgefuhrt wird, arbeiten wir an einem Lösungsvorschlag für dieses Problem (vgl. dazu Barzen et al. 2018 sowie Neuefeind et al. 2018). Gegenstand des Projekts ist die Adaption und Weiterentwicklung von Verfahren und Technologien aus dem Cloud-Deployment für die Digital Humanities (DH) mit dem Ziel, Management und Provisionierung von DH-Anwendungen zu optimieren und deren Sicherung und nachhaltigen Betrieb zu realisieren.

Das Projekt setzt hierbei auf den OASIS-Standard TOSCA (*Topology and Orchestration Specification for Cloud Applications*, siehe OASIS 2013, OASIS 2016) sowie dessen Open-Source-Implementierung OpenTOSCA (Binz et al. 2013, Breitenbücher et al. 2017). Mithilfe von TOSCA können (Forschungs-)Anwendungen mitsamt ihrer jeweiligen Laufzeitumgebung als Topologien von zusammenhängenden Software-Artefakten und Schnittstellen in standardisierter Weise modelliert werden. Daraufhin können diese Topologie-Modelle und damit alle benötigten Komponenten und Dateien der modellierten Applikation in sogenannten CSARs (*Cloud Service Archives*) paketiert werden, so dass diese portablen Archive von jeder TOSCA-Runtime interpretiert und automatisiert bereitgestellt werden können.

Methodisches Vorgehen

Eine wesentliche Voraussetzung für eine Lösung, die eine standardisierte, TOSCA-konforme Beschreibung von Softwaresystemen vorsieht, ist ein möglichst genaues Bild der technologischen Landschaft innerhalb der Digital Humanities. Grundlage für den im Projekt verfolgten Ansatz ist daher eine ausführliche Bedarfs- und Anforderungsanalyse, anhand derer die Spezifikation häufig eingesetzter Systemkomponenten sowie die Identifikation von Schlüsselkomponenten zur Erstellung von Anwendungsvorlagen erfolgen kann.

Im Projekt setzen wir hierfür im Wesentlichen auf zwei Instrumente: Zum einen quantitativ angelegte Umfragen in der Community in Form von Fragebögen, zum anderen die qualitativ orientierte Untersuchung von ausgewählten Beispielanwendungen auf Basis von gezielten Codeanalysen. Letzteres zielt auf eine genauere Analyse der quantitativ erhobenen Daten bezüglich der eingesetzten Technologien, um konkrete Systemkomponenten identifizieren zu können, die im weiteren Projektverlauf u. a. für die Modellierung von Usecases auf Basis des TOSCA-Standards eingesetzt werden sollen.

Eine erste Kurzumfrage wurde im Rahmen der DHd2018 in Köln unter den Teilnehmern der Konferenz durchgeführt (Barzen et al. 2018). Primäres Ziel des eingesetzten Fragebogens war es, eine explorative Typisierung von DH-Anwendungen vorzunehmen und das Spektrum der dabei eingesetzten Technologien zu konturieren. Die Umfrage ergab im Wesentlichen, dass sich Entscheidungen bezüglich der Technologienutzung vor allem am jeweiligen Anwendungstyp orientieren. Um eine möglichst zielgenaue Analyse typischer Technologien zu erreichen bietet es sich deshalb an, den Skopus der Befragung entsprechend auf einzelne Anwendungstypen hin einzugrenzen.

Technologienutzung im Kontext Digitaler Editionen

Nachdem sich der Zusammenhang zwischen der Art der Anwendung und den konkreten technologischen Entscheidungen besonders deutlich bei Digitalen Editionen zeigte, und da diese zugleich als besonders repräsentative Form digitaler Ergebnissicherung im Bereich der Digitalen Geisteswissenschaften angesehen werden können, haben wir uns dazu entschieden, hier den ersten Ausgangspunkt für eine gezielte Erhebung von Daten über die Technologienutzung zu setzen. Hierfür richteten wir gemeinsam mit dem Cologne Center for eHumanities (CCeH, siehe http://cceh.uni-koeln.de) sowie in Kooperation mit der Landesinitiative NFDI der Digitalen Hochschule NRW (siehe https://fdm-nrw.de) und der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste (AWK, siehe http://www.awk.nrw.de einen Workshop zum Thema „Nachhaltigkeit Digitaler Editionen“ aus, der am 17.9.2018 in den Räumen der AWK stattfand (siehe http://dch.phil-fak.uni-koeln.de/nde-workshop.html).

Der Workshop hatte das Ziel, vorhandene Lösungsansätze und Aktivitäten, die der nachhaltigen Bereitstellung von Digitalen Editionen gewidmet sind, mit der Fachcommunity zu diskutieren. So wird bereits seit einigen Jahren an verschiedenen Stellen an Konzepten gearbeitet, die dieses Problem adressieren. Bspw. wird in der Schweiz derzeit eine „Nationale Infrastruktur für Editionen“ (NIE-INE, siehe https://www.nie-ine.ch) aufgebaut, die auf die Homogenisierung von Editionsprojekten zielt, und in Österreich wurde mit dem „Kompetenznetzwerk Digitale Editionen“ (KONDE, siehe http://www.digitale-edition.at) ein umfangreiches Verbundprojekt eingesetzt, das ebenfalls den Aufbau und die Weiterentwicklung einer nationalen Forschungsinfrastruktur für Digitale Editionen zum Ziel hat. Auch in Deutschland gibt es vergleichbare, wenn auch lokaler ausgerichtete Initiativen, die auf die Schaffung einer gemeinsamen Infrastruktur für Editionsprojekte und damit längerfristig auf eine Standardisierung von Digitalen Editionen zielen. Im Rahmen des Workshops hatten wir die Gelegenheit viele der maßgeblichen Akteure als Sprecher zu gewinnen. Unter den Teilnehmern fanden sich ebenfalls viele Personen, die selbst an Digitalen Editionen arbeiten oder an der Nachhaltigkeitsproblematik anderweitig interessiert waren, z.B. aus Perspektive der Drittmittelgeber.

Begleitende Umfrage

Im Vorfeld des Workshops führten wir unter den insgesamt 80 Teilnehmern (70 Gäste und 10 eingeladene Sprecher) eine Umfrage durch, mit dem Ziel, Informationen über die Technologielandschaft speziell in diesem Bereich der DH zu gewinnen. Die Umfrage enthielt Fragen zu persönlichen Vorerfahrungen in der Arbeit mit Digitalen Editionen, zu den formalen Charakteristika von Editionsprojekten sowie zu den dabei eingesetzten Technologien. Schon anhand einer ersten Auswertung der Umfrage, die unmittelbar vor dem Workshop vorgenommen wurde und bei der die Angaben zu 31 verschiedenen Editionsprojekten berücksichtigt wurden, ließen sich deutliche Tendenzen in Bezug auf die eingesetzten Nachnutzungs- und Nachhaltigkeitsstrategien ablesen. So zeigen die in Abbildung 1 wiedergegebenen Angaben bezüglich der Datenmodellierung sehr deutlich, dass mit TEI-XML mittlerweile ein weit verbreiteter Standard zur nachhaltigen Sicherung von Daten vorliegt, der für speziellere Bedarfe durch weitere standardisierte Formate ergänzt wird.

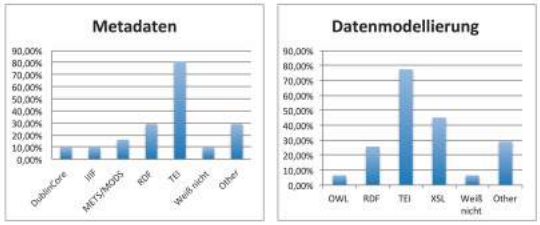


Abbildung 1: Nutzung von Standards für Metadaten und für die Modellierung von Primärdaten.

Aufseiten der Implementierungen dagegen fehlen solche Standards, was entsprechend zu einer deutlich höheren Heterogenität der eingesetzten Technologien führt (siehe Abbildung 2). Aus der Zusammenschau der verschiedenen Fragekategorien lässt sich ersehen, dass es zwar durchaus auch in Bezug auf die genutzten Technologien und Systemkomponenten Favoriten gibt, jedoch wird unsere Annahme einer technologischen und methodischen Heterogenität des Feldes dennoch klar bestätigt.

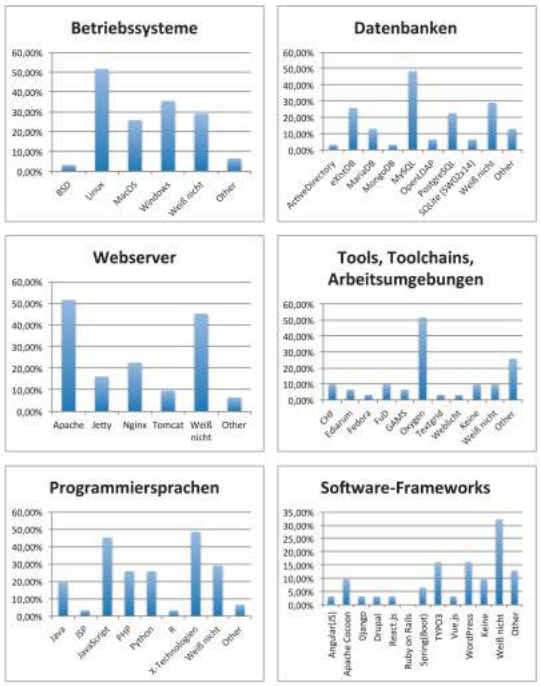


Abbildung 2: Technologienutzung im Kontext Digitaler Editionsprojekte.

Während sich also im Bereich der Primärdaten in Gestalt verschiedener TEI-Dialekte eine zunehmende Standardisierung abzeichnet, so gilt dies nicht für das Layout und die Präsentation von Editionen. Hinsichtlich Aussehen, Funktionalitäten und technischer Architekturen besteht hier weiterhin eine sehr große Heterogenität. Zudem deuten sich hier auch bereits verschiedene Technologie-Stacks an, die sich aus typischen Kombinationen von Systemkomponenten ergeben. Allerdings scheinen sich solche Stacks v. a. innerhalb verschiedener Trägerinstitutionen herauszubilden, was sich in Teilen auf personelle Überschneidungen zwischen Projekten oder auf institutionelle Lösungsansätze zur Bewältigung von Nachhaltigkeitsanforderungen zurückführen lässt. Diese Ergebnisse spiegelten sich auch in den Diskussionen auf dem Workshop wider. Während die Nachhaltigkeit der Daten weitestgehend gelöst zu sein scheint, brachte die Frage nach der Nachhaltigkeit der Software eine Reihe von Lösungsansätzen hervor. Die Vor- und Nachteile von verschiedenen Technologie-Stacks wurden kontrovers diskutiert. Weitestgehende Einigkeit bestand darin, dass eine Digitale Edition aus Daten plus Software besteht, dass also eine alleinige Konservierung der Daten für die Konservierung der Digitalen Edition an sich nicht ausreicht.

Fazit und Ausblick

In unserem Beitrag möchten wir die bisherigen Ergebnisse unserer Anforderungsanalyse an lebende Systeme in den Digital Humanities zur Diskussion stellen. Neben den hier gezeigten Kennzahlen fokussieren wir dabei verschiedene Technologie-Stacks, die typische Kombinationen von Systemkomponenten widerspiegeln. Die hier vorgestellten Ergebnisse der Anforderungsanalyse dienen uns als Basis für die weitere Projektarbeit. Unser Ziel ist es, häufig genutzte Schlüsselkomponenten und typische Anwendungsstrukturen digitaler Erkenntnisträger im Bereich der Digital Humanities zu identifizieren, sie anschließend in TOSCA zu modellieren und zu einer Komponentenbibliothek zusammenzufassen sowie in Form von Anwendungsvorlagen für weitere Modellierungen bereitzustellen, um dadurch konsistente und zukunftssichere Standards und Nachhaltigkeitsstrategien für Forschungssoftware zu etablieren. Über die konkrete Projektarbeit hinaus sind die hier vorgestellten Ergebnisse auch von generellem Interesse für die Community. So sind die erhobenen Daten u. a. auch für Datenzentren wie das DCH Köln von hohem Nutzen, um beispielsweise Strategien und Betreuungskonzepte für Forschungssoftware zu entwerfen.

Bibliographie

Barzen, J. / Blumtritt, J. / Breitenbücher, U. / Kronenwett, S. / Leymann, F. / Mathiak, B. / Neuefeind, C. (2018): *"SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften."* In: Book of Abstracts der 5. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2018), Köln 26.2.–2.3.2018, S. 471–474. https://kups.ub.uni-koeln.de/8085/1/boa-DHd2018-.pdf [letzter Zugriff 8.1.2019].

Binz, T. / Breitenbücher, U. / Haupt, F. / Kopp, O. / Leymann, F. / Nowak, A. / Wagner, S. (2013): *"OpenTOSCA - A Runtime for TOSCA-based Cloud Applications"*. In: ICSOC, 2013, S. 692–695.

Breitenbücher, U. / Barzen, J. / Falkenthal, M. / Leymann, F. (2017): *"Digitale Nachhaltigkeit in den Geisteswissenschaften durch TOSCA: Nutzung eines standardbasierten Open-Source Ökosystems"*, in: DHd 2017: Digitale Nachhaltigkeit, S. 235–238. http://www.dhd2017.ch/wp-content/uploads/2017/01/Abstractband_def_24.1.17-1.pdf [letzter Zugriff 8.1.2019].

Neuefeind, C. / Harzenetter, L. / Schildkamp, P. / Breitenbücher, U. / Mathiak, B. / Barzen, J. / Leymann, F. (2018): *"The SustainLife Project – Living Systems in Digital Humanities"*. In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing, 2018 (IBM Research Report RC25681), S.101-112.

OASIS (2013): *„Topology and Orchestration Specification for Cloud Applications Version 1.0"*. 25 November 2013. OASIS Standard. http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html . [letzter Zugriff 8.1.2019]

OASIS (2016): *„TOSCA Simple Profile in YAML, Version 1.0"*. Edited by Derek Palma, Matt Rutkowski, and Thomas Spatzier. 21 December 2016. OASIS Standard. http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.0/os/TOSCA-Simple-Profile-YAML-v1.0-os.html . [letzter Zugriff 8.1.2019]

OASIS (2018): *„Darwin Information Typing Architecture (DITA) Version 1.3 Errata 02"*. Edited by Robert D. Anderson and Kristen James Eberlein. 19 June 2018. OASIS Approved Errata. http://docs.oasis-open.org/dita/dita/v1.3/errata02/os/dita-v1.3-errata02-os.html [letzter Zugriff 8.1.2019].

Parnas, D. L. (1994): *"Software Aging"*. In: Proceedings of the 16th International Conference on Software Engineering (ICSE 1994). IEEE, Mai 1994, S. 279–287.

Sahle, P. und Kronenwett, S. (2013): *"Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities'"*. In: LIBREAS. Library Ideas 23, S. 76–96.

TEI Consortium, eds. (2018): *"TEI P5: Guidelines for Electronic Text Encoding and Interchange."* Version 3.4.0 vom 23.7.2018. http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 8.1.2019].

"The Bard meets the Doctor" – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Meyer, Selina
Selina.Meyer@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Schmidtbauer, Stephanie
Stephanie.Schmidbauer@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Molz, Johannes
johannes.molz@googlemail.com
Ludwig-Maximilians-Universität München, Deutschland

Einführung: Shakespeare und Intertextualität

In der jüngeren Literatur- und Kulturtheorie geht man davon aus, dass alle literarischen Texte immer auch durch eine reiche Tradition, ja ein ganzes Ökosystem, bestehender Literatur beeinflusst sind (Allen, 2000, S. 1). Dieser Einfluss, der sich im Text sowohl in impliziten als auch in expliziten Querverbindungen durch Zitate offenbart, wird gemeinhin als Intertextualität bezeichnet[1]. Als eine der zentralen Kulturtechniken der Postmoderne, lässt sich das Zitat an keinem Autor so gut studieren wie an William Shakespeare, dessen intertextuelles Nachwirken die gesamte westliche Kulturhemisphäre durchzieht (Maxwell & Rumbold, 2018, S. 1). Eine Untersuchung dieser allgegenwärtigen Spuren ist somit zwingend auch eine Geschichte unserer gesamten Kultur (Taylor, 1991) und beantwortet und eröffnet Fragen, die weit über Shakespeare hinausgehen. Dabei finden sich intertextuelle Bezüge auf das Werk Shakespeares nicht nur in unterschiedlichsten literarischen Genres, sondern in zunehmendem Maße auch im Massenmedium Film und Fernsehen (Malone, 2018).

Wenngleich Film in erster Linie ein visuelles Medium ist, so bietet sich über die Dialoge zusätzlich ein verbaler Analysezugang (vgl. Kozloff, 2000), Klarer (1998, S. 54) spricht in dem Zusammengang gar von Film als semi-textuelles Genre. Die den Filmen zugrunde liegenden Skripte lassen sich als Dramen lesen, da sie ausschließlich aus Figurennamen, Sprechakten und Bühnenanweisungen bestehen. Für die quantitative Analyse von Intertextualitätsphänomenen bei Filmen liegt ein großer Vorteil in der Verfügbarkeit vollständig transkribierter Dialoge die als Untertitel[2], Drehbücher[3] oder Fan-Transkripte[4] über diverse Online-Portale angeboten werden. Die maschinenlesbaren Dialoge können sodann mit bestehenden Methoden und Algorithmen aus dem informatischen Anwendungsbereich der *text reuse detection*, also der Identifikation einer Wiederverwendung bestimmter Textfragmente in anderen Texten, eingesetzt werden. Verwandte Arbeiten zur computergestützten Intertextualitätsforschung finden sich vor allem im Bereich historischer Sprachen, insbesondere Latein und Griechisch (vgl. Berti et al., 2013; Büchler et al. 2014; Scheirer et al., 2014; Forstall et al., 2015; Bamman & Crane, 2018), weniger im Bereich der Anglistik und fast gar nicht in der Shakespeareforschung. Wenngleich das Thema umfangreich mit qualitativ-hermeneutischen Methoden untersucht wurde und wird (vgl. etwa den Sammelband von Maxwell & Rumbold, 2018), so finden sich nur wenige Arbeiten, die quantitative, digitale Methoden einsetzen, bspw. das Hyper-Hamlet-Projekt[5], bei dem eine große elektronische Datenbank mit Hamlet-Referenzen aufgebaut wurde (Hohl Trillini & Quassdorf, 2010).

Vor diesem Hintergrund exploriert der vorliegende Beitrag intertextuelle Bezüge auf das Werk Shakespeares in der britischen Fernsehserie Dr. Who und setzt dabei

auf computergestützte Analysemethoden, die bislang vor allem im Bereich klassischer Altertumswissenschaften und historischer Sprachen eingesetzt wurden. Dr. Who eignet sich dabei in besonderer Weise, da die TV-Serie fester Bestandteil der britischen Kultur ist, was automatisch eine größere Nähe zum Werk ihres Landsmannes Shakespeare bedeutet. Andererseits gibt es einige Folgen bei Dr. Who, bei denen bereits im Titel explizite Shakespeare-Bezüge deutlich werden[6], was auf die Identifikation weiterer Zitate und Referenzen in anderen Folgen hoffen lässt. Unser Beitrag versteht sich im Wesentlichen als Fallstudie anhand derer überprüft werden soll, ob sich Standard-Methoden der *text reuse detection* prinzipiell auch für eine vergleichende Analyse von Shakespeare und TV-Serien eignen. Gleichzeitig soll am Beispiel von Dr. Who exemplarisch untersucht werden, welche Art und Menge von intertextuellen Shakespeare-Bezügen im Kontext von TV-Serien zu erwarten sind, um darauf aufbauend den computergestützten Ansatz zu optimieren und weitere Filme und Serien zu analysieren. Übergeordnetes und langfristiges Ziel ist es demnach, systematisch die Referenzierung Shakespeares in Film und Fernsehen mithilfe computergestützter Methoden zu erfassen und so die intertextuelle und intermediale Durchdringung von Shakespeares Werk über die klassischen Literaturgenres hinaus zu quantifizieren und zu kartieren.

Korpora: Shakespeare vs. Dr. Who

Im Rahmen unserer Fallstudie werden systematisch Referenzen aus zehn der bekanntesten Shakespeare-Stücke (*Hamlet*, *Macbeth*, *Merchant of Venice*, *Midsummer Night's Dream*, *Much Ado About Nothing*, *Julius Caesar*, *The Tempest*, *Romeo and Juliet*, *King Lear*, *Othello*) sowie aus 154 kurzen Sonetten identifiziert werden. Die Transkripte dieser Werke stammen von der Website *Open Source Shakespeare*[7] und umfassen etwa 244.000 Tokens. Gesucht werden die Shakespearebezüge in einem Vergleichskorpus, welches aus den transkribierten Dialogen der TV-Serie Dr. Who besteht. Doctor Who ist eine der populärsten Science Fiction-Serien der britischen Geschichte und wird seit 1969 vom BBC produziert. Mittlerweile umfasst die Serie 840 Folgen, verteilt auf über 36 Staffeln und 13 verschiedene Doktoren. Für die vorliegende Studie wurden die Serientranskripte der ersten fünf Doktoren analysiert. Die Transkripte wurden von der Fan-Seite *Chakoteya*[8] heruntergeladen und umfassen insgesamt 141 Folgen mit einem Gesamtumfang von etwa zwei Millionen Tokens. Die Transkripte enthalten einerseits die Dialoge und andererseits Metadaten zu den jeweiligen Szenen, die vergleichbar mit Bühnenanweisungen bei Dramen sind.

Methodik: Suche nach local alignments und keywords

Für die Identifikation von textuellen Übereinstimmungen und Textähnlichkeiten (*text reuse*) findet sich ein breites Spektrum an computergestützten Methoden, die im Überblickspapier von Bär et al. (2012) grundlegend hinsichtlich (1) inhaltlicher, (2) struktureller und (3) stilistischer Ähnlichkeit kategorisiert werden. In dieser Studie verwenden wir den Smith-Waterman-Algorithmus (Smith & Waterman, 1981), ein Standard-Verfahren zur Identifikation von *local alignments*, also textuellen Übereinstimmungen auf der Inhaltsebene, das u.a. auch im Bereich der Bioinformatik für die Analyse von Genomsequenzen eingesetzt wird. Dabei werden auch graduelle Unschärfen wie etwa *gaps*, also einzelne fehlende Worte in ansonsten größtenteils identischen Wortgruppen, berücksichtigt. Zur Analyse der Korpora wurden die R-Packages *TextMining*[9] für das *Preprocessing* der Daten und *TextReuse*[10] als Implementierung für den Smith-Waterman-Algorithmus verwendet. Beim *Preprocessing* wurde schnell deutlich, dass bspw. eine Stoppwortentfernung nicht sinnvoll ist, da diese teilweise sehr charakteristisch für Shakespeare-Zitate sind (Bsp.: „to be or not to be"). Auch von Stemming und Lemmatisierung wurde abgesehen, da dies zu sehr vielen falsch-positiven Treffern führt, die dann aufwendig durch manuelle Kontrolle ausgefiltert werden müssen.

Weiterhin wurde das Shakespeare-Korpus in jeweils überlappende 9-Gramme (*shingling*), also jeweils Gruppen von neun Wörtern zerlegt. Für jedes einzelne 9-Gramm wurde dann im Dr. Who-Korpus mithilfe des im TextReuse-Package implementierten Smith-Waterman-Algorithmus das optimale *local alignment*, also der beste Treffer ermittelt. Dabei ist erwähnenswert, dass für jedes einzelne Shakespeare-9-Gramm immer ein optimales *alignment* im Dr. Who-Vergleichskorpus berechnet wird, auch wenn dieses ggf. nur aus einem oder zwei gemeinsamen Wörtern besteht. Der Vorgang ist somit einerseits sehr rechenintensiv und liefert andererseits eine große Menge von *alignments*, deren ausschließlich manuelle Kontrolle enorm aufwendig wäre. Dabei ist offensichtlich, dass die Verwendung einzelner, noch dazu sehr allgemeiner Wörter (Artikel, Konjunktionen, etc.) längst keinen intertextuellen Bezug darstellt. Wir betrachten deshalb nur solche *alignments* näher, die aus mindestens drei oder mehr Wörtern bestehen.

Gleichzeitig sind allerdings sehr kurze Referenzen, bspw. die Nennung einer Figur wie „Othello" oder „Macbeth", nicht ausgeschlossen, gehen aber bei diesem Ansatz verloren. Wir ergänzen den *local alignment*-Ansatz deshalb um eine *keyword*-Suche, bei der wir eine manuell erstellte Liste mit etwa 140 Stichwörtern und sehr kurzen Phrasen verwendeten. Die Liste enthält alle *dramatis personae* mit eindeutigen Namen (bspw. Ophelia, Romeo, Caliban, etc.), die Titel aller Stücke und Apokryphen sowie allgemeine Referenzen zu Shakespeares Person und Biographie (bspw. Shakespeare, Shaksper, The Bard, Stratford upon Avon, Anne Hathaway, etc.).

Ergebnisse und Diskussion

Insgesamt wurden mit der *keyword*-Suche 201 Treffer im Dr. Who-Korpus identifiziert, wobei ein Großteil der Treffer in einer einzelnen Folge vorkommen, da hier zwei der Hauptfiguren Troilus (108) und Cressida (37) heißen (vgl. Tabelle 1).

Beispielreferenz: "TROILUS: I'm sorry, **Cressida** , but I must obey orders."

Bei genauerer Überprüfung der Treffer wird schnell klar, dass Figurennamen aus Shakespeares Historiendramen häufig auch in der ursprünglichen Bedeutung der zugrundliegenden historischen Figur zitiert werden (bspw. *Horatio [Nelson]*, *King John*, *Julius Caesar*, *Pericles*) und somit

keine genuine Shakespeareerereferenz vorliegt. Nach Entfernung dieser falsch-positiven Treffer verbleiben 185 echte Shakespeareerereferenzen. Hier muss allerdings angemerkt werden, dass die von uns verwendeten Transkripte neben den eigentlichen Dialogen auch umfangreiche Kommentare und Metadaten zu den Stücken enthalten, die häufig zusätzliche Shakespeare-Keywords enthalten. So kommt man ausschließlich über die Dialoge auf 77 Treffer, in den Metadaten finden sich zusätzliche 108 Treffer.

Staffel	Keywords	Matches insgesamt	Treffer Dialoge	Treffer Metadaten	Align-ments
Doctor 1	Troilus (108), Cressida (37), Shakespeare (13), Hamlet (5), Falstaff (2), Romeo and Juliet (1), Julius Caesar (1)	167	68	99	1
Doctor 2	Shakespeare (1), Richard III (1)	2	0	2	0
Doctor 3	*Horatio (2), King John (1)*	3	0	0	1
Doctor 4	Shakespeare (8), Hamlet (1), Henry V (1), Juliet (1), *Julius Caesar (1)*, Macbeth (1), Romeo (1), sonnets (1)	15	8	6	5
Doctor 5	*King John (11)*, Hamlet (1), Shakespeare (1), *Pericles (1)*	14	1	1	1
Gesamt (141 Einzelfolgen)		*201*	77	108	8

Tabelle 1 : Überblick zu den gefundenen Keywords und *alignments* im Dr. Who-Korpus. Kursiv gesetzte Treffer sind falsch-Positive, d.h. nach manueller Prüfung stellte sich heraus, dass sie keine unmittelbare Shakespeareerereferenz beinhalten.

Auffällig niedrig scheint im Vergleich zu den *keyword*-Treffern die Anzahl längerer *alignments* zu sein: So finden sich insgesamt nur acht längere Shakespeare-Zitate in unserem Dr. Who-Korpus, bspw.:

(1) "Friends, Romans, countrymen! Lend me your ears. I come to bury Caesar, not to praise him."
(Julius Caesar, Akt 3, Szene 2)

(2) "To be or not to be." (Hamlet, Akt 3, Szene 1)

(3) "By the pricking of my thumbs, something wicked this way comes." (Macbeth, Akt 4, Szene 1)

Gleichzeitig erhalten wir mit dem local aligment-Ansatz auch viele falsch-positive Treffer, die zwar längere, in beiden Texten vorkommende Sequenzen beschreiben, aber keine genuinen Shakespeare-Zitate sind, sondern eher hochfrequente Idiome, bspw.:

"You all know what you have to do." (kommt gleichermaßen bei Shakespeare und Dr. Who vor)

Methodenreflexion und Fazit

Der zweigleisige Ansatz einer *local alignment*- sowie einer gesonderten *keyword*-Suche bringt mit 193 intertextuellen Shakespearebezügen in unserem Dr. Who-Korpus eine beeindruckende Anzahl von Treffern, deren Identifikation manuell was Umfang und Arbeitsaufwand angeht, so nicht ohne Weiteres möglich gewesen wäre. Damit ergeben sich spannende Perspektiven für die Erweiterung des Ansatzes auf größere Korpora, mit weiteren Serien und Filmen. Gleichwohl muss angemerkt werden, dass der Löwenanteil der Treffer über den relativ trivialen Ansatz der *keyword*-Suche gefunden wurde. Von den insgesamt 193 Treffern stammen nur 8 aus dem *local alignment*-Ansatz. Bedenkt man den enormen Rechenaufwand (fast 200 Stunden mit Standard-Laptops) und den Aufwand der manuellen Filterung der *alignments*, so muss man den *local alignment*-Ansatz in dieser Form doch deutlich hinterfragen. Denkbar wären hier für künftige Experimente weitere Optimierungen auf algorithmischer Seite, bspw. die Berücksichtigung von Wortarten-Ngrammen oder die Verwendung alternativer Algorithmen aus dem Bereich der *content similarity* (vgl. Bär et al., 2012). Weitere Verbesserungen sind auf Seite der Ergebnispräsentation (vgl. Burghardt & Wolff, 2015) möglich, welche die Beurteilung der Treffer deutlich erleichtern und damit beschleunigen könnte. Sinnvoll scheint auch eine Vorfilterung durch die *keyword*-Suche, d.h. nur diejenigen Episoden in denen bereits ein Shakespeare- *keyword* erkannt wurde werden anschließend auch noch mit dem *local alignment*-Ansatz auf komplexere Zitate hin untersucht. Als nächste Schritte sollen deshalb systematisch Korrelationen zwischen komplexen *alignments* und einzelnen *keywords* in einem größeren Korpus untersucht werden.

Fußnoten

1. Für eine Einführung zu „Intertextualität" siehe Allen (2000).
2. Untertitel online verfügbar unter https://www.opensubtitles.org/de (alle URLs in diesem Artikel wurden zuletzt am 14.10.2018 überprüft)
3. > Drehbücher online verfügbar unter https://www.imsdb.com/
4. Fan-Transkripte online verfügbar unter http://transcripts.foreverdreaming.org/, https://www.springfieldspringfield.co.uk/, http://www.chakoteya.net/
5. HyperHamlet online verfügbar unter http://www.hyperhamlet.unibas.ch/
6. Bspw. „The Shakespeare Code", Dr. Who Folge 180, Transkription online verfügbar unter http://www.chakoteya.net/DoctorWho/29-2.htm
7. Open Source Shakespeare online verfügbar unter https://www.opensourceshakespeare.org/
8. Chakoteya Fan-Transkripte online verfügbar unter http://www.chakoteya.net/DoctorWho/index.html
9. Dokumentation online verfügbar unter https://cran.r-project.org/web/packages/tm/index.html
10. Dokumentation online verfügbar unter https://cran.r-project.org/web/packages/textreuse/index.html

Bibliographie

Allen, G. (2000): *Intertextuality*, London, New York: Routledge.

Bamman, D. / Crane, G. (2008) *The Logic and Discovery of Textual Allusion*, in: ACL Language Technology for Cultural Heritage, (1986).

Bär, D. / Zesch, T. / Gurevych, I. (2012): *Text Reuse Detection using a Composition of Text Similarity Measures*, in: Proceedings of COLING 2012, 1 (December), 167–184.

Berti, M. / Büchler, M. / Geßner, A. / Thomas, E. (2013): *Measuring the Influence of a Work by Text Reuse*, in: The Digital Classicist.

Burghardt, M. / Wolff, C. (2015): *Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik*, in: Book of Abstracts Workshop "Informatik und die Digital Humanities", Leipzig.

Büchler, M. / Burns, P. R. / Müller, M. / Franzini, E. / Franzini, G. (2014): *Towards a Historical Text Re-use Detection*, in **C. Biemann / A. Mehler (Eds.):** *Text Mining, Theory and Applications of Natural Language Processing*, Springer International Publishing.

Forstall, C. / Coffee, N. / Buck, T. / Roache, K. / Jacobson, S. (2015): *Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching*, in: Digital Scholarship in the Humanities, 30 (4), 503–515.

Hohl Trillini, R. / Quassdorf, S. (2010): *A "key to all quotations"? A corpus-based parameter model of intertextuality*, in: Literary and Linguistic Computing, 25 (3), 269–286.

Klarer, M. (1999): *An Introduction to Literary Studies*, London, New York: Routledge.

Kozloff, S. (2000): *Overhearing Film Dialogue*, Berkeley et al.: University of California Press.

Malone, T. (2018): *Quoting Shakespeare in Twentieth-Century Film*, in: **Maxwell, J. / Rumbold, K. (Eds.):** *Shakespeare and Quotation* (pp. 194–208). Cambridge: Cambridge University Press.

Maxwell, J. / Rumbold, K. (2018): *Shakespeare and Quotation*, Cambridge: Cambridge University Press.

Scheirer, W. / Forstall, C. / Coffee, N. (2016): *The sense of a connection: Automatic tracing of intertextuality by meaning*, Digital Scholarship in the Humanities, 31 (1), 204–217.

Smith, T. F. / Waterman, M. S. (1981): *Identification of Common Molecular Subsequences*, Journal of Molecular Biology, 147, 195–197.

Taylor, G. (1991): *Reinventing Shakespeare: A Cultural History from the Restoration to the Present*, Oxford: Oxford University Press.

Visualisierung zwischen Pluralismus und Fragmentierung: Zur Integration von multiplen Perspektiven auf kulturelle Sammlungen

Mayr, Eva
eva.mayr@donau-uni.ac.at
Donau Universität Krems, Österreich

Windhager, Florian
florian.windhager@donau-uni.ac.at
Donau Universität Krems, Österreich

Schreder, Günther
guenther.schreder@donau-uni.ac.at
Donau Universität Krems, Österreich

Einleitung

Die Digitalisierung der Sammlungen zahlreicher Museen, Archive und anderer Kulturinstitutionen ermöglicht den raschen Zugang zu historisch bedeutsamen Kulturgütern: Millionen von Bildern, Skulpturen, Musik- und Schriftstücken sind heutzutage mit nur wenigen Klicks als digitale Objekte erreichbar. Allerdings gelangen klassische Such-Interfaces rasch an ihre Grenzen, wenn das Ziel die freie Exploration und die Gewinnung eines effektiven Überblicks über eine kulturelle Sammlung sind. Bei hunderten von Objekten kann eine Sammlung in den üblichen Listen- und Rasteranordnungen kaum vollständig erfasst werden. Darüber hinaus stellt die produktive und aufschlussreiche Anordnung der Exponate auf dem Bildschirm eine konzeptuelle Herausforderung dar, da viele Arrangements möglich sind und die kulturellen Sammlungen eine Vielzahl von Metadaten-Dimensionen aufweisen. Wie kann man die Darstellung von solchen multidimensionalen Datenbanken und den darin befindlichen Objekten und damit die Gewinnung eines Überblicks über kulturelle Sammlungen verbessern?

Informationsvisualisierungen von kulturellen Sammlungen

In den letzten Jahren wurden vermehrt Informationsvisualisierungen entwickelt (vgl. Windhager, Federico et al., 2018 für einen Überblick), um jenseits von Listen und Rastern auch einen konzeptuellen Überblick über den Aufbau und die Struktur kultureller Sammlungen zu ermöglichen. Doch um einen "generösen" Zugang (Whitelaw, 2015) zu einer reichhaltigen Sammlung zu ermöglichen reicht eine Visualisierung alleine nicht aus (Doerk et al., 2017). In einer Analyse bestehender Informationsvisualisierungen von kulturellen Sammlungen (Windhager, Federico et al., 2018) zeigte sich, dass die Interfaces im Durchschnitt 2-3 - manche sogar bis zu 6 - verschiedene Visualisierungen einsetzen, um der informationellen Reichhaltigkeit (und damit auch der multidimensionalen kuratorischen und pädagogischen Komplexität) einer kulturellen Sammlung gerecht zu werden.

Doch solch eine Vielzahl von visuellen und konzeptionellen Perspektiven erzeugt neue Herausforderungen für die Besucher digitaler Sammlungen: Wenn Informationen über mehrere Darstellungen verteilt sind, wird es schwierig einen Gesamtverständnis der kulturellen Sammlung aufzubauen. Einzelne Visualisierungen geben jeweils nur einen spezifischen Blickwinkel auf die Sammlung (z.B. geographische Verteilung, historische Entwicklung, usw.). Es besteht die Gefahr, dass die einzelnen Blickwinkel nicht miteinander verknüpft werden und damit nur ein fragmentiertes Verständnis der Sammlung aufgebaut wird. Aber wie können die Besucher dabei unterstützt werden die Informationen in den verschiedenen Visualisierungen und Ansichten am besten zu verarbeiten und miteinander zu integrieren und so ein multidimensionales Verständnis - ein besser integriertes mentales Modell - der Sammlung aufzubauen? Dieser Frage widmete sich in den letzten 3 Jahren das Forschungsprojekt "PolyCube - Towards integrated mental models of cultural heritage data" (http://donau-uni.ac.at/de/polycube /).

Multiperspektivität und Informationsintegration in PolyCube

In PolyCube wurden multiperspektivische Informationsvisualisierungen von kulturellen Sammlungen entwickelt, die die Integration von Informationen auf verschiedenen Ebenen unterstützen: (1) Integration von abstrakten Überblicksdarstellungen auf Sammlungsebene ("distant reading" oder "distant viewing") mit Detaildarstellungen von Objekten ("close reading" oder "close viewing"), (2) Integration von mehreren Datendimensionen innerhalb einer Visualisierung, und (3) Integration von mehreren Visualisierungen mit verschiedenen Perspektiven auf die kulturelle Sammlung.

Überblick, Details, und Navigation als Integration

Dörk et al. (2011) betonen, dass auch in digitalen Informationsräumen die Benutzer "durch die Ausstellung flanieren" wollen auf der Suche nach interessanten Objekten, um sich anschließend in diesen zu vertiefen. Interfaces zu kulturellen Sammlungen ermöglichen zudem die fließende Navigation zwischen "Aggregationsebenen" einer Sammlung: Vom Überblick zum Einzelobjekt und vice versa (vertikale Immersion oder Abstraktion) - sowie die traditionelle Bewegung von Objekt zu Objekt ("horizontales Browsing"). Um dieses kinetische Spektrum abzudecken, können in den PolyCube-Visualisierungen einzelne Datenpunkte ausgewählt werden, um Detailansichten neben der Überblicksdarstellung einzublenden. So bleibt die Orientierung auf Sammlungsebene erhalten, während die Einzelobjekte in Detailansicht studiert und durchwandert werden können.

Multidimensionale Informationsvisualisierungen zur Integration von Zeit

PolyCube geht über eindimensionale (z.B. Zeitstrahlen) oder zweidimensionale Visualisierungen (z.B. Karten oder Netzwerkdarstellungen) hinaus und integriert die für kulturelle Sammlungen äußerst wichtige Zeitdimension auf multiple Weise in zweidimensionale Ansichten (vgl. Abbildung 1). In einer experimentellen Studie haben wir untersucht, welche von vier verschiedenen geo-temporal integrierten Visualisierungen Benutzer am besten dabei unterstützen multidimensionale Erkenntnisse über die kulturelle Sammlung zu gewinnen (Mayr et al., 2018): Die Ergebnisse belegen, dass die Techniken der Farbkodierung und des Raum-Zeit-Kubus das integrierte Verständnis von raum-zeitlichen Mustern in der Sammlung (z.B. Reisebewegungen des Künstlers) am besten unterstützen. Die anderen beiden Visualisierungen erschweren die Integration, da bei "coordinated multiple views" Informationen über größere visuelle Distanzen verknüpft und bei Animationen die vorangegangenen Datenpunkte im Arbeitsgedächtnis gespeichert werden müssen.

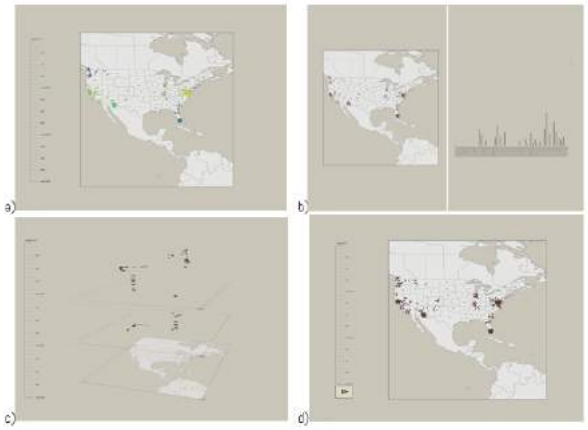


Abbildung 1. Multiple Möglichkeiten der raum-zeitlichen Visualisierung einer kulturellen Sammlung im PolyCube System. Zeitinformation wird a) als Farbkodierung, b) als "coordinated multiple view", c) als Zeitachse in einem Raum-Zeit-Kubus und d) als Bewegung in einer Animation enkodiert.

Verknüpfung mehrerer Visualisierungen

Doch oftmals sind es mehr als drei Dimensionen, die eine kulturelle Sammlung charakterisieren und daher von Interesse für die Benutzer einer Visualisierung sind. Im PolyCube-System kann Zeit in Zusammenschau mit geographischem, kategorialem und relationalem Raum dargestellt werden (vgl. Abbildung 2). Damit diese jedoch nicht nur isoliert voneinander betrachtet werden, bedarf es verschiedener Techniken, die deren Verknüpfung unterstützen: (a) Übergangsanimationen, (b) kohärenter Visualisierungstechniken, und (c) Koordination der Visualisierungen

Zunächst empfiehlt es sich, mehrere Visualisierungen mittels Übergangsanimationen ("seamless transitions") zu verknüpfen, die die Transformation der Datenpunkte von einer zur anderen Visualisierung veranschaulicht. Egal, ob die Visualisierungen seriell oder parallel präsentiert werden, wird dadurch die Zusammenführung der Informationen aus beiden Visualisierungen erleichtert. Kognitiv bedeutet dies, das zunächst nur ein mentales Modell von der ersten Visualisierung aufgebaut wird und die weiteren Informationen aus den folgenden Visualisierungen damit inkrementell verknüpft werden können (vgl. Schreder et al., 2016).

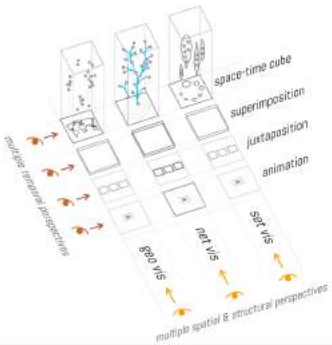


Abbildung 2. Das PolyCube System kombiniert multiple strukturelle (orange) and zeitliche Visualisierungstechniken (braun) für die reichhaltige, kontextuelle Visualisierung von Sammlungen des kulturellen Erbes.

Die Zuordnung und Interpretation der Informationen aus mehreren Visualisierungen wird darüber hinaus vereinfacht, wenn in allen die gleichen oder konsistente Gestaltungsprinzipien angewandt werden (z.B. konsistente Ausrichtung der Zeitachse, konsistentes Farbschema, konsistente Nutzung des Raum-Zeit-Kubus) (Qu & Hullman, 2018). Aus kognitiver Perspektive führen ähnliche visuelle Hinweisreize dazu, dass das System als ein Ganzes gesehen wird und daher auch die Information verknüpft und gemeinsam verarbeitet wird.

Parallel präsentierte Visualisierungen derselben Sammlung sollten nicht lose nebeneinander stehen, sondern als " coordinated multiple views " auch visuell miteinander koordiniert werden: Methoden wie linked highlighting, linking und brushing oder leader lines erzeugen visuelle Verbindungen zwischen den verschiedenen Repräsentationen derselben Datenpunkte in den verschiedenen Visualisierungen. Das ermöglicht es, die Positionierung eines Objektes im geographischem, relationalem und kategorialem Raum gemeinsam zu verstehen und kognitiv miteinander zu verknüpfen.

Diskussion und Ausblick

Generosität und perspektivischer Pluralismus, wie sie in PolyCube beispielhaft für kulturelle Sammlungen entwickelt wurden, können als Designstrategien von besonderer Relevanz auch für andere geistes- und kulturwissenschaftliche Gegenstände und Themen gelten, die immer schon durch Reichhaltigkeit (an informationellen aber auch interpretativen Dimensionen) ausgezeichnet sind. Mit dem PolyCube-Projekt wollen wir für die Designer von komplexen Visualisierungsumgebungen auch Möglichkeiten aufzeigen, wie kognitive und interpretative Folgekosten von dissonanten, inkohärenten oder fragmentierten Ansichten reduziert werden können (Windhager, Salisu et al., 2018). Wir plädieren neben der Bereitstellung von Technologien der "visuellen Analyse" in diesem Kontext für die Entwicklung eines neuen Instrumentariums der "visuellen Synthese" (Schreder et al., 2016), dass auch im Rahmen von komplexem Interface-Design den vermittelten Blick auf *Bäume und Wald* möglich macht.

Neben der Präsentation und Diskussion der entsprechenden Integrationstechniken wird ein Ausblick der zukünftigen Öffnung des Visualisierungssystems für ForscherInnen zu kulturellen Sammlungen gewidmet: Mit einem einfachen Spreadsheet-Editor können eigene Daten importiert werden, um die multi-perspektivische Exploration von beliebigen neuen Sammlungen von zeit-orientierten Kulturdaten zu gewährleisten.

Danksagung

Die beschriebene Arbeit wurde durch den Wissenschaftsfonds FWF P.No. P28363 gefördert.

Bibliographie

Dörk, M. / Carpendale, S. / Williamson, C. (2011): *The information flaneur: A fresh look at information seeking*, in: Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1215-1224). ACM.DOI: https://doi.org/10.1145/1978942.1979124

Dörk, M. / Pietsch, C. / Credico, G. (2017): *One view is not enough*. Information Design Journal, 23 (1), 39-47.

Mayr, E. / Schreder, G. / Salisu, S. / Windhager, F. (2018): *Integrated Visualization of Space and Time: A Distributed Cognition Perspective*. Manuscript in preparation.

Qu, Z / Hullman, J. (2018): *Keeping Multiple Views Consistent: Constraints, Validations, and Exceptions in Visualization Authoring*, IEEE Transactions on Visualization and Computer Graphics, 24 (1), p. 468-477. DOI: https://doi.org/10.1109/TVCG.2017.2744198

Schreder, G. / Windhager, F. / Smuc, M. / Mayr, E. (2016): *A Mental Models Perspective on Designing Information Visualizations for Political Communication.* JeDEM-eJournal of eDemocracy & Open Government, 8(3), 80-99.

Whitelaw, M. (2015): *Generous Interfaces for Digital Cultural Collections*. DHQ: Digital Humanities Quarterly, 9 (1).

Windhager, F. / Federico, P. / Schreder, G. / Glinka, K. / Dörk, M. / Miksch, S. / Mayr, E. (2018): *Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges*. IEEE Transactions on Visualization and Computer Graphics. DOI: 10.1109/TVCG.2018.2830759

Windhager, F. / Salisu, S. / Schreder, G. / Mayr, E. (2018): *Orchestrating Overviews: A Synoptic Approach to the Visualization of Cultural Collections.* Open Library of Humanities, 4 (2). DOI: http://doi.org/10.16995/olh.276

Vom Bild zum Text und wieder zurück

Donig, Simon
simon.donig@uni-passau.de
Universität Passau, Deutschland

Christoforaki, Maria
maria.christoforaki@uni-passau.de
Universität Passau, Deutschland

Bermeitinger, Bernhard
bernhard.bermeitinger@uni-passau.de
Universität Passau, Deutschland

Handschuh, Siegfried
siegfried.handschuh@unisg.ch
Universität St. Gallen, Schweiz

In den letzten Jahren hat die Anwendung von Verfahren der Computer Vision im Bereich der digitalen Kunstgeschichte und Objektforschung erheblich an Bedeutung gewonnen (Donig, Handschuh, Hastik, Kohle, Ommer, Radisch, Rehbein 2018). Dabei stellt das Schließen der semantischen Lücke eine zentrale Herausforderung für (teil-)automatisierte algorithmische Verfahren dar. Hier schlagen wir einen multimodalen Zugang vor, in dem wir eine fruchtbringende Lösung des Problems sehen und den wir im Kontext des Neoclassica-Projekts entwickeln.

Neoclassica ist ein Rahmenwerk zur Erforschung der ästhetischen Kultur des Klassizismus (ca. 1760-1860), das Methoden und Instrumente zur Erforschung von Architektur und Raumkunst bereitstellt (Donig, Christoforaki, Bermeitinger, Handschuh 2017). Dazu bedient es sich eines Ansatzes des Wissensrepräsentation in der Form einer eigenen Ontologie (Donig, Christoforaki, Handschuh 2016) sowie datengetriebener Forschungsinstrumente aus dem Bereich der künstlichen Intelligenz, hier insbesondere der Klassifizierung von Bildern und der semantischen Segmentierung von Bildinhalten mit Verfahren des Deep Learning (Donig, Christoforaki, Bermeitinger, Handschuh 2018).

Algorithmische Werkzeuge bedürfen qualitativer Metriken, um ihre Verlässlichkeit und Reproduzierbarkeit abzubilden. Was aber, wenn die Klassifizierung nicht auf einer Serie flacher Label beruht, sondern wenn die Grundlage für die Klassifizierung komplexe Konzepte sind, die durch semantische Hierarchien verbunden werden wie im Fall einer Annotation von Bilddaten mit einer Ontologie?

Bei unserer Arbeit an Neoclassica sind wir diesem Problem wiederholt begegnet. Wenn wir zum Beispiel einen Armlehnstuhl (Abb.1) in einem Bildwerk annotiert haben, dann weist das Konzept ohne Zweifel Gemeinsamkeiten mit dem eines Stuhls auf. In der Ontologie wird dieser Umstand dadurch ausgedrückt, dass *Stuhl* eine übergeordnete Klasse zur Klasse *Armlehnstuhl* ist (Abb.2).

Abb.1 Die Bounding-box um das Möbel spiegelt die Konfidenzrate des Algorithmus wieder.

Während für einen menschlichen Beobachter ein Armlehnstuhl eine spezielle Unterkategorie von Stühlen darstellt, ist für die von uns genutzten Algorithmen diese Zuordnung dagegen falsch - bezogen auf die von uns ursprünglich vorgenommenen Annotationen-, da sie die elementare semantische Beziehung zwischen den Konzepten nicht berücksichtigen.

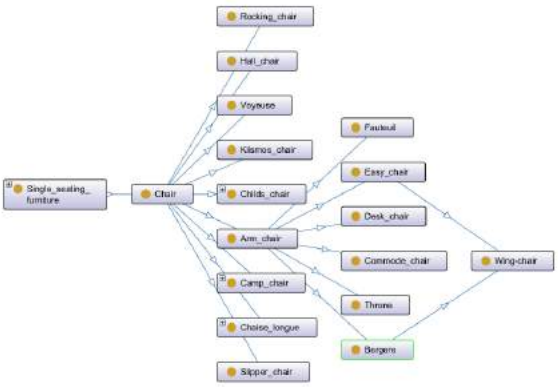


Abb.2 Neoclassica-Ontologie: Die *Chair* -Subhierarchie

Ein vergleichbares Problem stellt der Umgang mit Geschwisterklassen einer formalen Wissensrepräsentation dar. Armlehnstühle mit offenen (*Fauteuil*) und geschlossenen (*Bergère*) Armlehnen sind sich in vielen Aspekten ähnlich. Dennoch ist dem Klassifizierungsprozess dieses Verwandtschaftsverhältnis zunächst nicht zu eigen (Abb.3).

Abb.3 Geschwisterklasse mit hoher Konfidenzrate klassifiziert

Vorgehensweise

Wir stellen hier einen multimodalen Zugang vor, der einen Ansatz aus dem NLP, einem Bereich, wo solche Beziehungen schon lange eingehend studiert worden sind (Indurkhya, Damerau 2010: 120), (Miller 1995), mit einem Ansatz der Computer Vision verbindet - dem Deep Learning visueller Merkmale.

Dieser Zugang beruht auf der *distributional hypothesis*, die postuliert, dass eine Korrelation zwischen der Verteilung von Wörtern und ihrer semantischen Eigenschaften in einem Textkorpus besteht (Rubenstein & Goodenough 1965), was erlaubt, mit Hilfe ersterer die zweiten abzuschätzen (Sahlgren, 2008). Dies schließt somit auch Generalisierungen und Spezialisierungen o.ä. zwischen verschiedenen Klassen ein.

Die ausführlichste systematische Anwendung der Verteilungshypothese findet sich in Distributional Semantic Models (DSMs), die einen multidimensionalen Vektorraum bilden, in dem Wörter als Vektoren abgebildet werden (Lenci, 2018). Diese Vektoren bilden die Kookkurrenz eines Worts mit anderen Wörtern in einem Textkorpus ab, nähern sich so einem Kontext bzw. der Bedeutung dieses Wortes an. Diesen Prozess, in dem ein Wort auf einen 3Vektor abgebildet wird, bezeichnen wir als *word embeddings* (Mikolov, Chen, Corrado, Dean 2013), (Collobert, Weston 2008). Der Grad semantischer Nähe von zwei Wörtern kann 3durch die Anwendung mathematischer Formeln auf diese Vektoren repräsentiert werden (Budanitsky, Hirst 2006).

Für den hier vorgeschlagenen Beitrag haben wir ein DSM beruhend auf einem domänenspezifischen Textkorpus erzeugt.[1] Dazu benutzen wir das an unserem Lehrstuhl entwickelte Indra-Framework, das die Erzeugung, Verwendung und Evaluierung von Word Embedding-Modellen unterstützt (Sales, Souza, Barzegar, Davis, Freitas, Handschuh 2018).

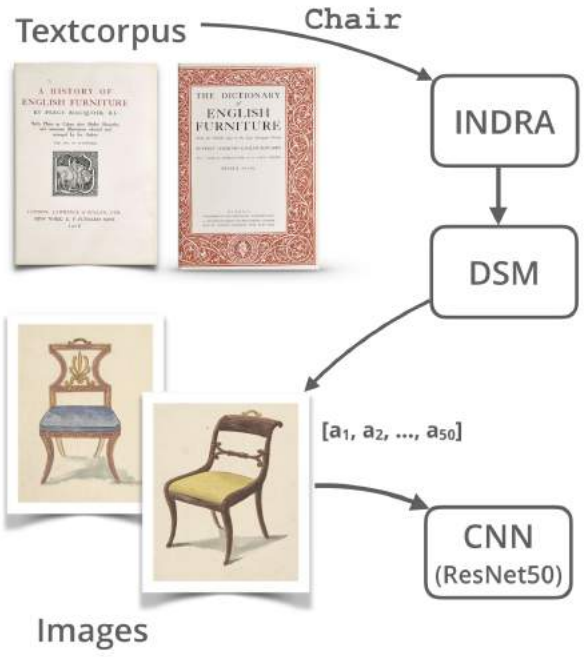


Abb.4 Trainingsprozess

Das DSM erlaubt es uns, Word Embeddings für die Klassen in unserem Neoclassica-Open-Korpus (Donig, Christoforaki, Bermeitinger, Handschuh 2018: 131;133) zu erstellen. Für den ersten Schritt dieses Experiments beschränken wir uns auf Bildwerke von einzelnen Objekten.

Anschließend trainieren wir ein Neuronales Netz (ResNet50 (He, Zhang, Ren, Sun 2015)) zur Bildklassifizierung statt mit herkömmlichen, flachen Labels mit den aus dem DSM hervorgehenden Vektoren. In Abb.4 illustrieren wir den Prozess, bei dem das Wort *Chair* mit dem Vektor [a1, a2, … a50] korrespondiert, der dann genutzt wird, um Bilder von Stühlen zu annotieren.

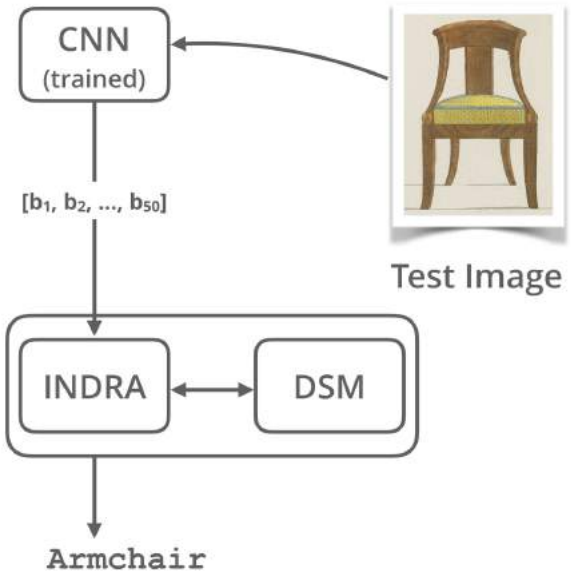


Abb.5 Testphase

Die Testphase wird in Abb.5 illustriert, dabei wird ein Testbild in das CNN eingespeist, dass es mit einem Vektor assoziiert.

Indra ermöglicht es uns nun, die nächsten Nachbarn dieses Vektors in Wörtern zu finden und diesen ein Text-Label zuzuordnen. Diese Relationen zwischen den Wörtern stellen zugleich eine semantische Beziehung der Bildinhalte her.

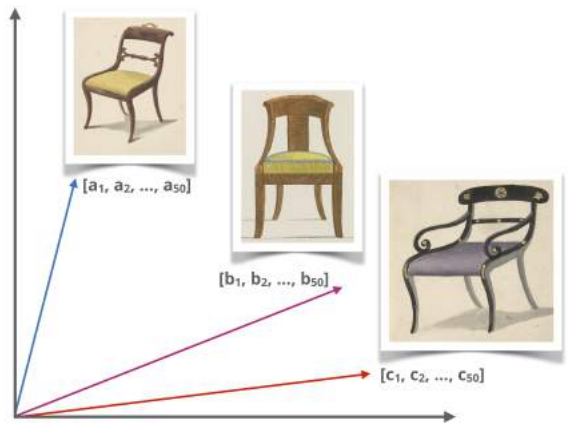


Abb.6 Idealtypische Repräsentation des reduzierten Vektorraums

Illustriert wird diese Beziehung in Abb.6, die eine idealtypische Visualisierung des 50-Dimensionalen Vektorraums, reduziert auf zwei Dimensionen, zeigt. Vektor [a_1, a_2, … a_{50}] entspricht dem Begriff *Stuhl,* wohingegen Vektor [c_1, c_2, … c_{50}] dem Begriff *Armlehnstuhl* entspricht. Das Testbeispiel eines Gondelstuhls entspricht Vektor [b_1, b_2, … b_{50}] , der größere semantische Nähe zu [c_1, c_2, … c_{50}] als zu [a_1, a_2, … a_{50}] aufweist, da der Rücken des Gondelstuhls einen armähnlichen Rahmen besitzt, der sich sanft nach vorne neigt und in die Armstützen übergeht.

Wir hoffen, dass diese semantische Beziehung zwischen mehreren Bildern natürlichsprachige Beziehungen zwischen den abgebildeten Artefakten besser reflektiert, als dies herkömmliche “simple” Klassifizierungsprozesse können.

Stand der Umsetzung & Teilergebnisse

Bildanalyse

Bislang haben wir Verfahren aus dem Bereich des Deep Learning eingesetzt, um Abbildungen einzelner Möbel (Bermeitinger, Donig, Christoforaki, Freitas, Handschuh 2017) sowie mit Möbelgruppen in Interieuransichten (Donig, Christoforaki, Bermeitinger, Handschuh 2018) zu klassifizieren. Wir konnten dabei zeigen, dass algorithmische Instrumente hervorragend in der Lage sind, Einzelobjekte zu identifizieren (0,94 aMP) - und dies, relativ unabhängig vom Vorliegen in einer bestimmten Technik und Materialität (Fotografie, Gemälde, Zeichnung, Druckgrafik). Für Darstellungen von Mobiliar in Interieurs können wir in unseren Experimenten immer noch gute Ergebnisse vorweisen (aMP 0.53; recall 0.51). Wie eine qualitative Analyse dieser Ergebnisse gezeigt hat, ist die Differenz zum vorausgegangenen Experiment nicht alleine auf die gestiegene Komplexität (z.B. hohe Zahl der Klassen, Überlappung von Objekten im Raum, generell Noise), sondern auch auf zahlreiche nominelle Fehlklassifizierungen zurückzuführen, die aus dem eingangs geschilderten Hierarchie-Problem resultieren.

Verteilungssemantik

Da es keine allgemeine Methode der Evaluierung eines domänespezifischen DSM gibt (Lenci, 2018), zeigen wir nachstehend, dass das Modell sinnvolle Ergebnisse produziert, wenn man diese mit Weltwissen sowie der Neoclassica-Ontologie vergleicht.

armchair :
['armchair', 'upholst',[2] 'sette', 'cane', 'mendlesham']

settee :
['sette', 'upholst', 'windsor', 'stool', 'armchair']

Da Begriffe mit sich selbst am nächsten verwandt sind, erscheinen sie an erster Stelle in der Begriffskette, was als ein Zeichen dafür gewertet werden kann, dass das DSM korrekt funktioniert. Beide Möbel gehören zu einer Klasse von gepolsterten Sitzmöbeln (‘*upholst*’) ; in einigen Fällen lagen Polsterungen auch lose auf einem Geflecht auf (‘*cane*’) . Weiter wird deutlich, dass es eine reziproke Beziehung zwischen beiden Begriffen gibt, denn sie referenzieren sich wechselseitig. Das Sofa weist in diesem Korpus außerdem eine enge Nachbarschaft zu einem weiteren Sitzmöbel, dem Hocker (‘stool’) auf. Insgesamt zeigen die Beispiele also bemerkenswerte semantische Nähe und Geschlossenheit.

Ein abschließendes Beispiel mag der Begriff des mehrarmigen Leuchters sein:

candelabra :
['candelabra', 'consol', 'torcher', 'girandol', 'candlestick']

Leuchter existieren in klassizistischen Interieurs für gewöhnlich in Paaren. Es macht daher Sinn, dass diese Leuchter auch im DSM als Mehrzahl auftreten ‘*candelabra*’. Für gewöhnlich stehen sie auf einem Möbel oder Kaminsims (daher ‘*consol*’ für einen Konsoltisch). In der Neoclassica-Ontologie hat die Klasse Candelabrum eine Reihe von ihr verwandten Klassen von Leuchtmitteln, die alle unabhängig von ihrem Vorliegen in der Ontologie auch innerhalb des DSM identifiziert worden sind.

Spezifisch sind dies die in der Ontologie auf einer Ebene angesiedelten Klassen *Candlestick*, *Torchere* und die etwas tiefer in der Hierarchie liegende *Girandole* (Abb.7).

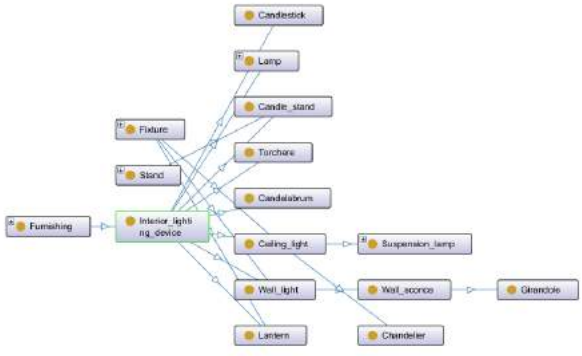


Abb.7 Neoclassica-Ontologie: Die *Interior_lighting_device*-Subhierarchie

An dieser Stelle ist noch einmal zu betonen, dass die Ordnung des DSM rein auf Statistik beruht und aus dem textuellen Korpus abgeleitet ist, während die hier

zum Vergleich herangezogene Ordnung der Ontologie eine menschengemachte Wissensrepräsentation ist.

Conclusum und nächste Schritte

Die vorliegende Kurzzusammenfassung unseres Vortrags schlägt ein Verfahren für eine Einbeziehung semantischer Kontextinformation in den Bildklassifizierungsprozess mit Deep Neural Networks vor. Das Verfahren entsteht im Rahmen des Neoclassica-Projekts und zielt insbesondere darauf, die Erkennung von Mobiliar in historischen Darstellungen von Innenräumen zu verbessern. Der multimodale Zugang wird, so die Hoffnung, dazu beitragen, Schwierigkeiten, denen wir in unserer bisherigen Arbeit begegnet sind - wie der Herausforderung unscharfer Konzepte oder dem Problem der Klassifizierung in semantischen Hierarchien - besser gerecht zu werden. Zukünftige Schritte werden sich auf drei Gebiete erstrecken. Erstens bedarf die Konstruktion eines domäne- und aufgabengerechten DSMs weiterer Verfeinerung. Es gilt zu evaluieren, ob der Umfang des Korpus für das beabsichtigte Ziel bereits ausreichend ist. Qualitätskriterien für eine Evaluierung müssen entwickelt werden, die nicht alleine nach NLP-Maßstäben, sondern auch im Domänezusammenhang sinnvoll sind. Zweitens gilt es eine adäquate Lösung für den Umgang mit zusammengesetzten Ausdrücken zu finden. (Zur Herausforderung semantischer Kompositionalität im Kontext des DSM vgl. Baroni et al., 2014).

Die vorläufigen Ergebnisse der beiden ersten Meilensteine in den Bereichen der Bildanalyse und der Transformation des Textkorpus in ein DSM geben uns die Hoffnung, dass die Einführung von Kontext in den Bildklassifizierungsprozess die *fuzzyness* des Domänegegenstands besser akkomodiert und damit letztlich auch zu einer Verbesserung der Trefferquote des Klassifikationsverfahrens beiträgt.

Fußnoten

1. Das Textkorpus umfasst 32 Quellen, die 1.987.544 Worte und 58.651 unique word forms repräsentieren. Es besteht aus englischsprachigen Fachpublikationen der Jahrhundertwende vom 19. zum 20. Jahrhundert (cf. Abschnitt Quellen im Literaturverzeichnis). Wir haben diese Texte ausgewählt, da sie stärker differenzierte Konzepte zur Beschreibung des Fachgebiets bieten und die Qualität der von uns durchgeführten optischen Zeichenerkennung (OCR) für diesen Zeitabschnitt deutlich höher war als für zeitgenössische Texte. Anders als moderne Fachtexte sind diese Publikationen zudem unter einer freien, permissiven Lizenz verfügbar.
2. Die Begriffe sind innerhalb des DSMs auf ihren Wortstamm zurückgeführt.

Bibliographie

Baroni, Marco / Bernardi, Raffaela / Zamparelli, Roberto (2014): *"Frege in Space: A Program of Compositional Distributional Semantics"*, LiLT (Linguistic Issues in Language Technology), 9: 241–346.

Bermeitinger, Bernhard / Donig, Simon / Christoforaki, Maria / Freitas, André / Handschuh, Siegfried (2017): *"Object Classification in Images of Neoclassical Artifacts Using Deep Learning."* Montreal, Canada. https://dh2017.adho.org/abstracts/590/590.pdf [Letzter Zugriff 25.09. 2018]

Bontempi, Gianluca (2017): *Handbook-Statistical Foundations of Machine Learning*. Bruxelles: Machine Learning Group Computer Science Department ULB Belgique.

Budanitsky, Alexander / Hirst, Graeme (2006): *"Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness"*, Computational Linguistics 32 (1): 13–47.

Collobert, Ronan / Weston, Jason (2008): *"A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning"*, in: Proceedings of the 25th International Conference on Machine Learning, 160–167.

Donig, Simon / Christoforaki, Maria / Handschuh, Siegfried (2016): *"Neoclassica - A Multilingual Domain Ontology. Representing Material Culture from the Era of Classicism in the Semantic Web"*, in: **Bozic, Bojan/Mendel-Gleason, Gavin/Debruyne, Christophe / O'Sullivan, Declan (eds.)**: *Computational History and Data-Driven Humanities*. CHDDH 2016 (=IFIP Advances in Information and Communication Technology, vol 482), Cham: Springer: 41–53, DOI 10.1007/978-3-319-46224-0_5.

Donig, Simon / Christoforaki, Maria /Bermeitinger, Bernhard/ Handschuh, Siegfried (2017): *"Neoclassica – an Open Framework for Research in Neoclassicism."* Montreal, Canada. https://dh2017.adho.org/abstracts/384/384.pdf [Letzter Zugriff 25. 09. 2018]

Donig, Simon / Christoforaki, Maria / Bermeitinger, Bernhard / Handschuh, Siegfried (2018): *"Bildanalyse durch Distant Viewing - zur Identifizierung von klassizistischem Mobiliar in Interieurdarstellungen"*, in: **Vogeler, Georg (ed.)**: *DHd 2018 - Kritik der digitalen Vernunft*. Köln: 130–137.

Donig, Simon / Handschuh, Siegfried / Hastik, Canan / Kohle, Hubertus / Ommer, Björn / Rehbein, Malte (2018): *"Der ferne Blick. Bildkorpora und Computer Vision in den Geistes- und Kulturwissenschaften - Stand - Visionen - Implikationen"*, in: **Vogeler, Georg (ed.)**: *DHd 2018 - Kritik der digitalen Vernunft*. Köln: 86–89.

He, Kaiming / Zhang, Xiangyu / Ren, Shaoqing /Sun, Jian (2016): *"Deep residual learning for image recognition"*, in: Proceedings of the IEEE conference on computer vision and pattern recognition: 770-778.

Indurkhya, Nitin / Damerau, Fred J. (2010): *Handbook of Natural Language Processing. Second Edition. Vol. 2. Machine Learning & Pattern Recognition Series*. Boca Raton, FL: Chapman & Hall/CRC Taylor & Francis Group.

Lenci, Alessandro (2018): *"Distributional models of word meaning"*, in: Annual review of Linguistics, 4 (1) :151-171.

Miller, George A. (1995): *"WordNet: A Lexical Database for English"*, in: Communications of the ACM 38 (11): 39–41.

Sales, Juliano Efson / Souza, Leonardo / Barzegar, Siamak / Davis, Brian / Freitas, André / Handschuh, Siegfried (2018): *"Indra: A Word Embedding and Semantic Relatedness Server."* In LREC. Miyazaki, Japan, 2018.

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): *"Efficient Estimation of Word Representations in Vector Space."* ArXiv:1301.3781 [Cs]. http://arxiv.org/abs/1301.3781. [Letzter Zugriff 25. 09. 2018]

Rubenstein, Herbert / Goodenough, John B. (1965): *"Contextual Correlates of Synonymy."* Communications of the ACM 8 (10): 627-6337. https://doi.org/10.1145/365628.365657. [Letzter Zugriff 25. 09. 2018]

Sahlgren, Magnus (2008): *"The Distributional Hypothesis."* Italian Journal of Linguistics 20, (1): 33–53.

Vom Digitalisat zum Kontextualisat – einige Gedanken zu digitalen Objekten

Türkoglu, Enes

enes.tuerkoglu@uni-koeln.de
Cologne Center for eHumanities, Theaterwissenschaftliche Sammlung der Universität zu Köln

Die Theaterwissenschaftliche Sammlung der Universität zu Köln zählt mit einer Objektanzahl im Millionenbereich zu einer der größten Einrichtungen dieser Art weltweit. Die immensen Bestände halten ihre ganz eigenen wissenschaftlichen und methodischen Herausforderungen bereit – darunter auch die Entwicklung geeigneter Werkzeuge, die maßgebende Eigenschaften der Objekte auf neuartige, einnehmende Weise vermitteln können.

Zu den Schätzen der Sammlung zählt ein Bestand türkischer Schattentheaterfiguren, in denen ein künstlerisches Spektakel ruht, welches die reglosen Figuren nur erahnen lassen können – die mit ihnen verbundene Spielpraxis bleibt abwesend. Das Vorhaben einer Digitalisierung dieser Bestände, die die Figuren lediglich in leblose Abbilder auf einem Bildschirm übersetzt, greift hier zu kurz. Vielmehr ist die Untersuchung der den Figuren innewohnenden Bewegungsmöglichkeiten unmittelbar forschungsrelevant und ermöglicht Erkenntnisse darüber, wie die Handlungsmöglichkeiten der Figuren die theatralen und narrativen Praktiken des türkischen Schattentheaters als Ganzes bestimmen.

Es gilt demnach eine Form von Digitalisierung und Bereitstellung auszuarbeiten, die der mit den Objekten verbundenen Spielpraxis Rechnung trägt. Als gangbare Lösung hierzu hat sich im Rahmen der laufenden Forschung ein interaktiver Zugang über die Spieleentwicklungsumgebung *Unity Engine* erwiesen, der die Erforschung der Expressivität der physikalischen Objekte und die Erfahrung ihrer Spielpraxis möglich gemacht hat. Unter anderem wurde auf diese Weise im vorliegenden Projekt die Figur des Frenk erarbeitet. Frenk spielt in den Stücken die Rolle des Frankophilen, des geizigen „Möchtegern-Europäers", der sich seinen Mitmenschen grundsätzlich überlegen fühlt, und dies auf herabwürdigende und angeberische Art auch beständig kundtut. Diesen antipathischen, und politisch durchaus geladenen, Grundeigenschaften wird aber darüber hinaus auch ein bestimmend komischer Zug verliehen. Die spezifischen Bewegungsmöglichkeiten der Schattenfigur korrelieren mit dieser Charakterisierung, v.a. in der Besonderheit einer unversehens ausklappbaren Hand, welche bspw. Geld verlangen kann. Dieser äußerst komische Effekt lässt sich im Programm deutlich erfahren. Es handelt sich hier also um eine interaktive Kontextualisierung, die sich aber explizit nicht etwa als Umgebung für „Digital Puppetry" oder als unterhaltsames Spiel mit kulturellem Hintergrund versteht, sondern als Projekt der Digital Humanities.

Zuvorderst wurden beim Erstellen des Programms die *Unity Best-Practices* der Spieleentwicklung beachtet, wonach es entsprechend diverser Entwurfsmustern aufgebaut werden konnte. Weiter wurde eine möglichst objektorientierte Gestaltung der Funktionen angestrebt, was bspw. Konzepte wie das Objekt Pooling, das Generieren von C#-Klassen für die XML Serialisierung oder das automatische Befüllen der dynamischen Menüs mit Daten einschloss. Über die Beschäftigung mit den Schattenfiguren hinaus versucht das hier vorgestellte Vorhaben mittels dieser Vorgehensweise auch generalisierbare Lösungen zur Verfügung zu stellen, die ebenso bei Objektgruppen Anwendung finden können, die in sich eine aktive, kulturelle Praxis tragen – so u.a. die Grundkonzepte des Input, die Bewegungen per Maus und Tastatur oder die Datenbank, für die Informationen in dem XML-Schema LIDO erfasst wurden.[1]

Momentan nutzen die Mehrzahl der großen Sammlungen digitaler Objekte die in diesem Projekt angeschnittenen Möglichkeiten der digitalen Forschungsräume jedoch kaum aus. Diese unengagierten Datenbanken bergen neben ungenutztem Potential in sich auch Problematiken, die sich in ihrer sturen Fortführung noch potenzieren werden (Egert, Goins & Phelps 2014). Denn zunehmend kommt es zu einer massenhaften Anhäufung digitaler Objekte der prinzipiell selben Signifikanz, deren Organisation, Erreichbarkeit und Nutzbarmachung ein immer größeres Problem wird und die blind bleiben gegenüber der Zentralität von Interpretation. Während nämlich in den traditionellen Institutionen des Erinnerns, also in Archiven, Bibliotheken und v.a. Museen, stets ein kulturelles System mitreflektiert wird, fehlt dieser nüchternen Form von Datenbanken ein entscheidendes Kriterium zum Verständnis geisteswissenschaftlicher Objekte: die Erfahrung ihres kulturellen Kontexts.

Dieser Feststellung entsprechend kann das Ziel eines Projektes der Digital Humanities dann nicht mehr eine optisch präzisere bzw. einprägsamere Dokumentation oder Abbildung sein, sondern die experimentelle Erforschung der den Objekten eingeschriebenen, kulturell-historischen Spannung unter besonderer Berücksichtigung der speziellen Handlungsmacht, welche im sinnlichen Überschuss der Objekte erfahrbar wird (Marx 2012; 2017). Eine solche experimentelle Erforschung ermöglicht dann einen Zugang nicht nur zu dem oft ästhetisch gefassten Eigenwert der Objekte, sondern zu der Prägekraft, die Objekte auf die menschliche Wahrnehmung und Vorstellungskraft als formierende Erfahrung entfalten. Der digitale Raum, mit seinen Möglichkeiten, Objekte gemäß ihren kulturellen Einschreibungen zu bespielen, ist so in der Lage, anstelle eines bloßen Digitalisates etwas zu produzieren, was tentativ „Kontextualisat" genannt werden kann.

In kulturtheoretischen Diskursen erfährt die angesprochene Prägekraft der Objekte oft eine übernatürliche oder magische Beiordnung; besonders prominent erscheinen hier die Abhandlungen Walter Benjamins zur „Aura" des Kunstwerks (Benjamin 1980). Aber auch bspw. Alfred Gell schreibt Kunstobjekten eine Handlungsmacht zu, die Rezipienten dazu führt, sie wie lebendige Wesen zu betrachten. Entstehen könne diese Handlungsmacht des Objekts durch eine „Verzauberung" durch Technologie, womit er im Wesentlichen eine Vereinnahmung des Rezipienten durch technische Virtuosität oder brillante Imagination meint (Gell 1998). Beachtenswert erscheint auch Steven Connors Theorie um den Begriff des „magical objects". Nach Connor liegt die

Magie eines Objekts in seiner distinktiven Aufforderung, auf spielerische Weise, oder sogar durch „praktische Träumerei“ über seine „affordance“ (Gibson 1986) zu reflektieren. Gemeint ist damit die von einem Gegenstand angebotene Gebrauchseigenschaft für Rezipienten. Auch Connors Verständnis einer Magie der Objekte meint damit keine esoterische oder übernatürliche Beeinflussung, sondern eben das durch die Gegenstände selbst evozierte Bewusstsein um ihre kulturellen Einschreibungen (Connor 2011). In welchen Begriffen der kulturtheoretische Diskurs den Einfluss der Objekte aber auch beschreiben mag, stets ist er auf die eine oder andere Weise verknüpft mit Effekten der Einschreibungen kultureller Systeme, die im materiellen, sinnlichen Überfluss der Objekte erfahrbar werden.

Mit dieser Einsicht stellt sich umgehend die Frage, wie digitale Objekte dieser speziellen, über Materialität vermittelte Wirkung von Dingen nachspüren können, transzendieren sie doch jede materielle Form in ihrer Existenz als rein symbolischer Inhalt. Alle Erkenntnisse, die ein digitales Objekt zulässt, resultieren demnach nicht aus Betrachtungen des Materials, sondern müssen zwangsweise aus Überlegungen zum jeweiligen symbolischen Inhalt erfolgen. Diese Ablösung der symbolischen Bedeutung von den Einschränkungen einer tatsächlich gegebenen Materialität (Kroker 2006) kann in Anlehnung an Benjamins erprobten Begriff „Aura der Information“ genannt werden. Die Aura der Information des digitalen Objekts fordert den Betrachter auf, eine bestimmte Repräsentation auf einem Monitor o.ä. zu ignorieren, und stattdessen über die Bedeutung und damit auch den kulturellen Kontext des jeweiligen Objekts nachzudenken. Daher ist die Aura der Information integraler Teil eines jeden Kontextualisates. Mit diesem direkten, unumgänglichen Bezug auf die kulturellen Einschreibungen des Objekts, in denen wie erläutert auch die spezielle Handlungsmacht, die „Magie“, begründet liegt, kann die Aura der Information eines erfolgreichen Kontextualisates eine eingängige Erfahrung generieren, die der Prägekraft des materiellen Objekts nicht nachsteht. Das digitale, immaterielle Kontextualisat kann so den Ersatzcharakter von Reproduktionen überkommen.

Ein solcher Satz lässt umgehend an die allzu präsente Kontroverse innerhalb des interdisziplinären Diskurses denken, nach der das Original sich einer vermeintlichen Bedrohung durch die digitalen Reproduktionen ausgesetzt sieht. Dieser Befürchtung, dass die „unsterblichen“ Digitalisate den Wert des materiellen Originals herabsetzen könnten, scheint einmal mehr die Abhandlung Benjamins inhärent zu sein. Im Licht des Informationszeitalters kann Benjamins pessimistischer Prognose aber spätestens mit Erschaffung der Kontextualisate entgegengestellt werden, dass die digitalen Reproduktionen, in denen sich nur die symbolischen Werte der Originale befinden, die Aura der Originale tatsächlich ausweitet und ihren Einfluss verstärkt. Denn gerade die immateriellen, symbolischen Inhalte der digitalen Reproduktion treten niemals in Konkurrenz zu der Materialität bzw. „Echtheit“ des Originals. Die Aura der Information ist kein Rivale der Aura des Originals. Das Kontextualisat wird vielmehr zum Vehikel der Aura – eine Erkenntnis, die zumindest dem Kulturtourismus nicht neu zu sein scheint (Schweibenz 2015).

Es ist also festzuhalten, dass Objekte sich nur in ihrer ganzen Komplexität öffnen können, wenn man sie angebunden an ihren kulturellen Kontext betrachtet und ihrer sich hier befindlichen aktiven, immersiven Dimension nachspürt. Mit der den Kontextualisaten integralen Aura der Information kann in einer Transzendierung der physischen Form direkter Zugang zu den symbolischen Inhalten der Dinge gewonnen werden. Besonders deutlich wird die Bedeutung der Kontextualisierung bei theaterhistorischen Archivalien wie den für dieses Projekt gewählten Schattenfiguren.[2] Gerade Puppen und Spielfiguren sind mit dem Kontext aufgeladen, aus dem sie stammen, v.a. mit ihren theatralen und narrativen Bezügen, die sie durch ihre eigenen materiellen Handlungsmöglichkeiten wiederum mitbestimmen. Will man den vollen Quellenwert dieser Objekte demnach nicht vernachlässigen, so muss man sie nicht nur in ihrer performativen, sondern auch in ihrer kulturgeschichtlichen Dimension wahrnehmen. Treten über diesen Weg dann auch die materiellen Bedingtheiten der theatralen Kultur in den Blick, so werden die Objekte nicht mehr bloß in illustrativem Sinne verwandt, sondern als wichtige Quellen, anhand derer die kulturellen Energien nachvollzogen werden können, die das Theater formieren.

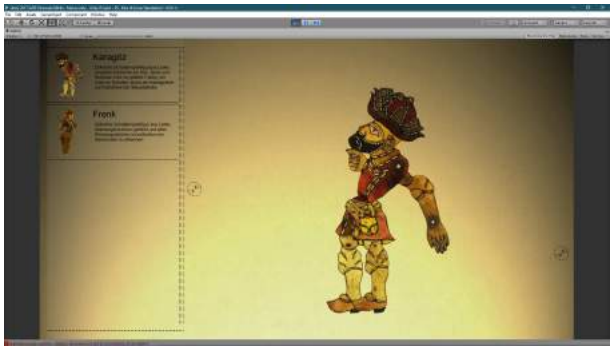


Abbildung 1. Screenshot des Archiv-Modus.

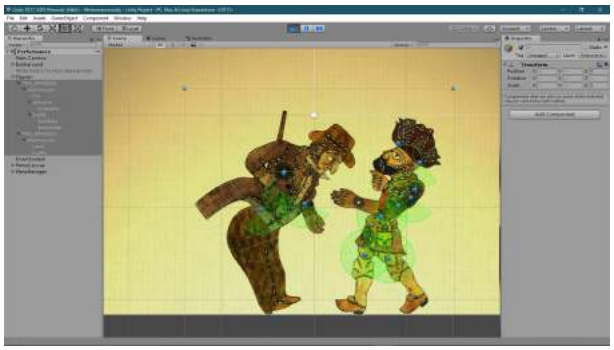


Abbildung 2. Screenshot des Performance-Modus mit hinzugefügten *Joints*.

Fußnoten

1. https://github.com/enestezi/shadowthing .
2. Für die genaue Darlegung der im Rahmen des Projekts gewonnen Erkenntnisse siehe „das türkische Schattentheater – eine Digitalisierungsversuch“ (Türkoglu 2017b)

Bibliographie

Ahrndt, Wiebke (2014): *Das Digitalisierte Museum – Sammlungen Und Museumsforschung Im Zeitalter Der Digitalisierung*. Eröffnungsvortrag „Das Museum von Babel", Panel „Sammlungen und Museumsforschung", Senckenberg Museum Frankfurt, http://www.senckenberg.de/files/forschung/projekte/museumvonbabel/babel_ahrndt_neu.pdf [letzter Zugriff 12 Oktober 2018].

And, Metin (2005): *Karagöz - Turkish Shadow Theatre*. Dost Yayinlari, Istanbul.

Benjamin, Walter (1980): *Das Kunstwerk Im Zeitalter Seiner Technischen Reproduzierbarkeit*, in: Walter Benjamin – Gesammelte Schriften, 471–508, 3rd ed., Suhrkamp, Frankfurt am Main, https://de.wikisource.org/wiki/Das_Kunstwerk_im_Zeitalter_seiner_technischen_Reproduzierbarkeit_(Dritte_Fassung) [letzter Zugriff 12 Oktober 2018].

Connor, Steven (2011): *Paraphernalia: The Curious Lives of Magical Things*. Profile Books, London.

Crosby, Daniel J. (2015): *Engaging with Digital Humanities: Becoming Productive Scholars Of The Humanities In A Digital Age*, Pacific Journal, no. 10, 57-67.

Dörk, Marian (2016): *Den Gehobenen Schatz Allen Zuganglich Machen- Marian Dörk Im Gesprach Mit Lydia Koglin*. Blog, Forschungsverbund Marbach Weimar Wolfenbuttel, http://www.mwwforschung.de/blog/blogdetail/den-gehobenen-schatz-allen-zugaenglich-machen/ [letzter Zugriff 12 Oktober 2018].

Egert, Christopher / Goins, Elizabeth S. / Phelps Andrew (2014): *Interactivity: New Rules of Engagement for The Humanities*. Journal of Interactive Humanities 2, no. 1, 27-29, doi: http://dx.doi.org/10.14448/jih.02.0001 [letzter Zugriff 12 Oktober 2018].

Finch, Julia (2017): *Digital Humanities, Art History, And Object Authenticity*. Field Guide, Mediacommons, http://mediacommons.org/fieldguide/question/what-role-digital-humanities-transforming-and-responding-arts/response/digital-humanities-a [letzter Zugriff 12 Oktober 2018].

Gell, Alfred (1998): *Art and Agency,* Oxford University Press, Oxford.

Gibson, James J. (1986): *Ecological Approach To Visual Perception*, Taylor & Francis Group, New York.

Grant, Ian John (o.D.): *Expressivity and The Digital Puppet: Mechanical, Digital and Virtual Objects In Games, Art And Performance*. Unveröffentlichte Dissertation.

Hazan, Susan (2001): *The Virtual Aura - Is There Space for Enchantment in A Technological World?*. Museums and the Web, https://www.museumsandtheweb.com/mw2001/papers/hazan/hazan.html [letzter Zugriff 12 Oktober 2018].

Jacob, Georg (1925): *Geschichte Des Schattentheaters Im Morgen- Und Abendland*, 2nd ed., Lafaire, Hannover.

König, Mareike (2016): *Was Sind Digital Humanities? Definitionsfragen Und Praxisbeispiele Aus Der Geschichtswissenschaft*. Blog, Digital Humanities Am DHIP, http://dhdhi.hypotheses.org/2642 [letzter Zugriff 12 Oktober 2018].

Kroker, Arthur / Kroker, Marilouise (2006): *The Aura Of The Digital*. Ctheory - 1000 Days Of Theory, no. 041, http://www.ctheory.net/ctheory_wp/the-aura-of-the-digital/ [letzter Zugriff 12 Oktober 2018].

Langner, Martin (2015): *Archaologische Datenbanken Als Virtuelle Museen*. Digital Classics Online 1, no. 1, doi: http://dx.doi.org/10.11588/dco.2015.1.20314 [letzter Zugriff 12 Oktober 2018].

Li, Tsai-Yen / Hsu, Shu-Wei (2006): *An Authoring Tool for Generating Shadow Play Animations with Motion Planning Techniques*.

LIDO Working Group. *What is LIDO*, http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/ [letzter Zugriff 12 Oktober 2018].

Marx, Peter W. (2012): *Scena Mundi. Prolegomena Zu Einer Anthropologie Der Imagination*, in: Raum- Maschine Theater. Szene und Architektur, Wienand, Köln.

Marx, Peter W. (2017): *Von Der Maßgabe Der Dinge*, unveröffentlicht.

McCarty, Willard (2014): *Getting There From Here. Remembering The Future Of Digital Humanities: Roberto Busa Award Lecture 2013*. Literary and Linguistic Computing 29, no. 3, 289-295, doi: http://dx.doi.org/10.1093/llc/fqu022 [letzter Zugriff 12 Oktober 2018].

Ritter, Hellmut (1924): *Karagos : Turkische Schattenspiele – I. Lafaire*, Hannover.

Sahle, Patrick (2015): *Digital Humanities? Gibt's Doch Gar Nicht!*. Grenzen und Möglichkeiten der Digital Humanities, Sonderband der Zeitschrift für digitale Geisteswissenschaften Nr. 1, doi: http://dx.doi.org/10.17175/sb001_004 [letzter Zugriff 12 Oktober 2018].

Schweibenz, Werner (2015): *Museum Analog, Museum Digital: Die Virtualisierung Des Museums Und Seiner Objekte*, in: Wenn Das Erbe In Die Wolke Kommt, Klartext Verlag, Essen.

Thaller, Manfred (2012): *Controversies Around The Digital Humanities: An Agenda*. Historical Social Research 37, no. 3, 12-13, letzter Zugriff 12 Oktober 2018, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-378617 [letzter Zugriff 12 Oktober 2018].

Türkoglu, E. (2017a): *Virtualisierung Historischer Schattenspiele mit der Hilfe von Unity*. Bachelorarbeit, Universität zu Köln, doi: http://dx.doi.org/10.13140/RG.2.2.30304.74249/1 [letzter Zugriff 12 Oktober 2018].

Türkoglu, E. (2017b): *das türkische Schattentheater – eine Digitalisierungsversuch*. doi: http://dx.doi.org/10.7767/muk.1970.16.34.345 [letzter Zugriff 12 Oktober 2018].

Unity Technologies *Unity - Game Engine*, letzter Zugriff 12 Oktober 2018, https://unity3d.com/de/ .

Vom Stellenkommentar zum Netzwerk und zurück: grosse Quellenkorpora und tief erschlossene Strukturdaten auf hallerNet

Stuber, Martin
martin.stuber@hist.unibe.ch
Universität Bern, Schweiz

Dängeli, Peter
p.daengeli@uni-koeln.de
Universität Bern, Schweiz; Cologne Center for eHumanities

Forney, Christian
christian.forney@hist.unibe.ch
Universität Bern, Schweiz

In Kooperation der Universität Bern und des *Cologne Center for eHumanities* CCeH entsteht seit August 2016 die Editions- und Forschungsplattform *hallerNet*, die im April 2019 veröffentlicht werden soll. Eine Spezialität von *hallerNet* sind die systematischen Verknüpfungen zwischen grossen Quellenkorpora und umfangreichen Metadaten, woraus sich ein grosses Analysepotenzial ergibt, das sich z.B. in explorativen Visualisierungen nutzen lässt.

Metadaten aus drei Jahrzehnten

In *hallerNet* integriert werden die umfangreichen prosopografischen und bibliografischen Strukturdaten, die in Form einer relationalen Verbunddatenbank seit anfangs der 1990er-Jahre im Rahmen der beiden SNF-Projekte zum Universalgelehrten Albrecht von Haller (1708-1777)[1] und zur aufgeklärten Reformsozietät *Oekonomische Gesellschaft Bern* (gegr. 1759)[2] erhoben worden waren. Ein mehrjähriges Transformationsprojekt (Kreditvolumen 1.18 Mio CHF) überführt diese untereinander vielfältig verknüpften Daten (vgl. Abb. 1) zu rund 48'000 Publikationen, 25'000 Personen, 20'000 Briefen, 3'000 Orten, 2'900 Pflanzenarten, 1'000 Versammlungen und 850 Institutionen in eine TEI-konforme XML-Datenstruktur (Recker-Hamm / Stuber 2015). Bemerkenswert ist dabei die grosse Tiefe der Metadatenerschliessung, die beispielsweise nicht nur biografische Eckdaten wie Geburts-/Sterbedatum und -ort oder die Hauptbeschäftigung umfasst, sondern die Ausbildungsstationen, Reiseziele, Ämter und Mitgliedschaften usw. einbezieht und auch das Beziehungsnetz (Verwandte, Briefpartner) soweit möglich erfasst.[3] Während die prosopografische Erhebung weitgehend auf biografischen Werken sowie auf edierten Universitätsmatrikeln und Mitgliederlisten von Akademien und Sozietäten basiert, stützen sich die bibliografischen Daten in erster Linie auf zwei autoptisch erarbeitete Grundlagenwerke (Monti 1983-1994, Steinke / Profos 2004).

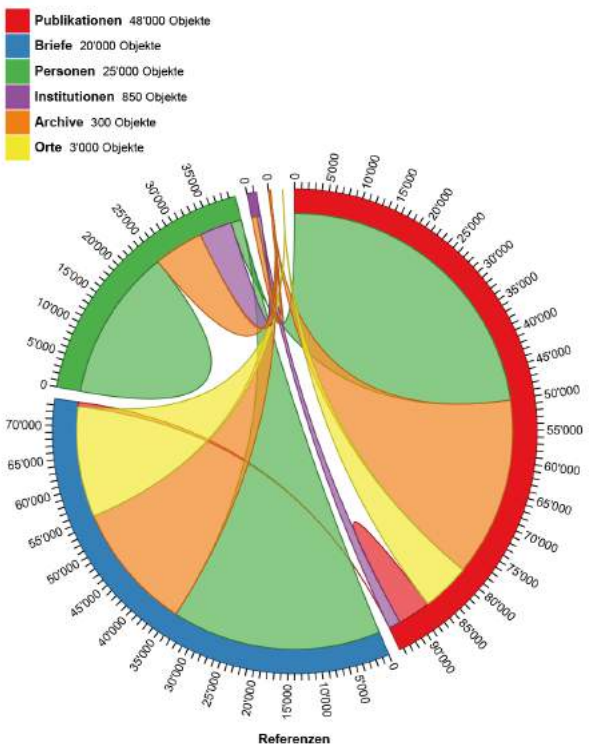


Abb. 1: Die rund 100'000 Metadatenobjekte sind vielfältig verknüpft, sowohl innerhalb des gleichen Entitätstyps als auch typenübergreifend.[4]

Auf der Grundlage der entstehenden Plattform startete anfangs 2018 ein vom *Schweizerischen Nationalfonds* SNF unterstütztes Editions- und Forschungsprojekt, das im geplanten Projektzeitraum von sechs Jahren zum einen die Gesamtedition der rund 9'000 Buchbesprechungen Albrecht von Hallers erarbeitet. Zum anderen soll eine begründete Auswahl von inhaltlich damit zusammenhängenden rund 8'000 Briefen ediert werden, dies als Zwischenetappe zur längerfristig angestrebten Gesamtedition von Hallers Korrespondenz, die insgesamt rund 17'000 Briefe umfasst.[5] In der systematischen Verknüpfung zweier komplementärer Quellenkorpora der privaten (und tw. im kleinen Kreis öffentlichen) Kommunikation (Briefe) und des öffentlichen Diskurses (Rezensionen) liegt eine erste Innovation der Plattform.[6]

"Datenzentrierte" versus "textzentrierte" digitale Editionen

Auf *hallerNet* werden die v.a. quantitativen und formalen Bezüge der älteren Forschungsdatenbank mit edierten Textinhalten in Verbindung gesetzt. Die über Jahrzehnte systematisch erhobenen und homogenisierten Metadaten erleichtern die Editionsarbeit erheblich, namentlich die Referenzierung von Entitäten wie Akteure, Orte, Publikationen und Institutionen. In dieser Verbindung tief erschlossener Strukturdaten mit edierten Textinhalten liegt die zweite Innovation der entstehenden Plattform. Bisher sind "datenzentrierte" digitale Editionen die Ausnahme gegenüber den vielen existierenden und sich in Entwicklung befindlichen

"textzentrierten" digitalen Editionen. Wo es sie gibt, verfolgen sie zumeist ganz spezifische Forschungsfragen und legen dazu ein entsprechend verengtes Datenmodell zugrunde.[7] Daneben existieren biografische, prosopografische oder bibliografische Datenbanken und Forschungsportale, die sich einzig auf Metadatenerhebung und -analyse konzentrieren.[8] Und schließlich gibt es digitale Editionsportale wie das Nowa Panorama Literatury Polskiej,[9] die relativ dichte biografische Daten erheben, diese aber nicht in strukturierter Form bereitstellen.

Selbstredend lagen auch der Datenerfassung in den verschiedenen Projekten um Albrecht von Haller seit den 1990er-Jahren immer spezifische Forschungsinteressen zugrunde. So folgte auf eine erste Phase, welche die Gesamterschliessung von Hallers Korrespondenz (Boschung et al. 2002) mit einer umfassenden prosopografischen Strukturanalyse (Stuber et al. 2005) kombinierte, eine zweite Phase, in der Netzwerkanalysen im Vordergrund standen und die in den Blick genommenen Akteure über Hallers Korrespondenznetz hinaus erweitert wurden (Dauser et al. 2008, Stuber / Krempel 2013). Die dieser Forschung zugrundeliegenden Daten wurden von Anfang an in der gleichen Datenbankanwendung (FAUST) und mit möglichst viel struktureller Homogenität gepflegt, die Datenpublikation erfolgte allerdings fast ausschliesslich in aggregierter Form und auf traditionelle Weise (Druck). Im Zuge der Datentransformation wird durch weitere Homogenisierung und die semantische Auszeichnung nach TEI-Richtlinien eine weitergehende Öffnung der Daten angestrebt, so dass diese auch in anderen Kontexten nutzbar werden.[10] Die Metadaten werden zu diesem Zweck über spezifische Formate wie das CMI-Format bzw. *CorrespSearch* zugänglich gemacht, daneben aber auch in vollem Umfang parallel im FAIR-Repositorium *Zenodo* veröffentlicht werden.[11]

Referenzannotation im Zentrum

Die Tatsache, dass die Briefe je nach Qualität der vorliegenden Grundlagen auf drei unterschiedliche Arten ediert werden – den Standards genügende Druckeditionen werden re-ediert, bei schlechteren Druckeditionen wird der Rohtext zur Arbeitserleichterung herangezogen, aber intensiv nachbearbeitet, und zuvor nicht edierte Dokumente werden ab dem Original neu ediert –, bedingte auch eine Reflektion über das Wesen der Fussnoten bzw. der Stellenkommentare beim Transfer vom Druck ins Digitale. Aus diesem Prozess resultierte ein Referenzsystem, das drei Annotationstypen vorsieht:

- Referenzannotationen: Referenzen auf Datenbankobjekte (`<rs key="person_00001">`), deren lokales Auftreten in einer `<note type="annotRef">` näher erläutert werden kann
- freie Annotationen zum Inhalt (`<note type="annotFree">`), d.h. inhaltliche Anmerkungen, die nicht in direktem Bezug zu einem bestimmten Datenbankobjekt stehen, sondern z.B. zu einem ganzen Absatz
- Annotationen zur Textkonstitution (`<note type="annotText">`)

Während für die letzten beiden Typen die digitale Umsetzung relativ nahe beim Druck bleibt (Anmerkungsziffer mit Tooltip bzw. Verweis zur Anmerkung), bieten die Referenzannotationen mehr Interaktivität und Zugangswege zu den Datenbankdaten. Wir erwarten, dass dieser Anmerkungstyp für künftige Editionen der *hallerNet*-Plattform noch an Bedeutung gewinnen wird. Es sind dann auch diese Referenzannotationen, die das Rückgrat der Entitätsauszeichnung bilden.

Doppelter Zweck der systematischen Entitätsauszeichnung

Innerhalb der Plattform verfolgt die Anbindung der digitalen Edition an die systematisierten Objekte der Datenbank einen doppelten Zweck: Zum einen macht sie die Quellentexte über plattformübergreifende Normdaten anschlussfähig zu anderen Ressourcen. Zum anderen lassen sich bei entsprechender Datenmodellierung aus solchen Referenzannotationen grosse Datenmengen für Netzwerkanalysen und räumliche Visualisierungen gewinnen. Diese Daten weisen erstens präzise Beziehungsqualitäten auf, so beispielsweise zum Verhältnis zwischen den Rezensenten und den Verfassern der rezensierten Publikationen. Zweitens sind die Nennungen der referenzierten Entitäten leicht den i.d.R. tagesdatierten Quellen Brief und Rezension zuzuordnen. Die generierten Forschungsdaten liefern damit Präzision sowohl in der raumzeitlichen Entwicklung als auch in der inhaltlichen Qualität der Netzbeziehungen und erlauben daher Antworten auf wichtige Postulate der aktuellen historischen Netzwerkforschung.[12] Dies möchten wir an einigen konkreten Beispiel andeuten. Ausgangspunkt ist die auf *hallerNet* als Prototyp erstellte digitale Edition der Münchhausen-Haller-Korrespondenz von Otto Sonntag, die ein erweitertes Briefsample von rund 850 Briefen (inkl. Beilagen) umfasst (Sonntag 2019). Dabei werden die vorhandenen prosopografischen Forschungsdaten angereichert mit standardisierten Angaben über soziale Position und Aufenthaltsort sämtlicher in diesem Briefsample erwähnten rund 340 Personen (ca. 1'500 Nennungen) zum Zeitpunkt des entsprechenden Briefdatums. Damit wird diese Korrespondenz als zeitlich und sozial differenzierter Kommunikationsraum darstellbar. Diese Informationen lassen sich einerseits für einzelne edierte Dokumente nutzen, etwa um eine Brieftranskription um eine Akteursliste zu ergänzen, die auch die zum Briefdatum jeweils aktuelle Position der genannten Individuen aufführt. Sie sind aber andererseits vor allem auf aggregierter Ebene interessant, da sich mit verschiedenen Visualisierungsformen bestimmte Interaktionsmuster eruieren lassen.

Die DHd-Konferenz fällt zeitlich direkt in die letzte Phase vor dem Launch der Plattform. Nachdem die Encodings der Münchhausen-Briefe in den letzten Monaten erstellt und mit den bestehenden und neuen Datenbankobjekten referenziert wurden, bietet sich bis dahin die Gelegenheit für erste analytische Auswertungen auf der neuen Datengrundlage.

Den beschriebenen Visualisierungsformen kommen in vereinfachter Form auf *hallerNet* auch Navigations- und Filterfunktionen zu. Der Weg von der einzelnen Textstelle in

der edierten Quelle zur Gesamtsicht in der Visualisierung ist also durchaus in beide Richtungen möglich.

Fußnoten

1. *Albrecht von Haller und die Gelehrtenrepublik des 18. Jahrhunderts* (1991–2003); Leitung: Urs Boschung (Institut für Medizingeschichte der Universität Bern); siehe allg. zu Albrecht von Haller: http://www.albrecht-von-haller.ch.
2. *Nützliche Wissenschaft, Naturaneignung und Politik. Die Oekonomische Gesellschaft Bern im europäischen Kontext (1750–1850)*; Leitung: André Holenstein, Christian Pfister (Historisches Institut der Universität Bern).
3. Siehe zum alten Haller-/OeG-Datenbankverbund: Flückiger / Stuber 2009, Steinke 2003.
4. Die Visualisierung zeigt die Verlinkung und die Anzahl der Referenzen zwischen sechs Entitätstypen. Die Daten wurden im Rahmen von Forney et al. (2018) berechnet und die Grafik mit *d3-chord* erstellt: https://github.com/d3/d3-chord.
5. Online-Edition der Rezensionen und Briefe Albrecht von Hallers. Expertise und Kommunikation in der entstehenden *Scientific community*. Leitung: André Holenstein (Historisches Institut), Hubert Steinke (Institut für Medizingeschichte), Oliver Lubrich (Germanistisches Institut).
6. Siehe zur Grundidee: Stuber 2004.
7. Beispiele hierfür sind etwa die Projekte *Proceedings of the Old Bailey* (https://www.oldbaileyonline.org/), *buckhardtsource.org* (http://burckhardtsource.org/), *Intoxicants and Early Modernity* (https://www.intoxicantsproject.org/), *Jahrrechnungen der Stadt Basel 1535 bis 1610* (https://gams.uni-graz.at/context:srbas), *Urfehdebücher der Stadt Basel* (https://gams.uni-graz.at/context:ufbas), *Sound Toll Registers* (http://soundtoll.nl) oder das leider dem digitalen Zerfall anheimgefallene *Alcalá account book*.
8. Als Beispiele zur Gelehrtenwelt lassen sich hier *Amore Scientiae Facti sunt Exules* (http://asfe.unibo.it) oder das *Repertorium Academicum Germanicum* (http://rag-online.org) nennen. Bemerkenswert sind auch die der *Symogih*-Plattform (Système modulaire de gestion de l'information historique, http://symogih.org) angegeliederten Projekte, die teilweise auch edierte Texte umfassen.
9. Nowa Panorama Literatury Polskiej / New Panorama of Polish Literature; http://nplp.pl und spezifischer http://tei.nplp.pl.
10. Zur Transformation ins TEI-Format vgl. Forney et al. 2018.
11. Diese Strategie verfolgt zwei Ziele: die langfristige Zugänglichkeit der Forschungsdaten, aber auch die Auffindbarkeit über spezialisierte Suchmaschinen wie *DataCite* oder die *Google Dataset Search*.
12. Siehe eine kritische Diskussion der Möglichkeiten und Grenzen der Netzwerkanalyse aus literaturwissenschaftlicher Perspektive (Baillot 2018).

Bibliographie

Baillot, Anne (2018): *Die Krux mit dem Netz. Verknüpfung und Visualisierung bei digitalen Briefeditionen*, in: **Bernhart, Toni et al. (eds.):** *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*, Berlin/Boston: De Gruyter 355-370 <halshs-01278211>.

Boschung, Urs et al. (2002): *Repertorium zu Albrecht von Hallers Korrespondenz* 1724-1777. 2 Bde. Basel.

Dauser, Regina et al. (2008): *Wissen im Netz. Botanik und Pflanzentransfer in europäischen Korrespondenznetzen des 18. Jahrhundert*, Berlin.

Flückiger, Daniel / Stuber, Martin (2009): *Vom System zum Akteur. Personenorientierte Datenbanken für Archiv und Forschung*, in: **Kirchhofer, André et al. (Hrsg.):** *Nachhaltige Geschichte. Festschrift für Christian Pfister*, Zürich, 253-269.

Forney Christian et al. (2018): *Vom geschützt zugänglichen Datenbankverbund zur offenen Editions- und Forschungsplattform: kritischer Rückblick auf halber Strecke*, DHd 2018 (Poster).

Monti, Maria Teresa (1983-1994): *Catalogo del Fondo Haller della Biblioteca Nazionale Braidense di Milano, a cura di Maria Teresa Monti*, 13 Bde. Milano.

Recker-Hamm, Ute / Stuber, Martin (2015): *Haller Online – Konzept für den Umbau, Ausbau und die langfristige Sicherung der Haller-/ OeG-Datenbank*, 25 Seiten (8.6.2015).

Sonntag, Otto (2019): *The Albrecht von Haller-Gerlach Adolph von Münchhausen Correspondence*, hallerNet (in Vorbereitung).

Steinke, Hubert (2003): *Archive databases as advanced research tools: the Haller Project*, in: **Antonio Vallisneri :** *L`edizione del testo scientifico d`età moderna, a cura di Maria Teresa Monti*, Firenze 2003, 191-204.

Steinke, Hubert / Profos, Claudia (2004): *Bibliographia Halleriana. Verzeichnis der Schriften von und über Albrecht von Haller*, Basel.

Stuber, Martin (2004): *Journal and letter. The interaction between two communication media in the correspondence of Albrecht von Haller*, in: **Lüsebrink, Hans-Jürgen / Popkin, Jeremy (eds.):** *Enlightenment, Revolution and the periodical press (Studies on Voltaire and the Eighteenth Century)*, 114-141.

Stuber, Martin et al. (2005): *Hallers Netz. Ein europäischer Gelehrtenbriefwechsel zur Zeit der Aufklärung*, Basel.

Stuber, Martin / Krempel, Lothar (2013): *Las redes académicas de Albrecht von Haller y la Sociedad Económica: un análisis de redes a varios niveles*, in: REDES. Revista hispana para el análisis. De redes sociales, 24/1 (2013), 1-26: https://doi.org/10.5565/rev/redes.450 / REDES Online English: The Scholarly Networks of Albrecht von Haller and the Economic Society of Bern – a Multi-Level Network Analysis: http://revista-redes.rediris.es/webredes/novedades/Stuber_Krempel_scholarly_networks.pdf.

Von IIIF zu IPIF? Ein Vorschlag für den Datenaustausch über Personen

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich

Vasold, Gunter
gunter.vasold@uni-graz.at
Universität Graz, Österreich

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Wien, Österreich

Gesellschaften fügen sich aus Individuen zusammen. Das gilt auch für die Vergangenheit, aus der die Mehrzahl der Individuen nur schlecht bis gar nicht dokumentiert ist. Es hat sich deshalb ein eigenständiger historischer Forschungsbereich entwickelt, die "Prosopographie", die sich der Aggregation von Einzelinformationen zu Individuen aus historischen Quellen und ihrer Auswertung widmet (Keats-Rohan 2007). Dieses Forschungsgebiet hat früh digitale Methoden eingesetzt. Der Beitrag widmet sich der Frage, ob die Methoden vergleichbar zu IIIF (International Image Interoperability Framework) in ein „International Proposography Interoperability Framework" (IPIF) integriert werden können.

Ein IPIF muss von den Personendatenbanken abweichen, die sich als kontrollierte Vokabularien und Referenzen für Linked Open Data in den Digital Humanities etabliert haben (GND/VIAF, deutsche-biographie), bzw. im Begriff sind, sich zu etablieren (wikidata). Diese berücksichtigen nämlich nicht den Vorgang, mit dem Informationen über eine Person aus historischen Quellen aggregiert werden. Der Ansatz weicht damit auch von der „personography" der TEI ab, die, wie die Linked-Data-Ressourcen, eine Person mit einer Liste an Eigenschaften beschreiben. Ein IPIF muss dagegen ein Modell realisieren, für das Bradley/Short (2005) die Bezeichnung „Factoid"-Model eingeführt haben. Es geht von drei Informationseinheiten aus: Quelle, Individuum und Aussagen der Quelle über das Individuum. John Bradley hat das Modell mehreren Projekten des King's College London zu Grunde gelegt (PASE, DPRR, CCEd). Auch das Persondendatenrepositorium (PDR) der Berlin-Brandenburgischen Akademie der Wissenschaften (Neumann et al. 2011) und Projekte, die die Software der BBAW weitergenutzt haben, verwenden das gleiche Modell, auch wenn das PDR nicht explizit auf Bradley referenziert. Ebenso verwendet das Repertorium Academicum Germanicum ein solches dreiteiliges Modell (Andresen 2008).

Das dreiteilige Modell impliziert auch, dass (auch widersprechende) Aussagen über dasselbe Individuum aus verschiedenen Quellen an verschiedenen Orten publiziert werden können. Es erscheint also als ein Paradebeispiel für das *Web of Data* des W3C. Das *Web of Data* ist die Fortführung der Semantic-Web-Aktivitäten des W3C. Es konzentriert sich auf die Öffnung von Datenbanken und erhebt insbesondere den Anspruch, individuelle kleine Datenmengen als RDF über das Semantic Web abfragbar zu machen. Technisch ist RDF, die Grundlage des *Web of Data*, eine weit verbreitete und gut unterstützte Technologie. Es ist deshalb auch eine Technologie, mit deren Hilfe immer häufiger Maschinen auf prosopographische Datenbanken zugreifen können. Deshalb haben Bradley/Pasin (2015) eine CIDOC-CRM basierte Version des Factoid-Modells vorgeschlagen und entsprechende Ontologien veröffentlicht (Bradley 2017). Das Basismodell ist aber auch mit anderen Vokabularien realisiert worden: SNAP verwendet z.B. Vokabularien aus dem Linking-Ancient-Wisdom-Projek[1]). Das King's Digital Lab hat jüngst mit Hilfe von *Ontop*[2] die prosopographische Datenbank zur römischen Republik als LOD-Ressource incl. eines SPARQL-Endpoints veröffentlicht.[3]

Diese Strategie teilt jedoch das Problem vieler RDF-Ressourcen: Die technische Pflege eines SPARQL-Endpoints ist sehr anspruchsvoll. SPARQL-Endpoints sind häufig nur unzuverlässig verfügbar. Nicht zuletzt deshalb stellen große Lieferanten von RDF-Ressourcen wie die Deutsche Nationalbibliothek (GND) bis dato keine SPARQL-Endpoints für ihre Daten zur Verfügung. Als alternative Technologie etabliert sich in den Digitalen Geisteswissenschaften zunehmend die Publikation von eigenen RESTful APIs, die zwar weit weniger flexible Abfragen erlauben, dafür aber deutlich stabiler funktionieren und einfacher implementiert werden können. RESTful APIs sind ein Quasi-Industriestandard und werden von jedem Webentwicklungsframework und jeder Programmiersprache unterstützt. Mit OpenAPI (ehemals swagger)[4] und Core API[5] liegen auch Vorschläge vor, derartige API-Definition standardisiert zu beschreiben, so dass die Implementation von einschlägigen API-Anbietern und API-Konsumenten teilweise sogar automatisiert werden kann.[6] Aus Sicht des Software-Engineering erscheint es also angemessen, auf eine eigene API-Definition statt auf einen SPARQL-Endpoint zurückzugreifen. Gleichzeitig wird es damit erschwert, Daten aus verschiedenen Datenquellen zu aggregieren, da für jeden Datenanbieter ein eigener API-Konsument programmiert werden müsste. Im Bereich der Bibliotheken hat sich deshalb für die Bereitstellung von Bildern von Büchern mit IIIF eine Kombination aus einem Datenstandard und einer Adressierungs-API durchgesetzt. Es ist an der Zeit, auch für personenbezogene Daten über einen solchen technischen Standard nachzudenken, der die Implementation von Anwendungen erleichtert und die Daten auch praktisch interoperabel macht.

Ein solcher Standard muss von konkreten Anwendungsszenarien ausgehen. Sie können unter den Überschriften „Biographical Lexicon", „Careers", „Source Editing", „Fact Checking", „New Interpretation", „Publish a Database", „Integrate Other Databases", „Analysis", „Tool User", „Tool Builder" zusammengefasst werden. Die Szenarien bilden sowohl Forschung mit prosopographischen Daten wie die Erzeugung solcher Daten ab. Zusätzlich achten die Szenarien darauf, nicht nur explizit prosopographische Workflows zu berücksichtigen, sondern schließen auch wissenschaftliches Edieren als Szenario mit ein, in dem der edierte Text als

Beleg für eine Person betrachtet werden kann. In einem Workshop in Wien im Februar 2017 haben Forscher aus dem Themengebiet der Prosopographie religiöser Orden solche Anwendungsszenarien diskutiert und einen Entwurf für eine API entwickelt.

Ein Ergebnis dieser Arbeit ist eine nach den Standards von OpenAPI beschriebenen Definition einer prosopographischen API.[7] Die API baut auf dem dreiteiligen Factoid-Modell auf und erlaubt den Zugriff auf Personen, Aussagen, Quellen und ihr Aggregat, einem „Factoid". Für alle diese Objekte gibt es eigene Pfade zur Suche und Ausgabe der Daten über die zu ihnen abgelegten IDs. Im Kern der API steht deshalb der Zugriff auf Factoide (/factoid). Sie können individuell über bekannte IDs adressiert werden (/factoid/id). Wichtiger sind aber inhaltliche Filtermöglichkeiten. Sie ergeben sich einfach aus den Eigenschaften des Factoids, als Aussage über eine Person. Die Parameter *s, st* und *f* lassen also die Suche in den Inhalten der mit dem Factoid verknüpften Quellen (source), Aussagen (statement) und den Metadaten des Factoids selbst (factoid) zu. Dabei ist der Standard eine Volltextsuche. Ebenso lassen sich die Quellen und Personen abfragen. Als Parameter können aber auch Identifikatoren für die einzelnen Informationsgruppen übergeben werden, also z.B. mit /statement/?p_id=Placidus_Seiz alle Aussagen über die Person mit einem Identifikator „Placidus_Seiz" in einem beliebigen Kontext. Die Anwendung liefert dann ein JSON-Objekt zurück, in dem diese Aussagen formalisiert sind. Zu jeder Aussage gehört eine ID, mit der Entwickler z.B. über die API überprüfen können, woher die jeweilige Aussage stammt.

Als Rückgabewert der API-Definition sind JSON-Serialisierungen vorgesehen. Die Statements können Daten als Text (z.B. der Quelle) ebenso wie strukturiert als Graph enthalten. Die Graphen sollen den Spezifikationen von JSON-LD folgen. Damit können zwei Ziele erreicht werden: Erstens ist damit die Ausgabe der API direkt in Linked-Open-Data-Umgebungen nutzbar, kann prinzipiell auch in einer FROM-Klausel einer SPARQL-Abfrage integriert werden oder in Caching-Mechanismen wie im 2011 als Linked Data Middleware von Virtuoso vorgeschlagenen URI-Burner verwendet werden. Zweitens wird damit ein Standard verwendet, der die Referenzierung der verwendeten Vokabularien und ihre formale Beschreibung mit RDFS und OWL ermöglicht.

Der Workshop in Wien hat als Kernproblem eines echten Datenaustausches die divergierenden Datenmodelle für die Einzelaussagen über die Individuen identifiziert. Während die Individuen selbst im Factoid-Modell keine beschreibenden Metadaten tragen und damit kaum Probleme beim Datenaustausch erzeugen, sind für die Aussagen über die Individuen je nach Projekt, Verwendungszweck und Forschungsdomäne eine Vielfalt von Vokabularien im Einsatz. Einen Ausweg aus dieser Situation bietet die 2017 gegründete dataforhistory-Initiative.[8] Die Initiative arbeitet daran projekt- und domänenspezifische Modellierungen zu erleichtern, die zum CIDOC CRM kompatibel sind. Die derzeitige API-Definition sieht deshalb vor, dass die zurückgegebenen Daten eine Referenz auf ein Schema (in JSON-LD als @context) enthalten müssen, das die verwendeten Klassen und Eigenschaften auf Definitionen im CIDOC CRM abbildet, der es der die API konsumierenden Anwendung erlaubt, die Daten als CIDOC CRM zu interpretieren und darauf aufbauende Operationen durchzuführen. Ergänzend dazu ist ein Parameter format=json/cidoc-crm vorgesehen, bei dem die Transformation serverseitig stattfindet. Die Abbildung auf CIDOC CRM soll insbesondere die grundlegenden Suchoperationen ermöglichen, die Katerina Tzompanaki und Martin Doer 2012 formuliert haben und die im Projekt researchspace[9] realisiert werden. Die API definiert die Objekteigenschaft graph für die strukturierte Repräsentation der Daten über Personen.

John Bradley und Michele Pasin haben 2015 eine OWL basierte Ontologie vorgestellt, in der eine „temporal entity documented" (TED) als Ereignis (E4 und E5 im CIDOC-CRM) oder als eine zeitliche Einheit oder klare zeitliche Grenzen (E3: condition, state) modelliert sind. Das entspricht dem Stand der Diskussion über prosopographische Datenmodelle (Lind 1994, Andresen 2008, Tuominen / Hyvönen / Leskinen 2018).

Nicht zuletzt der Erfolg von IIIF belegt, dass eine solche API aber auch Referenzimplementationen benötigt. Dabei ist entsprechend der oben beschriebenen Benutzungsszenarien sowohl an Ressourcen zu denken, die Daten bereitstellen, als auch an Anwendungen, die diese Daten konsumieren. Die Nachnutzung des „Archiveditors", eines zunächst projektinternen Werkzeugs der BBAW, in anderen Projekten zeigt, dass dabei nicht nur an Datenextraktion und –anzeige sondern auch an Datengenerierung zu denken ist. Im Rahmen der Arbeit an der Personendatenbank der Österreichischen Akademie der Wissenschaften ist deutlich geworden, dass gerade automatische Informationsextraktion von „Personenrelationen" (Schlögl et al. forthcoming, Schlögl et al. 2018) von einer solchen API profitieren kann. Die automatisch generierten Aussagen können als eigenständige Factoide in die Personendatenbanken eingehen. Die Metadaten des Factoids und die Referenz auf die verwendete Quelle stellen sicher, dass sie als automatisch generierte Daten identifizierbar bleiben. Der Vortrag wird Beispiele für Datenangebote aus dem Umfeld mittelalterlicher Urkunden (Register der Urkundenempfänger von Papsturkunden nach den Regesten von August Potthast, Daten aus monasterium.net) und Steuererhebungen (England) vorstellen, und Prototypen für Anwendungen benennen, welche die mit der API bereitgestellten Daten konsumieren können.

Fußnoten

1. http://lawd.info/ontology/
2. https://github.com/ontop/ontop
3. http://romanrepublic.ac.uk/rdf , Dokumentation von John Bradley: http://romanrepublic.ac.uk/rdf/doc
4. https://www.openapis.org/
5. http://www.coreapi.org
6. z.B. das Python-Framework Flask in Verbindung mit https://github.com/zalando/connexion , vgl. weitere Tools: https://swagger.io/tools/open-source/open-source-integrations/
7. https://github.com/GVogeler/prosopogrAPhI
8. http://dataforhistory.org
9. https://www.researchspace.org/

Bibliographie

Andresen, Suse (2008): *Das 'Repertorium Academicum Germanicum'. Überlegungen zu einer modellorientierten*

Datenbankstruktur und zur Aufbereitung prosopographischer Informationen der graduierten Gelehrten des Spätmittelalters, in: **Sigrid Schmitt u. Michael Matheus (eds.):** *Städtische Gesellschaft und Kirche im Spätmittelalter* (Geschichtliche Landeskunde 62). Stuttgart: Steiner 17-26.

Bradley, John (2017): *Factoids. A site that introduces Factoid Prosopograph*, http://factoid-dighum.kcl.ac.uk/ und https://github.com/johnBradley501/FPO

Bradley, John / Pasin, Michele (2015): *Factoid-based Prosopography and Computer Ontologies. Towards an integrated approach*, in: DSH 30,1: 86-97.

Bradley, John / Short, Harold (2005): *Texts into databases. The Evolving field of New-style Prosopography*, in: LLC 20, suppl. 1: 3-24.

CCEd: *Clergy of the Church of England Database*, King's College London http://theclergydatabase.org.uk/

DPRR: *Digital Prosopography of the Roman Republic*, King's College London

Keats-Rohan, Katherine S.B. (ed.) (2007): *Prosopography. Approaches and Applications. A Handbook* (Prosopographica et genealogica 13). Oxford: P&G.

Lind, Gunner (1994): *Data Structures for Computer Prosopography*, in: Yesterday: Proceedings from the 6th International Conference of the Association of History and Computing, Odense 1991. Odense: University Press of Southern Denmark. 77-82.

Neumann, Gerald / Körner, Fabian / Roeder, Torsten / Walkowski, Niels-Oliver (2011): *Personendaten-Repositorium*, in: Berlin-Brandenburgische Akademie der Wissenschaften. Jahrbuch 2010: 320-326.

PASE: *Prosopography of Anglo-Saxon England*, King's College London, URL: http://www.pase.ac.uk/jsp/index.jsp

Schlögl, Matthias / Katalin Lejtovicz (2018): *A Prosopographical Information System (APIS)*, in: **Antske Fokkens / ter Braake Serge / Sluijter, Ronald / Arthur, Paul / Wandl-Vogt, Eveline (eds.):** *BD-2017. Biographical Data in a Digital World 2017. Proceedings of the Second Conference on Biographical Data in a Digital World 2017*. Linz, Austria, November 6-7, 2017. Budapest: CEUR (CEUR Workshop Proceedings 2119): 53-58.

Schlögl, Matthias / Lejtovicz, Katalin / Bernád, Ágoston Zénó / Kaiser, Maximilian / Rumpolt, Peter (2018): *Using deep learning to explore movement of people in a large corpus of biographies*. Zenodo. http://doi.org/10.5281/zenodo.1149023

Tuominen, Jouni / Hyvönen, Eero / Leskinen, Petri (2018): *Bio CRM. A Data Model for Representing Biographical Data for Prosopographical Research*, in: BD-2017. Biographical Data in a Digital World 2017, hg. v. Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt, Budapest: CEUR (CEUR Workshop Proceedings 2119): 59-66.

Tzompanaki, Katerina / Doerr Martin (2012): *Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories*, Technical Report ICS-FORTH/TR-429, April 2012, <http://cidoc-crm.org/docs/TechnicalReport429_April2012.pdf>

Von Wirtschaftsweisen und Topic Models: 50 Jahre ökonomische Expertise aus einer Text Mining Perspektive

Wehrheim, Lino
lino.wehrheim@ur.de
Universität Regensburg, Deutschland

Text Mining für historische Fragestellungen

Aufgrund der zunehmenden Verfügbarkeit digitaler Textsammlungen kommt Methoden des Text Mining (Heyer 2009) auch für zeit- und wirtschaftshistorische Fragestellungen eine immer größere Bedeutung zu. Der Beitrag geht der Frage nach, ob und wie ein spezifischer Ansatz, die automatisierte Inhaltsanalyse mittels Topic Modelling (Blei 2012), für die Untersuchung eines wirtschaftshistorisch bedeutsamen Korpus eingesetzt werden kann: der Jahresgutachten des Sachverständigenrats zur Begutachtung der gesamtwirtschaftlichen Entwicklung (SVR), eines der prominentesten Gremien in der ökonomischen Politikberatung. Ihr erstes Jahresgutachten veröffentlichten die „Wirtschaftsweisen" 1964, seither hat sich ein beachtliches Korpus angesammelt, dessen Umfang für Untersuchungen mit einer zeitlich oder thematisch begrenzten Fragestellung zwar ein weniger großes Hindernis darstellt (Schanetzky 2007; Strätling 2001); wird jedoch eine diachrone Fragestellung verfolgt, bieten Topic Models einen vielversprechenden Ansatz. Diese haben in den vergangenen Jahren in verschiedensten Wissenschaftsbereichen Anwendung gefunden (Boyd-Graber et al. 2017; Wehrheim 2018), wobei ihre Popularität maßgeblich auf das von Blei et al. (2003) eingeführte Model der *Latent-Dirichlet-Allocation* (LDA) zurückzuführen ist.

Korpus, Vorverarbeitung und Modellerstellung

Das untersuchte Korpus besteht aus den zwischen 1965 und 2015 veröffentlichten Publikationen des SVR, d.h. 51 Jahresgutachten, 23 Sondergutachten und 7 Expertisen.[1] Hinzu kommen die seit 1968 veröffentlichten Jahreswirtschaftsberichte, in denen die Bundesregierung ihre Sicht der Wirtschaftslage erörtert und zum Jahresgutachten Stellung bezieht und die ein naheliegendes Vergleichskorpus bilden. Die Texte wurden über die Parlamentsdokumentation des Deutschen Bundestags bezogen.[2] Die dort bereitgestellten OCR-Scans weisen eine sehr hohe Textqualität auf. Längere

Dokumente wurden anhand der jeweiligen Gliederung in Einzelkapitel geteilt (Jockers 2013). Dabei sollten möglichst gleich lange Texte entstehen und diese möglichst einheitliche Themenbereiche enthalten. Daher wurden Texte teilweise erneut auf einer tiefer liegenden Gliederungsebene geteilt. Daraus resultierte ein Korpus aus 1.150 Dokumenten mit einem Textumfang von 8,3 Millionen Wörtern (Vorworte, Inhalts- und Literaturverzeichnisse sowie Anhänge wurden entfernt). Dieses Vorgehen besitzt den Vorteil, dass so thematisch homogenere Texte entstehen, welche wiederum abgrenzbare, spezifische Topics produzieren. Für die Auswertung wurden anschließend die Topic-Anteile der einzelnen Teil-Dokumente aggregiert, indem der Mittelwert über alle Teile des jeweiligen Dokuments gebildet wurde, was die Ergebnisse zwar anfällig für statistische Ausreißer macht, aber eine sich jeweils zu eins addierende Topic-Verteilung ergibt, was für spätere Anwendungen notwendig ist.

Das Korpus wurde um häufig auftretende Wörter (Stopwörter) bereinigt (Boyd-Graber et al. 2015). Das Topic Model wurde mit MALLET[3] erstellt (2.000 Iterationen, Optimierung alle 20 Iterationen) und die dort integrierte Stopliste genutzt. Zudem wurden alle Wörter mit einer absoluten Häufigkeit größer-gleich 1.000 entfernt, wobei nach manueller Durchsicht bestimmte bedeutungstragende Begriffe davon ausgenommen wurden. Die Stopliste wurde nach einem ersten Testlauf um weitere häufig auftretende Wörter ergänzt. Zudem wurden aus Kompatibilitätsgründen Umlaute und Eszett entfernt. Die optimale Anzahl an Topics wurde zunächst mittels der R-Erweiterung *ldatuning* von Murzintcev (2016) geschätzt, allerdings lieferte der daraus resultierende Wert von etwa 30 nur sehr unspezifische Topics. Daher wurde die Topic-Anzahl sukzessive erhöht und schließlich ein Wert von 70 festgelegt, welcher einen Kompromiss zwischen Spezifität und Redundanz der Topics darstellt.

Ergebnisse

Die meisten der 70 Topics können unmittelbar als kohärente ökonomische Themen interpretiert werden.[4] Da einige Topics unterschiedliche Teilaspekte eines Oberthemas betreffen, wurden diese zu „Obertopics“ zusammengefasst, indem die entsprechenden Topic-Anteile addiert wurden. Als Kriterium wurde ihre begriffliche Ähnlichkeit, ausgedrückt in der Kosinus-Ähnlichkeit (Aletras und Stevenson 2014; Wiedemann 2016), herangezogen, wobei auch Topics gruppiert wurden, die zwar nur eine geringe Ähnlichkeit, dafür aber eine hohe inhaltliche Übereinstimmung aufwiesen. Damit reduzierte sich die Topic-Anzahl auf 10 Obertopics, von denen 8 große ökonomische Themenbereiche umfassen, sowie 12 Einzeltopics (Tabelle 1).[5]

Obertopic	Nr. der einbezogenen Topics	Token-Anteil		Einzeltopics (Topic-Nr.)	Token-Anteil
Allgemein-sprache	06, 09, 11, 26, 29, 44	25,08%		Europa (58)	0,87%
Arbeitsmarkt	08, 13, 20, 21, 24, 53, 40, 57	9,59		Methodik (10)	0,78
Finanz- und Eurokrise	03, 19, 27, 41, 69	4,71		Wettbewerb (14)	0,76
Geldpolitik	05, 23, 46, 48, 50, 54, 63, 66, 67	10,09		Expertise Demographie (38)	0,70
Öffentliche Finanzen	01, 15, 31, 32, 35, 42, 52, 60, 64, 68	12,54		Agrarpolitik (12)	0,62
Ost-/ Westdeutschland	07, 18, 37	3,38		Erneuerbare Energien (45)	0,56
Soziale Sicherung	00, 22, 39, 65	3,83		Produktivität & Innovation (49)	0,54
Verteilung	30, 47	1,77		Energie-versorgung (16)	0,52
Weltwirtschaft	28, 33, 43, 55, 59, 62	9,34		Expertise Lebensqualität (04)	0,47
Wirtschafts-entwicklung	02, 17, 25, 36, 56	12,63		Strukturwandel (51)	0,45
				Immobilien-markt (61)	0,39
				Bildung (34)	0,38
Summe		92,96			7,04

Tabelle 1: Topics der Jahresgutachten

Die Topic-Verteilungen können für einige grundsätzliche Aussagen über das Korpus herangezogen werden. Zum Vergleich des Inhalts zweier Dokumente wird in der Literatur die Jensen-Shannon-Divergenz (JSD) herangezogen, welche die Differenz der Topic-Verteilungen misst (Steyvers und Griffiths 2007). Abbildung 1 zeigt die thematische Divergenz zwischen einem Jahresgutachten und dem jeweiligen Vorjahresgutachten. Die JSD ist mit durchschnittlich 0,37 gering (maximale Divergenz besteht bei JSD=1), die Jahresgutachten weisen also eine hohe thematische Kontinuität auf. Allerdings ergeben sich einige thematische Brüche, wie etwa durch die Wiedervereinigung im Jahresgutachten 1990 oder die Finanzkrise im Jahresgutachten 2009. Die thematische Vielfalt lässt sich mittels eines in der Ökonomik genutzten Maßes zur Messung der Konzentration der Marktmacht von Unternehmen abbilden: dem Herfindahl-Hirschman-Index (HHI). Dieser ist definiert als die Summe der quadrierten Marktanteile aller Unternehmen, welche sich mit den Topic-Anteilen eines Dokuments vergleichen lassen. Somit kann auf Basis der Topic-Anteile eines Jahresgutachtens dessen „thematische Konzentration“ ermittelt werden. Wie anhand von Abbildung 1 deutlich wird, weisen die Jahresgutachten eine schwache und zudem abnehmende thematische Konzentration auf (maximale Konzentration, d.h. ein Monopol, besteht bei HHI=1).

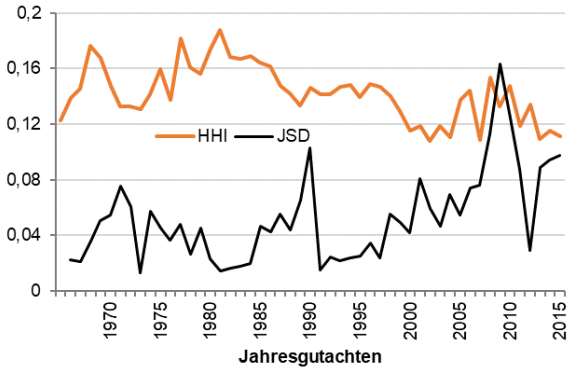
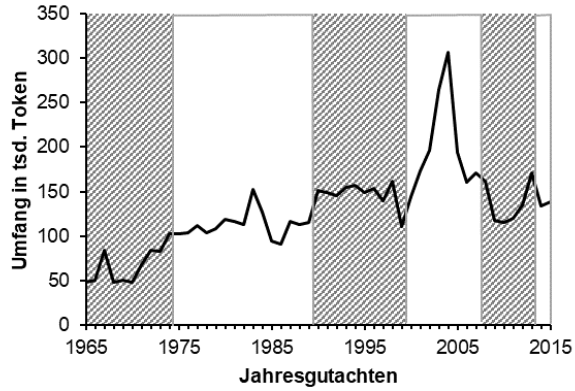


Abbildung 1. Topic-Divergenz und -Konzentration der Jahresgutachten

Dass die Gutachten nur eine geringe jährliche Themenveränderung aufweisen, erscheint angesichts der institutionellen und personellen Kontinuitäten des Rates durchaus plausibel (Schanetzky 2007). Nichtsdestotrotz stellt sich die Frage, ob und inwieweit nicht zumindest ein längerfristiger, eher gradueller Wandel zu beobachten ist. Um einen solchen zu bestimmen, wurden die Jahresgutachten analog zur Vorgehensweise bei Wiedemann (2016) auf Basis der Topic-Verteilungen in Cluster eingeteilt, wobei sich die in Abbildung 2 dargestellten sechs Cluster ergaben. Im zweiten Schritt wurden die durchschnittliche Topic-Verteilung für jedes Cluster berechnet.

	Cluster/ Periode	Token
1.	1965-1974	665.068
2.	1975-1989	1.685.157
3.	1990-1999	1.471.786
4.	2000-2007	1.610.708
5.	2008-2013	820.253
6.	2014-2015	271.820



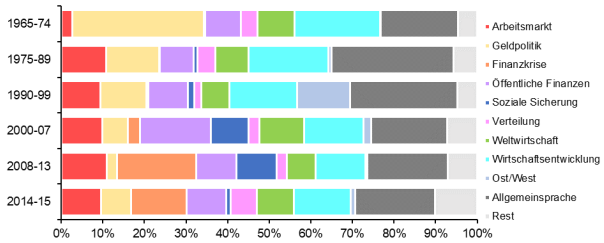


Abbildung 2: Topic-Verteilungen der Jahresgutachten nach Topic-Clustern
Quelle: eigene Darstellung in Anlehnung an Wiedemann (2016).

Zur Illustration seien drei Obertopics herausgegriffen. Die Entwicklung des Obertopics *Arbeitsmarkt* weist einen positiven Trend und eine hohe Korrelation mit der Entwicklung der Arbeitslosenquote auf (Korrelationskoeffizient: 0,6), besonders bis zu Beginn der 1990er Jahre. In anderen Worten korreliert der Umfang, in dem der SVR über das Thema Arbeitsmarkt schrieb, positiv mit der tatsächlichen Entwicklung auf dem Arbeitsmarkt. Der Ausschlag des Jahresgutachten 1977 erklärt sich mit einer ausführlichen Diskussion der Ursachen der Arbeitslosigkeit, vor allem mit einem hohen Anteil von Topic 40. Dieses enthält viele mit einer angebotsorientierten Erklärung von Arbeitslosigkeit[6] verbundene Begriffe wie „Investition“, „Kostenniveau“ oder „rentabel“.

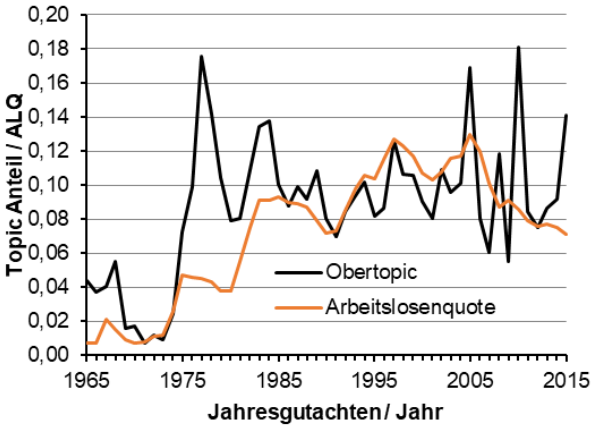


Abbildung 3: Obertopic Arbeitsmarkt Anmerkungen: Arbeitslosenquote bis 1991 früheres Bundesgebiet. Quelle: eigene Darstellung, Bundesagentur für Arbeit.

Auch das Obertopic *Geldpolitik* weist mit seinem fallenden Trend einen der entsprechenden ökonomischen Variablen, der Inflationsrate, ähnelnden Verlauf auf. Allerdings ist hier der Zusammenhang weniger stark als im Falle des Arbeitsmarkts (0,45), da das Obertopic neben der Preisniveauentwicklung auch andere Bereiche der Geldpolitik betrifft (erstere ist durch das Untertopic 67 repräsentiert, hier beträgt der Korrelationskoeffizient 0,63).

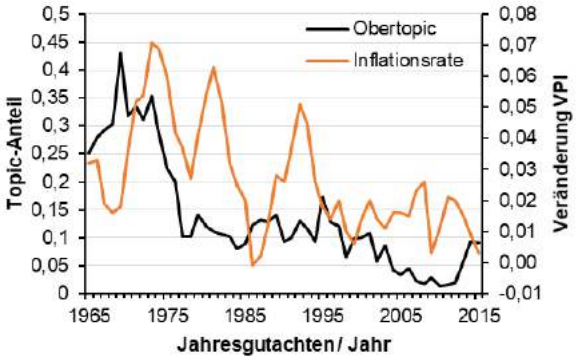


Abbildung 4: Obertopic Geldpolitik Anmerkungen: VPI = Verbraucherpreisindex des statistischen Bundesamts (bis 1990 Preisindex für die Lebenshaltung aller Privathaushalte), Veränderung ggü. Vorjahr. Quelle: eigene Darstellung, statistisches Bundesamt.

Als drittes Beispiel sei das Obertopic *Soziale Sicherung* genannt. Hier ist die Sozialleistungsquote eine naheliegende Vergleichsvariable, mit welcher der Topic-Anteil leicht positiv korreliert (0,51). Allerdings wird anhand der Entwicklung deutlich, dass sich der SVR erst seit Ende der 1990er Jahre in größerem Umfang mit der Sozialpolitik befasste, einer Zeit, die mit dem Antritt der rot-grünen Regierung und deren Reformpolitik ganz im Zeichen der Sozialpolitik stand. Den Obertopics *Öffentliche Finanzen*, *Wirtschaftsentwicklung* und *Weltwirtschaft* ließen sich ebenfalls ökonomische Variablen zuordnen, beispielsweise die Staatschuldenquote (0,44), die Wachstumsrate des Bruttoinlandsprodukts (0,28) und der Leistungsbilanzsaldo (-0,06), für die verbleibenden Obertopics erscheint eine solche Zuordnung hingegen wenig sinnvoll.[7]

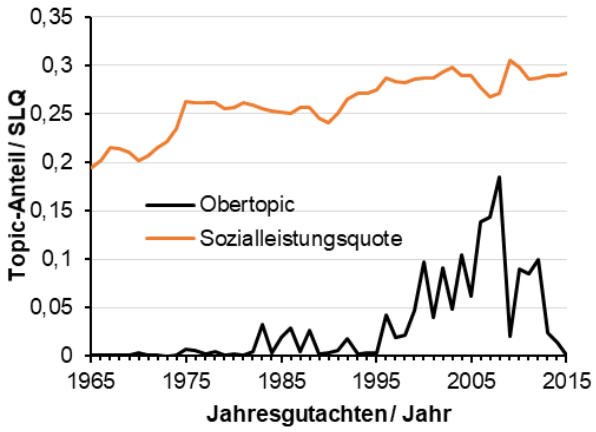


Abbildung 5: Obertopic Soziale Sicherung Anmerkungen: Sozialleistungen in Prozent des BIP, bis 1990 unrevidierte Werte, ab 2009 einschließlich privater Krankenversicherung, bis 1991 früheres Bundesgebiet. Quelle: eigene Darstellung, Bundesministerium für Arbeit und Soziales

Mit der Arbeit von Wegner (1985) liegt eine Vergleichsstudie zwischen den Jahresgutachten und Jahreswirtschaftsberichten vor, in denen die Bundesregierung u.a. zu den Jahresgutachten Stellung bezieht. Mittels der JSD kann dieser Vergleich aus einer neuen Perspektive angestellt und zeitlich erweitert werden. Da die JSD einer Topic-Verteilung über dieselben Topics bedarf, wurde ein weiteres Model für die Jahresgutachten und Jahreswirtschaftsberichte entworfen, aus dem vergleichbare Topics resultierten.[8] Die in Abbildung 6 dargestellte Entwicklung zeigt, dass beide Subkorpora grundsätzlich eine mittlere bis schwache Divergenz aufweisen, sich also thematisch ähneln, wobei durchaus Schwankungen zu beobachten sind. Zusätzlich wurde die sprachliche Ähnlichkeit zwischen Jahresgutachten und Jahreswirtschaftsbericht auf Basis der Kosinus-Ähnlichkeit bestimmt. Nach Abzug der Stopwörter ergibt sich eine recht hohe, tendenziell abnehmende Kosinus-Ähnlichkeit, d.h. Jahresgutachten und Jahreswirtschaftsbericht nutzen ein ähnliches Vokabular. Die Korrelation zwischen JSD und Kosinus-Ähnlichkeit ist leicht negativ (-0,3), was plausibel erscheint: Je größer die Divergenz, desto kleiner ist die Ähnlichkeit.

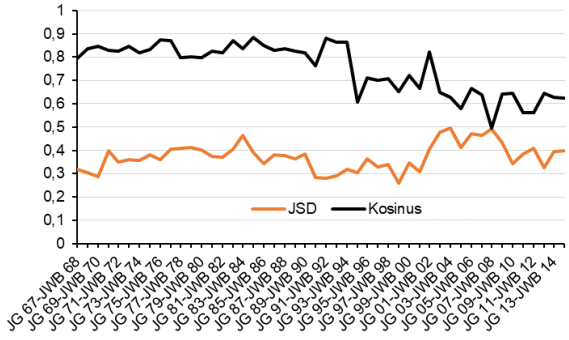


Abbildung 6: JS-Divergenz und Kosinus-Ähnlichkeit von Jahresgutachten und Jahreswirtschaftsbericht

Fazit und Ausblick

Die Frage, ob sich die Jahresgutachten des SVR für eine Untersuchung mit Topic Models eignen, kann eindeutig bejaht werden. Für die gewählten Modellparameter ergeben sich interpretierbare, sinnvolle Topics, deren zeitliche Entwicklung auf Basis eines Vergleichs mit entsprechenden ökonomischen Variablen und generellen historischen Entwicklungen plausibel erscheint. Der forschungspraktische Nutzen ist in diesem Fall der folgende: (1) können die für ein bestimmtes Thema relevanten Textpassagen identifiziert werden, was das für historische Forschungsfragen unabdingbare *close reading* erheblich vereinfacht. Die Makroperspektive des Topic Models hilft, Relevantes von weniger Relevantem abzugrenzen und den Blick auf strukturelle Veränderungen des Korpus zu lenken. (2) ermöglicht bzw. erleichtert diese Makroperspektive diachrone Fragestellungen. Beispielsweise lassen sich durch die verschiedenen Bestandteile der Obertopics Veränderungen in der Betrachtungsweise eines ökonomisches Themengebiets, etwa des Arbeitsmarkts, identifizieren. Der empirische Zugang erlaubt zum einen den Abgleich mit Ergebnissen historisch-qualitativer Arbeiten zum Sachverständigenrat. Zum anderen werden die Jahresgutachten quantitativen Forschungsansätzen zugänglich gemacht. So ließe sich etwa untersuchen, in welcher zeitlichen bzw. kausalen Beziehung ökonomische Realität und deren Analyse durch den SVR zueinander stehen (Lüdering und Winker 2016); (3) eröffnen sich neue Möglichkeiten des Vergleichs mit anderen Quellen, wie den Jahreswirtschaftsberichten oder der Presseberichterstattung, was gerade für die Frage nach dem Beratungserfolg der Wirtschaftsweisen vielversprechend erscheint.

Fußnoten

1. Das erste Jahresgutachten sowie einige SG wurden aus hier nicht zu thematisierenden Gründen ausgeschlossen.
2. Siehe http://pdok.bundestag.de/ Zuletzt geprüft am 04.09.2018.
3. Siehe http://mallet.cs.umass.edu/ Zuletzt geprüft am 04.09.2018.
4. Einsehbar unter: https://drive.google.com/open?id=1H3RTw-wu-ggIXFocFwFDJahresgutachten9vwlypFfko
5. Eine Topic-Karte ist einsehbar unter https://drive.google.com/open?id=1p4KBCAahA4zq4z4iTQ5UzcGXqp6qG35t
6. Diese lautet verkürzt: Arbeitslosigkeit resultiert aus schlechten Angebotsbedingungen, insbesondere aus einem Mangel an Investitionen seitens der Unternehmen aufgrund zu hoher Arbeitskosten. Demnach kann Arbeitslosigkeit durch Lohnzurückhaltung reduziert werden.
7. Grundsätzlich stellt sich hier in Anbetracht des allgemeinen Charakter der Obertopics natürlich die Schwierigkeit der Auswahl geeigneter ökonomischer Variablen.
8. Einsehbar unter: https://drive.google.com/open?id=1qNLV3j2B2yE14QiH06SihSh_rCZwvYrL . Ein Dispersionsplot der Suchbegriffe „Sachverständigenrat“ und „Jahresgutachten“ im JWB-Korpus findet sich unter https://drive.google.com/open?id=1yiOT4NOrchueh2FXgsHN3V3wUzSDsZyR

Bibliographie

Aletras, Nikolaos; Stevenson, Mark 2014: *Measuring the Similarity between Automatically Generated Topics*. In: *Proceedings of the 14th Conference of the European Chapter of the ACL*, S. 22–27.

Blei, David; Ng, Andrew Y.; Jordan, Michael I. 2003: *Latent Dirichlet Allocation*. In: *Journal of Machine Learning Research* 3, S. 993–1022.

Blei, David M. 2012: *Probabilistic Topic Models*. In: *Communications of the ACM* 55 (4), S. 77–84.

Boyd-Graber, Jordan; Hu, Yuening; Mimno, David 2017: *Applications of topic models*. Boston: now (Foundations and trends in information retrieval).

Boyd-Graber, Jordan; Mimno, David; Newman, David J. 2015: *Care and Feeding of Topic Models*. In: **Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva und Stephen E. Fienberg (Hg.)**: *Handbook of Mixed Membership Models and Their Applications*. Boca Raton, S. 225–274.

Heyer, Gerhard 2009: *Introdution to TMS 2009*. In: **Gerhard Heyer (Hg.)**: *Text mining services.* Leipzig: LIV (Leipziger Beiträge zur Informatik, Bd. 14), S. 1–14.

Jockers, Matthew Lee 2013: *Macroanalysis. Digital Methods and Literary History*. Urbana Ill.

Lüdering, Jochen; Winker, Peter 2016: *Forward or Backward Looking? The Economic Discourse and the Observed Reality*. In: *Journal of Economics and Statistics* 236 (4), S. 483–515.

Murzintcev, Nikita 2016: *ldatuning (R package)*. Online verfügbar unter https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf, zuletzt geprüft am 19.03.2018.

Schanetzky, Tim 2007: *Die große Ernüchterung. Wirtschaftspolitik Expertise und Gesellschaft in der Bundesrepublik 1966 bis 1982.* Berlin.

Steyvers, Mark; Griffiths, Tom 2007: *Probabilistic Topic Models.* In: **Thomas K. Landauer, Danielle S. McNamara, Simon Dennis und Walter Kintsch (Hg.)**: *Handbook of Latent Semantic Analysis.* Hoboken, S. 427–448.

Strätling, Ansgar 2001: *Sachverständiger Rat im Wandel. Der theoretische Argumentationshintergrund des Sachverständigenrats zur Begutachtung der Gesamtwirtschaftlichen Entwicklung zur Beschäftigungspolitik von 1964-1999.* Marburg.

Wegner, Klaus 1985: *Im Blickpunkt: Sachverständigenrat und Konjunktur- und Wachstumspolitik der Bundesregierung seit 1964.* Frankfurt.

Wehrheim, Lino 2018: *Economic History Goes Digital: Topic Modeling the Journal of Economic History.* In: *Cliometrica. DOI:* 10.1007/s11698-018-0171-7.

Wiedemann, Gregor 2016: *Text Mining for Qualitative Data Analysis in the Social Sciences. A Study on Democratic Discourse in Germany*. Wiesbaden.

Wandel in der Wissenschaftskommunikation? Ergebnisse der Umfrage bei den Bloggenden von de.hypotheses.org

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

Die einfache Zugänglichkeit von Wissenschaftsblogs ab der Jahrtausendwende ermöglicht es Forschenden, selbst zu entscheiden wann, wo und was sie veröffentlichen wollen. Diese selbstbestimmte Übernahme eines wissenschaftlichen Publikationsraums ist ein spektakulärer Schritt, ähnlich wie die Erfindung von "Essays" durch Montaigne im 16. Jahrhundert oder die Entstehung der "Gelehrtenrepublik" ab der Mitte des 17. Jahrhunderts (König 2015: 58). Wissenschaftliche Blogs sind Orte, in denen aus laufenden Forschungsprojekten kommuniziert und mit der Fachcommunity diskutiert werden kann. Sie ermöglichen Einblicke in das Labor oder die Werkstatt der Forschenden und zeigen damit "Wissenschaft im Entstehen" (Mounier 2013). Der wissenschaftliche Austausch über Blogartikel, Kommentare und Links ist interaktiv, schnell und direkt, in einer einzigartigen Weise, die in anderen Publikationsformaten wie Mailinglisten oder Zeitschriften nicht möglich ist oder nicht praktiziert wird. Blogs können als das fehlende Bindeglied zwischen mündlicher Kommunikation auf Konferenzen oder in Universitätsseminaren und schriftlicher Kommunikation in traditionellen Artikeln oder Rezensionen angesehen werden. Sie ermöglichen es Forschenden, eine direkte Verbindung zugleich zu ihren Peers und zu ihren Studierenden herzustellen und darüber hinaus mit Journalisten und der interessierten Öffentlichkeit in Kontakt zu treten.

Seit 2012 sorgt das Blogportal für die Geistes- und Sozialwissenschaften de.hypotheses.org für eine florierende Blogpraxis im deutschsprachigen Raum. Das Portal ist Teil der europäischen Plattform hypotheses und bietet als zentraler Einstiegsort kostenlose und werbefreie Blogs an für Forschende, die technische Updates, Hosting und Sicherheitsfragen nicht selbst übernehmen können oder wollen und Teil einer Community sein möchten. Community Management und Redaktion bewerben die besten aktuellen Beiträge in den sozialen Medien und auf der Startseite der Plattform. Sie bieten außerdem technischen und graphischen Support und beantragen bei den Nationalbibliotheken die Zuteilung einer ISSN für die Blogs. Die Blogbeiträge sind mit Permalinks versehen und die Inhalte der Plattform werden von BnF und DNB archiviert. Derzeit sind auf der deutschsprachigen Seite rund 350 Wissenschaftsblogs vereint.

Obwohl es sich um ein relativ neues Phänomen handelt, ist die Forschung zur wissenschaftlichen Nutzung von sozialen Medien in den letzten Jahren stark gewachsen (für einen umfassenden Überblick siehe Sugimoto et

al. 2017). Die empirische Forschung zur Nutzung von sozialen Medien erfolgt durch Fragebögen, qualitative Interviews und teilnehmende Beobachtungsstudien. Diese Studien untersuchen Praktiken von innen heraus und fragen Forschende nach ihren Methoden, Vorlieben oder Widerständen in Bezug auf die wissenschaftliche Nutzung von sozialen Medien (siehe z.B. Ponte und Simon (2011), hauptsächlich für Großbritannien; Bader, Fritz und Gloning (2012) und Pscheida et al. (2013) für Deutschland). Andere Forschungsbereiche untersuchen digitale Praktiken und Online-Communities von außen, z.B. über die Analyse von Inhalten und Sprache oder über die Analyse von Netzwerken anhand von Links und Kommentaren. Beide Arten von Studien zeigen zumeist eine Vielzahl von unterschiedlichen Nutzungen, Zwecken und Motiven der sozialen Medien, je nach Plattform, akademischem Rang, Status, Geschlecht und Alter der Forschenden sowie unterschiedlich nach Disziplinen, Ländern und geografischen Regionen.

Statistiken und Beobachtungen der geisteswissenschaftlichen Blogs geben einen Einblick in die Motivationen der Bloggenden, in ihre Blogpraktiken wie auch in ihre Kommunikation über Kommentare und Verlinkungen. Abgesehen von den beiden erwähnten bereits gealterteten Umfragen allgemein zu sozialen Medien in der Wissenschaft (Bader, Fritz und Gloning, 2012, sowie Pscheida et al. 2013), gab es anders als im angelsächsischen Raum (Jarreau 2015) im deutschsprachigen Raum noch keine spezifische Befragung geisteswissenschaftlicher Bloggenden im größeren Ausmaß. Diese Lücke wird durch eine Umfrage geschlossen, die im Herbst 2018 bei den rund 350 Blogs von de.hypotheses sowie auch darüber hinaus durchgeführt wird und deren Ergebnisse der Vortrag vorstellen möchte.

Die Umfrage zielt in erster Linie darauf, Gründe für das Wissenschaftsbloggen sowie konkrete Praktiken des geisteswissenschaftlichen Bloggens abzufragen und darüber mögliche Änderungen im Publikations- und Kommunikationsverhalten von Geisteswissenschaftlerinnen und -wissenschaftlern empirisch gestützt zu ermitteln. Als empirische Studie mit medientheoretischem Bezug knüpft der Vortrag damit direkt an das Tagungsthema an.

Die Hauptfragebereiche der Umfrage beziehen sich auf Motivationen für das Bloggen, auf Inhalte, Zeitaufwand, Publikum und formale Gestaltung sowie auf den „Erfolg“ der Blogs in Bezug auf Kommentare, Rückmeldungen und Zugriffsstatistiken. Das Abfragen von Personendaten soll ermöglichen, diese Antworten mit akademischem Rang, Alter und Geschlecht der Bloggenden zurückzukoppeln und darüber etwa gender- und statusspezifische Praktiken ausmachen zu können. Dagegen geht es nicht um die Abfrage der Zufriedenheit der Bloggenden mit der Plattform de.hypotheses selbst. Folgende Themenblöcke werden u.a. angesprochen:

Die Nutzung wissenschaftlicher Blogs ist je nach Strategie und Zielen der Forschenden sehr unterschiedlich. In der Kategorie "Über das Blog" auf den Wissenschaftsblogs von de.hypotheses bekommt man dazu einen Einblick. Bloggende Forschende nennen als erste Motivation den Wunsch, ihr Forschungsthema zu diskutieren, ihre Online-Reputation zu verbessern, sich in der Wissenschaft zu positionieren und Netzwerke zu pflegen (König 2015: 59). Darüber hinaus wollen die Forschenden das Schreiben üben oder ihren Schreibstil verbessern und sich kreativ ausdrücken. Andere Untersuchungen deuten darauf hin, dass Bloggen Forschenden das Gefühl vermittelt, in ihrer Arbeit mit anderen verbunden zu sein (Mewburn und Thomson 2013: 1107). Blogs können als Dokumentation für Forschungsprojekte dienen, als eine Art digitaler Zettelkasten, der über Schlagwörter und Kategorien strukturiert und öffentlich zugänglich ist. Diese Angaben zur Motivation und Gründe des Bloggens werden in der Umfrage abgefragt, wobei es zugleich auch darum gehen wird, ob diese Ziele subjektiv nach Empfinden der Einzelnen erreicht werden.

Ein weiterer Fragenblock der Umfrage behandelt die internen Abläufe bei der Veröffentlichung auf Wissenschaftsblogs. Einige Blogs funktionieren ähnlich wie Zeitschriften: Sie haben eine Redaktion, die Autorinnen und Autoren einlädt, Artikel redaktionell bearbeitet und sicherstellt, dass Blogbeiträge in traditionellen Bibliographien und Bibliothekskatalogen katalogisiert werden. Aber auch in Einzelblogs publizieren Autorinnen und Autoren oftmals erst, nachdem die Beiträge von einer anderen Person gegengelesen worden sind. Dies schließt an die Beobachtung an, dass Forschende ihre Blogs als Orte der Selbstpublikation nicht leichtsinnig befüllen, sondern sich viele Gedanken machen, was sie wann, wie und in welcher Form publizieren. Vielen Forschenden fällt es schwer, unfertige oder aufkommende Ideen zu veröffentlichen. Sie haben Angst davor, sich zu irren und wollen vermeintliche Sackgassen nicht öffentlich machen. Die Angst vor Plagiaten hindert sie ebenso daran, über aktuelle Erkenntnisse und aktuelle Projekte zu bloggen. Welche strategischen und konzeptionellen Grundideen geisteswissenschaftliche Bloggende verfolgen soll ebenso wie die Organisation der redaktionellen Zwischenschritte vor der Veröffentlichung durch die Umfrage deutlich werden.

Es gibt eine große Vielfalt an Inhalten, die in wissenschaftlichen Blogs veröffentlicht werden. Einige Bloggende schreiben grundsätzlich nur über ihr Forschungsthema. Andere diskutieren wissenschaftliche Arbeitspraktiken, geben Karriereberatung oder nutzen ihr Blog zur Begleitung der Lehre. Häufig wird in wissenschaftlichen Blogs die akademische Kultur allgemein kritisiert (Mewburn und Thomson 2013: 1110). Alles in allem lassen Blogs den Forschenden als hybride Person erscheinen und zeigen, dass die akademischen Interessen von Wissenschaftlerinnen und Wissenschaftlern viel breiter sind als die in klassischen Medien veröffentlichte Forschung das widerspiegelt (Mewburn und Thomson 2013: 1114). Darüber hinaus unterstreicht der unterschiedliche und informelle Stil der Blogartikel die Vielfalt des wissenschaftlichen Schreibens und die Vielfalt der Perspektiven. In wissenschaftlichen Blogs ist es möglich, in der ersten Person Singular zu schreiben, engagiert, witzig, kreativ und essayistisch zu schreiben, Smileys oder Strikes zu verwenden, Code, Bilder und Videos einzubetten und damit multimedial zu publizieren (König 2015: 64-65). Die Umfrage soll Aufschluss geben, ob und in welchem Umfang die bloggenden Geisteswissenschaftlerinnen und Geisteswissenschaftler von den stilistischen Freiheiten des Genres profitieren oder ob sie sich freiwillig an traditionellere Formate und Publikationsrhythmen anpassen.

Einige explorative Studien deuten darauf hin, dass sich Blogs weder bei der Sprachwahl noch bei der Themenwahl an ein Laienpublikum wenden (Mahrt und Puschmann 2014: 4; Mewburn und Thomson 2013: 1113). In der Umfrage wird gezielt bei der deutschsprachigen Community von hypotheses abgefragt, an welches Publikum sich die Forschenden in der Regel wenden, ob der Transfer von Forschungsergebnissen in die Öffentlichkeit zu den Zielen gehört und ob sich die

Bloggenden sprachlich auf ein Laienpublikum einstellen oder ob sie auf Medienaufmerksamkeit zielen. Es gibt Rückmeldung von Bloggenden, wonach Blogbeiträge als Vorstufe für Peer Review-Artikel gesehen werden und Forschende aufgrund von Blogbeiträgen aufgefordert worden sind, diese zu vollständigen Artikeln auszubauen. Wie verbreitet dieses Phänomen ist, soll die Umfrage empirisch zeigen.

Bloggen wird als eine hochgradig interaktive Praxis angesehen (Mahrt und Puschmann 2014: 6), obwohl Kommentare zu geistes- und sozialwissenschaftlichen Blogs knapper geworden sind, u.a. weil die Diskussion von Blogartikeln auf Twitter, Facebook oder andere soziale Medien verschoben wurde (König 2015: 72). Auf dem Blog-Hub SciLogs mit mehrheitlich Bloggenden aus den Naturwissenschaften erhalten Artikel durchschnittlich fünf Kommentare. Blogging-Stars wie der österreichische Astronom Florian Freistätter wiederum erhalten regelmäßig zwischen 50 und 100 Kommentare pro Artikel (Lobin 2017: 226). Die aktuellen Statistiken für die Plattform de.hypotheses liegen zwar vor, bei den Bloggenden wird aber nachgefragt, wie sie mit den Kommentaren umgehen, ob sie selbst welche schreiben, ob die Inhalte der Kommentare überwiegend positiv, neutral oder negativ sind, welche andere Form von Rückmeldungen sie für ihre Blogbeiträge erhalten und wie wichtig ihnen diese für das Einschätzen des eigenen Erfolgs sind.

Bibliographie

Bader, Anita / Fritz, Gerd / Gloning, Thomas (2012): Digitale Wissenschaftskommunikation 2010-2011. Eine Online-Befragung. Gießen, http://geb.uni-giessen.de/geb/volltexte/2012/8539/ .

Jarreau, Page (2015): *All the Science That Is Fit to Blog. An Analysis of Science Blogging Practices*. LSU Doctoral Dissertations 1051, https://digitalcommons.lsu.edu/cgi/viewcontent.cgi?article=2050&context=gradschool_dissertations .

König, Mareike (2013): *Die Entdeckung der Vielfalt: Geschichtsblogs auf der internationalen Plattform hypotheses.org*, in: **Peter Haber / Eva Pfanzelter (eds.):** *Historyblogosphere. Bloggen in den Geschichtswissenschaften*. München: Oldenbourg 181–197.

König, Mareike (2015): *Herausforderung für unsere Wissenschaftskultur: Weblogs in den Geisteswissenschaften*, in: **Wolfgang Schmale (ed.):** *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität*. Stuttgart: Steiner 57-74.

Lobin, Henning (2017): *Aktuelle und künftige technische Rahmenbedingungen digitaler Medien für die Wissenschaftskommunikation*, in: **Peter Weingart / Holger Wormer / Andreas Wenninger / Reinhard F. Hüttl (eds.):** *Perspektiven der Wissenschaftskommunikation im digitalen Zeitalter*. Weilerswist: Velbrück 223-258.

Mahrt, Merja / Cornelius Puschmann (2014): *Science Blogging: an exploratory study of motives, styles, and audience reactions*, in: Journal of Science Communication 13/3: A05. https://jcom.sissa.it/archive/13/03/JCOM_1303_2014_A05 (accessed August 31, 2018).

Mewburn, Inger / Pat Thomson (2013): *Why Do Academics Blog? An Analysis of Audiences, Purposes and Challenges*, in: Studies in Higher Education 38/8: 1105–1119.

Mounier, Pierre (2013): *Die Werkstatt öffnen: Geschichtsschreibung in Blogs und sozialen Medien*, in: **Peter Haber / Eva Pfanzelter (eds.):** *Historyblogosphere. Bloggen in den Geschichtswissenschaften*. München: Oldenbourg 51-59.

Ponte, Diego / Simon, Judith (2011): *Scholarly Communication 2.0: Exploring Researchers' Opinions on Web 2.0 for Scientific Knowledge Creation*, Evaluation and Dissemination, Serials Review 37(3): 149-156.

Pscheida, Daniela / Albrecht, Steffen / Herbst, Sabrina / Minet, Claudia / Köhler, Thomas (2013): *Nutzung von Social Media und onlinebasierten Anwendungen in der Wissenschaft. Erste Ergebnisse des Science 2.0-Survey 2013 des Leibniz-Forschungsverbunds "Science 2.0"*, https://d-nb.info/1069096679/34.

Sugimoto, Cassidy R. / Sam Work / Vincent Larivière / Stefanie Haustein (2017): *Scholarly Use of Social Media and Altmetrics: a Review of the Literature*, in: Journal of the Association for Information Science and Technology, 68/9: 2037–2062.

Wie katalogisiert man eigentlich virtuelle Realität? Überlegungen zur Dokumentation und Vernetzung musealer Objekte und digitaler Vermittlungsformate

Diehr, Franziska

f.diehr@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz, Deutschland

Glinka, Katrin

k.glinka@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz, Deutschland

Forschung im Bereich der Digital Humanities ist qua Selbstverständnis von interdisziplinären Ansätzen geprägt. Nur durch die Kombination von vielfältigen fachwissenschaftlichen Expertisen kann es gelingen, Erkenntnisse zu gewinnen, welche bei Beschränkung auf nur eine disziplinäre Herangehensweise und Methode nicht - oder nicht im gleichen Maße - möglich gewesen wäre. So jedenfalls die Verheißung der digitalen Geisteswissenschaften. Ob die Potenziale digitaler Forschungsansätze und Methoden in den DH bereits vollends verwirklicht werden, vermag dieser Beitrag nicht zu bewerten. Dass digitale Technologien jedoch auch außerhalb der Hochschulen und Forschungseinrichtungen fraglos den Raum des Möglichen entscheidend erweitern, ist das Grundverständnis, welches für das deutschlandweite Verbundprojekt 'museum4punkt0' richtungsweisend ist. Digitale Technologien eröffnen Museen neue Formate der

Interaktion, Interpretation und Kommunikation. Im Vergleich zu objektbezogener Forschung im Museum, die durch den Einsatz digitaler Erschließungssysteme, der Nutzung und Publikation von digitalen Reproduktionen und der Integration von digital gestützten Untersuchungsmethoden eine methodische Erweiterung erfährt, wird die Entwicklung digitaler Kommunikations- und Vermittlungsformate bisher eher selten als transdisziplinärer Forschungsauftrag im Museum verstanden (Glinka 2018b).

Experimentierlabor für digitale Anwendungen im Museum

Mit museum4punkt0 wurde 2017 erstmals in Deutschland ein museales Forschungsprojekt initiiert, welches Kulturinstitutionen verschiedener Sparten, Größen und institutionellen Strukturen mit dem Ziel der Entwicklung und Beforschung digitaler Anwendungen in einem Verbund vereint.[1] Zentrales Merkmal des Verbundes ist die Vernetzung und gegenseitige Unterstützung der beteiligten Institutionen bei der Entwicklung und Evaluation von digitalen Vermittlungs- und Kommunikationsangeboten (Glinka 2018a). Untersucht wird, wie neueste digitale Technologien effektiv für die Aufgaben von Museen, insbesondere in der Wissensvermittlung, nutzbar gemacht werden können. In modular strukturierten Teilprojekten entstehen digitale Kommunikations- und Vermittlungsformate mit Fokus auf Virtual- und Augmented Reality, 3D-Digitalisierung, Gamification sowie weiteren Formen der digitalen Kommunikation.[2]

Die Teilprojekte fokussieren die museale Praxis der jeweiligen Institutionen und entwickeln themenspezifische Angebote, zugeschnitten auf ihre Publikumsgruppen und Vermittlungsziele. Dabei greifen die Fallstudien auf Sammlungsobjekte und das inhärente Wissen der jeweiligen Museen zurück und erschaffen neue Arten der digitalen Narration, um Besucher*innen und Nutzer*innen einen multimedialen, -modalen und -perspektivischen Zugang zu Museumsbeständen zu ermöglichen. Die narrativen Formate nutzen Spezifika einzelner Objekte und Wissensbestände, um komplexe Inhalte für unterschiedliche Besucher- und Nutzergruppen verständlich zu vermitteln (Navarro/ Fuhrmann 2018). Hier fließen kreativ-künstlerisches Design und wissenschaftlich-kuratorische Konzeption ineinander, die von museumspädagogischen Fragestellungen geleitet werden. Begleitend werden zudem Ansätze der Nutzer- und Rezeptionsforschung entwickelt, welche methodische Ähnlichkeiten zwischen klassischer Besucherforschung und Ansätzen des human-centred Design und Forschung aus dem Bereich der Mensch-Maschine Interaktion (HCI) zum Ausgangspunkt nimmt, um diese im Kontext der Forschung im Museum in einen Dialog zu bringen (Bauer 2018).

Von den spezifischen Fallstudien losgelöst, werden forschungsgetriebene Ansätze wie u. a. webbasierte Methoden der Wissensvermittlung, Strategien der digitalen Kommunikation sowie Potenziale der Informationsvernetzung und Wissensgenerierung in dem Teilprojekt 'V - Virtueller Raum' erprobt. Hier werden zusätzliche Themen und Ansätze, welche bisher nicht von den museal-fokussierten Teilprojekten abgedeckt werden, teilweise in Kooperation mit assoziierten Partnern, bearbeitet.

Neben weiteren Formaten entsteht im Teilprojekt V ein gemeinsamer Objektkatalog, welcher als öffentlich zugängliche Webanwendung primär die Funktion hat, die im Verbund neu entstehenden Digitalisate und digitalen Produkte zu aggregieren, zu vernetzen und mit Wissen anzureichern. Darüber hinaus wird ein weiteres Forschungsinteresse verfolgt, welches im Fokus des vorliegenden Beitrags steht: Zentral beschäftigt uns die Frage, wie heterogene (digitale) Objekttypen, -formate und Sammlungsbestände in einer virtuellen Umgebung aggregiert, vernetzt und für weitere Nutzungskontexte, z.B. der Wissensvermittlung und der musealen Forschung, eingesetzt werden können. Dass dies mit Blick auf die Vielzahl an Technologien und Medien, welche im Verbundprojekt zum Einsatz kommen, grundlegende Fragen der Wissensrepräsentation aufwirft, möchten wir im Folgenden ausführen.

Eine gemeinsame Wissensbasis für übergreifende Fragestellungen

Die Heterogenität der beteiligten Museen schlägt sich auch in der Vielfalt der Objektarten nieder: Von Lebendtieren über Oral-History Interviews, filmischen Repräsentationen von Fastnachtsbräuchen bis hin zu Gemälden und technischen Geräten ist die ganze Bandbreite musealer Sammlungs- und Forschungsgegenstände vertreten. Diese Vielfalt setzt sich auch in den generierten digitalen Formaten fort: diese reichen von Virtual und Augmented Reality über Chatbots, webbasierte Interaktionsformate und Citizen Science Anwendungen bis hin zu Browsergames. Diese Diversität soll bei der Modellierung des Objektkatalogs berücksichtigt werden und auch erhalten bleiben. Aus informationswissenschaftlicher Sicht stellen sich Fragen zur Konstruktion eines solchen Modells: Wie kann anhand heterogener Objekttypen eine gemeinsame Wissensbasis für transdisziplinäre Fragestellungen modelliert werden? Wie können disziplinspezifische Beschreibungskategorien beibehalten werden und wie müssen diese aufbereitet sein, um für übergreifende museale Fragestellungen, z. B. der Museumspädagogik oder der Rezeptions- und Medienforschung, nutzbar zu sein? Wie können Vermittlungsstrategien zu einer direkten Generierung neuer Erkenntnisse führen und wie kann multiperspektivisches Wissen in den Objektkatalog eingebunden werden?

Potenziale musealer Objektdokumentation für erweiterte Forschungszugriffe

Der Internationale Museumsrat ICOM definiert die musealen Kernaufgaben als Sammeln, Bewahren und Dokumentieren, Forschen und Vermitteln (ICOM 2010: 29). Dabei spiele das Sammlungsmanagement für die Erfüllung der Mission der Museen eine zentrale Rolle, so Wallace:

> "Collections management must be understood as a means to an end — not the end in itself. It exists as a process to enable museums to fulfil their broader social remit by

> maximising the use of collections, giving collections and users a voice and reflecting diverse experiences." (Wallace 2001: 83)

Allerdings ist in Bezug auf jene Systeme, die im Sammlungsmanagement genutzt werden, festzustellen, dass kaum eine Verschränkung mit den weiter gefassten musealen Aufgaben stattfindet. Die gängige Museumspraxis zeigt, dass Sammlungs- und Objektdokumentationssysteme vor allem dem Nachweis des Museumsbestandes dienen (Srinivasan et al. 2009: 265). Konventionelle Sammlungsmanagementsysteme bieten primär Zugriffe zur Objektverwaltung, vorrangig für Restaurierungs- und Ausstellungsvorhaben (Cameron 2003: 331). Forscher*innen im Museum greifen für Recherchen auf die Bestandsdokumentation zu, um beispielsweise Informationen zum Zustand des Objekts oder seines aktuellen Verbleibs zu erhalten. Zur Durchführung von objekt- und sammlungsbasierter Forschung werden zusätzliche Materialien wie Archivalien und Sekundärliteratur herangezogen. Das durch die Forscher*innen generierte Wissen findet aber meist keinen Rückfluss in die Sammlungsdokumentation und verbleibt isoliert in Ausstellungskatalogen, fachwissenschaftlichen Beiträgen und anderen Publikationsformaten. Dies gilt ebenfalls für in Vermittlungsangeboten generierte Objektkontexte: Zwar wird die Verwendung von Musealien in Ausstellungen dokumentiert, jedoch finden begleitende Texte, Tonspuren aus Audioguides oder andere multimediale Formate meist keinen Weg in die Dokumentationssysteme. Sie entziehen sich somit zukünftigen Zugriffen und sind für die Erstellung neuer Angebote schwer nachnutzbar.

Weitaus komplexer gestaltet es sich bei Formaten wie Virtual Reality Anwendungen: Digitalisate, 3D-Scans, Videosequenzen und Kontextinformationen werden miteinander kombiniert, arrangiert und inszeniert. Konventionelle Museumsdokumentationssysteme wären bei der Katalogisierung komplexer VR-Anwendungen vor große Hürden gestellt, da sie auf die Erfassung grundlegender Metadaten traditioneller Musealien ausgerichtet sind (Srinivasan et al. 2009:268, Cameron / Robinson 2007:166).

Diese Beispiele zeigen, wie unzulänglich die bisherigen Systeme für übergreifende Kontexte sind. Cameron und Robinson fordern dazu auf, traditionelle Dokumentationsformen zu überdenken:

> "The challenge is to revisit the current epistemological foundations on which documentation is formulated and to consider how diverse cultural and theoretical ideas, for example polysemic interpretive models (ones that recognize the inherent pluralistic meanings of objects), might revise documentation, taking account of [...] technological potentialities." (Cameron / Robinson 2007:169)

Eine Möglichkeit, konventionelle Katalogisierungsmechnismen zu erweitern, besteht in der Reintegration extern erzeugter Wissensbestände und Kontexte in die museale Bestandsdokumentation. Mit der Entwicklung des Objektkatalogs im Rahmen von museum4punkt0 wird erprobt, wie Museumsdokumentationssysteme mit Forschungs- und Vermittlungswerkzeugen verschränkt werden können, um extern generierte Informationsressourcen und Wissensbestände für eine erweiterte Objektdokumentation nachzunutzen. Cameron und Robinson fragen "what considerations are important in reassessing traditional museum documentation models in the context of Web-based digital access?" (2007:166). Dazu evaluieren wir, welche Schnittstellen ein solcher webbasierter Objektkatalog benötigt und welche Mapping-Szenarios dafür entwickelt werden müssen. Aus Sicht der Wissensrepräsentation untersuchen wir, wie Daten modelliert und strukturiert sein müssen, damit sie zwischen verschiedenen Anwendungen semantisch verlustfrei austauschbar sind.

Mit dem Objektkatalog als Schnittstelle zwischen Forschungsmethoden und -werkzeugen der digitalen Geisteswissenschaften und den Daten des kulturellen Erbes untersuchen wir, wie museale Datenbestände aufbereitet sein müssen, um als Forschungsdaten für (inter-)disziplinäre Fragestellungen nutzbar zu sein. Weiterhin erproben wir, wie bereits entwickelte Methoden und Werkzeuge der digitalen Geisteswissenschaften z. B. Text- und Bildannotationen sowie algorithmische Bilderkennungsverfahren im Kontext musealer Fragestellungen eingesetzt werden können.

Ausblick

Auf Grundlage der vorliegenden Konzeption des Objektkatalogs haben wir technische Bedarfe an ein nachnutzbares System abgeleitet. Hierbei werden insbesondere graphbasierte Technologien fokussiert: Anhand des Konzepts zeichnet sich ab, dass die zu schaffende Wissensbasis einen ontologisch-vernetzten Charakter aufweisen wird. Deren Modellierung, Repräsentation und Abfrage lässt sich am besten durch eine Graphstruktur realisieren.

Aktuell analysieren wir konkretisierte Anforderungen an das Datenmodell und ermitteln gemeinsam mit den Verbundpartnern Schemata für die Beschreibung der digitalen Objekte und Anwendungen. Im Zuge dessen entwickeln wir ein Metadatenschema, das erstmals auch komplexe digitale Formate, wie z.B. VR-Anwendungen, als Teil eines im musealen Kontext entstehenden Objektkatalogs erfassbar macht. Die Schemata bilden die Grundlage für die konzeptionelle Modellierung der gemeinsamen Wissensbasis des Objektkatalogs, die dann in einem maschinenlesbaren Datenmodell umgesetzt werden, welches in das gewählte System implementiert wird. Im Sinne des iterativen Vorgehens, welches die Arbeit des Verbundprojektes prägt, wird auch der Objektkatalog schrittweise entwickelt, evaluiert und angepasst. Dazu zählen testmäßige Ingests von Objektdaten sowie auch deren Export in andere Systeme, um die Anbindungsfähigkeit des Katalogs an digitale Forschungswerkzeuge zu erproben. Die dabei entstehenden Mapping-Szenarios, entwickelte oder integrierte Schnittstellen, werden dokumentiert und für die breite Nachnutzung aufbereitet. Alle Ergebnisse sowie auch der Objektkatalog selbst, werden mit offenen Lizenzen versehen und frei zugänglich publiziert.

Fußnoten

1. Beteiligt sind die Stiftung Preußischer Kulturbesitz und ihre Staatlichen Museen zu Berlin, die Stiftung Humboldt Forum im Berliner Schloss, das Deutsche Auswandererhaus Bremerhaven, das Deutsche Museum, die Fastnachtsmuseen

Schloss Langenstein und Narrenschopf Bad Dürrheim mit weiteren Museen der schwäbisch-alemannischen Fastnacht und das Senckenberg Museum für Naturkunde Görlitz. museum4punkt0 wird gefördert von der Beauftragten der Bundesregierung für Kultur und Medien aufgrund eines Beschlusses des Deutschen Bundestages.
2. Für eine Übersicht über die Themen und Inhalte der Teilprojekte siehe http://museum4punkt0.de/teilprojekte

Bibliographie

Bauer, Nadja (2018): *Audiences first, digital second*, in: museum4punkt0 Blog August 2018 http://www.museum4punkt0.de/blogpost/audiences_first_digital_second/ [letzter Zugriff: 12.10.2018].

Cameron, Fiona (2003): *Digital Futures I: Museum Collections, Digital Technologies, and the Cultural Construction of Knowledge*, in: Curator: the Museum Journal 46 (3) 325–40.

Cameron, Fiona / Robinson, Helena (2007): *Digital Knowledgescapes: Cultural, Theoretical, Practical, and Usage Issues Facing Museum Collection Databases in a Digital Epoch*, in: **Cameron, Fiona / Kenderdine, Sarah (eds.):** *Theorizing Digital Cultural Heritage: A Critical Discourse*. MIT Press 165–191.

Glinka, Katrin (2018a): *Ein Verbund ist mehr als die Summe seiner Teilprojekte*, in: museum4punkt0 Blog Juli 2018 http://www.museum4punkt0.de/blogpost/ein-verbund-ist-mehr-als-die-summe-seiner-teilprojekte/ [letzter Zugriff: 12.10.2018].

Glinka, Katrin (2018b): *The Process Is Part of the Solution: Insights from the German Collaborative Project museum4punkt0*, in: Museum International, Volume 70, Special Issue: Museums in a Digital World. International Council of Museums (ICOM) and Blackwell Publishing 90–103 10.1111/muse.12195.

ICOM - Internationaler Museumsrat (2010): *Ethische Richtlinien für Museen von ICOM*. ICOM Deutschland (eds.).

Navarro, Cristina / Fuhrmann, Dietmar (2018): *Ideen auf dem multiperspektivischen Prüfstein: Ein fruchtbarer Workshop*, in: museum4punkt0 Blog Juli 2018 http://www.museum4punkt0.de/blogpost/ideen-auf-dem-multiperspektivischen-pruefstein-ein-fruchtbarer-workshop/ [letzter Zugriff: 12.10.2018].

Srinivasan, Ramesh / Boast, Robin / Furner, Jonathan / Becvar, Katherine M. (2009): *Digital Museums and Diverse Cultural Knowledges: Moving Past the Traditional Catalog*, in: The Information Society 25:4 265–278 10.1080/01972240903028714.

Wallace, Amanda (2001): *Collections Management and Inclusion*, in: **Dodd, J. / Sandell, R.:** *Including Museums: Perspectives on Museums, Galleries and Social Inclusion*. Leicester: University of Leicester, Department of Museum Studies 80–83.

Wissenschaftliche Rezeption digitaler 3D-Rekonstruktionen von historischer Architektur

Messemer, Heike
heike.messemer@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Abstract

Digitale 3D-Rekonstruktionen von historischer Architektur werden seit über 30 Jahren im wissenschaftlichen Kontext erstellt und als Präsentationsmedien und Forschungswerkzeuge verwendet. Sie tragen nicht nur dazu bei, Erkenntnisse und Hypothesen zu einem Bauwerk zu visualisieren, sondern sie können auch im Laufe des Erstellungsprozesses neue Forschungsfragen aufwerfen sowie neue Erkenntnisse befördern. Insofern stellen sie vor dem Hintergrund der Digital Humanities ernstzunehmende Forschungsbeiträge dar, die entsprechend wissenschaftlich rezipiert werden sollten. Allerdings zeigt sich bei der kunsthistorischen Analyse von Projekten zur digitalen Rekonstruktion von historischer Architektur, dass nicht immer eine Rezeption der 3D-Modelle in der Wissenschaftscommunity erfolgt, beispielsweise bei einer späteren Rekonstruktion desselben Bauwerks. Anhand von zwei Beispielen wird dies veranschaulicht, wobei mögliche Ursachen erörtert und Lösungsmöglichkeiten vorgeschlagen werden. Ziel ist es, damit zur Erhöhung der internationalen Sichtbarkeit und Rezeption von wissenschaftlichen 3D-Rekonstruktionen in der Wissenschaftscommunity beizutragen und so einen disziplinübergreifenden Diskurs anzuregen.

Definition

Bei digitalen 3D-Modellen historischer Architektur handelt es sich um mit Hilfe des Computers und auf historischen Quellen basierende Rekonstruktionen von erhaltener, teilweise noch existierender, nicht mehr bestehender oder nie erbauter Architektur (Messemer 2018). Im wissenschaftlichen Kontext werden sie vor allem als Präsentationsmedien und/oder Forschungswerkzeuge verwendet. Digitale Rekonstruktionen dienen dazu Erkenntnisse, Hypothesen und Wissenslücken zu den untersuchten Bauwerken im dreidimensionalen Raum zu visualisieren. Insbesondere im Falle von Forschungswerkzeugen steht die Ergründung von bestimmten Forschungsfragen zum visualisierten Bauwerk im Vordergrund. Der Erstellungsprozess kann jedoch oftmals auch dazu beitragen, neue Fragen aufzuwerfen und neue Erkenntnisse zu generieren.

Grundlage und Methodologie

Grundlage der hier vorgestellten Untersuchung bildet die kunsthistorische Analyse von neun herausragenden, wegweisenden 3D-Projekten – Projekte, in denen die Erstellung einer 3D-Rekonstruktion von historischer Architektur im Mittelpunkt steht – in der von der Autorin 2018 abgeschlossenen Dissertation im Fach Kunstgeschichte (Messemer 2018).[1] Im Rahmen dieser Arbeit wurde unter anderem untersucht, inwiefern die wegweisenden, mit wissenschftlichem Anspruch erarbeiteten 3D-Projekte in Berichten zu später angefertigten digitalen Rekonstruktionen der betreffenden Gebäude Erwähnung finden. Exemplarisch seien hierzu zwei Projekte vorgestellt.

3D-Projekte zur Klosterkirche Cluny III

Als eines der ersten umfangreichen 3D-Projekte im deutschsprachigen Raum, kann die 1989 durchgeführte digitale Rekonstruktion der Klosterkirche Cluny III durch *asb baudat* gesehen werden (Cramer / Koob 1993; Messemer 2016: 27-29). Unter der Leitung des Architekten Manfred Koob (gest. 2011) wurde das heute nur mehr als Fragment existierende Bauwerk für den 1991 ausgestrahlten Dokumentarfilm *Auf den Spuren der Salier. Nomaden auf dem Kaiserthron* des SWF (heute SWR) mit wissenschaftlichem Anspruch am Computer 3D-modelliert, der 2008 auch als DVD veröffentlicht wurde. Eine wichtige Grundlage für die digitale Rekonstruktion bildete dabei die umfangreiche architekturhistorische Forschung Kenneth John Conants. Ausführlich dokumentiert in Wort und Bild ist der gesamte Entstehungsprozess des 3D-Modells in einer 1993 herausgegebenen Buchpublikation von Koob und Horst Cramer, die im Kontext von 3D-Projektberichten bis heute ihresgleichen sucht (Cramer / Koob 1993). Insgesamt stellt dieses 3D-Projekt einen wesentlichen Beitrag für die Forschung zu Cluny III dar, indem es erstmals die wissenschaftlichen Erkenntnisse Conants in Form einer digitalen Visualisierung zeigt und vor allem räumlich erfahrbar macht. Darüberhinaus konnten während des Rekonstruktionsprozesses auch Unstimmigkeiten in der bis dahin vorliegenden Forschung erkannt werden (Cramer / Koob 1993: 78).

Zwischen 1990 und 2010 wurde unabhängig von diesem 3D-Projekt Cluny III in unterschiedlichen Kontexten digital rekonstruiert. Bereits ein Jahr nach der Fertigstellung des 3D-Modells durch *asb baudat*, fertigten die Studenten Christian Père und Philippe Marécaux an der Pariser Ecole Nationale Supérieure d'Arts et Métiers (heute: Arts et Métiers ParisTech) eine digitale Rekonstruktion des Bauwerks an (Dorozynski 1993). Auch sie verwendeten hierfür die von Conant zusammengetragenen Erkenntnisse. Unterstützt wurden sie von der Kunsthistorikerin Dominique Vingtain sowie *IBM France*. Der dazu im Juli 1993 erschienene Artikel in *Science* enthält keinen Verweis auf Koobs Projekt (Dorozynski 1993).[2]Dies mag damit zusammenhängen, dass die vor 1993 zur ersten digitalen Rekonstruktion von Cluny III erschienenen Artikel allesamt auf Deutsch verfasst waren. Teils handelte es sich dabei um Fachmagazine, Regionalzeitungen, aber auch überregionale Zeitungen (z.B. ohne Autor 1990; Dechau 1990; ohne Autor 1991; Behringer 1991). Diese von Dritten verfassten Publikationen über Koobs 3D-Projekt sind damit für Fachkreise im nicht-deutschsprachigen Ausland kaum fassbar.

Möglicherweise ist dies auch ein Grund, weshalb in den Projektberichten und Artikeln zu weiteren 3D-Projekten, die Cluny III zum Gegenstand hatten, die Arbeit von *asb baudat* nicht erwähnt wird. So entwickelte ein Wissenschaftlerteam unter Leitung von Luc Genevriez die Rekonstruktion der Studenten in Form einer Virtual Reality-Anwendung weiter und zeigte diese 1993 auf der Konferenz *Imagina* in Monte Carlo (Joscelyne 1994). Der darüber in *WIRED* 1994 publizierte Artikel enthält keine Information zu dem 1989 realisierten 3D-Modell (Joscelyne 1994).

Eine erneute Weiterentwicklung der von den Studenten erstellten Rekonstruktion erfolgte schließlich 2008 anlässlich des 1100. Jahrestags des Klosters von Cluny im Jahr 2010 (Petty / Landrieu / Coulais / Père / de Ganay 2012: 71; Père / Landrieu / Rollier-Hanselmann 2010; Landrieu / Père / Rollier-Hanselmann / Castandet / Schotte 2012).[3] In Form des Projekts *Abbaye de Cluny – Projet GUNZO 2010 (Gunzo)* erarbeitete unter Koordination von Christian Père ein multidisziplinäres Team an der Arts et Métiers ParisTech in Cluny mit dem Centre des monuments nationaux und der in Frankreich ansässigen Firma *on-situ* eine digitale 3D-Rekonstruktion der Klosterkirche. In den dazugehörigen Publikationen wird das Modell von *asb baudat* ebenfalls nicht genannt. Hingegen findet das Projekt von Luc Genevriez in dem von dem Archäologen Paul Reilly erstellten Überblick zu 3D-Projekten in den 1980er- und 1990er-Jahren eine kurze Erwähnung (Reilly 1996: 45). Zusammenfassend lässt sich somit feststellen, dass die Arbeit von *asb baudat* in der internationalen Forschungslandschaft nicht rezipiert wurde.

3D-Projekte zur Basilika Santa Maria Maggiore in Rom

Ein weiteres wegweisendes 3D-Projekt stellt die am *Cultural Virtual Reality Lab (CVRLab)* an der University of California Los Angeles (UCLA) zwischen 1998 und 2000 erarbeitete digitale Rekonstruktion der Basilika Santa Maria Maggiore in Rom dar (Frischer / Favro / Liverani / De Blaauw / Abernathy 2000; Messemer 2016: 37-38). Sie entstand unter dem Projektleiter und Archäologen Bernard Frischer in interdisziplinärer Zusammenarbeit mit dem Kunsthistoriker Sible De Blaauw (damaliger Vizedirektor des Koninlijk Nederlands Instituut, Rom), dem Direktor der Abteilung für byzantinische, mittelalterliche und moderne Kunst an den Vatikanischen Museen, Arnold Nesselrath, dem damaligen Kurator am Department für Klassische Antike an den Vatikanischen Museen, Paolo Liverani, und der Architekturhistorikerin Diane Favro (Frischer 2004: 68). Ziel des Projekts war es, 3D-Visualisierungen für ein Video zur Erläuterung der Baugeschichte der Basilika für die Ausstellung *Aurea Roma* im *Palazzo delle Esposizioni* in Rom zu erstellen.[4] Für die digitale Rekonstruktion griffen sie auf die aktuelle archäologische und kunsthistorische Forschung zur Basilika zurück. Aus wissenschaftlicher Sicht hat das 3D-Modell eine besondere Bedeutung, als es eine von zwei Hypothesen zur Gestaltung der Apsis wiedergibt, die in der Forschung zur Kirche nach wie vor diskutiert werden

(Frischer / Favro / Liverani / De Blaauw / Abernathy 2000: 156). Frischer und seine Kollegen präsentierten ihre Arbeit auf dem *Festival of Virtual Reality in Archaeology*, das 1998 parallel zur Konferenz *CAA 1998 (Computer Applications and Quantitive Methods in Archaeology)* in Barcelona stattfand. Eine ausführliche Dokumentation des gesamten Erstellungsprozesses liefert der zugehörige im Jahr 2000 erschienene Projektbericht (Frischer / Favro / Liverani / De Blaauw / Abernathy 2000).

Nur wenige Jahre später publizierte die in Italien zu Kunstgeschichte des Mittelalters lehrende Maria Andaloro 2006 ein mehrbändiges Werk, in dem sie unter anderem auch eine digitale Rekonstruktion des Innenraums von Santa Maria Maggiore vorstellt (Andaloro 2006: 269-294; Visconti 2006). Erarbeitet wurde diese innerhalb des Forschungsprojekts *PRIN 2004 Banche Dati e Sistemi Digitali di rappresentazione visiva. Pittura (dipinti murali, mosaici, tavole), arredi liturgici a contesto monumentale, in Italia e a Bisanzio (IV-XV secolo)* an der Fakultät für Architektur der Università degli Studi „G. d'Annunzio" Chieti sowie an der Fakultät für Konservierung von Kulturgütern der Università degli Studi della Tuscia in Viterbo. Ziel war es, anhand der Erstellung von digitalen Rekonstruktionen von Sakralbauten in Rom die zugehörigen Wandmalereien und Mosaike im Raum virtuell zu verorten. Kunsthistoriker und Architekten zeichneten für die 3D-Modelle verantwortlich, die auf historischen Quellen sowie Vermessungsdaten basierten. Allerdings wird auf die unter Frischer erfolgte digitale Rekonstruktion nicht verwiesen. Dies ist verwunderlich, da sie in der ständigen Ausstellung in der Basilika präsentiert wird. Zudem ist sie auch online frei zugänglich auf *YouTube* zu finden – ein Umstand, der für 3D-Projekte der 1990er-Jahre nicht selbstverständlich ist. Ein möglicher Grund, warum das von Frischer et al. durchgeführte Projekt bis heute vor allem in der Kunstgeschichte wenig rezipiert wurde, könnte darin zu finden sein, dass der Projektbericht im Kontext der Archäologie veröffentlicht wurde und der beteiligte Kunsthistoriker Sible De Blaauw bislang keinen Aufsatz zur 3D-Rekonstruktion in einer einschlägigen kunsthistorischen Publikation verfasst hat.

Fazit und Lösungsvorschläge

Unter wissenschaftlicher Arbeitsweise erstellte digitale 3D-Rekonstruktionen historischer Architektur können als eigenständige wissenschaftliche Arbeiten erachtet werden, denn sie geben jeweils die von den Erstellern visualisierten Interpretationen und Hypothesen zum visualisierten Bauwerk wieder. Entsprechend wäre es vor dem Hintergrund der Digital Humanities für die archäologische, architektur-, bau- und kunsthistorische Forschung von immensem Mehrwert, wenn 3D-Modelle unter den genannten Voraussetzungen auch als wissenschaftliche Arbeiten wahrgenommen werden. Hierzu ist es notwendig Berichte über die 3D-Projekte zu publizieren, um deren Rezeption in der Wissenschaftscommunity zu ermöglichen.

Da es sich bei den hier vorgestellten 3D-Projekten um zwei bedeutende Arbeiten im Kontext der digitalen Rekonstruktion von historischer Architektur handelt, kann angenommen werden, dass im Falle von weniger bekannten Projekten bzw. weniger intensiv von der kunsthistorischen Forschung fokussierten Bauwerken die Rezeption noch weniger stark ausgeprägt ist.

Die hier vorgestellten Beispiele legen nahe, dass eine Kombination aus unterschiedlichen Veröffentlichungsorten und -medien eine höhere Sichtbarkeit gewährleistet, und damit eine breite wissenschaftliche Rezeption ermöglicht. Dies sollte möglichst auch auf internationaler Ebene erreicht werden. So scheint es notwendig Projektberichte auf Englisch zu publizieren, die bestenfalls online frei zugänglich zu finden sind. In Bezug auf eine disziplinär ausgerichtete Sichtbarkeit wäre es hilfreich Artikel in einschlägigen Publikationsreihen, Tagungsbänden und Zeitschriften jeweils unterschiedlicher Fächer zu veröffentlichen – schließlich sind die Projekte vielfach interdisziplinär. Umgekehrt kann der Blick über den disziplinären Tellerrand Berichte zu relevanten Projekten in fremden Fachbereichen zutage fördern. Eine vielfältigere Zugänglichkeit würde auch die Diskussion um Best Practice Beispiele im Bereich der 3D-Modellierung anregen. Zudem würde sie dazu beitragen eine intensivere Auseinandersetzung sowohl mit Konventionen zur Erstellung digitaler 3D-Rekonstruktionen historischer Architektur als auch mit Fragen der Langzeitarchivierung von 3D-Projekten zu führen.

Fußnoten

1. Die Dissertation wurde betreut von Prof. Dr. Stephan Hoppe und Prof. Dr. Hubertus Kohle am Institut für Kunstgeschichte der Ludwig-Maximilians-Universität München.
2. Das von Cramer und Koob herausgegebene Buch erschien erst im Dezember 1993.
3. Bereits 2004 wurde eine überarbeitete Version des von den Studenten erstellten 3D-Modells in Form des Films *Maior Ecclesia* veröffentlicht. Vgl.: Petty / Landrieu / Coulais / Père / de Ganay 2012: 73.
4. Vgl. Video auf *YouTube*: https://www.youtube.com/watch?v=ciTZq8beKhA [letzter Zugriff 06. Januar 2019].

Bibliographie

Andaloro, Maria (2006): *La pittura medievale a Roma 312-1431. Atlante percorsi visivi* 1 (Suburbio, Vaticano, Rione Monti). Viterbo / Mailand: Università della Tuscia / Jaca Book.

Behringer, Anja (1991): *„Cluny III. Auferstanden aus Ruinen: Cluny, das glühende Herz des Abendlandes"*, in: ´Frankfurter Allgemeine Magazin 90: 62-68.

Cramer, Horst / Koob, Manfred (1993): *Cluny. Architektur als Vision*. Heidelberg: Edition Braus.

Dechau, Wilfried (1990): *„Cluny IV"*, in: Deutsche Bauzeitung. Zeitschrift für Architekten und Bauingenieure 12: 114-115.

Dorozynski, Alexander (1993): *„Computers Bring Back a Long-Lost French Abbey"*, in: Science 261 (5121): 544-545.

Frischer, Bernard (2004): *„Mission and recent projects of the UCLA Cultural Virtual Reality Laboratory"*, in: Tiré-à-part des Actes du colloque. Virtual Retrospect 2003 (= Archéovision 1). Bordeaux: Editions Ausonius 65-74.

Frischer, Bernard / Favro, Diane / Liverani, Paolo / De Blaauw, Sible / Abernathy, Dean (2000): *„Virtual Reality and Ancient Rome: The UCLA Cultural VR Lab's Santa Maria Maggiore Project"*, in: **Barceló, Juan A. / Forte, Maurizio / Sanders, Donald H. (Hrsg.)**: *Virtual Reality in Archaeology*

(= British Archaeological Reports, International Series 843). Oxford: Archaeopress 155-162.

Joscelyne, Andrew (1994): *„Cluny Abbey has been rebuilt“*, in: WIRED 01.01.1994. https://www.wired.com/1994/01/virtual-cluny/ [letzter Zugriff 06. Januar 2019].

Landrieu, Jérémie / Père, Christian / Rollier-Hanselmann, Juliette / Castandet, Stéphanie / Schotte, Guillaume (2012): *„Digital rebirth of the greatest church of Cluny Maior Ecclesia: From optronic surveys to real time use of the digital model“*, in: **Remondino, F. / El-Hakim, S. (Hrsg.)**: *4th ISPRS International Workshop 3D-ARCH 2011: "3D Virtual Reconstruction and Visualization of Complex Architectures"* XXXVIII-5/W16: 31-37.

Messemer, Heike (2016): *„The Beginnings of Digital Visualization of Historical Architecture in the Academic Field“*, in: **Hoppe, Stephan / Breitling, Stefan (Hrsg.)**: *Virtual Palaces, Part II. Lost Palaces and their Afterlife. Virtual Reconstruction between Science and Media* (= PALATIUM e-Publications 3). München: arthistoricum.net 21-54.

Messemer, Heike (2018): *Entwicklung und Potentiale digitaler 3D-Modelle historischer Architektur – Kontextualisierung und Analyse aus kunsthistorischer Perspektive*. Dissertation, Ludwig-Maximilians-Universität München (unveröffentlicht).

ohne Autor (1990): *„Zu Fuß durch ein Gebäude, das es gar nicht gibt. Wie in Bensheimer Computern aus einer Fata Morgana Cluny III wurde / Teil eines Fernsehfilms“*, in: Bergsträßer Anzeiger, 06.10.1990.

ohne Autor (1991): *„Cluny III mit dem Computer rekonstruiert“*, in: Stein. Steinmetz und Bildhauer 5: 65-69.

Père, Christian / Landrieu, Jérémie / Rollier-Hanselmann, Juliette (2010): *„Reconstitution virtuelle de l'église abbatiale Cluny III : des fouilles archéologiques aux algorithmes de l'imagerie“*, in: Vergnieux, R. / Delevoie, S. (Hrsg.): Actes du Colloque Virtual Retrospect 2009 (= Collection Archéovision 4). Bordeaux: Editions Ausonius 151-159.

Petty, Zoé / Landrieu, Jérémie / Coulais, Jean-François / Père, Christian / de Ganay, Osmond (2012): *„Space and time scaling issues in data management: the virtual restitution of Cluniac heritage“*, in: Applied Geomatics 6 (2): 71-79.

Reilly, Paul (1996): *„Access to Insights: stimulating archaeological visualisation in the 1990s“*, in: **Márton, Erzsébet (Hrsg.)**: *The Future of Our Past '93-'95. International Conference of Informatics*. Budapest: Hungarian National Museum 38-51.

Visconti, Manuel (2006): *„Introduzione“*, in: **Andaloro, Maria (2006.)**: *La pittura medievale a Roma 312-1431. Atlante percorsi visivi* 1 (Suburbio, Vaticano, Rione Monti). Viterbo / Mailand: Università della Tuscia / Jaca Book 14-15.

Poster

Annotation gesprochener Daten mit WebAnno-MM

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Remus, Steffen
remus@informatik.uni-hamburg.de
Language Technology Group, Universität Hamburg, Deutschland

Ferger, Anne
anne.ferger@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Bührig, Kristin
kristin.buehrig@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Biemann, Chris
biemann@informatik.uni-hamburg.de
Language Technology Group, Universität Hamburg, Deutschland

Einleitung

Wir präsentieren WebAnno-MM[1], eine Erweiterung des populären web-basierten Annotationstools WebAnno[2] (Yimam et al., 2013; Eckart de Castilho et al., 2014), welches die Annotation transkribierter Daten ermöglicht und dabei parallel synchronisierte Ansichten auf die Daten zur Verfügung stellt. Die Erweiterung wurde im Rahmen eines Projekts zur Erarbeitung und Erprobung innovativer Lehrmethoden entwickelt (vgl. Remus et al, 2018), um Annotation von Videodaten als Unterstützung für die Analyse und Reflexion authentischer Diskursbeispiele einsetzen zu können. Eine entscheidende Rolle spielen dabei das kollaborative Annotieren und die systematische Überprüfung der Übereinstimmung – einerseits zwischen Studierenden, als Diskussionsgrundlage im Seminar, und anderseits durch einen Studierenden zu verschiedenen Zeitpunkten, um so Einblicke in die Aneignung neuer Konzepte zu bekommen. Die Auswertung von Inter- und Intra-Annotator-Agreement wird durch die elaborierte Nutzerverwaltung und das bereits existierende Kurationsmodul von WebAnno ermöglicht. Ein weiterer Vorteil von WebAnno in Hinblick auf die Diskursannotation ist der bessere Überblick über größere Abschnitte des Diskurses verglichen mit gängigen Werkzeugen, die in erster Linie für die Transkription und Annotation auf Äußerungsebene entwickelt worden sind, wie z.B. der EXMARaLDA Partitur-Editor (Schmidt und Wörner, 2014) oder ELAN (Sloetjes, 2014). Auch hinsichtlich der Struktur der Annotationen bietet WebAnno große Flexibilität, da neben einfacher Spannenannotation auch relationale und feature-basierte Annotationen möglich sind. Ein wichtiger Aspekt für den Einsatz in der Lehre ist zudem die einfache Handhabung einer web-basierten Plattform.

WebAnno-MM Plugin

WebAnno bietet in der Basisversion nur geringe Funktionalität für transkribierte gesprochene Daten. WebAnno-MM bietet daher nun als Plugin für WebAnno sowohl die Möglichkeit, alignierte Mediendateien abzuspielen, als auch die Transkriptionsdaten nach bestimmten Konventionen und dem vorgesehenen Layout für die qualitative Analyse zu visualisieren. Durch ein Open-Source-Projekt wie WebAnno wird nicht nur der Einsatz der Lehrmethode und der neuen WebAnno-Version in verschiedenen Kontexten ermöglicht, sondern auch die Ergänzung und Weiterentwicklung zusätzlicher Funktionalitäten für ähnliche Vorhaben. In Hinblick auf Interoperabilität und Nachhaltigkeit der Lösung wurde dementsprechend der auf den TEI-Richtlinien[3] basierende ISO-Standard „Transcription of Spoken Language“[4] als Grundlage gewählt, für die eine Konvertierung aus gängigen Transkriptionswerkzeugformaten bereits möglich ist (vgl. Schmidt et al., 2017).

Datenmodellierung in WebAnno

Die Grundlage in WebAnno‘s Backend ist UIMA (Unstructured Information Management Architecture)[5] (Ferrucci und Lally, 2004). UIMA speichert Textinformationen, d.h. den Text selbst und die Annotationen, in sogenannten CASs (Common Analysis Systems). Eine große Herausforderung ist die Darstellung von Transkriptionen und den (zeitlichen) Annotationen basierend auf mehreren Sprechern. Die Textsequenz soll dabei die Wahrnehmung einer Konversation möglichst wenig behindern und die einzelnen segmentierten Äußerungen der Sprecher und ihrer kontinuierlichen Annotationen beibehalten. Dazu parsen wir ISO/TEI-XML-Inhalte und speichern Äußerungen einzelner Sprecher in verschiedenen sogenannten Views (verschiedene CASs des gleichen Dokuments) und behalten Zeitstempel als Metadaten innerhalb einer CAS. Wir verwenden das jeweils auf einzelnen Sprecherbeiträgen basierende *annotationBlock*-XML-Element als eine unterbrechungsfreie Einheit, da wir in diesem Fall davon ausgehen können, dass einzelne Annotationen innerhalb der Zeitgrenzen des Sprecherbeitrags liegen. Äußerungen und vorhandene Spannenannotationen werden auf diese Weise in die vorhandene Annotationsansicht von WebAnno übernommen. Eigenständige Ereignisse wie beispielsweise non-verbale Äußerungen hingegen, die als sogenannte *Incidents* auch über andere, sprecherabhängige Äußerungen hinweg auftreten können, werden nicht in die sequentielle WebAnno Annotationsansicht konvertiert, sondern nur in der zusätzlichen Partituransicht dargestellt.

Visualisierung der Transkriptionsdaten

Die neu integrierte Partituransicht basiert auf einer Visualisierung von Äußerungen und Annotationen im bekannten Partiturlayout[6]. Abbildung 1 (rechts) zeigt einen Screenshot von der Partituransicht. Die Annotationsansicht und die Partituransicht sind synchronisiert, d.h. durch Klicken auf die entsprechenden Zeitmarkierungen im jeweiligen Browserfenster ändert sich der Fokus in dem anderen. Durch Klicken wird auch die Medienwiedergabe gestartet oder pausiert. Darüber hinaus kann in der Partituransicht unter den verknüpften Medienformaten die aktuell abzuspielende Datei festgelegt und die Breite der Partiturflächen nach Bedarf eingestellt werden. Es stehen außerdem die Metadaten der Aufnahme und den Sprechern zur Verfügung.

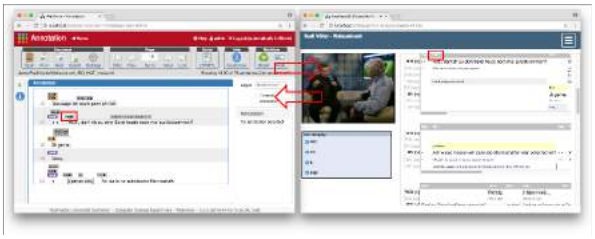


Abbildung 1. Annotationsansicht (links), Partituransicht (rechts)

Präsentation

Wir stellen auf der Grundlage unseres Posters die theoretischen Hintergründe sowie die technische Implementierung der Erweiterung vor und zeigen das Plugin und seine Anwendung in einer Live-Demonstration.

Fußnoten

1. https://github.com/webanno/webanno-mm
2. https://webanno.github.io
3. http://www.tei-c.org/guidelines/p5/
4. http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338
5. https://uima.apache.org/
6. Beispiel der Partituransicht: http://hdl.handle.net/11022/0000-0000-4F70-A

Bibliographie

Eckart de Castilho, Richard / Biemann, Chris / Gurevych, Iryna / Yiman, Seid Muhie (2014): *WebAnno: a flexible, web-based annotation tool for CLARIN*, in: Proceedings of the CLARIN Annual Conference 2014 1-3.

Ferrucci, David / Lally, Adam (2004): *UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment*, in: Natural Language Engineering, 10(3–4): 327-348.

Remus, Steffen / Hedeland, Hanna / Ferger, Anne / Bühring, Kristin / Biemann, Chris (2018): *WebAnno meets EXMARaLDA*, in: Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018. Linköping: University Electronic Press.

Schmidt, Thomas / Wörner, Kai (2014): *EXMARaLDA*, in: **Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (eds.):** *Handbook on Corpus Phonology*. Oxford: University Press 402-419.

Schmidt, Thomas / Hedeland, Hanna / Jettka, Daniel (2017): *Conversion and annotation web services for spoken language data in CLARIN*, in: Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016. Linköping: University Electronic Press 113–130.

Sloetjes, Han (2014): *ELAN: Multimedia annotation application*, in: **Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (eds.):** *Handbook on Corpus Phonology*. Oxford: University Press 305–320.

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 the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations , Sofia, Bulgaria 1-6.

Auf alles gefasst? Metadaten im Virtuellen Kupferstichkabinett

Rössel, Julia
roessel@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Maus, David
maus@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Einleitung

In mehreren Kooperationsprojekten der Herzog August Bibliothek Wolfenbüttel und des Herzog Anton Ulrich-Museums wurden von 2007 bis heute rund 100.000 Datensätze zu Kulturobjekten aus den Graphischen Sammlungen beider Institutionen erzeugt. Das Poster stellt Konzepte und Strategien eines Metadaten Assessment vor, das die Daten des Virtuellen Kupferstichkabinetts (www.virtuelles-kupferstichkabinett.de) analysiert. Die Analyse ist Teil einer nachhaltigen Qualitätssicherung für Metadaten beider Kulturinstitutionen und erweist sich als vermittelnde Praxis zwischen medialen Kontexten und aktuellen wie zukünftigen NutzerInnen.

Problemstellung

Das Virtuelle Kupferstichkabinett wurde von Beginn an als öffentlich zugängliches Online-Portal konzipiert. Der Blick auf die erzeugten Informationsobjekte wird so von einer Auffassung der Objekte als Komposite von Bildern und zugehörigen Metadaten innerhalb eines im World Wide Web verfügbaren Informationssystems dominiert. Diese visuelle Präsentation kann als Remediation gedruckter Kataloge gesehen werden, an denen sich Erwartungen der NutzerInnen zu orientieren scheinen.

Neben der Präsentation und Publikation von Informationen verpflichten sich Kulturinstitutionen dazu, eine langfristige Perspektive in den Blick zu nehmen. Diese erfordert es, die eigenen Informationsobjekte bzw. -produkte jenseits aktueller Trends der Präsentation im Internet und jenseits einer bestimmten maschinellen Verarbeitung zu denken und entsprechend mit ihnen umzugehen. Sie müssen so konzipiert werden, dass sie innerhalb eines Curation Lifecycle dauerhaft genutzt werden können.

Metadaten Assessment geht vom dynamischen Wesen dieser Informationsobjekte aus, die innerhalb eines Nutzungsprozesses verschiedene Formen annehmen, welche jeweils unterschiedliche Praktiken und Epistemologien mit sich bringen. Mangelndes Verständnis für diesen Formwandel ist als ein Hemmschuh für die Metadatenqualität beispielsweise von der Task Force Metdatenqualität der Europeana ausgemacht worden (Dangerfield/Kalshoven 2015: 42). Gerade deshalb und trotz der individuellen Voraussetzungen, die unsere Projektstruktur mit sich bringt, zeigen wir, dass unsere Herangehensweise im Bereich der Kulturdaten im Prinzip übertragbar ist, denn gerade in Übertrag- bzw. Wiederholbarkeit ist Ziel und Grundidee von Metadaten Assessments.

Nach nunmehr zehnjähriger Digitalisierungs- und Erschließungsarbeit wird an Informationsobjekten des Virtuellen Kupferstichkabinetts deutlich, dass sich die Praxis der Verzeichnung und Konzepte der Digitalisierung weiter entwickelt haben. Ähnlich wie sich eine stets wandelnde Nutzung auf die Struktur materieller Sammlungen niederschlägt, verändern sich unter anderem auch die Auffassungen darüber, welche Informationen auf welche Weise als Metadaten aufgenommen werden sollen, was ein eigenständiges Informationsobjekt konstituiert oder wie eine adäquate Datenhaltung und -präsentation aussehen sollte. Unterschiedliche Wissenskontexte in den Sammlungen (z.B. wissenschaftliche Herangehensweisen an Zeichnungen und Druckgraphiken) spielen bei ihrer Formung ebenso eine Rolle, wie Unterschiede in der Erfassungstiefe oder Weiterentwicklungen der zugrundeliegenden technischen Systeme. Hinzu kommt die Verschiedenartigkeit der digitalisierten Objekte, die neben Einzelblättern auch Serien, Sammelbände und Graphik in Büchern mit einschließt, deren Bestandteile allerdings nicht in hierarchischer Form verzeichnet werden konnten. All diese Komponenten beeinflussen die Struktur eines „Objektes", dessen (Feld-)Struktur nur auszugsweise im Poster präsentiert werden wird.

Vorgehen

Im Hinblick auf ihre signifikanten Eigenschaften werden die Metadaten des VKK in ihren verschiedenen Dimensionen analysiert, die Literatur aus bibliothekarischen, aber auch ökonomisch orientierten Bereichen entnommen sind. Einzelnen Sektoren im Curation Lifecycle können im Poster beispielhaft Qualitätsdimensionen zugeordnet werden. Dabei gelten für die Nutzbarkeit u.a. die Qualitätsparameter der Korrektheit, Vollständigkeit, Relevanz und Aktualität (Olson 2011: 24-27). Zur Vollständigkeit etwa gehört die Anreicherung durch Normdaten, wo sie noch nicht vorliegen. Im Poster werden die diesbezüglichen Methoden, Fragestellungen und erste Ergebnisse präsentiert werden. Einblicke in den Grad der Interoperabilität und Vollständigkeit können etwa durch die Analyse der Nutzung von Normdaten gewonnen werden. Objekt-Objekt-Beziehungen, die zwischen einzelnen graphischen Blättern hergestellt werden und bislang unterschiedliche Formen des Bezugs abdecken, sind die größeren Problemfelder des Projektes und müssen v.a. unter dem Blickwinkel der Konsistenz analysiert werden.

Ein Metadaten Assessment fragt darüber hinaus nach ökonomisch-institutionellen Kontexten der Datenerzeugung, -nutzung und -haltung. Entsprechend kann im Poster eine Workflowanalyse vorgestellt werden, die aufzeigt, worin bisher angewandten Maßnahmen zur Qualitätssicherung bei der Datenerzeugung bestehen und wo sie möglicherweise zu kurz greifen. Dabei können zudem NutzerInnengruppen ermittelt werden, die verschiedene Anforderungen an die interne Nutzung (z.B. bibliothekarische Auskunft) haben. Zu diesen gehört unter anderem eine fachspezifische Recherche, deren Parameter aus in den 1990er Jahren durch die International Federation of Library Associations and Institutions (IFLA) aufgestellten Anforderungen an bibliographische Informationsobjekte abgeleitet werden können (Arbeitsstelle für Standardisierung 2006: 8).

Das Poster wird an dieser fachspezifischen Recherche Theorie und Praxis des Metadaten Assessment erläutern, sowie erste Analyseergebnisse vor und mögliche Qualitätssicherungsmaßnahmen zur Diskussion stellen.

Bibliographie

Arbeitsstelle für Standardisierung (Hrsg.) (2006): *Funktionelle Anforderungen an Bibliographische Datensätze – Abschlussbericht der IFLA Study Group on the Functional Requirements for Bibliographic Records*. Frankfurt am Main: Deutsche Nationalbibliothek.

Bolter, Jay David / Grusin, Richard (1999): *Remediation – understanding new media*, Cambridge: MIT Press.

Bruce, Thomas R. / Hillmann, Diane I. (2004): *"The Continuum of METADATA Quality: Defining, Expressing, Exploiting"*, in: Metadata in Practice, ALA Editions 238-256.

Dangerfield, Marie-Claire / Kalshoven, Lisette (2015): *Report and Recommendations from the Task Force on Metadata Quality*, https://pro.europeana.eu/post/metadata-quality-task-force-report [letzter Zugriff: 24. September 2018].

Digital Curation Center (2018): http://www.dcc.ac.uk/resources/curation-lifecycle-model [letzter Zugriff: 24. September 2018].

Dushay, Naomi / Hillmann, Diane I. (2003): *"Analyzing Metadata for effective Use and Reuse"*, in: Proceedings of DCMI International Conference on Dublin Core and Metadata Applications: 161-170.

Foulonneau, Muriel / Riley, Jenn (2008): *Metadata for digital resources: implementation, systems design and interoperability*. Oxford: Chandos.

Janert, Philipp K. (2011): *Data Analysis with Open Source Tools*. Sebastopol: O'Reilly Media.

McCallum, Q. Ethan (2012): *Bad Data Handbook*. Sebastopol: O'Reilly Media.

Neuroth, Heike / Oßwald, Achim / Scheffel, Regine u. a. (2010) Hrsg.: *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Frankfurt am Main: Nestor.

Olson, Jack E. (2011): *Data Quality – The Accuracy Dimension*. San Francisco: Morgan Kaufmann Publishers.

Schöch, Christof (2013): *"Big? Smart? Clean? Messy? Data in the Humanities"*, in: Journal of Digital Humanities 2: 2-13.

Augmentierte Notizbücher und Natürliche Interaktion – Unterstützung der Kulturtechnik Handschrift in einer digitalen Forschungswelt

Schwappach, Florin
florin.schwappach@ur.de
Universität Regensburg, Deutschland

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Kontextualisierung: Wissenschaftliche Notizbücher

In der jüngeren Geschichte sind und waren Notizbücher ständige Begleiter von Schriftstellern, Künstlern, Geschäftsleuten und Wissenschaftlern (Yeo, 2014). Ein besonderes Faszinosum stellen dabei wissenschaftliche Notizbücher dar, da sie posthum Einblicke in die Gedankenwelt ihrer Autor_Innen erlauben und dadurch die Genese wissenschaftlicher Theorien und Methoden nachvollziehbar machen. Zu den bekannteren Beispielen wissenschaftlicher Notizbücher und deren nachträglicher Exegese zählen etwa die Aufzeichnungen von Isaac Newton (vgl. McGuire & Tamny, 1983), Paul Dirac (vgl. Galison, 2000), Alexander von Humboldt (vgl. Lenz, 2015) und Theodor Fontane[1].

Wenngleich sich die Art und Weise des Notierens im Laufe der Zeit stark verändert hat, zeigt sich als Konstante doch stets das inhärente Bedürfnis nach einer Materialisierung von Denkprozessen durch Aufschreiben und Visualisieren. Dabei nimmt das Anfertigen von Notizen im wissenschaftlichen Arbeiten eine besondere Stellung ein und ist Teil eines mehr oder weniger chaotischen Entwurfsprozesses (Krauthausen, 2010), weichen doch die letztendlichen Ergebnisse oftmals stark von den ursprünglich gesetzten Zielen ab. Ernst Mach umschreibt das Schaffen von Wissen gar als geleiteten Findeprozess, der letztlich auf einen günstigen „psychischen Zufall" abzielt (Krauthausen, 2010). Ein solcher Findeprozess umfasst das Skizzieren, Notieren und Sammeln von Ideen und bestehendem Wissen. Das Notieren mit Papier und Stift unterstützt einen geleiteten Findeprozess in idealer Weise indem es eine breite Palette von Interaktionsmöglichkeiten, auch *affordances* (Gibson, 1977) genannt, eröffnet. Dazu gehören u.a. das Knüllen, Reißen, Falten, oder auch das Markieren, Streichen, Färben und Zeichnen. Als Antipode zum Notieren auf Papier stehen im digitalen Zeitalter eine Vielzahl digitaler Annotations- und Notiztools zur Verfügung, welche die genannten *affordances* von Papier und Stift aber allenfalls simulieren. Während der Gebrauch der analogen Werkzeuge nur von der Kreativität der Nutzer begrenzt wird, müssen entsprechende Funktionen in digitalen Programmen explizit erdacht, implementiert und getestet werden. Schnittstellen zu anderen Anwendungen müssen entworfen und standardisiert werden, während analoge Objekte beliebig kombinierbar sind. Der entscheidende Vorteil handschriftlichen Notierens auf Papier liegt somit in der großen individuellen Gestaltungsfreiheit und Flexibilität und der damit einhergehenden Verstärkung der eigenen kognitiven Fähigkeiten (Norman 1993; Piolat et al. 2005). Gleichzeitig bieten allerdings auch digitale Notizwerkzeuge spezifische Vorteile gegenüber der analogen Variante: So sind sie meist cloud-basiert, auf mehreren Endgeräten verfügbar und mit Backup-Funktionen versehen. Verknüpfungen mit Literaturdatenbanken, einfaches Sortieren der Einträge sowie Textverarbeitungsfunktionen wie z.B. eine Volltextsuche und nicht zuletzt Copy & Paste unterstützen den meist digital geprägten Arbeitsprozess.

Das hier vorgestellte Projekt positioniert sich im Spannungsfeld von traditioneller Kulturtechnik und innovativen Informationssystemen und zielt darauf ab, die Qualitäten von Papier- und Stift-Interaktion zu erhalten, indem hilfreiche digitale Funktionen mittels Augmented Reality (AR) in analoge Notizbücher integriert werden. Ziel ist es also, das beste aus der analogen sowie auch der digitalen Welt in einer augmentierten Notizanwendung zu vereinen.

Technisches Umsetzungskonzept

AR-Anwendungen erlauben es, durch eine Kombination aus Tracking- und Display-Technologien, Nutzern virtuelle Informationen anzuzeigen, die in der realen Welt verankert zu sein scheinen. In Verbindung mit Steuergeräten oder Gestenerkennung können so potenziell beliebige Gegenstände ohne integrierte technische Vorrichtungen zu Interfaces von Informationssystemen werden (siehe z.B. Martedi et al., 2012; Biefang et al., 2017). Hier setzt auch unser AR-Notizbuch-Projekt an: Anwender sollen mit ihren persönlichen

Notizbüchern und Stiften arbeiten können und möglichst natürlich und intuitiv per AR zusätzliche digitale Funktionen angeboten bekommen. Im Unterschied zu kommerziell bereits verfügbaren Entwicklungen, die mit speziellen Stiften oder digitalem Papier auf eine technische Modifikation der Schreibwerkzeuge setzen, steht hier die Untersuchung eines nicht-invasiven Ansatzes im Vordergrund.

Geplante Funktionen (vgl. Abbildung 1) sind dabei etwa:

- Volltextsuche in den handschriftlichen Notizen
- Verschlagwortung einzelner Einträge
- Verknüpfung von Literaturquellen und deren Anzeige
- Sicherung der handschriftlichen Notizen in vektorisierter Form

Das Hauptaugenmerk der technischen Umsetzung liegt auf dem Stift als primärem Interaktionswerkzeug. Somit dienen Stiftgesten im Sinne der *Natural Interaction* (Nielsen, 1993) als Interaktionsmethode mit dem AR-System. Gleichzeitig soll über Handschriftenerkennung das Geschriebene digitalisiert und so bspw. der oben genannten Volltextsuche zugänglich gemacht werden.

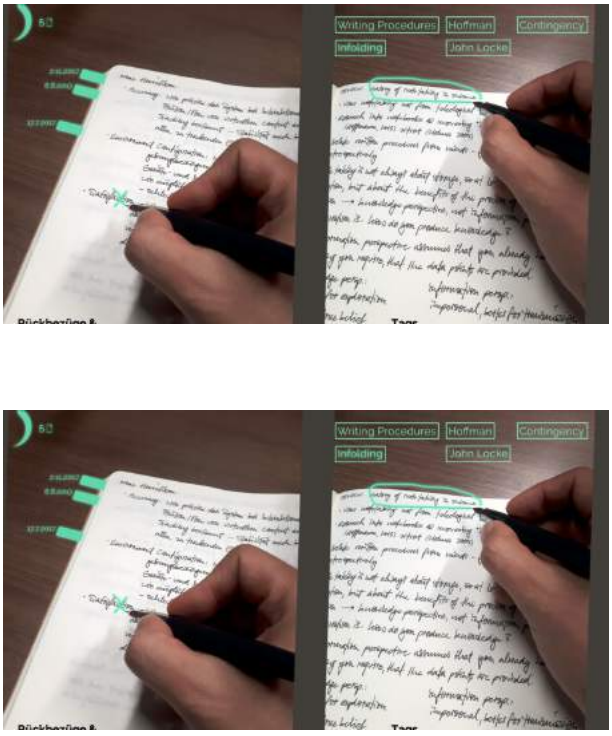


Abbildung 1. – Mockups von geplanten Funktionen des AR-Notizbuchs.
Links: Markierung eines Suchbegriffs durch eine Stiftgeste und Anzeige der Einträge, die diesen Begriff enthalten.
Rechts: Einkreisen eines Begriffs per Stiftgeste, zu dem Tags angezeigt werden sollen. Über die Tags kann auf andere Einträge mit entsprechenden Schlagwörtern zugegriffen werden.

Um Stiftbewegungen zu erkennen, wird ein Deep Learning-Ansatz verfolgt. Unter Einsatz eines *Mask-RCNN*-Netzes (He et al., 2017) wird dabei ein Kamerabild segmentiert und dazu genutzt, *Convolutional Pose Machines* (Wei et al., 2016) zu trainieren, welche *Keypoints* des Stifts in der Hand der Nutzer liefern. Das Ziel dieser Erkennung ist ein vektorisiertes Schriftbild mit Bewegungsinformationen, das an bestehende Systeme zur Online-Handschrifterkennung (Keysers et al., 2017) übergeben werden kann. Als Trainingsdaten für die neuronalen Netze dienen das *Google Open Image Dataset v4* (Krasin et al., 2017) sowie eigene Aufnahmen. Zudem soll die Eignung von Tiefenkameras untersucht werden, um mehr Informationen über die vorliegende Notizszene in den Machine Learning-Ansatz einfließen zu lassen. Außerdem ist eine Erweiterung des Trainingskorpus geplant, um die Genauigkeit und Robustheit der Detektion der Stiftpose zu erhöhen. Gleichzeitig wird eine Client-Server-Architektur entworfen, um die für eine Echtzeit-Erkennung notwendige Rechenleistung außerhalb des AR-Headsets zur Verfügung zu stellen.

Status Quo und Ausblick

Das hier vorgestellte Konzept zur computergestützten Augmentierung handschriftlicher Notizbücher mithilfe von Deep Learning, Natural Interaction und Augmented Reality ist Teil des laufenden Dissertationsprojekts von Florin Schwappach. Im Zuge dessen wurde bislang eine interdisziplinäre (Wissenschaftsgeschichte, Medienwissenschaft, Psychologie, UX Design), theoretische Fundierung des Ansatzes erarbeitet. Parallel dazu wurden erste Prototypen zum Tracking von Stiftbewegungen und zur automatischen Handschriftenerkennung entwickelt, die bereits grundlegend funktionabel sind und im Laufe des Projekts stetig verbessert werden sollen. Die nächsten Schritte beinhalten die Umsetzung der AR-Komponente für die Notizbücher, also bspw. das Durchsuchen der digitalisierten Handschriften und die Visualisierung der Ergebnisse. Das AR-Notizbuch wird im Sinne des *user centered design*-Prozesses (siehe ISO 9241-210, 2010) während des Projekts iterativ anhand von Nutzerstudien auf Usability und Funktionalität hin evaluiert und optimiert.

Über die Entwicklung eines funktionsfähigen Prototypen und entsprechende Nutzertests werden im Rahmen des vorgestellten Projekts einerseits Erkenntnisse über natürliche Interaktionsmodalitäten im AR-Kontext gewonnen. Andererseits werden konkrete technische Ansätze dazu entwickelt, wie sich innovative AR-Systeme mit State-of-the-Art Machine Learning-Methoden verbinden lassen, um ihren Funktionsumfang zu erweitern. Das Projekt leistet weiterhin einen wichtigen Forschungsbeitrag für die Digital Humanities indem es Möglichkeiten der multimodalen Integration von analogen und digitalen Methoden im Kontext des wissenschaftlichen Arbeitens untersucht.

Fußnoten

1. Vgl. die digitale genetisch-kritische und kommentierte Edition von Fontanes Notizbüchern (herausgegeben von G. Radecke), online verfügbar unter https://fontane-nb.dariah.eu/index.html, abgerufen am 02.01.2019

Bibliographie

Biefang, K. / Kunkel, J. / Loepp, B. / Ziegler, J., (2017): *Eine Sandbox zur physisch-virtuellen Exploration von Ausgrabungsstätten*, in: **Burghardt, M. / Wimmer, R. / Wolff, C. / Womser-Hacker, C. (Hrsg.):** *Mensch und Computer 2017 - Workshopband*. Regensburg: Gesellschaft für Informatik e.V..

Galison, P. (2000): *The suppressed drawing: Paul dirac's hidden geometry*. Representations, (72), 145–166. Abgerufen von http://www.jstor.org/stable/2902912

Gibson, J. (1977): *The theory of affordances*, in: **R. Shaw / J. Bransford (Hrsg.):** *Perceiving, Acting, and Knowing: Toward and Ecological Psychology* (S. 62–82). Hillsdale, NJ: Erlbaum.

He, K. / Gkioxari, G. / Dollár, P. / Girshick, R. (2017, Oktober): *Mask R-CNN*, in: Computer Vision (ICCV), 2017, IEEE International Conference on (S. 2980-2988). IEEE.

ISO 9241-210. (2010): *Ergonomics of human-system interaction – Part 210: Human-centred design process for interactive systems.*

Keysers, D. / Deselaers, T. / Rowley, H. A. / Wang, L. L. / Carbune, V. (2017): *Multi-Language Online Handwriting Recognition*. IEEE Trans. Pattern Anal. Mach. Intell, 39(6), 1180-1194.

Krasin I. / Duerig T. / Alldrin N. / Ferrari V. / Abu-El-Haija S. / Kuznetsova A. / Rom H. / Uijlings J. / Popov S. / Kamali S. / Malloci M. / Pont-Tuset J. / Veit A. / Belongie S. / Gomes V. / Gupta A. / Sun C. / Chechik G. / Cai D. / Feng Z. / Narayanan D. / Murphy K. (2017): *OpenImages: A public dataset for large-scale multi-label and multi-class image classification*. Verfügbar unter https://storage.googleapis.com/openimages/web/index.html .

Krauthausen, K. (2010): *Vom Nutzen des Notierens: Verfahren des Entwurfs*, in: **K. Krauthausen / O. Nasim (Hrsg.):** *Notieren, Skizzieren: Schreiben und Zeichnen als Verfahren des Entwurfs* (S. 7–26). Zürich.

Lenz, M. (2015): *Bewegte Systematik. Alexander von Humboldts „Amerikanische Reisetagebücher* als Problemfelder der Literaturgeschichte und historischen Epistemologie. HiN - Alexander von Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien, 16(31), 78-104.

Martedi, S. / Uchiyama, H. / Enriquez, G. / Saito, H. / Miyashita, T. / Hara, T. (2012): *Foldable augmented maps.* IEICE TRANSACTIONS on Information and Systems, 95(1), 256-266.

McGuire, J. / Tamny, M. (1983): *Certain philosophical questions: Newton's trinity notebook.* Newton's Trinity Notebook. Cambridge University Press.

Nielsen, J. (1993): *Noncommand user interfaces.* Commun. ACM, 36(4), 83–99.

Norman, D. A. (1993): *Things that make us smart: Defending human attributes in the age of the machine*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Piolat, A. / Olive, T. / Kellogg, R. T. (2005): *Cognitive effort during note taking.* Applied Cognitive Psychology, 19(3), 291–312.

Wei, S. E. / Ramakrishna, V. / Kanade, T. / Sheikh, Y. (2016): *Convolutional pose machines*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (S. 4724-4732).

Yeo, R. (2014): *Introduction*, in: Notebooks, english virtuosi, and early modern science (S. 1–35). Chicago/London: University of Chicago Press.

Automatische Übersetzung als Bestandteil eines philologischen B.A.-Curriculums mit DH-Schwerpunk

Baillot, Anne
anne.baillot@univ-lemans.fr
Le Mans Université, Frankreich

Wottawa, Jane
jane.wottawa@univ-lemans.fr
Le Mans Université, Frankreich

Barrault, Loïc
loic.barrault@univ-lemans.fr
Le Mans Université, Frankreich

Bougares, Fethi
fethi.bougares@univ-lemans.fr
Le Mans Université, Frankreich

Neue Ko-AutorInnen: Mélissa Emorine (melissa.emorine.etu@univ-lemans.fr) und Ludovic Gervais (ludovic.gervais.etu@univ-lemans.fr)

In diesem Poster wird das im WS 2018/19 an der Universität Le Mans umgesetzte Konzept der Einbettung eines Seminars zur halbautomatischen Übersetzung in ein reguläres philologisches Curriculum dargestellt und dessen Ergebnisse kritisch beleuchtet.

Seit diesem Semester wird ein Digital Humanities-Modul in die Pflichtveranstaltungen des Germanistik-B.A.s eingebunden, das u.a. ein Seminar zur Einführung in automatischer Übersetzung beinhaltet. Dieser Schwerpunkt wurde gewählt, da Übersetzungen für Studenten der Philologie zum Studienalltag gehören und an der Universität entwickelte Kompetenzen im späteren Arbeitsleben der Studienabsolventen gehören können. Außerdem gibt es an der Universität Le Mans thematische Interessensüberschneidungen in der Forschung zwischen der Informatik und den Philologien. Sie stützt sich auf das starke Profil der lokalen Informatik in diesem Bereich, insbesondere im Bereich der multimodalen Übersetzung (Elliott *et al.*, 2017; Caglayan, Barrault, & Bougares, 2016; Afli, Barrault & Schwenk, 2016; Caglayan *et al.*, 2016).

Die Einbindung des DH-Moduls in das Germanistik-Curriculum ermöglicht es, das Konzept zuerst in einem einzelnen Studienbereich mit vergleichsmäßig geringem Effektiv zu erproben. Ziel ist eine Erweiterung auf die Anglistik ab 2020 im Rahmen eines Vertiefungsmoduls zur Übersetzungstheorie und -praxis. Das Poster stellt die erste Lehrveranstaltung dieser Versuchsreihe im Seminar „Einführung in die automatische übersetzung" vor, die mit 5 französischen, 4 deutschen und einer

englischen MuttersprachlerInnen durchgeführt wurde. Bei den TeilnehmerInnen handelt es sich um alle in dem Seminar eingeschriebenen Studierende. Das Poster wird drei Schwerpunkte beinhalten: der Aufbau und die Durchführung der Lehrveranstaltung, der verwendete informatische Hintergrund und schließlich eine Präsentation unseres Projekts, die philologischen Studiengänge an der Universität Le Mans Stück für Stück in DH-Curricula umzuformen, die sich vom B.A. bis zur Promotion erstrecken sollen.

Das Seminar bietet eine Einführung zur Handhabung des Web-Interface „matecat", das es den NutzerInnen erlaubt die einzelnen Sätze der Nachrichten, oder Untertitel automatisch vorzuübersetzen, ein Translation Memory anzulegen, oder auf eines zurückzugreifen, und die Übersetzungen nachzubearbeiten. Dieses Interface ermöglicht eine größere Kohärenz der Übersetzung (Serpil, Durmuşoğlu-Köse, Erbek, Öztürk, 2016) sowie ein zeiteffizienteres Arbeiten.

Der Aufbau des Seminars sieht zunächst eine Übungsphase vor, in der die üblichen Tools (online-Lexika) und ihre Grenzen evaluiert werden. In einem zweiten Schritt wird das Interface „MateCat" eingeführt, wobei jeder Kursteilnehmende ein eigenes Projekt angelegt bekommt, das er im Laufe des Semesters nicht nur mit Übersetzungsinhalten, sondern auch mit einem projektübergreifenden Glossar anreichern muss. Außerdem wird zum einen die Arbeitsweise mit dem Interface kritisch reflektiert und zum anderen in die informatischen Grundlagen der angewendeten Technologie eingeführt (statistische Methoden des Systems und neurale Netzwerke). Potentiale und Grenzen der automatischen Übersetzung sollen den Studierenden damit vermittelt werden (Hussein, 2015 ; Viehhauser, 2018).

Die Frage nach den DH-Curricula wird in diesem Kontext durch eine deutliche Verankerung in der philologischen Praxis beantwortet. Dargestellt wird das gesamte Einführungsmodul, in das diese Lehrveranstaltung eingebettet ist, sowie die Kompetenzen, die dadurch eingeworben werden sollen. Auf den nationalen französischen Kompetenzreferenzrahmen wird dabei hingewiesen wie auf dessen Bedeutung für die Konzeption von DH-Curricula in Frankreich.

Bibliographie

Afli, H., Barrault, L., & Schwenk, H. (2016): *Building and using multimodal comparable corpora for machine translation*. Natural Language Engineering, 22(4), 603-625.

Caglayan, O., Barrault, L., & Bougares, F. (2016): *Multimodal attention for neural machine translation*. arXiv preprint arXiv:1609.03976.

Caglayan, O., et al. (2016): *Does multimodality help human and machine for translation and image captioning?*. arXiv preprint arXiv:1605.09186.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017): *Findings of the second shared task on multimodal machine translation and multilingual image description*. arXiv preprint arXiv:1710.07177.

Hussein, A. (2015): *The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined?*. Journal of comparative social work, 4(1).

Serpil, H., Durmuşoğlu-Köse, G., Erbek, M., & Öztürk, Y. (2016): *Employing computer-assisted translation tools to achieve terminology standardization in institutional translation: Making a case for higher education*. Procedia-Social and Behavioral Sciences, 231, 76-83.

Viehhauser, G. (2018): *Digital Humanities als Geisteswissenschaften. Zur Auflösung einer Tautologie. Digital Humanities: Perspektiven der Praxis*, 1, 17.

Bauanleitung für einen Forschungsraum mit institutionellem Fundament: Erfahrungen aus fünf Jahren Infrastrukturentwicklung im Forschungsverbund MWW

Dogunke, Swantje
swantje.dogunke@klassik-stiftung.de
Forschungsverbund MWW | Klassik Stiftung Weimar, Deutschland

Steyer, Timo
steyer@hab.de
Forschungsverbund MWW | Herzog August Bibliothek Wolfenbüttel

Ein Schwerpunkt des Forschungsverbunds Marbach Weimar Wolfenbüttel besteht im Aufbau einer digitalen Infrastruktur, um bestandsbezogene Forschung von der Korpusbildung über Analyse- und Auswertungsverfahren bis zur Veröffentlichung von Forschungsergebnissen digital zu unterstützen. Das infrastrukturelle Profil konzentriert sich auf den Bereich der Sammlungsforschung und führt über diese thematische Ausrichtung Erschließung und Forschung zusammen. Durch die gemeinsame Forschungsinfrastruktur sollen die Sammlungen und die damit verbundenen Forschungsaktivitäten virtuell zusammenwachsen.

Die Forschungsinfrastruktur MWW ist modular aufgebaut; die zentralen Komponenten bilden ein verlässlicher Speicher und ein virtueller Forschungsraum. Die Gesamtarchitektur des virtuellen Forschungsraums folgt der Taxonomie TaDiRAH (Borek 2014) und konnte in einen mehrschichtigen Architekturplan überführt werden. Dieser bildet über eine bestandsübergreifenden Suche, durch Instrumente für kollaboratives Arbeiten und durch Services zur Analyse sowie Modifizierung von Forschungsdaten signifikante Bestandteile der geisteswissenschaftliche Forschungsprozesses ab. Innerhalb des Forschungsraumes steht den Projekten jeweils eine spezifische Arbeitsumgebung zur Verfügung, in denen über ein Basis-Set an Services hinaus die projektrelevanten Tools und Services aus dem Portfolio des Forschungsraums angeboten werden.

Der Beitrag widmet sich primär jedoch nicht den Funktionen des Forschungsraums, sondern möchte die Kriterien des Aufbaus ausgehend von dem über Umfragen ermittelten Bedarf thematisieren und dabei auch auf nicht realisierbare Vorstellungen und Ansprüche eingehen. Darauf aufbauend

soll das Konzept der Realisierung vorgestellt werden, welches sich weniger auf Neuprogrammierung von Services denn auf die Adaption und Modifizierung bestehender Tools sowie auf die Vernetzung mit anderen Infrastrukturen fokussiert.

Den zweiten Schwerpunkt bildet die entworfene Anforderungsliste für die Integration von digitalen Sammlungen in den Forschungsraum. In diesem Feld ergaben sich aufgrund der unterschiedlichen Bestände und technischen Ausrichtungen der beteiligten Häuser ein Bedarf an gemeinsamer Standardisierung und der Entwicklung von Anwendungsprofilen für Metadaten, um die verteilten Sammlungen zu einem interoperablen Datenpool zu transformieren (Gradl 2015). Die Wichtigkeit von Normdaten- und Forschungsdatenmanagement wird dabei an konkreten Beispielen thematisiert werden (Kraft 2017).

Zudem wird der Beitrag anhand von bestandsbezogenen Forschungsprojekten die Nutzung und den Mehrwert der aufgebauten Forschungsinfrastruktur erläutern (Beyer 2017). Anhand dieses Beispiels wird deutlich werden, welchen Mehrwert digitale Arbeitsmethoden für das jeweilige Projekt dargestellt haben und welche Forschungsaufgaben sonst nur eingeschränkt umsetzbar gewesen wären. Dabei spielen im Sinne des Tagungsthema auch die Integration unterschiedlicher medialer Inhalte in eine Publikationsform eine wesentlich Rolle.

Kritisch wird aber auch gefragt, nach dem Aufwand an Vermittlung und Unterstützung im Bereich der Digital Humanities, welche die Realisierung solcher Projekte mit sich bringt und welche Grenzen trotzdem weiterhin bestehen. In diesem Kontext wird auch auf die Frage eingegangen, wie eine Forschungsinfrastruktur sich innerhalb der beteiligten Einrichtungen legitimiert und welche Erfolgskriterien entscheidend sind (Dogunke 2018).

Abschließend wird ein Modell vorgestellt, um die Weiterentwicklung von aufgebauter Infrastruktur und die parallele Planung multidisziplinärer Forschungsprojekte zu koordinieren. Der Mehrwert dieses Vorgehens besteht in der exakten und frühen Ermittlung des Nutzerbedarfs, so sind z.B. Gesamtkostenpläne, Arbeitspläne sowie Synergieeffekte durch eine bessere Personalverteilung und die gemeinsame Entwicklung fehlender Komponenten zu nennen. Das Moderationsverfahren wurde bereits in mehreren Projekten angewendet und soll gemeinsam mit einer Dokumentation der Community zur Verfügung gestellt werden.

Das Poster möchte die positiven und negativen Erfahrungen, welche das Projektteam beim Aufbau des Forschungsraums gemacht hat, visualisieren und zur Diskussion anregen. Da Infrastrukturaufbau im Kontext der digitalen Geisteswissenschaften nach wie vor ein aktuelles Thema dargestellt, wie es z. B. die Diskussion um Software-as-a-Service zeigen (Cremer 2018), erhoffen sich die Einreichenden die Community damit ansprechen zu können.

Bibliographie

Beyer, Hartmut / Münkner, Jörn / Schmidt, Katrin / Steyer, Timo (2017): *Bibliotheken im Buch: Die Erschließung von privaten Büchersammlungen der Frühneuzeit über Auktionskataloge*, in: **Hannah Busch, Franz Fischer und Patrick Sahle [Hrsg.]**: *Kodikologie und Paläographie im digitalen Zeitalter 4 (Codicology and Palaeography in the Digital Age)*. Norderstedt 2017. S. 43-70.

Borek, Louise (2014): *TaDiRAH – Taxonomy of Digital Research Activities in the Humanities*, in: DHd-Blog. https://dhd-blog.org/?p=3073 [letzter Zugriff: 27.09.2018].

Cremer, Fabian /Wübbena, Thorsten: *The next big thing will be a lot of small things – Serviceorientierung als Modell für die Infrastrukturlandschaft*, in: DHd-Blog https://dhd-blog.org/?p=10480 [letzter Zugriff: 27.09.2018].

Dogunke, Swantje, Steyer, Timo und Mayer, Corinna: *"Barcamp Data and Demons: von Bestands- und Forschungsdaten zu Services. Treffen sich ein Bibliothekar, eine Archäologin, ein Informatiker, ...".* In:LIBREAS. Library Ideas, Nr. 33 (2018). https://libreas.eu/ausgabe33/dogunke/ [letzter Zugriff: 6.09.2018].

Gradl, Tobias, Henrich, Andreas, Plutte, Christoph: *Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen*. In: **Constanze Baum und Thomas Stäcker (Hrsg.)**: *Grenzen und Möglichkeiten der Digital Humanities*. 2015. DOI: 10.17175/sb001_020 [letzter Zugriff: 6.9.2018].

Kraft, Angelina: *The FAIR Data Principles for Research Data*. In: TIB-Blog – Weblog der Technischen Informationsbibliothek (TIB). Beitrag vom 12. September 2017. Online:

https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data / [letzter Zugriff: 6.09.2018].

Bildbezogenes Machine Learning anhand der Glasmalereien des Corpus Vitrearum Medii Aevi

Kolodzie, Lisa
lisa.kolodzie@adwmainz.de
Akademie der Wissenschaften und Literatur | Mainz, Deutschland

Farbe ist ein essentieller Teil alltäglicher Kommunikation ohne Worte. Sie kann Gefühle und Gedanken ausdrücken und beeinflussen, ist Teil des Kulturkreises in dem wir aufwachsen und ist daher ein Schlüsselelement in der Vermittlung von Informationen.

Die Glasmalerei als ein zentrales Kommunikationsmedium des Mittelalters, stellt einen wichtigen Gegenstand historischer Forschung dar, lebt von diesem Konzept. Besonders auffällige Merkmale der Glasmalerei sind die Strahlkraft der verwendeten Farben sowie die sakrale Bedeutung, die mit dem Einsatz bestimmter Pigmentierungen einher geht.

Diese Präsentation beschäftigt sich damit, die ausführlich aufgearbeiteten Standfigurenbilder und Ornamentschieben der Elisabethkirche in Marburg hinsichtlich der verwendeten Farben mit Digitalen Methoden zu untersuchen. Dazu werden Techniken aus den Bereichen *Machine Learning* und *Computer Vision* im Rahmen einer Masterarbeit herangezogen. Ziel der Arbeit ist das Clustering einer Vereinfachung der Farbwerte von Glasfenstern mit einem Rückbezug auf die Metadaten der Bilder.

Das Projekt Corpus Vitrearum Medii Aevi erfasst und ediert auf internationaler Ebene den Gesamtbestand mittelalterlicher Glasmalereien in Europa und Nordamerika sowie die jeweiligen Restaurierungszustände der Objekte. Die Ergebnisse dieser Analysen werden in gedruckten Bildbänden publiziert. Die Druckeditionen widmen sich einer ausführlichen ikonographischen, stilistischen und handwerklichen Analyse, die Rückschlüsse auf Baufortschritt von Kirchen oder den Netzwerken von Stiftern und Werkstätten zulässt. Die digitalen Bilder des CVMA Deutschland werden über Metadaten mit Normdatensätzen verknüpft und im Onlinebildarchiv zugänglich gemacht.

Einen besonders interessanten Datensatz stellen die Standfigurenbilder und Ornamentscheiben der Elisabethkirche in Marburg dar. Bei diesen beiden Arten der Glasmalerei überschneiden sich technologische Einheit (Fotografie einer einzelnen Scheibe) und thematisch-motivische Einheit. Im Gegensatz zum ausgewählten Beispieldatensatz bestehen Glasmalereien häufig aus großen Fenstern, deren Motive aus mehreren Einzelscheiben bestehen. Diese beziehen sich hinsichtlich der Farbigkeit der abgebildeten Motive aufeinander und qualifizieren sich daher nicht für eine Einzelscheibenanalyse, die im Rahmen dieser Arbeit vollzogen wird. Zudem bietet das Buch *Die Elisabethkirche zu Marburg in ihrer ursprünglichen Farbigkeit* für die im Vorfeld ausgewählten Fenster einen manuell ausgewerteten Vergleichsdatensatz. Michler analysiert die Glasmalereien hier ausführlich und fügt jeder Untersuchung eine Grafik der verwendeten Farben, sowie ihren Anteil am jeweiligen Motiv bei.

Automatische Bildanalyse mithilfe von Machine Learning wird in den Digitalen Geisteswissenschaften als Methode häufig für das Erkennen handschriftlicher Texte angewendet oder um beispielsweise Textdaten aus retrodigitalisierten Editionen zu gewinnen. Der erste Schritt in einem solchen Prozess ist in vielen Fällen das Reduzieren von Farbwerten. Diese Reduktion auf Schwarz-Weiß-Bilder vereinfacht es der verwendeten Software Linien zu erkennen und Buchstaben zuzuordnen. Daher gilt es, andere Tools und Techniken für den Bereich der im Vorfeld als zentral identifizierten Eigenschaft der Glasfenster ausfindig zu machen die eine Analyse der Farbwerte in den Vordergrund stellen.

In filmwissenschaftlichen Bildanalysen finden auch Graustufen und Farbwerte Beachtung, beispielsweise in quantitativer Farb- und Perspektivenanalyse von Szenen oder sogar ganzen Filmen. Ein Beispiel dazu bildet der Aufsatz *Dead and Beautiful: The Analysis of Colors by Means of Contrasts in Neo-Zombie Movies* von Pause und Walkowski. Daher ist der Bearbeitungsansatz bisher eher in dieser Fachdisziplin zu suchen als in der Kunstgeschichte.

Für die Analyse ist es zunächst wichtig, wie Farbdaten in Bildern abgespeichert werden. Dafür ist eine genauere Betrachtung sowohl von Bildspeicherungsformaten, Farbräumen (CMYK, RGB, HSV, etc.) und -profilen (sRGB, Adobe RGB, etc.), als auch von Visualisierungen der Pixelangaben vonnöten. Ziel ist hierbei herauszufinden, welche Grundlagen die Farbdaten der Bilder vergleichbar und für Machine Learning-Algorithmen verarbeitbar machen und diese Erkenntnisse auf den Datenbestand anzuwenden.

Nach der Auswertung der Bilddaten werden die Grundlagen von Machine Learning erklärt. Zunächst geht es um die Unterschiede zwischen Supervised und Unsupervised Machine Learning sowie der Auswahl des richtigen Algorithmus nach dem im Vorfeld gesteckten Ziel. Dazu wird insbesondere die Bibliothek TensorFlow eingesetzt, welche frei über Github verfügbar und ausführlich dokumentiert ist.

Basierend auf den genannten Grundlagen wird ein zweistufiges Verfahren angewandt: In der Datenaufbereitung werden die Bilder mithilfe von Unsupervised Algorithmen vergleichbar gemacht. Dieser Schritt ermöglicht das manuelle Festlegen einer Clustermenge, die die Farbwerte der unterschiedlich großen Bilddateien vergleichbar macht. An dieser Stelle werden eine Ausgabe von Farbwerten von drei, fünf und sieben Farbclustern pro Bild, die in einer Tabelle ausgegeben werden, gegenübergestellt. Im zweiten Schritt werden die so vergleichbar gemachten Daten mit einem ebenfalls Unsupervised arbeitenden Clustering-Algorithmus gruppiert. Auch hier ist es wichtig zu betrachten, welche Aussage getroffen werden soll und inwiefern das Verwenden der vorher beschriebenen Cluster die Ergebnisse der Algorithmen beeinflusst.

Die Ergebnisse werden in den kunsthistorischen Kontext der erfassten Glasmalerei des Corpus Vitrearum Medii Aevi analytisch eingearbeitet. Der Text beschreibt das geplante Vorgehen im Rahmen einer Masterarbeit, deren Ergebnisse sich nur aus einem geringen Teil des Bildmaterials des Corpus Vitrearum Medii Aevi zusammensetzt. Von Interesse ist nach Abschluss der Arbeit das Ausweiten auf zusätzliche Bestände.

Bibliographie

Binding, Günther / Speer, Andreas (Hrsg.) (1994): *Mittelalterliches Kunsterleben nach Quellen des 11. Bis 13. Jahrhunderts*. Stuttgart- Bad Cannstatt : Frommann-Holzboog.

Bühler, Peter (u.a.) (2018): *Digitale Farbe. Farbgestaltung – Colormanagement – Farbverarbeitung*. Berlin, Heidelberg : Springer Berlin Heidelberg.

Frodl-Kraft, Eva (1970): *Die Glasmalerei. Entwicklung, Technik, Eigenart*. Wien [u.a.] : Schroll.

Géron, Aurélien (2017): *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. Beijing; Boston; Farnham; Sebastopol; Tokyo : O'Reilly

Kurz, Susanne (2016): *Digital Humanities. Grundlagen und Technologien für die Praxis*. Wiesbaden : Springer Vieweg.

Michler, Jürgen (1984): *Die Elisabethkirche zu Marburg in ihrer ursprünglichen Farbigkeit*. Marburg : Elwert.

Open Source for Science (2018): *TensorFlow-Course*. https://github.com/osforscience/TensorFlow-Course [letzter Zugriff: 12.01.2019].

Pause, Johannes / Walkowski, Niels-Oliver (2017): *Dead and Beautiful: The Analysis of Colors by Means of Contrasts in Neo-Zombie Movies*, in: Digital Humanities 2017. Conference Abstracts. https://dh2017.adho.org/abstracts/095/095.pdf [letzter Zugriff: 12.01.2019].

CMIF Creator - digitale Briefverzeichnisse leicht erstellt

Müller-Laackman, Jonas
jonas.mueller-laackman@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Dumont, Stefan
dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Grabsch, Sascha
grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Der Webservice correspSearch[1] wird an der Berlin-Brandenburgischen Akademie der Wissenschaften entwickelt, um digitale und gedruckte Briefeditionen editionsübergreifend zu vernetzen und durchsuchbar zu machen. Darüber hinaus sollen die Briefmetadaten mit correspSearch für neue digitale Forschungsmethoden (wie z.B. aus der Historischen Netzwerkforschung) bereitgestellt werden. Mittlerweile sind über 44.000 edierte Briefe in correspSearch nachgewiesen.

Ein wesentlicher Bestandteil von correspSearch ist die Möglichkeit, den Datenbestand mit Metadaten zu eigenen - digitalen oder gedruckten - Briefeditionen zu erweitern. Durch die damit einhergehende Standardisierung der Metadaten auf der Grundlage der Richtlinien der Text Encoding Initiative (TEI) werden Editionen für die digitale Vernetzung erschlossen. Je mehr Briefmetadaten bereitgestellt und aggregiert werden, desto größer ist der Nutzen dieser Daten für die Recherche und Forschung. Mit dem CMIF Creator erfasste Daten werden im „Correspondence Metadata Interchange Format" (CMIF) kodiert, das von der TEI Correspondence Special Interest Group entwickelt und gepflegt wird. Anschließend werden die TEI XML-Daten auf Anbieterseite online verfügbar gemacht und vom Webservice per URL abgerufen. Die Metadaten der Briefeditionen werden somit dezentral aggregiert und erfordern nicht den Betrieb einer zentralen Speicherlösung. Der Webservice correspSearch, sowie das zugrundeliegende CMIF und das TEI-Elementset correspDesc wurden 2018 mit dem *Rahtz Prize for TEI Ingenuity*[2] ausgezeichnet.

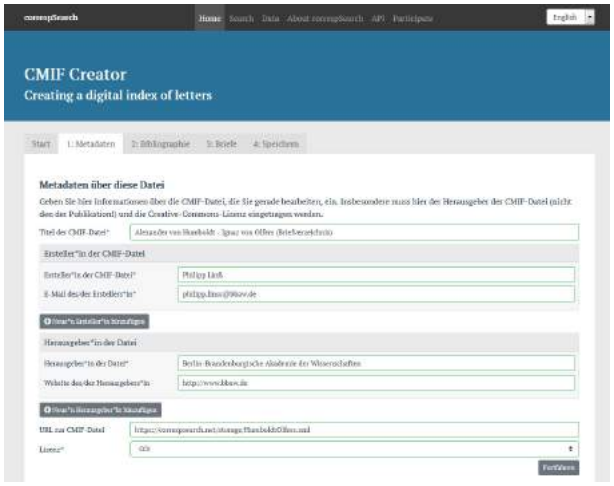


Abbildung 1. Oberfläche des CMIF Creator, Schritt 1: Metadaten.

Abbildung 2. Listenansicht der Briefdatensätze.

War es bis vor kurzem nötig, das CMIF manuell zu kodieren, bietet der *CMIF Creator*[3] nun eine benutzerfreundliche grafische Oberfläche, um eigene Brief-Metadaten in das nötige Format zu übertragen.[4] Der mittlerweile in seiner zweiten Version verfügbare CMIF Creator bietet eine einfach zugängliche Möglichkeit, online und browserbasiert Briefeditionen in TEI-konforme CMIF-Dateien zu übertragen. Diese können dann in das Datenbanksystem von correspSearch überführt werden. Durch die browserbasierte Anwendung, die nicht auf serverseitige Speicherroutinen angewiesen ist, können Nutzer*innen die gesamte Verarbeitung lokal ausführen. Das bedeutet, dass die Nutzer*innen jederzeit vollständige Kontrolle über ihre Daten haben, da diese ausschließlich lokal gespeichert werden. Das Zwischenspeichern und Laden begonnener CMIF-Dateien ermöglicht weiterhin eine Bearbeitung über einen längeren Zeitraum. Der CMIF Creator unterstützt die Abfrage von Normdaten der *Gemeinsamen Normdatei* (GND) der Deutschen Nationalbibliothek[5] über die von *lobid* bereitgestellte API[6] , sowie das Abrufen von ortsbezogenen Normdaten aus der Datenbank des Ortsnamendienstes *GeoNames*[7] , und ermöglicht eine in das User Interface integrierte Auswahl und Zuweisung von Normdaten zu Korrespondent*innen und Orten. Die

Unterstützung von Normdaten ist essentiell, da correspSearch für die Auflösung der Ambiguitäten von Personen- und Ortsnamen wesentlich auf die Nutzung von eindeutigen Normdaten-IDs zurückgreift. Die Oberfläche des Editors ist bewusst einfach und gleichzeitig so flexibel wie möglich gestaltet, um das CMIF Format im Frontend möglichst gut abzubilden und auch für weniger technisch versierte Nutzer*innen verfügbar zu machen.

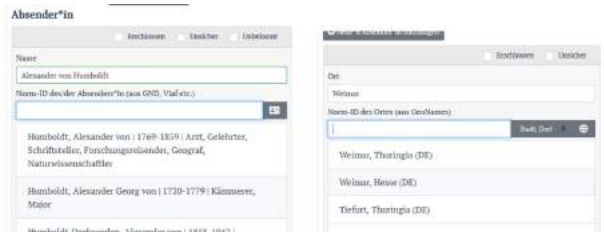


Abbildung 3. Listenauswahl von Normdaten.

In den letzten Jahren hat sich JavaScript durch die breite Verwendung von hochentwickelten JavaScript-Frameworks zu einer der zentralen Skriptsprachen im Bereich des Frontend-Development entwickelt. Das zugrundeliegende Datenformat JSON bietet seinerseits eine bewährte und einfache Möglichkeit der Datenspeicherung im Frontend. Da XML-Frameworks oder -Bibliotheken für einen browserbasierten Editor dieses Umfangs als weitestgehend veraltet[8] und unzureichend performant evaluiert wurden, erfolgte die Entscheidung schließlich zugunsten einer JavaScript- und JSON-basierten Anwendung. Durch die Crossplatform-Tauglichkeit von JavaScript bietet sich in der Perspektive darüber hinaus großer Spielraum für die Weiterentwicklung des CMIF Creators analog zum Format und den Bedürfnissen der Nutzer*innen.

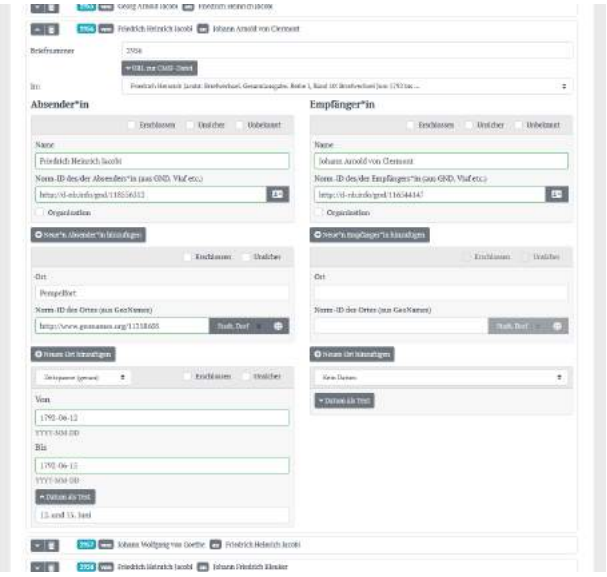


Abbildung 4. Ansicht eines Briefdatensatzes.

Während der Arbeit im Editor verbleiben die Daten im JSON-Format und ermöglichen so eine dynamische Nutzung im Kontext des dem Editor zugrundeliegenden reaktiven JavaScript-Frameworks *vue.js*[9] . Durch die Verwendung von *vue.js* in Verbindung mit dem CSS-Framework *Bootstrap*[10] wird jede Eingabe und Änderung der Nutzer*innen unmittelbar im JSON-Datenmodell abgebildet. Beim Speicherungsvorgang werden die JSON-Daten über die correspSearch-eigene API in valides XML serialisiert, das dann später von correspSearch abgerufen werden kann. Im Falle von invaliden Eingaben enthält der CMIF Creator interne Validierungsfunktionen, die sowohl nach der Eingabe in einzelnen Feldern, als auch vor dem abschließenden Speichervorgang, die Eingaben auf technische Korrektheit überprüft. Finden sich Fehler, so werden diese nicht nur lokalisiert, sondern auch verlinkt, sodass Nutzer*innen direkt zum entsprechenden Eintrag springen können. Beim Laden von validen XML-Dateien in den Editor findet die Umwandlung in JSON ebenfalls über die correspSearch-API statt. Der CMIF Creator wird als OpenSource-Software entwickelt und unter der GNU LGPL lizensiert. Die Quellen sind auf der GitHub-Seite von correspSearch[11] frei zugänglich.

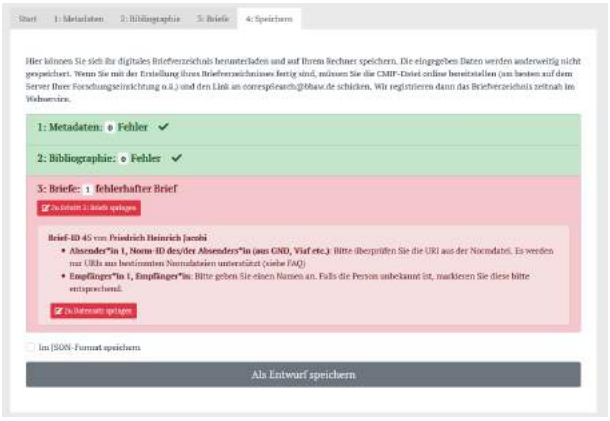


Abbildung 5. Bei fehlerhaften Einträgen kann die CMIF-Datei nur als JSON-Entwurf gespeichert werden.

Abbildung 6. Gibt es keine Fehler, können die Daten als valide .xml-Datei gespeichert werden.

Fußnoten

1. https://correspsearch.net/
2. http://www.bbaw.de/presse/pressemitteilungen/pressemitteilungen-2018/Rahtz-Prize-for-TEI-Ingenuity-2018
3. https://correspsearch.net/creator/index.xql

4. Der CMIF Creator basiert auf der schon in Andert et al. 2015 geäußerten Annahme, dass die Eingabe von Briefmetadaten möglichst effizient gestaltet sein sollte. Im Gegensatz zu dem von Andert et al. skizzierten Tool arbeitet der CMIF Creator allerdings standardbasiert. Darüber hinaus überlässt der CMIF Creator die Daten in der Obhut ihrer Ersteller*innen, die jederzeit vollständige Einsicht in die und volle Kontrolle über die (lokal) gespeicherten XML-Daten haben.
5. http://www.dnb.de/
6. http://lobid.org/gnd/api
7. http://www.geonames.org/
8. So lief z.B. die Unterstützung für betterForm in existdb Anfang 2018 aus, siehe hierzu auch den Eintrag in der Mailingliste https://sourceforge.net/p/exist/mailman/message/36239166/ bzw. den entsprechenden Issue https://github.com/eXist-db/exist/pull/1736
9. https://vuejs.org/
10. https://bootstrap-vue.js.org/
11. https://github.com/correspSearch

Bibliographie

Andert, Martin / Frank Berger / Paul Molitor / Jörg Ritter (2015): *An Optimized Platform for Capturing Metadata of Historical Correspondence*, Literary and Linguistic Computing Advance Access 30 (4): 471–480. https://doi.org/10.1093/llc/fqu027.

Dumont, Stefan (2016): *CorrespSearch – Connecting Scholarly Editions of Letters*, Journal of the Text Encoding Initiative, Nr. Issue 10 (Dezember). https://doi.org/10.4000/jtei.1742.

Stadler, Peter (2012): *Normdateien in der Edition*, editio 26: 174–83.

TEI Consortium, Hrsg. o. J. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0 vom 23.07.2018.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/. [zuletzt abgerufen: 12. Oktober 2018]

TEI Correspondence SIG, Hrsg. (2015): *Correspondence Metadata Interchange Format (CMIF)*, https://github.com/TEI-Correspondence-SIG/CMIF.

Das Latin Text Archive (LTA) – Digitale Historische Semantik von der Projektentwicklung an der Universität zur Institutionalisierung an der Akademie

Geelhaar, Tim
geelhaar@em.uni-frankfurt.de
Goethe Universität Frankfurt am Main, Deutschland

Einleitung

Die Arbeit zur digitalen historischen Semantik in Frankfurt am Main erreicht mit der neuen webbasierten Plattform und Datenbank „Latin Text Archive“ (LTA)[1] im Rahmen der Dateninfrastrukturen der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) ein neues Level. Nach über zehn Jahren Entwicklungsarbeit an den Datenbanken „Historical Semantics Corpus Management (HSCM)“[2], „Frankfurt Latin Lexicon (FLL)“ sowie an der Webseite „www.comphistsem.org“[3] ergaben sich drei grundlegende Herausforderungen: (1) Nachhaltigkeit und Verfügbarkeit der geleisteten Arbeit mussten gesichert, (2) Benutzerfreundlichkeit und Funktionalität verbessert sowie (3) Akzeptanz und Verankerung innerhalb der Fachwissenschaft gesteigert werden. Das LTA antwortet auf diese Herausforderungen und führt mit historischen Referenzkorpora neue Arbeitsinstrumente ein. Das LTA ist auch das Ergebnis einer neuen strategischen Partnerschaft, um die bisherige Arbeit aus der stets zeitlich begrenzten Projektentwicklung an der Universität in einen dauerhaften Betrieb an einer Akademie zu überführen.

Zur Standortbestimmung der Frankfurter Arbeit lässt sich Michael Piotrowskis Definition von *digital humanities* (Piotrowski 2018: 2) anwenden: Danach ist sie den *applied digital humanities* zuzuorten, da ein konkretes geschichtswissenschaftliches Forschungsziel mit digitalen Techniken verfolgt wird. Der Historiker Bernhard Jussen hat zur Umsetzung seines Postulates einer „kulturellen Semantik“ nach Wegen gesucht, wie computerlinguistische Methoden helfen können, Bedingungen, Mittel und Formen multimedialer Sinnproduktion in vergangenen Gesellschaften zu erforschen (Jussen 2000: 24ff.). Es geht um die kontrollierte Analyse von semantischem Wandel innerhalb der lateinischen Textproduktion in poströmischer Zeit. Ein Anwendungsgebiet ist die politische Geschichte. So lässt sich mittels der computergestützten Semantik danach fragen, welche impliziten politischen Ordnungsmodelle in Texten sichtbar werden, bevor mit der Wiederentdeckung der Politik des Aristoteles der Begriff des Politischen aufkommt? Außerdem kann die Begriffs- und Ideengeschichte Verwendungszusammenhänge von zentralen Vokabeln untersuchen (Geelhaar 2015; Schwandt 2018). Am Ende sollen aber nicht nur Forschungsergebnisse und -methoden, sondern auch Arbeitsinstrumente und Forschungsdaten einem breiten fachwissenschaftlichen Publikum ohne Programmierkenntnissen zur Weiterverwendung bereitstehen. Hierzu hat Bernhard Jussen den Computerlinguisten Alexander Mehler und dessen Text Technology Laboratory[4] für eine Kooperation gewonnen, in der die texttechnologische und die geschichtswissenschaftliche Seite ihre jeweiligen Agenden verfolgen und vom interdisziplinären Austausch profitieren. Aus dieser Konstellation ergibt sich, dass die Vorstellung des LTA aus geschichtswissenschaftlicher Perspektive den Fokus nicht auf technische bzw. technologische, sondern auf programmatische und anwenderbezogene Aspekte legt.

Das Latin Text Archive: Teil des DTA-Markenstamms

Das LTA ist eine frei zugängliche, webbasierte Plattform zu Korpusaufbau und Korpusanalyse sowie auch eine Datenbank. Technisch baut es auf den am Deutschen Textarchiv (DTA, DFG-gefördert zwischen 2007 und 2016, siehe Geyken et al. 2018) entwickelten Komponenten zur Textpräsentation auf und erweitert somit den Markenstamm des DTA um eine lateinische Textkomponente. Diese umfasst die Textproduktion im lateinischsprachigen Europa von (zunächst) 400 bis 1500. Die versammelten Texte basieren auf kritischen und somit für die Geschichtswissenschaft validen Editionen, soweit sie in Open Access verfügbar sind. Sie werden zum Zweck des Text Mining um den kritischen Apparat gekürzt[5], im TEI-P5-Format nach dem für HSCM entwickelten Datenmodell aufbereitet, vollständig lemmatisiert und mit einer aufwändigen Metadaten-Annotierung angereicht, die zugleich die Vernetzung zu anderen digitalen Ressourcen herstellt – nicht zuletzt zu den textgebenden Institutionen selbst.[6] Hierbei handelt es sich u. a. um die *Monumenta Germaniae Historica* (MGH). Dieses wichtige deutsche Editionsunternehmen für mittelalterliche Texte stellt seine Editionen im „openMGH"-Projekt unter Creative-Commons-Lizenz in TEI-konformen Versionen zur Verfügung.[7] Doch erst durch die Datenintegration ins LTA können auch reine Anwender vom openMGH-Projekt profitieren, da die Editionen nun erst wortstatistisch vergleichend analysiert werden können. Des Weiteren ist das LTA auf kontrollierte Datenerweiterung durch die gezielte Aufbereitung und Übernahme aus projektexternen Quellen ausgelegt, um in Zukunft ein repräsentatives Korpus historischer, lateinischer Textproduktion analysierbar zu machen. Hierzu können die Daten entweder als Gesamtkorpus und nach freier Auswahl durch den Anwender an Analysemodule weitergereicht werden, von denen die Voyant Tools bereits verfügbar sind.[8] Die in den Vorgängerprojekten entwickelten Analysetools werden als weiteres, externes Modul zugänglich gemacht. Dabei handelt es sich um Konkordanz- und Kookkurrenzanalysen sowie die Berechnung semantischer Netzwerke.

Zusammenhang zwischen LTA und HSCM

Das LTA unterscheidet sich in mehrfacher Hinsicht von seinen Vorgängeranwendungen. Die Überführung des Datenbestandes aus HSCM in das LTA als Teil der von der BBAW betreuten Dateninfrastruktur dient dem Zweck der nachhaltigen Verfügbarkeit. Zudem wird das LTA als explizites Parallelangebot zum DTA vom Renommee der BBAW profitieren, die durch eigene Datenprojekte nicht nur eine ausgezeichnete Expertise in den DH vorweisen kann, sondern auch bereits hohe Anerkennung in den Geisteswissenschaften genießt. Zudem gewinnt das LTA durch die Anlehnung an das DTA, da es einen Wiedererkennungseffekt in der Benutzerführung gibt, der das Arbeiten mit dem LTA erleichtert. Die wesentlichen Neuerungen gegenüber HSCM bestehen aber nicht nur in der verbesserten Benutzerführung; wichtiger noch ist die Trennung der Primärdatenaufbereitung von der Datenverwaltung, indem die Daten in HSCM kuratiert und im LTA zur Verfügung gestellt werden. Das Preprocessing neuer Texte wird weiterhin über HSCM als Teil des eHumanities- Desktops[9] laufen und über den eigens vom Text Technology Lab entwickelten TT Lab Tagger (vor der Brück/Mehler 2016; Eger/Gleim/Mehler 2016) für die automatische Lemmatisierung lateinischer Texte, über das dafür nötige morphologische Lexikon („Frankfurt Latin Lexicon") sowie über Editoren zur kontrollierenden, manuellen Nachlemmatisierung und zur nachträglichen Korrektur des TEI-Codes. Die vollständig bearbeiteten Texte werden anschließend in das LTA überführt, wo es nicht mehr möglich sein wird, in den jeweiligen Source-Code des Textes einzugreifen. Dies erlaubt eine feste Indexierung (auch der Lemmatisierungsinformationen), wodurch die Schnelligkeit bei der Verarbeitung von Suchanfragen bedeutend gesteigert wird. Darüber hinaus ist durch den Datentransfer die Analyse des Materials nicht mehr auf die in HSCM vorhandenen Tools beschränkt, sondern können im Grunde von allen denkbaren Toolkits wie eben den Voyant Tools oder Diacollo[10] weiterverwendet werden.

Geschichtswissenschaftliche Referenzkorpora

Die dritte wesentliche Neuerung sind die unter geschichtswissenschaftlichen Aspekten kontrollierten Referenzkorpora[11], um Veränderungen im Sprachgebrauch zeitlich wie genrespezifisch berechnen und visualisieren zu können. Wie die DTA-Referenzkorpora beinhalten diese Referenzkorpora ganze Werke und nicht nur Samples wie linguistische Korpora (z. B. das British National Corpus). Das Clustering von Texten wird nach Vierteljahrhunderten und nicht nach Zehn-Jahres-Schritten wie im DTA geschehen[12], weil eine präzisere zeitliche Zuordnung aufgrund fehlender Datierungen und mitunter komplizierten Überlieferungsgeschichten nicht möglich ist. Dieser Arbeitsschritt erforderte zudem eine erneute klassische Quellenkritik zur Chronologie einzelner Texte. Die Textmengen pro Zeiteinheit sollen quantitativ nicht zu sehr voneinander abweichen, was angesichts der teilweise eklatanten Disparität historischer Schriftproduktion eine große Herausforderung darstellt. Außerdem wird, soweit möglich, der Verbreitungsgrad handschriftlicher Überlieferung berücksichtigt, wenngleich das LTA ansonsten an der Idee des Textes als abstrakter Größe aus konzeptionellen Gründen festhalten muss.[13] Das erste Referenzkorpus besteht aus narrativen Texten, die aus den Scriptores-Reihen der MGH stammen und historiographische wie hagiographische Texte beinhalten. Künftige Korpora werden Briefe bzw. Urkunden und vor allem auch theologische bzw. juristische Traktate umfassen, um somit Sprachgebrauch in verschiedenen Genres vergleichen zu können.

Fußnoten

1. http://lta.bbaw.de
2. Jussen/Mehler/Ernst 2007; Cimino/Geelhaar/Schwandt 2015. HSCM wurde zwischen 2008 und 2014 aus den Mitteln

des Gottfried Wilhelm Leibniz-Preises der DFG sowie aus den Mitteln des LOEWE-Schwerpunktes „Digitale Humanities“ finanziert, um im BMBF-Projekt „Computational Historical Semantics“ (2013-2016) weiterentwickelt zu werden.
3. Eine Präsentation der Frankfurter Projekte für das CCeH 2017 findet sich unter: http://www.geschichte.uni-frankfurt.de/43013259/geelhaart (alle folgenden Links wurden eingesehen am 6.10.2018)
4. https://www.texttechnologylab.org/
5. Kritisch hierzu Fischer 2017: S266.
6. Es gibt Verlinkungen, soweit möglich, zum Repertorium Fontium Mediae Aevi (www.geschichtsquellen.de) , zu VIAF (Personen und Werke) und zum Katalog der Staatsbibliothek zu Berlin für bibliographische Angaben.
7. http://www.mgh.de/dmgh/openmgh/
8. https://voyant-tools.org/ Zu dessen Anwendung in der Geschichtswissenschaft siehe Schwandt 2018: 125-133.
9. HSCM ist ein Modul der VRE „eHumanities Desktop 2.2“ (www.hudesktop.hucompute.org) des Text Technology Laboratory.
10. https://clarin-d.de/de/kollokationsanalyse-in-diachroner-perspektive
11. Zu den Schwierigkeiten mit historischen Korpora und von Historikern organisierten Korpora siehe Geelhaar 2015: 11f.
12. Unser Kooperationspartner IRHT/CNRS verfolgt im Corpus-Building-Project VELUM (http://www.agence-nationale-recherche.fr/Project-ANR-17-CE27-0015) eine sehr viel gröbere Stratifikation.
13. Eine Korpusanreicherung mittels digital edierter Handschriften ist technisch in HSCM/LTA realisierbar, würde aber zu Inkonsistenzen im Materialbestand führen. Hierzu auch Fischer 2017: S280.

Bibliographie

Vor der Brück, Tim / Mehler, Alexander (2016): *„TLT-CRF: A Lexicon-supported Morphological Tagger for Latin Based on Conditional Random Fields”*, in: “Proceedings of the 10th International Conference on Language Resources and Evaluation”.

Eger, Steffen/ Gleim, Rüdiger / Mehler, A. (2016): *„Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art”*, in: “Proceedings of the 10th International Conference on Language Resources and Evaluation”.

Cimino, Roberta / Geelhaar, Tim / Schwandt, Silke (2015): *“Digital Approaches to Historical Semantics: new research directions at Frankfurt University”*. In: Storicamente 11. http://storicamente.org/historical_semantics [letzter Zugriff 12.10.2018] 7. DOI: 10.12977/stor594

Fischer, Franz (2017): *„Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Critism”*, in: Speculum 92/S1: S266-S287. https://doi.org/10.1086/693823

Geelhaar, Tim (2015): *“Talking About christianitas at the Time of Innocent III (1198–1216): What Does Word Use Contribute to the History of Concepts?”* in: Contributions to the history of concepts 10/2: 7–28. https://doi.org/10.3167/choc.2015.100202

Geyken, Alexander / Boenig, Matthias / Haaf, Susanne / Jurish, Bryan / Thomas, Christian / Wiegand, Frank (2018): *„Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN“*, in: **Lobin, Henning/ Schneider, Roman / Witt, Andreas (Hgg.)**: *Digitale Infrastrukturen für die germanistische Forschung* (= Germanistische Sprachwissenschaft um 2020, Bd. 6). Berlin/Boston: De Gruyter, 219–248. https://doi.org/10.1515/9783110538663-011

Jussen, Bernhard, Mehler, Alexander / Ernst, Alexandra (2007): *„A Corpus Management System for Historical Semantics. Sprache und Datenverarbeitung“*, in: International Journal for Language Data Processing 31/2: 81-87.

Jussen, Bernhard (2000): *Der Name der Witwe. Erkundungen zur Semantik der mittelalterlichen Bußkultur*. (VMPIG, Bd. 158). Göttingen.

Piotrowski, Michael (2018): *„Digital Humanities – An explication”*, in: **Burghardt, Manuel, Müller-Birn, Christian (eds.)**: *INF-DH 2018 – Workshopband*, 25. Sept. 2018, Berlin https://doi.org/10.18420/infdh2018-07

Schwandt, Silke (2018): *„Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora“*, in: Geschichte und Gesellschaft 44: 107-134. https://doi.org/10.13109/gege.2018.44.1.107

Mehler, Alexander / vor der Brück, Tim / Gleim, Rüdiger / Geelhaar, Tim (2015): *„Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger”*, in Text Mining: From Ontology Learning to Automated text Processing Applications, C. Biemann and A. Mehler, Eds., Berlin/New York: Springer, 2015, pp. 87-112.

Mehler, Alexander / Schwandt, Silke / Gleim, Rüdiger / Jussen, Bernhard: *„Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien”"*, Journal for Language Technology and Computational Linguistics (JLCL), vol. 26, iss. 1, pp. 97-117, 2011.

Das Niklas-Luhmann Archiv - Ein multimediales Dokumenten-Netz

Zimmer, Sebastian
sebastian.zimmer@uni-koeln.de
Universität zu Köln, Deutschland

Gödel, Martina
mgoedel@uni-koeln.de
Universität zu Köln, Deutschland

Persch, Dana
dana.persch@uni-koeln.de
Universität zu Köln, Deutschland

Niklas Luhmann (1927-1998) zählt zu den bedeutendsten Soziologen des 20. Jahrhunderts. Sein wissenschaftlicher Nachlass umfasst u.a. einen ca. 90.000 (größtenteils handschriftliche) Notizzettel umfassender Zettelkasten, den Luhmann zwischen 1953 und 1996 gepflegt hat. Daneben finden sich annähernd 200 bislang unveröffentlichte Manuskripte von teils erheblichem Umfang. Im Rahmen

des Projekts erfolgt eine archivarische Sicherung und theoriegenetische Erschließung des Nachlasses sowie im Anschluss daran eine Überführung des Materials in ein Internetportal, welches im Rahmen der DHd 2019 veröffentlicht wird. Ergänzend zur Präsentation des Nachlassmaterials wird das Portal eine Sammlung von Audio- und Videoaufnahmen präsentieren, die Vorlesungen, Seminaren, Vorträgen sowie Radio- und Fernsehinterviews Luhmanns dokumentieren.

Ein Ziel des Projektes ist es, sowohl die Dokumente innerhalb eines Bestandes als auch die verschiedenen Dokumentenbestände selbst auf der Basis der editorischen Bearbeitung miteinander in Beziehung zu setzen. Das Webportal soll es dem Nutzer ermöglichen, Querverbindungen zwischen den Materialbeständen nachzuvollziehen und so das Gesamtwerk Luhmanns in seiner Vernetztheit kennenzulernen. Dies soll mit verschiedenen multimedialen und multimodalen Einstiegs- und Navigationsmöglichkeiten erreicht werden: Facettierte Suchmöglichkeiten, interaktive und editorisch aufbereitete Visualisierungen, bis hin zu 3D-Umgebungen. Die Visualisierungen bieten verschiedene Perspektiven auf das Werk: sowohl Gesamtüberblicke als auch dokumentenzentrierte Ansichten.

Exemplarisch lässt sich das an der Visualisierung des Zettelkastens zeigen: Eine besondere editorische Herausforderung besteht beim Zettelkasten darin, dass die Sammlung durch zwei Merkmale gekennzeichnet ist, deren Kombination das besondere theoretische Kreativitätspotential der Sammlung begründet, zugleich aber ihre lineare Lesbarkeit erschwert, wenn nicht verunmöglicht: (a) eine **nichthierarchische Ordnungsstruktur** aufgrund eines ausschließlich lokalen Anschlussprinzips bei der Einstellung neuer Zettel, so dass ursprünglich direkt hintereinander stehende Zettel durch später eingestellte Zettel getrennt werden; (b) ein **Verweisungssystem**, bei dem die thematisch oder konzeptionell miteinander zusammenhängenden, aber eben verstreut in der Sammlung stehenden Zettel aufeinander verweisen, indem auf den Zetteln jeweils die entsprechenden Zettelnummern notiert werden.

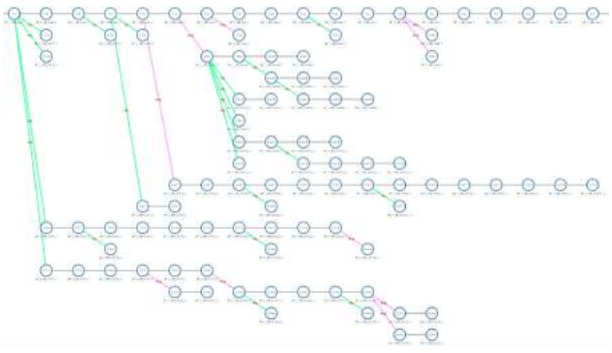


Abbildung 1: Ausschnitt einer Visualisierung der im Rahmen der Edition erstellten inhaltlich-logischen Einordnungs- und Navigationsstruktur des digitalen Zettelkastens.

Um eine Lesbarkeit der Sammlung zu ermöglichen, werden bei der fachwissenschaftlichen Edition die entsprechenden Argumentationsstränge identifiziert, wobei erstens eine Verknüpfung von (ursprünglichen) Zettelfolgen vorgenommen und zweitens später eingeschobene oder ergänzende Diskussionsstränge als davon abgehende Zettelfolgen entsprechend platziert werden. Um aber eine globale Perspektive auf die Sammlung zu erhalten, die insbesondere auch die Verweisungen auf andere Zettel berücksichtigt und die damit die Netzwerkstruktur der Sammlung deutlich macht, wurde eine entsprechende dreidimensionale Visualisierung entwickelt, die sowohl die inhaltlich-logischen Zettelfolgen in einer zweidimensionalen Ebene enthält als auch abschnittsübergreifende Querverweise, die aus dieser Ebene herausragen. Diese digitale 3D-Umgebung kann auch mittels VR-Hardware erkundet werden. So kann sich der Benutzer seine Perspektive auf den Bestand stufenlos selbst auswählen: Von der Vogelperspektive über den gesamten Zettelkasten bis hin zur Ego-Perspektive ausgehend von einem bestimmten Zettel.

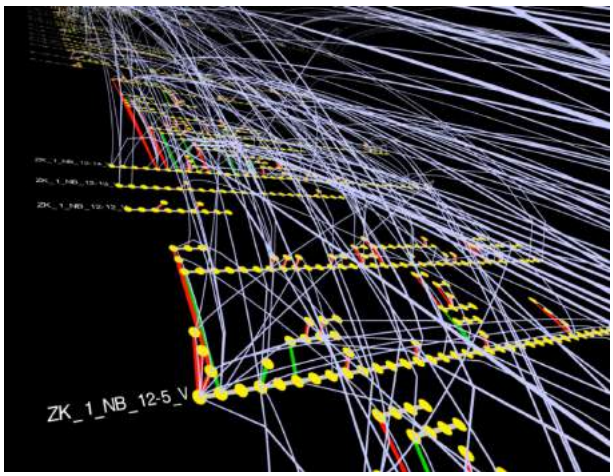


Abbildung 2: Gedankengänge aus der Vogelperspektive: Die verschiedenen Argumentationsstränge des Zettelkastens und deren Verweise in einer 3D-Umgebung

Ein besonders wichtiges verbindendes Element zwischen den zu erschließenden Materialien aus dem Nachlass Niklas Luhmanns sind darüber hinaus die bibliographischen Informationen im Nachlassmaterial. Ihre Modellierung, Zusammenführung und Visualisierung ermöglicht detaillierte Einblicke in Luhmanns Arbeitsweise, aber auch in die Genese seiner Werke. Um die Berührungspunkte sichtbar zu machen, wird eine umfassende bibliographische Datenbank aufgebaut. Die Datensätze bündeln Informationen zur Erwähnung und Zitation einzelner Werke durch Luhmann und machen so die Rezeptionsstruktur innerhalb des Luhmannschen Werks deutlich. Zusätzlich können sie als Brücke vom Zettelkasten zu den Manuskripten genutzt werden; auf diese Weise ermöglichen sie den Editoren bestandsübergreifende Verbindungen zu etablieren **:** zwischen Luhmanns Manuskripten, deren verschiedenen Fassungen und den veröffentlichten Werken, den Audio- und Videodokumenten **,** auf denen Vorträge und Vorlesungen Luhmanns dokumentiert sind, und dem Zettelkasten.

Bibliographie

Gfrereis, Heike / Strittmatter, Ellen (2013): *Zettelkästen. Maschinen der Phantasie.* Ausstellungskatalog. Deutsche Schillergesellschaft. Marbach a.N.

Gödel, Martina / Schmidt, Johannes / Zimmer, Sebastian (2018): *Digitale Differenz. Luhmanns Zettelkasten als physisch-historisches Objekt und als vernetzter Navigationsraum (Vortragsabstract)*, in: **Georg Vogeler (ed.)**: *DHd 2018: Kritik der digitalen Vernunft*. Köln, 178-181 (http://dhd2018.uni-koeln.de/)

Gödel, Martina / Zimmer, Sebastian (2017): *Niklas Luhmanns Werk und Lesekosmos - DH in der bibliographischen Dimension (Vortragsabstract)*, in: **Georg Vogeler (ed.)**: *Konferenzabstracts. DHd 2017 Bern: Digitale Nachhaltigkeit*. Bern, 180-184 (http://www.dhd2017.ch/)

Krajewski, Markus (2011): *Paper Machines. About cards & catalogs*, 1548-1929. Cambridge: MIT Press.

Schmidt, Johannes F.K. (2018): *"Niklas Luhmann´s Card Index: Thinking Tool, Communication Partner, Publication Machine"*, in: **Alberto Cevolini (ed.)**: *Forgetting Machines. Knowledge Management Evolution in Early Modern Europe*. Leiden: Brill, 289-311.

Watts, Duncan (2004): *"The »new« science of networks", in: Annual Review of Sociology* 30, 243-270.

Dependenzbasierte syntaktische Komplexitätsmaße

Proisl, Thomas

thomas.proisl@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Konle, Leonard

leonard.konle@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg

Evert, Stefan

stefan.evert@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg

Die Beschreibung der Komplexität von (literarischen) Texten muss für jeden Aspekt, also Vokabular, Satz/Syntax, uneigentliche Rede, Intertextualität usw., gesondert vorgenommen werden. Im Folgenden beschäftigen wir uns mit dem Aspekt Satz/Syntax, der lange Zeit vor allem über die durchschnittliche Satzlänge erfasst wurde (Sherman 1893, Flesch 1948, Best 2005). Dabei bleibt aber, so eine naheliegende Vermutung, die interne syntaktische Komplexität eines Satzes unberücksichtigt. Die meisten Leser würden z. B. einen stark verschachtelten Satz als syntaktisch komplexer einstufen als eine gleich lange parataktische Konstruktion. Unsere Arbeit zielt darauf, diesen Aspekt unter Verwendung der im *Natural Language Processing* (NLP) weitverbreiteten dependenzbasierten Syntaxmodelle messen zu können. Kontext unserer Überlegungen ist das Unterfangen, Textkomplexität quantitativ zu erfassen. So können Annahmen in der Literaturwissenschaft und Linguistik über die unterschiedliche Komplexität der Texte bestimmter Gattungen, Autoren oder gar von Teilsystemen, z. B. populäre Literatur vs. Hochliteratur, empirisch überprüft werden. Bislang wird die syntaktische Komplexität überwiegend auf Phrasenstrukturbäumen ermittelt (für eine Übersicht siehe Vajjala Balakrishna 2015: 51–52), allerdings fehlen dafür in vielen Sprachen verlässliche NLP-Werkzeuge. Auf der anderen Seite stehen mit dem Universal-Dependencies-Projekt (Nivre u. a. 2016)[1] bereits mehr als 100 manuell erstellte Baumbanken in über 60 Sprachen (darunter auch ältere Sprachstufen) in einer sprachübergreifend konsistenten Annotation zur Verfügung und es gibt computerlinguistische Pipelines wie etwa UDPipe (Straka und Straková 2017),[2] die Texte in allen diesen Sprachen tokenisieren, taggen, lemmatisieren und parsen können. Von daher liegt es nahe, die syntaktische Komplexität von Texten auch mit dependenzbasierten Maßen zu messen. Einen ersten Vergleich von dependenzbasierten Komplexitätsmaßen hat Oya (2012) durchgeführt.

Für unsere Untersuchung verwenden wir ein deutschsprachiges Korpus von knapp 1.000 Romanen aus den letzten 60 Jahren. Bei etwa 85% der Texte handelt es sich um Heftromane (Romanzen (13%), Science Fiction (65%) und Horror (7%)), bei den restlichen 15% um Hochliteratur (kanonische Texte und/oder Literaturpreisträger). Alle Texte wurden mit dem DARIAH-DKPro-Wrapper (Jannidis u. a. 2016)[3] verarbeitet.

Syntaktische Komplexitätsmaße sind typischerweise auf Satzebene definiert. Wir berechnen für jeden Satz die folgenden Maße:[4]

- *Average dependency distance* (= durchschnittlicher Abstand zweier durch eine Dependenzrelation verbundener Tokens (Liu 2008; Oya 2011))
- *Closeness centrality* des Wurzelknotens (= Kehrwert der durchschnittlichen Länge der kürzesten Pfade vom Wurzelknoten zu allen anderen Knoten); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- *Closeness centralization* (= Erweiterung der closeness centrality von einem einzelnen Knoten auf einen ganzen Graphen (Freeman 1978)); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- *Outdegree centralization*, die Erweiterung der *outdegree centrality* (= Anzahl der von einem Knoten ausgehenden Kanten) von einem einzelnen Knoten auf einen ganzen Graph (Freeman 1978); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- Durchschnittliche Anzahl von Dependenten pro Token
- Höhe des Dependenzbaums (= der längste kürzeste Pfad vom Wurzelknoten zu einem anderen Knoten)

Zum Vergleich ermitteln wir zusätzlich die Satzlänge, d. h. die Anzahl Tokens pro Satz. Um einen Wert für die

syntaktische Komplexität eines gesamten Textes zu erhalten, bilden wir jeweils die Mittelwerte über alle Sätze.

Die Ergebnisse sind in den folgenden sechs Grafiken als Boxplots dargestellt (der weiße Kreis markiert zusätzlich das arithmetische Mittel):

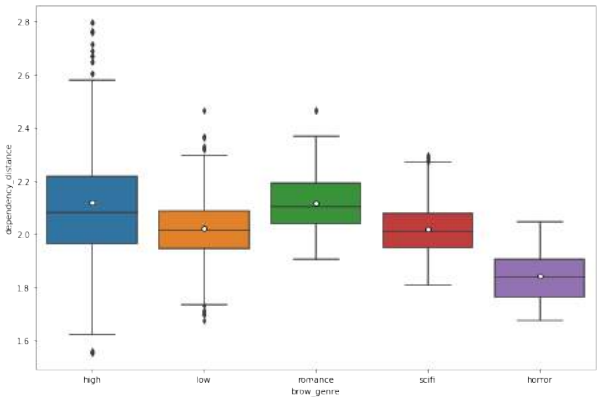


Abbildung 1. Average dependency distance

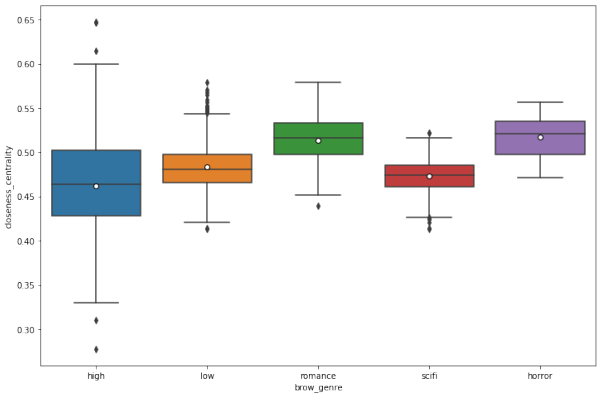


Abbildung 2. Closeness centrality

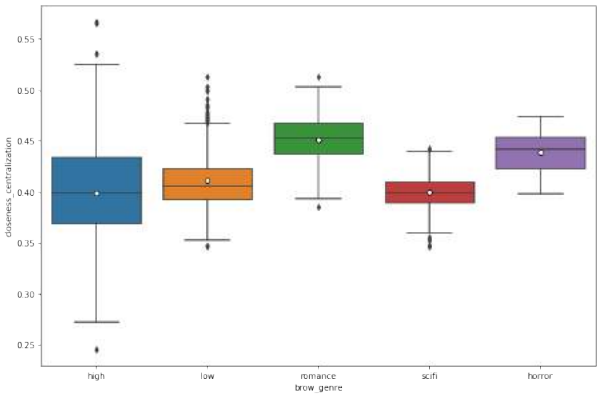


Abbildung 3.Closeness centralization

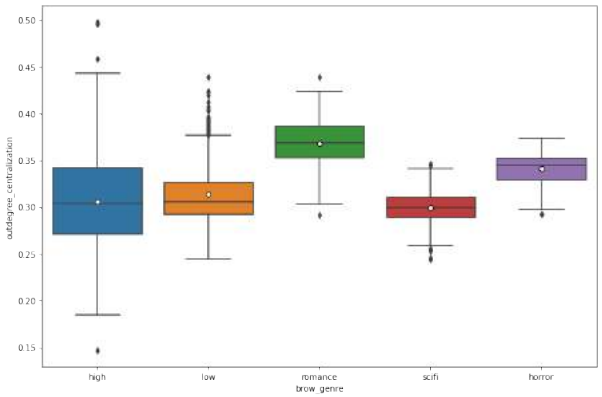


Abbildung 4. Outdegree centralization

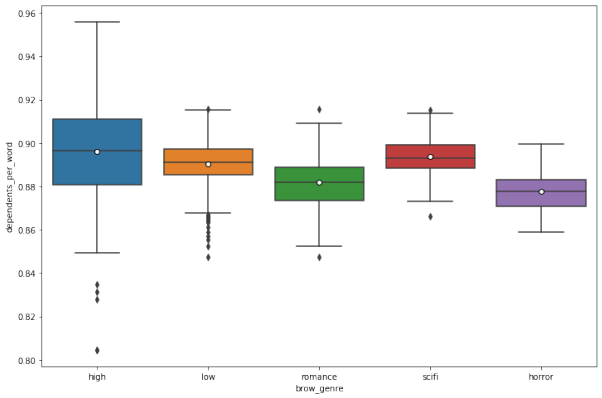


Abbildung 5. Dependenten pro Token

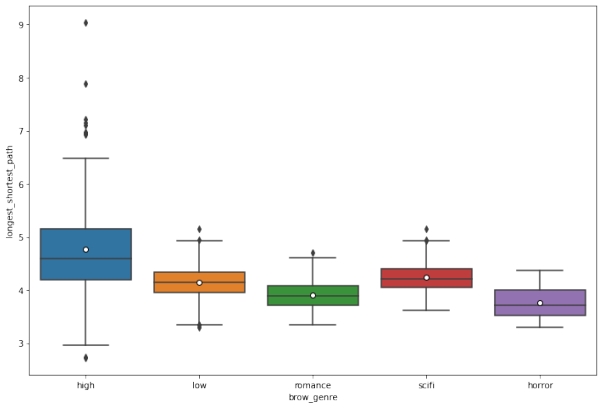


Abbildung 6. Höhe des Dependenzbaums

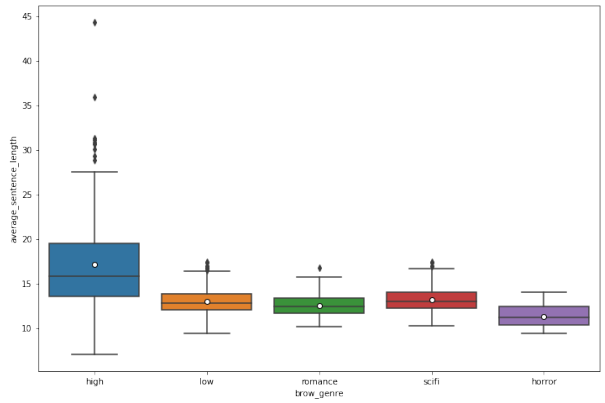


Abbildung 7. Satzlänge

Die Boxplots für Hoch- und Schemaliteratur insgesamt würden nahelegen, dass es für die untersuchten Maße keinen statistisch signifikanten Unterschied zwischen Hoch- und Schemaliteratur gibt. Die Detailansicht für die einzelnen Unterkategorien der Schemaliteratur bringt jedoch Interessantes zu Tage. Zwischen den einzelnen Kategorien untereinander gibt es deutlich ausgeprägtere Unterschiede als zwischen Hoch- und Schemaliteratur insgesamt. Besonders auffällig ist, dass fast alle Maße eine signifikant höhere Komplexität für Science-Fiction-Literatur anzeigen als für Romanzen oder Horrorhefte. Wahrscheinlich liegt das daran, dass das SF-Teilkorpus aus Romanen der Serie ‚Perry Rhodan' besteht, der auch von literaturwissenschaftlicher Seite eine Sonderrolle innerhalb der Heftromane zugeschrieben wird (Nast 2017). Ebenfalls auffällig ist, dass alle Maße eine viel größere Streuung für die Hochliteratur aufweisen als für die zahlenmäßig viel stärker vertretene Schemaliteratur. Dafür bieten sich zwei Erklärungsmodelle an: a) die Hochliteratur besteht eigentlich aus mehreren Gattungen, die sich wiederum deutlich voneinander unterscheiden; b) der Unterschied lässt sich auf die unterschiedlichen Eigenschaften der literarischen Teilfelder zurückführen – In der Hochliteratur dominiert der Wert Variation/Überraschung, in den populären Genres der Wert Erwartbarkeit, wahrscheinlich sogar erzwungen durch Lektoren.

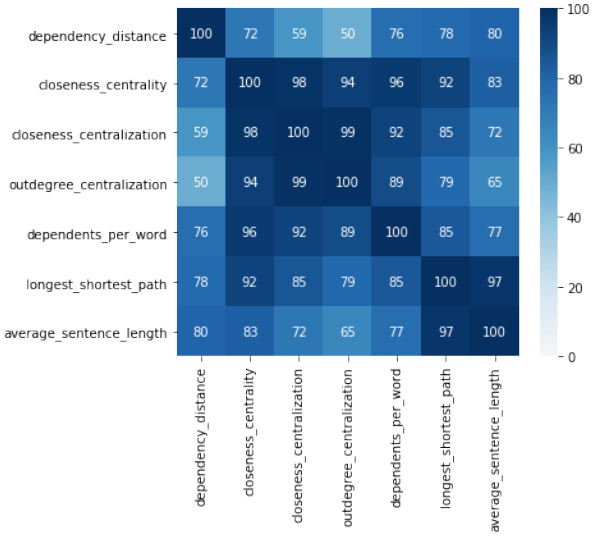


Abbildung 8.Pearson-Korrelationen zwischen den Komplexitätsmaßen

Eine Analyse der Pearson-Korrelationen zwischen den Komplexitätsmaßen zeigt, dass einige davon nahezu perfekt korrelieren.[5] So beispielsweise *closeness centralization* und *outdegree centralization* (r = 0,99), *closeness centrality* und *closeness centralization* (r = 0,98), die Höhe des Dependenzbaums und die Satzlänge (r = 0,97) und *closeness centrality* und die Anzahl Dependenten pro Wort (r = 0,96). *Average dependency distance* ist den anderen dependenzbasierten Maßen am unähnlichsten (0,50 ≤ r ≤ 0,78). Insgesamt betrachtet, korrelieren die verwendeten Maße zwar recht robust mit der durchschnittlichen Satzlänge (0,65 ≤ r ≤ 0,83), scheinen sich aber (mit Ausnahme der Höhe des Dependenzbaums) doch auch ausreichend stark von ihr zu unterscheiden um den durch das Parsen der Texte und Berechnen der dependenzbasierten Maße entstehenden Mehraufwand zu rechtfertigen. Zusätzlich könnte es durch die gezielte Entwicklung längenkorrigierter Maße gelingen, unterschiedliche Aspekte syntaktischer Komplexität getrennt voneinander zu erfassen.

Fußnoten

1. http://universaldependencies.org/
2. http://ufal.mff.cuni.cz/udpipe
3. https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper
4. https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity
5. *Closeness centrality*, *closeness centralization* und *outdegree centralization* wurden für diese Analyse mit −1 multipliziert, damit für alle Maße größere Werte eine höhere Komplexität anzeigen.

Bibliographie

Best, Karl-Heinz (2005): *Satzlänge*, in: **Köhler, Reinhard / Altmann, Gabriel / Piotrowski, Rajmund G.:** *Quantitative Linguistik / Quantitative Linguistics*. Berlin: de Gruyter-Mouton 298–304.

Flesch, Rudolf (1948): *A New Readability Yardstick*, in: Journal of Applied Psychology 32 Nr. 3: 221–233. 10.1037/h0057532 [letzter Zugriff am 8. Januar 2019].

Freeman, Linton C. (1978): *Centrality in Social Networks. Conceptual Clarification*, in: Social Networks 1, Nr. 3: 215–239. 10.1016/0378-8733(78)90021-7 [letzter Zugriff am 15. Oktober 2018].

Jannidis, Fotis / Pernes, Stefan / Pielström, Steffen / Reger, Isabella / Reimers, Nils / Vitt, Thorsten (2016): *DARIAH-DKPro-Wrapper Output Format (DOF) Specification*, in: DARIAH-DE Working Papers 20. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2016-6-2 [letzter Zugriff am 15. Oktober 2018].

Liu, Haitao (2008): *Dependency Distance as a Metric of Language Comprehension Difficulty*, in: Journal of Cognitive Science 9, Nr. 2: 159–191. http://cogsci.snu.ac.kr/jcs/issue/vol9/no2/JCS_Vol_09_+No_2+p.+159+- +191+Dependency+Distance+as+a+Metric+of+Language+Comprehension+Difficulty.pdf [letzter Zugriff am 15. Oktober 2018].

Nast, Mirjam (2017): *„Perry Rhodan" lesen. Zur Serialität der Lektürepraktiken einer Heftromanserie*. Bielefeld: transcript.

Nivre, Joakim / >de Marneffe, Marie-Catherine / Ginter, Filip / Goldberg, Yoav / Hajic, Jan / Manning, Christopher D. / McDonald, Ryan / Petrov, Slav / Pyysalo, Sampo / Silveira, Natalia / Tsarfaty, Reut / Zeman, Daniel (2016): *Universal Dependencies v1: A Multilingual Treebank Collection*, in: **Calzolari, Nicoletta / Choukri, Khalid / Declerck, Thierry / Goggi, Sara / Grobelnik, Marko / Maegaard, Bente / Mariani, Joseph / Mazo, Hélène / Moreno, Asunción / Odijk, Jan / Piperidis, Stelios (eds.):** *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portorož: European Language Resources Association 1659–1666. http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf [letzter Zugriff am 15. Oktober 2018].

Oya, Masanori (2011): *Syntactic Dependency Distance as Sentence Complexity Measure*, in: Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics. 313–316. https://www.researchgate.net/profile/Masanori_Oya2/publication/266584664_Syntactic_Dependency_Distance_as_Sentence_Complexity_Measure/links/54fe480f0cf2672e223ed842.pdf [letzter Zugriff am 15. Oktober 2018].

Sherman, Lucius Adelno (1893): *Analytics of literature, a manual for the objective study of English prose and poetry*. Boston: Ginn. https://archive.org/details/analyticsofliter00sheruoft/page/n3 [letzter Zugriff am 8. Januar 2019].

Straka, Milan / Straková, Jana (2017): *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*, in: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver: Association for Computational Linguistics 88–99. http://www.aclweb.org/anthology/K17-3009 [letzter Zugriff am 15. Oktober 2018].

Vajjala Balakrishna, Sowmya (2015): *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Dissertation, Eberhard-Karls-Universität Tübingen. https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/64359/THESIS-FINAL.pdf [letzter Zugriff am 15. Oktober 2018].

DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde. Drei Forschungsansätze.

Lorenz, Anne Katrin
anne.lorenz@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

Müller-Dannhausen, Lea
lea.mueller-dannhausen@gmx.de
Johannes Gutenberg-Universität Mainz

Trautmann, Marjam
marjam.trautmann@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

Einleitung

Digitale Editionen gehören zu den „prominentesten Themen" (Sahle 2017: 237) in den Digital Humanities. Sie leisten Grundlagenarbeit für die geisteswissenschaftliche Forschung und bilden ein eigenes Forschungsfeld in den Digital Humanities. Hier verortet sich auch das digitale Editions- und Forschungsprojekt „DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde" (https://sturm-edition.de).

Gegenstand ist das 1910 gegründete Berliner Kunstunternehmen „Der Sturm" um den Publizisten, Komponisten und Kritiker Herwarth Walden, der zahlreichen Künstlerinnen und Künstlern unterschiedlicher Kunstgattungen eine Plattform bot. Das Unternehmen umfasste die Zeitschrift „Der Sturm", die „Sturm"-Galerie, die „Sturm"-Bühne sowie den „Sturm"-Verlag. Neben den Veröffentlichungen des „Sturm" selbst bezeugen die überlieferten Briefe an das Ehepaar Walden den internationalen Einfluss des Unternehmens.

Unser Editionsprojekt, in dem bereits digital verfügbares Material aus dem „Sturm"-Kontext transkribiert, standardkonform nach XML/TEI P5 aufbereitet und mit Normdaten versehen wird, führt diese Quellen erstmals zentral zusammen und setzt sie mittels digitaler Methoden in Relation zueinander.

Drei Forschungsansätze im STURM-Projekt

Integraler Teil des Editionsprojektes ist die Forschung mit den STURM-Quellen. Auf dem Poster stellen wir neben der Quellenedition die folgenden drei Forschungsansätze vor, die sich explizit mit den im Projekt edierten Materialien beschäftigen.

Modellierung der STURM-Domäne

Ein Ansatz arbeitet mit den in der digitalen Quellenedition DER STURM zusammengeführten Quellen und modelliert deren Verknüpfungen und Beziehungen untereinander. Die Grundlage bildet hierbei das CIDOC Conceptual Reference Model als Domain-Ontologie im Bereich Cultural Heritage (Doerr 2009), die im Bereich der musealen Sammlungen und der bildenden Kunst weit verbreitet ist. Ergänzt wird diese Ontologie durch fachspezifische Vokabulare wie den Getty Art & Architecture Thesaurus (AAT). Die semantischen Verknüpfungen zwischen den einzelnen Quellen und Quellen-Typen beziehen sich auf die in den bereits edierten Quellen annotierten Entitäten (Personen, Orte und Werke) sowie auf weitere Entitäten, die noch in den Editionsprozess aufgenommen werden: auf Körperschaften, Ereignisse und Themen. In der digitalen Quellenedition des STURM sind für diese Entitäten bereits URIs vorhanden bzw. vorgesehen. Durch Anreicherung der Entitäten mit Normdaten werden Verknüpfungen mit weiteren Ressourcen ermöglicht.

Die semantische Modellierung der STURM-Domäne – bestehend aus den Quellen der Edition und dem, was sie bezeugen – dient der Weiterentwicklung der digitalen Quellenedition DER STURM sowie perspektivisch der Weiternutzung der dabei gewonnenen Daten, denn sie bildet die Grundlage für eine Verfügbarmachung der Daten und ihrer Zusammenhänge in Form von Linked Open Data. In dieser Form können die Daten beim Ermitteln und Zeigen von Zusammenhängen helfen und somit die Basis bilden für weitere Forschungen in einzelnen Fachwissenschaften, insbesondere in der Kunst- und in der Literaturwissenschaft.

Historische Netzwerkforschung zum „Sturm“

Ein weiterer Ansatz im Projekt beschäftigt sich mit der historischen Netzwerkforschung (Düring et al. 2016). In der bisherigen „Sturm“-Forschung steht aufgrund der Komplexität und Dezentralität der Quellen eine dezidiert „kunsthistorische Beschäftigung mit dem *Sturm*“ aus (van Rijn 2013: 11). Insbesondere an einer ‚Gesamtschau‘ zum „Sturm“ fehlt es – ein Desiderat, an das hier mit der Methode der Historischen Netzwerkforschung angeknüpft werden soll. Dafür wird für das zu untersuchende Phänomen des historischen „Sturm“ ein Gesamtnetzwerk modelliert und anhand algorithmischer sowie hermeneutischer Methoden analysiert (Brandes et al. 2013: 4). Datengrundlage der Netzwerkerhebung bilden archivalische Metadaten und die im STURM-Projekt maschinenlesbar modellierten Quellen.

Das Netzwerk ist multimodal, bestehen die einzelnen zueinander in Relation stehenden Entitäten doch aus im Kontext des „Sturm“ aufkommenden Personen, multimedialen Werken, Briefdaten und einigem mehr. Diese Daten gilt es in der Visualisierung des erhobenen Netzwerkes anschaulich zu machen (Baillot 2018: 357). Darüber hinaus offenbart die computergestützte Historische Netzwerkforschung auch in der Analyse komplexer Netzwerke ihre Stärken. Durch eine rein klassische Untersuchung des facettenreichen „Sturm“-Netzwerkes, so die These, würden Darstellung und Analyse unübersichtlicher und damit fehleranfälliger werden.

Die kritische Untersuchung der Quellen selbst – und damit einhergehend die Interpretation des erhobenen Gesamtnetzwerkes „Sturm“ – bildet einen weiteren wichtigen Schritt in der Gesamtanalyse des Kunstunternehmens hinsichtlich eines möglichen ‚gesamtgesellschaftlichen Spiegels‘. Ziele der historischen Netzwerkstudie sind das Ausmachen und die Analyse von *broken ties* (Jannidis 2017: 158) im komplexen „Sturm“-Netzwerk und die Einordnung des „Sturm“ in den zeitgenössischen kultur- und soziopolitischen Kontext.

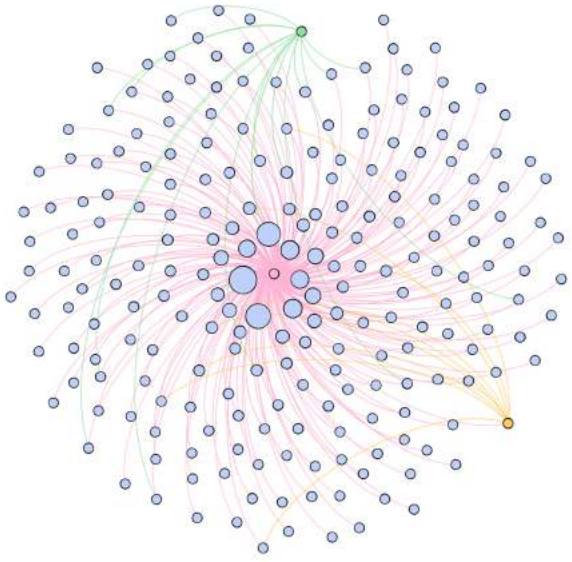


Abbildung 1. Gerichtetes Korrespondenznetzwerk um Herwarth Walden (rot), Else Lasker-Schüler (gelb) und Nell Walden (grün); Layout: ForceAtlas 2; Knotengröße: *outdegree;* Knotenfarbe: *indegree*; Kantendicke: *weight* (Menge der überlieferten Briefe).

Diskursanalyse des Simultaneitätsbegriffs im „Sturm“

Ein Beispiel für die fachspezifische Nutzung digitaler Editionen gibt die Studie zu avantgardistischen Simultaneitätskonzepten im „Sturm“. Als ein die Avantgarde bestimmendes Strukturprinzip kommt Simultaneität in verschiedenen Kunstrichtungen vor, die innerhalb ihrer Programmatik mitunter konkurrierende Modelle entwickeln. F. M. Marinettis „Technisches Manifest der futuristischen Literatur“, die kubistischen „Fenster“-Bilder Robert Delaunays oder die Simultangedichte der Dadaisten ähneln sich im grundlegenden Bestreben, nach dem avantgardistischen Primat von „Transgression und Diffundierung“ (Asholt / Fähnders 2000: 17) beim Rezipienten gleichzeitig unterschiedliche Wirklichkeitsansichten und -ebenen zu erzeugen. Im medialen Komplex des „Sturm“, mit seinen

vielfältigen Publikations- und Distributionswegen sowie seinen dezidiert antimimetischen Gestaltungsprinzipien, finden diese bildkünstlerischen wie literarischen Werke trotz Gattungs- und Stilunterschieden gleichermaßen eine adäquate Präsentationsform. Eingebettet in den spezifischen historischen Kontext des „Sturm"-Netzwerks werden sie von Para- und Metatexten begleitet, die ihre Rezeption lenken und kommentieren – und die nicht zuletzt in den privaten Korrespondenzen an Walden vorbereitet und verhandelt werden.

Die webbasierte Zusammenführung der verschiedenen Quellen des „Sturm" erlaubt es nun, wechselseitige Bezüge zwischen den unterschiedlichen Textsorten offenzulegen und so diskursiv etablierte sprachlich-rhetorische Muster zu rekonstruieren. Mit Hilfe korpuslinguistisch informierter diskursanalytischer Verfahren wird der Frage nachgegangen, inwiefern solche multimodalen Diskurspraktiken die Bedeutung von Gleichzeitigkeit im historischen „Sturm"-Netzwerk unterschiedlich konstituieren und auf diese Zeitkonstruktion als typisch avantgardistischen Diskurs verweisen. Um die entsprechenden „Diskursfragmente" (Jäger 2012: 98ff.) im Netzwerkkontext situieren und den einzelnen Künstlern und Kunstrichtungen zuordnen zu können, sollen zusätzlich netzwerkanalytische Zugänge berücksichtigt werden. Schließlich zielt die Auswertung der im Projekt erarbeiteten Daten darauf ab, den Begriff der Simultaneität im Problemfeld von Scheitern (Habermas) und Überleben (Luhmann) der Avantgarde zu konturieren.

Bibliographie

Asholt, Wolfgang / Fähnders, Walter (2000): *„Einleitung"*, in: **Asholt, Wolfgang / Fähnders, Walter (eds.)**: *Der Blick vom Wolkenkratzer. Avantgarde – Avantgardekritik – Avantgardeforschung*. Amsterdam / Atlanta: Editions Rodopoi 9–27.

Baillot, Anne (2018): *„Die Krux mit dem Netz. Verknüpfung und Visualisierung bei digitalen Briefeditionen"*, in: **Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (eds.)**: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven.* Open Access: De Gruyter 355–370.

Brandes, Ulrik / Robins, Garry / McCranie, Ann / Wasserman, Stanley (2013): *Editorial. What is network science?* Network Science 1: 1–15.

Chytraeus-Auerbach, Irene / Uhl, Elke (2013): *„Vorwort"*, in: **Chytraeus-Auerbach, Irene / Uhl, Elke (eds.)**: *Der Aufbruch in die Moderne. Herwarth Walden und die europäische Avantgarde.* Berlin: LIT Verlag 7–11.

Doerr, Martin (2009): *„Ontologies for Cultural Heritage"*, in: **Staab, Steffen / Studer, Rudi (eds.)**: *Handbook on Ontologies.* Berlin / Heidelberg: Springer Verlag 463–486.

Düring, Marten / Eumann, Ulrich / Stark, Martin / Keyserlingk, Linda v. (2016) (eds.): *Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen.* Berlin: LIT Verlag.

Jäger, Siegfried (2012): *Kritische Diskursanalyse. Eine Einführung.* 6. vollständig überarbeitete Aufl. Münster: Unrast.

Jannidis, Fotis (2017): *„Netzwerke"*, in: **Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.)**: *Digital Humanities. Eine Einführung.* Stuttgart: J. B. Metzler 147–161.

Pirsich, Volker (1985): *Der Sturm. Eine Monographie*. Herzberg: Traugott Bautz 1985.

Rijn, Maaike van (2013): *Bildende Künstlerinnen im Berliner „Sturm" der 1910er Jahre*. Tübingen: TOBIAS-lib, http://tobias-lib.uni-tuebingen.de/volltexte/2013/6987.

Sahle, Patrick (2017): *„Digitale Edition"*, in: **Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.)**: *Digital Humanities. Eine Einführung.* Stuttgart: J. B. Metzler 234–249.

Der TextImager als Front- und Backend für das verteilte NLP von Big Digital Humanities Data

Hemati, Wahed
hemati@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

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

Uslu, Tolga
uslu@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Abrami, Giuseppe
abrami@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Immer mehr Disziplinen benötigen Natural Language Processing (NLP) Werkzeuge, um automatische Textanalysen auf verschiedenen Ebenen der Sprache durchzuführen. Die Anzahl der NLP-Werkzeuge wächst rasant[1]. Auch die Anzahl der frei oder anderweitig zugänglichen Ressourcen wächst. Angesichts dieser wachsenden Zahl an Werkzeugen und Ressourcen ist es schwierig, den Überblick zu behalten; gleichzeitig ist ein Computational-Linguistic-Framework, das große Datenmengen aus verschiedenen Quellen verarbeiten kann, noch nicht etabliert. Ein solches Framework sollte in der Lage sein, Daten verteilt zu verarbeiten und gleichzeitig eine standardisierte Programmier- und Modellschnittstellebereitzustellen. Darüber hinaus sollte es modular und leicht erweiterbar sein, um die ständig wachsende Palette neuer Ressourcen und Tools zu integrieren. Das Framework muss offen genug für Erweiterungen Dritter sein, wobei jede Erweiterung für die gesamte Community zugänglich bleibt. Das Framework sollte es zudem Dritten ermöglichen, den Zugang zu ihren Erweiterungen zu beschränken, wenn dies beispielsweise durch Urheberrecht, geistiges Eigentum oder Datenschutz erforderlich ist. Um diesen Anforderungen gerecht zu werden, haben wir den TextImager (Hemati 2016, Mehler et al. 2018) um ein verteiltes Serversystem mit Cluster-Computing-Funktionen auf der Basis von UIMA (Ferrucci andLally 2004) weiterentwickelt.

UIMA ist ein Framework zur Verwaltung von Datenflüssen zwischen Komponenten. Es bietet standardisierte Interfaces zur Erstellung von Komponenten an. Dabei können die Komponenten einzeln oder im Verbund in einer Pipeline-Struktur ausgeführt werden. UIMA bietet weitgehende Möglichkeiten der sequenziellen Ordnung von NLP-Werkzeugen und verspricht, auch in Zukunft von der Community weiterentwickelt zu werden: Prozess-Management auf der Basis von UIMA erscheint nach derzeitigem Stand daher als erste Wahl im Bereich von NLP und DH.

TextImager bietet eine Vielzahl von UIMA-basierten NLP-Komponenten an, darunter unter anderen einen Tokenizierer, einen Lemmatisierer, einen Part-Of-Speech-Tagger, einen Named-Entity-Parser und einen Dependency Parser, und zwar für eine Vielzahl von Sprachen, darunter Deutsch, Englisch, Französisch und Spanisch. Dieses Spektrum an Werkzeugen besteht allerdings nicht ausschließlich aus Eigenentwicklungen, sondern wird maßgeblich um Entwicklungen Dritter erweitert, wozu unter anderem die Tool-Palette von Stanford CoreNLP (Manning 2014), OpenNLP (OpenNLP 2010) und DKpro (Eckart de Castilho 2014) zählen.

In Zeiten von Big Data wird es immer relevanter, Daten schnell zu verarbeiten. Ausdiesem Grund ist TextImager als Multi-Server- und zugleich als Multi-Instanz-Clusteraufgebaut, um das verteilte Verarbeiten von Daten zu ermöglichen. Dafür setzt TextImager auf UIMAs Cluster-Management-Dienste UIMA-AS[2] und UIMA-DUCC[3] auf.

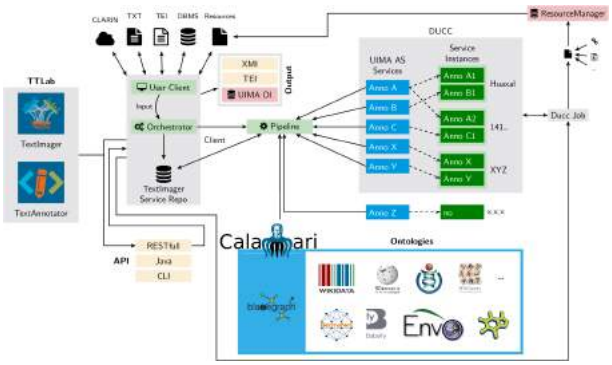


Abbildung 1

Abbildung 1 zeigt eine schematische Darstellung von TextImager. Jede NLP-Komponente läuft als UIMA-AS Webservice auf dem Computing-Cluster des TextImager. Dabei können mehrere Instanzen einer Komponente instanziiert (s. Abbildung1, Service Instances) werden und dennoch über eine Webservice-Schnittstelle (s. Abbildung1, UIMA AS Services) angesprochen werden. Dazu wird das Java Messaging Service (JMS) verwendet, das die Kommunikation zwischen verschiedenen Komponenten einer verteilten Anwendung ermöglicht. JMS implementiert ein Point-to-Point-Kommunikationssystem. Dieser Kommunikationstyp basiert auf dem Konzept der message queues (Warteschlangen), senders (Sender) und receivers (Empfänger). Jedem Dienst ist eine Eingabewarteschlange und eine Ausgabewarteschlange zugeordnet. Um mehrere Instanzen einer Komponente zu verteilen, verbinden sich die Instanzen mit der gleichen Service-Eingangswarteschlange. Die Instanzen erhalten aus dieser Warteschlange Arbeitseinheiten. Nach der Verarbeitung wird das Ergebnis an eine Ausgabewarteschlange zurückgegeben. Die Ausgabewarteschlange eines Dienstes kann an eine Eingabewarteschlange eines anderen Dienstes angeschlossen werden, um eine Pipeline zu erstellen. Aufgrund dieser Ein- und Ausgabewarteschlangen-Systematik kann jeder Service Arbeitseinheiten asynchron bearbeiten. Durch diese Architektur ist TextImager eine Multi-Server-, Multi-Service- und Multi-Service-Instanz-Architektur.

Darüber hinaus bietet TextImager ein Toolkit, das es jedem Entwickler ermöglicht, einen eigenen TextImager-Cluster aufzusetzen und Services im TextImager-System hinzuzufügen. Entwickler können den Zugriff auf die Dienste einschränken, wenn dies wie oben beschrieben erforderlich ist, was mittels die Integration des ResourceManagers (Gleim 2012) und des AuthorityManagers (Gleim 2012) realisiert wird.

Durch Freigabe des Quellcodes des TextImager und die Bereitstellung von Leitlinien für dessen Erweiterung wollen wir es Dritten ermöglichen, ihre NLP-Software über die Webservices von TextImager zu vertreiben, so dass die gesamte wissenschaftliche Gemeinschaft davon profitiert.

Installationsanweisungen und Beispiele für die Einrichtung eines TextImager-Servers finden Nutzer in folgendem GitHub-Repository: https://github.com/texttechnologylab/textimager-server.

Der Beitrag erörtert die Möglichkeiten und Grenzen des NLP von Big Data, stellt den TextImager als Werkzeug für diesen Bereich zur Diskussion und zeigt anhand von drei Nutzungsszenarien Einsatzmöglichkeiten in den DH auf.

Fußnoten

1. https://github.com/topics/nlp
2. https://uima.apache.org/doc-uimaas-what.html
3. https://uima.apache.org/doc-uimaducc-whatitam.html

Bibliographie

de Castilho, Richard Eckart / Gurevych, Iryna (2014): *A broad-coverage collection of portable NLP components for building shareable analysis pipelines*, in: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT 1–11.

Ferrucci, David / Lally, Adam (2004): *UIMA: an architectural approach to unstructured information processing in the corporate research environment.*, in: Natural Language Engineering 10(3-4) 327–348.

Gleim, Rüdiger / Mehler, Alexander / Ernst, Alexandra (2012): *SOA implementation of the ehumanities desktop*, in: Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts.

Mehler, Alexander / Hemati, Wahed / Gleim, Rüdiger / Baumartz, Daniel (2018): *VienNA: Auf dem Weg zu einer Infrastruktur für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprache*, in Forschungsinfrastrukturen und digitale Informationssysteme in der germanistischen Sprachwissenschaft , H. Lobin, R. Schneider, and A. Witt, Eds., Berlin: De Gruyter, 2018, vol. 6.

Hemati, Wahed / Uslu, Tolga / Mehler, Alexander (2016): *Textimager: a distributed uima-based system for nlp*, in:

Proceedings of the COLING 2016 System Demonstrations. FederatedConference on Computer Science and Information Systems.

Manning, Christopher / Surdenu, Mihai / Bauer, John / Finkel, Jenny / Bethard, Steven / McClosky, David (2014): The stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations 55–60.

OpenNLP (2010): *Apache OpenNLP*, in: http://opennlp.apache.org [letzter Zugriff 05. Oktober 2018]

Die PARTHENOS Training Suite

Wuttke, Ulrike
wuttke@fh-potsdam.de
FH Potsdam, Deutschland

Neuroth, Heike
neuroth@fh-potsdam.de
FH Potsdam, Deutschland

Digitale Forschungsinfrastrukturen spielen eine zunehmende Rolle im geistes- und kulturwissenschaftlichen Bereich (z.B. ESF 2011, Benardou, Champion, Dallas & Hughes (eds.) 2017). Damit ihre Potentiale ausgeschöpft und die Früchte der ressourcenintensiven Entwicklungsforschungsarbeit geerntet werden können, müssen sich diejenigen, die mit ihnen forschen, und diejenigen, die an ihrer Entwicklung und ihrem Ausbau beteiligt sind, zusätzliches theoretisches Wissen und praktische Fähigkeiten aneignen.

Als das EU-Projekt PARTHENOS[1], das die wichtigsten europäischen eHumanities und eHeritage Infrastrukturen umfasst, startete, boten geistes- und kulturwissenschaftliche Infrastrukturen bereits Trainings- und Ausbildungsangebote an und trugen zur Entwicklung, Ausbau oder Unterhalt von Plattformen für Online-Trainingsmaterialien bzw. -Kursübersichten bei (z.B. #dariahTeach[2], CLARIN Videolectures[3]). Viele dieser Angebote sind jedoch projektspezifisch bzw. thematisch und methodisch spezialisiert und setzen Vorwissen über die Rolle und Funktion von Forschungsinfrastrukturen voraus, das potentiellen neuen Nutzer*innen meist fehlt. Ausreichendes Vorwissen über infrastrukturelle Aspekte ist jedoch eine Grundvoraussetzung für die erfolgreiche Auseinandersetzung mit thematisch und methodisch spezialisierten Themengebieten der Digital Humanities. PARTHENOS adressiert diesen Bedarf einerseits durch die Entwicklung eines disziplinübergreifenden inhaltlichen Curriculums zu infrastrukturell breit angelegten Themengebieten und andererseits zielgruppengeeigneter Vermittlungswege, die neue technologische Vermittlungsmöglichkeiten nutzen.

Die Entwicklung der Trainingsangebote erfolgt durch das PARTHENOS Training Team (Leitung Trinity College Dublin) gemeinsam mit den an PARTHENOS beteiligten Partnern auf der Basis der Analyse der Nutzeranforderungen und bestehender Trainingskonzepte im Bereich eHumanities und eHeritage Infrastrukturen in der ersten Projektphase (Drude et al. 2016, Oltersdorf et al. 2016). Ausgehend von diesen Erkenntnissen entwickelt PARTHENOS auf der Grundlage des gemeinsamen Wissensschatzes der Projektpartner grundlegende Trainings- und Lehrmaterialien. Die disziplinübergreifend breit angelegten Materialien sollen Interesse an infrastrukturell getriebener Forschungs- und Entwicklungsarbeit wecken, ein Grundverständnis für die inhärenten Potentiale und Herausforderungen schaffen sowie eine Grundlage für weiterführende Auseinandersetzungen bilden. Schwerpunkte sind dabei die Wissensvermittlung über die Rollen, Funktionen und Potentiale geistes- und kulturwissenschaftlicher digitaler Infrastrukturen für Wissenschaftler*innen, Praktiker*innen, Entwickler*innen, Mitarbeiter*innen in Rechenzentren und Entscheidungsträger*innen und die didaktischen Ziele Bewusstseinsbildung (*awareness raising*) und Erweiterung von Fertigkeiten (*skills building*) (Edmond, Garnett 2017).

Vor diesem Hintergrund entwickelt PARTHENOS Online-Materialien für das eigenständige Selbststudium und die Verwendung durch Lehrpersonen (insbesondere durch die Bereitstellung ergänzender Übungsmaterialien und eines exemplarischen Workshop-Curriculums) und bietet (virtuelle) Veranstaltungstypen an (Edmond et al. 2016, Spiecker et al. 2017). Die Materialien werden auf der PARTHENOS Training Suite[4] bereitgestellt, einer WordPress-basierten eLearning-Plattform, die aus einer Reihe von Modulen besteht (z.B. „Introduction to Research Infrastructures“, „Management Challenges in Research Infrastructures“, „Open Up your Research and Data“). Lehrende und Lernende werden zur besseren Orientierung linear durch die Inhalte der Module geführt. Sie können jedoch zu jedem Zeitpunkt über die Modulnavigation gezielt einzelne Inhalte ansteuern und so ihren Lernprozess und den verschiedene Sinne ansprechenden Modus der Vermittlung und Rezeption individuell steuern. Vereinzelt können sie auch zwischen unterschiedlichen Formen der Vermittlung derselben Inhalte wählen (z.B. Text oder Video).

Im Gegensatz zu printbasierten Lehrmaterialien nutzt die PARTHENOS Training Suite zeitgemäße technologische Möglichkeiten und erlaubt die Einbindung multimedialer Inhalte und multimodaler Formen der Aneignung und Vermittlung durch Lernende und Lehrende. Innerhalb der Module reicht die Spannbreite der bereitgestellten Ressourcen von Video-Vorträgen, Interviews, Erklärfilmen, über Präsentationsmaterialien, Übungen und Erläuterungen zu Grundprinzipien, bis zu Literatur- und Linksammlungen. Des Weiteren hat PARTHENOS erfolgreich die fünfteilige “PARTHENOS eHumanities and eHeritage Webinar Series” durchgeführt (Drenth & Wuttke 2018, Wuttke 2019). Alle Webinarmaterialien (wie Präsentationsfolien als PDF und Powerpoint und Videoaufnahmen der Webinare) stehen zur Nachnutzung zur Verfügung.[5]

Mit der Bereitstellung der Materialien als Open Educational Resources (OER), greift PARTHENOS einen Grundwert der auf einer Kultur des Teilens und der Nachnutzung basierenden Digital Humanities auf. Ebenso werden bei der Entwicklung die Möglichkeiten des partizipativen digitalen Erschaffens ausgelotet („collaboration as creation“, Burdick et al. 2016: 84). Um neue Entwicklungen aufzugreifen, findet ein permanenter Austausch mit den Projektpartnern über die Weiterentwicklung der Trainingsangebote statt. So wird einerseits die Erfüllung der Bedarfe der Fachcommunities sichergestellt (Bottom-Up-Approach) und andererseits die Innovationskraft sowie die Qualität der Produkte erhöht.

Zusätzlich werden Kooperationen mit externen Partnern (wie z.B. Europeana[6] oder Foster[7]) eingegangen, sowohl bezüglich der Einbindung externer Materialien, aber auch der Nachnutzung der PARTHENOS-Materialien.

Unser Poster vermittelt die Grundprinzipien und wichtigsten Erkenntnisse bei der Entwicklung der PARTHENOS Trainingsangebote, die so multimedial und multimodal sind, wie die Digital Humanities selbst. Es stellt die PARTHENOS Trainingsangebote mit Hilfe von Text und Graphiken vor, um ihre Bekanntheit und Nachnutzung durch die deutschsprachige Digital Humanities Community zu fördern.

Danksagung: Die im Poster vorgestellten Ergebnisse beruhen auf der gemeinsamen Arbeit von PARTHENOS Arbeitspaket 7 „Skills, Professional Development and Advancement“ und den PARTHENOS Projektpartnern.

PARTHENOS has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 654119.

Fußnoten

1. PARTHENOS steht für „Pooling Activities, Resources and Tools for Heritage e-Research, Optimization and Synergies“, Projektlaufzeit 2015-2019, Projektwebseite: http://www.parthenos-project.eu/. Siehe auch Wuttke, Neuroth, Spiecker 2019.
2. Projekt-Webseite: https://teach-blog.dariah.eu/.
3. Projekt-Webseite: http://videolectures.net/clarin/.
4. Die Einstiegsseite der PARTHENOS Training Suite ist zu erreichen über: http://training.parthenos-project.eu/
5. Mehr Informationen zur PARTHENSO-Webinarserie sowie Links zu den einzelnen Webinaren und die Webinarmaterialen sind zu finden über: http://training.parthenos-project.eu/sample-page/ehumanities-eheritage-webinar-series/.
6. Projekt-Webseite: https://www.europeana.eu/portal/.
7. Projekt-Webseite: https://www.fosteropenscience.eu/.

Bibliographie

Burdick, Anne / Drucker, Johanna / Lunenfeld, Peter / Presner, Todd / Schnapp, Jeffrey (eds.) (2016): *Digital Humanities*. Cambridge, MA: MIT Press.

Benardou, Agiatis / Champion, Erik / Dallas, Costis / Hughes, Lorna (eds.) (2017): *Cultural Heritage Infrastructures in Digital Humanities*. Digital Research in the Arts and Humanities. London, New York: Routledge.

ESF (2011): *Research Infrastructures in the Digital Humanities*. Science Policy Briefing, 42. European Science Foundation.

Drenth, Petra / Wuttke, Ulrike (2018): *“Successful PARTHENOS e-Humanities and eHeritage Series concluded”*. PARTHENOS News Beitrag, 28.05.2018, http://www.parthenos-project.eu/succesfull-parthenos-ehumanities-and-eheritage-webinar-series-concluded [letzter Zugriff 03. Januar 2019].

Drude, Sebastian / Di Giorgio, Sara / Ronzino, Paola / Links, Petra / Van Nispen, Annelies / Verbrugge, Karolien / Degl'Innocenti, Emiliano / Stiller, Juliane / Oltersdorf, Jenny / Spiecker, Claus (2016): *Report on user requirements, PARTHENOS deliverable D2.1.* Veröffentlicht am 20.10.2016. http://www.parthenos-project.eu/Download/Deliverables/D2.1_User-requirements-report-v2.pdf [letzter Zugriff 03. Januar 2019].

Edmond, Jennifer / Garnett, Vicky (2017): *„Soft Skills in hard places: the changing face of DH training in European research infrastructures. Pre-print as presented at the DH Benelux Conference, 2017“*. http://hdl.handle.net/2262/85444 [letzter Zugriff 03. Januar 2019].

Edmond, Jennifer / Garnett, Vicky / Burr, Elisabeth / Laepke, Stefanie / Oltersdorf, Jenny / Goulis, Helen (2016): *Initial Training Plan*. Veröffentlicht am 07.06.2016. http://www.parthenos-project.eu/Download/Deliverables/D7.1_Initial_Training_Plan.pdf [letzter Zugriff 03. Januar 2019].

Oltersdorf, Jenny / Edmond, Jennifer / Garnett, Vicky / Henriksen, Lina / Mergoupi, Eirini-Savaidou / Povlsen, Claus (2016): *Report on the assessment of the education and training plans and activities. PARTHENOS deliverable D2.2.* Veröffentlicht am 20.20.2016. http://www.parthenos-project.eu/Download/Deliverables/D2.2_Report_Assessment_Education_Training.pdf[letzter Zugriff 03. Januar 2019].

Spiecker, Claus / Oltersdorf, Jenny / Wuttke, Ulrike / Edmond, Jennifer / Garnett, Vicky / Läpke, Stefanie (2017): *Report on training and education activities and updated planning*. Veröffentlicht am 22.04.2017. Verfügbar unter: http://www.parthenos-project.eu/Download/Deliverables/D7.2_Training_Plan_FINAL.pdf [letzter Zugriff 03. Januar 2019].

Wuttke, Ulrike (2019): *„The “PARTHENOS Training Webinar Series”: Webinars as a means of delivering successful research infrastructure training in eHumanities and eHeritage“*, in Liber Quarterly (im Erscheinen).

Wuttke, Ulrike / Spiecker, Claus / Neuroth, Heike (2019): *„PARTHENOS – Eine digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften“*, in: Bibliothek Forschung und Praxis (im Erscheinen).

Digitales Publizieren im Spiegel der Zeitschrift für digitale Geisteswissenschaften - ZfdG: Eine Standortbestimmung

Fricke-Steyer, Henrike
henrike.fricke@hab.de
Forschungsverbund MWW, Deutschland

Klaffki, Lisa
klaffki@hab.de
Forschungsverbund MWW, Deutschland

Die Zeitschrift für digitale Geisteswissenschaften - ZfdG[1] ist ein Open Access-Forschungsperiodikum, das sich Themen an der Schnittstelle von geisteswissenschaftlicher und digitaler Forschung widmet. Adaptionen von Informatik und Informationswissenschaft eröffnen der Gesamtheit der Geisteswissenschaften neue Wege der Wissenserschließung, tragen zur Etablierung neuer Forschungsansätze bei und liefern neue Möglichkeiten der Auf- und Nachbereitung von Quellen, Dokumenten, Daten und Medien. Die Verknüpfung von technischen Innovationen und geisteswissenschaftlichen Forschungsfragen bildet die Grundlage zu einer Standortbestimmung der digitalen Geisteswissenschaften.

Mit der Zeitschrift für digitale Geisteswissenschaften – ZfdG bietet der Forschungsverbund Marbach Weimar Wolfenbüttel (MWW) in Zusammenarbeit mit dem Verband Digital Humanities im deutschsprachigen Raum (DHd) seit 2015 ein Forum zur Präsentation und Diskussion von Forschungsergebnissen im Kontext der Digital Humanities. Die Geisteswissenschaften richten ihr Augenmerk zunehmend auf Fragestellungen, die digitale Möglichkeiten in ihre Überlegungen einbeziehen oder diese vermehrt zum Ausgangspunkt ihrer Forschungen und Projekte machen. Auch lassen sich alte Fragestellungen mit Hilfe digitaler Methoden neu bearbeiten, überprüfen oder auf wesentlich größere Korpora beziehen. Von der Digitalisierung der Primärquellen bis zur Änderung der Publikationskultur und Fachkommunikation unter digitalen Bedingungen reichen die Möglichkeiten, auf denen solche Fragestellungen basieren oder von denen sie ausgehen können. Die Zeitschrift für digitale Geisteswissenschaften – ZfdG versteht sich als Organ, das all diese Entwicklungen disziplinübergreifend begleitet und auch die philosophischen, politischen, sozialen und kulturellen Implikationen und Konsequenzen beleuchtet, die der digitale Wandel mit sich bringt. Sie setzt sich für eine Geisteswissenschaft im digitalen Zeitalter ein, die die entscheidenden Fragen und Themen auf dem Weg zu digitalen Geisteswissenschaften verhandelt und auch kritischen Einwänden in diesem Feld Raum für Debatten bietet. Durch ein klares Bekenntnis zu Open Access sind die Beiträge für alle zugänglich, durch die Verfügbarkeit der Beiträge als XML stehen auch sie als potentielles Quellenmaterial für weitere Forschungen zur Verfügung.

Da digitale Veröffentlichungsformen zunehmend als vollwertige wissenschaftliche Publikation an Bedeutung gewinnen und neben die traditionellen Publikationsformate gedruckter Monographien oder Zeitschriftenartikel treten (Kohle 2017: 199), liegt es nahe, die digitale Publikationsform selbst zum Gegenstand des Erkenntnisinteresses zu machen. Das Poster möchte daher im Spiegel der bisher veröffentlichten Artikel der Zeitschrift für digitale Geisteswissenschaften – ZfdG auf die Landschaft des Digitalen Publizierens blicken und dabei folgende Aspekte thematisieren:

- Autorschaften (kollaborativ (Heller und Bartling 201 4) oder einzeln)
- Einbindung multimedialer Inhalte
- Publikation / Verbindung mit Forschungsdaten
- Fachliche Zuordnung(en) der Artikel
- Gewähltes Peer Review-Verfahren (Post- (Amsen 2014) oder Prepublication)
- Metriken (Zugriffszahlen und Downloads)
- Wissenschaftliche Rezeption der Artikel

Dazu sollen die bis zur Konferenz erschienenen Beiträge bzw. deren Metadaten auf die genannten Aspekte hin quantitativ ausgewertet werden und die Ergebnisse dem Konferenzmedium Poster angemessen visualisiert werden, etwa durch den Einsatz von Diagrammen und Wortwolken. Diese Bestandsaufnahme versteht sich auch als Beitrag zur Diskussion um die Entwicklung des Bereichs des Digitalen Publizierens (DHd-Arbeitsgruppe 2016). Denn das Potential digitaler Veröffentlichungen liegt gerade auch in der Interaktionsfähigkeit dieser mit anderen medialen Formen, die Einbindung multimedialer Inhalte (Maciocci 2017) bis hin zu sogenannten Enhanced Publications (Degwitz 2015), dem Vernetzen mit anderen Online-Ressourcen durch Linked Open Data oder der parallelen Publikation von Artikel und Forschungsdaten bzw. sogenannten Data Papers. Deshalb soll hier exemplarisch geprüft werden, inwieweit diese Möglichkeiten, die technisch umsetzbar sind, von den AutorInnen auch bei der Konzeption ihrer Artikel genutzt werden.

Fußnoten

1. http://www.zfdg.de/

Bibliographie

Amsen, Eva (2014): *What is post-publication peer review?*, in: Blogpost auf F1000 Research Blog. https://blog.f1000.com/2014/07/08/what-is-post-publication-peer-review/ [letzter Zugriff: 14.10.2018].

Degkwitz, Andreas (2015): *Enhanced Publications Exploit the Potential of Digital Media*, in: Evolving Genres of ETDs for Knowledge Discovery. Proceedings of ETD 2015 18th International Symposium on Electronic Theses and Dissertations 51–59.

DHd-Arbeitsgruppe (2016): *Digitales Publizieren*, in: **DHd-Arbeitsgruppe (eds.):** Working Paper *Digitales Publizieren* http://diglib.hab.de/ejournals/ed000008/startx.htm [letzter Zugriff: 14.10.2018].

Heller, Lambert / The, Ronald / Bartling, Sönke (2014): *Dynamic Publication Formats and Collaborative Authoring*, in: **Bartling S., Friesike S. (eds):** *Opening Science.* Cham: Springer 191–211. DOI: 10.1007/978-3-319-00026-8.

Kohle, Hubertus (2017): *Digitales Publizieren*, in: **Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.):** *Digital Humanities. Eine Einführung.* Stuttgart: Metzler Verlag 199–205.

Maciocci, Giuliano (2017): *Designing Progressive Enhancement Into The Academic Manuscript: Considering a design strategy to accommodate interactive research articles*, in: Blogpost auf eLife Sciences. https://elifesciences.org/labs/e5737fd5/designing-progressive-enhancement-into-the-academic-manuscript [letzter Zugriff: 14.10.2018].

Eine Basis-Architektur für den Zugriff auf multimodale Korpora gesprochener Sprache

Batinic, Josip
josip.batinic@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Frick, Elena
frick@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Gasch, Joachim
gasch@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Das Projekt ZuMult – „Zugänge zu multimodalen Korpora gesprochener Sprache – Vernetzung und zielgruppenspezifische Ausdifferenzierung“ (zumult.org) – hat sich zum Ziel gesetzt, eine Architektur zu entwickeln, die einen einheitlichen Zugriff auf verschiedene Korpora gesprochener Sprache (Audio- und Videoaufzeichnungen mündlicher Interaktion mit zugehörigen Metadaten, Transkripten, Annotationen) an verschiedenen Standorten ermöglicht, und auf deren Basis Zugangswege gestaltet werden können, die für die Bedarfe spezifischer Nutzergruppen (z.B. Sprachlehrforschung, Variationslinguistik) optimiert sind. Mit unserem Poster stellen wir das technische Konzept und eine prototypische Implementierung einer solchen Basisarchitektur vor.

Ausgehend von einer vergleichenden Analyse vorhandener Plattformen (u.a. Datenbank für Gesprochenes Deutsch, Schmidt 2016; GeWiss-Korpus-Interface, Fandrych, Meißner & Wallner 2017; Repositorium des Hamburger Zentrums für Sprachkorpora, Hedeland et al. 2014; sowie mehrere Lösungen, die außerhalb des deutschsprachigen Raums entwickelt wurden, z.B. Eshkol-Taravella et al. 2012, Komrsková et al. 2018) und einer Bestandsaufnahme existierender Standards im Bereich multimedialer Daten (vgl. dazu auch Schmidt 2014 und Schmidt et al. 2010) haben wir eine Dreiebenen-Lösung entwickelt, die so weit wie möglich auf etablierte (De Facto-)Standards aufbaut und anschlussfähig an existierende Lösungen ist. Damit wird eine transferfähige Basis für einen flexiblen Zugriff auf multimodale Korpora geschaffen.

Kern der Architektur ist zum einen eine objektorientierte Modellierung der Korpus-Bestandteile (Aufnahmen, Metadaten zu Sprechereignissen und Sprechern, Transkripte, Annotationen und Zusatzmaterialien) und ihrer Beziehungen zueinander. Für deren digitale Repräsentation (Serialisierung) werden Standards verwendet, soweit sie existieren. Für Medienobjekte können wir auf industrielle Standards insbesondere aus dem Kontext der Moving 133_final-* Expert Group (MPEG) zurückgreifen. Die Repräsentation von Transkripten und Annotation folgt dem in ISO (2016) definierten und auf den Richtlinien der Text Encoding Initiative (TEI) basierenden Format für „Transcriptions of Spoken Language“. Metadaten werden grundsätzlich in XML repräsentiert; in Ermangelung eines echten Standards, der in der Lage wäre, der Bandbreite und Komplexität von Metadaten im Bereich multimodaler Korpora vollständig gerecht zu werden, orientieren wir uns in diesem Bereich an CMDI-Profilen, die im CLARIN-Kontext für solche Korpora entwickelt wurden (z.B. Hedeland & Wörner 2012).

Zum anderen beinhaltet die Architektur ein vereinheitlichtes Konzept zur Query auf Transkriptions- und Annotationsdaten. Dieses baut auf Überlegungen zu einer „Corpus Query Lingua Franca“ (Banski et al. 2016, ISO 2018) auf und berücksichtigt somit in der Korpuslinguistik verbreitete Suchsprachen wie CQP, ANNIS-QL, Poliqarp und weitere, die allerdings für die Besonderheiten angepasst werden müssen, die spontansprachliche Daten gegenüber schriftsprachlichen Korpora aufweisen.

Die Basisarchitektur besteht somit aus zwei gleichberechtigten Komponenten: Aus der Modellierung der Korpus-Bestandteile ergeben sich Zugriffs- und Navigationsmöglichkeiten für ganze Objekte bzw. Objekthierarchien, die auf Nutzerseite vor allem für ein exploratives Browsing auf den Daten eingesetzt werden. Die Query-Komponente ermöglicht hingegen eine gezielte Auswahl von (Teilen) von Objekten und damit systematische Recherchen im Sinne einer korpuslinguistischen Methodik. Beide Komponenten werden technisch als „Locators“ bzw. „Filters“ in einer REST API umgesetzt. Diese wird in der weiteren Projektarbeit die Basis darstellen, um zielgruppenspezifisch optimierte Zugänge zu den Daten zu entwickeln.

Neben einem Überblick über diese Basis-Architektur wird unser Poster auch auf die konkrete Implementierung eingehen, die am Institut für Deutsche Sprache für den Zugriff auf die Daten aus dem Archiv für Gesprochenes Deutsch entwickelt wurde. Diese setzt auf ein vorhandenes Backend auf, das die Grundlage für die Datenbank für Gesprochenes Deutsch bildet und XML-basierte Daten in einer objektrelationalen Oracle-Datenbank hält. Für die Arbeiten in ZuMult wird dieses Backend für die im Projekt definierten Bedarfe angepasst und erweitert. Prototypische Applikationen, die den Einsatz der REST API illustrieren, werden als Software-Demonstrationen die Posterpräsentation ergänzen.

Bibliographie

Banski, Piotr / Frick, Elena / Witt, Andreas (2016): *"Corpus Query Lingua Franca (CQLF)".* Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia 2804-2809. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-50405

Eshkol-Taravella, I. / Baude, O. / Maurel, D. / Hriba, L. / Dugua, C. / Tellier, I., (2012): *"Un grand corpus oral ‚disponible' : le corpus d'Orléans 1968-2012. "* In: Ressources linguistiques libres, TAL. 52,3/2011, 17-46.

Fandrych, Christian / Meißner, Cordula / Wallner, Franziska (eds.) (2017): *"Gesprochene Wissenschaftssprache – digital Verfahren zur Annotation und Analyse müdlicher Korpora."* Deutsch als Fremd- und Zweitsprache. Tübingen: Stauffenburg.

Hedeland, Hanna / Wörner, Kai (2012): *"Experiences and Problems creating a CMDI profile from an existing Metadata Schema".* Proceedings of LREC-Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, Istanbul, European Language Resources Association (ELRA) 37-40. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): *"Multilingual Corpora at the Hamburg Centre for Language Corpora". In: Ruhi, Şükriye/Haugh, Michael/Schmidt, Thomas/Wörner, Kai (Hrsg.):* Best Practices for Spoken Corpora in Linguistic Research. Newcastle: Cambridge Scholars Publishing, 2014. S. 208-224.

ISO (ed.) (2016): *ISO 24624:2016 Language resource management – Transcription of spoken language.* https://www.iso.org/standard/37338.html

ISO (ed.) (2018): *ISO 24623-1:2018 Language resource management – Corpus query lingua franca (CQLF) -- Part 1: Metamodel.* https://www.iso.org/standard/37337.html

Komrsková, Zuzana / Kopřivová, Marie / Lukeš, David / Poukarová, Petra / Goláňová, Hana (2018): *"New Spoken Corpora of Czech: ORTOFON and DIALEKT."* Journal of Linguistics 68:2, 219-228.

Schmidt, Thomas (2014): *"(More) Common Ground for Processing Spoken Language Corpora?"* In: **Ruhi, Şükriye/Haugh, Michael/Schmidt, Thomas/Wörner, Kai (eds.)**: *Best Practices for Spoken Corpora in Linguistic Research.* Newcastle: Cambridge Scholars Publishing, 2014 249-265. http://pub.ids-mannheim.de/autoren/divers/3119.html

Schmidt, Thomas (2017): *"DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim."* In: Zeitschrift für Germanistische Linguistik 45(3), S. 451-463.

Schmidt, Thomas / Elenius, Kjell / Trilsbeek, Paul (2010): *"Multimedia encoding and annotation"*. In: **Hinrichs, Erhard (ed.)**: *Interoperability and standards*. Utrecht: Utrecht University, 2010 121-124. http://www.exmaralda.org/files/CLARIN_Standards.pdf

Ein Editionsportal (nicht nur) für Thüringen

Prell, Martin

martin.prell@uni-jena.de
Friedrich-Schiller-Universität Jena, Deutschland

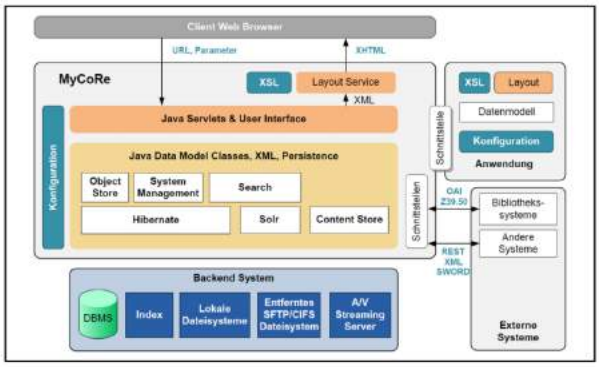
Projektübersicht

Ziel des an der Universität Jena durchgeführten Projekts ist die Entwicklung eines uneingeschränkt zugänglichen Onlineportals zur Publikation digitaler historisch-kritischer Editionen von vorrangig handschriftlichen Quellen der Neuzeit. Es wird an der Thüringer Universitäts- und Landesbibliothek (ThULB) gehostet und in Zusammenarbeit mit zahlreichen Partnern (darunter READ/Transkribus, Deutsches Textarchiv, Herzog August Bibliothek Wolfenbüttel, Landesarchiv Thüringen, Sammlungs- und Forschungsverbund Gotha und viele weitere) realisiert. Mit TEI-basierten Meta- und Volltextdaten, Digitalisaten, Registern, Paratexten, Visualisierungen, gezielten Such-, Sortierungs- und Filteroptionen wird das Editionsportal Thüringen eine wissenschaftlich hochwertige digitale Publikations- und Rechercheumgebung anbieten. Das Portal spricht in erster Linie Projekte an, die ihre Edition dauerhaft verfügbar und intuitiv zugänglich machen wollen, aber kein eigenes Onlineportal entwickeln und nachhaltig betreiben können oder möchten. Damit diese Projekte den Heraus- und Anforderungen exzellenter digitaler Editionen [s. bspw. Sahle 2014] entsprechen, legt das Portal Schwerpunkte auf die nachfolgend erläuterten Aspekte.

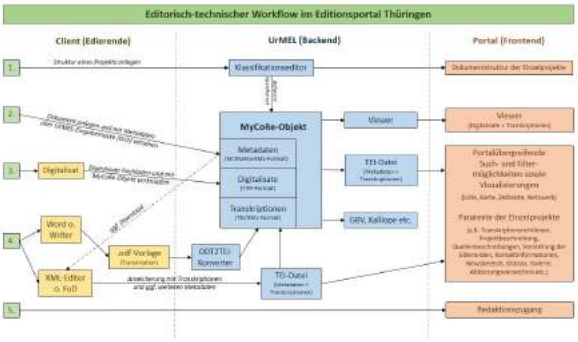
Fokus Nachnutzung und Langzeitverfügbarkeit

Um die im Portal befindlichen Editionen nachnutzen und in andere digitale Wissensbasen einbinden zu können, liegt ein Fokus auf der Bereitstellung von Datenschnittstellen, der Verwendung offener und standardisierter Datenformate und der Nachnutzung bewährter Open-Source-Software. So basiert nicht allein das technische Backend-System MyCoRe auf entsprechenden Technologien (Abb. 1). Auch das eigens für das Editionsportal entwickelte TEI-Basisformat (ThULBBf) ist eng an das TEI-Subset des Deutschen Textarchivs (DTABf) angelehnt und erweitert dieses handschriftenspezifisch (bspw. in den Bereichen Normdatenanreicherung, Materialität, Textzeugen-Wiedergabe, Transkriptions- und Auszeichnungsrichtlinien, auto-generierte Register etc., Abb. 2.), um die Ausspielung der Portaldaten in das DTA zu ermöglichen. Damit geht sowohl die Speicherung und Verwendung der Forschungsdaten in der CLARIN-D-Infrastruktur als auch die Nachnutzung etablierter linguistischer DTA-Tools einher, ohne dass (ggf. redundante) Neuentwicklungen erforderlich werden.

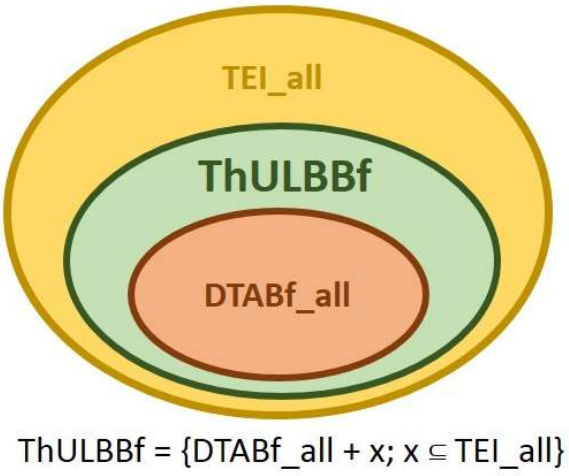
Damit ist bereits ein weiteres zentrales Portalmerkmal, die Langzeitverfügbarkeit der darin dargebotenen Informationen, benannt. Wenngleich die Frage der Langzeitspeicherung noch ungelöst ist [Carusi / Reimer 2010: 42], so folgt das Portal doch derzeitigen Nachhaltigkeitsempfehlungen[1]. Dazu gehört unter anderem, dass die nachhaltige Speicherung der Portaldaten im Rahmen der von EFRE geförderten Langzeitarchivierungsplattform des Landes Thüringen auf Grundlage des Digitalen Archives NRW (Kooperation mit dem Landschaftsverband Rheinland) erfolgt [Mutschler 2017: 316].



MyCoRe-Architektur (http://www.mycore.de/features/index.html)



Der verbindliche Workflow für Editionsprojekte des Portals



Das TEI-Basisformat des Editionsportals (ThULBBf)

Fokus multimodale editorische und fachwissenschaftliche Benutzung

Das Portal verfolgt einen generischen Ansatz mit fest definiertem Arbeitsworkflow (Abb. 3), um Editionen projektübergreifend multimodal recherchier-, visualisier- und erforschbar zu machen. Zugleich wird jede Edition des Portals als eigenständige Forschungsleistung in Form einer projektspezifischen Publikation sichtbar und zitierbar. Editoren können zudem zwischen verschiedenen Werkzeugen zur Editionserstellung wählen. Je nach Ziel und Kenntnisstand können Microsoft Word, die Trierer Editionsforschungsumgebung FuD oder ein beliebiger XML-Editor eingesetzt werden. Für NutzerInnen bietet das Portal multimodale Recherche- und Visualisierungsinstrumente für disziplinübergreifende Fragestellungen (Netzwerke, Zeitleisten, Diagramme, Karten etc.). Durch die Einbettung des Editionsportals in ein umfassenderes Cultural-Heritage-Internetportal[2] können die Quellenbestände dutzender Thüringer Institutionen und Partner außerhalb Thüringens in die Recherche und Analyse unmittelbar einbezogen werden.

Verortung innerhalb der editorischen Portallandschaft

Innerhalb der editorischen Portallandschaft verortet sich das Editionsportal Thüringen zwischen Einzeleditionsprojekte weitestgehend nivellierenden Textsammlungen wie bspw. dem Deutschen Textarchiv oder dem TextGrid-Repository einerseits und Einzeleditionsportalen mit zu projektspezifisch ausgerichteten Eigenschaften und Funktionalitäten ohne editionsübergreifende Such- und Analysemöglichkeiten andererseits. Zu letzteren sind bspw. die Editionen der Wolfenbütteler Digitalen Bibliothek oder des Geisteswissenschaftlichen Asset Management Systems (GAMS) zu rechnen. Auch die Repositorien (trans)nationaler Infrastrukturen wie bspw. DARIAH oder CLARIN bleiben aufgrund ihrer wiederum zu unspezifischen Forschungsdatenausrichtung hinter den anwendungsorientierten Datenmodellen und Funktionalitäten des Editionsportals Thüringen zurück. Zum Novum des Editionsportals gehört ferner seine betont breite Ausrichtung auf möglichst zahlreiche Handschriftengattungen, unterschiedliche Zielgruppen, disparate Editionswerkzeuge und eine priorisierte Softwarenachnutzung.

Das Portal wird in einem ersten Schritt die bereits existierenden und in Arbeit befindlichen Editionen der digitalen ThULB-Bibliothek UrMEL aufnehmen und an diesen evaluiert. Es ist aber als Plattform konzipiert, die zur digitalen Aufbereitung von Quellenbeständen verschiedenster, auch kleinerer Einrichtungen (Archive, Bibliotheken, Museen, Vereine) anregen soll und keinesfalls an Thüringer Quellen, Institutionen o.ä. gebunden ist.

Zielgruppenorientierung

Das Portal versteht sich in erster Linie als wissenschaftliches Angebot, das schwer zugängliche Handschriften aus fünf Jahrhunderten verfügbar macht und hochwertige Volltexte für die computergestützte Weiterverarbeitung liefert. In editionswissenschaftlicher Hinsicht leistet es einen Forschungsbeitrag zu der Frage, wie ein generisches Editionsportal sowohl projektspezifische, aus der Heterogenität der Quellen und den Erwartungen der

Edierenden, Rezipienten und Förderer resultierende, als auch gesamtportalische, editionsübergreifende Interessen und Ansprüche gleichermaßen verbinden kann [s. bspw. Dogunke 2017]. Das Editionsportal richtet sich aber nicht ausschließlich an die Wissenschaft, sondern auch an Schulen und Bildungseinrichtungen und die interessierte Öffentlichkeit. Diesen wird mit der Integration des e-Learning-Tools „TranskribusLearn“ ein Werkzeug zum Erlernen altdeutscher Schrift angeboten sowie die Möglichkeit gegeben, das Portal mit eigenen Transkriptionen anzureichern. Eine hohe Qualität der Editionsinhalte und -daten wird durch Empfehlungen und Anleitungen, Reviewing, verbindliche Workflows (Abb. 3) und Dateneingaben sowie verschiedene technische Validierungsinstanzen gewährleistet.

Projektgenese und Arbeitsstand

Das Projekt ist aus einem Editionsprojekt [Prell / Schmidt-Funke 2017] hervorgegangen, für das nur sehr geringe Ressourcen zur Verfügung standen. Da dies eine häufig zu beobachtende Rahmenbedingung von Editionen darstellt, werden die im Editionsprojekt entwickelten Lösungen ausgebaut, institutionalisiert und anderen ForscherInnen kostenfrei zugänglich gemacht. Aktuell befindet sich das Portal in der zweiten Förderphase, nachdem in den Jahren 2017 und 2018 konzeptuelle und anpassende Maßnahmen des Backend-Systems vorgenommen worden. Derzeit findet die Entwicklung des TYPO3-Frontends statt.

Fußnoten

1. Vgl. beispielsweise die durch Sustainability-Zertifikate wie das Data Seal of Approval bzw. CoreTrustSeal formulierten Kriterien für nachhaltige Datenrepositorien (https://www.coretrustseal.org/ , https://www.datasealofapproval.org (letzter Zugriff 27.09.2018) sowie Buddenbohm et al. 2014.
2. Das „Digitale Kultur- und Wissensportal Thüringen“ wird im 1. Quartal 2019 online geschaltet.

Bibliographie

Buddenbohm, Stefan / Enke, Harry / Hofmann, Matthias / Klar, Jochen / Neuroth, Heike / Schwiegelshohn, Uwe (2014): *„Erfolgskriterien für den Aufbau und nachhaltigen Betrieb Virtueller Forschungsumgebungen“*. Göttingen. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2014-7.pdf [letzter Zugriff 27.09.2018].

Carusi, Annamaria / Reimer, Torsten (2010): *“Virtual Research Environment Collaborative Landscape Study. A JISC funded project.”* Oxford/London: JISC http://www.jisc.ac.uk/publications/reports/2010/vrelandscapestudy.aspx [letzter Zugriff 27.09.2018].

Dogunke, Swantje (2017): *„Tagungsbericht: Editionsportale, 03.08.2017 – 04.08.2017 Jena“*, in: H-Soz-Kult, 10.10.2017, <www.hsozkult.de/conferencereport/id/tagungsberichte-7350> [letzter Zugriff 28.09.2018].

Mutschler, Thomas (2017): *„Neue Wege der Kulturgutdigitalisierung in Thüringen“*, in: Bibliotheksdienst 51, 310-321.

Prell, Martin / Schmidt-Funke, Julia (Hg.) (2017): *„Digitale Edition der Briefe Erdmuthe Benignas von Reuß-Ebersdorf (1670-1732)“*. Jena http://erdmuthe.thulb.uni-jena.de [letzter Zugriff 28.09.2018].

Sahle, Patrick (2014): *„Kriterienkatalog für die Besprechung digitaler Editionen, Version 1.1“* (unter Mitarbeit von Georg Vogeler und den Mitgliedern des IDE) https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ [letzter Zugriff 27.09.2018].

Ein GUI-Topic-Modeling-Tool mit interaktiver Ergebnisdarstellung

Severin, Simmler
severin.simmler@stud-mail.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

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

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

Zu den Verfahren der digitalen Textanalyse, die in den vergangenen Jahren in den textbasierten digitalen Geisteswissenschaften etabliert wurden, gehört das LDA (Latent Dirichlet Allocation) Topic Modeling (Blei 2012; Steyvers und Griffiths 2006). Diese Methode eignet sich zur Analyse der Verteilung semantischer Wortgruppen in Textsammlungen, und kann sowohl für die computergestützte Textklassifikation als auch für die explorative Betrachtung der Inhalte eines Corpus herangezogen werden. Um dem zunehmenden Interesse am Topic Modeling von Seiten der DH-Community Rechnung zu tragen, entwickelt DARIAH-DE seit 2017, basierend auf den Python-Bibliotheken "LDA" von Allan Riddell und "DARIAHTopics" (Jannidis et al. 2017), den DARIAH-TopicsExplorer, eine Software, mit der interessierte Forschende Topic Modeling auf ihren eigenen Rechnern an ihren eigenen Texten ausprobieren können, und die den gesamten Analyseprozesses, vom unverarbeiteten Text bis zum Ergebnis, durch eine graphische Nutzeroberfläche (GUI) unterstützt.

Der TopicsExplorer bietet nicht die Leistungsfähigkeit und vor allem die Flexibilität bisheriger Lösungen, die entweder, wie das weit verbreitete MALLET (McCallum 2002), als Kommandozeilenprogramme, oder aber, wie Gensim (Rehurek und Sojka 2010), als Bibliothek für eine Programmiersprache konzipiert wurden. Dafür kann er aber ohne Kenntnis irgendeiner Programmier- oder Skriptsprache eingesetzt werden und muss nicht einmal über die Kommandozeile aufgerufen werde. Damit schließt

der TopicsExplorer eine wichtige Lücke: Forschende, die selbst nicht, oder noch nicht, programmieren, können sich hier einen Eindruck davon verschaffen, wie die Methode funktioniert und was sie theoretisch leisten kann. Das befähigt auf der einen Seite, Forschungsarbeiten, die auf Topic Modeling basieren, und ihre Ergebnisse informiert einzuschätzen. Auf der anderen Seite kann das Werkzeug für einfache Fragestellungen, die Topic Modeling erfordern, direkt eingesetzt werden, und bei komplizierteren Ansätzen eine informierte Entscheidung darüber ermöglichen, ob sich die Aneignung der notwendigen technischen Fähigkeiten für die Verwendung einer fortgeschrittenen Lösung für die jeweilige Forschungsfrage überhaupt lohnt.

Der TopicsExplorer wurde in einer ersten Version 2017 und 2018 im Rahmen mehrerer Workshops und Konferenzen verschiedenen Gruppen von Forschenden und Studierenden vorgestellt. Die Erfahrungen aus dem Umgang mit Nutzerinnen und Nutzern in solchen Workshops und vor allem ihr direktes Feedback sind seither umfangreich in die Weiterentwicklung eingeflossen und die daraus resultierenden Änderungen gehen weit über die Beseitigung von Bugs und die Sicherstellung der nachhaltigen Funktionalität hinaus. Aus einem anfänglichen "GUI-Demonstrator" (Simmler et al. 2018), der noch die Installation einer Python-Bibliothek erforderte, auf einem lokalen Server lief und im Browser angezeigt wurde, ist eine vollwertige Stand-Alone-Software geworden, die nach dem Herunterladen ohne weitere Vorbereitung auf gängigen Windows-, MacOS- und Linux-Systemen gestartet werden kann. Die Visualisierungen können interaktiv manipuliert, und die Ergebnisse im csv-Format exportiert werden. Zahlreiche kleinere, von der Testcommunity gewünschte Features, wie z.B. eine Fortschrittsanzeige mit Abbruchbutton (Abb. 1), haben die Usability wesentlich verbessert.

Zur Zeit wird eine Version in grundlegend überarbeitetem Design vorbereitet, deren Ziel es ist, auch komplizierteren, aus den Testcommunies heraus formulierten Ansprüchen an die Interaktivität der Software gerecht zu werden. Auf technischer Ebene wird dabei die Visualisierung der Ergebnisse statt mit der bisher verwendeten Python-Bibliothek "Bokeh" direkt in Javascript realisiert, um zusätzliche Flexibilität für die Umsetzung neuer Funktionalitäten zu gewinnen. Äußerlich wird es damit möglich, Ergebnisse auch nachträglich interaktiv umzusortieren und in mehreren Fenstern darzustellen. Mit dem Ziel, die User-Experience im explorativen Umgang mit den erzeugten Modellen zu verbessern, wird darüber hinaus ein völlig neues Visualisierungskonzept auf Basis der Vorschläge von Chaney und Blei (2012) umgesetzt. Dieses Konzept erlaubt es nicht nur, einzelne Dokumente im Corpus mitsamt den dazugehörigen Analyseergebnissen zu betrachten, es werden darüber hinaus automatisch andere Texte mit ähnlichen inhaltlichen Schwerpunkten vorgeschlagen (Abb. 2).

Wir hoffen, dass der TopicsExplorer mit den angestrebten Verbesserungen und Erweiterungen dazu beiträgt, eine mittlerweile doch recht verbreitete Forschungsmethode aus der Nische derjenigen DH-Verfahren heraus zu holen, die nur von Programmiererinnen und Programmierern verwendet, verstanden und kritisch diskutiert werden. Die neuen Features, die für das kommende Release entwickelt werden, sollten dazu einen wesentlichen Beitrag leisten.

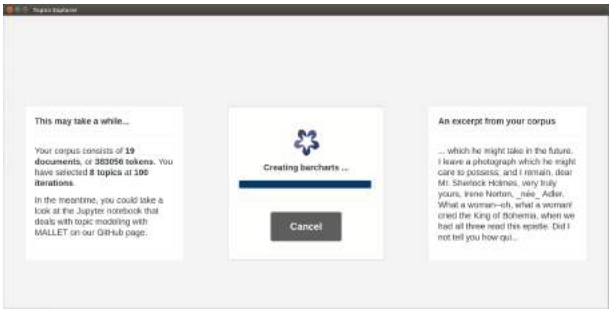


Abbildung 1. Fortschrittsanzeige für die laufende Modellierung im aktuellen TopicsExplorer

Abbildung 2. Übersicht für ein Dokument im Prototypen der Version 2

Bibliographie

Blei, David M. (2012): *Probabilistic Topic Models*, in: Communication of the ACM55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

Chaney, Allison J.B. / Blei, David M. (2012): *The Visualizing Topic Models*, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media 419-422.

Jannidis, Fotis/ Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2017): *Making topic modeling easy: a programming library in Python*, in: Proceedings of the Digital Humanities 2017 Conference.

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.*

Rehurek, Radim/ Sojka, Petr (2010): *Software framework for topic modelling with large corpora.* In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Simmler, Severin / Vitt, Thorsten / Pielström. Steffen (2018): *LDA Topic Modeling über eine graphisches Interface*, in: Konferenzabstracts der 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. 428-429.

Steyvers, Mark/ Griffiths, Tom (2006): *Probabilistic Topic Models*, in: Latent Semantic Analysis: A Road to Meaning, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

Ein Web Annotation Protocol Server zur Untersuchung vormoderner Wissensbestände

Tonne, Danah
danah.tonne@kit.edu
Karlsruher Institut für Technologie, Deutschland

Götzelmann, Germaine
germaine.goetzelmann@kit.edu
Karlsruher Institut für Technologie, Deutschland

Hegel, Philipp
philipp.hegel@fu-berlin.de
Freie Universität Berlin, Deutschland

Krewet, Michael
m.krewet@fu-berlin.de
Freie Universität Berlin, Deutschland

Hübner, Julia
julia.huebner@fu-berlin.de
Freie Universität Berlin, Deutschland

Söring, Sibylle
ssoering@cedis.fu-berlin.de
Center für Digitale Systeme

Löffler, Andreas
andreas@loeffler-bermayer.de
Karlsruher Institut für Technologie, Deutschland

Hitzker, Michael
mail@hitzker.de
Karlsruher Institut für Technologie, Deutschland

Höfler, Markus
Markus.Hoefler@live.de
Karlsruher Institut für Technologie, Deutschland

Schmidt, Timo
timo.schmidt2@student.kit.edu
Karlsruher Institut für Technologie, Deutschland

Der Sonderforschungsbereich 980 „Episteme in Bewegung“

Der Sonderforschungsbereich 980 (SFB 980) „Episteme in Bewegung“ untersucht Prozesse des Wissenstransfers und Wissenswandels in europäischen und nicht-europäischen Kulturen vom 3. Jahrtausend vor Christus bis etwa 1750 nach Christus. In 27 Teilprojekten aus 21 Disziplinen wird gezeigt, wie gerade dort, wo in den Selbstbeschreibungen der vormodernen Kulturen und aus der Perspektive der Moderne Kontinuität und Stabilität im Vordergrund stehen, Neukontextualisierungen von Wissen fassbar werden. Die konkreten Möglichkeiten digitaler Methoden sowie Auswirkungen auf den Forschungsprozess werden an Hand zweier Anwendungsfälle dargelegt.

Anwendungsfall 1: Prozesse der Traditionsbildung in der de interpretatione-Kommentierung der Spätantike

Die Erforschung der handschriftlichen Überlieferung einer Schrift kann Erkenntnisse über beispielsweise Verwandtschaftsverhältnisse oder Verwendungskontexte einzelner Exemplare liefern. Im Falle von Aristoteles' Schrift *de interpretatione*, von der 150 griechischsprachige Handschriften erhalten sind, stößt die analoge Untersuchung jedoch an ihre Grenzen (Montanari 1984, Reinsch 2001). Vielfach hatte ein Kopist eines Textes gleich mehrere Exemplare mit unterschiedlichen Textversionen vor sich. Beim Kopieren konnte er diese vermischen oder sogar Vorlagen im Sinne einer für besser befundenen Version korrigieren. Ebenso konnte er beim Kopieren verschiedene Erklärungen unterschiedlichster Provenienz an den Rand des Textes schreiben. Diese Praktik macht es immens schwierig, den genauen Weg, wie die Schrift überliefert wurde, und Traditionen unter den Erklärungen zu erforschen.

Anwendungsfall 2: Vermittlung kommunikativer Alltagsroutinen im Kontext sprachlicher Diversität in der Frühen Neuzeit

Im Zentrum stehen 315 Lehrwerke für Moderne Fremdsprachen mit Deutsch-Anteil (Glück 2002) aus der Frühen Neuzeit (Mitte des 15. Jahrhunderts bis Anfang des 18.Jahrhunderts). Diese Sprachlehrwerke waren in der Regel mehrsprachig angelegt und richteten sich an Reisende aller Art. Einige dieser Drucke wurden über Jahrzehnte immer wieder in überarbeiteter Form herausgegeben und eignen sich daher besonders gut für die Untersuchung von Wandelprozessen. Eine weitere spannende Fragestellung bilden die Autorisierungsstrategien in diesen Lehrwerken: wer ist autorisiert, Sprache zu beschreiben / zu unterrichten und mit welchen Mitteln legitimieren die Autoren ihre Werke?

Modellierung fachwissenschaftlichen Wissens als Annotation

Zentrale Methodik zur Untersuchung der Forschungsfragen in beiden Anwendungsfällen ist die Anreicherung von Bilddigitalisaten mit zusätzlichen Informationen. Dieses Wissen wird auf technischer Ebene als digitale Annotation modelliert, wobei das verwendete Modell für beliebige referenzierbare Daten nutzbar sein soll. Seit Februar 2017 steht dazu mit dem *Web Annotation Data Model* (WADM) des W3C-Konsortiums (Young et al. 2017) der Nachfolger zum weitverbreiteten *Open Annotation Data Model* zur Verfügung. Mit diesem standardisierten Modell und Format soll der Austausch von Annotationen über Disziplin- und Systemgrenzen ermöglicht werden.

In der Infrastruktur des SFB 980 finden diese Empfehlungen ihre Anwendung, um der großen Heterogenität der Disziplinen, Erkenntnisinteressen und Arbeitsweisen und damit auch der Vielfältigkeit der Annotationstypen, -formate und -praktiken Rechnung zu tragen. Schon bei den zwei betrachteten Anwendungsfällen entstammen die Annotationen drei unterschiedlichen Quellen:

1. *Adobe Acrobat:* Forschende annotieren Digitalisate innerhalb von PDFs, die Annotationen können als XFDF exportiert werden.
2. *Automatische Layoutanalyse:* Algorithmen vermessen Seiten-, Text- und Bildbereiche, die Informationen werden im PAGE-Standard abgelegt.
3. *Projektspezifische grafische Annotationsoberfläche:* Informationen werden direkt als Annotationen gemäß des Web Annotation Data Models gespeichert.

Um eine disziplinübergreifende Auswertung zu ermöglichen, werden die XML-Dateien aus 1. und 2. mit Hilfe eines Parsers in das Web Annotation Data Model überführt und im Annotationsspeicher zugreifbar abgelegt. Abbildung 1 zeigt Auszüge der Modellierung für die automatische Layoutanalyse sowie manueller fachwissenschaftlicher Annotation. Die vermessenen bzw. eingegebenen Informationen werden im so genannten *body* aufgeführt. Zentral ist hierbei die Nutzung des Feldes *purpose* zur Klassifizierung der einzelnen *bodies* mit Hilfe von Kategorien aus dem Web Annotation Data Model (z.B. *classifying*, *tagging*) und der TaDiRAH (Borek et al. 2016) Taxonomie (z.B. *translation*, *transcription*), um eine automatische Auswertung und Visualisierung zu ermöglichen. Auch die aus der Layoutanalyse stammende Unterscheidung von Seiten-, Text- und Bildbereichen wird im *body* modelliert. Die annotierten Bilder werden im *target* referenziert, Bildausschnitte konform mit dem SVG-Standard definiert. Durch die Nutzung des Feldes *creator* sowohl auf Annotationsebene als auch in den *bodies* kann die Herkunft der Informationen nachverfolgt werden und können beispielsweise verschiedene Algorithmusversionen der Layoutanalyse verglichen werden.

Abbildung 1: JSON-Repräsentation einer Annotation (links) mit Auszügen möglicher bodies aus fachwissen-schaftlicher Annotation (Mitte) bzw. automatischer Layoutanalyse (rechts) im Web Annotation Data Model

Ein Blick unter die Haube: RDF-Server und SPARQL-Anfragen

Nach der Modellierung ist die verlässliche Sicherung und die standardisierte Abrufbarkeit der Annotationen von immenser Bedeutung. Das gemeinsam mit dem *Web Annotation Data Model* veröffentliche *Web Annotation Protocol* (Sanderson 2017) definiert eine REST-Schnittstelle für einen Annotationsspeicher und stellt so eine erfolgreiche Kommunikation zwischen Servern und Clients sicher. Ein im Rahmen der SFB-Infrastruktur entwickelter, generischer Annotationsserver setzt alle obligatorischen und viele der optionalen Vorgaben des Protokolls sowie die für Annotationen relevanten Teile der *Linked Data Platform* (LDP)-Empfehlungen vollständig um. Die javabasierte Serverarchitektur bindet dabei modular einen RDF Triple Store (derzeit Apache Jena TDB2, aber prinzipiell austauschbar) an. Auf diese Weise ist ein standardkonformer und damit an verbreitete Annotationsprogrammen anbindbarer Server entstanden, dessen lose gekoppelte Speicherkomponente bei beispielsweise Softwareobsoleszenz oder gestiegenen Skalierbarkeitsanforderungen mit geringstem Aufwand ausgetauscht werden kann. Nach aktuellem Stand sind ca. 15500 Bildannotationen aus automatischer Layoutanalyse und manueller Transkription und Klassifizierung im WADM-Format in Form einzelner RDF-Graphen abgelegt.

Neben der REST-Schnittstelle ist mit dem Triple Store zugleich auch ein SPARQL 1.1 Endpunkt verbunden, der semantische Anfragen an die Annotationsdaten ermöglicht. Hier rücken neben den formalen Vorgaben der Annotationen zur Erzeugung, Bearbeitung und Anzeige (Annotationstyp, Selektoren, Links zu Ressourcen, etc.), die Annotationsinhalte in den Blick. Diese sind je nach Forschungsinteresse unterschiedlich ausgestaltet und in einem oder mehreren *bodies* abgelegt. Mit inhaltsspezifischen SPARQL-Anfragen können diese *bodies* miteinander oder mit den *targets* in Beziehung gesetzt und ausgewertet werden. Über die eigenen Serverinhalte hinaus können mit Hilfe sogenannter föderierter SPARQL-Anfragen weitere Metadaten innerhalb der *Linked Open Data Cloud* zur Analyse hinzugezogen werden, ohne dass diese redundant abgelegt werden müssen. So eröffnen sich Möglichkeiten zur Nachnutzung von Norm- und Geodaten sowie zur projektübergreifenden Nutzung von Vokabularen und Taxonomien.

Ergebnisse und Ausblick

Bei *de interpretatione* wurden Textbereiche, besondere Textvarianten, Erklärungen, Diagramme und Interlinearglossen mit zusätzlichen Informationen wie Beschreibungen, Transkriptionen, Übersetzungen und vielfältigen Tags angereichert. Mit Hilfe von SPARQL-Anfragen werden beispielsweise die Transkriptionen von Interlinearglossen auf verschiedenen Digitalisatseiten dahingehend abgefragt, zu welchem Satz des Aristoteles-Textes und zu welchem Wort oder Satzfragment sie gehören (realisiert durch *tagging*). Auf diese Weise können verschiedene Variationen der Interlinearglossen verglichen, gezählt und visualisiert werden. Schon bei Betrachtung eines einzelnen Satzes (vergleiche Abbildung 2) lassen sich im Handschriftenkorpus Gemeinsamkeiten zwischen Manuskriptgruppen aufzeigen. Für die anderen Formen der antiken Kommentierung sowie der Textvarianten gestaltet sich die Auswertung analog. Die so durchgeführte Untersuchung von signifikanten Gemeinsamkeiten und Variationen *aller* Handschrifteninhalte ergänzt die herkömmliche textkritische Methode und stellt damit eine unerlässliche Hilfe für die Erforschung der komplexen Austauschprozesse dar, die bei der Textüberlieferung stattgefunden haben.

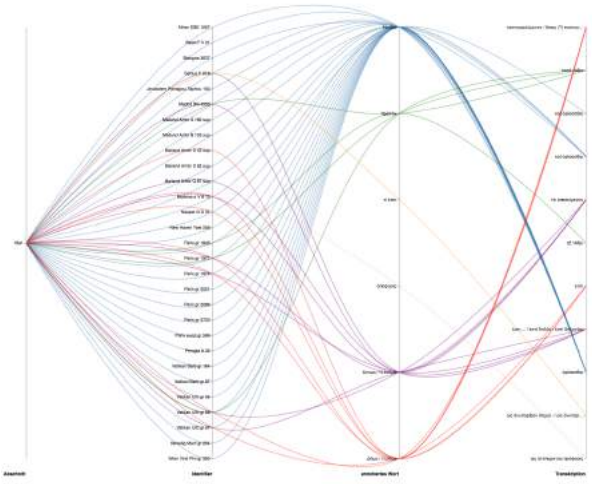


Abbildung 2: Übersicht der Interlinearglossen des Satzes 16a1 aus *de interpretatione* (1. Achse: Satz des Referenztextes, 2. Achse: Signatur der Handschrift, 3. Achse: glossiertes Wort im Referenzsatz, 4. Achse: Transkription der Glosse)

Im Falle der frühneuzeitlichen Sprachlehrwerke dient ein Annotationstool in einem ersten Schritt dem explorativen Annotieren und dem damit einhergehenden Aufdecken unterschiedlichster Prozesse des Wissenswandels. Im Weiteren wird das Tool dann genutzt, um Einzelanalysen vorzunehmen und um Teile des Korpus quantitativ auszuwerten. So werden Autorisierungsstrategien ausfindig gemacht und in verschiedene Kategorien unterteilt. Diese Ergebnisse werden anschließend mit den Metadaten wie Autor, Zielsprache, und didaktischer Ausrichtung des Lehrwerks korreliert.

Die vorgestellten Anwendungsfälle nutzen bereits die entwickelte Annotationsinfrastruktur und konnten eine neue Herangehensweise an ihre spezifischen Fragestellungen erproben. Die Infrastruktur erlaubt es, Text- und Bilddaten in ihren wechselseitigen Bezügen in den Blick zu nehmen und gemeinsam zu betrachten. Das *Web Annotation Data Model* hat sich als vielseitiges Modell bewiesen, um die heterogenen Annotationen zu modellieren. Dabei sind die Anwendungsmöglichkeiten jedoch weder auf die beiden Anwendungsfälle noch auf den Sonderforschungsbereich in seiner Gesamtheit beschränkt, sondern im Gegenteil Potentiale für alle Disziplinen und digitalen Datensätze erkennbar. Unterschiedliche, disziplinspezifische Praktiken und Anforderungen können im Modell abgebildet werden, die Annotationen verbleiben nach der zugehörigen Ontologie (Ciccarese 2017) strukturiert und damit auswertbar. Für die Inhalte ist allerdings kein semantisches Modell vordefiniert und auch nicht a priori festlegbar. Der Aushandlungsprozess eines solchen Modells ist daher für jede Anwendergruppe erneut und jeweils iterativ durchzuführen. Ebenso verbleibt die manuelle Annotation für die Forschenden arbeitsintensiv, da vielfältige Informationen abgelegt werden. Grafische Benutzeroberflächen können hier unterstützen, den Prozess so komfortabel wie möglich zu gestalten. Zusätzlich werden durch die Verwendung des *Web Annotation Data Models* die Kollaboration mit anderen Fachwissenschaftlerinnen und Fachwissenschaftlern und die Nachnutzbarkeit ermöglicht, so dass die Arbeitslast reduziert werden kann. Größter Vorteil ist jedoch die Möglichkeit, Informationen mit stabilen, in einem Repositorium abgelegten Daten zu verknüpfen und mit dynamisch im Forschungsprozess erlangtem Wissen anzureichern. Auf diese Weise werden Objektmetadaten, automatische und fachwissenschaftliche Annotationen gemeinsam auswertbar und damit gänzlich neue Erkenntnisse möglich.

Bibliographie

Borek, Luise / Perkins, Jody / Schöch, Christof / Dombrowski, Quinn (2016): *„TaDiRAH: A Case Study in pragmatic classification“* in: DHQ: Digital Humanities Quarterly 10/1.

Ciccarese, Paolo / Young, Benjamin / Sanderson, Robert (2017): *„Web Annotation Vocabulary. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/ [letzter Zugriff 19.12.2018].

Glück, Helmut (2002): *Deutsch als Fremdsprache in Europa vom Mittelalter bis zur Barockzeit*. Berlin/New York: De Gruyter.

Linked Data Platform (LDP) 1.0: https://www.w3.org/TR/ldp/ [letzter Zugriff 05.10.2018].

Montanari, Elio (1984): *La sezione linguistica del Peri Hermeneias di Aristotele*. Florenz.

PAGE XML-Schema: https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd [letzter Zugriff 05.10.2018].

Reinsch, Diether Roderich (2001): *„Fragmente einer Organon-Handschrift des zehnten Jahrhunderts aus dem Katharinenkloster auf dem Berg Sinai“* in: Philologus 151: 151-177.

Sanderson, Robert (2017): *„Web Annotation Protocol. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-protocol-20170223/ [letzter Zugriff 08.11.2018].

Sonderforschungsbereich 980 „Episteme in Bewegung“: http://www.sfb-episteme.de/ [letzter Zugriff 05.10.2018].

SPARQL: https://www.w3.org/TR/sparql11-query/ [letzter Zugriff 05.10.2018].
SVG: https://www.w3.org/TR/SVG11/ [letzter Zugriff 05.10.2018].
TDB2: https://jena.apache.org/documentation/tdb2/ [letzter Zugriff 05.10.2018].
XML Forms Data Format Specification: https://web.archive.org/web/20160408204348/https://partners.adobe.com/public/developer/en/xml/XFDF_Spec_3.0.pdf [letzter Zugriff 05.10.2018].
Young, Benjamin / Ciccarese, Paolo / Sanderson, Robert (2017): *„Web Annotation Data Model. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-model-20170223/ [letzter Zugriff 08.11.2018].

Erneuerung der Digitalen Editionen an der Herzog August Bibliothek Wolfenbüttel

Schaßan, Torsten
schassan@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Baumgarten, Marcus
baumgarten@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Steyer, Timo
steyer@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Fricke-Steyer, Henrike
henrike.fricke@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Iglesia, Martin de la
iglesia@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Kampkaspar, Dario
kampkaspar@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Klaffki, Lisa
klaffki@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Parlitz, Dietrich
parlitz@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Die Anfänge der Wolfenbütteler Digitale Bibliothek (WDB)

Die Wolfenbütteler Digitale Bibliothek (WDB) wurde an der Herzog August Bibliothek (HAB) vor fast 20 Jahren auf der Grundlage der damals verfügbaren Technologien zur Anzeige von Objektdigitalisiaten konzipiert: PHP-Skripte, die auf einem Apache-Server ausgeführt werden; Frame-Technologien zur Darstellung unterschiedlicher Inhalte nebeneinander auf dem Bildschirm; Datenablage in einfachen Ordnerstrukturen.

Im Hintergrund wurde die Präsentationsoberfläche zum Workflow-Tool zur Objektdigitalisierung ausgebaut. Ebenfalls PHP-Skript-gestützt wurden Eingabeoberflächen für die Eingabe von Bestellungen von Digitalisaten und die Prüfergebnisse der Restaurierung, für die Dokumentation des Bearbeitungsstandes in der Fotowerkstatt und der Veröffentlichung von Digitalisaten programmiert und an die bestehenden Struktur angebunden.

Nach und nach erweiterte sich die Bandbreite der Inhalte, die über die WDB zugänglich gemacht werden sollten. Neben die Repräsentation von vollständigen Objektdigitalisaten, die von Cover zu Cover digitalisiert wurden, traten Digitalisierungen, die primär in anderen Systemen zur Anzeige kommen sollten wie beispielsweise die Digitalisate des Virtuellen Kupferstichkabinetts. Es entstanden die ersten digitalen Editionen, in denen komplexe digitale Inhalte gemeinsam auf den Bildschirm gebracht werden mussten. E-Books und Rundum-Digitalisate von Handschriften, die in besonderer Weise zur Anzeige gebracht werden mussten, sind weitere Beispiele. Dadurch verkomplizierte sich nach und nach die Struktur der Programmierung der WDB abermals.

Neue Herausforderungen

Mit der Zeit haben sich Sehgewohnheiten geändert und die Anforderungen an das Design digitaler Inhalte sind gestiegen. Das Aufkommen mobiler Endgeräte, die zunehmend zum Konsum dieser Inhalte genutzt werden, zeigten auf, dass die Darstellung, die im Rahmen der WDB genutzt wird, teilweise nicht mehr funktional ist. Insbesondere das Eigenleben, das Daten in Zeiten des Linked Open Data zu pflegen führen, zeigt die Limitationen der WDB auf. Insbesondere die Option, digitale Inhalte aus der WDB in unterschiedlicher Granularität verlässlich zu adressieren und in unterschiedlichen Kontexten anzeigen zu können, erfordert eine grundlegende Neukonzeption der Mechanismen zur (persistenten) Adressierung der Seiten.

Andererseits läßt die Digitalisatpräsentation grundlegende Funktionalitäten vermissen, die sich in anderen Bibliotheken durchgesetzt haben. So sind beispielsweise ein Drehen der Bilder oder andere Bildmanipulationen nicht möglich. Auch Navigationsmöglichkeiten, die sich auf mobilen Endgeräten durchgesetzt haben, wie das Wischen von Seite zu Seite sind in der WDB nicht möglich. Die Navigation wie der Zoom in den Digitalisaten sind noch mit altertümlich anmutenden Steuerungselementen gelöst.

Die Erneuerung

In der HAB wurde nach Feststellung dieser Mangel eine temporäre Arbeitsgruppe zur Erneuerung der WDB ins Leben gerufen. Dabei wurde unter anderem die Aufteilung der Anforderungen in drei Säulen festgelegt: Objektdigitalisate, digitale Editionen und digitale Publikationen. Die Darstellung der reinen Objektdigitalisate inklusive der Workfloworganisation sollen getrennt werden von der Anzeige digitaler Editionen und einem möglichen, neu aufzubauenden Publikationsserver. Im Poster werden vor allem die Überlegungen zur Erneuerung der Säule „Digitale Editionen“ präsentiert.

Die Arbeitsgruppe hat eine Liste an Fragen und Funktionalitäten identifiziert, für welche je separat untersucht, der State-of-the-Art skizziert und die Anforderungen der HAB formuliert wurden:

- Anforderungen an die Startseite der WDB
- aktuelle Suchtechnologien
- Layoutoptimierung
- Versionierung und persistente Adressierung
- Downloadmöglichkeiten
- (externe) Verwaltung von Literaturlisten
- Visualisierungsstrategien
- Schnittstellen
- Schemaunterstützung und -dokumentation
- IIIF
- Langzeitarchivierung nach OAIS
- Trennung von Arbeitsumgebung, Archiv und Publikation
- bevorzugte Arbeitsumgebung (z.B. XML-Datenbanken)
- Nutzerstatistik
- Linked Open Data

Für jedes der Felder wurde eine Bedarfsanalyse und eine Marktsondierung durchgeführt und konzeptionelle Antworten für die Erneuerung der WDB in diesem Bereich formuliert.

In dem Poster sollen nicht nur die technische Erneuerung adressiert werden, sondern auch die Einbettung der neuen Infrastruktur in einen organisatorischen Workflow und die Frage, warum Infrastruktur altert und wie man dem entgegenwirken kann.

FormIt: Eine multimodale Arbeitsumgebung zur systematischen Erfassung literarischer Intertextualität

Schlupkothen, Frederik
schlupko@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Nantke, Julia
nantke@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Zielsetzung

Eine der großen Herausforderungen sowohl im Studium als auch in der Erfassung und Beschreibung geisteswissenschaftlicher Praktiken liegt in deren mangelnder Beobachtbarkeit (vgl. Spoerhase 2015: 59–61). Dieser Umstand wirkt sich nicht nur auf die wissenschaftstheoretische Erfassung der Forschungspraxis aus. Er schlägt sich ebenfalls signifikant in der unzureichenden Systematisierung zentraler Phänomene literarischer Kommunikation nieder. Ein prominentes Beispiel hierfür bildet die Kategorie literarischer Intertextualität: Bisherigen Ansätzen zur strukturierten Erfassung fehlt es durchweg an einer systematischen Koppelung von induktiv beobachteten Textmerkmalen (etwa in Heilmann 1998; Bauer Lucca 2001; Buß 2006) und übergeordneten Taxonomien. Letztere werden bislang lediglich punktuell durch konkrete Beispiele belegt (etwa Schulte-Middelich 1985, Genette 1993, Holthuis 1993). Zudem fehlt eine Beschreibungssprache, die den sich dabei vollziehenden Abstraktionsprozess nachvollziehbar macht.

Im Projekt *FormIt* kommen verschiedene Repräsentations-Modi zum Einsatz, um den beiden wechselwirkenden Herausforderungen der mangelnden Sichtbarkeit und Konkretisierung zu begegnen. Die Praxis der Herstellung und Interpretation intertextueller Beziehungen zwischen literarischen Texten wird mithilfe digitaler Methoden und Werkzeuge präzisiert und für Dritte beobachtbar gemacht. Dies geschieht anhand der konsequenten Verknüpfung der Ebenen, auf denen sich intertextuelle „Schreibweisen“ (Verweyen/Wittig 2010: 38) im Text manifestieren, mit den Schlussfolgerungen darüber, „was das Charakteristikum ‚Intertextualität‘ konkret für einen Text bedeutet“ (Kocher 2007: 180). Ziel ist es, anhand einer multidirektionalen Verbindung zwischen Texten und Deutungen den Prozess der Abstraktion von intertextuellen Funktionsweisen zu konkretisieren und zu explizieren.

Auf dem Poster wird der hierfür bislang implementierte Workflow präsentiert und dabei anhand von Beispielen gezeigt, wie sich mittels dieser Vorgehensweise sowohl intertextuelle Beziehungen zwischen literarischen Texten als auch zwischen unterschiedlichen medialen Erscheinungsformen strukturiert beschreiben und darstellen lassen.

Vorgehensweise

Zunächst wurde im Rahmen des Projekts eine *digitale Methode* zur Erfassung literaturwissenschaftlicher Bedeutungsbildung erarbeitet (vgl. Nantke/Schlupkothen 2018), die stetig weiterentwickelt wird. Dabei werden mithilfe eines situationstheoretischen Formalismus (vgl. Barwise/Perry 1983; Devlin 1990) analytisch-interpretative Lektürepraktiken mathematisch beschrieben, was eine Übersetzung in ein maschinenlesbares Format ermöglicht. Dieser Ansatz wurde im Sinne einer *digitalen Praxis* dahingehend weiterentwickelt, dass die formale Beschreibung in eine Linkbase (Maler u.a. 2010: 2.3) integriert wurde. Diese Arbeitsumgebung dient der multimodalen Erfassung sowie der intuitiven Repräsentation der analytisch-

interpretativen Textarbeit. Perspektivisch soll dieser Prozessablauf durch ein Abfrage-Modell ergänzt werden, welches materialbasierte Aussagen bezüglich übergeordneter Kategorien und Funktionsweisen von Intertextualität auf der Basis von SPARQL (Seaborne/Harris 2013) ermöglicht (vgl. Abbildung 1).

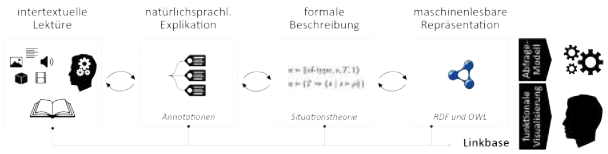


Abbildung 1. Schematische Darstellung des vollständigen Workflows

Das Projekt basiert in Erfassung, Abfrage und Darstellung der Daten konsequent auf W3C-Standards (XML-Editoren, SPARQL-Interpreter, Webbrowser). Dies gewährleistet die Einbindung in ein existierendes Software-‚Ökosystem' und somit die Nachhaltigkeit und Interoperabilität der erzeugten Datensätze.

Multimodale Arbeitsumgebung

Zur Analyse der intertextuellen Phänomene wurde eine Linkbase erzeugt, in der auf der Grundlage von XLink (Maler u.a. 2010) die bei der Feststellung und Deutung intertextueller Schreibweisen erfolgenden Arbeits- und Abstraktionsschritte erfasst sowie modular und multimodal repräsentiert werden können. Auf diese Weise werden die interpretatorischen Prozesse bei der Herstellung intertextueller Beziehungen in systematischer Form nachvollziehbar gemacht.

Die Linkbase ermöglicht die dynamische Verknüpfung der Textstellen mit der natürlichsprachlichen Erläuterung und der formalen Beschreibung der intertextuellen Beziehung (s. Abbildung 2).

(a)

```
<li xl:type="extended">
    <span xl:type="locator" xl:label="t1" xl:href="Katz.htm#f1"></span>
    <span xl:type="locator" xl:label="t2" xl:href="Owen.htm#f1"></span>
    <span xl:type="locator" xl:label="na" xl:href="Anno.htm#a1"></span>
    <span xl:type="locator" xl:label="fa" xl:href="Form.htm#s1"></span>
    <span xl:type="locator" xl:label="fa" xl:href="Form.htm#s3"></span>
    <span xl:type="arc" xl:from="t1" xl:to="t2"></span>
    <span xl:type="arc" xl:from="t2" xl:to="t1"></span>
    <span xl:type="arc" xl:from="na" xl:to="t1"></span>
    <span xl:type="arc" xl:from="na" xl:to="t2"></span>
    <span xl:type="arc" xl:from="t1" xl:to="na"></span>
    <span xl:type="arc" xl:from="t2" xl:to="na"></span>
    <span xl:type="arc" xl:from="na" xl:to="fa"></span>
    <span xl:type="arc" xl:from="fa" xl:to="na"></span>
</li>
```

(b)

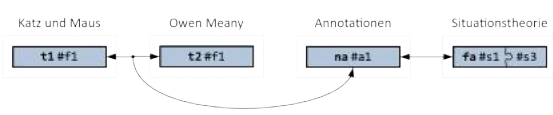


Abbildung 2. XLink-Beschreibung einer intertextuellen Relation in der Linkbase (a) und ihre schematische Darstellung (b)

In der Webbrowser-Darstellung der Linkbase werden die gesteigerten Möglichkeiten genutzt, die eine digitale Repräsentation hinsichtlich der Integration verschiedener Lektüre-Modi bietet. Die variabel kombinierbaren Segmente auf der Benutzeroberfläche ermöglichen die Repräsentation der Ausgangstexte in ihrer jeweils spezifischen Modalität. Hierbei können entweder die Texte sowie die formale Beschreibung einzeln gelesen oder gezielt spezifische Links traversiert werden (s. Abbildung 3).

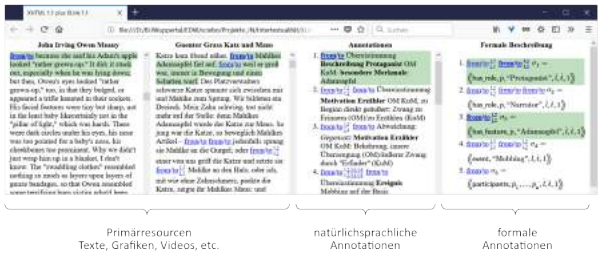


Abbildung 3. Multimodale Darstellung der intertextuellen Relation

Die Gestaltung der Benutzeroberfläche trägt der Annahme Rechnung, dass eine solche multimodale Kombination aus Lesen, visueller Erfassung und hypertextueller Nachverfolgung der textuellen Relationen sowie aus natürlichsprachlicher und formaler Beschreibung einen maßgeblichen heuristischen Mehrwert sowohl gegenüber einer rein beschreibenden Darstellung in schriftlicher Form als auch gegenüber einer rein auf Maschinenlesbarkeit gerichteten formalen Repräsentation bietet.

Fazit und Ausblick

Der dargestellte Workflow dient nicht nur der systematischen Modellierung intertextueller Beziehungen, sondern dokumentiert gleichzeitig den Prozess der dabei vollzogenen literaturwissenschaftlichen Praktiken.

Multimodalität spielt hierbei in doppelter Hinsicht eine zentrale Rolle:

Erstens werden die digitale Repräsentation der Ausgangstexte, die natürlichsprachliche und die formale Beschreibung der intertextuellen Relationen sowie diagrammatische Visualisierungen (vgl. Huber/Krämer 2018: 30) der hergestellten Zusammenhänge zwischen den einzelnen Segmenten auf der Benutzeroberfläche kombiniert.

Zweitens unterstützt der Workflow die Erfassung von Relationen zwischen reinen Schriftdokumenten ebenso wie die Modellierung von Beziehungen zwischen modal differenten Dokumenttypen wie etwa zwischen Texten, Comics und Filmen.

Das Projekt zielt gleichermaßen auf einen Einsatz im hochschuldidaktischen Bereich wie auf die Nutzung im Rahmen der computergestützten Forschung auf dem Feld der literarischen Intertextualität. Um dieses zweifache Ziel zu erreichen, werden aktuell Lösungen für die funktionale Visualisierung der in der Linkbase gespeicherten Ergebnisse in einem adäquaten Präsentationsformat sowie für die Überführung der formalen Beschreibungen in OWL erarbeitet (vgl. Abbildung 1; erste Ansätze hierzu in Kokar/Matheus/Baclawski 2009).

Bibliographie

Barwise, Jon / Perry, John (1983): *Situations and Attitudes*. Cambridge: Bradford Book, MIT Press.

Bauer Lucca, Eva (2001): *Versteckte Spuren: eine intertextuelle Annäherung an Thomas Manns Roman "Doktor Faustus"*. Wiesbaden: Deutscher Universitätsverlag.

Buß, Angelika (2006): *Intertextualität als Herausforderung für den Literaturunterricht. Am Beispiel von Patrick Süßkinds Das Parfum*. Frankfurt a. M. u.a.: Peter Lang.

Devlin, Keith J. (1990): *Logic and Information*. Cambridge: Cambridge University Press.

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

Heilmann, Iris (1998): *Günter Grass und John Irving. Eine transatlantische Intertextualitätsstudie.* Frankfurt a. M. u.a.: Peter Lang.

Huber, Martin / Krämer, Sybille (2018): *"Dimensionen digitaler Geisteswissenschaften"*, in: **Huber, Martin / Krämer Sybille (eds.)**: *Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften; 3) DOI: 10.17175/sb003_013.

Kocher, Ursula (2007): *"Im Gewirr der Fäden: Intertextualitätstheorie und Edition"*, in: **Falk, Rainer / Mattenklott, Gert (eds.)**: *Ästhetische Erfahrung und Edition*. Tübingen: Max Niemeyer, 175–185.

Kokar, Mieczyslaw M. / Matheus, Christopher J. / Baclawski, Kenneth (2009): *"Ontology-based situation awareness"*, in: Information Fusion 10, 1. St. Louis, MO, USA: Elsevier, 83–98.

Maler, Eve / Walsh, Norman / Orchard, David / DeRose, Steven (2010): *„XML Linking Language (XLink) Version 1.1"*. W3C Recommendation. https://www.w3.org/TR/2010/REC-xlink11-20100506/.

Nantke, Julia / Schlupkothen, Frederik (2018): *"Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches' Semantic Web"*, in: Konferenzabstracts DHd 2018 Kritik der digitalen Vernunft http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf , 345–349.

Schulte-Middelich, Bernd (1985): "Funktionen intertextueller Textkonstitution", in: Ulrich Broich, Manfred Pfister (eds.): *Intertextualität. Formen, Funktionen, anglistische Fallstudien*. Tübingen: Niemeyer, 197–242.

Seaborne, Andy / Harris, Steven (2013): *"SPARQL 1.1 Query Language. W3C Recommendation"*. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

Spoerhase, Carlos (2015): *"Das ‚Laboratorium' der Philologie? Das philologische Seminar als Raum der Vermittlung von Praxiswissen (circa 1850–1900)"*, in: **Andrea Albrecht, Lutz Danneberg, Olav Krämer und Carlos Spoerhase (eds.)**: *Theorien, Methoden und Praktiken des Interpretierens*. Berlin/München/Boston: De Gruyter, 53–80.

Verweyen, Theodor / Wittig, Gunther (2010): *Einfache Formen der Intertextualität: Theoretische Überlegungen und historische Untersuchungen*. Paderborn: mentis.

Forschung öffnen: Möglichkeiten, Potentiale und Grenzen von Open Science am Beispiel der offenen Datenbank "Handke: in Zungen"

Hannesschlaeger, Vanessa
vanessa.hannesschlaeger@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

In diesem Poster werden die Möglichkeiten und Grenzen der Öffnung eines Forschungsprojekts für andere Forschende, aber auch für die interessierte Öffentlichkeit dargestellt. Das Open Science-Projekt *Handke: in Zungen* dient dabei als Beispiel, anhand dessen verschiedene Aspekte der "Öffnung" von digitaler Geisteswissenschaft diskutiert werden. Nach einer kurzen Vorstellung des Projekts widmet sich der Beitrag den darin angewandten Methoden der Offenen Wissenschaft, aber auch den Anforderungen, die diese mit sich bringen sowie der Frage danach, wie und wozu Offene Wissenschaft konsequent umgesetzt werden kann.

Seit den beginnenden 1980er Jahren haben Fremdsprachen in den Bühnentexten des österreichischen Schriftstellers Peter Handke (*1942) zunehmend an Bedeutung gewonnen. In den frühen, sprachkritischen Stücken der 1960er und 70er Jahre spielen die Sprache, ihre Gemachtheit und die Reflexion darüber die zentrale Rolle – mit einem Umschwung, der sich am "dramatischen Gedicht" Über die Dörfer (1981) festmachen lässt, werden die Bühnenarbeiten Handkes zunehmend "erzählend" (Kastberger/Pektor 2012: 5), gewinnen zunehmend an "Handlung". Mit dieser "Wende" (Höller 2013), mit der auch der Beginn von Handkes Tätigkeit als Übersetzer einhergeht, halten auch die fremden Sprachen Einzug in die Stücke des Autors.

Im vorgestellten Projekt werden sämtliche fremdsprachigen Wörter und Textteile in den beinah 30 Bühnentexten Handkes erhoben und untersucht. Die Leitfragen dabei sind, ob und in welcher Weise bestimmte Sprachen für bestimmte semantische Felder und Themenbereiche eingesetzt werden, welche Sprachen vorherrschen, ob und wie sich die Wichtigkeit einzelner Sprachen im Lauf der Zeit verändert und wie die verschiedenen einfließenden Fremdsprachen miteinander in Beziehung stehen. Für die Analyse dieser Fragen werden die relevanten Textstellen in der relationalen Datenbank *Handke: in Zungen* gesammelt, wo sie sortier-, durchsuch- und auswertbar gemacht werden. Die Datenbank und das Projekt, in dessen Rahmen sie entsteht, sind der Offenen Wissenschaft verpflichtet und dienen daher als Ausgangspunkt für den geplanten Beitrag.

Die Umsetzung eines Offenen Ansatzes in Forschungsprojekten bringt eine Reihe an Themen mit sich, mit denen sich Forschende der traditionellen Geisteswissenschaften nicht vorrangig beschäftigen müssen, die aber in den Digital Humanities von zentraler Bedeutung sind. Zu diesen gehören etwa die Frage nach offener

Lizenzierung von Daten, Code und Forschungsergebnissen wie Aufsätzen und Präsentationen, aber auch jene nach deren (langfristiger) Aufbewahrung und Verfügbarmachung, nach adäquater Dokumentation und nach Kommunikation und Vermittlungsarbeit. Das *Handke: in Zungen*-Projekt eignet sich für eine Diskussion dieser verschiedenen Aspekte von Open Science deshalb in besonderer Weise, weil es unter anderem dank Unterstützung durch Wikimedia Deutschland im Rahmen des Wikimedia-Fellowship-Programms zu Freiem Wissen und Offener Wissenschaft umgesetzt wurde. Aus diesem Grund hat das Projekt neben der eigentlichen Datenbank-Web-App mehrere online-Präsenzen, die zur Öffnung des Projekts und des darin gesammelten Wissens beitragen: Zwei GitHub-Repositories machen Projekt-Informationen und -Logbuch sowie den Code der Web-App verfügbar, eine Wikiversity-Seite versammelt den Datenmanagementplan sowie alle weiteren relevanten Informationen und Berichte zum Projekt, in einer offenen Zotero-Gruppe sind die Quellenangaben der bearbeiteten Primärtexte verfügbar und auf einem Twitter-Account werden alle Interessierten über Neuigkeiten aus dem Projekt auf dem Laufenden gehalten.

Die zahlreichen Kommunikations- und Distributionskanäle, die von diesem Projekt bespielt werden, werden in diesem Beitrag vorgestellt und ihre jeweils spezifischen Vor- und Nachteile diskutiert. Ebenso werden die Voraussetzungen und Rahmenbedingungen des Projekts (rechtliche Voraussetzungen, Personal- und Zeitressourcen), die seinen Grad an Öffnung beeinflusst haben, zum Thema gemacht. Ebenfalls thematisiert werden die Notwendigkeit und Rolle von Publikumsveranstaltungen in Offenen Forschungsprojekten. Diese Bereiche werden dabei den Aktionsfeldern von Open Science zugeordnet, wie sie das Open Science Network Austria OANA definiert (Open Access, Open Research Data, Open Evaluation, Citizen Science, Open Methodology).

Es soll dabei vorgeschlagen werden, "Open Science" nicht als eine strikt definierte Methode mit einem fixen Satz an verpflichtenden Elementen der Öffnung zu verstehen. Vielmehr sollte "Open" als eine Skala gesehen werden, auf der Projekte, die offene Methoden anwenden wollen, den für sie jeweils angemessenen Platz finden müssen, der von den oben erwähnten Rahmenbedingungen mitbestimmt wird. Grundsätzlich jedoch, so das abschließende Argument dieses Beitrags, sollte sich die geisteswissenschaftliche Forschung - insbesondere die digitale - konsequent auf ihre eigene Öffnung hin orientieren. Dafür sprechen neben praktischen auch ideologische Argumente. So formulieren es auch Pomerantz und Peek in ihrem Aufsatz *Fifty shades of open*, in dem die gesellschaftliche Bedeutung von Offener Wissenschaft thematisiert wird: "As the number of open resources of all types increases, the more open resources will be created using them and derived from them, and the more open resources there will be. This snowballing growth of openness is socially beneficial, and, we believe, will make the world a better place." (Pomerantz/Peek 2016)

Bibliographie

Höller, Hans: Eine ungewöhnliche Klassik nach 1945. Das Werk Peter Handkes. Berlin: Suhrkamp 2013.

Kastberger, Klaus / Pektor, Katharina: Vorwort, in: Dies. (Hg.): Die Arbeit des Zuschauers. Peter Handke und das Theater. Salzburg/Wien: Jung und Jung 2012.

Open Science Network Austria (OANA): Über Open Science. https://www.oana.at/ueber-open-science/

Pomerantz, Jeffrey / Peek, Robin: *Fifty shades of open*, in: *First Monday* 4/2016. http://firstmonday.org/ojs/index.php/fm/article/view/6360/5460

Ressourcen zum Projekt:

Web-App: https://handkeinzungen.acdh.oeaw.ac.at/

Projekt-Logbuch: https://github.com/vanyh/handkeinzungen

Github-Repository: https://github.com/vanyh/handkeinzungen-app

Wikiversity-Seite: https://de.wikiversity.org/wiki/Wikiversity:Fellow-Programm_Freies_Wissen/Einreichungen/Dramatische_Sprachen:_Fremdsprachen_in_den_B%C3%BChnentexten_von_Peter_Handke

Zotero-Gruppe: https://www.zotero.org/groups/1840645/peter_handke_stage_texts

Twitter-Account: https://twitter.com/HandkeinZungen

Gattungserkennung über 500 Jahre

Calvo Tello, José
jose.calvo@morethanbooks.eu
Universität Würzburg, Deutschland

Fragen

Wie gut lassen sich Gattungen und Untergattungen durch Maschinelles Lernen über eine längere Periode erkennen? Obwohl eine Reihe von Artikeln die Frage hauptsächlich für Englisch (Kessler, Numberg, und Schütze 1997; Petrenz und Webber 2011; Underwood 2014) und Deutsch (Hettinger et al. 2016) beantwortet hat, befasst sich wenig Forschung mit diesem Thema aus einer diachronischen Perspektive oder wird auf spanischen Texte angewendet (Henny-Krahmer 2018). Welche Gattungen sind leichter zu erkennen, welche komplizierter? Welche Algorithmen, Transformationen und Anzahl der lexikalischen Einheiten funktionieren am besten?

Datensatz: CORDE 1475-1975

Zur Beantwortung ob verschiedene Gattungen durch Maschinelles Lernen erkannt werden können, wurde das umfangreichste historische Korpus des Spanischen analysiert, CORDE. Dieses Korpus wurde von der Real Academia Española kompiliert (Rojo Sánchez 2010; Sánchez Sánchez und Domínguez Cintas 2007) und ist ein standard-Tool in der Hispanistik über das online Such-Interface (Kabatek und Pusch 2011). Für die Analyse wurden die Frequenzen der Tokens und die Metadaten jedes Texts an Forscher weitergegeben. Das Korpus beinhaltet ca. 300 Millionen

Tokens (34.000 Texte) und die Texte sind mit expliziten Metadaten über Jahrhunderte, Länder und Gattungen markiert.

Die Daten der mittelalterlichen Sektion des Korpus präsentieren mehrere Probleme (Rodríguez Molina und Octavio de Toledo y Huerta 2017), wie beispielsweise ausgeprägte Unausgewogenheit der Anzahl der Texten im Vergleich zu anderen Jahrhunderten oder schwankende philologische Qualität. Deswegen wurden für diese Analyse nur die Texte der letzten 500 Jahre des Korpus selektiert, die länger als 100 Tokens sind. Somit beinhaltet das analysierte Korpus über 22.000 Texte (über 244 Millionen Tokens). Die Metadaten unterscheiden:

- Fachtexte in Themen (Jura, Geschichte, Geisteswissenschaften…)
- Gattungen und Untergattungen (lyrischer Vers, kurzer dramatischer Vers…)
- oder Medien (journalistische Texte, Briefe…).

Eine komplette Liste der Gattungen ist auf den Abbildungen zu finden.

Methoden der Evaluation

Die Klassifikation wurde binarisiert durchgeführt, d. h. jeder Text könnte zu jeder Gattung gehören oder nicht. Verschiedene Parameter wurden evaluiert:

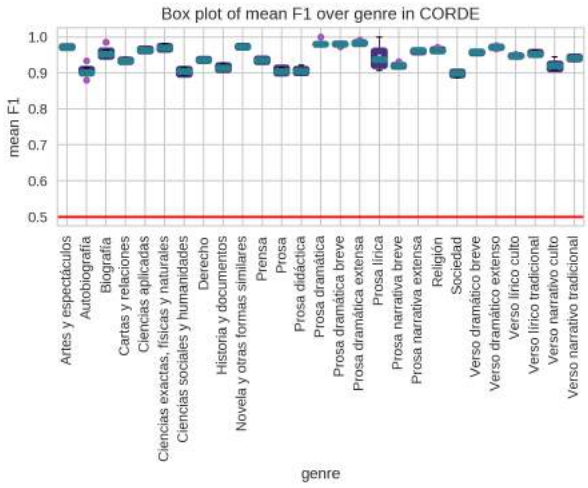
- Transformation der lexikalischen Information: relative Frequenz, binäre Frequenz, z-scores, TF-IDF, logarithmierte relative Frequenz
- Algorithmen: k-Nearest Neighbors, Random Forest, Logistic Regression und Support Vector Machine
- Anzahl der Tokens: 10, 50, 100, 500, 1.000, 2.000, 3.000, 4.000, 5.000 und 6.000

Das Korpus wurde für jede Gattung undersampled: die gleiche Anzahl an positiven wie an negativen Fällen wurden für jede Gattung gesamplet. Die Evaluation wurde mit Hilfe von Cross-Validation (10 folds) durchgeführt und der Mittelwert der F1 Scores berechnet. Der Code wird als Python Notebook über GitHub zugänglich sein.

Ergebnisse und Diskussion

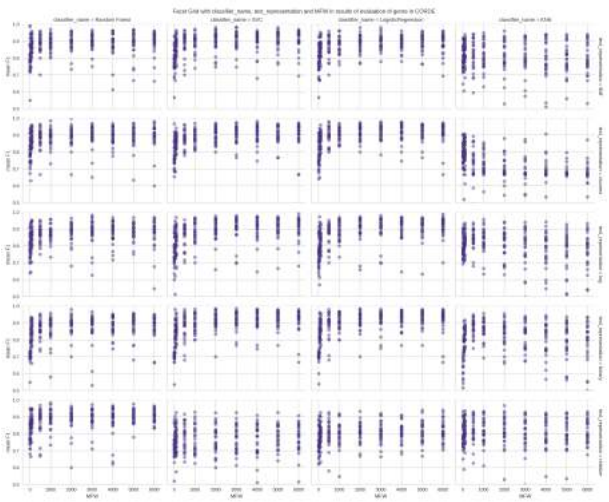
Die höchsten F1 Scores der Kombinationen von Parametern für jede Gattung lagen zwischen 0,9 und 1,0 mit einem Mittelwert der verschiedenen Gattungen von 0,96 (Standardabweichung von 0,03). Diese sehr hohen Ergebnisse ähneln sich denen von Underwood (2014), der an einem sehr großen Datensatz forschte. Die häufigsten Parameter bei den besten Ergebnissen waren Logistic Regression (16 Fälle von 27), binäre Häufigkeit (16, was nicht zu erwarten war) und 6.000 MFW (9).

Auf den nächsten Boxplots sind die 10 besten Kombinationen zu sehen. Jeder Punkt entspricht dem Mittelwert der F1 Scores der Cross-Validation der 10 besten Kombinationen von Parametern. Diese sind nach Gattung differenziert aufgelistet:



Folgende Gattungen wurden am besten erkannt: Theater (Vers und Prosa), Romane, lyrischer Vers, und Fachtexte über Naturwissenschaften und Kunst. Lyrische Prosa zeigt heftige Schwankungen, außerdem wurden die niedrigsten Ergebnisse von folgenden Gattungen erreicht: Autobiografie, narrativer Vers, Essay, lyrische Prosa, Prosa sowie Fachtexte über Gesellschaft, Geschichte und Geisteswissenschaften.

Ein interessanter Aspekt ist, welche die allgemeinen Tendenzen der Parameter und dessen Kombinationen sind. Dafür eignet sich ein Facet Grid Scatter Plot mit den Algorithmen als Spalten und den Transformationen als Reihen (einzelne Punkte entsprechen den Mittelwert der F1 Scores pro Gattung):



Hinsichtlich der Transformation (Reihen) zeigen die relative und die logarithmierte Häufigkeit niedrigere Ergebnisse als TF-IDF, z-scores und die binäre Häufigkeit. Bei den Algorithmen (Spalten) ist KNN merklich schlechter als die anderen drei. Zuletzt ist noch zu erkennen, dass die Qualität der Ergebnisse bis zu einer Anzahl von 2.000 Tokens zunimmt, und mit Schwankungen bis 6.000 stabil bleibt. Ein interessanter Aspekt ist die Tatsache, dass spezifische Kombinationen (SVC + TF-IDF, binäre + Logistic Regression,

relative + Random Forest) von Vorteil im Vergleich zu anderen sind.

Bibliographie

Henny-Krahmer, Ulrike (2018): *"Exploration of Sentiments and Genre in Spanish American Novels."* In DH Conference. Mexico City: ADHO.

Hettinger, Lena / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas (2016): *"Classification of Literary Subgenres."* In DHd Konferenz, 154–58. Leipzig: Universität Leipzig.

Kabatek, Johannes / Pusch, Claus D. (2011): *Spanische Sprachwissenschaft: eine Einführung*. Tübingen: Narr.

Kessler, Brett/ Numberg, Geoffrey / Schütze, Hinrich (1997): *"Automatic Detection of Text Genre."* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 32–38. ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics.

Petrenz, Philipp / Webber, Bonnie (2011): *"Stable Classification of Text Genres."* Computational Linguistics 37 (2): 385–93.

Rodríguez Molina, Javier / Octavio de Toledo y Huerta, Álvaro Sebastián (2017): *"La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística."* Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques, no. 6: 5–68.

Rojo Sánchez, Guillermo (2010): *"Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA."* Lingüística, no. 24: 11–50.

Sánchez Sánchez, Mercedes / Domínguez Cintas, Carlos (2007): *"El banco de datos de la RAE: CREA y CORDE."* Per Abbat: boletín filológico de actualización académica y didáctica, no. 2: 137–48.

Underwood, Ted (2014): *"Understanding Genre in a Collection of a Million Volumes*, Interim Report."

Gutenberg Biographics - der Mainzer Professorenkatalog

Gerhards, Donata
dgerhard@students.uni-mainz.de
Johannes Gutenberg-Universität, Deutschland;
Universitätsarchiv Mainz

Hüther, Frank
F.Huether@ub.uni-mainz.de
Johannes Gutenberg-Universität, Deutschland;
Universitätsarchiv Mainz

Das Projekt

Gutenberg Biographics ist ein webbasiertes System zur schwerpunktmäßigen Erfassung und Präsentation der Biographien von Mainzer Gelehrten in der Zeit von 1477-1798/1802 und seit der Wiedergründung der Johannes Gutenberg-Universität 1946 bis in die Gegenwart. Das Digital Humanities Projekt wird in Kooperation der Universitätsbibliothek Mainz, des Instituts für Geschichtliche Landeskunde an der Universität Mainz e.V sowie der Akademie der Wissenschaften und der Literatur | Mainz durchgeführt. Die Schwerpunkte liegen hierbei in der quellenbasierten Erschließung, der Forschungsdatenmodellierung im Bereich der Digitalen Biographik sowie im Aufbau und der Präsentation der webbasierten Plattform. Seit 2013 wird das Projekt vom Forschungsverbund Universitätsgeschichte finanziell unterstützt.

Gutenberg Biographics verzeichnet die Personen in ihren Rollen als Handelnde innerhalb unterschiedlicher historischer Ereignisse. Die Biographien der einzelnen Gelehrten werden als Summe ihrer Ereignisse erfasst, sodass ein Zugang sowohl über die Person selbst als auch über die Ereignisse ermöglicht wird. Die stark strukturierten Daten der einzelnen Professoren_innen und Lehrenden der Universität Mainz dienen hierbei als Basis für eine Vielzahl von interdisziplinären Fragestellungen (Vgl. Auge 2013): Welche Karrierewege sind typisch oder treten gehäuft auf? Lässt sich ein Karrieremuster feststellen? Welche wissenschaftlichen oder verwandtschaftlichen Netzwerke innerhalb der einzelnen Fachbereiche und/oder auf universitärer Ebene lassen sich identifizieren? Durch welche Mitgliedschaften in Vereinen, Interessensverbänden oder politischen Parteien nehmen die Professoren_innen Einfluss auf die Gesellschaft jenseits des akademischen Feldes (Beispielhaft hierfür Göllnitz 2013)? Durch die ereignisbasierte Modellierung der Personendaten profitieren also nicht nur die regionalgeschichtliche und biographische Forschung, sondern auch die Universitäts- und Wissenschaftsgeschichte von *Gutenberg Biographics*.

Datenstruktur

Verwendet wird das Cultural Heritage Framework der Akademie der Wissenschaften Mainz (Schrade 2017). Dieses ermöglicht eine hohe Flexibilität der Datenmodellierung, sodass es auch an zukünftige Fragestellungen oder Ereignisse angepasst werden kann. Für die insgesamt knapp 1650 Personendatensätze[1] (vgl. Eckert 2017) werden zum einen die Stammdaten bestehend aus einer System-ID, dem Namen und Namensvarianten, Lebensdaten, Konfession und Fachgebieten erfasst. Zum anderen wird jeder Personendatensatz sowie alle Orte und Körperschaften mit der jeweiligen eindeutigen Identifikationsnummer verknüpft. Personen werden zusätzlich auch über die VIAF identifiziert. Die einzelnen Ereignisse werden nach der Ereignisart, dem Datum und einer zugehörigen Rolle, welche der Handelnde einnimmt, strukturiert. Ergänzt werden diese Informationen um in Beziehung stehende Entitäten, Orte, Personen und Fachbereiche bzw. Institute sowie um archivische Quellen und Literatur. Durch die Verwendung einer GND-BEACON-Datei[2] werden verschiedene Projekte und bibliographische Informationen miteinander verknüpft und aufeinander bezogen. Die verwendeten Personendatensätze der GND werden zu Tp-Sätzen des Kategorisierungslevels 1 aufgearbeitet und fehlende Einträge innerhalb des

Normdatenkatalogs ergänzt. So verfügen alle verzeichneten Professoren nach Abschluss des Projekts über eine eindeutige standardisierte ID, die im Rahmen von Semantic Web und Linked Open Data weiterverwendet werden kann.

FAIRness

Gutenberg Biographics legt großen Wert auf Datenstandards und Normdaten sowie auf die Offenheit und gute Nachnutzbarkeit aller im Projekt erschlossenen Forschungsdaten. Hierzu werden die Daten in standardkonformen Formaten über mehrere Schnittstellen (BEACON, TEI/XML, OpenSearch) zur freien Nachnutzung angeboten und somit die FAIR-Prinzipien[3] angewendet.

F INDABLE:

Dank persistenter Identifikatoren und Permalinks können die einzelnen Uniform Resource Identifiers (URI) der Datensätze von *Gutenberg Biographics* sowohl von Maschinen als auch durch den Menschen langfristig im Web gefunden werden. Zur besseren Auffindbarkeit von Ressourcen wird eine BEACON-Datei, welche mit den Normdaten der Gelehrten versehen ist, verwendet.

A CCESSIBLE:

Gutenberg Biographics besitzt verschiedene offene Schnittstellen, über welche die Forschungsdaten bereitgestellt werden. Neben einer RESTful-API steht eine Open Search-Schnittstelle zur Verfügung.[4]

I NTEROPERABLE:

Die Forschungsdaten liegen zum Austausch in standardkonformen TEI/XML vor. Auch die Präsentationsschicht der Anwendung bietet die Forschungsdaten in strukturierter Form unter Zuhilfenahme der schema.org[5] Ontologie an. Dank der Verwendung sowohl von Normdaten in Form von GND und VIAF als auch einer BEACON-Datei, ist *Gutenberg Biographics* zu anderen Professorenkatalogen, Lexika oder bibliographischen Katalogen kompatibel.

R EUSABLE:

Abgesehen von den verwendeten Fotographien stehen alle Daten unter der Creativ Commons Lizenz CC-BY 4.0. Die offenen Daten können somit unter Nennung des Urhebers durch Dritte weiterverwendet und bearbeitet werden.

Perspektiven

Geplant, aber bisher noch nicht umgesetzt, ist ein weiterer Zugang mittels Einbindung von Geoinformationssystemen und historischen Kartenmaterialien, der eine weitere Herangehensweise zur Beantwortung von historischen Fragestellungen mit digitalen Methoden ermöglicht. Weiterhin ist angedacht, die Forschungsdaten zukünftig auch in einer CIDOC CRM Modellierung zur Verfügung zu stellen (unter Verwendung des Erlangen CRM/OWL)[6]. *Gutenberg Biographics* leistet somit einen wertvollen Beitrag als Provider von strukturierten (Forschungs-) Daten für die Geschichts- und Kulturwissenschaften im Sinne der Digitalen Geisteswissenschaften. Ein nächster und notwendiger Schritt wäre sicherlich, möglichst viele digitale Gelehrtenverzeichnisse im Rahmen einer gemeinsamen Plattform zusammenzuführen. Hierfür muss jedoch eine einheitliche Basis der Modellierung bspw. in Form von katalogübergreifenden Standards und Schnittstellen gefunden werden, welche eine Zusammenführung der unterschiedlichen methodischen Grundlagen der einzelnen Kataloge (vgl. Schrade 2016: 19) ermöglicht. Durch einen interoperablen Zusammenschluss aller Professorenkataloge kann das Potential der in ihnen enthaltenen Biographien noch besser für die (historische) Forschung genutzt werden.

Fußnoten

1. Die Zahl setzt sich zusammen aus rund 750 Professoren_innen, die bis 1973 an die Johannes Gutenberg Universität Mainz berufen worden sind und weiteren rund 900 Professoren aus der frühneuzeitlichen Mainzer Universität (1477-1798/1823). Bis Januar 2019 konnten bereits mehr als 700 Datensätze veröffentlicht werden.
2. https://github.com/digicademy/beaconizer
3. https://www.force11.org/group/fairgroup
4. http://gutenberg-biographics.ub.uni-mainz.de/daten.html
5. https://schema.org/
6. http://erlangen-crm.org/

Bibliographie

Auge, Oliver / Piotrowski, Swantje (2013): *Professorenkataloge 2.0 – Ansätze und Perspektiven webbasierter Forschung in der gegenwärtigen Universitäts- und Wissenschaftsgeschichte*, in: Jahrbuch für Universitätsgeschichte 16: 143-339.

Eckert, Karin / George, Christian / Hüther, Frank (2017): *Gutenberg Biographics: Eine biographische Online-Datenbank zur Mainzer Universitätsgeschichte*, in: ABI Technik – Zeitschrift für Automation, Bau und Technik im Archiv-, Bibliotheks- und Informationswesen 37, 3: 171-178 https://doi.org/10.1515/abitech-2017-0041 [letzter Zugriff 08. Januar 2019].

Göllnitz, Martin (2013): *Das ‚Kieler Gelehrtenverzeichnis' in der Praxis: Karrieren von Hochschullehrern im Dritten Reich zwischen Parteizugehörigkeit und Wissenschaft*, in: **Auge, Oliver / Piotrowski, Swantje (eds.):** *Professorenkataloge 2.0 – Ansätze und Perspektiven webbasierter Forschung in der gegenwärtigen Universitäts- und Wissenschaftsgeschichte*. Jahrbuch für Universitätsgeschichte 16. Stuttgart: Franz Steiner Verlag 291-312.

Pfeiffer, Barbara (2013): *Über Zweck und Nutzen der Gemeinsamen Normdatei*, in: **Auge, Oliver / Piotrowski, Swantje (eds.):** *Professorenkataloge 2.0 – Ansätze und Perspektiven webbasierter Forschung in der gegenwärtigen Universitäts- und Wissenschaftsgeschichte*. Jahrbuch für Universitätsgeschichte 16. Stuttgart: Franz Steiner Verlag 251-259.

Schrade, Torsten (2016): *Deutsche Professorenkataloge. Perspektiven, Möglichkeiten, Potentiale zur Interoperabilität* https://digicademy.github.io/presentation-catalogus-professorum/#/step-1 [letzter Zugriff 08. Januar 2019].

Schrade, Torsten (2017): *Sammlungs- und Editionsportale mit dem Cultural Heritage Framework der Digitalen Akademie. Ein Werkstattbericht.* https://digicademy.github.io/2017-editionsportale-jena/#/step-1 [letzter Zugriff 08. Januar 2019].

Gute Wörter für Delta: Verbesserung der Autorschaftsattribution durch autorspezifische distinktive Wörter

Dimpel, Friedrich
mail@dimpel.de
Universität Erlangen

Proisl, Thomas
thomas.proisl@fau.de
Universität Erlangen

1. Einleitung

Autorschaftsattributionsverfahren sind längst etabliert (vgl. Burrows 2002, Jannidis et al. 2015). Auf DHd-Jahrestagungen wurden Techniken vorgestellt, die die Attributionsverfahren optimieren (Büttner et al. 2016, Dimpel 2017a, 2018a/b). Die Optimierung der Verfahren ist deshalb wichtig, weil ein literaturwissenschaftliches Interesse besteht, die Frage nach der Autorschaft auch bei schwierigen Bedingungen zu klären. Gute Bedingungen, bei denen in Evaluationstests über hohe Erkennungsquoten berichtet wurde, sind dann gegeben, wenn viele Texte aus der gleichen Gattung vorliegen, wenn die Texte eine ausreichende Länge aufweisen (mindestens 5.000 Wortformen), wenn die Texte chronologisch nicht zu weit auseinander liegen und wenn die Texte einen möglichst normierten Sprachstand hinsichtlich Orthographie und Standardsprache aufweisen (Schöch 2014, Eder 2013a und 2013b).

All diese Bedingungen sind nicht erfüllt, wenn man mittelhochdeutsche Kleinepik untersuchen möchte – gerade etwa mit Blick auf die Kleinepik des Strickers wäre eine stilometrische Klärung bei einigen Texten interessant. Mittelhochdeutsche Texte folgen keiner geregelten Orthographie, es liegen sowohl mehr oder weniger normalisierte als auch nicht-normalisierte digitale Texte vor, sie sind oft dialektal geprägt und viele Texte aus dem Bereich der Kleinepik übersteigen kaum eine Anzahl von 2.000 Wortformen. Anhand der ‚Halben Birne' (umstrittene Autorschaft Konrads von Würzburg; 2.469 Wortformen) wurde ein Verfahren vorgestellt, wie mit Burrows' Delta die distinktiven Wörter ermittelt werden können, die Konrads Wortschatz von anderen Autoren unterscheiden (Dimpel 2018a). Damit konnte die Erkennungsqualität so optimiert werden, dass Delta auf die ‚Halbe Birne' angewendet werden konnte. Nun wird anhand eines weniger problematischen Korpus (Romane ca. 19. Jahrhundert[1]) evaluiert, wie sich Erkennungsquote und False-Positives verändern, wenn man Delta nicht auf allen Most Frequent Words (MFWs) berechnet, sondern nur auf „guten" MFWs.

2. Gute-Wörter und die Level-2-Differenz

Für Burrows' Delta ermittelt man relative Worthäufigkeiten – hier für zwei Texte von Wolfram (Willehalm, Parzival) sowie den Tristan von Gottfried. Weiter werden Mittelwert, Standardabweichung und *z*-Werte (Spalten Z1, Z2, Z3) sowie die absolute Differenz der *z*-Werte (Spalten AbsDiff) berechnet. Der Mittelwert der absoluten *z*-Wert-Differenzen ergibt das Abstandsmaß Delta – hier wäre Delta für Parzival–Willehalm der Mittelwert der drittletzten Spalte.

									Gleiche Autoren Abs Diff	Versch. Autoren Abs Diff	
F	Wh1	Pz2	Trist3	Mean	StDev	Z1 Wh	Z2 Pz	Z3 Trist	Z1 Z2	Z1 Z3	L2Diff
1 der	3,19	2,62	2,13	2,65	0,53	1,02	-0,05	-0,98	1,07	2,00	0,93
2 daz	2,02	1,93	2,77	2,24	0,46	-0,48	-0,67	1,15	0,19	1,63	1,44
3 und	1,97	1,01	3,52	2,17	1,27	-0,15	-0,91	1,07	0,76	1,22	0,46
4 ir	1,33	1,81	2,14	1,76	0,41	-1,06	0,12	0,93	1,18	1,99	0,81
5 er	1,51	1,63	1,97	1,70	0,24	-0,80	-0,32	1,12	0,48	1,92	1,44
6 in	1,30	1,25	1,66	1,40	0,23	-0,48	-0,67	1,15	0,19	1,63	1,44
7 ich	1,29	1,43	1,30	1,34	0,08	-0,67	1,15	-0,48	1,82	0,18	-1,63
...											

Die guten Wörter beruhen auf den Level-2-Differenzen: Die *z*-Wert-Differenzen zwischen den Autortexten sollen kleiner sein als die *z*-Wert-Differenzen zwischen Autortexten und Distraktortexten: Man zieht die Differenzen zwischen Autor- und Distraktortexten von den Differenzen zwischen den Autortexten untereinander ab. Im Beispiel bildet man die Differenz aus der *z*-Wert-Differenz zwischen Wolframs ‚Willehalm' und Gottfrieds ‚Tristan' (für „der": 2,0) und zwischen Wolframs 'Willehalm' und Wolframs 'Parzival' (für „der" 1,07). Die Level-2-Differenz für „der" beträgt 0,93. Bei einer positiven Level-2-Differenz kann ein Wort die Autortexte von den Distraktortexten unterscheiden; „ich" mit negativer Level-2-Differenz ist hier ungeeignet. In dem Beispiel mit nur drei Texten ist dies auch anhand der Worthäufigkeitswerte und *z*-Werte nachvollziehbar. Sonst sind jedoch weder erhöhte Worthäufigkeitswerte noch erhöhte *z*-Werte allein hinreichend, um ein Wort als distinktiv zu identifizieren (Dimpel 2017b).

Formal ausgedrückt werden die Level-2-Differenzen anhand von zwei Teilkorpora ermittelt: Einem Teilkorpus A = a 1 , … , a n von Autortexten und einem Teilkorpus D = d 1 , … , d n von Distraktortexten anderer Autoren. Dabei ist jeder Text ein Vektor von *z*-Werten. Es werden für jeden Autortext a i die absoluten *z*-Wert-Differenzen zu jedem Distraktortext d und zu jedem anderen Autortext a j ermittelt, daraus die Level-2-Differenzen berechnet und anschließend gemittelt. Für die Mittelwerte der Level-2-Differenzen gilt:

$$l = \frac{\sum_{a_i \in A} \sum_{a_j \in A \setminus a_i} \sum_{d \in D} (\mathrm{abs}(a_i - d) - \mathrm{abs}(a_i - a_j))}{|A|\,|D|\,(|A| - 1)}$$

3. Setting und Korpus

Wir prüfen, ob mit „guten Wörtern“ Texte des Zielautors besser von denen anderer Autoren abgegrenzt werden können als mit dem üblichen Delta-Ansatz, der rein auf den n häufigsten Wörtern basiert. Wir fassen die Autorschaftsattribution als ein Zwei-Klassen-Problem auf: Für jeden Text im Ratekorpus wollen wir wissen, ob er vom Zielautor oder von einem anderen Autor ist.

Für die Berechnung der guten Wörter und zur Evaluation werden ein Ratekorpus und ein Vergleichskorpus verwendet. Im Vergleichskorpus befinden sich ein Autorvergleichstext (Text vom Zielautor), je nach Testreihe zwanzig oder mehr Distraktortexte von Autoren, die nicht im Ratekorpus vertreten sind, sowie in manchen Experimenten Texte von den übrigen im Ratekorpus vertretenen Autoren. Wenn der Delta-Abstand zwischen Ratetext und Autorvergleichstext der niedrigste ist, gilt dieser Text als dem Zielautor zugeordnet; ist der Abstand zu einem Distraktortext von einem anderen Autor am niedrigsten, gilt der Text als falsch erkannt.[2] Für die Evaluation werden Precision, Recall, F_1-Score und Accuracy berechnet und über alle Autoren hinweg gemittelt.

Die guten Wörter werden jeweils auf den alphabetisch ersten drei Texten des Autors für jeden Autor in vier Varianten berechnet. Ausgangspunkt sind jeweils die häufigsten 1.200 Wörter:

1. Als gute Wörter werden Wörter gewählt, bei denen die Mittelwerte aller Level-2-Differenzen ≥ 0,4 sind.
2. wie A, jedoch Mittelwert ≥ 0,2.
3. Wie A, zusätzlich müssen 3 von 3 oder 5 von 6 Autortexten zu mindestens je einem Distraktortext eine Level-2-Differenz > 1,64 aufweisen.
4. Wie C, jedoch muss die Level-2-Differenz nur > 1,2 statt 1,64 sein; zudem mit Mittelwerten ≥ 0,4. Die zwei schlechtesten Level-2-Differenzen werden ignoriert, da ein Wort einen Text womöglich gut von vielen Distraktortexten unterscheiden kann, bei einigen wenigen Distraktortexten jedoch womöglich sehr schlecht funktioniert.[3]

Wir verwenden stets 400 gute Wörter, da auch kurze Texte mit 2.000 Token getestet werden sollen. Wenn eine Variante mehr als 400 gute Wörter liefert, verwenden wir die 400 Wörter mit den größten Level-2-Differenz-Mittelwerten; wenn sie weniger als 400 gute Wörter liefert, füllen wir mit MFWs auf.

4. Experimente

Wir führen drei Testreihen durch:

1. Bootstrapping der Ratetexte. Im Vergleichskorpus werden die vollen Texte verwendet; um die Robustheit zu erhöhen, wird ein Bootstrapping-Verfahren auf die Ratetexte angewendet. Pro Ratetext werden 100 Stichproben gezogen, wobei für jede Stichprobe zufällig so viele Wörter mit Zurücklegen aus dem Ratetext gezogen werden, wie der Ratetext umfasst.
2. Subsampling der Ratetexte. Im Vergleichskorpus werden die vollen Texte verwendet, aus den Ratetexten werden jedoch zufällige Stichproben von 2.000 Wörtern gezogen, um das Verhalten auf kurzen Ratetexten zu simulieren.
3. Subsampling der Rate- und der Vergleichstexte. Sowohl aus den Rate- als auch aus den Vergleichstexten werden zufällige Stichproben von 2.000 Wörtern gezogen, um das Verhalten auf kurzen Textsorten zu simulieren.

Dadurch werden wichtige Einsatzszenarien (lange Texte, kurze Texte mit langen Vergleichstexten und kurze Texte mit kurzen Vergleichstexten) abgedeckt.

Testreihe 1: Bootstrapping der Ratetexte

Hier evaluieren wir die Methoden bei vollständigen Romanen. Einmal enthält das Vergleichskorpus Texte von allen Autoren im Ratekorpus, im zweiten Schritt enthält das Vergleichskorpus zwar einen Text des Zielautors, jedoch keinen Text der anderen Autoren im Ratekorpus.

Für das erste Experiment verwenden wir ein Vergleichskorpus, das je einen Text von allen 32 Autoren enthält. Für jeden Autor ermitteln wir die guten Wörter auf Basis des Vergleichskorpus und zwei weiteren Texten des Autors (also insgesamt drei Autortexten und 31 Distraktortexten). Die Ratekorpora für jeden Autor umfassen drei Texte des Autors und drei zufällige Texte von den übrigen Autoren.

Mittelwerte über alle 32 Autoren:

	Zielautor			nicht Zielautor			global	
Methode	Prec.	Recall	F_1	Prec.	Recall	F_1	Acc.	F_1
MFW400	**0,9994**	0,7976	0,8633	0,8623	**0,9994**	0,9181	0,8985	0,8907
MFW1200	0,9992	0,8249	0,8795	0,8831	0,9992	0,9298	0,9120	0,9046
A	0,9774	0,9689	0,9693	0,9753	0,9716	0,9699	0,9702	0,9696
B	0,9804	**0,9750**	**0,9744**	**0,9804**	0,9752	**0,9746**	**0,9751**	**0,9745**
C	0,9803	0,9746	0,9742	0,9800	0,9751	0,9744	0,9748	0,9743
D	0,9799	0,9742	0,9738	0,9796	0,9751	0,9744	0,9746	0,9741

Die klassischen Delta-Varianten (MFW400, MFW1200) haben in Bezug auf den Zielautor eine hohe Precision und einen niedrigeren Recall, während sich das Verhältnis für Texte anderer Autoren umdreht: Niedrige Erkennungsquote, dafür kaum False Positives (Texte, die fälschlicherweise dem Zielautor zugeschrieben werden).

Die Gute-Wörter-Varianten führen zu deutlichen Recall-Verbesserungen beim Zielautor auf Kosten einer etwas geringeren Precision. Die Precision für Texte anderer Autoren steigt zu Lasten des Recalls. Mit den guten Wörtern werden also mehr Texte des Autors richtig erkannt (>97%), dafür gibt es minimal mehr False Positives (2,5%).

Mehr häufigste Wörter funktionieren besser als weniger. Alle vier Gute-Wörter-Varianten funktionieren besser als die reinen Delta-Varianten (Verbesserungen von 6–8 Punkten bei Accuracy und F_1).

Für das zweite Experiment wird lediglich das Vergleichskorpus verändert. Es enthält jetzt jeweils einen Text des Zielautors und je einen Text von 24 zusätzlichen Autoren (die sich nicht mit den 32 getesteten Autoren überschneiden). Die wahren Autoren für drei der sechs Texte in den Ratekorpora befinden sich nicht mehr im

Vergleichskorpus, was die Klassifikation der entsprechenden Texte erschwert.

	Zielautor			nicht Zielautor			global	
Methode	Prec.	Recall	F_1	Prec.	Recall	F_1	Acc.	F_1
MFW400	0,9438	0,8497	0,8657	0,8982	0,9225	0,8908	0,8861	0,8782
MFW1200	**0,9453**	0,8929	0,9008	0,9216	**0,9280**	0,9099	0,9105	0,9053
A	0,9111	0,9827	0,9383	0,9861	0,8751	0,9120	0,9289	0,9251
B	0,9112	0,9830	0,9384	0,9864	0,8750	0,9120	0,9290	0,9252
C	0,9131	0,9829	0,9395	0,9865	0,8781	0,9141	0,9305	0,9268
D	0,9134	**0,9854**	**0,9410**	**0,9883**	0,8777	**0,9145**	**0,9316**	**0,9278**

Tatsächlich bewegen sich alle Ergebnisse auf einem etwas niedrigeren Niveau als im ersten Experiment. Auch hier: Mittels guter Wörter steigt der Recall für Texte des Zielautors deutlich, während die Precision leicht sinkt. Für Texte anderer Autoren verhält es sich umgekehrt. Auch hier führen die guten Wörter zu besseren Accuracy- und F_1-Werten (2–5 Punkte).

Testreihe 2: Subsampling der Ratetexte

In dieser Testreihe wollen wir prüfen, wie gut die einzelnen Methoden für kurze Rate- und lange (vollständige) Vergleichstexte funktionieren. Dazu ziehen wir für jeden Ratetext 100 Stichproben à 2.000 Wörter; der übrige Versuchsaufbau ist identisch zu Testreihe 1.

Test 2a) Alle Autoren aus den Ratekorpora sind im Vergleichskorpus vertreten:

	Zielautor			nicht Zielautor			global	
Methode	Prec.	Recall	F_1	Prec.	Recall	F_1	Acc.	F_1
MFW400	**0,9823**	0,6993	0,7936	0,7888	**0,9835**	0,8685	0,8414	0,8310
MFW1200	0,9816	0,6704	0,7555	0,7816	0,9781	0,8573	0,8243	0,8064
A	0,9025	0,9365	**0,9081**	0,9460	0,8694	**0,8886**	**0,9029**	**0,8983**
B	0,8955	**0,9368**	0,9032	**0,9479**	0,8569	0,8777	0,8968	0,8905
C	0,8935	0,9358	0,9019	0,9463	0,8544	0,8750	0,8951	0,8884
D	0,9035	**0,9368**	**0,9081**	0,9472	0,8685	0,8851	0,9027	0,8966

Die Ergebnisse sind auf den 2.000-Wort-Samples spürbar schlechter. Das Gesamtbild ist jedoch dasselbe: Die guten Wörter verbessern die Klassifikation und führen zu deutlich höheren Recall-Werten für den Zielautor bei etwas niedrigerer Precision.

Test 2B) Die Nicht-Zielautoren sind nicht im Vergleichskorpus vertreten:

	Zielautor			nicht Zielautor			global	
Methode	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Acc.	F_1
MFW400	**0,9345**	0,7601	0,8177	0,8197	**0,9358**	**0,8648**	**0,8480**	**0,8413**
MFW1200	0,9303	0,7546	0,8003	0,8261	0,9116	0,8435	0,8331	0,8219
A	0,7844	0,9670	0,8565	0,9605	0,6806	0,7570	0,8238	0,8067
B	0,7823	**0,9682**	0,8560	0,9618	0,6783	0,7562	0,8233	0,8061
C	0,7860	0,9692	0,8586	**0,9627**	0,6829	0,7589	0,8260	0,8087
D	0,7906	0,9670	**0,8605**	0,9615	0,6925	0,7678	0,8297	0,8141

Auch hier verbessern die guten Wörter die Erkennung der Zielautortexte (Erkennungsquote >96%). Größere Einbußen gibt es bei der Erkennung von Texten, die nicht vom Zielautor stammen, so dass die guten Wörter zwar die Erkennung von Texten des Zielautors enorm verbessern (Reduktion der Fehlerrate >80%), die Ergebnisse insgesamt aber leicht schlechter sind.

Testreihe 3: Subsampling der Rate- und der Vergleichstexte

Nunmehr werden die guten Wörter mit zwei Ratetexten und einem Autorvergleichstext berechnet; wenn jedoch mehr als sechs Texte pro Autor vorliegen, werden entsprechend mehr Ratetexte verwendet. Da diese drei (oder mehr) Texte nicht wie oben reihum als Rate- und Autorvergleichstext verwendet werden, ist die Qualität der Wortlisten etwas schlechter. Im Evaluationstest werden nicht nur im Ratekorpus kleine Bag-of-Words mit 2.000 Wortformen verwendet, sondern auch im Vergleichskorpus, wodurch die Erkennungsquoten deutlich schlechter werden. Die Parameter für die Author-Recall-Ermittlung sind: 250 Stichproben bei 400 MFWs, im Vergleichskorpus befinden sich neben dem Autorvergleichstext 24 Texte von anderen Autoren. Die False-Positives werden hier mit je zwei Texten von allen anderen 31 Autoren als Ratetexte in 32 Tests gegen das Vergleichskorpus ermittelt, in dem neben 24 Texten von anderen Autoren reihum alle 32 Autoren mit einem Autorvergleichstext vertreten waren (Parameter: Bag-of-Words mit 2.000 Wortformen, 400 MFWs, 100 Stichproben).

	Author Recall	Other Recall 1. Ratetextautor im Vergleichskorpus vorhanden	Other Recall 2. Ratetextautor im Vergleichskorpus nicht vorhanden
MFW400 (Baseline)	0,48	0,89	0,89
A	0,55	0,91	0,86
B	0,56	0,92	0,89
C	0,69	0,77	0,67
D	0,57	0,88	0,85

Auch hier übertrifft der Gewinn bei der Verbesserung der Erkennungsquote die Verschlechterung bei den False-Positives. Wenn sich der fragliche Autor im Vergleichskorpus befindet, ergibt sich sogar bei den False-Positives bei A) und B) eine leichte Verbesserung. Bei C) ist die Verschlechterung bei den False-Positives problematisch.

5. Fazit

Die Gute-Wörter-Verfahren führen zu besseren Erkennungsquoten; in geringem Umfang treten mehr False-Positives auf, doch sind die positiven Effekte größer als die negativen. Die detaillierten Zahlen, für die hier kein Platz verfügbar ist, zeigen, dass die False-Positives je nach Autor erheblich schwanken. In einem Anwendungsfall kann vorab getestet werden, ob es bei dem fraglichen Autor zu erhöhten False-Positives kommt – um zu prüfen, ob das Verfahren bei diesem Autor gut anwendbar ist. Künftig wollen wir untersuchen, ob ein mittels Maschinellem Lernen erstelltes autorspezifisches Vokabular zu weiteren Verbesserungen führt.

Fußnoten

1. Verwendet werden alle Romane aus dem Textgrid-Repository bzw. von gutenberg.org, bei denen mindestens sechs Texte eines Autors vorliegen. Aussortiert wurden Novellensammlungen etc. sowie dialektal geprägte Texte. Es verbleiben 280 Texte von 32 Autoren.
2. Eine ausführliche Darstellung des Evaluationsverfahrens findet sich in Dimpel 2017b.
3. Diese Parameter haben sich in Prätests als vielversprechend erwiesen. Mit Blick auf die Rechenzeit muss es zunächst bei vier Varianten bleiben – für die False-Positives-Experimente in Testreihe 3 war pro Wert eine mehrtägige Rechenzeit nötig.

Bibliographie

Büttner, Andreas / Dimpel, Friedrich Michael / Evert, Stefan / Jannidis, Fotis / Pielström, Steffen / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (2016): *„Delta“ in der stilometrischen Autorschaftsattribution, in: Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma*. Konferenzabstracts zur DHd-Tagung 2016 in Leipzig, http://dhd2016.de/, S. 61–74

Büttner, Andreas / Dimpel, Friedrich Michael / Evert, Stefan / Jannidis, Fotis / Pielström, Steffen / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (2017): *„„Delta“ in der stilometrischen Autorschaftsattribution“*, in: Zeitschrift für digitale Geisteswissenschaft 2017. DOI: 10.17175/2017_006.

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 (2016): *„Burrows' Delta im Mittelalter: Wilde Graphien und metrische Analysedaten“*, in: Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts zur DHd-Tagung 2016 in Leipzig, http://dhd2016.de/: 65–70.

Dimpel, Friedrich Michael (2017a): *„Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch“*, in **Stolz, Michael (Hrsg.)**: *Konferenzabstracts DHd 2017 Bern. Digitale Nachhaltigkeit*. Bern: 100-103. http://www.dhd2017.ch/programm .

Dimpel, Friedrich Michael (2017b): *„Ein Delta-Rätsel: Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an?“*, in: Dariah-de-Working Papers 25, 2017, resolver.sub.uni-goettingen.de/purl/, http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2017-5-1 .

Dimpel, Friedrich Michael (2018a): *Die guten ins Töpfchen: Zur Anwendbarkeit von Burrows' Delta bei kurzen mittelhochdeutschen Texten nebst eines Attributionstests zu Konrads ‚Halber Birne'*, in: **Georg Vogeler (Hg.)**: *Kritik der digitalen Vernunft. Konferenzabstracts DHd 2018*, Köln 2018, http://dhd2018.uni-koeln.de, S. 168–173.

Dimpel, Friedrich Michael (2018b): *„Stabile Autorschaft trotz handschriftlicher Varianz? Die Erfolgsquote von Burrows' Delta bei nicht-normalisierten mittelhochdeutschen Texten optimieren“*, in: ZfdA 147, 2018, S. 341-363.

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.

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, CO: Association for Computational Linguistics: 79–88. 10.5281/zenodo.18177. http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [Abruf 20.8.2015].

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Reger, Isabella/ Schöch, Christof / Vitt, Thorsten (2016): *„Burrows Delta verstehen“*, in: *Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma*. Konferenzabstracts zur DHd-Tagung 2016 in Leipzig, http://dhd2016.de/: 61–65.

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*. New York: 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_empirical_.html .

Schöch, Christof (2014): *„Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“* in: **Christof Schöch und Lars Schneider (Hrsg.)**: *Literaturwissenschaft im digitalen Medienwandel*, Berlin (Philologie im Netz, Beiheft 7): 130–157.

Herausforderungen für die Klassifikation historischer Buchillustrationen Überlegungen am Beispiel retrodigitalisierter Kinder- und Jugendsachbücher des 19. Jahrhunderts

Helm, Wiebke
wiebke.helm@uni-leipzig.de
Universität Leipzig, Deutschland

Im, Chanjong
imchan@uni-hildesheim.de
Universität Hildesheim, Deutschland

Mandl, Thomas
mandl@uni-hildesheim.de
Universität Hildesheim, Deutschland

Schmideler, Sebastian
sebastian.schmideler@uni-leipzig.de
Universität Leipzig, Deutschland

Zusammenfassung

Die maschinelle Verarbeitung einer großen Anzahl von Abbildungen eröffnet unter anderem für Forschungen im Bereich der Kunstgeschichte neue Chancen. Obwohl eine Vielzahl von Buchillustrationen bereits digitalisiert vorliegt, werden diese in den Digital Humanities bisher noch wenig beachtet. Das „Distant Viewing"-Verfahren soll es ermöglichen, nach Ähnlichkeiten von Bildern zu suchen oder Bildinhalte mittels Objekterkennung zu identifizieren. Der folgende Beitrag stellt grundlegende technische Herausforderungen ins Zentrum, die sich bei der automatischen Klassifikation von Illustrationen in einem Korpus von historischen Kinder- und Jugendbüchern des 19. Jahrhunderts ergeben haben. Neben der genauen Lokalisierung, also dem Erkennen der Lage von Illustrationen auf einer Buchseite erwies sich die korrekte Bestimmung einzelner Bildobjekte als schwierig. Für die Objekterkennung stehen zwar derzeit leistungsfähige Systeme zur Verfügung, doch wurden diese mit Fotografien natürlicher Objekte optimiert, wodurch sie sich nur bedingt für die Anwendung auf die im 19. Jahrhundert vorherrschenden Drucktechniken oder die präzise Zuweisung von fiktionalen Bildinhalten eignen. Eine weitere Schwierigkeit ergab sich durch die unterschiedlichen Formate, in denen Illustrationen angelegt sind. Tafelbilder oder Abbildungen im Text können gerahmt sein oder fließende Übergänge aufweisen und Ornamente oder Textbausteine enthalten. Darüber hinaus können auch in Initialen und ornamentalen Schmuckleisten weitere Bildmotive vorkommen. Nachfolgend sollen die Herangehensweisen an die genannten Herausforderungen umrissen, Testergebnisse diskutiert und Lösungsansätze vorgeschlagen werden.

Einleitung

Die Digital Humanities entwickeln sich kontinuierlich weiter und verarbeiten längst nicht mehr nur ausschließlich Texte (Kohle 2013). Doch widmen sich nur wenige Studien der massenhaften Analyse von Abbildungen, was unter anderem auf die innerhalb der (kunstgeschichtlichen) Forschungscommunity bestehenden Vorbehalte und Kontroversen, inwieweit digitale Unterstützung das Selbstverständnis der Disziplinen verändert, zurückzuführen ist.

Abbildung aus Körner, Friedrich (Hrsg.): Das Buch der Welt: Wanderungen nach Nord und Süd, Ost und West, zu den Wohnstätten der Gesittung und den Bewohnern der Wildniß. Bd. 1: Die alte Welt, Leipzig: Spamer 1855: 101; urn:nbn:de:gbv:084-18938

Die automatische Erkennung von Objekten auf Abbildungen bietet ein nicht zu unterschätzendes Potential für künftige Forschungsvorhaben, kann doch mithilfe von Software ein weitaus größerer Korpus von Bildern untersucht werden als es dem menschlichen Betrachter allein möglich ist. Beispielsweise können so Analysen zur Häufigkeit von Objekten oder quantitative Erkenntnisse zur Motiv-Geschichte einer bestimmten Illustration in größerem Umfang durchgeführt werden.

Die Verbreitung von visuellem Material erlebte im 19. Jahrhundert einen rasanten Anstieg. Die Zunahme von Wissen vermittelnden Illustrationen in Kinder- und Jugendsachbüchern jener Epoche bildet diesen Prozess, der die Popularisierung von neuen Erkenntnissen und Entdeckungen beförderte, eindrücklich ab (vgl. Ries 1992, Schmideler 2014). Das Bildmaterial dieser Publikationen bietet daher eine ideale Voraussetzung für die Erforschung von Illustrationen mit Hilfe digitaler Methoden.

Abbildung aus Voltz, Johann Michael (Ill.) (1815): Bilderbogenbuch, Nürnberg: Campe; urn:nbn:de:gbv:084-12012713339

In den vergangenen Jahren wurden verschiedene Werkzeuge zur Klassifikation von Objekten in Abbildungen entwickelt und optimiert. Vor allem Ansätze aus Deep Learning Prozessen trugen dazu bei, für die Bilderkennung relevante Merkmale von Bildmotiven in Abhängigkeit von der Aufgabenstellung selbstständig herauszufiltern. Hier sind es insbesondere die aus mehreren Schichten bestehenden neuronalen Netzwerke (Convolutional Neural Networks, CNNs) (Krig 2016), die bei vielen Benchmarks hervorragende Werte aufweisen. Auf CNNs basieren auch die im Projekt getesteten Werkzeuge Yolo und SDS, die per Download installiert und lokal genutzt werden können. Sie sind auf Fotografien trainiert, können zahlreiche Arten von Objekten erkennen und weisen die untersuchten Objekte durch eine Markierung, ein Rechteck, aus (z.B. Redmon et al. 2016).

Fragestellungen zu Illustrationen aus der DH-Forschung

Innerhalb der Digital Humanities beschäftigen sich, wie eingangs erwähnt, nur wenige Forschungsprojekte mit Fragestellungen zu Illustrationen. Aufsehen erregte die erfolgreiche Klassifikation von zahlreichen Gemälden nach Künstler und Genre, bei der neuronale Netzwerke eingesetzt wurden (Saleh / Elgammal 2015).

In einem weiteren Ansatz wurde versucht, der Stil und der Inhalt von Gemälden zu trennen, um den Stil eines Künstlers auf andere Inhalte übertragen zu können (Gatys et al. 2015). Die Studie erlaubt allerdings nur eine subjektive Bewertung der Ergebnisse.

Tiefergehende, auf Algorithmen gesteuerte Analysen wie in der Studie von Yarlagadda et al. (2013) versuchen auf einer Illustration die Geste einer Person zu erkennen, um daraus Rückschlüsse auf den damit verbundenen Rechtsakt eines Herrschers ziehen zu können.

Ein von der Projektgruppe erstellter Korpus von illustrierten Kinder- und Jugendsachbüchern des 19. Jahrhunderts repräsentiert einen Ausschnitt von verschiedenen Reproduktionsverfahren, die im Untersuchungszeitraum zur Herstellung von Buchillustrationen zur Anwendung kamen. Der Fokus wurde auf einige wenige ausgewählte Druckverfahren - Holzstich, Kupferstich und Feder- bzw. Kreidelithographie - gelegt. Es wurde ein System zur automatischen Klassifizierung dieser Verfahren erstellt, das allerdings noch kein zufriedenstellendes Ergebnis liefern konnte (Im et al. 2018).

Herausforderungen und Lösungsansätze

Werkzeuge zur Bilderkennung sollten zur korrekten Identifikation von Objekten bei einer großen Anzahl von Illustrationen führen. Aber wie verhält sich beispielsweise auf Fotografien trainierte Software bei der Anwendung auf Illustrationen, die in einer davon abweichenden reprografischen Technik angefertigt wurden und/oder weniger bekannte beziehungsweise untypische Bildobjekte ausweisen?

Bei einem Test an ca. 200 illustrierten historischen Kinderbüchern aus der retrodigitalisierten Kollektion des Sammlers Karl Hobrecker - digital bereitgestellt von der UB der TU Braunschweig - wurden alle als Bild erkannten Objekte mit dem Tool Yolo untersucht. Das Analyseergebnis machte deutlich, dass die Qualität des Werkzeuges zur Objekterkennung derzeit noch nicht ausreicht, um Bildmotive mit hoher Wahrscheinlichkeit für eine größere Kollektion zu verfolgen, da der Domain-Shift von Fotos zu Illustrationen zu einer Verschlechterung der Erkennungsleistung führt. Diese könnte durch umfangreiche Lerndaten verbessert werden, doch scheint dies keine tragfähige Lösung für Forschungsarbeiten im Rahmen der Digital Humanities zu sein, da selten einfache Objekte im Vordergrund stehen, sondern in der Regel Kompositionen oder komplexe Bildstrukturen sowie fiktionale Motive vorliegen. Somit kann die eindeutige Deklaration von Einzelmotiven lediglich der Klärung von Zwischenfragen im Verlauf eines Forschungsprozesses dienen. Auch aus Kostengründen können oftmals keine neuen Trainingsmengen für heterogene Fragestellungen erzeugt werden. Deshalb bietet sich das Nachtrainieren von bereits an Fotografien optimierten Systemen wie VGG-19 oder GoogleNet an.

Von Problemen bei der Erkennung von Objekten in Gemälden berichten beispielsweise auch Crowley und Zisserman (2016), welche ein CNN-Erkennungssystem auf Kunstwerke des 19. und 20. Jahrhunderts angewendet haben. Sehr kleine Objekte wie zum Beispiel winzige Tiere oder Flugzeuge im Bildhintergrund wurden kaum erkannt. Erst die Entwicklung eines zweistufigen Verfahrens führte zu verbesserten Ergebnissen.

In der Wissen vermittelnden Literatur - in Fachbüchern ebenso wie in populärwissenschaftlicher Literatur oder Kinder- und Jugendsachbüchern des 19. Jahrhunderts - werden Tiere teils in ungewöhnlichen Perspektiven dargestellt, die sich so nicht auf Fotos wiederfinden. Beispielsweise wird eine Fledermaus in der Draufsicht mit ausgestreckten Flügeln gezeigt. Auch werden Tiere selten in einem realistischen Größenverhältnis wiedergegeben. Das kann u.a. kompositionspragmatische Gründe haben. Somit ist das implizite Wissen eines vortrainierten Klassifikationssystems über Größen von Tieren in diesem Fall nicht weiterführend.

Die genutzte vortrainierte Version von Yolo eignet sich für die Erkennung von ca. 9.000 Klas-sen. Davon wurden in 1.891 Abbildungen aus 168 Titeln mit ca. 16.000 Seiten der digitalen Hobrecker-Kollektion insgesamt 4.600 Objekte aus 200 Klassen erkannt. Die häufigsten zeigt die nachfolgende Tabelle.

Objekt	Anzahl
Person	1.932
Organism	248
Artefact	188
Living thing	160
Entertainer	129
Animal	120
Worker	117
Bird	112
Horse	96

Tabelle: Frequenz der am häufigsten erkannten Bildobjekte mit Yolo

Es ist geplant, die Bilder zu einzelnen Klassen einer Inspektion zu unterziehen, um so einen Einblick in die Qualität der Klassifikation zu gewinnen.

In Kinderbüchern treten häufig anthropomorphisierte Tierfiguren auf, die den mit Fotografien trainierten Algorithmen nicht bekannt sind. Moderne Objekte wie Autos oder Baseball-Schläger werden hingegen gut erkannt, sie sind aber irrelevant für historische Illustrationen. Andererseits können Treffer in solchen Klassen als Fehler betrachtet werden, die bei der Verbesserung und Weiterentwicklung der Werkzeuge helfen können.

Die Erkennung von Objekten in Buchillustrationen birgt aber über die bereits genannten noch weitere technische Herausforderungen. In den digitalisierten Bibliotheksbeständen können Illustrationen durch Werkzeuge der Schrifterkennung identifiziert und ihre Position ausgezeichnet werden. Bei der Hobrecker-Kollektion wie auch bei einzelnen Buchdigitalisaten aus anderen Institutionen zeigte sich aber, dass diese Auszeichnung zu ungenau ist und häufig nicht die notwendige Qualität erreicht. Darum muss die Identifikation von Illustrationen einer Buchseite zunächst technisch gelöst werden. Dafür hat sich der Einsatz von CNNs bewährt. Kleinere Trainingsmengen sind in diesem Fall ausreichend, wie beispielsweise eine Data Challenge alter Handschriften belegt, aufgrund derer die Erkennung unterschiedlicher Inhaltsbereiche verbessert werden konnte (Mehri et al. 2017).

Eine weitere Herausforderung stellt bereits der Aufbau einer Trainingsmenge zur Erkennung von Abbildungen auf digitalisierten Buchseiten dar: So werden beispielsweise rechteckige Abbildungen mit und ohne Rahmen getrennt in zwei Klassen erfasst. Doch treten auch Abbildungen auf, die kein einfaches quadratisches Format aufweisen, wie die nachstehenden beiden Abbildungen belegen.

Abbildung aus Braun, Ferdinand: Braun, Ferdinand: Der junge Mathematiker und Naturforscher : Einführung in die Geheimnisse der Zahl und Wunder der Rechenkunst ; eine Anleitung zu aufmerksamer Naturbetrachtung, begleitet von zahlreichen Aufgaben zur Uebung des Urtheils und der Anschauung, Leipzig: Spamer 1876: 259; urn:nbn:de:gbv:084-14052311261

Eine weitere Schwierigkeit bei der Klassifikation von Bildern in retrodigitalisierten Kinder- und Jugendsachbüchern des 19. Jahrhunderts stellen doppelseitige oder faltbare Bildseiten dar. Für deren automatische Erfassung ist ein zusätzlicher Arbeitsschritt notwendig, der je zwei Druckseiten zusammenfasst und prüft, ob es sich um ein doppelseitiges Bild handelt, da die Auszeichnung der digitalen Daten auch hierzu keine Anhaltspunkte liefert.

Bei der automatischen Bilderkennung von Buchillustrationen ist der ornamentale Schmuck, der die Ränder der Buchseiten ziert oder die Abbildungen rahmt, ebenfalls zu berücksichtigen, da auf diese Weise nicht nur Aussagen zur Verbreitung von Bildmotiven erfasst werden. Auch für Fragen zur Stilistik von Bildern erweist sich dieses Element als relevant. Eher selten kommen in der untersuchten Kollektion gestaltete Initialen vor. Hierbei handelt es sich um farbig abgehobene Buchstaben (Lombarden) oder um künstlerisch aufwändig gestaltete Initialen (z.B. Figureninitiale). Gerade letztgenannte Bildelemente sollten einbezogen werden, da sie beispielsweise in ABC-Büchern eine wichtige Funktion bei der Wissensvermittlung übernehmen.

Coverabbildung von Hoffmann, Franz (1873): Land- und See-Bilder in Erzählungen für die reifere Jugend: Zwei Theile in einem Bande, Stuttgart: Schmidt und Spring; https://nbn-resolving.org/urn:nbn:de:gbv:084-14110314057

Personifizierter Buchstabe. Abbildung aus: Das allergrösste Bilder-ABC, Berlin: Winckelmann [1828]; urn:nbn:de:gbv:084-13932

Ausblick

Wichtig für den zukünftigen Einsatz von „Distant Viewing"-Verfahren ist die Entwicklung von Evaluierungskonzepten, die stärker auf die Arbeitsweise und Bedarfe der Digital Humanities ausgerichtet sind und sich nicht nur an den Bewertungsmaßstäben der Bildverarbeitung orientieren. Zudem müssen geeignete Benutzungsoberflächen entwickelt werden, die den Umgang mit Ergebnissen erleichtern und Möglichkeiten zur Interaktion anbieten. Auf diese Weise könnten fehlklassifizierte Bilder sehr schnell und einfach von kompetenten Bildbetrachtern identifiziert und deren Informationen zur Verbesserung bestehender Systeme beitragen. Somit sind die vorgestellten Fehlklassifikationen der automatischen Bilderkennung nicht ausschließlich als Negativkriterium zu bewerten, sondern vielmehr als eine Chance zur Weiterentwicklung des „Distant Viewing"-Verfahrens wahrzunehmen.

Danksagung

Die Autoren danken für die Förderung durch die Fritz Thyssen Stiftung für das Projekt „Entwicklung der Bildikonographie in Wissen vermittelnder Kinder- und Jugendliteratur und Schullehrbüchern des 19. Jahrhunderts: ein Distant Viewing Ansatz". Ebenso wird der UB der TU Braunschweig für die Bereitstellung der Daten der Hobrecker-Sammlung gedankt.

Bibliographie

Crowley, Elliot / Zisserman, Andrew (2016): *"The Art of Detection"*, in: Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9913. Springer, Cham. S. 721-737

Gatys, Leon A. / Ecker, Alexander S. / Bethge, Matthias (2015): *A neural algorithm of artistic style* arXiv preprint arXiv:1508.06576.

Heuwing, Ben / Mandl, Thomas / Womser-Hacker, Christa (2016): *"Combining contextual interviews and participative design to define requirements for text analysis of historical media"*, in: Proceedings of ISIC, the Information Behaviour Conference, Zadar, Croatia, 20-23 September, 2016: Part 1. Information Research, 21 (4) http://www.informationr.net/ir/21-4/isic/isic1606.html

Im, Chanjong / Ghauri, Junaid / Rothman, John / Mandl, Thomas (2018): *"Deep Learning Approaches to Classification of Production Technology for 19th Century Books"*, in: Lernen. Wissen. Daten. Analysen (LWDA 2018) Workshop Fachgruppe Information Retrieval (FGIR 2018) August 22-24, Mannheim: 150-158. http://ceur-ws.org/Vol-2191

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

Krig, Scott (2016): *Computer Vision Metrics*. Cham: Springer.

Mehri, Maroua / Heroux, Pierre / Gomez-Krämer, Petra / Mullot, Rémy (2017): *"Texture feature benchmarking and evaluation for historical document image analysis"*, in: International Journal on Document Analysis and Recognition (IJDAR) 20 (1): 325-364.

Pech, Klaus-Ulrich (2008): *Der ökonomisch-technische Prozess und die Entwicklung der Kinder- und Jugendbuchliteratur*, in: **Brunken, Otto / Brüggemann, Theodor / Hurrelmann, Bettina / Michels-Kohlhage, Maria / Wilkending, Gisela (Hrsg.)**: *Handbuch zur Kinder- und Jugendliteratur. Von 1850 bis 1900*. Stuttgart, Weimar: J. B. Metzler 15-22.

Redmon, Joseph / Divvala, Santosh / Girshick, Ross / Farhadi, Ali (2016): *"You only look once: Unified, real-time object detection"*, in: Proceedings of the IEEE conference on computer vision and pattern recognition: 779-788.

Ries, Hans (1992): *Illustration und Illustratoren des Kinder- und Jugendbuchs im deutschsprachigen Raum 1871–1914*. Osnabrück: Wenner.

Saleh, Babak / Elgammal, Ahmed (2015): *"Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature"*, in: Computer Science > Computer Vision and Pattern Recognition, Arbeitsberichte. http://arxiv.org/abs/1505.00855

Schmideler, Sebastian (2014): *„Das bildende Bild, das unterhaltende Bild, das bewegte Bild – Zur Codalität und Medialität in der Wissen vermittelnden Kinder- und Jugendliteratur des 18. und 19. Jahrhunderts"*, in: **Weinkauff, Gina u.a. (Hrsg.)**: *Kinder- und Jugendliteratur in Medienkontexten. Adaption – Hybridisierung – Intermedialität – Konvergenz*. Frankfurt a. M.: Peter Lang 13-26.

Yarlagadda, Pradeep / Monroy, Antonio / Carque, Bernd / Ommer, Björn (2013): *"Towards a Computer-based Understanding of Medieval Images"*, in: **Bock, Hans Georg / Jäger, Willi / Winckler, Michael J. (eds.)**: *Scientific Computing and Cultural Heritage. Contributions in Computational Humanities*. Berlin: Springer 89-97.

Herausforderungen für Thementhesauri und Sachregister-Vokabularien zur Erschließung im Kontext des digitalen Editionsprojekts Cisleithanische Ministerratsprotokolle

Kurz, Stephan
stephan.kurz@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Zaytseva, Ksenia
ksenia.zaytseva@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Einführung

Gegenstand des umliegenden Projekts ist die Umstellung eines Langzeitvorhabens der Österreichischen Akademie der Wissenschaften (ÖAW) auf eine XML-basierte digitale Edition nach den Maßgaben der Vorschläge der Text Encoding Initiative (TEI).[1] Der Ministerrat war das zentrale Organ der Regierungstätigkeit in der Habsburgermonarchie beziehungsweise (nach 1867) in Österreich-Ungarn.[2] Seine Sitzungsprotokolle präsentieren alle Facetten staatlichen Lebens, von Fragen der Struktur und der Organisation des Staates bis zu gesellschaftlichen, wirtschaftlichen und technischen Entwicklungen sowie kulturellen und sozialen Problemen. Die Edition der Ministerratsprotokolle, die seit den 1970er Jahren unter wechselnder editorischer Verantwortung erscheint und in der ersten Serie 28 gedruckte Bände mit rund 18000 Druckseiten vorzuweisen hat, wird aktuell am Institut für Neuzeit- und Zeitgeschichtsforschung der ÖAW auf ein TEI-basiertes Editionsverfahren umgestellt, das die zu edierenden Bände in einer Print- und einer Online-Komponente aus einer Quelle speist.[3]

Die wissenschaftliche Edition der Ministerratsprotokolle zeichnet sich aus durch:

- den Umfang der zu edierenden Quellen
  - die große abzudeckende Zeitspanne von 1848 bis 1867 für die retrodigitalisierte erste Serie; von 1867 bis 1918 für die digitale Edition;
  - als Folge des umfassenden Zeitraums die Varianz etwa in den Schreibungen (mit Konsequenzen für die Textkritik, aber auch für die Arbeit mit Named Entities)
- die unterschiedliche Provenienz der Quellen
  - teilweise durch den Justizpalastbrand 1927 beschädigte Protokolle („Brandakten")
  - Ergänzungsmaterial aus verschiedenen Archivbeständen
- vielfältige interne und externe Bezüge
  - zwischen Tagesordnungspunkten und Sitzungen, aber auch z.B. bandübergreifend
  - zwischen ediertem Text und wissenschaftlichem Kommentar
  - zu den Registern und Verzeichnissen, ergänzt durch Normdaten und Linked (Open) Data

Scope der Posterdokumentation

Das Poster präsentiert vor dem Hintergrund des Gesamtworkflows für die TEI-Edition eine Schlüsselstelle auf dem Weg zu einer digitalen Edition, die den mit dem Begriff verbundenen Anforderungen gerecht werden soll – die „Registerdaten":

- Einerseits werden die Register der bereits edierten Bände rückerschlossen und ergänzt um neue Editionsdaten als Named Entities in einer relationalen Datenbank abgelegt, die das APIS-Projekt[4] als biographische, prosopographische und mit Start- und Enddaten versehene Entitäten modelliert;
- andererseits werden die Sachregister in Form eines Thementhesaurus modelliert, der Politikbereiche sowie die administrative Gliederung der Habsburgermonarchie berücksichtigt.

Herausforderungen

Prüfsteine bestehen betreffend die Modellierung und Erstellung von notwendigen Auxiliardaten, wie etwa Named Entities mit diachroner Varianz (Beispiel: Namensänderung durch Personenstands- oder Standesänderungen; Umgang mit Mehrsprachigkeit bei Ortsnamen, wenn sie

Zuordnungsveränderungen unterliegen), vor allem aber die Komplexität von Sachregister-Vokabularien und -Thesauri.

Methoden und Werkzeuge

Bei der Bearbeitung werden gängige Standards aus den Bereichen Semantic Web und Linked Open Data befolgt. Alle kontrollierten Vokabularien folgen dem Datenmodell von Simple Knowledge Organization System (SKOS),[5] um Wiederverwendbarkeit und Interoperabilität mit Daten in anderen Projekten zu gewährleisten. Die generierten Daten werden mit Dublin Core-Metadaten versehen und in ARCHE (A Resource Centre for the HumanitiEs),[6] dem geisteswissenschaftlichen Repositorium der Österreichischen Akademie der Wissenschaften, archiviert.

Linked-Data-Anbindung

Folgende Sets von Linked Open Data-Datensätzen werden, so vorhanden, zur Anreicherung der Ministerratsprotokolltexte verwendet:

- Orte: GeoNames-IDs[7]
- Normdaten für Personen, Körperschaften, Ministerkonferenzen und Veranstaltungen, Geografische Entitäten, aus der Gemeinsamen Normdatei (GND)[8]
- Sachregister: kontrollierte Vokabularien, nach Möglichkeit Anbindung an den GEneral Multilingual Environmental Thesaurus (GEMET)[9]
- Prosopographische und biographische Daten: APIS-Datensatz zum Österreichischen Biographischen Lexikon (ab 2019 unter CC-Lizenz).

Dokumentiert und im Poster zur Diskussion gestellt wird der zum Zeitpunkt der DHd 2019 aktuelle Stand der Arbeiten am Editionsprojekt, inklusive der bisherigen Recherchen zu parallel gelagerten Editionsprojekten (u.a.: Die Protokolle des Bayerischen Ministerrats 1945–1962 Online)[10] und einer Einladung zur Nachnutzung der entstehenden Daten.

Fußnoten

1. https://www.tei-c.org/ . [XML: eXtended Markup Language]
2. Vgl. den Einleitungsband zur Serie 1, insbes. darin Rumpler, Helmut: Ministerrat und Ministerratsprotokolle 1848 bis 1867. Wien: ÖBV 1970, S. 11-108.
3. Zum aktuellen Stand bei der Edition vgl. https://www.oeaw.ac.at/inz/forschungsbereiche/kulturelles-erbe/forschung/ministerratsprotokolle-habsburgermonarchie/ .
4. Vgl. http://apis.acdh.oeaw.ac.at/ ; APIS steht für Austrian Prosopographical Information System; das 2019 auslaufende Projekt wurde in Kooperation zwischen dem Austrian Centre for Digital Humanities (ACDH), dem Institut für Stadt- und Regionalforschung (ISR) und dem Institut für Neuzeit- und Zeitgeschichtsforschung (INZ) – alle an der Österreichischen Akademie der Wissenschaften – entwickelt und bildete in einem ersten Schritt prosopographische Daten aus dem Österreichischen Biographischen Lexikon (ÖBL) als Set von Relationen zwischen persons, places, institutions, events und works ab. Modelliert als relationale Datenbank, ermöglicht APIS die Darstellung der Entitäten und Relationen in mehreren Exportformaten, die Linked-Open-Data-fähig sind, darunter neben diversen Serialisierungen auch RDF und [TEI]-XML.
5. Vgl. Miles/Bechhofer 2009.
6. https://www.oeaw.ac.at/acdh/tools/arche/ ; das Repositorium wird betrieben von ÖAW-ACDH und dem Rechenzentrum der Österreichischen Akademie der Wissenschaften (ARZ) und basiert auf Fedora Commons.
7. Vgl. https://www.geonames.org/ . Vorgesehen ist auch eine Verknüpfung mit einem weiteren am ÖAW-ACDH angesiedelten Projekt, das geografische Informationsdaten mit einer historischen Tiefendimension erweitert: HistoGIS, siehe http://histogis.acdh.oeaw.ac.at/ .
8. Die von vielen überregionalen Bibliotheksverbünden des deutschsprachigen Raums betriebene Datei (verwaltet von der Deutschen Nationalbibliothek, http://portal.d-nb.info/) bietet einerseits Ressourcen, andererseits ±eindeutige Identifikatoren zur Disambiguierung. Für den deutschsprachigen Raum des 19. Jahrhunderts sind in der Mehrzahl der Fälle auch für die zweite Reihe der administrativen Rangordnung bereits zumindest Rumpfdaten vorhanden, und auch das Virtual International Authority File (VIAF) der Library of Congress (ein Superset von GND) hätte diesen gegenüber keine Vorteile zu bieten. Eine Online-Abfrage bietet auch das Bibliotheksservicezentrum Baden-Württemberg, siehe http://swb.bsz-bw.de/DB=2.104/ .
9. Online-Zugang siehe http://www.eionet.europa.eu/gemet , Vorteile von GEMET sind die Verbreitung und die Verfügbarkeit (auch als SKOS/RDF).
10. Siehe http://www.bayerischer-ministerrat.de/ .

Bibliographie

Viele Herausgeber und Bearbeiter (1970–2015): *Die Protokolle des österreichischen Ministerrates 1848–1867*. Wien: mehrere Verlage. [https://hw.oeaw.ac.at/ministerrat/].

Viele HerausgeberInnen und BearbeiterInnen (1973–2018): *Die Habsburgermonarchie 1848–1918*, 12 Bände, Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Bernád, Ágoston Zénó / Gruber, Christine / Kaiser, Maximilian (2017): *Europa baut auf Biographien*. Wien: new academic press.

Fokkens, Antske / ter Braake, Serge / Ockeloen, Niels / Vossen, Piek / Legêne, Susan / Schreiber, Guus / de Boer, Victor (2017): *„BiographyNet: Extracting Relations Between People and Events“*, in: Europa baut auf Biographien, 193–223.

Heath, Tom / Bizer, Christian (2011): *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. San Rafael: Morgan & Claypool

Miles, Alistair / Bechhofer, Sean (eds.) (2009): *SKOS Simple Knowledge Organization System*, W3C Recommendation, 18 August 2009, http://www.w3.org/TR/skos-reference .

Eine ausführliche historiographische Literaturliste kann bei Bedarf zur Verfügung gestellt werden.

I like to PROV it! Ein Data Object Provenance Tool für die Digital Humanities

Mühleder, Peter
peter.muehleder@uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Hoffmann, Tracy
tracy.hoffmann@uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Rämisch, Florian
raemisch@ub.uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Der Umgang mit Forschungsdaten ist inzwischen ein wichtiger Bestandteil geisteswissenschaftlicher Forschungsprojekte (nicht nur in Digital Humanities (DH) Projekten). Diese können in zahlreichen Formen und Formaten - von einfachen Tabellendokumenten, Protokollen bis hin zu komplexen Datensets und Visualisierungen - vorliegen. Die Daten unterliegen während der Forschung häufig Veränderungsprozessen: Sie müssen aufwendig bereinigt, transformiert, kombiniert, angereichert und/oder korrigiert werden. Dieses 'Preprocessing' der Daten stellt sich oft als explorativer, dynamischer und iterativer Prozess dar, der beispielsweise immer wieder unterschiedliche Algorithmen oder Mappings einsetzt. Diese Bearbeitungsschritte erzeugen damit jeweils neue Versionen eines Datenobjekts.

Um einen Überblick zu behalten, kommen oft individuelle Lösungsstrategien zum Einsatz, welche in vielen Fällen für andere Personen schwer nachzuvollziehen sind. Die Nachvollziehbarkeit ist aber entscheidend dafür, ob die Daten später von anderen Forschenden nachgenutzt werden können. Um dieses Ziel zu erreichen sollte die Provenance von Daten (Informationen zu Herkunft, Verarbeitungsprozesse, etc.) während des Datenlebenszyklus mit erfasst werden. Sie hilft dabei, die Herkunft der Daten und deren Bearbeitung und auch eventuelle Fehler bis zum aktuellen Zustand zu verstehen. Mit anderen Worten, Provenance kann auch eine kritische Perspektive auf Daten ermöglichen (D'Ignazio and Klein 2016).

Die Provenance-Erfassung von Daten ist in verschiedenen (meist naturwissenschaftlichen und sozialwissenschaftlichen) Disziplinen bereits etabliert und Teil des Forschungsprozesses (vgl. Freire et al. 2008, Oliveira et al. 2018). Für die Nutzung im Kontext geisteswissenschaftlicher Forschung zeigt sich jedoch, dass die bereits bestehenden Lösungen oft überkomplex sind, da sie primär die Reproduzierbarkeit der Daten sicherstellen sollen (vgl. Pasquier, T. et al. 2017). Im Fall der Digital Humanities steht jedoch das Datenobjekt oft im Vordergrund: "Recording provenance information in the digital humanities [...] has to concentrate on the systematic recording of input and output data within the workflows" (Küster et al. 2011:321). Dabei ist es nicht zwingend notwendig den gesamten technischen Prozess (mit dem Ziel der Reproduktion) zu erfassen, sondern eine nachvollziehbare Erklärung der unterschiedlichen Iterationen des Datenobjekts zu liefern. In diesem Sinne betrachten wir Provenance in den DH in erster Linie als Dokumentationsaufgabe. Diese soll dabei eine Ergänzung zu oder einen ersten Schritt in Richtung komplexerer Workflow Management und Provenance Tracking Systeme darstellen.

Dieses Poster beschäftigt sich mit der Frage, wie eine derartige Dokumentation durch Forschungsdaten-Provenance in einem Digital Humanities aussehen kann und stellt ein Tool vor, welches im diggr Projekt entwickelt und eingesetzt wird. Damit werden exemplarisch mögliche Antworten auf zwei wesentliche Aspekte der Forschungsdaten-Provenance in Digital Humanities Projekten formuliert:

- Welche Informationen werden/sollten erfasst werden?
- Wie können die Informationen einfach erfasst und (menschenlesbar) wieder ausgegeben werden?

Provit (Rämisch & Mühleder 2018) wurde als ein Tool entwickelt, das einzelnen Forschenden und kleinen Gruppen helfen soll, in datengetriebenen, experimentellen Projekten den Überblick über Datentransformationsschritte zu behalten. Das Ziel von provit ist es, ein einfach zu bedienendes Tool sowohl für die manuelle als auch für die (semi-)automatische Erfassung von Provenance zur Verfügung zu stellen. Diese Provenance-Informationen sollen dabei langfristig (unabhängig von projektspezifischen Infrastrukturen) verfügbar sein. Es erstellt dafür für das jeweilige Datenobjekt (Datei) einen auf das PROV-O Vokabular (Labo et al. 2013) basierenden RDF Graphen, der Akteure, Aktivitäten und ihren Einfluss auf ein spezifisches Datenobjekten retrospektiv beschreibt. Zusätzlich werden weitere Metadaten wie Zeitpunkt der Bearbeitung und Ursprungsort der Daten erfasst. Eine Browser Applikation steht für einen einfachen und übersichtlichen Zugang zu den Provenance-Informationen bereit, welche diese aus einer Datei und der dazugehörigen Quelldatei als dynamische Timeline und Netzwerk dargestellt. Diese Darstellung dient dazu einen Überblick über die Zusammenhänge und Details der Bearbeitungsschritte der Datei zu erhalten (siehe Abb. 1) und unterstützt das Verständnis der Provenance-Daten auch für technisch weniger versierte Forschende.

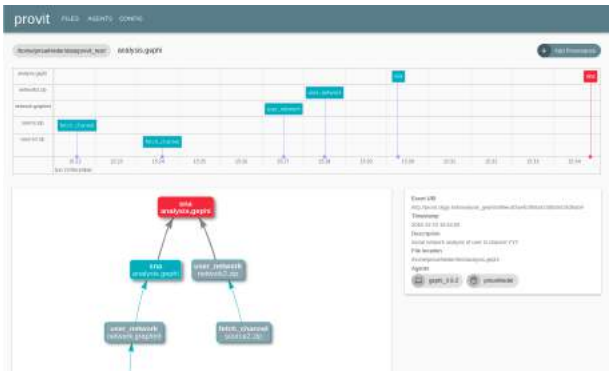


Abbildung 1. Visualisierung der Provenance-Informationen eines Datenobjektes (Provit Version 1.0 Prototyp)

Provit ermöglicht es somit auf eine standardisierte Weise, Provenance zu Forschungsdatenobjekten (unabhängig von Dateiformat und Bearbeitungsform) zu erfassen, ohne dabei auf eine komplexe technische Infrastruktur angewiesen zu sein. Die Verwendung des PROV-O Vokabulars stellt dabei die Interoperabilität sicher. Mit Provenance versehene Forschungsdaten können so dem geforderten Anspruch der Nachvollziehbarkeit Rechnung tragen und bei dem Verständnis und der Kritik von Forschungsdaten unterstützen.

Bibliographie

D'Ignazio, Catherine / Klein, Lauren F. (2016): *"Feminist Data Visualization"*, in: Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE.

Freire, Jualina / Koop, D. / Santos, E. / Silva, C. T. (2008): *"Provenance for Computational Tasks: A Survey"*, in: Computing in Science & Engineering May/June 2008, 11-21, DOI 10.1109/MCSE.2008.79.

Lebo, Timothy / Sahoo, Satya / McGuinness, Deborah (2013): *PROV-O: The PROV Ontology*. W3C Recommendation. https://www.w3.org/TR/prov-o/ [letzter Zugriff 12. Oktober 2018].

Rämisch, Florian / Mühleder, Peter (2018): *Provit* (Version v0.2.3). Zenodo, DOI http://doi.org/10.5281/zenodo.2268521

Oliveira, Wellington / De Oliveira, Daniel / Braganholo, Vanessa (2018): *"Provenance Analytics for Workflow-Based Computational Experiments: A Survey"*, in: ACM Computing Surveys 51/3, 53:1-53:25, DOI 10.1145/3184900.

Pasquier, T. / Lau, M. K. / Trisovic, A. / Boose, E. R. / Couturier, B. / Crosas, M. / Seltzer, M. (2017): *"If these data could talk"*, in: Scientific Data 4, DOI 10.1038/sdata.2017.114.

010 Jahre IDE-Schools – Erfahrungen und Entwicklungen in der außeruniversitären DH-Ausbildung

Fritze, Christiane
christiane.fritze@onb.ac.at
Institut für Dokumentologie und Editorik (IDE); Österreichische Nationalbibliothek

Fischer, Franz
franz.fischer@uni-koeln.de
Institut für Dokumentologie und Editorik (IDE); Universität zu Köln

Vogeler, Georg
georg.vogeler@uni-graz.at
Institut für Dokumentologie und Editorik (IDE); Karl-Franzens-Universität Graz

Schnöpf, Markus
schnoepf@i-d-e.de
Institut für Dokumentologie und Editorik (IDE); Berlin-Brandenburgische Akademie der Wissenschaften

Scholger, Martina
martina.scholger@uni-graz.at
Institut für Dokumentologie und Editorik (IDE); Karl-Franzens-Universität Graz

Sahle, Patrick
sahle@uni-koeln.de
Institut für Dokumentologie und Editorik (IDE); Universität zu Köln

Die Digital Humanities haben sich im Verlauf der letzten zehn Jahre aus einem randständigen Thema an den deutschen Universitäten zu einem etablierten Ausbildungsbereich verwandelt. Die seit Jahren anhaltende Diskussion um konvergente Curricula zeugt von dieser Entwicklung (Sahle 2013). Digitale Editionen waren vor zehn Jahren im deutschsprachigen Raum selten anzutreffen. In ihren Ausprägungsformen waren sie noch sehr unterschiedlich und trugen den Charakter vereinzelter Leuchtturmprojekte, die die Grenzen neuer Verfahren in den Geisteswissenschaften ausloteten. Mittlerweile ist die "digitale Editorik" ein eigener Forschungsbereich. Während Fachkenntnisse in diesem Bereich vor zehn Jahren nur in außeruniversitären Sonderveranstaltungen wie Summer Schools und Workshops erworben werden konnten, gibt es heutzutage an einigen deutschsprachigen Universitäten regelmäßige Lehrveranstaltungen zum Thema, die im Kontext der bisher entstandenen Lehrstühle[1] der Digital Humanities verortet sind.

Obwohl sich die universitäre Ausbildung in den letzten Jahren merklich und kontinuierlich verbessert hat, ist dennoch der Bedarf nach den Schools des Instituts für Dokumentologie und Editorik (IDE) ungebrochen; dies ist ein deutliches Zeichen, dass die Digital Humanities weiterhin stärker an den Universitäten verankert werden müssen. Daneben bieten die IDE-Schools ein gutes Angebot für InteressentInnen sowohl des außeruniversitären, als auch des postdoktoralen Sektors.

Deshalb bleiben komprimierte Angebote jenseits von Studiengängen wie die Veranstaltungsreihe ESU Leipzig[2] oder eine Vielzahl vereinzelter Workshops oder Summer Schools[3] die einzige Möglichkeit, sich grundlegende Kompetenzen für die von individuellen Forschungsfragen angetriebene Arbeit in den Digital Humanities anzueignen. Daneben wurden auch verschiedene Online-Angebote für E-Learning entwickelt. Für den Bereich der digitalen Editorik seien hier beispielsweise Kurse bei #dariahTeach, Schulungsmaterialien von DiXiT und DARIAH-DE oder Dokumentationen auf Webseiten oder GitHub genannt.[4]

Das IDE bietet seit 2008 regelmäßig einwöchige Schools an, die sich auf Themen rund um digitale Editionen konzentrieren. Bis 2018 wurden insgesamt 13 Schools in Wien (4), Köln (2), Chemnitz (2), Graz (2), Berlin (1), Weimar (1) und Rostock (1) mit insgesamt fast 300 TeilnehmerInnen durchgeführt.[5] Sie vermitteln wesentliche Kenntnisse für AnfängerInnen und Fortgeschrittene auf dem Gebiet der XML-basierten digitalen Editorik. In der Regel organisieren dabei lokale InteressentInnen die finanziellen und örtlichen Rahmenbedingungen und können grobe inhaltliche Vorgaben machen. Das IDE übernimmt die inhaltliche Ausgestaltung und die Auswahl des Lehrpersonals. Dabei wird in die Planung neuer Schools immer die Auswertung von Evaluationsbögen der vorangegangenen Schools einbezogen.

Das IDE legt Wert darauf, dass die TeilnehmerInnen an ihren eigenen Editionsprojekten arbeiten, um die Motivation, die eigene Arbeit konsequent auf den neuen Methoden aufzubauen, zu erhöhen. Es hält jedoch auch eigene Übungsmaterialien bereit, um den Einstieg durch gemeinsames Erarbeiten der jeweils neuen Lernstoffe sowohl an der "Tafel" und zeitgleich am eigenen Arbeitsgerät zu erleichtern.

Der erfolgreiche Besuch einer School wird stets durch ein Zertifikat bescheinigt, das in manchen Fällen als "credit points" in Studiengängen angerechnet werden konnte. Besonderes Augenmerk wird auf eine gute personelle Betreuung der TeilnehmerInnen durch zusätzliche TutorInnen und einen hohen Praxisanteil für Übungen gelegt. Die Kurse behandeln Basistechnologien wie XML, XSL, XQuery, Python, kontrollierte Vokabularien und Normdaten, editionsrelevante Kapitel der TEI, Metadaten, Text Mining, sowie allgemeine Webtechnologien wie HTML und JavaScript oder neuere Ansätze wie Graphentechnologien. Die konsequente Online-Bereitstellung von Vortragsfolien und Übungsaufgaben auf der Website des IDE ermöglicht auch nachträglich, sich Inhalte der Schools anzueignen und fügt das Angebot in die wachsende Zahl von online verfügbaren Tutorials ein (s.o.). Im Zuge der Schools wurden essentielle Technologien in Flyerform[6] kurz zusammenzufassen. Das auf den Schools vermittelte Wissen lässt sich so einerseits direkt nachnutzen, andererseits können diese Angebote auch zeitlich versetzt in andere Schools eingebunden werden.

Das Poster wird die mit den Schools gewonnenen Erfahrungen der letzten Jahre zusammenfassen. Es vergleicht die IDE-Schools mit thematisch benachbarten Veranstaltungen,[7] analysiert Trends und Konstanten der curricularen Struktur, erhebt statistische Angaben über die BesucherInnen, präsentiert die Sicht der TeilnehmerInnen durch die Auswertung einer umfassenden Befragung und verortet das Angebot im Gesamtfeld der digitalen Editorik bzw. der Digital Humanities im Allgemeinen. Es leistet damit einen Beitrag für die Untersuchung der Vermittlungsformen und Lehrinhalte in der Ausbildungslandschaft der Digital Humanities außerhalb ordentlicher Studiengänge und die Auswirkungen dieser Ausbildungsformen auf Forschung und Karriere.

Fußnoten

1. https://dhd-blog.org/?p=6174
2. http://www.culingtec.uni-leipzig.de/ESU_C_T/node/97
3. Siehe z.B. Digital Humanities at Oxford Summer School (https://digital.humanities.ox.ac.uk/dhoxss/), Digital Humanities Summer Institute der University of Victoria (http://www.dhsi.org/courses.php).
4. Siehe https://teach.dariah.eu/ ; zu den Schulungsmaterialien im Rahmen des Marie Sklodowska Curie Doktorandenprogramms DiXiT: http://dixit.uni-koeln.de/programme/materials/ ; zu den Lehrmaterialien von DARIAH-DE siehe exemplarisch "Digitale Textedition mit TEI" von Christof Schöch: https://de.dariah.eu/tei-tutorial ; eine Workshop-Dokumentation unter https://www.lib.ncsu.edu/workshops/introduction-to-xml-and-digital-scholarly-editing-using-the-text-encoding-initiative-tei-1 , ein Github-Repository unter https://github.com/slstandish/lrbs-scholarly-editing .
5. Zur Dokumentation der Schools siehe https://www.i-d-e.de/aktivitaeten/schools/ .
6. https://www.i-d-e.de/publikationen/weitereschriften/xml-kurzreferenzen/ .
7. Zu digitalen Editionen siehe z.B. die Reihe „Edirom" in Paderborn 2013-2018 (https://ess.uni-paderborn.de/) oder als Einzelveranstaltungen Madrid "Edición digital académica" 2015 (https://extension.uned.es/actividad/idactividad/9408) und 2016 (https://formacionpermanente.uned.es/tp_actividad/idactividad/8680 , München "Digital Humanities" 2017 (https://dhmuc.hypotheses.org/summerschool-2017), Prag 2017 (https://praguebeast.hypotheses.org/program) und Grenoble 2018 (https://edeen.sciencesconf.org/).

Bibliographie

Digital Humanities als Beruf. Fortschritte auf dem Weg zu einem Curriculum. Akten der DHd-Arbeitsgruppe " Referenzcurriculum Digital Humanities ". Graz 2015.

Digital Humanities Course Registry. Dariah/Clarin 2014-2018. https://registries.clarin-dariah.eu/courses/

Fritze, Christiane / Rehbein, Malte (2012): *Hands-On Teaching Digital Humanities: A Didactic Analysis of a Summer School Course on Digital Editing*, in: **Hirsch, Brett D. (ed.):** *Digital Humanities Pedagogy: Practices, Principles and Politics [Online]*. Cambridge: Open Book Publishers. http://books.openedition.org/obp/1617

Henny, Ulrike (2012): *Digitale Editionen – Methoden und Technologien für Fortgeschrittene* [Tagungsbericht zur IDE-School, Chemnitz 2012], in: H-Soz-Kult, 11.12.2012, www.hsozkult.de/conferencereport/id/tagungsberichte-4540

Locke, Brandon T. (2017): *Digital Humanities Pedagogy as Essential Liberal Education: A Framework for Curriculum Development*, in: DHQ 11.3 (2017). http://www.digitalhumanities.org/dhq/vol/11/3/000303/000303.html

Neuber, Frederike (2015): *Spring in Graz – Sunshine and X-technologies* [Bericht zur IDE-School Graz 2015], in: DiXiT Blog 26.4.2015. https://dixit.hypotheses.org/633

Sahle, Patrick (2008): *Digitale Editionen – Methodische und technische Grundfertigkeiten* [Tagungsbericht zur IDE-School, Köln 2008], in: H-Soz-Kult, 21.11.2008, www.hsozkult.de/conferencereport/id/tagungsberichte-2353

Sahle, Patrick (2013): *DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities*.

(= DARIAH-DE Working Papers Nr. 1). Göttingen: GOEDOC. http://nbn-resolving.de/urn.nbn.de.gbv:7-dariah-2013-1-5

Korrektur von fehlerhaften OCR Ergebnissen durch automatisches Alignment mit Texten eines Korpus

Bald, Markus
markusbald92@gmail.com
Universität Würzburg, Deutschland

Damiani, Vincenzo
vincenzo.damiani@uni-wuerzburg.de
Universität Würzburg, Deutschland

Essler, Holger
holger.essler@uni-wuerzburg.de
Universität Würzburg, Deutschland

Eyeselein, Björn
bjoern.eyselein@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Bei der Transkription historischer Texte liefert die OCR (Optical Character Recognition) trotz deutlicher Verbesserungen mit Open-Source-Tools wie OCRopus 1.3 und Tesseract 3.5 bzw. 4.0 meist keine perfekten Ergebnisse. Oft ist der zu transkribierende Text jedoch schon an anderer Stelle verfügbar, ohne dass der genaue Fundort bekannt ist. Um die Nachkorrektur zu vereinfachen, muss zu dem zu transkribierenden Text der Vergleichstext gefunden und aligniert werden, um dann Abweichungen zu korrigieren (bei historischen Dokumenten ändert sich oft der Wortlaut eines Textes je nach Überlieferung in verschiedenen Quellen, so dass man nicht immer davon ausgehen kann, das jede Abweichung ein OCR-Fehler ist). Dafür stellen wir das Open-Source-Tool "OCR-Textkorpus-Aligner" vor. Es orientiert sich an der Vorgehensweise des Passim-Tools, geht aber darüber hinaus, indem u. a. der alignierte Text so aufbereitet wird, dass er auch zum Training der OCR-Software benutzt werden kann. Die Vorgehensweise besteht darin, zeilenweise mit N-Grammen (aktuell 5-Gramme) Kandidaten für ähnliche Zeilen im Vergleichstext zu generieren und die Zeile mit der höchsten Übereinstimmung als Ergebnis zurückzuliefern. Zusätzlich wird ein komfortabler Editor zur Nachkorrektur bereitgestellt. Der OCR-Textkorpus-Aligner wurde in zwei Teilprojekten im Kallimachos-Verbund-Projekt (www.kallimachos.de) erfolgreich eingesetzt: im Anagnosis-Projekt (http://www.kallimachos.de/kallimachos/index.php/Anagnosis:Main) und im Narragonien-Projekt (http://www.kallimachos.de/kallimachos/index.php/Narragonien:Main) für die Transkription von frühen griechischen Drucken und einer Druckausgabe des Narrenschiffes.

In Kap. 2 wird kurz der Stand der Forschung dargestellt und in Kap. 3 die Methoden und die Benutzungsoberfläche des Alignment-Tools "OCR-Textkorpus-Aligner" präsentiert. Kap. 4 beschreibt die Evaluationsergebnisse aus zwei Anwendungsdomänen, die in Kap. 5 diskutiert werden und Kap. 6 gibt einen Ausblick mit beabsichtigen Weiterentwicklungen.

Stand der Forschung

Das „Sequence Alignment“, eine musterbasierte Suche von ähnlichen Zeichenketten in großen Sequenzen, wird als wesentlicher Bestandteil der Bioinformatik hauptsächlich dazu verwendet, um ähnliche DNA-, RNA- und Proteinstränge zu finden, die auf strukturelle, funktionelle oder evolutionäre Beziehungen hindeuten können. Dementsprechend fokussieren sich die meisten bereits existierenden Alignment-Programme wie „Lalign“ (https://embnet.vital-it.ch/software/LALIGN_form.html) oder „BLAST“ (Altschul et al. 1997) auf die naturwissenschaftliche Anwendung, wobei sich durch die geringe Anzahl an Nukleotid-Buchstaben kaum Möglichkeiten zur Nutzung bei Textkorpora bieten. Eine Anwendung von BLAST für Textkorpora wird z. B. in (Vesanto et al. 2017) vorgestellt. Angepasste Versionen dieser Tools wie Passim (Smith et al. 2014) aus dem Leipziger "Open Philology Project", sind primär auf das Matching von längeren ähnlichen Textpassagen ausgerichtet, sodass häufig einzelne, insbesondere kurze Zeilen nicht gefunden werden können. Die Software verwendet als zentrale Komponente ein Framework namens „jAligner“ (Ahmed 2018), welches die am weitesten verbreiteten Sequence-Alignment-Algorithmen implementiert und dabei den zur Verfügung stehenden Zeichenvorrat nicht einschränkt.

Methoden

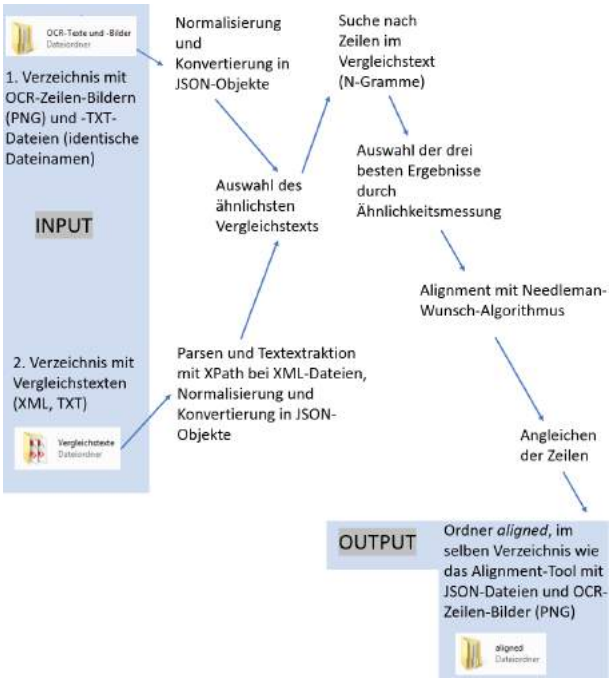


Abbildung 1: Workflow-Diagramm des Alignment-Tools vom Input (links) bis zum Output (rechts).

Der Workflow des Tools "OCR-Textkorpus-Aligner" ist in Abbildung-1 veranschaulicht. Vor dem "Sequence Alignment" werden die Zeilen durch Entfernen der Diakritika normalisiert, um die Chance zu erhöhen, zu den OCR-Zeilen jeweils passende Entsprechungen im Vergleichstext zu finden. Durch eine Ähnlichkeitsmessung der Textanfänge wird zunächst das Vergleichsdokument mit der höchsten Entsprechung vorausgewählt. Anschließend wird jede zu transkribierende Zeile (im folgende "OCR-Zeile" genannt) in N-Gramme aus fünf Zeichen segmentiert und diese im Vergleichstext (im folgenden Ground-Truth bzw. GT-Zeile genannt) gesucht. Aus lokalen Clustern von Treffern bei der N-Gramm-Suche werden Kandidaten generiert, die hinsichtlich der Anzahl der gefundenen N-Gramme und einer Ähnlichkeitsmessung bewertet werden. Aus dieser Einschätzung ergibt sich der jeweils beste Kandidat. Der globale (über die volle Länge der Zeilen alignierende) Needleman-Wunsch-Algorithmus (Needleman und Wunsch 1970) richtet die OCR-Zeile und den besten GT-Kandidaten so aufeinander aus, dass möglichst viele Zeichen übereinstimmen. Fehlende Zeichen (z. B. ein Komma) in der OCR-Zeile im Vergleich zur GT-Zeile werden durch Trennstriche (-) aufgefüllt, die Lücken markieren. Dabei wird die Länge der OCR-Zeile an die Länge der GT-Zeile angepasst. Die Ergebnisse des Alignments dienen anschließend als Input für das Korrektur-Tool.

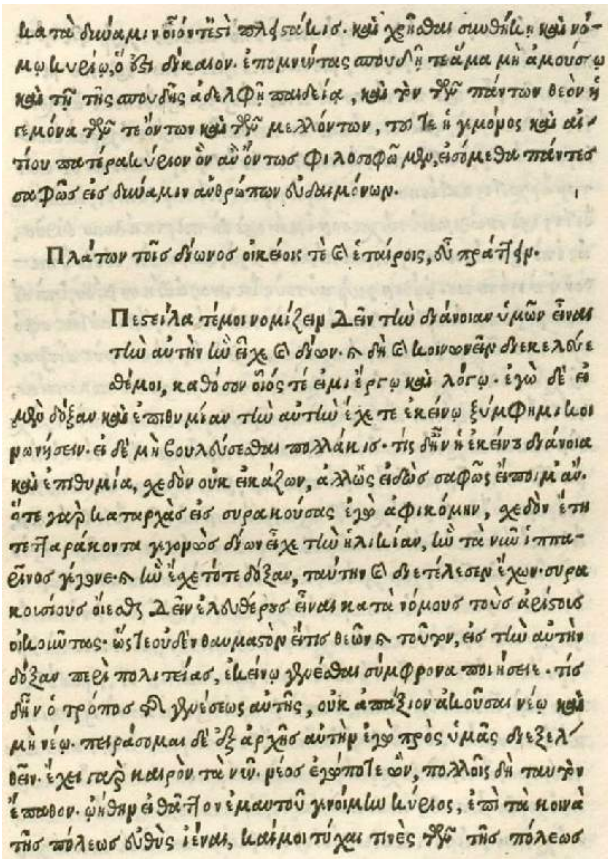


Abbildung 2: Eine Seite aus dem Band *Epistulae diversorum philosophorum, oratorum, rhetorum*, Venedig, A. Manutius, 1499 = GW 9367 [Platon, *Briefe*, 323c7-324c1].

Ein Beispiel für eine Input-Seite findet sich in Abbildung-2. Abbildung-3 zeigt die Bedienung des OCR-Textkorpus-Aligner mit Alignment anhand der ersten beiden Zeilen von Abbildung-2. Unter der Anzeige des Orginal-Scans wird oben die GT-Zeile und darunter die OCR-Zeile angezeigt. Abweichungen einzelner Zeichen sind durch rote Kästchen markiert. Die Nutzer können dann entweder das Zeichen aus der GT-Zeile in die OCR-Zeile übernehmen oder umgekehrt oder mittels einer virtuellen, konfigurierbaren Tastatur beide Zeichen durch ein anderes ersetzen. Nach einer Markierung springt die Auswahl zur nächsten abweichenden Stelle weiter. Durch Tastenkombinationen lässt sich die Markierung der Fehler noch beschleunigen. Auch lassen sich alternative Alignment-Kandidaten auswählen und komplette Zeilen editieren. Die korrigierten Zeilen lassen sich herunterladen und als Ground Truth zum Training der OCR einsetzen.

Abbildung 3: Oben: Startseite des OCR-Textkorpus-Aligner mit Auswahl des OCR-Ergebnisses und dem Ordner mit Vergleichsdateien. Mitte: Status des Alignment-Prozesses. Unten: Auswahl von den oberen beiden Zeilen aus Abbildung-2 zur Nachkorrektur in einem dreizeiligen Editor mit Originalzeile, durch Alignment ausgewählter Vergleichszeile ("GT") und OCR Zeile, wobei durch farbige Markierungen die Ground Truth festgelegt wird (ganz unten ein virtueller Editor mit domänenspezifischen Sonderzeichen).

Evaluation

Bei den beiden Evaluationen wurde jeweils der Prozentsatz der korrekt alignierten OCR-transkribierten Zeilen zu allen Zeilen des OCR-Textes berechnet. Der erste Datensatz beinhaltet eine frühe Druckausgabe von 13 Briefen Platons in altgriechischer Sprache[1], verteilt auf zwölf Seiten. Die Scanbilder (Beispiel s. Abbildung-2) wurden in 302 Zeilen segmentiert, auf die anschließend eine OCR-Erkennung mit einer Fehlerrate von ca. 15% angewandt wurde. Als Vergleichstexte dienten 793 Transkriptionen der Perseus Digital Library. Diese ständig erweiterte Open-Source-Volltextdatenbank wird von der TUFTS University/Universität Leipzig auf Initiative von Prof. Gregory R. Crane seit 1987 entwickelt. Bei Nutzung der Software Passim lag bei unseren Experimenten die Erkennungsrate im besten Fall bei 136 von 302 Zeilen und damit bei 45% (mit N-Gramm-Größe auf Wortebene = 2 und Deaktivierung von einschränkenden Parameter wie eine Mindestanzahl an N-Gramm-Matches zwischen den Texten sowie eine Mindestlänge des Alignments). Demgegenüber hatte das „OCR-Textkorpus-Aligner"-Tool 292 von 302 Zeilen (ca. 96,7%) bei einer N-Gramm-Größe von fünf Zeichen korrekt zugeordnet. Eine Fehleranalyse der zehn nicht gefundenen Zeilen ergab, dass sich darunter vier zu kurze Zeilen (mit weniger oder knapp über fünf Zeichen) und sechs Zeilen mit sehr vielen OCR-Fehlern (wegen häufigen Großbuchstaben bzw. generell schlechter OCR-Qualität) befanden,

Der zweite Datensatz entstammt einer frühen Druckausgabe des mittelalterlichen „Narrenschiff"-Text (Ausgabe Basel vom 1.3.1497 = "GW 5061"), der im Rahmen des Würzburger „Narragonien"-Projekts digitalisiert wird, wobei die OCR-Fehlerrate ca. 18% betrug. Hier wurden von 10834 OCR-Zeilen 9384 GT-Zeilen im Vergleichstext einer anderen Druckausgabe des Narrenschiffs gefunden. Dies entspricht 86,62%, was angesichts von 1758 Zeilen mit kurzen Marginalien und 312 Zeilen, die lediglich aus Seiten-Nummern bestehen, ein gutes Ergebnis ist. Die Texte wurden vor dem Alignment durch Konvertierung in Kleinbuchstaben normalisiert.

Diskussion

Je weniger Zeichen eine Zeile enthält und je höher die OCR-Fehlerrate ist, desto schlechter ist das Alignment. Um diesen Zusammenhang zu quantifizieren, haben wir Erwartungswerte unter der idealisierten Annahme der Gleichverteilung der OCR-Fehler und konstanter Länge der Zeilen (10 bzw. 20 Zeichen) berechnet (s. Abbildung-4).

In den beiden Datensätzen der Evaluation enthielten die Zeilen meist mehr als 20 Zeichen, was bei einer OCR-Fehlerrate von 18% zu einem Erwartungswert von ca. 6% bzw. 33% mit keinem bzw. nur höchstens einem N-Gramm der Länge fünf zur korrekten Vergleichszeile führt. Bei einer OCR-Fehlerrate von 15% reduzieren sich diese Erwartungswerte auf 3% bzw. 23%. Wie erwartet beziehen sich die Fehler meist auf sehr kurze Zeilen, bei denen die Erwartungswerte deutlich höher sind (s. linke Kurve in Abbildung-4). Das Alignment lässt sich durch Normalisierung (z. B. bei Sonderzeichen oder Großbuchstaben) deutlich verbessern. Wir haben auch mit 4-Grammen statt 5-Grammen experimentiert, aber das ergab in den Beispieldomänen keine Verbesserungen, sondern nur höhere Laufzeiten, kann aber in anderen Domänen sinnvoll sein.

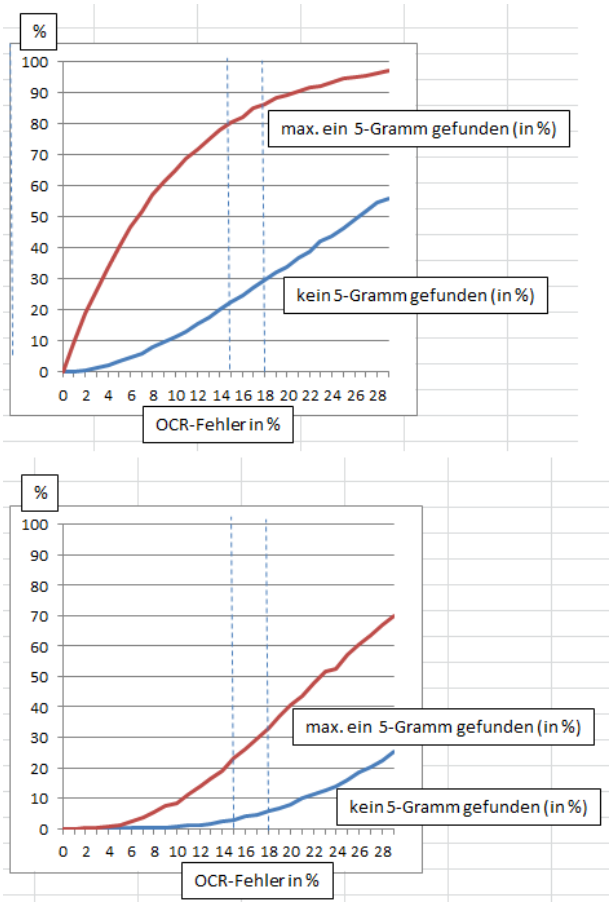


Abbildung 4: Erwartungswerte für das Finden von keinem (untere blaue Linie) bzw. höchstens einem (obere rote Linie) 5-Gramm in einem Text mit 10 (linke Kurve) bzw. 20 (rechte Kurve) Zeichen, wenn die Fehlerquote der OCR-Erkennung für die 10 bzw. 20 Zeichen von 0% bis 29% variiert wird. Die gestrichelten Linien zeigen die Erwartungswerte bei einer Fehlerquote der OCR-Erkennung von 15% und 18% an.

Zusammenfassung und Ausblick

Trotz relativ hoher OCR-Fehlerraten von 15% bzw. 18% konnten längere OCR-Zeilen relativ zuverlässig in GT-Vergleichstexten gefunden werden. Dafür reichen meist schon ein oder zwei passende N-Gramme aus. Um auch kürzere Zeilen zuordnen zu können, wollen wir die Tatsache ausnutzen, dass die Reihenfolge der Zeilen in OCR-Text und Vergleichstext meist übereinstimmt. Somit können aus der relativen Position von nicht erkannten Zeilen zu erkannten Zeilen Kandidaten für die korrekte Zuordnung generiert werden, die dann durch den Needle-Wunsch-Algorithmus auf Zeilenebene überprüft werden. Für den praktischen Gebrauch ist eine Einbindung in OCR Workflows wichtig. Dazu muss nur die Schnittstelle eingehalten werden, die auf Ordnern mit Dateien in Standard-Formaten sowie auf Namenskonventionen beruht (vgl. Abb. 1). Wir haben dies für das OCR-Framework „OCR4all“ umgesetzt, welches Algorithmen für alle Schritte von der Vorverarbeitung über die Segmentierung und die OCR einschließlich Nachtraining mit einem Editor zur Nachkorrektur bereitstellt.

Fußnoten

1. *Epistulae diversorum philosophorum, oratorum, rhetorum*, Venedig, A. Manutius, 1499 = GW 9367.

Bibliographie

Altschul, Stephen F. / Madden, Thomas L. / Alejandro A. Schäffer / Zhang, Jinghui / Zhang, Zheng / Miller, Webb / Lipman, David J. (1997): *“Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”* in Nucleic Acids Res 25 (3389-3402).

Moustafa, Ahmed (2018): *“JAligner: Open source java implementation of Smith-Waterman”* in http://jaligner.sourceforge.net .

Needleman, Saul B. / Wunsch, Christian D. (1970): *“A general method applicable to the search for similarities in the amino acid sequence of two proteins”* in Journal of Molecular Biology 48, (443–453).

Smith, David A. / Cordell, Ryan / Dillon Maddock Elizabeth / Stramp, Nick / Wilkerson, John (2014): *“Detecting and modeling local text reuse”* in Proceedings of the JCDL’14 (Joint Conference on Digital Libraries; 183–192), IEEE Press.

Vesanto, Aleksi / Nivala, Asko / Salakoski, Tapio / Salmi, Hannu / Filip, Ginter (2017): *“A system for identifying and exploring text repetition in large historical document corpora”* in Proceedings of 21. NoDaLiDa (Nordic Conference on Computational Linguistics).

Leistungsfähige und einfache Suchen in lexikografischen Datennetzen Ein interaktiv-visueller Query Builder für Property-Graphen

Meyer, Peter
meyer@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Einleitung: Property-Graphen für lexikografische Ressourcen

Klassische XML-basierte lexikografische Ressourcen können durch Graphenstrukturen mit zusätzlichen Vernetzungen und Informationen angereichert werden (Měchura 2016).[1] Dabei werden die Artikel eines Wörterbuchs zunächst durch eigenständige XML-Dokumente repräsentiert; bestimmte

XML-Elemente in diesen Dokumenten – die in typischen Anwendungsfällen z.B. den im Artikel gebuchten Wörtern oder deren Bedeutungsdefinitionen entsprechen – können dann zusätzlich in einer Graphdatenbank für Property-Graphen (vgl. Robinson / Eifrem / Webber 2013) durch Knoten verschiedener Typen repräsentiert werden. Im Redaktionsprozess können auch z.B. weitere Knoten hinzugefügt werden, um zusätzliche Information abzubilden. Kanten zwischen solchen Knoten können nicht nur bereits vorhandene relationale Informationen aus den Quelldokumenten, sondern auch zusätzliche, insbesondere auch dokumentübergreifende, Relationen zum Ausdruck bringen. Der resultierende Graph fungiert dann als ausdrucksstarke zusätzliche Navigations- und Repräsentationsebene.

Ein Query Builder für die Graphensuche

Lexikograf/innen ebenso wie Endnutzer/innen einer solchen Ressource benötigen eine Zugriffsstruktur, die Suchen nach komplexen Konstellationen in solchen Graphen ermöglicht. Für Graphendatenbanken stehen zahlreiche Abfragesprachen zur Verfügung, deren Verwendung jedoch sehr voraussetzungsreich ist. Die Entwicklung von interaktiv-visuellen Systemen zur endnutzerfreundlichen Graphenabfrage ist aktuelles Forschungsgebiet (vgl. z.B. Bhowmick / Choi / Li 2018; Pienta / Navathe / Tamersoy / Tong / Endert / Chau 2016).

Als eine auf die Bedürfnisse der digitalen Lexikografie zugeschnittene, sich insbesondere auch an interessierte Endnutzer sowie Lexikografen ohne IT-Vorkenntnisse richtende Lösung präsentiert das Poster einen visuellen Query Builder, der von den Komplexitäten der in vielen gängigen Property-Graphendatenbanksystemen implementierten Open Source-Abfragesprache Apache TinkerPop Gremlin (Rodriguez 2015; http://tinkerpop.com) abstrahiert.[2] Das Poster illustriert die Verwendung des Systems anhand einer Datenbank zu lexikalischen Entlehnungen aus dem Deutschen in andere Sprachen, die die mitunter verwickelten Entlehnungswege von Wörtern als Pfade in einem Graphen abbildet.

Abfragen werden im Browser durch das visuelle Zusammenstellen eines Baumes von Abfragekomponenten erzeugt, die Eigenschaften von Knoten beschreiben. Die so erstellten Abfragen sind zu jedem Zeitpunkt semantisch konsistent. Nach jeder Änderung an der Abfrage wird diese serverseitig in einer für Administratoren frei konfigurierbaren Weise in eine Gremlin-Graphtraversierungsanweisung umgesetzt und die Suchresultate in Echtzeit zurückgegeben. Angesichts der Mächtigkeit von Gremlin und der Möglichkeit, Traversierungen mit beliebigen Seiteneffekten zu verknüpfen, ist die direkte, manuelle Eingabe von Gremlin-Anweisungen nur in einer separaten, für Administratoren bestimmten Konsole möglich.

Im allgemeinen Fall werden *n*-Tupel von Knoten gesucht, die bestimmte Attribute aufweisen und zwischen denen nutzerdefinierte Pfade bestehen sollen. Entsprechend werden die Resultate tabellarisch als sortierbare *n*-Tupel präsentiert.

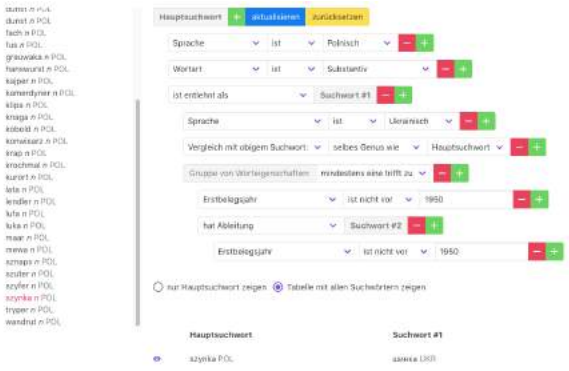


Abbildung 1. Beispiel für eine Query Builder-Suchanfrage in einem lexikografischen Netzwerk für Entlehnungsbeziehungen: Suche polnische Substantive, die als Lehnwort ohne Genuswechsel ins Ukrainische gewandert sind, wobei das Lehnwort oder eine Ableitung dazu nicht vor 1950 belegt ist.

Komplexere Abfragen

Eine Relation zwischen zwei Knoten (z.B. direkte Kante mit einem bestimmten Attribut; ein Pfad mit maximal 3 Kanten; ein Pfad beliebiger Länge) wird in einer speziellen Abfragekomponente als "relationales Quasi-Attribut" eines der beiden Knoten eingegeben; die weiteren Eigenschaften des jeweils anderen Knoten erscheinen dann auf der hierarchisch nächsttieferen Ebene unterhalb dieser Abfragekomponente, wie aus Abb. 1 ersichtlich. Das Kombinieren von Suchkriterien durch eine Boolesche Abfragekomponente ist nicht nur für echte Knotenattribute, sondern auch für solche relationalen Quasiattribute erlaubt. So sind alternative oder verbotene Pfade beschreibbar, die in Gremlin als Sub-Traversierungsroutinen verarbeitet werden müssen und in einer rein graphischen visuellen Metapher nicht mehr ohne weiteres darstellbar wären. Referenzieren anderer Knoten ist über ein sich automatisch aktualisierendes Nummerierungsschema möglich, um Sachverhalte wie "Knoten B hat einen anderen Wert für Attribut X als Knoten A" oder auch Zyklen und andere nicht-baumartige Konstellationen im Graph auszudrücken.

Schon bei kleinen Graphen können komplexere Abfragen leicht zu nicht akzeptablen Suchlaufzeiten führen (vgl. Wood 2012; Bonifati / Fletcher / Voigt / Yakovets 2018), die über Zeitbeschränkungen in der Graphtraversierung gekappt werden müssen. Durch geeignete Maßnahmen kann in vielen Fällen die Existenz weiterer Suchergebnisse festgestellt und autorisierten Nutzern über eine Warteschlange die Möglichkeit gegeben werden, ihre Suchabfrage vollständig abarbeiten zu lassen.

Für jedes gefundene Knoten- *n*-Tupel kann ein sie enthaltender Ausschnitt (Subgraph) des Gesamtgraphen angezeigt und bei entsprechender Autorisierung von lexikografischen Bearbeitern in einem frei konfigurierbaren Editor visuell redigiert werden (vgl. Abb. 2).

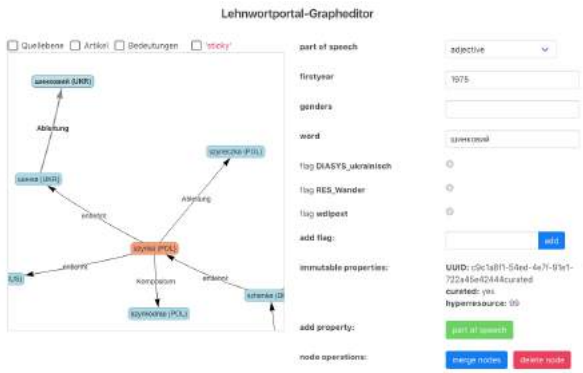


Abbildung 2. Suchergebnis mit passendem Ausschnitt aus dem Graphen und Editorfunktionalität.

Fußnoten

1. Die im Umfeld von Linked (Open) Data verwendeten Verfahren (vgl. Gracia / Kernerman / Bosque-Gil 2017) verwenden üblicherweise Graphendarstellungen lexikografischer Daten im RDF-Format, für die das hier vorgestellte, speziell für Property-Graphen und deren Abfragesprache Gremlin entwickelte Werkzeug nicht geeignet ist. Viele Arbeiten zu nutzerfreundlichen Suchwerkzeugen auf RDF-Netzen (z.B. Ferré 2017) sind jedoch für die hier behandelte Problematik sehr wohl von grundsätzlichem Interesse, weil sie in vergleichbarer Weise eine endnutzerfreundliche Zugriffsschicht über die RDF-Abfragesprache SPARQL legen.
2. Der Query Builder ist Komponente eines derzeit in Entwicklung befindlichen Open-Source-Softwaresystems zur Verwaltung und Online-Publikation graph-erweiterter lexikografischer Ressourcen (Meyer / Eppinger 2018), das im Rahmen des von der Fritz Thyssen Stiftung geförderten Projektes "Das Lehnwortportal Deutsch als Forschungs- und Publikationsplattform" entwickelt wird.

Bibliographie

Bhowmick, Sourav S. / Choi, Byron / Li, Chengkai (2018): *Human Interaction with Graphs: A Visual Querying Perspective*. San Rafael, CA: Morgan & Claypool Publishers.

Bonifati, Angela / Fletcher, George / Voigt, Hannes / Yakovets, Nikolay (2018): *Querying Graphs*. San Rafael, CA: Morgan & Claypool Publishers.

Ferré, Sébastien (2017): *"Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language"*, in: Semantic Web: Interoperability, Usability, Applicability 8(3): 405-418.

Gracia, Jorge / Kernerman, Ilan / Bosque-Gil, Julia (2017): *"Toward Linked Data-Native Dictionaries"*, in: **Kosem, Iztok / Tiberius, Carole / Jakubíček, Miloš / Kallas, Jelena / Krek, Simon / Baisa, Vít (eds.)**: *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno: Lexical Computing 550-559 https://elex.link/elex2017/proceedings-download/ [letzter Zugriff 12. Oktober 2018].

Měchura, Michal (2016): *"Data structures in lexicography: from trees to graphs"*, in: **Horák, Aleš / Rychlý, Pavel / Rambousek, Adam (eds.)**: *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016*. Brno: Tribun EU 97-104.

Meyer, Peter / Eppinger, Mirjam (2018): *"fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data"*, in: **Čibej, Jaka / Gorjanc, Vojko / Kosem, Iztok / Krek, Simon (eds.)**: *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17-21 July, Ljubljana*. Ljubljana: Znanstvena založba 1017-1022.

Pienta, Robert / Navathe, Shamkant / Tamersoy, Acar / Tong, Hanghang / Endert, Alex / Chau, Duen Horng (2016): *"VISAGE: Interactive Visual Graph Querying"*, in: AVI: Proceedings of the Workshop on Advanced Visual Interfaces 272–279.

Robinson, Ian / Eifrem, Emil / Webber, Jim (2013): *Graph Databases*. Sebastopol, CA: O'Reilly & Associates.

Rodriguez, Marko A. (2015): *"The Gremlin Graph Traversal Machine and Language"*, in: **Cheney, James / Neumann, Thomas (eds.)**: *Proceedings of the 15th Symposium on Database Programming Languages (DBPL 2015).* New York: The Association for Computing Machinery 1-10.

Wood, Peter T. (2012): *"Query Languages for Graph Databases"*, in: SIGMOD Record 41(1): 50-60.

Linked Biondo - Modelling Geographical Features in Renaissance Text and Maps

Görz, Günther
guenther.goerz@fau.de
Bibliotheca Hertziana, Roma, Italien; FAU Erlangen-Nürnberg, Deutschland

Seidl, Chiara
chiara.seidl0510@gmail.com
Bibliotheca Hertziana, Roma, Italien; FAU Erlangen-Nürnberg, Deutschland

Thiering, Martin
martin.thiering@campus.tu-berlin.de
Bibliotheca Hertziana, Roma, Italien; TU Berlin, Berlin, Deutschland

Bibliotheca Hertziana's project "Historical spaces in texts and maps" (http://biblhertz.it/en/research/research-projects-of-the-institute/historical-spaces-in-texts-and-maps-biondo-project/ ; 19.12.2018) aims at a cognitive-semantic analysis of Flavio Biondo's "Italia Illustrata" (1474) linking with contemporary maps (Goerz et al., 2018). At focus are relations between historical maps and texts aiming to explore the historical understanding of space and the knowledge associated with it. Our research combines cognitive-semantic parameters such as toponyms, landmarks, spatial frames of reference, geometric relations, gestalt principles and different perspectives with computational and cognitive linguistic analysis (Thiering, 2015). With contributing to Spatial Humanities (Bodenhamer et al., 2010)

we are convinced that generally, all maps are cognitive maps, depicting culture-specific spatial knowledge and practices (Blakemore/Harley, 1980). Biondo's mention of using non-identifiable maps gives a reason for comparing toponyms in his text and in 15th-century maps.

Recogito is being used as the main tool for static annotations of places and persons/peoples in both, text and maps. These annotations are complemented by cognitive-linguistic spatial role markups by means of the brat tool. Moreover, special emphasis is put on the narrative aspect of Biondo's text which indicates an event-based representation of movement. To achieve a deeper and more generic semantic level of linguistic and map-related annotations, we pursue the transition to an ontology-based representation. For this purpose, we are actually defining a domain ontology based with the event-based CIDOC Conceptual Reference Model (CRM, ISO standard 21127) and its spatio-temporal extension CRMgeo (http://cidoc-crm.org ; 19.12.2018) in the framework of our CRM implementation in OWL-DL, as well as appropriate mapping rules to be applied to the annotations exported in CSV and RDF formats. Using the CIDOC CRM opens up a wide spectrum of interoperability and linking to many web resources, such as the gazetteers being used with Recogito. Ontological enrichment with CRM as the top conceptual model would provide a generic "assignment event" which has open positions to be filled or linked with the semantic roles, resp., for agent, (material and immaterial) constituents, time-span, and place. This allows a semantic interpretation of annotations such that, e.g., for each tagged PlaceName we can generate an instantiated CRM description in RDF/OWL triple format, ready for publication as Linked Open Data. In the same fashion, mappings are applied to the results of spatial role labeling: These triples encode cognitive parameters, primarily "figure - trajectory/path[= spatial_relation] – ground" constructions. As a next step, the ontologically enriched representations of places and spatial relations will be combined into more complex MOVE events.

In addition to the described analytic perspective, we also pursue a synthetic view in the sense that we will use the data found by the analytic steps to reconstruct plausible cognitive sketch maps in future work.

Bibliography

Blakemore, Michael / Harley, Brian J. (1980): *Concepts in the History of Cartography - A Review and Perspective*, in: Cartographica. International Publications on Cartography, 17/4, Monograph 26. University of Toronto Press: Toronto.

Bodenhamer, David J. / Corrigan, John / Harris, Trevor M. (eds.) (2010): *The Spatial Humanities. GIS and the Future of Humanities Scholarship*. Bloomington & Indianapolis: Indiana University Press.

Görz, Günther / Geus, Klaus / Michalsky, Tanja / Thiering, Martin (2018): *Spatial Cognition in Historical Geographical Texts and Maps: Towards a cognitive-semantic analysis of Flavio Biondo's ``Italia Illustrata''*, in: **Boutoura, Ch., Tsorlini, A. (Ed.):** *Digital Approaches to Cartographic Heritage.* 13th Conference of the International Cartographic Association Commission on Cartographic Heritage into the Digital, Madrid, 29-44.

Thiering, Martin (2015): *Spatial Semiotics and Spatial Mental Models: Figure-Ground Asymmetries in Language*. Berlin: De Gruyter Mouton.

Linked Open Travel Data: Erschließung gesellschaftspolitischer Veränderungen im Osmanischen Reich im 19. Jahrh. durch ein multimediales Online-Portal

Wettlaufer, Jörg
jwettla@gwdg.de
Georg-August-Universität Göttingen, Deutschland

Kilincoglu, Deniz
dkilinc@uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Einleitung

Das Osmanische Reich war im 19. Jahrhundert ein im Übergang zur Moderne begriffenes Staatsgebilde, in dem, wie auch in Europa, politische Konzepte wie Nationalismus zunehmend Bedeutung gewannen. Aufgrund dieser Verschiebungen bildeten sich neue Identitäten aus, die weniger religiös, sondern vielmehr nationalstaatlich geprägt waren. In diesem Projekt sollen Informationen zur politischen Situation und zur Wahrnehmung dieser Transformation durch Reisende im Osmanischen Reich in diesem Zeitraum zusammengetragen und erschlossen werden. Dabei wird ein Linked Open Data Ansatz verfolgt, der die extrahierten Entitäten und Informationen als Ressourcen für multimodale Forschungsansätze zur Verfügung stellt.

Methode

Das Projekt konzentriert sich auf die Gattung der Reiseberichte (vgl. Wettlaufer 2007) europäischer Reisender im Osmanischen Reich, um die Außenperspektive auf die Verhältnisse im inneren des Osmanischen Reichs zu erfassen (Kilincoglu 2015, 2017a,b;). Hierzu werden die publizierten Berichte europäischer Reisender (vgl. Calikbasi 2004, Genc 2015; Schiffer 1999; Spackman 2017) zunächst identifiziert und in eine Zotero Bibliographie mit Primär- und Sekundärliteratur eingepflegt. Bislang konnte ca. 450 individuelle Berichte identifiziert werden.[1] Die gefundenen Berichte wurden in Hinblick auf die darin Reisenden

ausgewertet. Digitalisate stehen nach ersten Stichproben für ca. 95% der Berichte über Google Books und Hathi Trust[2] zur Verfügung und werden daraus für das Projekt nachgenutzt. Ein im Aufbau befindliches Online-Portal bietet zunächst einen Überblick zu den Reisenden, ihren Berichten und erlaubt Zugriff auf die rechtefreien Volltexten ihrer Reisebeschreibungen.[3]

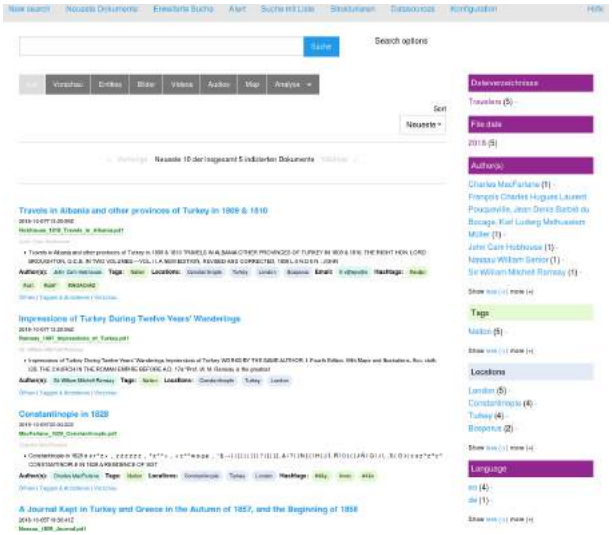


Abbildung 1. Open Semantic Desktop Search User Interface (Django) mit Beispieldaten aus dem Projekt.

Für die inhaltliche Erschließung der Berichte werden die Texte in einen Suchindex eingepflegt und relevante Entitäten extrahiert, die eine facettierte Suche in dem Apache/Solr-Index möglich machen. Hierfür wird die Suchumgebung Open Semantic Search (OSS)[4] verwendet, die eine integrierte NER[5] bietet. In einer weiteren Ausbaustufe sollen die Linked Data Kapazitäten von Open Semantic Search verwendet und zusätzliche Metadaten über die Reisenden oder die von Ihnen besuchten Orte eingebunden werden. Dabei können zielgerichtet Ortsnamen automatisiert über wikidata in die NER eingebracht werden (Abb. 2). Weil die automatisierte Generierung von Facetten und Metadaten nicht fehlerfrei möglich sein wird, ist eine manuelle Nachbearbeitung und Anreicherung der Texte notwendig. Das Projekt ist über einen Zeitraum von drei Jahre geplant und soll anschließend nachhaltig über ein geisteswissenschaftliches Forschungsdatenzentrum im Rahmen der NFDI gehostet werden.

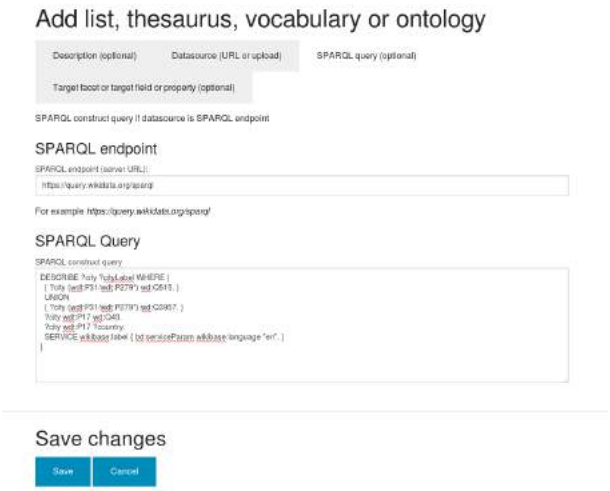


Abbildung 2. Generierung von Facetten aus Listen oder Ontologien bzw. direkt über SPARQL Abfragen.

Die so aufbereiteten Texte, Abbildungen und Metadaten sollen anschließend multimodal und -medial über die GIS Funktionalität der Plattform (ebenfalls in OSS integriert) präsentiert werden. Damit können z.B. die Itinerare der Reisenden visualisiert werden (vgl. auch Wettlaufer 2010, 2011). Zugeordnet zu einen einzelnen Stationen werden Abbildungen, die sich in den Digitalisaten finden und rechtefrei sind, angezeigt bzw. verlinkt. Dadurch wird die Reiseerfahrung plastisch nachvollziehbar und die textuelle Überlieferung in einen geographischen und medialen Kontext gesetzt.

Ausblick auf die Ergebnisse

Auf der Grundlage dieser Datenaufbereitung sollen anschließend relevante Beschreibungen der politischen Situation im Osmanischen Reich des 19. Jahrhunderts identifiziert und vergleichend analysiert werden. Einige händisch gezogene Stichproben lassen einen reichen Ertrag an Fundstellen erhoffen, die über die nationalstaatlichen Strömungen und Überlegungen dieser Zeit aus der Perspektive von Reisenden Auskunft geben.[6] Die extrahierten Metadaten zu Personen, Orten, Publikationen und Reiserouten sollen schließlich mit anderen Ressourcen wie Normdatensätzen und wikidata verknüpft und die gewonnen Erkenntnisse so als Linked Open (Travel) Data maschinenlesbar zur Verfügung gestellt werden. Darüber hinaus soll das Portal offene, reichhaltige und verlässliche Ressource für Wissenschaftler und Forschungsgruppen bieten, die sich für den mittleren Osten im Spiegel von Beschreibungen aus erster Hand im langen 19. Jahrhundert interessieren. Schließlich ist auch eine Nachnutzung der spezifischer Anpassungen in der NLP Toolchaine der OSS Suchumgebung für andere Projekte zur Erforschung von Reiseliteratur geplant.[7]

Fußnoten

1. https://www.zotero.org/groups/2185870/travelers_in_the_19th_century_ottoman_empire . In der Gesamtzahl von ca. 700 Einträgen finden sich noch eine Reihe von Dubletten (unterschiedliche Ausgaben, Übersetzungen etc.), die in einem zukünftigen Arbeitsschritt noch konsolidiert werden müssen.
2. https://books.google.de/ und https://www.hathitrust.org/ . Stichproben weisen darauf hin, dass die OCR Qualität der Digitalisate in der Regel ausreichend für Text-Mining und NER ist.
3. Siehe einen Prototypen mit einigen Beispielen und Vorarbeiten, aber ohne die geplante Text-Mining Funktionalität unter: http://middle-east-travelers.de
4. https://www.opensemanticsearch.org/
5. Spacy NER und Stanford NER können einzeln oder gemeinsam genutzt werden.
6. MacFarlane (1829), 484 über Griechenland: „and even then, on retiring from the village, and taking a solitary path that led to my abode, my ears have been delighted with the sounds of men and women's voices, of the violin or guitar attuned to strains of jollity, or to their national or patriotic airs. * Their boldness astonished me. Their favourite songs were the invocation of the unfortunate Riga, “the Sword of Colocotroni,” the “Death of Marco Bozzari,” the brave “Canaris,” &c. &c. And these I have frequently heard them singing on the Bosphorus, when Turks were within ear-shot!” Vgl. auch Senior (1859), 211.
7. Zum Beispiel für das Portal www.digiberichte.de , auf dem seit 2007 eine KWIC Suche auf Volltexten zu spätmittelalterlichen Reiseberichten angeboten wird.

Bibliographie

Calikbasi, Durdu (2004): *Das Osmanische Reich in der Darstellung deutschsprachiger Reisberichte um die Jahrhundertwende 1900*, Norderstedt: Books on Demand.

Genc, Kaya (2015): *An Istanbul Anthology: Travel Writing through the Centuries*, Cairo: The American University in Cairo Press.

Kilincoglu, Deniz T. (2015): *Economics and Capitalism in the Ottoman Empire*, London: Routledge.

Kilincoglu, Deniz T. (2017a): *Islamic Economics in the Late Ottoman Empire: Menâpirzâde Nuri Bey's Mebâhis-i İlm-i Servet.* The European Journal of the History of Economic Thought 24 (3): 528–54.

Kilincoglu, Deniz T. (2017b): *Studying Economics as War Effort: The First Economic Treatise in the Ottoman Empire and Its Militaristic Motivations*, in: **Ikeda, Yukihiro / Rosselli, Annalisa (eds.):** *War in the History of Economic Thought: Economists and the Question of War*, 78–99. Abingdon, Oxon: Routledge.

MacFarlane, Charles (1829): *Constantinople in 1828; a Residence of Sixteen Months in the Turkish Capital and Provinces: With an Account of the Present State of the Naval and Military Power, and of the Resources of the Ottoman Empire*, London: Saunders and Otley.

Schiffer, Reinhold (1999): *Oriental Panorama: British Travelers in 19th Century Turkey*, Amsterdam: Rodopi.

Senior, Nassau William (1859): *A Journal Kept in Turkey and Greece in the Autumn of 1857, and the Beginning of 1858*, London: Longman, Brown, Green, Longmans, and Roberts.

Spackman, Barbara (2017): *Accidental Orientalists: Modern Italian Travelers in Ottoman Lands*, Oxford: Oxford University Press.

Wettlaufer, Jörg (2007): *Reise- und Gesandtschaftsberichte als Quellen der Hof- und Residenzenforschung*, in: **Paravicini, Werner (hg.):** *Höfe und Residenzen im spätmittelalterlichen Reich*, Textband, , bearb. von Jan Hirschbiegel und Jörg Wettlaufer, Ostfildern: Thorbecke, 361-372 (= Residenzenforschung 15, III).

Wettlaufer, Jörg (2010): *Europäische Reiseberichte des Späten Mittelalters. Das Projekt einer Digitalisierung der Editionen und eines Themenportals im Internet [Les récits de voyageurs européens à la fin du Moyen Âge. Le projet de digitalisation des éditions et d'un portail de recherche sur internet]*, in: **Guenée, Bernard / Moeglin, Jean-Marie (hg.):** *Relations, échanges et transferts en Europe dans les derniers siècles du Moyen Âge*, Paris: Éditions de Boccard, 539-555.

Wettlaufer, Jörg (2011): *Poster zu www.digiberichte.de auf der Tagung .hist2011*, in Berlin, Humboldt-Universität, 14.-15.9.2011. http://digihum.de/scripts/download.php?File=SnJnZ3luaHNyZV8yMDExX3F2dHZvcmV2cHVnci5jcXM= [zuletzt abgerufen 12.10.2018].

Mapping the Moralized Geography of 'Paradise Lost': Prototypenbildung am Beispiel einer geo-kritischen Erschließung von Miltons Werk

Schaeben, Marcel
m.schaeben@uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland

El Khatib, Randa
elkhatib.randa@gmail.com
Electronic Textual Cultures Lab, University of Victoria, Canada

Seit Anbeginn des digitalen Zeitalters untersuchen GeisteswissenschaftlerInnen mögliche Weiterentwicklungen und Alternativen zum Medium der Printedition. Bis heute entstanden und entstehen zahlreiche innovative und komplexe digitale Editionen von Texten, die angereichert durch Hyperlinks, Suchfunktionen, umfangreiche Register u.a. neue kritische Zugänge zu etwa literarischen Werken, historischen Quellen oder verschiedensten Materialarten wie z.B. Briefkorrespondenzen oder Tagebüchern bieten. Der primäre Zugangsweg, den digitale Editionen zu ihrem Gegenstand bieten, ist jedoch nach wie vor

das Medium *Text*: ein oder mehrere Textquellen sowie der kritische oder Sachapparat. Digitale Methoden bieten GeisteswissenschaftlerInnen jedoch die Möglichkeit, mit alternativen Zugängen zum Editionsgegenstand zu experimentieren.

Das 1667 veröffentlichte epische Gedicht *Paradise Lost* von John Milton ist eine Neuerzählung des biblischen Buch Genesis, welches die Geschichte des Höllensturzes, der Versuchung von Adam und Eva und ihre Vertreibung aus dem Garten Eden enthält. Darin erwähnt Milton biblische, aber auch zahlreiche nicht-biblische Orte. Diese illustriert er durchweg mit Kommentaren und Anspielungen, die sich aus klassisch-mythologischen, biblischen sowie (Milton-)zeitgenössischen Ansichten zu diesen Orten speisen und deren Quellen weit über den biblischen Text hinausgehen. Milton zeigt eine tiefe Kenntnis „alter" sowie „neuer" Geographie, mithilfe derer er seinem Werk eine enorme raum-zeitliche Komplexität und Tiefe verleiht. Charakteristisch ist, dass Milton den Orten eine moralische Wertigkeit zuweist: die meisten Erwähnungen von Ortsnamen werden durch seinen literarischen Kommentar moralisch kontextualisiert und lassen eine positive oder negative Konnotation erkennen.

Der Prototyp einer Web-App *Exploring the Moralized Geography of Paradise Lost* soll, aufbauend auf Praktiken der *Literary Geography* im Allgemeinen sowie der *Literary Cartography* im Speziellen (Piatti 2016: 89), diese unterschiedlichen räumlichen und zeitlichen Dimensionen der literarischen Geographie *Paradise Losts* erstmals visuell explorativ zugänglich machen und dabei den Kern einer geo-kritischen Edition bilden, die sowohl im pädagogischen, als auch im geisteswissenschaftlich-epistemologischen Kontext Wirkung entfalten kann.

Abbildung 1. Hauptansicht mit eingeblendetem Kommentar und Textpassagen.

Den zentralen Zugang zum Werk bildet hierbei der räumlich-geographische. Erwähnte Orte wurden manuell geokodiert und als Marker auf eine digitale Karte gelegt. Ein besonderes Augenmerk liegt auf Miltons moralischer Kontextualisierung: positiv, negativ sowie neutral konnotierte Orte werden farblich markiert; mehrere sich widersprechende Erwähnungen eines Ortes werden durch Farbtonnuancen oder alternativ durch Kreisdiagramme, die optional an die Stelle eines Markers treten können, visualisiert (Abb. 2).

Abbildung 2. Kreisdiagramme zur Illustration mehrerer Erwähnungen eines Ortes mit unterschiedlichen moralischen Konnotationen.

Zur Illustration der unterschiedlichen Temporalitäten, die Miltons Geographie zugrunde liegen, können verschiedene historische Karten eingeblendet und über die Basis-Weltkarte gelegt werden, die durch die Technik der Georektifizierung an das Mercator-Koordinatensystem angepasst wurden. Gruppen thematisch verwandter Orte sowie die Anzahl der Erwähnungen der einzelnen Orte werden visuell hervorgehoben. Weiterhin enthält der Prototyp für jeden Ort die entsprechenden Passagen des Gedichts, in denen er erwähnt wird. Es kann zwischen vier Basis-Weltkarten gewählt werden, die je unterschiedliche topographische oder politische Charakteristiken hervortreten lassen. Ortsmarker für solche Orte, die im Buch der Genesis erwähnt werden, lassen sich optional zusätzlich einblenden, um einen Vergleich biblischer Geographie mit Miltons literarischer Geographie zu ermöglichen. Um die Dichte der Vielzahl visualisierter Faktoren zu entzerren, lassen sich bis zu vier synchronisierte Instanzen der Karte nebeneinander anzeigen, die unabhängig voneinander konfiguriert werden können (Abb. 3).

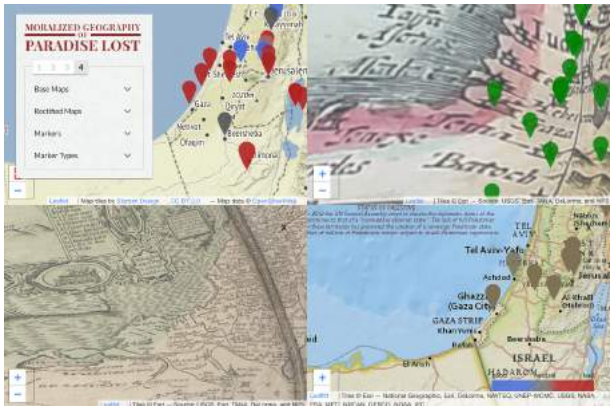


Abbildung 3. Vier unterschiedlich konfigurierte, synchronisierte Karten.

Galey und Ruecker (2010: 406) argumentieren, dass ein experimenteller, digitaler Prototyp, ähnlich wie kritische Printeditionen und andere Produkte geisteswissenschaftlicher Forschung, als eine Art Verdinglichung einer Theorie bzw. einer Hypothese verstanden werden kann. Mit unserem Prototyp wollen wir verdeutlichen, dass digitale Kartographie weit mehr ermöglicht als die reine Illustration geographischer

Zusammenhänge (Cooper und Gregory 2011: 90). Der Entwicklungsprozess eines digitalen kartenbasierten Prototyps ist ein Prozess kritischer Interpretation; sein Ergebnis wiederum versteht sich als Mittel zur weiteren kritischen Interpretation.

In unserem Prototyp wird die durch den kartographischen Ansatz zunächst postulierte Beziehung zwischen realem und literarischem Ort durch editorische Kommentare sowie unterschiedliche Visualisierungen unmittelbar problematisiert. Im Sinne einer *Deep Map* (Bodenhamer 2016: 212ff) sollen die diskursiven und ideologischen Dimensionen der *literarischen Topographie* des Werks, und damit die sozialen und kulturellen Implikationen, die in Miltons literarische Geographie eingegangen sind (Hess-Lüttich 2017: 9), zugänglich gemacht werden. Damit soll ein Beitrag zur Bildung von Argumenten, Hypothesen und Fragestellungen auf der Grundlage des räumlich-zeitlichen Aspekts des Werks geleistet werden. Die in die Entwicklung des Prototyps selbst eingegangenen Vorannahmen sollen transparent dargestellt und diskutiert werden. Durch flexible Konfigurationsmöglichkeiten im Benutzerinterface wird es BenutzerInnen ermöglicht, selbst mit unterschiedlichen Visualisierungsmöglichkeiten einzelner Faktoren zu experimentieren.

Auf unserem Poster präsentieren wir den aktuellen Stand unseres Prototyps und illustrieren Probleme und Herausforderungen, die sich bei der Visualisierung der unterschiedlichen Faktoren ergeben. Beispielhaft wird gezeigt, wie sich anhand des Prototyps für bestimmte Thesen über Miltons literarische Geographie argumentieren lässt.

Bibliographie

Bodenhamer, David J. (2016): *Making the Invisible Visible: Place, Spatial Stories and Deep Maps*, in: **Cooper, David / Donaldson, Christopher / Murrieta-Flores, Patricia (eds.):** *Literary Mapping in the Digital Age*. London / New York: Routledge 207-220.

Cooper, David / Gregory, Ian N. (2011): *Mapping the English Lake District: A Literary GIS*, in: Transactions of the Institute of British Geographers 36 (1): 89-108.

Galey, Alan / Ruecker, Stan (2010): *How a Prototype Argues*, in: Literary and Linguistic Computing 25 (4): 405-24.

Hess-Lüttich, Ernest W. B. (2017): *Spatial Turn: On the Concept of Space in Cultural Geography and Literary Theory*, in: Meta-Carto-Semiotics 5 (1): 27-37 http://ojs.meta-carto-semiotics.org/index.php/mcs/article/view/21 [letzter Zugriff 15. Oktober 2018].

Piatti, Barbara (2016): *Mapping Fiction: The Theories, Tools and Potentials of Literary Cartography*, in: **Cooper, David / Donaldson, Christopher / Murrieta-Flores, Patricia (eds.):** *Literary Mapping in the Digital Age*. London / New York: Routledge 106-119.

„Medialization follows function!“ Multimodaler Hypertext als Publikationsmedium (nicht nur) für die Geschichtswissenschaft

Wachter, Christian
christian.wachter@stud.uni-goettingen.de
Universität Göttingen, Deutschland

Wissenschaftliches Publizieren: Knowledge Design wird im Mediendesign sichtbar

Vermittlung von Wissen ist immer an den Gebrauch von Medien gebunden. Dabei haben selbige stets Einfluss auf die Semantik der Botschaften. Gerade im wissenschaftlichen Kontext gilt es, diese Wirkung der *Medienästhetik* (zum Begriff s. Schnell 2000, 2001) beim Gestalten von Publikationen zu beachten und reflektiert einzusetzen. Hier gilt schließlich Klaus Krippendorffs Leitsatz: „Design is making sense of things” (Krippendorff 2006: xiii).

So benutzen Geisteswissenschaftler*innen sinnvollerweise traditionell typografische Publikationsformen. Immerhin ermöglichen tendenziell linear gestaltete Narrative eine präzise Vermittlung logischer Argumentationsverläufe. Diese werden isomorph im Textfluss symbolisch repräsentiert und können von Leser*innen eng nachvollzogen werden. Die textsprachliche Modalität macht diese direkte Vermittlung möglich.

Hypertext erweitert den Möglichkeitsspielraum, Wissen zu medialisieren

Wenn jedoch komplexere, pluralistisch angelegte Zusammenhänge vermittelt werden sollen, leistet *Hypertext* als ein unlineares Medium eindeutig mehr: Gleichzeitigkeiten, Verflechtungen, Perspektivenpluralismus oder Anschlussfähigkeit an weitere Methoden sind insbesondere zu nennen. Sie können mit Hypertext auf eine explizite Weise vermittelt werden, wie es mit der Typografie nicht möglich ist. Dies ist die Kernaussage meines Promotionsprojektes, das ich mit dem präsentierten Poster vorstelle. Das Projekt ist als *medientheoretischer Beitrag zur Grundlagenforschung des E-Publishing* gedacht.

Dabei liegt der Fokus auf der Geschichtswissenschaft, da ich Strategien und Ziele der Wissenserzeugung innerhalb

meiner eigenen Disziplin reflektiere und davon abhebend frage, welche mediale Vermittlungsform sich im Einzelfall als adäquat erweist. Diese grundlegende Blickrichtung motiviert den Titel „Medialization follows function!“[1]. Weil in anderen Disziplinen ganz ähnliche Strategien und Ziele der argumentativ-logischen Wissensgestaltung und -vermittlung bestehen, soll das präsentierte Poster auch zu einem Ausblick über die Geschichtswissenschaft hinaus anregen.

In der historischen Metatheorie wurden nicht nur die textuell-narrativen Bedingungen der Repräsentation von Geschichte bereits besprochen (s. einführend Crivellari et al. 2004; Haas 2004; Rüsen 2013: 191-220; Stopka 2018). Darüber hinaus wurden etwa auch Visualität als produktive Erweiterung der Ausdrucksmöglichkeiten (v.a. Haas 2006; Staley 2014) sowie Hypertext als Medium zur pluralistischen Darstellung von Geschichte(n) (v.a. Krameritsch 2007, 2009) ins Spiel gebracht. Krameritschs Beiträge stellen essenzielle Grundlagenarbeiten dar, an der sich mein Dissertationsprojekt methodisch und inhaltlich stark orientiert. Anders als Krameritsch geht es mir jedoch weniger um kollaborative Praxen der Wissenserzeugung in der Postmoderne oder um die Abbildung einer geschichtswissenschaftlichen Diskurs- und Wissenslandschaft mithilfe von netzwerkartigen Hypertexten. Vielmehr untersuche ich, wie einzelne Wissensangebote für sich hypertextuell publiziert werden können. Wo sich Krameritsch doch hierauf bezieht, sind es wiederum netzwerkartige, von Benutzer*innen relativ frei zu navigierende Hypertexte, für die er als innovative Vermittlungsformen votiert. Auf der Basis meiner theoretischen Reflexionen gelange ich hingegen zu visualisierten, *multilinear* gestalteten Hypertexten als vielversprechendste Varianten. Sie lösen m.E. deutlich stärker die historiografische Kernaufgabe ein, Geschichte(n) als (pluralistische) Zusammenhänge narrativ strukturiert zu vermitteln.

Theoretische Fundierung

Den Beginn macht eine epistemologische Grundlegung, die von einem non-dualistischen konstruktivistischen Wissensbegriff (vgl. zusammenfassend Weber 2010: 183-184) ausgeht. Dieses Wissensverständnis verlangt, in Publikationen stets Material und Verfahren zu explizieren, mit denen konstruiert wird; nur so werden die Wissensprodukte epistemisch überhaupt nachvollziehbar (vgl. Ceccato nach Zitterbarth 1991: 77). Im Kern geht es dabei natürlich nicht um Auflistungen parzellierter Einzelinformationen, sondern um das interpretierende, logisch-argumentative Herausstellen Sinn-voller Beziehungen zwischen Informationen. *Sinnzusammenhänge* mit ihren *Kohärenzstrukturen* stellen so die eigentlichen Wissensangebote (nicht nur) der Historiografie dar (vgl. Haas 2004: 233). Weil das Knowledge Design beim Publizieren symbolisch in ein Mediendesign überführt wird, folgen in der Dissertation medientheoretische Reflexionen, die schwerpunktmäßig die besondere Eignung, aber auch Grenzen von Hypertext für die Repräsentation explizit von Zusammenhängen (vgl. Storrer 2000, 2004; Winko 2005, 2008) adressieren. Insgesamt operationalisiere ich transdisziplinär Forschungsergebnisse v.a. aus der Literaturwissenschaft, den Medienwissenschaften sowie der Informatik. Die Multimodalitätsforschung (anschließend an Kress / van Leeuwen 2001) wird insbesondere in bild-linguistischer Prägung einbezogen. Texte, Bilder und Text-Bild-Kombinationen lassen sich daran ansetzend als Komplexe aus verknüpften Propositionen fassen (s. zusammenfassend Große 2011: 118-122). Narrative Texte geben die Propositionen semantisch explizit und hauptsächlich monosequenziert an, was medienästhetisch zu einer entsprechend sukzessiven Rezeption (z.B. von Argumentationslinien) führt. Sämtliche Zusammenhänge werden jedoch erst am Ende der Lektüre oder bei hoher Komplexität doch lediglich partiell erkennbar. Dementgegen können Bilder pluralistisch verknüpfte Propositionen simultan darstellen, wobei die Zusammenhänge semantisch unterdeterminiert bleiben. Sie werden allerdings ganzheitlich und nicht sukzessive wahrgenommen (vgl. Stöckl 2011: 48-49). Beide Modalitäten bieten gewissermaßen Vor- und Nachteile für die Wissensvermittlung.

Multimodaler Hypertext kombiniert konstruktiv verschiedene Medienästhetiken

Genau hier erweist sich die multimodale Kombination aus *textueller Darstellung in Knoten und Kanten* mit einem *Mapping der Gesamtstruktur* eines Hypertextes als vielversprechend. Logische Zusammenhänge werden nämlich über Knoten und Kanten hinweg als narrative Pfade repräsentiert – mehrfach nebeneinander sowie verknüpft, daher multilinear. Gleichzeitig wird ein ikonischer Überblick über die gesamte Kohärenzstruktur gegeben. Die Nachteile der einen Modalität werden weitreichend durch die Vorteile der anderen kompensiert, was besonders relevant ist, wenn Zusammenhänge aufgrund der in ihnen angelegten Pluralität kaum linear darzustellen sind. Auch die Einbindung von Video- oder Bildelementen erweitert den medienästhetischen Gestaltungsspielraum. Wird das Mapping im User Interface als interaktives navigatorisches Mittel genutzt, können Knoten und Kanten vom visualisierten Kohärenzgerüst ausgehend direkt angesteuert werden. Ein derartiges Prinzip haben E-Publishing-Tools wie *Scalar* (The Alliance for Networking Visual Culture: https://scalar.me/anvc/scalar/) längst umgesetzt, weswegen ich sie im Promotionsprojekt evaluativ heranziehe.

Als eine Hauptaussage des Projektes kann gelten, dass derlei Hypertexte mitnichten allein medienpädagogisch bedeutsam sind, sondern auch eine genuin epistemische Relevanz besitzen. Schließlich verrät das multimodale Mediendesign direkt etwas über die Konstruktionsweisen und -bedingungen von komplexen, pluralistischen Wissensangeboten. In dieser Hinsicht möchte ich mit dem Promotionsprojekt einen theoretischen Beitrag leisten, um ein reflektiertes E-Publishing sowie die (Weiter-)Entwicklung von E-Publishing-Tools zu unterstützen.

Fußnoten

1. Er ist an den architektonischen Gestaltungsgrundsatz „Form follows function“ angelehnt, der von Louis H. Sullivan berühmt gemacht wurde (Sullivan 1896).

Bibliographie

The Alliance for Networking Visual Culture: *Scalar* https://scalar.me/anvc/scalar/ [letzter Zugriff 29. Dezember 2018].

Crivellari, Fabio / Kirchmann, Kay / Sandl, Marcus / Schlögl, Rudolf (2004): *Einleitung. Die Medialität der Geschichte und die Historizität der Medien*, in: **Crivellari, Fabio / Kirchmann, Kay / Sandl, Marcus / Schlögl, Rudolf (eds.):** *Die Medien der Geschichte. Historizität und Medialität in interdisziplinärer Perspektive* (= Historische Kulturwissenschaft 4), Konstanz: UVK Verlagsgesellschaft 9-45.

Große, Franziska (2011): *Bild-Linguistik. Grundbegriffe und Methoden der linguistischen Bildanalyse in Text- und Diskursumgebungen* (= Germanistische Arbeiten zu Sprache und Kulturgeschichte 50), Frankfurt a.M. / Berlin / Bern / Brüssel / New York / Oxford / Wien: Peter Lang.

Haas, Stefan (2004): *Designing Knowledge. Theoretische und pragmatische Perspektiven der medialen Bedingungen der Erkenntnisformulierung und -vermittlung in den Kultur- und Sozialwissenschaften*, in: **Crivellari, Fabio / Kirchmann, Kay / Sandl, Marcus / Schlögl, Rudolf (eds.):** *Die Medien der Geschichte. Historizität und Medialität in interdisziplinärer Perspektive* (= Historische Kulturwissenschaft 4), Konstanz: UVK Verlagsgesellschaft 211-236.

Haas, Stefan (2006): *Vom Schreiben in Bildern. Visualität, Narrativität und digitale Medien in den historischen Wissenschaften*, in: zeitenblicke 5, vol. 3: http://www.zeitenblicke.de/2006/3/Haas*index_html [letzter Zugriff 5. November 2013].

Krameritsch, Jakob (2007): *Geschichte(n) im Netzwerk. Hypertext und dessen Potenziale für die Produktion, Repräsentation und Rezeption der historischen Erzählung*, Munster / Munchen: Waxmann.

Krameritsch, Jakob (2009): *Die fünf Typen des historischen Erzählens – im Zeitalter digitaler Medien*, in: Zeithistorische Forschungen / Studies in Contemporary History 6, vol. 3: 413-432.

Kress, Gunther / van Leeuwen, Theo (2001): *Multimodal Discourse The Modes and Media of Contemporary Communication*, London / New York: Arnold / Oxford University Press.

Krippendorff, Klaus (2006): *The Semantic Turn. A New Foundation for Design*, Boca Raton / London / New York: Taylor & Francis.

Rüsen, Jörn (2013): *Historik. Theorie der Geschichtswissenschaft*, Köln / Weimar / Wien: Böhlau.

Schnell, Ralf (2000): *Medienästhetik. Zu Geschichte und Theorie audiovisueller Wahrnehmungsformen*, Stuttgart: J.B. Metzler.

Schnell, Ralf (2001): *Medienästhetik*, in: **Schanze, Helmut (ed.):** *Handbuch der Mediengeschichte*, Stuttgart: Kröner 72-95.

Staley, David J. (2014): *Computers, Visualization, and History. How New Technology Will Transform Our Understanding of the Past*, Armonk / London: M.E. Sharpe.

Stöckl, Hartmut (2011): *Sprache-Bild-Texte lesen. Bausteine zur Methodik einer Grundkompetenz*, in: **Diekmannshenke, Hajo / Klemm, Michael / Stöckl, Hartmut (eds.):** *Bildlinguistik. Theorien – Methoden – Fallbeispiele* (= Philologische Studien und Quellen 228), Berlin: Erich Schmidt Verlag 45-70.

Stopka, Katja (2018): *Geschichte und Literatur*, in: **Busse, Laura / Enderle, Wilfried / Hohls, Rüdiger / Meyer, Thomas / Prellwitz, Jens / Schuhmann, Annette (eds.):** *Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften* (= Historisches Forum 23 / Veröffentlichungen von Clio-online 2), Berlin: Clio-online und Humboldt-Universität zu Berlin E.5-1-E.5-19.

Storrer, Angelika (2000): *Was ist 'hyper' am Hypertext?*, in: **Kallmeyer, Werner (ed.):** *Sprache und neue Medien* (= Institut für deutsche Sprache. Jahrbuch 1999). Berlin: De Gruyter 222-249.

Storrer, Angelika (2004): *Kohärenz in Hypertexten*, in: Zeitschrift für germanistische Linguistik 31, vol. 2: 274-292.

Sullivan, Louis H. (1896): *The Tall Office Building Artistically Considered*, in: Lippincott's Magazine 57: 403-409.

Weber, Stefan (2010): *Konstruktivistische Medientheorien*, in: **Weber, Stefan (ed.):** *Theorien der Medien. Von der Kulturkritik bis zum Konstruktivismus* (= UTB 2424). Konstanz: UVK Verlagsgesellschaft 170-188.

Winko, Simone (2005): *Hyper – Text – Literatur. Digitale Literatur als Herausforderung an die Literaturwissenschaft*, in: **Segeberg, Harro / Winko, Simone (eds.):** *Digitalität und Literalität. Zur Zukunft der Literatur*, Munchen: W. Fink 137-157.

Winko, Simone (2008): *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* (= Studien und Texte zur Sozialgeschichte der Literatur 71). Tübingen: Max Niemeyer Verlag 511-533.

Zitterbarth, Walter (1991): *Der Erlanger Konstruktivismus in seiner Beziehung zum konstruktiven Realismus*, in: **Peschl, Markus F. (ed.):** *Formen des Konstruktivismus in der Diskussion. Materialien zu den "Acht Vorlesungen über den Konstruktiven Realismus"* (= Cognitive Science 2). Wien: Wiener Universitätsverlag 73-87.

Mit neuen Suchstrategien vom isolierten Text zu 'illuminierten Urkunden'

Bürgermeister, Martina

martina.buergermeister@uni-graz.at
Universität Graz, Österreich

Bartz, Gabriele

Gabriele.Bartz@oeaw.ac.at
OÄW, Wien

Gneiß, Markus

Markus.Gneiss@oeaw.ac.at
OÄW, Wien

Immer mehr mittelalterliche Urkunden sind heute digital verfügbar, doch viele nur als Text in Urkundenbüchern. Die hohe Anzahl von online verfügbaren Abbildungen hat die Erforschung von ‚illuminierten Urkunden', die sich in

unterschiedlichsten Archiven befinden, erst möglich gemacht. Unter ‚illuminierte Urkunden' versteht man Urkunden mit bildlichem Dekor, die in spezifischen Fällen einen starken wechselseitigen Bezug von Bild und Text aufweisen. Eine Suchabfrage nach Illuminationen führt aber kaum zu Treffern. Das heißt, mit klassischen Suchstrategien sind ‚illuminierte Urkunden' wenn überhaupt nur mit sehr hohem Zeitaufwand auffindbar.

Deshalb scheint es zielführende, dass Expertinnen und Experten der Diplomatik, Kunstgeschichte und Digital Humanities zusammenarbeiten, um neue Suchstrategien für ‚illuminierte Urkunden' zu erproben. Urkunden mit Bildschmuck als interdisziplinäres Forschungsfeld sind noch nicht lange im Fokus der Geisteswissenschaften (Roland/Zajic 2013, Bartz/Gneiß 2018). Das erklärte Ziel besteht darin, Hinweise auf Illuminationen in den Metadatentexten aufzufinden. Dazu werden in einem mehrstufigen Textanalyseprozess aus bereits erschlossenen Urkunden automatisiert Stichwörter extrahiert („keyword extraction"), gerankt und zur API-Abfrage von online Archivdatenbanken aufbereitet, um zu neuen Treffern zu führen.

Der Einsatz von textstatistische Methoden zur Beantwortung von Forschungsfragen der digitalen Diplomatik ist state of the art (vgl. Tilahun/Feuerverger/Gervers 2012, Perreaux 2014, HIMANIS-Projekt: www.himanis.org). Neu jedoch ist deren Anwendung, um über Metadaten zu einer innovativen Suchstrategie nach Urkunden mit Bildschmuck zu gelangen. Unser Beitrag ist als Reaktion auf das wahrgenommene Defizit der isolierten Betrachtung von Text- und Bilddaten zu verstehen.

Daten

In unserem Fokus steht eine weit verbreitete Gruppe von ‚illuminierten Urkunden': den Wappenbriefen. Bei dieser Urkundengattung gewährt ein Herrscher das Recht ein bestimmtes Wappen zu tragen, das normalerweise auf der Urkunde gemalt ist. Wappenbriefe sind in großer Zahl ausgestellt worden und sind deshalb einheitlich in ihren Textbestandteilen. In Vorgängerprojekten (FWF, Go!Digital) wurden bereits etwa 150 Wappenbriefe mit historischen Methoden gesammelt, beschrieben und über das Urkundenarchiv „monasterium.net" (monasterium.net/mom/collection/illuminierteUrkunden) veröffentlicht.

Zum ersten Mal sollen nun unter Anwendung textstatistischer Methoden eine große Zahl an Wappenbriefen gefunden, in Hinblick auf die gesamteuropäische Entwicklung untersucht und über monasterium.net veröffentlicht werden. Die multimodalen Erschließungsmöglichkeiten in monasterium.net schaffen die Voraussetzung zur komparativen Untersuchung von textlichen Eigenschaften und Illuminationen von zentral-, süd- und westeuropäischen Beispielen. Insgesamt werden durch das Projekt neue Impulse für die Erforschung von Wappenbriefen gesetzt.

Methoden

Die geringe Anzahl an bisher gesammelten und tiefenerschlossenen Wappenbriefen, sowie der Umstand, dass es sich um Kurztexte handelt, ist für das Verfahren der „keyword extraction" eine besondere Herausforderung. Um jene Wörter in den bereits erschlossenen Dokumenten zu finden, die konstitutiv sind, werden ausschließlich statistische Methoden, welche an kein umfangreiches Trainingsset gebunden sind, eingesetzt. Grundlage der Verfahren ist die numerische (vektorisierte) Repräsentation aller zu untersuchenden Wörter eines Dokuments. In einem ersten Schritt soll das TF-IDF-Maß gewichtete Stichwörter liefern. Dann werden zwei weitere methodische Ansätze (clustering, entropy), die sich zur Stichwortextraktion eignen (vgl. Herrera/Pury 2008, Carretero-Campos et al. 2013, Jamaati/Mehri 2018) auf Texte der Wappenbriefe angewandt. Zur Bestimmung der Wortrelevanz spielt sowohl beim Clustering also auch bei der Entropie-Methode die Verteilung der Wörter im Text eine entscheidende Rolle. Beim Clustering wirkt sich die Streuung gemessen an unterschiedlichen Wortabständen (Vorkommen, Position) auf das Ranking der so generierten Stichwörter aus. Dabei gilt, je größer die Abweichung, desto relevanter das Wort (vgl. Zhou/Slater 2003, Carretero-Campos et al. 2013). Die mithilfe Shannons Entropie (Shannon/Weaver 1963) errechneten relevanten Wörter werden auch an ihrer Verteilungsheterogenität gemessen: Je ungleichmäßiger ihre Verteilung desto relevanter der Inhalt (vgl. Herrera/Pury 2008). Schließlich werden die Ergebnisse aller Messungen evaluiert und für die Datenbankabfrage aufbereitet. Die ermittelten Schlüsselwörter sollen die Qualität des Information Retrieval erhöhen und über die Abfrage von öffentlichen Datenbank-APIs, wie Regesta Imperii (www.regesta-imperii.de) und Archivportal-D (www.archivportal-d.de/) zum Auffinden von weiteren Wappenbriefen führen.

Schluss

Die Suchstrategie geht auf eine neue Art mit isoliertem Text um. Die Multimedialität von ‚illuminierten Urkunden' wird durch die automatisierte Extraktion von Stichwörtern zu einem serialisierten Abfragemuster für öffentliche Archiv-APIs, die so zu umfangreicheren Quellenfunden führen. Die quantitative, statistische Methodologie des Projektes bringt zudem eine neue Perspektive auf Wappenbriefe und deren formelhafte Struktur, die über Zeit und Ort verglichen werden kann. Was im vorliegenden Projekt an Wappenbriefen getestet wird kann auf andere Urkundentypen ausgedehnt werden und einen grundlegenden DH-Beitrag zum Thema „keyword extraction" bei Kurztexten liefern.

Bibliographie

Bartz, Gabriele / Gneiß, Markus (2018) (eds.): *Illuminierte Urkunden. Beiträge aus Diplomatik, Kunstgeschichte und Digital Humanities/Illuminated Charters*. Essays from Diplomatic, Art History and Digital Humanities (= Archiv für Diplomatik, Schriftgeschichte, Siegel-und Wappenkunde 16), Wien: Böhlau Verlag.

Carretero-Campos, Concepcion / Bernaola-Galvan, Pedro / Coronado Jiménez, Ana Victoria / Carpena, Pedro (2013): *Improving statistical keyword detection in short texts: Entropic and clustering approaches* in: Physica A 392: 1481-1492.

Herrera, Juan P. / Pury, Pedro A. (2008): *Statistical keyword detection in literary corpora* in: The European Physical Journal B 63/1: 135-146.

Jamaati, Maryam / Mehri, Ali (2018): *Text mining by Tsallis entropy*, in: Physica A 490: 1368-1376.

Perreaux, Nicolas (2014): *De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques* in: **Ambrosio/ Barret/Vogeler (eds.)** *Digital diplomatics. The computer as a tool for the diplomatist?* (= Archiv für Diplomatik Beihefte 14), Köln/Weimar/Wien: Böhlau Verlag 187-210.

Roland, Martin / Zajic, Andreas (2013): *Illuminierte Urkunden des Mittelalters in Mitteleuropa*, in: Archiv für Diplomatik 59: 241-432.

Shannon, Claude Elwood / Weaver, Warren (1963): *The Mathematical Theory of Communication* Champaign: Illinois University Press.

Tilahun, Gelila / Feuerverger, Andrey / Gervers, Michael (2012): *Dating Medieval English Charters*, in: The Annals of Applied Statistics 6/4: 1615-1640.

Zhou, H. / Slater, G.W. (2003): *A metric to search for relevant words*, in: Physica A 329 (2003) 309–327.

Modellprojekt "eHumanities - interdisziplinär": Forschungsdatenmanagement im Rahmen des "Digitalen Campus Bayern"

Schulz, Julian

julian.schulz@lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Unabhängig davon, ob man die unter dem Begriff *Digital Humanities* zusammengefassten Forschungsansätze und Methoden als eigenständige, neuartige Fachdisziplin oder als digital transformierte, geisteswissenschaftliche Einzeldisziplinen versteht: die zunehmende Proliferation digitaler Daten und Datenverarbeitung in den Geisteswissenschaften erfordert ein langfristiges Forschungsdatenmanagement und die Bereitstellung von unterstützenden Diensten (DHd-AG Datenzentren 2017: 4; Krefeld/Lücke 2017: 2).

Während in diesem Punkt Konsens besteht, zeigen doch die aktuellen Diskussionen bezüglich der Ausgestaltung einer zu schaffenden Nationalen Forschungsdateninfrastruktur (NFDI), dass der Denkprozess darüber, wie Forschungsdatenmanagement (FDM) in den Geisteswissenschaften organisiert werden sollte, noch keinen Abschluss gefunden hat.[1] Hinzu tritt ein bislang sehr uneinheitlicher Umgang mit Forschungsdaten in den Geistes- und Sozialwissenschaften (RfII 2016: 10).

In der Diskussion über das Organisationsmodell der NFDI ist jedoch eine Gemeinsamkeit identifizierbar – und aus Sicht der Einreichenden unabdingbar: Eine wie auch immer ausgestaltete Konsortienstruktur kann nur dann auf fruchtbaren Boden in den geisteswissenschaftlichen Fachdisziplinen fallen, wenn eine doppelseitige Rückkopplung mit den einzelnen Wissenschaftsstandorten gegeben ist (DHd-AG Datenzentren 2018). Einer regionalen bzw. lokalen Verankerung des FDM kommt dabei eine wichtige Rolle zu. Dies bringt zum einen Herausforderungen auf organisatorischer Ebene mit: Wie können vor Ort Workflows und Prozesse, Beratungs- und Schulungsangebote etabliert werden? Wie kann der Austausch zwischen lokalen Akteuren mit der zu schaffenden Konsortienstruktur gestaltet werden? Andererseits ergeben sich Herausforderungen auf technischer Ebene, insbesondere in Fragen der Adressierbarkeit von Forschungsdaten in übergeordneten Repositorien.

Hier setzt das vom Bayerischen Ministerium für Wissenschaft und Kunst im Rahmen der Förderlinie „Digitaler Campus Bayern" aufgelegte Projekt *eHumanities – interdisziplinär* an. Das Projekt, welches zugleich als Modellvorhaben „Forschungsdatenmanagement Bayern" fungiert[2], sieht sich an der Schnittstelle zwischen übergeordneten Strukturen und lokalen, geisteswissenschaftlichen Akteuren. Projektbeteiligte sind die Universitätsbibliotheken der Friedrich-Alexander-Universität Erlangen-Nürnberg (UB FAU) und der Ludwig-Maximilians-Universität München (UB LMU) sowie die IT-Gruppe Geisteswissenschaften (ITG) der Ludwig-Maximilian-Universität München.

Ziel des Projekts ist es, abgestimmt auf die Bedürfnisse der digitalen Geisteswissenschaften modellhaft Konzepte, Services und Workflows sowie konkrete technische Lösungen für das Forschungsdatenmanagement (zunächst) an den Standorten Erlangen-Nürnberg und München zu erarbeiten. Aufbauend auf die dabei gewonnenen Erfahrungen werden Schulungs- und Lehrangebote konzipiert.

Das einzureichende Poster möchte einen Überblick zu den im Laufe des ersten Projektjahres in fünf Arbeitspaketen erzielten Ergebnissen geben und diese in den Gesamtkontext der aktuellen Diskussionen um das FDM in den Geisteswissenschaften einbetten. Anhand eines Anwendungsfalles (Projekt *VerbaAlpina*[3]), der im Rahmen von *eHumanities – interdisziplinär* als Pilotprojekt fungiert, sollen die organisatorischen und technischen Abläufe anschaulich präsentiert und zur Diskussion gestellt werden sowie zukünftige Arbeitsschritte bzw. Desiderate aufgezeigt werden:

Aktuell erfolgt im Rahmen des Projekts eine Bestandsanalyse gängiger Metadatenmodelle und ihre Anwendung auf den umfangreichen Projektdatenbestand der ITG. In diesem Kontext findet ein Austausch mit anderen FDM-Akteuren (z.B. TU München, Leibniz-Rechenzentrum) statt, um eine interdisziplinäre Vernetzung und Abstimmung über die gemeinsame Verwendung von Standards und Modellen zu gewährleisten. Neben der formalen Auszeichnung der Daten wird gleichzeitig nach einem flankierenden inhaltlichen Erschließungsmodell für die ausgesprochen heterogene Datenlandschaft der Digital Humanities gesucht (vgl. Abb. 1). Damit eng verbunden ist die Etablierung von Arbeitsabläufen für die standardisierte Weitergabe der Forschungsdaten aus dem jeweiligen Einzelprojekt – hier *VerbaAlpina* – seitens der ITG in das zentrale Repositorium der Universitätsbibliothek (*Open Data LMU*[4]). Durch die Vergabe von DOIs wird die Adressierbarkeit und Zitierbarkeit der Forschungsergebnisse bei Bedarf bis auf die Einzeldatensatzebene hinunter dauerhaft gewährleistet. Gleichzeitig werden die Daten durch die UB für übergeordnete Repositorien – im vorliegenden

Fall für das Forschungsdateninfrastrukturprojekt *GeRDI*[5] – sowie für Kataloge bzw. Ressource Discovery Systeme (z.B. *BASE*) zugänglich gemacht. Um auf diese Weise eine mehrdimensionale Vernetzung und interdisziplinäre Kooperation zu ermöglichen, werden im Rahmen eines weiteren Arbeitspaketes standardunabhängige und flexible Schnittstellenlösungen (auf Basis von *OAI-PMH*) entwickelt.

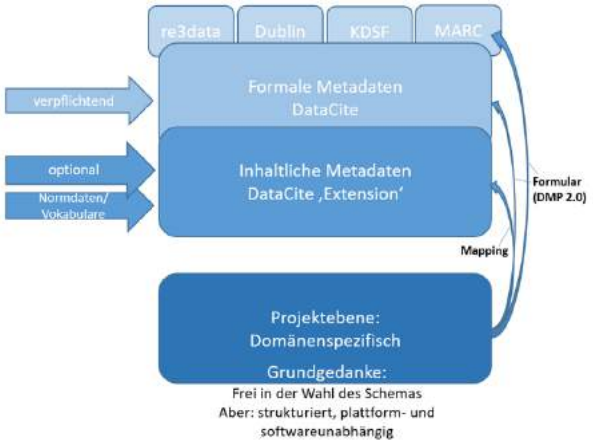


Abbildung 1. Metadatenmodell und Prozess der Metadatenanreicherung im Rahmen des FDM

Die Kooperation von UB LMU und einer eng an die Geisteswissenschaften angebundenen Einrichtung (ITG) beschreitet in diesem Anwendungsfall als Konzept für ein domänenspezifisches Datenzentrum einen erfolgsversprechenden Weg. Es könnte Modellcharakter für andere Universitätsstandorte besitzen. Durch die enge Verzahnung eines auf diese Weise operierenden Datenzentrums mit der Fachcommunity besteht die Möglichkeit, als lokaler Kooperationspartner für ein zu bildendes geisteswissenschaftliches NFDI-Konsortium zu agieren (vgl. Abb. 2). Als Bindeglied zwischen den Geisteswissenschaften und dem übergeordneten Konsortium kann es Beratungstätigkeiten übernehmen, Bedarfe ermitteln und kommunizieren. Damit würde – dem jüngst veröffentlichten RfII-Positionspapier folgend – ein Beitrag geleistet, dass dieses Konsortium „auf Dauer dynamisch" bleibt (RfII 2018: 3).

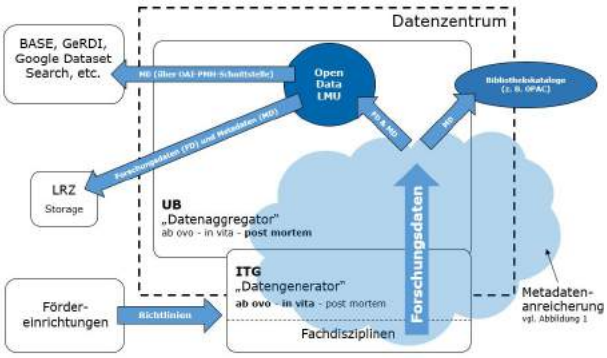


Abbildung 2. Rollenverteilung beim FDM am Beispiel der LMU als Prototyp für ein Datenzentrum

Um die Nachnutzung der im Rahmen eines Forschungsprojekts entstehenden Daten zu vereinfachen, kommt der Beratung vor Projektbeginn (*ab ovo*) und der steten Begleitung des Vorhabens durch ein domänenspezifisches Datenzentrum eine entscheidende Rolle zu. Für die datenbezogene Planung von Forschungsvorhaben, die strukturierte Datenerzeugung sowie -dokumentation (Metadaten, Abbild des Entstehungskontextes) und eine sinnvolle Planung der Datenaufbereitung hat sich die Erstellung von Datenmanagementplänen (DMP) etabliert. Ein Arbeitspaket, angesiedelt an der UB FAU, verfolgt daher die Zielsetzung, verschiedene bestehende Softwarelösungen zu evaluieren und das Werkzeug der Wahl für die Geisteswissenschaften anzupassen. Das Projekt reagiert damit auf die wachsenden Anforderungen seitens der maßgeblichen Förderinstitutionen, die eine detaillierte Datenmanagementplanung bereits bei Beantragung eines Projektes fordern. Da die Anforderungen in den einzelnen Wissenschaftsbereichen sehr unterschiedlich sind, werden die Tools insbesondere auf ihre Modularität hin untersucht. Durch die Trennung von allgemeinen und fachspezifischen Inhalten ist es später möglich, im Projekt entstehende generische DMP-Vorlagen auch außerhalb der Geisteswissenschaften zu verwenden.

Um Datenproduzenten frühzeitig für Verfahren zur datenbezogenen Planung von Forschungsvorhaben zu sensibilisieren, finden die Ergebnisse der oben genannten Arbeitsbereiche schließlich Eingang in ein abgestimmtes Schulungspaket, das wiederum an der FAU Erlangen-Nürnberg konzipiert wird. Neben fachunabhängigen Lehrinhalten (Datenorganisation, Datenstruktur, Datensicherung, Datendokumentation, Datenzitation, Datenrecherche) werden auch fachspezifische Kursinhalte (beispielsweise Datenschutz, Datenpublikation) Berücksichtigung finden. Die fachunabhängigen Lehreinheiten eignen sich für eine Verwendung über die Digital Humanities hinaus und können – ergänzt um disziplinspezifische Schwerpunkte – in anderen Fachkontexten nachgenutzt werden.

Das Projekt *eHumanities – interdisziplinär* versteht sich als regionale Initiative zur Bündelung bestehender und Schaffung neuer Dienstleistungen und Infrastrukturen sowie konzeptioneller und technischer Lösungen für das Forschungsdatenmanagement in den Geisteswissenschaften. Durch seine enge Anbindung an die forschenden Wissenschaftler/-innen wird ein Beitrag für die Verankerung der zu bildenden NFDI-Konsortienstruktur in diesem Fachbereich geleistet. Auf Grundlage der im Rahmen des Projekts erarbeiteten bzw. noch zu erzielenden Ergebnisse, die Modellcharakter für die Geisteswissenschaften besitzen, kann das Staatsministerium für Wissenschaft und Kunst in Rückkopplung mit dem bayerischen CIO-Gremium Prozesse für eine Adaptierung in anderen Wissenschaftsdomänen in die Wege leiten. Es ist ein erklärtes Ziel, dadurch Doppelstrukturen zu vermeiden und den interdisziplinären Austausch zwischen den einzelnen Wissenschaftszweigen zu befördern.

Fußnoten

1. Vgl. hierzu die Workshop-Reihe sowie die Auflistung der Positionspapiere verschiedener Verbände auf http://forschungsinfrastrukturen.de/.
2. Im Rahmen des Projekts wurde eine Webseite eingerichtet, die sich über das Projekt hinaus als zentrale Anlaufstelle für FDM-Aktivitäten (in Bayern) sieht: www.fdm-bayern.org.
3. Projektwebseite: https://www.verba-alpina.gwi.uni-muenchen.de/.
4. Open Data LMU: https://data.ub.uni-muenchen.de/.
5. Projektwebseite: https://www.gerdi-project.eu/.

Bibliographie

DHd-AG Datenzentren (2017): *„Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum. Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten“* https://doi.org/10.5281/zenodo.1134760.

DHd-AG Datenzentren (2018): *„Vorschlag der AG Datenzentren im DHd zur Bildung und Strukturierung eines NFDI-Konsortiums für die Geisteswissenschaften“* https://dhd-ag-datenzentren.github.io/nfdi-vorschlag.

Krefeld, Thomas / Lücke, Stephan (2017): *„Nachhaltigkeit – aus der Sicht virtueller Forschungsumgebungen“*. Korpus im Text. Version 7 (10.03.2017, 12:27) http://www.kit.gwi.uni-muenchen.de/?p=5773&v=7.

Rat für Informationsinfrastrukturen (2016): *„Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland“*. Göttingen, 160 S. http://www.rfii.de/?p=1998.

Rat für Informationsinfrastrukturen (2018): *„In der Breite und forschungsnah: Handlungsfähige Konsortien. Dritter Diskussionsimpuls zur Ausgestaltung einer Nationalen Forschungsdateninfrastruktur (NFDI) für die Wissenschaft in Deutschland“*. Göttingen, 6 S. http://www.rfii.de/?p=3509.

Multimedia aus Rezipientenperspektive: Wirkungsmessung anhand von Biofeedback

Schlör, Daniel
daniel.schloer@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Veseli, Blerta
blerta.veseli@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Die Wirkung, die Medien wie Literatur, Musik oder Film auf Rezipienten haben, ist häufig Gegenstand kultur- und geisteswissenschaftlicher Arbeiten. Das Rezipieren von beispielsweise Literatur oder Film involviert emotionale Prozesse (Schwarz-Friesel, 2013), welche sich durch physiologische Reaktionen, wie die Veränderung von Atmung, Herzfrequenz oder Schweißabsonderung, ausprägen (Hergovich, 2018). Eben diese Ausprägung in Form von klar messbarer Kenngrößen erlaubt die Digitalisierung der Rezipientenwirkungen durch körpernahe Daten. Diese Digitalisate ermöglichen den digitalen Kultur- und Geisteswissenschaften gezielte Rückschlüsse auf die Wahrnehmung inhaltlicher, narrativer oder stilistischer Aspekte rezipierter Medien zu ziehen.

In dieser Arbeit stellen wir eine einfache Methode vor, Rezipientenwirkung in ihrer physischen Ausprägung zu digitalisieren und auszuwerten. Um leicht eine ausreichend große Datenbasis für quantitativ belegbare Ergebnisse zu schaffen, stellen wir Nutzern und Forschern dazu unsere mobile Applikation “BioReader” als Werkzeug zur Verfügung, das beim Lesen und Medienkonsum den Lesefortschritt sowie physiologische Reaktionen über Fitnessuhren erfasst.

In einer Machbarkeitsstudie wurden auf diese Weise körpernahe Sensordaten von 15 Probanden in einer multimedialen Versuchsreihe beim Lesen von vier Kurzgeschichten und Schauen eines Kurzfilms aufgenommen, die wir in diesem Beitrag vorstellen.

Verwandte Arbeiten

Medienkonsum wird im großen Stil vor allem aus wirtschaftlichen Interessen, beispielsweise von Videostreaming bzw. eBook Anbietern wie Netflix (Tassi, 2018) bzw. Jellybooks (Buchreport, 2017) erfasst und analysiert und misst meist lediglich das Konsumverhalten. Aus wissenschaftlicher Sicht gibt es einige Studien, die physiologische Reaktionen auf multimedialen Konsum untersuchen:

So haben (Riese et al., 2014) die Reaktionen der Pupille auf Spannung im Text untersucht und dabei festgestellt, dass es signifikante Korrelationen zwischen dem Durchmesser der Pupille und dem Verlauf der Spannung in einem Text über die Zeit gibt. Eine Studie von (Richter et al., 2011) untersuchte die Reaktion von Lesern auf schmerzassoziierte Wörter wie “bohrend” oder “krampfartig” und stellte fest, dass bei der Verarbeitung derartiger Wörter dieselben Hirnareale aktiviert werden, die für die Verarbeitung realer Schmerzreize zuständig sind. In einer Studie von (Bar-Haim et al., 2004) wurde die Herzfrequenz bei Kindern gemessen, denen sechs Geschichten mit verschiedenen Themen präsentiert wurden. Das Ergebnis dieser Studie zeigte, dass relativ zum zuvor gemessenen Ruhepuls, die Probanden bei einigen Geschichten eine signifikant schnellere Herzfrequenz aufwiesen.

In ihrer Studie untersuchten (Baldaro et al., 2001) physiologische Reaktionen auf Filme. Dafür wurden den Probanden zwei zehn-minütige Filmausschnitte gezeigt: ein *Surgery film* über eine Operation am Brustkorb und ein *Neutral film* über Landschaften. Während des Schauens

der Filme wurden Körperparameter wie beispielsweise Atemfrequenz und Herzrate erfasst. Es zeigte sich im Vergleich zum *Neutral film* ein stärkeres Absinken der Herzfrequenz während des *Surgery* Films.

"BioReader" Tracking-Applikation

Die von uns entwickelte Tracking App "BioReader", ist als eReader konzipiert und erlaubt das Lesen von Texten aus der eigenen Nutzerbibliothek. Diese werden seitenweise auf dem Smartphone dargestellt. Eine Eigenentwicklung war nötig, um die aufgezeichneten Daten den gelesenen Stellen automatisch zuordnen zu können.

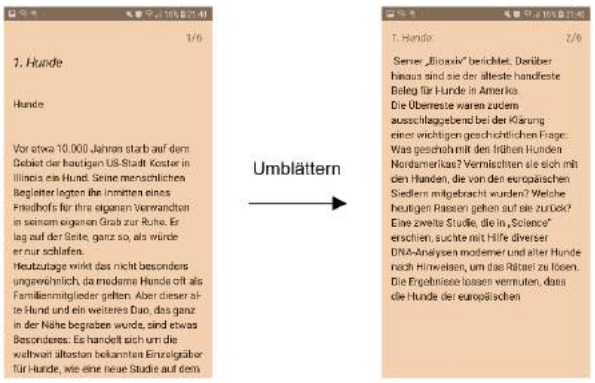


Abbildung 1. Seitendarstellung in App "BioReader"

Für die Messung nutzt die App alle verfügbaren Sensoren, die in der über Android Wear gekoppelten Smartwatch verbaut sind. Ist die Uhr verbunden, startet die Messung automatisch mit dem Öffnen eines Textes. Alle Events während der Aufzeichnung, beispielsweise das Drücken von Home-Button, Fokus-Wechsel zu anderen Apps, Schließen der Anwendung, Abschalten des Bildschirms oder Anzeigen des Lock-Screens werden protokolliert, um potentielle Ablenkungen beim mobilen Lesen (Kuzmicova et al., 2018) zu berücksichtigen.

Der Nutzer kann sich seinen Lesefortschritt und jeweils aufgezeichneten Sensorwerte für jede Lese-Session auf einer Auswertungsseite visualisieren lassen.

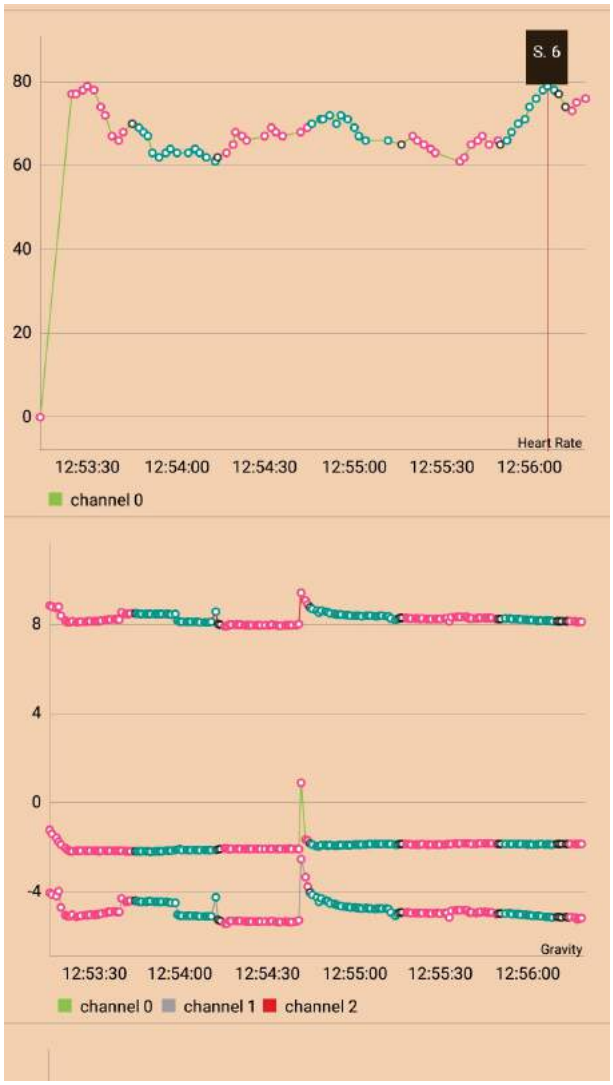


Abbildung 2. Datenvisualisierung mit Lesefortschritt-Anzeige am Beispiel des Gravity und Heart Rate Sensors

Um Messwerte während der Rezeption anderer Medien wie beispielsweise Filme oder Musik aufzuzeichnen, verfügt die App über die Möglichkeit Messungen manuell zu starten und zu beenden. Über eine Export-Funktion können alle von der App aufgezeichneten Daten im CSV-Format exportiert werden.

Experiment

Mit dieser App wurde eine Machbarkeitsstudie durchgeführt, um zu untersuchen, ob mit verbreiteten Fitnessuhren physiologische Reaktionen der Rezipienten auf verschiedene Medien messbar sind.

Korpus Zusammenstellung

Für die im Folgenden präsentierte Studie wurden vier kurze Textausschnitte und ein Kurzfilm ausgewählt, um die Veränderung von Körperparametern während der Rezeption zu untersuchen.

Dabei sollen die Texte bzw. das Video zum einen leicht verständlich und schnell zu lesen bzw. zu schauen sein, zum anderen sollen sie starke emotionale Reaktionen und damit potenziell auch körperliche Veränderungen hervorrufen. Hinsichtlich der Textauswahl wurden dazu unterschiedliche Textarten mit verschiedenen Wirkungszielen verwendet, darunter ein Sachtext (Arbeitstitel: *Hunde*[1]), welcher in einem nüchternen Sprachstil gehalten ist und zur Eingewöhnung der Probanden dienen soll, eine jugendsprachlich verfasste, narrative Kurzgeschichte (Arbeitstitel: *Praxis*[2]) mit komischen Erzählelementen und zwei negative Texte: ein Kurzartikel (Arbeitstitel: *Genie*[3]), der sich durch die Verwendung eines stark negativ konnotierten Vokabulars auszeichnet und ein Textausschnitt (Arbeitstitel: *James*[4]), welcher den Tathergang während eines Mordes objektiv und chronologisch beschreibt.

Bezüglich des Videos fiel die Wahl auf den dreiminütigen Horror-Kurzfilm *Lights Out*[5], der auf kurze, intensive Filmsequenzen mit starken Schockmomenten und die Vermeidung von direkter Gewaltdarstellung setzt.

Durchführung

Jede Versuchsperson wurde zunächst über den Studienablauf, d.h. Reihenfolge der Inhalte und Fragebögen, aufgeklärt. Im weiteren Verlauf bedienten die Teilnehmer selbstständig und abgesehen vom Ablaufplan ohne weitere Anleitung die BioReader App, die zum Lesen der Texte, Starten und Stoppen der Messung und zur Visualisierung der Messwerte verwendet werden sollte. Für die Messung wurde eine Polar M600 Pulsuhr verwendet, die die Probanden während der Studie am Handgelenk trugen. Diese zeichnet unter anderem Sensorwerte für Accelerometer, Gyroskop und Herzfrequenz auf, wobei wir im Rahmen dieser Studie unseren Fokus auf die Herzfrequenz setzen.

Nach dem Lesen aller Texte bzw. Schauen des Videos wurden die Probanden in einem Fragebogen bezüglich der Wahrnehmungsintensität, Immersion bzw. Transportation in die narrative Welt (Appel, Richter 2010), Gefühle bei der Rezeption und Intensität einzelner Teile befragt.

Schließlich füllten die Versuchspersonen noch einen Fragebogen zur Usability der BioReader App aus und beantworteten einige demographische Fragen.

Ergebnisse

Zur Bewertung der Usability und User Experience der App wurde der *UEQ - User Experience Questionnaire* (Laugwitz et al., 2008) eingesetzt, welcher den Gesamteindruck der Nutzer in Bezug auf die Applikation misst. Die Auswertung des Fragebogens ergab, dass die App von den Probanden als sehr nutzerfreundlich und einfach zu bedienen eingeschätzt wurde. Im Folgenden werden die Ergebnisse der Machbarkeitsstudie vorgestellt, die einen Ausblick auf mögliche Anwendungen der App geben soll. Daher beschränken wir uns in diesem Beitrag auf eine explorative Auswertung zur Gewinnung von Hypothesen, die in weiteren Arbeiten statistisch bewertet werden müssen.

Die Versuchspersonen zeigen im Verlauf der Studie zum Texterlebnis eine mittlere Herzfrequenzschwankung von 23,27 Schläge pro Minute, wobei sie wenigstens um 13, höchstens um 33 Schläge pro Minute im Verlauf schwankt. Wir normieren die Sensordaten pro Teilnehmer, um Abweichungen im Textverlauf zu erkennen und visualisieren diese neben der dazugehörigen Textseite. Ein Punkt repräsentiert auf der Y-Achse den Lesezeitpunkt, auf der X-Achse den Messwert der Herzfrequenz. Die Verbindung der Punkte approximiert linear den normierten Herzratenverlauf eines Probanden.

Um eine Eingewöhnung an die Studiensituation zu erreichen, beginnen alle Teilnehmer mit dem Sachtext. Entsprechend Abbildung 3 messen wir, dass einige Probanden zu Beginn der Studie einen relativ hohen Puls haben, der sich im Verlauf der ersten halben Seite normalisiert. Betrachtet man die Herzfrequenz über alle Textseiten, wird deutlich, dass Probanden an unterschiedlichen Stellen im Text mit einer Veränderung des Pulses reagierten. Bei manchen Probanden scheinen z.B. Gewaltdarstellungen (siehe Abbildung 4) zu einer Pulserhöhung (z.B. H1User4, H2User6 oder H2User7) zu führen, was sich bei einigen Probanden (z.B. H2User7, H1User2, H1User6, H1User4) auch bei komischen Elementen (siehe Abbildung 5) beobachten lässt.

Zu Beginn des Videos “Lights Out”, in einer Szene, in der sich die Darstellerin unter einer Bettdecke versteckt und bei der Schockszene am Ende des Kurzfilms zeigen einige Probanden deutliche Veränderungen in der Herzfrequenz (siehe Abbildung 6). Diese intensiven Szenen zeichnen sich durch besonders tiefe und laute Frequenzen in der Audiospur aus, die zur Wirkung der Szene beitragen. Die Bewertung der Szenen durch die Probanden deckt sich mit der Analyse der Messwerte und motiviert die Hypothese, dass besonders intensiv erlebte Spannungselemente durch Veränderung in der Herzrate messbar werden.

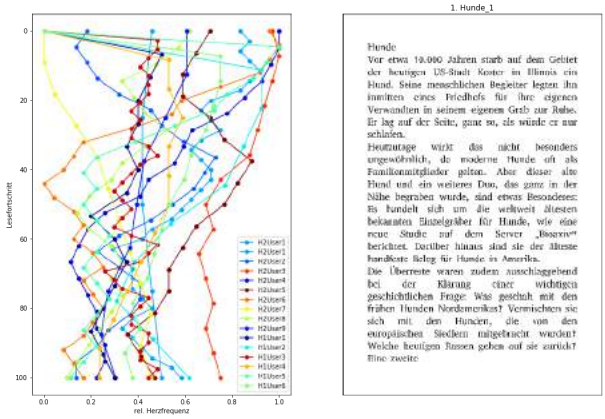


Abbildung 3. Herzfrequenzverläufe für den ersten Text (Sachtext) im Studienverlauf

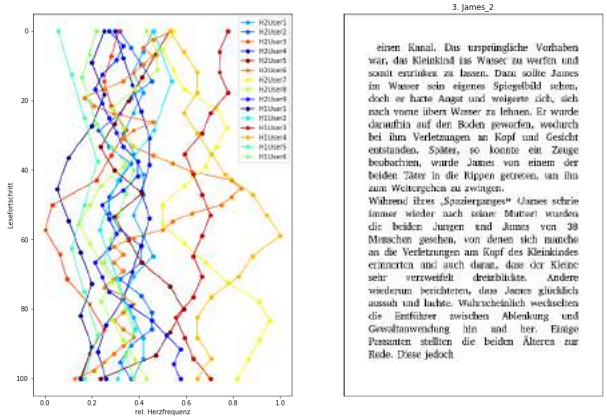


Abbildung 4. Herzfrequenzverläufe für einen Textabschnitt mit Gewaltdarstellungen

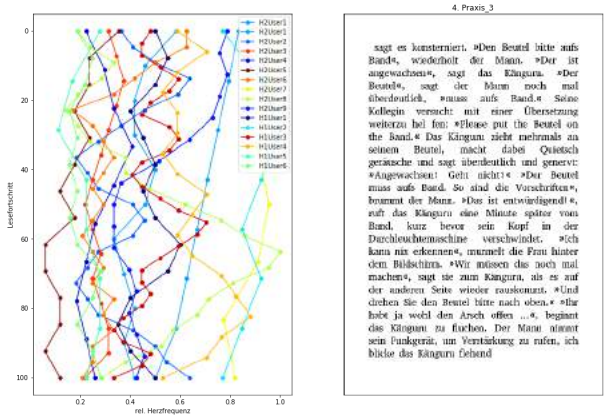


Abbildung 5. Herzfrequenzverläufe für einen komischen Textabschnitt

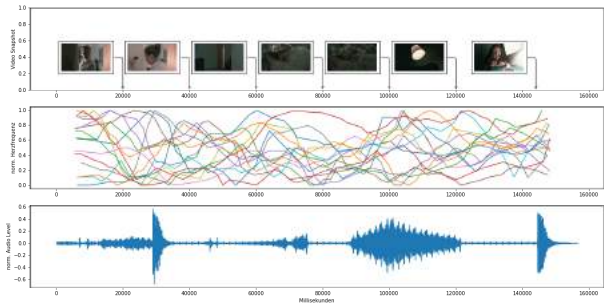


Abbildung 6. Video Snapshots, Herzfrequenzen und Audio Waveform für den Film "Lights out"

Fazit und Ausblick

Die von uns entwickelte App ermöglicht Nutzern bequem Körperreaktionen bei Multimedia Konsum zu tracken und bietet für Studien ein Werkzeug, um Zusammenhänge zwischen Nutzerwahrnehmung, Körperreaktionen und Medieninhalten zu erfassen und zu untersuchen.
In unserer Teststudie konnten wir zeigen, dass die Usability der App von Probanden als sehr gut bewertet wurde und die App in der Lage ist, Sensordaten wie Herzfrequenz zuverlässig zum Lesefortschritt aufzuzeichnen und darzustellen.

Damit eröffnet "BioReader" die Möglichkeit in großangelegten Studien physiologische Reaktionen und Rezipientenwirkung zu digitalisieren und diese zu untersuchen, um damit generalisierbare Erkenntnisse zu gewinnen. Diese können darüber hinaus mit inhaltlichen Analysen, beispielsweise Sentiment-Analyse, kombiniert werden und neue Zusammenhänge zwischen Wahrnehmungsebene und Inhaltsebene aufdecken.

In zukünftigen App-Versionen planen wir zusätzliche Sensorik wie beispielsweise Hautleitwertsensor oder die Front-Kamera für eine bildbasierte Emotionserkennung zu integrieren.

Fußnoten

1. Wei-Haas, Maya: Wohin verschwanden die ersten Hunde Amerikas? In: National Geographic. Link: https://www.nationalgeographic.de/wissenschaft/2018/07/wohin-verschwanden-die-ersten-hunde-amerikas (zuletzt aufgerufen: 28.09.2018).
2. Kling, Marc-Uwe: Theorie und Praxis. In: Die Känguru-Chroniken. Ansichten eines vorlauten Beuteltiers, Berlin 2009.
3. WDR: 4. November 1970 - Entdeckung des Wolfskinds "Genie". Link: https://www1.wdr.de/stichtag/stichtag-554.htm l (zuletzt aufgerufen: 28.09.2018)
4. Wikipedia: Mord an James Bulger. Link: https://de.wikipedia.org/wiki/Mord_an_James_Bulger (zuletzt aufgerufen: 28.09.2018)
5. Lights Out, in: Youtube. Link: https://www.youtube.com/watch?v=kNbJE0y29_c&t=7s (zuletzt aufgerufen: 28.09.2018)

Bibliographie

Appel, Markus / Richter, Tobias (2010): *Transportation and Need for Affect in Narrative Persuasion: A Mediated Moderation Model*, Media Psychology, 13:2, 101-135, DOI:10.1080/15213261003799847

Baldaro, Bruno / Codispoti, Maurizio / Mazetti, Michela / Tuozzi, Giovanni (2001): *Autonomic reactivity during viewing of an unpleasant film*, in: Perceptual and Motor Skills, Seite 797-805.

Bar-Haim, Yair / Fox, Nathan A. / VanMeenen, Kirsten M. / Marshall, Peter J.(2004): *Children's narratives and patterns of cardiac reactivity*, in: Developmental Psychobiology Volume 44, Seite 238.

Buchreport (2017): *Mit Jellybooks lernen lernen Verlage und Autoren die Leser kennen*, Link: https://www.buchreport.de/2017/11/06/leser-machen-unglaublich-gerne-mit-bei-jellybooks/ [zuletzt aufgerufen: 01.10.2018]

Kuzmicova, Anezka / Schilhab, Theresa / Burke, Michael (2018): *m-Reading: Fiction reading from mobile phones*, in: Convergence: The International Journal of Research into New Media Technologies, Seite 1-17.

Laugwitz, B. / Schrepp, M. / Held, T. (2008): *Construction and evaluation of a user experience questionnaire*, in: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

Richter, M. / Miltner, W. / Weiss, T. (2011): *Schmerzwörter aktivieren schmerzverarbeitende Hirnareale*, in: Schmerz Volume 25, Seite 322.

Riese, Katrin / Bayer, Mareike / Lauer, Gerhard / Schacht, Annekathrin (2014): *In the eye of the recipient - pupillary responses to suspense in literary classics*

Schwarz-Friesel, Monika (2013): *Sprache und Emotion*, UTB GmbH, Stuttgart

Andreas Hergovich (2018): *Allgemeine Psychologie - Wahrnehmung und Emotion*, facultas.wuv Universitäts, Wien, S. 137.

Tassi, Paul (2018): *The ten horror movies netflix says are so scary, viewers can't finish them*, Link: https://www.forbes.com/sites/insertcoin/2018/03/14/the-ten-horror-movies-netflix-says-are-so-scary-viewers-cant-finish-them/ [zuletzt aufgerufen 01.10.2018].

Multimediale Modelle multimodaler Kommunikation Motion-Capturing in der computergestützten Gestenforschung

Schüller, Daniel

schueller@humtec.rwth-aachen.de
RWTH Aachen, Deutschland

Mittelberg, Irene

mittelberg@humtec.rwth-aachen.de
RWTH Aachen, Deutschland

Unter redebegleitende Gestik werden hier kinetische Arm-, Hand-, und Kopfbewegungen und –konfigurationen verstanden, wie sie von Sprechenden und ihren zuhörenden und zuschauenden Dialogpartnern mehr oder weniger bewusst im Rahmen der mündlichen Face-to-Face-Kommunikation eingesetzt werden (vgl. Kendon 2004, Müller 1998, Müller et.al. 2013, 2014). Bei redebegleitenden Gesten handelt es sich stets um performative, temporale und spontane Vollzüge. Im Unterschied zur Lautsprache lassen sich zunächst keine expliziten syntaktischen und semantischen Regeln erkennen, die das Phänomen unterfüttern. Somit erscheint redebegleitende Gestik (im Unterschied zu sog. Emblemen) als idiosynkratischer, aber dennoch integraler und auch musterhaft unterfütternder Bestandteil von kommunikativer Interaktion vermittels sog. *natürlicher* (körpereigener) *Medien* . Gesprochene Sprache wird somit als multimodaler Vollzug beschrieben, von dem sich die diskursintegrierten Gesten nicht im Sinne einer eigenen medialen Spur abtrennen lassen.

Neben den etablierten, überwiegend rein qualitativ arbeitenden Methoden zur Untersuchung redebegleitender Gestik entwickeln sich zunehmend computerbasierte Verfahren, welche zusätzlich zur qualitativen Analyse von Videomaterial auch quantitative, auf numerische Daten gestützte Analyseperspektiven eröffnen. Im Natural Media Lab der RWTH Aachen werden beispielsweise Gesprächspartner mittels eines markerbasierten, optischen 3D- *Motion-Capture* -Systems aufgenommen und aus dem Verbund von MoCap-, Video- und Audiodaten ein Korpus erstellt. Die multimediale Transkription (Jäger 2004, 2012) versetzt das Korpus in den theoretischen Status eines digitalen Modells (Schüller und Mittelberg, 2017), das für GestenforscherInnen im Sinne eines empirischen Relativs an die Stelle der (Summe der) realen, raumzeitlichen Bewegungen der Sprechenden tritt und die verschiedenen Modalitäten (gesprochene Sprache, Gestik) der Kommunikationssituation erfasst.

Der Vorteil dieses in den letzten Jahren mit Informatikern der RWTH Aachen entwickelten Verfahrens gegenüber der klassischen Videoanalyse ergibt sich hierbei aus der zusätzlichen multimedialen Verschränkung der Annotationen des Gestenforschers mit den numerischen Daten des MoCap-Systems. Diese Methodik unterstützt und ergänzt die Analyse redebegleitender Gesten insofern, als diskursintegrierte Gestik nun auch probandenübergreifend und quantitativ untersucht werden kann, indem auf den MoCap-Daten ein digitaler Algorithmus als operative Instanz eines Ähnlichkeitsmodells arbeitet (Beecks et. al. 2015, 2016; Schüller et.al. 2017), der mittels Gesten-Signaturen algorithmisch nach Ähnlichkeiten im kinetischen Verhalten der Probanden sucht. Weiterhin eröffnet der notationale Charakter (Goodman 1997) des digitalen Modells die Möglichkeit zur Erstellung von Diagrammen (Schüller und Mittelberg 2016) zur Visualisierung quantitativer Daten wie Beschleunigung, Abstände von Händen und Armen zum Körper, Nutzung des Gestenraumes (McNeill 1992) etc., während der multimodalen Artikulation gesprochener Sprache. Drittens gibt es die Möglichkeit der Projektion von Bewegungsspuren auf die Videodaten, sodass auch eine visuelle Repräsentation der ansonsten nicht sichtbaren numerischen Daten erfolgt und die Videodaten anreichert.

Erkenntnis- und wissenschaftstheoretisch betrachtet ist eine Reflexion des theoretischen Status solcher Modelle hochinteressant, da es sich um mittels digitaler, technischer Verfahren erzeugte, zeichenbasierte Transkriptionen realer Ereignisse handelt, die letztlich den Untersuchungsgegenstand empirischer Forschung bilden und somit zentral an der Gegenstandskonstitution beteiligt sind (Jäger 2012). Welchen Einfluss hat die Anwendung verschiedener Abstraktionslevel auf die Gegenstandskonstitution? Welche Zeichenprozesse sind an diesem transkriptiven Verfahren beteiligt?

In unserer Posterpräsentation wollen wir zunächst den eigentlichen Forschungsgegenstand – multimodalen Sprachgebrauch in körpereigenen Medien am Beispiel des Verbundes aus gesprochener Sprache und redebegleitender Gestik – erfassen. Hierauf folgt eine zeichentheoretische Beschreibung der technischen, transkriptiven Verfahren der Korpuskompilation im Natural Media Lab der RWTH Aachen. Darauf aufbauend soll ein Fokus auf die wechselseitige multimediale Verschränkung von Annotationen, Sprach- und Videodaten, numerischen MoCap-Daten, Algorithmus, und den hier jeweils involvierten Abstraktionsleveln gelegt werden.

Bibliographie

Beecks, C. / Hassani, M. / Hinnell, J. / Schüller, D. / Brenger, B. / Mittelberg, I. / Seidl, T. (2015): *Spatiotemporal Similarity Search in 3D Motion Capture Gesture Streams*, in: Proceedings of the 14th International Symposium on Spatial and Temporal Databases (SSTD), Seoul, South Korea, August 26-28, 2015. S.355–372.

Beecks, C. / Hassani, M. / Hinnell, J. / Schüller, D. / Brenger, B. / Mittelberg, I / Seidl, T. (2016): *Efficient Query Processing in 3D Motion Capture Databases via Lower Bound Approximation of the Gesture Matching Distance*, in: International Journal of Semantic Computing, 10(1), 5-25.

Goodman, Nelson (1997): *Sprachen der Kunst. Entwurf einer Symboltheorie*. Frankfurt a. M.: Suhrkamp.

Jäger, Ludwig (2004): *Transkription - zu einem medialen Verfahren an den Schnittstellen des kulturellen Gedächtnisses*, in: TRANS Internet-Zeitschrift für Kulturwissenschaften, 15 Nr., September 2004.

Jäger, Ludwig (2012): *Transkription*, in: **Christina Bartz / Ludwig Jäger / Marcus Krause / Erika Linz (eds.):** *Handbuch der Mediologie. Signaturen des Medialen*. München: Fink, S.306-315.

McNeill, David (1992): *Hand and mind: What Gestures Reveal about Thought*. Chicago: Chicago University Press.

Müller, Cornelia (1998): *Redebegleitende Gesten. Kulturgeschichte - Theorie – Sprachvergleich*. Berlin: Berliner Wissenschafts-Verlag.

Müller, Cornelia / Alan Cienki / Ellen Fricke / Silva Ladwig / David McNeill / Sedinha Teßendorf (eds.) (2013): *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (Handbooks of Linguistics and Communication Science 38.1). Berlin, Boston: Mouton de Gruyter.

Müller, Cornelia / Alan Cienki / Ellen Fricke / Silva Ladwig / David McNeill / Jana Bressem (eds.) (2014): *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (Handbooks of Linguistics and Communication Science 38.2). Berlin, Boston: Mouton de Gruyter.

Schüller, D. / Beecks, C. / Hassani, M. / Hinnell, J. / Brenger, B. / Seidl, T. / I. Mittelberg (2017): *Automated Pattern Analysis in Gesture Research: Similarity Measuring in 3D Motion Capture Models of Communicative Action*, in: Digital Humanities Quarterly, 2017:11.2.

Schüller, D. / Mittelberg, I. (2016/ erschienen 2018): *Diagramme von Gesten: Eine zeichentheoretische Analyse digitaler Bewegungsspuren*, in: Zeitschrift für Semiotik 38, 3-4, S.7-38.

Schüller, D. / Mittelberg, I. (2017/ erschienen 2018): *Motion-Capture-gestützte Gestenforschung. Zur Relevanz der Notationstheorie in den Digitalen Geisteswissenschaften*, in: Zeitschrift für Semiotik 39, 1-2, S.109-146.

Multimedia Markup Editor (M3): Semi-Automatische Annotationssoftware für statische Bild-Text Medien

Moisich, Oliver
moisich@gmail.com
Universität Paderborn, Deutschland

Hartel, Rita
rst@upb.de
Universität Paderborn, Deutschland

Dieses Poster stellt ein Annotationssystem für graphische Narrative vor. Mittels einer auf Java basierten graphischen Oberfläche sind Annotator_innen in der Lage, graphische Narrative und andere statische Bild/Text-Medien in XML zu annotieren. Grundlage für die hier verwendete XML-basierte Annotationssprache, „Graphic Novel Markup Language", kurz GNML, ist die von John Walsh entwickelte, auf TEI basierende, „Comic Book Markup Language", kurz CBML. Abbildung 1 zeigt einen Überblick über die in GNML darstellbaren Objekt-Typen und deren Beziehungen.

Das System ermöglicht Annotator_innen die Annotation verbaler und visueller Repräsentation auf der Comicseite; dazu zählen unter anderem Panels, Charaktere, Sprechblasen, Captions, Erzähltext, Sprechtext, Onomatopoeia und diegetischer Text. Figurenkonstellationen können über verschiedene spezialisierte Interaktionsmuster festgehalten werden. Darüber hinaus erfassen narratologische Tools eine Übersicht über subjektive Filterung einer Erzählsituation (Fokalisierung), Diegese einer Erzählsituation sowie die Hierarchisierung von Erzählwelten.

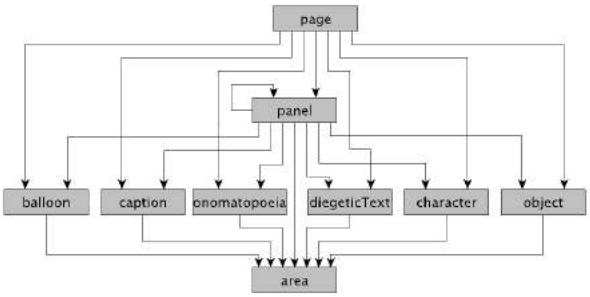


Abbildung 1. Überblick über die verschiedenen Objekt-Typen in GNML

Features

Das Annotieren wird Nutzern durch die Automatisierung einiger Prozesse erleichtert. So lokalisiert das Annotationssystem auf Basis von Verfahren, wie z.B. dem aus der Computergraphik bekannten Marching Squares Algorithmus, die Konturen der Panels. Einfachere Strukturen

wie z.B. Sprechblasen oder Captions müssen vom Nutzer nur durch einen einfachen Klick auf den Hintergrund selektiert werden, so dass danach mit Hilfe eines Floodfill-Verfahrens die komplette Sprechblase erkannt wird. Komplexere Repräsentationen wie Charaktere können mithilfe einer auf der Livewire Segmentation (Mortensen & Barrett, 1995), die beim ungenauen Umranden eines Charakters sich an dessen kontrastreiche Kanten heftet und so die Kontur des Charakters genau erfasst, effizient graphisch annotiert werden. Eine automatische Texterkennung ist derzeit noch nicht integriert, da diese aufgrund der schwer zu erfassenden, oft verwendeten pseudo-handschriftlichen Fonts noch keine befriedigenden Ergebnisse liefert. Dieser wird aber hoffentlich für spätere Versionen in einer ausreichenden Qualität entweder direkt integriert in den Editor oder als optionaler Vorverarbeitungsschritt verfügbar sein. Um aber dennoch die Fehlerrate bei der manuellen Eingabe der Texte und Charakter-Namen möglichst gering zu halten, sind Mechanismen wie eine automatische Rechtschreibprüfung oder Autovervollständigung eingebaut.

Die Beta-Version des Annotationssystems wurde zusammen mit einer Gruppe von Studierenden am „Graphic Narrative Corpus“ (GNC), dem ersten digitalen Korpus für englischsprachige graphische Narrative, getestet (Dunst, Hartel, & Laubrock, 2017). Eines der Hauptziele des Annotationssystems ist die quantitative Analyse, welche auch und vor allem etablierte narratologische Terminologie auf empirische Evidenzen hin prüft. Die Operationalisierung bestehender narratologischer Diskurse in einer für digitale und quantitative Forschung optimierten Testumgebung erfordert die umfassende Gestaltung einer narratologischen Annotationsebene. Grundlage dieser Annotationsebene sind sowohl Theorien der kognitiven und transmedialen Narratologie als auch empirische Daten, die mithilfe des vorgestellten Annotationssystems erhoben und anschließend analysiert wurden.

Abbildung 2 zeigt die Oberfläche des Annotationssystems mit einer Beispielannotation.

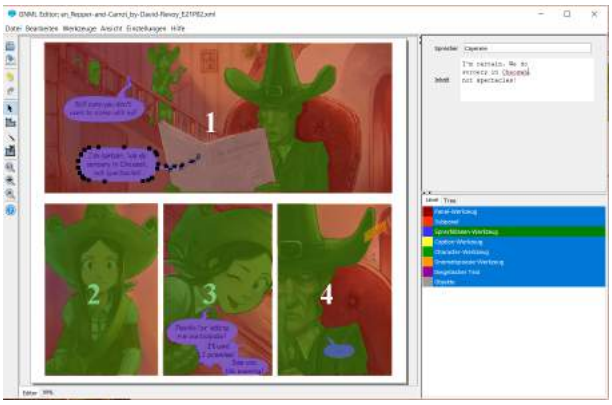


Abbildung 2. Graphische Oberfläche und Beispielannotation des Comics „Pepper & Carrot“ (Revoy, 2017)

Ausblick

Durch die quantitative Operationalisierung etablierter qualitativer narratologischer Terminologie fördert das Annotationssystem die Erweiterung digitaler und quantitativer Methodologien in den Geisteswissenschaften. Komplexe Text-Bild-Interaktionen in graphischen Narrativen werden so mithilfe des Annotationssystems der Analyse von größeren Korpora eröffnet und ermöglichen eine weitreichende digitale Analyse multimodaler Narrative. Gleichzeitig trägt die evidenzbasierte Untersuchung dieser Elemente in zweierlei Hinsicht zu einem besseren Verständnis theoretischer Modelle bei: Zum einen zeigen die annotierten Daten die Möglichkeiten und Grenzen traditioneller Begrifflichkeiten an konkreten Analysebeispielen auf; zum anderen bietet die Auseinandersetzung von Annotator_innen mit komplex erzählten graphischen Narrativen (beispielsweise im Seminarkontext) einen hohen didaktischen Wert insofern, als dass sich Annotator_innen anhand von Fallbeispielen mit narratologischen Konzepte und Problemen kritisch auseinandersetzen.

Die letzte Version des Annotationssystems ist frei verfügbar unter:

http://graphic-literature.upb.de/?page_id=3592

Eine FAQ zum Annotieren mit dem Programm ist zu finden unter:

http://graphic-literature.upb.de/?page_id=4123

Bibliographie

Dunst, A. / Hartel, R. / Laubrock, J. (2017): *The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities*. 2nd International Workshop on coMics Analysis, Processing, and Understanding, 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan.

Mortensen, E. / Barrett, W. (1995): *Intelligent scissors for image composition*. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques (pp. 191-198). ACM.

Revoy, D. (2017): *Pepper & Carrot. Episode 21: The Magic Contest*. (Pepper and Carrot) Retrieved 2018 10, from https://www.peppercarrot.com/en/article400/episode-21-the-magic-contest

Multimodale Sentimentanalyse politischer Tweets

Ziehe, Stefan
stefan.ziehe@stud.uni-goettingen.de
GCDH, Universität Göttingen, Deutschland

Sporleder, Caroline
csporled@gwdg.de
GCDH, Universität Göttingen, Deutschland

Die automatische Analyse von Tweets mit politischem Inhalt kann Sozial- und Politikwissenschaftlern Aufschluss über die Prozesse politischer Meinungsbildung geben. Beiträge in den sozialen Medien spiegeln oft Tendenzen in der Zufriedenheit mit politischen Parteien wider und helfen

umstrittene oder viel diskutierte Themen zu identifizieren. Auch das Kommunikationsverhalten verschiedener Gruppen lässt sich aus ihrer Interaktion bei Twitter analysieren (z.B. unterschiedliche Dominanz von Echokammereffekten (Colleoni et al., 2014) oder Verbreitung von Gerüchten und Fake News (Ma et al., 2018)).

Eine wichtige Rolle spielt dabei die Sentimentanalyse, die es erlaubt zu identifzieren, ob ein Akteur Zustimmung oder Ablehnung zu einem bestimmten Inhalt signalisiert. Für Textdaten lässt sich das Sentiment meist recht gut bestimmen. Tweets sind aufgrund ihrer Kürze jedoch zum einen oft schwer recht kryptisch, zum andern enthalten sie häufig weitere Materialien, insbesondere Bilder, die nennenswert zur Aussage beitragen. Z.B. ist der Text „und wieder ein neuer Morgen" neutral formuliert, gewinnt aber eine negative Bedeutung, wenn er um eine Bild bereichert wird, das eine lange Autoschlange zeigt. Ebenso kann es sein, dass ein relativ neutrales Bild lediglich über den Text ein positives oder negatives Sentiment zugewiesen bekommt.

Die meisten bisher existierenden Sentimentanalyseverfahren beschränken sich auf die Verarbeitung entweder von Text- oder von Bilddaten. Modelle, die beide Modalitäten

berücksichtigen, sind noch vergleichsweise selten, können jedoch eine signifikant höhere Genauigkeit bei der Sentiment-Vorhersage erreichen als solche, die dies nicht tun (You et al. 2016). Fast alle multimodalen Verfahren nutzen eine Deep-Learning-Architektur. Solche Verfahren sind herkömmlichen Lernverfahren zwar oft überlegen, sie sind aber aufgrund der Vielzahl der möglichen Architekturen auch relativ schwer zu optimieren. Das Ziel dieser Arbeit ist es, verschiedene multimodale Sentimentanalyseverfahren und -architekturen systematisch zu vergleichen und auf ihre Vor- und Nachteile hin zu untersuchen.

Das grundsätzliche Schema der Modelle orientiert sich am "Latent Multimodal Mixing" (Bruni et al. 2014); hierbei werden zunächst Text- und Bild-Features extrahiert, als Vektoren kodiert

und anschließend in einem dritten Schritt auf einen gemeinsamen (multimodalen) Vektorraum abgebildet (Fusion). Aus diesen Vektoren kann dann mit Methoden des maschinellen Lernens das Sentiment berechnet werden. Innerhalb dieses Schemas können beliebige und auch neuartige Kombinationen verschiedener Methoden zur Feature-Extraktion und Fusion verwendet werden. Hierfür gibt es unter anderem folgende Möglichkeiten:

- Texte können mit dem Doc2Vec-Verfahren auf einen latenten Vektorraum abgebildet werden. Dieses Verfahren erzielte in der Vergangenheit bereits gute Ergebnisse bei der Sentimentanalyse. (Le et al. 2014)
- Basierend auf einem existierenden Word-Embedding-Modell (z.B. GloVe (Pennington et al. 2014)) können die Word-Embeddings aller Wörter eines Textes auf verschiedene Arten zu einem Text-Embedding aggregiert werden (z.B. gewichteter Mittelwert, elementweises Minimum/Maximum). (De Boom et al. 2016)
- Für die Extraktion visueller Features können bereits existierende Deep Learning-Modelle zur Bild-Klassifikation in leicht modifizierter Form wiederverwendet werden. (Campos et al. 2017)
- Aus dem Farbhistogramm eines Bildes können können statistische Features erster Ordnung berechnet werden.
- Der Fusionsschritt besteht aus einer einfachen Verkettung der Text- und Bild-Vektoren. Zusätzlich kann auch eine affine Projektion auf einen latenten multimodalen Vektorraum gelernt werden. (Chen et al. 2017)

Die Datengrundlage für das Training der Modelle bilden manuell annotierte multimodale Tweets u.a. aus dem Photo Tweet Sentiment Benchmark (Borth et al. 2013), sowie das Columbia MVSO Image Sentiment Dataset (Dalmia et al. 2016). Aufgrund der unterschiedlichen Größe der Datensätze wird ein Transfer Learning-Ansatz verfolgt: Die Modelle werden zunächst auf den MVSO-Daten vortrainiert und anschließend mithilfe der Twitter-Daten feinadjustiert.

Erste Testergebnisse bestätigen, dass Modelle, die Text- und Bild-Features fusionieren, eine höhere Genauigkeit erreichen können als unimodale Modelle. Allerdings haben die bisher getesteten Modelle derzeit noch Schwierigkeiten damit, negatives Sentiment korrekt zu klassifizieren.

Bibliographie

You, Quanzeng / Luo, Jiebu / Jin, Hailin / Yang, Jianchao (2016): *Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia*, in: The Ninth International Conference on Web Search and Data Mining, February 2016, San Francisco, CA, USA.

Bruni, Elia / Tran, Nam Khanh / Baroni, Marco (2014): *Multimodal distributional semantics*, in: Journal of Artificial Intelligence Research 49, 1 (Januar 2014), 1-47

Colleoni, Elanor/ Rozza, Alessandro / Arvidsson, Adam (2014): *Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data*, in: Journal of Communicatio 64 (2014) 317–332

De Boom, Cedric / Van Canneyt, Steven / Demeester, Thomas / Dhoedt, Bart (2016): *Representation learning for very short texts using weighted word embedding aggregation*, in: Pattern Recognition Letters 80, C (September 2016), 150-156.

Campos, Victor / Jou, Brendan / Giró-i-Nieto, Xavier (2017): *From Pixels to Sentiment: Fine-uning CNNs for Visual Sentiment Prediction*, in: Image and Vision Computing 65 (September 2017), 15-22

Chen, Xingyue / Wang, Yunhong / Liu, Qingjie (2017): *Visual and Textual Sentiment Analysis Using Deep Fusion Convolutional Neural Networks*, in: IEEE International Conference on Image Processing (ICIP), Beijing, China

Dalmia, Vaidehi / Liu, Hongyi / Chang, Shih-Fu (2016): *Columbia MVSO Image Sentiment Dataset*, in: CoRR, abs/1611.04455

Borth, Damian / Ji, Rongrong / Chen, Tao / Breuel, Thomas / Chang, Shih-Fu (2013): *Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs*, in: The 21st ACM International Conference on Multimedia, October 2013, Barcelona, Spain

Le, Quoc / Mikolov, Tomas (2014): *Distributed Representations of Sentences and Documents*, in: The 31st International Conference on Machine Learning, June 2014, Beijing, China

Ma, Jing / Gao, Wei / Wong, Kam-Fai (2018) *Rumor Detection on Twitter with Tree-Structured Recursive Neural Networks*, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1980–1989, Melbourne, Australia, July 15 - 20, 2018

Pennington, Jeffrey / Socher, Richard / Manning, Christopher (2014): *GloVe: Global Vectors for Word Representation*, in: Empirical Methods in Natural Language Processing (EMNLP), 1532-1543

Multimodales Zusammenspiel von Text und erlebter Stimme – Analyse der Lautstärkesignale in direkter Rede

Guhr, Svenja
svenjasimone.guhr@stud.uni-goettingen.de
GCDH, Universität Göttingen, Deutschland

Varachkina, Hanna
hanna.varachkina@stud.uni-goettingen.de
GCDH, Universität Göttingen, Deutschland

Lee, Geumbi
geumbi.lee@stud.uni-goettingen.de
GCDH, Universität Göttingen, Deutschland

Sporleder, Caroline
csporled@gwdg.de
GCDH, Universität Göttingen, Deutschland

Die Betrachtung von Erzählformen und Figurenrede stellt in der Literaturwissenschaft ein wichtiges Analysekriterium dar. Bereits 1969 setzte sich Gérard Genette mit dem repräsentationslogischen Verhältnis in Prosatexten auseinander. Dabei betrachtete er die Zeit, den Raum und die Figuren, jedoch ließ er die Art und Weise des Sprechakts außen vor (Genette, 1969). Einige Jahrzehnte später entwickelte sich in der Narratologie ein neuer Forschungszweig, der sich mit der Figurenrede auseinandersetzt. Die Audionarratologie analysiert den Zusammenhang zwischen Geräuschen und narrativen Formen in Texten und befasst sich mit dem geräuschvollen Erlebnis des leisen Lesevorgangs in der Vorstellungskraft der Leserinnen und Leser (i.F. Leser) 8Mildorf / Kinzel, 2016, Kuzmǐcová, 2013).

Autorinnen und Autoren (i.F. Autoren) nutzen die direkte Rede als Mittel, um den Figuren eine Stimme zu geben und sie in den Köpfen ihrer Leser sprechen zu lassen (Nord, 1997). Als Möglichkeiten bestehen Geräuschbeschreibungen aber auch die redeeinleitenden Verben, Verba Dicendi, die bei der Verschriftung der Figurenrede eine relevante Rolle spielen. Diese Verbgruppe beschreibt die Art und Weise, wie die Leser sich die Konversation der Figuren vorzustellen haben, d.h. ob die Romanfiguren z.B. schreiend oder flüsternd kommunizieren. Verba Dicendi können anhand ihrer multimodalen Eigenschaft das Inhaltsverständnis unterstützen. Sie beschreiben die Sprechsituation und die Realisierung der direkten Rede und dienen dem multimodalen Zusammenspiel von Text und erlebter Stimme.

Katsma analysierte die in Romanen dargestellte Lautstärke von Dialogen computergestützt, indem er die Verba Dicendi in Lautstärkeniveaus einteilte und daran den Dynamikverlauf aufgeteilt auf Romankapitel verfolgte und die Dialogbeiträge von Figuren in ein Lautstärkediagramm einordnete (Katsma, 2014). Auch Chapman (1984) und Nord (1997) beschäftigten sich mit Lautstärke in Prosatexten. Hierbei nannten sie den Zusammenhang zwischen beschreibenden adverbialen Bestimmungen und lautstärkeneutralen Verba Dicendi wie z.B. in „called loudly“ (Chapman, 1984: 95) oder „said [...] hastily“ (Nord, 1997: 114).

Im Rahmen einer Pilotstudie haben wir die Verba Dicendi und die sie beschreibenden Adverbien näher betrachtet mit dem Ziel stichprobenartig zu untersuchen, inwiefern die Lautstärkesignale eines Textes mit literaturwissenschaftlich relevanten Kategorien korrelieren. Eine schon von Katsma (2014) geäußerte Hypothese ist, dass Lautstärke mit „Emotionalität“ zusammenhänge. Eine literarische Strömung wie der Expressionismus, der durch eine Betonung des inneren Ausdrucks und eine Ablehnung des Rationalen charakterisiert ist, müsste sich daher durch große Schwankungen zwischen lauten und leisen Passagen auszeichnen.

Um diese Hypothese zu untersuchen wurden drei Korpora (insg. 161 Texte) zusammengestellt. Das erste besteht aus 57 Prosatexten um 1900 von Autoren, die der expressionistischen Literaturströmung zugeordnet werden können (Anz, 2016: 5f.). Als Vergleichsbasis dient ein zweites Korpus desselben Zeitraums von Autoren, die verschiedenen literarischen Strömungen zugeordnet werden (Fin de Siècle, Exilliteratur, Naturalismus, Realismus). Das dritte Korpus, bestehend aus 14 kurzen Texten von Autoren um 1900, dient als manuell annotiertes Kontrollkorpus der Evaluierung der Ergebnisse.

Als erster Schritt im methodischen Vorgehen wurden aus den zu betrachtenden Korpora mit einem regelbasierten Verfahren die verwendeten Verba Dicendi automatisch identifiziert und extrahiert. Als Grundlage für die Entwicklung des verwendeten Algorithmus wurden die morphologischen Eigenschaften der regelmäßigen und unregelmäßigen standarddeutschen Verba Dicendi im Präteritum und Präsens sowie die sie umgebenden Satzzeichen beachtet. Daraufhin wurden die Verben mit den Ergebnissen aus einer vorhergehenden Studie zum Lautstärkeempfinden von deutschen Verba Dicendi (Onlineumfrage, 2018) abgeglichen. Hierbei generierte Prozentwerte ergaben Abstufungen von schreien (100%) über rufen (91.4%), sagen (25.5%) und flüstern (14%) bis denken (2%). Die daraus gezogenen Werte der extrahierten redeeinleitenden Verben wurden in Abhängigkeit zur Textlänge aufsummiert, sodass ein Lautstärkeprofil pro Prosatext erstellt werden konnte.

Zudem wurden Adjektive und Adverbien, die die Verba Dicendi als Träger der Beschreibungsinformationen der direkten Rede umgeben, in einem Fenster von vier Tokens vor und nach dem Verb in die Betrachtung miteinbezogen. Dabei wurde herausgefunden, dass vor allem die Adverbien und Kollokationen (z.B. „mit lauter Stimme“) für das Lautstärkeprofil relevant sind.

Mithilfe der Erkenntnisse aus der Analyse der lautstärkevermittelnden Verba Dicendi und der sie

beschreibenden Adverbien soll im weiteren Vorgehen untersucht werden, ob Lautstärkehinweise in Prosatexten neue Erkenntnisse für die Literaturwissenschaft im Hinblick auf Einordnungskriterien in verschiedene literarische Strömungen liefern können. Auch soll der Einfluss, den Beschreibungen von Figurenrede auf das Leseempfinden von Diskurspassagen haben, analysiert werden. Hierbei sollen die expressionistischen Prosatexte als Vergleichsbasis dienen, um Gemeinsamkeiten und Unterschiede im Hinblick auf die Lautstärkekriterien zwischen den Werken des expressionistischen Kontrollkorpus und den Werken aus dem heterogenen Korpus mit Prosatexten um 1900 herauszustellen.

Bibliographie

Anz, Thomas (2016 2): *Literatur des Expressionismus*, Springer-Verlag GmbH Deutschland.

Bakhtin, Mikhail (1981): *Forms of Time and of the Chronotope in the Novel*, in: The Dialogic Imagination, Texas: University of Texas Press.

Chapman, Raymond (1984): *The Treatment of Sounds in Language and Literature*, Basil Oxford: Blackwell Publisher Limited.

Dahl, Alva (2014): *The separation of voices in a literary utterance: a dialogical approach to discourse presentation, viewpoint, focalization – and punctuation*, Online Proceedings of the Annual Conference of the Poetics and Linguistics Association (PALA).

Delazari, Ivan (2016): *Voicing the Split Narrator: Readers' Chores in Toni Morrison's "Recitatif"*, in: **Mildorf, Jarmila / Kinzel, Till (2016):** *Audionarratology. Interfaces of Sound and Narrative*, De Gruyter.

Genette, Gérard (1969): *Figures*. Essais, Paris: Editions du Seuil.

Katsma, Holst (2014): *Loudness in the Novel*, Stanford, US: Standford Literary Lab.

Kuzmičová, Anežka (2013): *Outer vs. Inner Reverberations: Verbal Auditory Imagery and Meaning-Making in Literary Narrative*, Journal of Literary Theory 7 (1-2): 111-134, https://philpapers.org/archive/KUZOVI.pdf [zuletzt abgerufen 04.01.2019].

LimeSurvey (2018): *LimeSurvey Manual*, https://manual.limesurvey.org/LimeSurvey_Manual.

Mildorf, Jarmila / Kinzel, Till (2016): *Audionarratology: Prolegomena to a Research Paradigm Exploring Sound and Narrativey*, in: **Mildorf, Jarmila / Kinzel, Till (2016):** *Audionarratology. Interfaces of Sound and Narrative*, De Gruyter.

Nord, Christiane (1997): *Alice abroad: Dealing with descriptions and transcriptions of paralanguage in literary translation*, in: **Poyatos, Fernando (1997):** *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 107-130.

Scarry, Elaine (2001): *Dreaming by the Book*, Princeton: Princeton UP.

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

Winkler, Edeltraud (1987): *Syntaktische und semantische Eigenschaften von Verba Dicendi*, Akademie der Wissenschaften, DDR.

Museum Analytics: Ein Online-Tool zur vergleichenden Analyse musealer Datenbestände

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland

Kohle, Hubertus
hubertus.kohle@lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Burg, Severin
severin.jo.burg@gmail.com
Università di Bologna, Italien

Küchenhoff, Helmut
kuechenhoff@stat.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland

Museen produzieren in den letzten Jahren verstärkt digitale Inventare ihrer Bestände, die sie zuweilen auch im Internet veröffentlichen: Manche stellen Teilbestände zur Verfügung, andere den ganzen Besitz; wie etwa das Metropolitan Museum in New York, das über 375.000 Objekte datenbankmäßig erschlossen präsentiert.[1] Üblicherweise werden einzelne Objekte aus diesen Inventaren gefiltert; entweder solche, die bekannt sind, oder andere, auf die man über erschließende Metadaten stößt. „Museum Analytics", kurz „MAX", ein an der Ludwig-Maximilians-Universität München am Institut für Kunstgeschichte und Institut für Statistik entwickeltes Online-Tool, sieht etwas anderes vor: Es ermöglicht es, im Unterschied zu anderen Systemen (Stack 2018), die vorhandenen Metadaten als „Massendaten" statistisch zu erschließen, zu analysieren und zu visualisieren – ohne programmieren zu können. Damit eignet es sich besonders für die Lehre in geisteswissenschaftlichen Kontexten.[2]

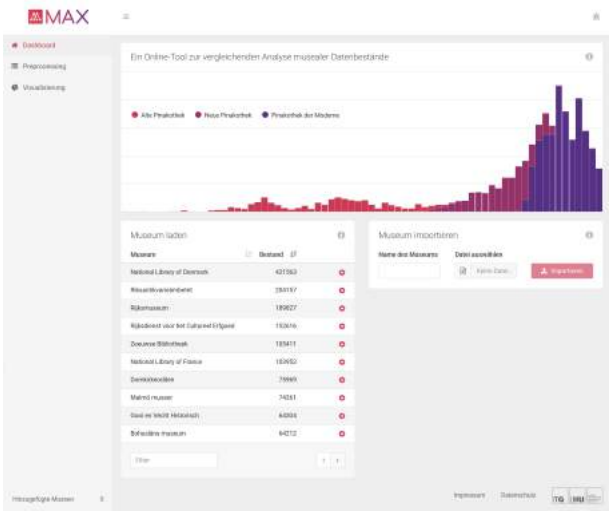


Abbildung 1: Screenshot des Moduls

Dashboard, das das Laden und Importieren von Museen ermöglicht.

„MAX" implementiert folgende Komponenten:

1. Ein *Graphical User Interface*, das responsiv ist und ohne exzessive Einarbeitung schnelle Fortschritte und eine flüssige Analyse ermöglicht. Es ist mit der Open-Source-Programmiersprache *R* und dem auf *R* aufsetzenden Webapplikationspaket *Shiny* realisiert, die eine leichte Modularisierung im *Backend* garantieren (R Core Team, 2017; Chang et al., 2017). Hilfestellungen werden durch eine allgemeinverständliche Dokumentation und den jeweiligen Funktionen zugeordnete *Screencasts* gegeben, die einzelne Verfahrensschritte näher erläutern.
2. Eine Schnittstelle zu musealen *Application Programming Interfaces* (*APIs*), um bestehende Daten möglichst einfach in das Tool importieren und weiterverarbeiten zu können. Momentan bereitgestellt werden fast 4.000.000 Objekte aus 200 Institutionen. Auch eigene, etwa per *Web Scraping* extrahierte und als CSV- oder RDS-Datei vorliegende Bestände können eingespeist und bearbeitet werden. Daher löst sich die Beschränkung auf museale Inventare: Wer sich für die Entwicklung von Bevölkerungszahlen interessiert, kann „MAX" ebenfalls nutzen.
3. Komplementiert werden diese Komponenten mit der Javascript-Bibliothek *Highcharts* durch dynamische und interaktive Grafiken, die bei *Mouseover* weitere Kennzahlen anzeigen oder mittels *Zoom* interessierende Bereiche vergrößern oder selegieren können.[3] Sie bereichern die statistische Analyse, indem sie komplexe Zusammenhänge attraktiv abbilden und die Nutzerinnen und Nutzer unmittelbar in die Analyse einbinden.

Diese Komponenten sind in drei Modulen integriert: *Dashboard*, *Preprocessing* und *Visualisierung*. *Dashboard* dient der Auswahl und dem Import der jeweils interessierenden Museen, die in *Preprocessing* diversen statistischen Operationen unterzogen werden können, um bspw. heterogene Datierungsangaben zu vereinheitlichen. *Visualisierung* präsentiert die Ergebnisse in grafischer Form, wobei die verschiedenen Diagrammarten, z. B. Histogramm oder *Bubble Chart*, automatisch je nach Datentyp der zu visualisierenden Spalte eines oder mehrerer Museen angewandt werden. Ein besonderes *Feature* ist mit der *Historie* verbunden: Einzelne Arbeitsschritte – oder ganze Arbeitsschrittketten – können mit ihr reversibel gemacht oder auf andere zu bearbeitende Museen übertragen werden. Es wurden sowohl Standardmethoden der Statistik implementiert als auch für die geisteswissenschaftliche Disziplin spezifische Schnittstellen, um bspw. die Entstehungsländer der Kunstwerke einer Sammlung in einer Karte visualisieren zu können.

Je nach Erschließungstiefe der Bestände kann untersucht werden, welche Sammlungskonjunkturen es in bestimmten Museen gegeben hat, welche Gattungen zu welcher Zeit besonders beliebt waren oder ob es Zusammenhänge zwischen der Sammlungstätigkeit und gesellschaftspolitischen Bedingungen gab. Bei hinreichend repräsentativer Datenlage lassen sich auch Fragen adressieren, die die Kunstgeschichte als solche betreffen: Welche künstlerischen Techniken waren im historischen Wandel vorherrschend? Welche Themen zu welcher Zeit beliebt? Ist die Säkularisierung an der Entwicklung der Ikonographie der Kunstwerke abzulesen? Die Werke bzw. deren Digitalisate können auch direkt adressiert und statistisch analysiert werden: Ist die Verwendung von bestimmten Farben bzw. Farbzusammenstellungen historisch beschreibbar? Lässt sich Heinrich Wölfflins Bildanalyseinstrumentarium, in dem etwa das Renaissance-Kunstwerk als linear und klar, das barocke als malerisch und unklar beschrieben wurde, automatisieren und auf seine historische Berechtigung befragen?

Die Funktionalitäten des Tools wurden in einem Seminar mit geisteswissenschaftlichen Studierenden im Sommersemester 2018 an der Ludwig-Maximilians-Universität überprüft und evaluiert, worauf mit der Integration der genannten *Screencasts* reagiert wurde. „MAX" ist ein Teil des „Digital Humanities Virtual Laboratory" („DHVLab"), einer modularen Lehr- und Forschungsinfrastruktur zur Ausbildung von Studierenden der Kunst-, Geschichts- und Sprachwissenschaften in Anwendungen und Methoden der *Digital Humanities*.[4]

Fußnoten

1. https://www.metmuseum.org/art/collection (23.09.2018).
2. https://www.max.gwi.uni-muenchen.de/ , das Online-Tool selbst unter https://dhvlab.gwi.uni-muenchen.de/max/ (beide 23.09.2018). Es wird im Rahmen des Programms Lehre@LMU zur Stärkung der Forschungsorientierung in der Lehre gefördert.
3. https://www.highcharts.com/ (23.09.2018).
4. https://dhvlab.gwi.uni-muenchen.de/ (24.09.2018); weitere Einblicke in Klinke (2018).

Bibliographie

Chang, Winston / Cheng, Joe / Allaire, JJ / Xie, Yihui / McPherson, Jonathan (2017): *shiny: Web Application Framework for R*, https://cran.r-project.org/package=shiny .

Klinke, Harald (Hrsg., 2018): *#DigiCampus. Digitale Forschung und Lehre in den Geisteswissenschaften.* München: Universitätsbibliothek der Ludwig-Maximilians-Universität.

R Core Team (2017): *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, https://www.r-project.org/ .

Ridge, Mia (2012): *Mia Ridge Explores the Shape of Cooper-Hewitt Collections*, https://labs.cooperhewitt.org/2012/exploring-shape-collections-draft/ .

Stack, John (2018): *Exploring Museum Collections Online. Some Background Reading,* https://lab.sciencemuseum.org.uk/exploring-museum-collections-online-some-background-reading-da5a332fa2f8 .

Netzwerkanalyse für Historiker. Probleme und Lösungen am Beispiel eines Promotionsvorhabens

Toscano, Roberta
toscano-r@outlook.de
Universität Stuttgart, Deutschland

Einleitung

Methoden der Digital Humanities erfreuen sich zunehmender Akzeptanz in den Geisteswissenschaften - auch in traditionellen Disziplinen wie den Geschichtswissenschaften. Die Werkzeuge der Digital Humanities können hilfreich sein, um die eigenen Daten zu strukturieren, unerwartete Zusammenhänge zu erkennen und Erkenntnisse hervorzubringen, die man ohne den Einsatz von digitalen Methoden nicht erlangt hätte. Vor allem im Bereich der historischen Netzwerkanalyse gibt es seit einigen Jahren neue interessante Entwicklungen. Während man sich anfangs noch an Werkzeugen aus der Sozialwissenschaft bediente (Jannidis / Kohle / Rehbein 2017: 14), gibt es mittlerweile Software, die nicht nur statistische Ergebnisse ausgibt, sondern auch für historische Fragestellungen geeignet ist. Die Vorteile liegen auf der Hand: einerseits kann man komplexe Beziehungen, deren Reichweite und Auswirkungen ordnen und darstellen; andererseits durch Visualisierungen die Ergebnisse einer breiten Öffentlichkeit zugänglich machen (Düring / Kerschbaumer 2016: 31). Wenn man anfangs mit dem Gedanken spielt, ob man für die Bearbeitung einer geisteswissenschaftlichen Forschungsarbeit überhaupt digitale Hilfsmittel anwenden sollte, wird man mit verschiedenen Fragen konfrontiert:

Wie und wo fange ich an? Welches Tool ist sinnvoll? Welche Daten habe ich überhaupt und wie möchte ich diese darstellen? Welchen Mehrwert verspreche ich mir von dem Einsatz digitaler Methoden? Muss ich meine Fragstellungen anpassen um ein zufriedenstellendes Ergebnis zu erhalten? Welche Kenntnisse und Fähigkeiten muss ich mir aneignen? Lohnt sich das überhaupt? Der Posterbeitrag soll an diese Fragen anknüpfen.

Netzwerkanalyse für Historiker

Die eigene Recherche hat gezeigt, dass das Angebot und die Initiativen (national wie international) von "Social Network Analysis" (kurz SNA) für Historiker auf den ersten Blick enorm zu sein scheint. Allerdings stellt sich bei näherer Betrachtung meistens heraus, dass nur einige der Tools für das eigene Thema geeignet sind. Schließlich variieren bei geschichtlichen Forschungsfragen die gewählten Zugänge sowie die Quellenlage und der Blickwinkel. Dieses Poster soll einerseits die meist verbreiteten Techniken für Historiker in der Netzwerkanalyse vorstellen; andererseits soll an konkreten Schritten der Entscheidungsprozess für ein geeignetes Tool zur Netzwerkanalyse beispielhaft am eigenen Dissertationsvorhaben nachvollzogen werden. Aufkommende Probleme und mögliche Lösungswege sollen beleuchtet werden.

Anwendung in der Praxis mit „Gephi“

Mein Forschungsthema befasst sich mit der Darstellung von Palästina in württembergischen Medien des 19. Jahrhunderts. Dafür werte ich politische, religiöse und pädagogische Medien aus. Eine der Hauptfragestellungen ist unter anderem welche Netzwerke gebildet wurden, in denen die Hauptakteure eine gewissen Einfluss übten und dadurch die gegenseitigen Beziehungen beider Länder prägten und förderten. Eine große Rolle spielt die Wissensverbreitung über das damals relativ unbekannte "Heilige Land": wie wurde spezielles Wissen - unter anderem Agrarmethoden, kulturelles und religiöses Leben in Palästina - weitergegeben? Neben den sozialen Beziehungen bestand auch ein reger wirtschaftlicher Austausch zwischen Palästina und Württemberg, der sich unter anderem im Handel äußerte. In diesem Zusammenhang plane ich meine Ergebnisse, die Verbindungen und Verflechtungen, sowie den Transfer zwischen beiden Länder mit der Hilfe von digitalen Tools zu visualisieren. Knotenpunkte wie Personen, Handelsbeziehungen, etc. sollen bei der Analyse im Fokus stehen und die ursprünglich aufgeworfenen Fragen ergänzend bereichern.

Diese Forschungsarbeit und Fragestellungen bieten den Rahmen für die nähere Betrachtung der Möglichkeiten der historischen Netzwerkanalyse mit der Open Software "Gephi". Die zahlreichen Tutorials (https://gephi.org/users/), die aktive Nutzercommunity (https://gephi.wordpress.com/), die intuitive Nutzermaske und die vielfältigen Projekten, die bereits umgesetzt wurden, sind nur einige der Gründe, die für die Anwendung von „Gephi“ sprechen. Diese und andere Entscheidungskriterien, die in diesem Beispiel zur Wahl von „Gephi“ geführt haben, werden im Posterbeitrag veranschaulicht.

Zweck des Posters

Der Beitrag soll einen Überblick zu den existierenden Angeboten von Netzwerkanalysen geben sowie Vor-und Nachteile zur Diskussion stellen. Am Beispiel des eigenen

Promotionsvorhabens werden Anregungen gegeben, Schritte erläutert und der Entscheidungsprozess begleitet. Das Ziel ist es aus der Perspektive eines Anfängers, Möglichkeiten aufzuzeigen wie man anfängliche Herausforderungen meistern kann. Der Beitrag wendet sich an Ein-und Quereinsteiger und soll durch eine konzeptionelle Gestaltung vermitteln, welches Potenzial und welcher Mehrwert in Netzwerkanalysen steckt. Damit stellt dieses Poster zu gleichen Teilen eine Absichtserklärung und einen Erfahrungsbericht dar.

Bibliographie

Bastian M. / Heymann S., Jacomy M. (2009): *Gephi: an open source software for exploring and manipulating networks*, International AAAI Conference on Weblogs and Social Media. https://gephi.org/publications/gephi-bastian-feb09.pdf

Düring, Marten. (2017): *Historical Network Research. Network Analysis in the Historical Disciplines.* **http://historicalnetworkresearch.org/.**

Düring, Marten / Kerschbaumer, Florian (2016): *Quantifizierung und Visualisierung. Anknüpfungspunkte in den Geschichtswissenschaften*, in: **Düring, Marten / Eumann, Ulrich / Stark, Martin / von Keyserlingk, Linda (eds.):** *Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen*. Münster: Lit Verlag 31- 44.

Grandjean, M. (2015): *GEPHI – Introduction to Network Analysis and Visualization*. http://www.martingrandjean.ch/gephi-introduction/

Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (2017): *Digital Humanities. Eine Einführung*. Stuttgart: J. B. Metzler.

Milligan, I. (2015): *From Dataverse to Gephi: Network Analysis on our Data, A Step-by-Step Walkthrough*. https://ianmilligan.ca/2015/12/11/from-dataverse-to-gephi-network-analysis-on-our-data/

Scott, John. *What Is Social Network Analysis?* London: Bloomsbury Academic, 2013.

OCR Nachkorrektur des Royal Society Corpus

Klaus, Carsten
s8caklau@stud.uni-saarland.de
Universität des Saarlandes, Saarbrücken, Deutschland

Fankhauser, Peter
fankhauser@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Klakow, Dietrich
dklakow@lsv.uni-saarland.de
Universität des Saarlandes, Saarbrücken, Deutschland

Einleitung

Linguistische Analysen historischer Texte stellen Forscher oftmals vor große Herausforderungen. Im Gegensatz zur Digitalisierung moderner Dokumente kann es bei jahrhundertealten Texten zu Schwierigkeiten kommen. Diese weisen oftmals eine geringere Qualität auf, sodass es beim Einlesen zu Fehlern kommt. Solche können schwerwiegende Störfaktoren für weitere Analysen sein. In diesem Beitrag beschreiben wir den **Noisy Channel Spell Checker**, ein Verfahren zur automatisierten Korrektur von Optical Character Recognition (OCR) induzierten Rechtschreibfehlern in historischen Texten, genauer dem **Royal Society Corpus**.

Beim Royal Society Corpus (RSC) handelt es sich um eine Sammlung wissenschaftlicher Texte von 1665 bis 1869, veröffentlicht im Journal *Philosophical Transactions of the Royal Society of London.* Das Korpus umfasst ungefähr 10.000 Dokumente mit insgesamt 35.000.000 Tokens. Die Texte wurden mithilfe von Optical Character Recognition digitalisiert, bedingt durch das alte Material der Dokumente wurden jedoch Worte falsch erkannt und somit Rechtschreibfehler eingestreut. Diese sollen in einer Nachkorrektur berichtigt werden. (UdS Fedora Commons o.J.)

State of the Art

Das Korpus wird einer strikten Versionskontrolle unterzogen. Fortschritte bzgl. Formatierung oder Fehlerkorrektur werden in aufsteigenden *corpusBuild* Versionen festgehalten. Derzeit wird das Royal Society Corpus durch einen **Pattern**-basierten Ansatz bereinigt (Knappen, 2017). Hierbei werden Ersetzungsregeln auf die Texte angewendet um Fehler mit ihrer richtigen Form auszutauschen, wie beispielsweise *tbe* → *the*. Der große Nachteil dieses Verfahrens ist jedoch, dass nur ein Bruchteil der induzierten OCR Fehler abgedeckt wird, was in einer geringen Fehlererkennung resultiert. Im Folgenden erläutern wir unseren Ansatz, welcher mit einem statistischen Lernverfahren deutlich bessere Ergebnisse erzielt.

Methodik

Der *Noisy Channel Spell Checker* basiert auf dem **Noisy Channel Model** (Shannon, 1948). Ein potentiell fehlerhaftes Wort w wird wie folgt korrigiert: Aus einer Vorauswahl an geeigneten Kandidaten c aus C wird abgeschätzt welcher am ehesten als Korrektur ŵ in Frage kommt.

$$\hat{w} = \underset{c \in C}{\operatorname{argmax}}\ P(c)^{\lambda} P(w|c)$$

Das Noisy Channel Model besteht zum einen aus dem **Sprachmodell** P(c) und zum anderen dem **Fehlermodell** P(w|c). Es werden hierbei zwei intuitive Gedanken kombiniert: Das Sprachmodell schätzt die Wahrscheinlichkeit des Kandidaten in seinem Wortkontext ab. Hochfrequentierte

Worte sind demnach sehr wahrscheinlich. Das Gegengewicht hierzu bildet das Fehlermodell. Diese Verteilung gibt an wie sicher w eine fehlerhafte Variante von c ist, schätzt also ab, wie wahrscheinlich einzelne Korrekturschritte von w nach c sind. λ ist ein frei wählbarer Parameter, mithilfe dessen man das Sprachmodell gewichten kann. (Jurafsky 2016: 61-73)

Training des Modells

Die Besonderheit unseres Ansatzes besteht darin, dass Sprach-, sowie Fehlermodell **korpusspezifisch** trainiert werden. Es sind keine aufwändigen Trainingsdatenannotationen notwendig, denn es werden lediglich die Korpusdateien verwendet.

- Das **Sprachmodell** wurde mithilfe der aktuellsten *corpusBuild* Version des Royal Society Corpus trainiert. Diese Texte sind durch die Patterns bereits best möglich bereinigt worden. Somit wurde versucht das Rauschen innerhalb der Verteilung zu reduzieren.
- Zum Trainieren des **Fehlermodells** wurden die bereits erwähnten Patterns als Wissensbasis hinzugezogen. Die Idee war hier aus der Korrektur durch die Patterns eine Wahrscheinlichkeitsverteilung zu erzeugen, also das Fehlerverhalten im Korpus zu generalisieren. Anhand eines Beispiels lässt sich dies veranschaulichen: Gegeben die Ersetzungsregel fuch → such. Diese wird in folgende Sequenz von edit Operationen aufgebrochen: f|s + u|u + c|c + h|h. Der Trainingsprozess erfasst nun wie oft edit Operationen angewendet wurden und leitet daraus eine Verteilung ab.

Resultate und Diskussion

Als Testmenge haben wir 26 Dokumente aus dem Korpus extrahiert. Diese wurden eigens korrigiert um einen Gold Standard zu erhalten. Als Evaluationsmetriken wählten wir *Precision* (Anteil der validen Korrekturen), *Recall* (Abdeckung der einzelnen Fehler) und daraus den *F1-Score* (harmonisches Mittel aus Pre. und Rec.). Um die Ergebnisse unserer Arbeit zu vergleichen, haben wir zwei weitere Methoden auf die Testdaten angewendet. Dies waren zum einen die **Pattern s** und zum anderen nutzten wir als Referenzkorrektur für das Noisy Channel Model eine Implementierung von **Peter Norvig** (Norvig, 2009). Die Ergebnisse sind in Abbildung 1 aufgetragen.

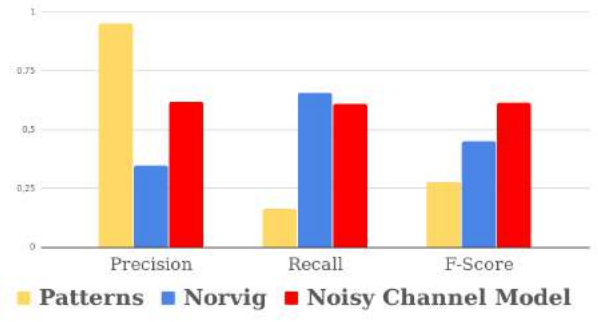


Abbildung 1: Resultate einzelner Korrekturmethoden angewendet auf den Testdatensatz

Man kann erkennen, dass die Pattern korrektur (gelb) die beste Precision erziel t . Dies ist ein typisches Verhalten regelbasierte r Systeme. Im Gegensatz dazu decken die beiden anderen Verfahren eine größere Menge an Fehlern ab, dies wird am höheren Recall deutlich. Besonders Norvigs Variante (blau) ist hier führend, jedoch tendiert diese auch zur Überkorrektur von richtig erfassten Wörtern. Wir waren bestrebt, dass unser Spell Checker (rot) dies weitestgehend vermeidet, indem es Precision und Recall möglichst balanciert. Es werden also viele OCR Rechtschreibfehler korrigiert und gleichzeitig wird die Rate an Falsch Positiven gering gehalten . Hierbei w ar das Optimieren der Gewicht u ng λ des Language Models ein essentieller Bestandteil der Arbeit, sodass unser Modell schlussendlich einen F-Score von 0.61 2 erzielte. Bei der Überlegung unseren Ansatz auf andere historische, unaufbereitete Texte anzuwenden empfiehlt es sich das Fehlerverhalten in diesen Texten bestmöglich zu generalisieren . Deshalb sollte bereits eine Wissensbasis in Form von Ersetzungspatterns vorliegen um das Error Model korpusspezifisch zu trainieren, das heißt genauso wie in diesem Beitrag beschrieben.

Zusammenfassung

Im Vergleich zur derzeitigen pattern-basierten Methode verbesserte der *Noisy Channel Spell Checker* die Korrekturqualität um mehr als das Doppelte. Es werden nun Fehler berichtigt, die die Patterns nicht einmal als solche erkennen. Die Hauptmotivation zum Aufbau des Royal Society Corpus sind Untersuchungen der diachronischen Entwicklung von wissenschaftlichem Englisch (UdS Fedora Commons o.J.). Die Bereinigung der Texte macht es möglich, dass diese Analysen in Zukunft weitaus genauer und verlässlicher werden.

Bibliographie

Jurafsky Daniel / Martin James H. (2016): *"Spelling Correction and the Noisy Channel"* In: Speech and Language Processing, 3. Edition, S. 61-73.

Kermes, Hannah / Degaetano-Ortlieb, Stefania / Khamis, Ashraf / Knappen, Jörg / Teich, Elke (2016): *"The Royal Society Corpus: From Uncharted Data to Corpus"*, in: Proceedings of the Tenth International Conference on

Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA).

Knappen, Jörg / Fischer, Stefan / Kermes, Hannah / Teich, Elke / Fankhauser, Peter (2017): *"The Making of the Royal Society Corpus"*, in ListLang@NoDaLiDa.

Norvig, Peter (2008): *"Natural Language Corpus Data: Beautiful Data"*. [online] http://norvig.com/ngrams/ [letzter Zugriff 08. November 2017].

Shannon, Claude E. (1948): *"A Mathematical Theory of Communication"*, in Bell System Technical Journal.

UdS Fedora Commons Repository (o.J.): *"The Royal Society Corpus (RSC)"*, https://fedora.clarin-d.uni-saarland.de/rsc/ . [letzter Zugriff 29. März 2018].

Paleocoran: Virtuelle Rekonstruktion von Korankodizes mit IIIF

Pohl, Oliver
opohl@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Marx, Michael
marx@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Franke, Stefanie
stfranke@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Artika, Farah
artika@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Schnöpf, Markus
schnoepf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Mahmutovic, Edin
mahmutovic@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Das Projekt „Paleocoran“ untersucht Koran-Handschriftenfragmente aus dem siebten bis zehnten Jahrhundert aus der Amr ibn al-ʿĀs- Moschee in al-Fusṭāt (Alt-Kairo), die heute in verschiedenen Sammlungen weltweit verteilt aufbewahrt werden. Paleocoran sammelt kodikologische (buchgeschichtliche) und paläographische (schriftgeschichtliche) Daten zu den Handschriftenfragmenten, um die zu rekonstruieren, welche Fragmente ursprünglich einen Kodex gebildet haben. Die digitale Rekonstruktion der Kodizes – einem Puzzle mit ca. 25.000 Teilen zu vergleichen – wird durch IIIF digital ermöglicht.

Das von F. Déroche (Paris) und M. Marx (Potsdam) geleitete DFG-ANR-Projekt „Paleocoran“ greift inhaltlich und methodisch an das Akademievorhaben „Corpus Coranicum“ der BBAW an: Datensätze aus den Corpus-Coranicum-Datenbanken zu Handschriften und Koran-Textvarianten werden in einem System zur wortgenauen Textstellenverortung von Koranpassagen verwendet. Die in „Paleocoran“ und „Corpus Coranicum“ generierten philologischen Daten werden in einem web-basierten System auf Grundlage des PHP-Frameworks Laravel in einer MySQL-Datenbank aufgezeichnet.

Für die Rekonstruktion der Korankodizes aus Alt-Kairo werden die Seiten der einzelnen Fragmente in der Datenbank erfasst und mit Textstellenkoordinaten (Sure-Vers-Wort) ausgezeichnet. Derzeit befinden sich ca. 1100 Koranhandschriften mit insgesamt über 25.000 Manuskriptseiten in der aus mehr als 40 Sammlungen Corpus Coranicum-Datenbank. Die hier vorgehaltenen Bilder werden über den IIIF-kompatiblen Bildserver digilib ausgeliefert.

Anhand der Textstellenangaben wird der Text auf der Manuskriptseite mit dem Text der Koran-Druckausgabe Kairo 1924 nach Schreibvarianten, Textvarianten und Verszählung untersucht. In vielen Fällen enthalten die verschiedenfarbigen Tinten der Handschrift Vokalzeichen, die Textvarianten in die Handschriften eintragen. Auch diese werden als unterschiedliche Lesarten verzeichnet.

Außerhalb der philologischen Daten stellen Illuminationen und Ornamente eine wichtiges Kennzeichen für Manuskriptfragmente dar, die ursprünglich zusammengehörten. Die Form und farbliche Gestaltung der funktionalen Ornamente (Verstrenner und Kapitelüberschriften). Form und Layout (z.B. Pflanzenornamente oder geometrische Muster) der Ornamente oder deren farbliche Gestaltung lassen dabei auf eine gemeinsame Herkunft der Fragmente schließen.

Für den Vergleich und die Darstellung von Orthographiedifferenzen im Vergleich zum Text der Druckausgabe Kairo 1924 wurde im Rahmen des Paleocoran-Projekts die Programmierbibliothek „Rasmify“ entwickelt. Rasmify entfernt sämtliche buchstabendifferenzierenden Zeichen (Diakritika) und Vokalzeichen aus arabischen Zeichenketten, sodass nur das Konsonantenskelett (arabisch: rasm) der verarbeiteten Strings übrig bleibt. Dies ist wichtig, da die frühen Koranhandschriften sehr häufig Buchstaben undifferenziert schreiben. Durch die reduzierte Wortform wird es einfacher, Unterschiede zwischen einzelnen Koranhandschriftenfragmenten und der Kairiner Druckausgabe zu identifizieren. Das Programm „Rasmify“ wurde in PHP, Python 3 und JavaScript als freie Software veröffentlicht und kann einfach über die jeweiligen Abhängigkeitsverwaltungen composer, pip und npm nachgenutzt werden.

Ähnliche Muster bei Abweichungen in Textvarianten (Lesarten), Schreibvarianten- und Verssegmentierung weisen darauf hin, dass die betreffenden Fragmente ursprünglich demselben Korankodex stammen. Bei genügend vorliegenden Indizien werden die einzelnen Handschriftenfragmente bzw. Teile der Fragmente einem virtuellen Kodex zugeordnet.

Durch die den einzelnen Handschriftenseiten zugeordneten Textstellen werden dann die dem virtuellen Kodex zugeordneten Handschriftenseiten nach Textkoordinate sortiert, sodass letztendlich ein IIIF-Manifest für den virtuellen Kodex erstellt werden kann.

Auf der Projektwebseite paleocoran.eu kann mittels des IIIF-Manifests der virtuell rekonstruierte Korankodex in seiner ursprünglichen Form im IIIF-Viewer Mirador digital abgebildet werden. Mirador bietet darüber hinaus Lichttischfunktionalitäten, sodass sowohl einzelne Seiten desselben virtuell rekonstruierten Korankodex als auch unterschiedliche virtuell rekonstruierte Korankodizes miteinander gezielt verglichen werden können.

Zusätzlich werden Metadaten der zugeordneten Manuskriptfragmente sowie Metadaten zur kodikologischen und paleographischen Einordnung des rekonstruierten Kodex angezeigt. Weiterhin werden die Lesarten- und Orthographievarianten sowie Ornamente samt Wortkoordinate und zitierfähigem IIIF-Bildausschnitt angegeben, sodass Forschende die virtuelle Rekonstruktion nachvollziehen können.

Zum aktuellen Zeitpunkt wurden 338 virtuell rekonstruierte Kodizes bzw. Kodexteile angelegt und über 1500 Lesartenvarianten sowie über 2500 Orthographieunterschiede identifiziert. Durch die Anbindung an das Corpus Coranicum Projekt ist die langfristige Sicherung und Nachnutzung gewährleistet. Der Launch der Paleocoran-Projektwebseite soll Ende 2018 erfolgen.

Pfälzische Burgen und ihre Umgebung im Mittelalter, modelliert anhand von Neo4j, QGIS und 3D Modellen

Pattee, Aaron

aaron.pattee@iwr.uni-heidelberg.de
Ruprecht-Karls-Universität, Heidelberg

Kuczera, Andreas

andreas.kuczera@adwmainz.de
Akademie der Wissenschaften und der Literatur, Mainz

Volkmann, Armin

armin.volkmann@asia-europe.uni-heidelberg.de
Stiftung Preußischer Kulturbesitz, Berlin

Abstract:

Das interdisziplinäre Projekt an der Schnittstelle der Baugeschichte und angewandter Informatik untersucht sechs Burgen der mittelalterlichen Pfalz um den Zusammenhang zwischen Sozialhierarchie, Architektur und Landschaft zu forschen. Das Ziel ist neue Kenntnisse zur chronologischen Baugeschichte zu gewinnen anhand der Verbindung heterogenen Daten in einer Graphendatenbank. Drei Komponente bezogen auf verschiedenen Methoden der Informatik und Geoinformatik dienen dazu Ergebnisse in einer *Neo4j*-basierten Datenbank zu verbinden: 1. Räumliche Landschaftanalysen in Geographischen Informations-Systemen (GIS), basieren auf der systematischen Auswertung digitaler Geländemodelle (DGM) mit 32 weiteren geo-referenzierten historischen Landkarten (1540-1799 n.Chr.), und werden durchgeführt, um neue Kenntnisse zur strategischen und politischen Funktionen der Bauten zu erhalten; 2. Die Fallstudien werden fotogrammetrisch sowie mit einem terrestrischen Laserscanner (TLS) hochgenau vermessen und zur detaillierten Landschaftsanalyse in den GIS verknüpft; 3. Eine Netzwerkanalyse von 310 überlieferten Urkunden (882-1589 n.Chr.) der Inhaber der Burgen um das Projekt in den historischen Kontext der Pfalz einzuordnen und Verbindungsstrukturen aufzuzeigen.

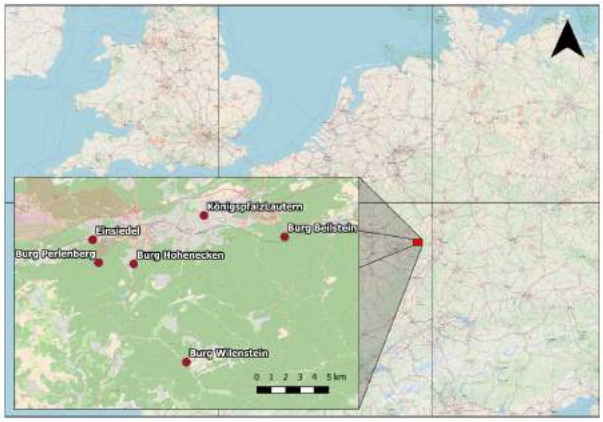


Abbildung 1: Lage der Burgen als Fallstudien des Projektes CITADEL.

Einführung:

Das CITADEL Projekt erstellt einen integrativen Forschungsansatz in der drei Komponenten, bezogen auf drei Methoden, vereint werden können. Die Methoden bestehen aus der Anwendung einer Graphendatenbank in *neo4j*, 3D terrestrischen Laserscans und fotogrammetrischen Modellen der ausgewählten Burgen, die zusammen im GIS mit weiterführenden, georeferenzierten historischen Landkarten zur Landschaftsanalyse der Pfalz kontextualisiert ausgewertet werden. Der Zweck dieser Integration verschiedener Datensätze ist es, neuartige Kenntnisse zur Konstruktion mittelalterlicher Burgen im Zusammenhang mit der Landschaft und dem politischen Umfeld mit einhergehender Sozialhierarchie und wechselseitiger Beeinflussung zu gewinnen. Dahingehend wurden zur weiteren Methodenentwicklung die Burgen der mittelalterlichen Pfalz in der Umgebung von Kaiserslautern als Fallstudien erforscht (Abb. 1).

Die Fallstudien—Die Königspfalz Lautern (oft Kaiserspfalz genannt), die Deutschherrenkommende Einsiedel (heutiger Einsiedlerhof) und die Burgen Beilstein, Hohenecken, Perlenberg und Wilenstein—bilden die pfälzische Kulturlandschaft des Mittelalters am Nordrand des Pfälzerwaldes ab, und sie dienten aufgrund der relativ guten Quellenlage, schon oft als Untersuchungsobjekte für Archäologen und Historiker. Leider gibt es nur wenig

Information über die Burg Perlenberg und fast keine Information über das Netzwerk, d.h. die Interaktionen der Burgen miteinander sowie ihre Funktionen im Mittelalter. Die Fragen nach dem Status der Burgherren, der Netzwerkverbindungen und der herrschaftlich-organisatorischen Leistungsfähigkeit dieser Strukturen in Bezug auf die topographische Lage in der Landschaft sind weitgehend unbeantwortet (vgl. Liddiard, 2011). Diese Fragen können nur von dem verfolgten integrativen Forschungsansatz weiterführend beantwortet werden.

Abbildung 2: 3D-Modell der Königspfalz in Kaiserslautern in der Seitenansicht.

Datenquellen und Methoden:

Die heterogenen Datenquellen von CITADEL bestehen aus historischen Urkunden der Ministerialenfamilien der Burgen, den hochdetaillierten selbst aufgenommenen 3D-TLS und den fotogrammetrischen Modellen (Abb. 2), historischen Landkarten der Pfalz und digitalen Geländemodellen (DGM). Die überlieferten Urkunden der Ministerialenfamilien *von Beilstein*, *von Hohenecken*, *von Lautern/de Lutra* und *von Wilenstein* versorgen das Projekt mit zentralen, historischen Informationen in Bezug auf Bauphasen, sowie dem Status und Amt der Mitglieder der obengenannten Familien. Der gesellschaftliche Stand des sogenannten ‚Adels' oder ‚niederen Adels' prägte die Form der jeweiligen Burg, insbesondere zum Thema Macht und der Repräsentation der Macht (Untermann, 2007). Allerdings ist die Auswirkung dieser Prägung nicht explizit erforscht worden, und nur vier bis sechs Bauphasen sind in den Urkunden deutlich erwähnt, was nicht ausreicht, um alle Bauphasen der sechs Burgen über die Jahrhunderte feinchronologisch zu erfassen. Des Weiteren sind alle Burgen heute Ruinen, die aufgrund der fragmentarischen Erhaltung mehr oder weniger Ansatzpunkte bieten, die Bauphasen der Strukturen klar zu bestimmen. Burg Hohenecken stellt dabei eine noch heute recht gut erhaltene Ausnahme dar. Die Netzwerkanalyse verbindet dabei den politischen Einfluss zur jeweiligen Zeitebene in Bezug auf die erforschten Bauphasen der Burgen. Damit können architektonische Elemente der Bauphasen mit dem entsprechenden gesellschaftlichen Status der Personen verbunden werden.

Die Burgen wurden durch TLS (Terrestrische Laserscans) mit einer Präzision von drei Millimetern im Rahmen des Projektes aufgenommen. Dieselben Burgen wurden zusätzlich fotogrammetrisch erfasst. Durch die Verbindung von TLS und Fotogrammmetrie werden hochgenau vermessene und gleichzeitig fotorealistische Modelle für alle sechs Burgen erzeugt (Pattee, 2016), die als herrschaftliche Schnittstellen, Knotenpunkte der räumlichen Landschaftsanalyse im GIS darstellen. Gleichzeitig werden bauhistorische und strukturelle Untersuchungen ausgeführt, bzw. zu Bauphasenanalysen, Bauquerschnitten und Mauerneigungen, die mit den vorherigen Bauuntersuchungen verglichen werden, um neue Kenntnisse zu generieren.

Die historischen Landkarten und Urkunden dienen dazu, das Projekt in den geographisch-geschichtlichen Kontext der Pfalz einzuordnen. Die bisher genutzten Karten konnten von der *David Rumsey Map Collection* (Stanford University Library) kostenfrei heruntergeladen und im GIS georeferenziert werden. Damit können Straßen und Wege in der Umgebung der Burgen in der frühen Neuzeit durch auf den historischen Karten ausgewiesene Kreuzungen lokalisiert werden. Leider ist der Datensatz der historischen Landkarten jedoch unvollständig erhalten geblieben, um genauer festzulegen wie die Straßen und Wege die einzelnen Burgen miteinander vernetzten. Daher werden *Least-Cost-Path*-Berechnungen im GIS erzeugt, wobei plausible hypothetische Wegverbindungen aufgrund der Oberflächengestalt des Reliefs und weiteren „physischen Kostenfaktoren" kalkuliert werden.

Ergebnisse:

Die Graphendatenbankanalyse in *neo4j* hat deutliche Tendenzen betreffend des Zusammenhangs zwischen Ministerialenfamilien und benachbarten Klöstern aufgedeckt. Leitende Mitglieder der Familien, die sich am Anfang ihres Amtes als Ministerialen der stauferischen Könige und Kaiser mit der Kirche eng verbunden haben, konnten ihre Familien vor einem politischen Untergang nach Amtsrücktritt erfolgreicher schützen, als Familien die sich nicht frühzeitig mit der Kirche verbanden. Die Macht der Ministerialen hat sich in der zweiten Hälfte des 13. Jahrhunderts stark reduziert, und viele dazugehörige Familien haben geliehene Ländereien rasch verkauft, um sich von der königlichen Herrschaft, der sie vertraglich verpflichtet waren, zu trennen (Spieß, 1975). Andere Familien fingen aber schon am Ende des 12. Jahrhunderts an, geliehene Ländereien an Klöster zu verschenken, um ein starkes Verhältnis mit der Kirche zu entwickeln und gleichzeitig eine fortschreitende Trennung von der kaiserlichen Gewalt zu vollziehen—dies ist bisher weitgehend unerforscht (Bihrer, 2011).

In *neo4j* konnten wir feststellen, wann und an wen die vier Ministerialenfamilien der obengenannten Burgen Ländereien verschenkt oder verkauft haben, und zwar im zeitlichen Verlauf. Im Zusammenhang mit der Entwicklung der Burgen wird es deutlich, dass die Familien die schon früh anfingen mit der Kirche zu handeln, ihre Burgen des Weiteren prächtiger ausgebaut haben. Die kirchliche Verbindung in der Form von familienzugehörigen Geistlichen, Pfründen, Klerikern und sogar einem Bischof hielten die betreffenden Familien auf hohem Stand, auch nach der politischen Veränderung der Macht der Ministerialen am Ende des 13. Jahrhunderts. Die Bauphasen der Burgen unterstützten diese These, da nur die Burgen der kirchlich stark verbundenen Familien nach 1300 regelmäßig noch ausgebaut wurden. Nur die Burg Hohenecken wurde in den folgenden Jahrhunderten sogar als Schloss etabliert. Die erkennbaren Bauphasen der Burg Hohenecken zeichnen gleichzeitig die Phasen der politischen Macht der Inhaber ab und zwar hauptsächlich nachdem keiner der Familie noch als Ministeriale diente. D.h. die Familie konnte sich immer noch auf hohem Kostenniveau ihren Stammsitz ausbauen, da sie noch über beträchtliche Finanzmittel verfügte. Welche architektonischen Elemente genau zu diesen Zeiten gebaut wurden, wird noch bauhistorisch untersucht, anhand der präzisen Laserscans und Foto realistischen, fotogrammetrischen Modelle.

Die Landschaft der Burgen und die dazu angeliehenen Ländereien, nahegelegenen Straßen und Seen prägten die Entwicklung einer Burg ebenso, wie die politischen Einflüsse. Zu diesen Landschaftsuntersuchungen gehörten *Least-Cost-Path*-Analysen, um energetisch effiziente Wegführungen zwischen den Ausgangs- und Endpunkten der Burgen zu berechnen, vgl. Herzog 2017. Darüber hinaus wurden im GIS Sichtfeldanalysen durchgeführt, um festzustellen, was oder wie weit man von einer bestimmten Höhe, z.B. vom Bergfried einer Burg, sehen konnte. Die 3D-Laserscanmodelle der Burgen dienen im GIS als Ausgangslagen der Knotenpunkte der *Least-Cost-Path*-Analysen, und sie liefern gleichzeitig genau vermessene Höhen der Burgen für die Sichtfeldanalysen (Abb. 3). Interessanterweise konnten die Burgmannen der sechs Burgen sich gegenseitig nicht sehen, zumindest in der heutigen bewaldeten Lage. Dies spricht dafür, dass die Bergfriede der Höhenburgen einem anderen Zweck dienten, nämlich um Ländereien im eigenen Besitz und nahegelegene Straßen zu beobachten, statt benachbarte Burgen zu überwachen.

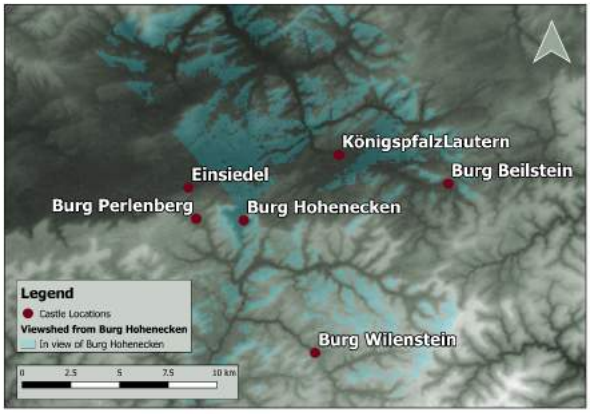


Abbildung 3: 12 km-Radius im Rahmen der Sichtfeldanalyse vom 29-Meter hohen Bergfried der Burg Hohenecken; hellblau: sichtbar; Rest: unsichtbar.

Aussicht

Neue Kenntnisse werden auf Mikro- und Makroebene der Studie entstehen, in denen die Fragen vom Status der Burgherren, der Netzwerkverbindungen und der herrschaftlich-infrastrukturellen Leistungsfähigkeit bezüglich der topographischen Lage in der Landschaft hinterfragt und neu beantwortet werden können. In einem folgenden Arbeitsschritt werden Straßen- und Wegverläufe der historischen Landkarten mit den berechneten Verbindungen aus den *Least-Cost-Path*-Analysen und mit Entwässerungsanalysen der Bäche und Flüsse für potentielle Wegverläufe, bzw. Netzwerke verglichen. Dabei soll herausgefunden werden, welche Straßen und Wege im Mittelalter überhaupt existieren konnten, wenn man bedenkt, dass durch den bekannten mittelalterlichen Wasserbau und die einhergehende Fischteichanlage vielerorts Täler unpassierbar waren. Folgende Analysen der Wassereinzugsgebiete der Täler des Waldgebietes werden der Berechnung von Entwässerungsmustern und Abflussregimen in den Hanglagen dienen und damit diese Fragestellung detailliert beantworten. Die Vereinigung dieser Methoden ermöglicht den Gewinn neuer Kenntnisse, um zu einem holistischen Bild der Funktion der mittelalterlichen Burgen der Pfalz zu gelangen.

Bibliographie

Bihrer, Andreas (2011): *"Research on the Ecclesiastical Princes in the Later Middle Ages—State-of-the-Art and Perspectives"* in: **Huthwelker, Thorsten / Peltzer, Jörg / Wenhöner, Maximilian (eds.)**: *Princely Rank in Medieval Europe—Trodden Paths and Promising Avenues*. Ostfildern: Thorbecke 49-70.

Herzog, Irmela (2017): *"Reconstructing Pre-Industrial Long Distance Roads in a Hilly Region in Germany, Based on Historical and Archaeological Data"* in Studies in Digital Heritage, 1(2), 642-660.

Liddiard, Robert (2011): *"English Castle Building"* in: **Huthwelker, Thorsten / Peltzer, Jörg / Wenhöner, Maximilian (eds.)**: *Princely Rank in Medieval Europe—Trodden Paths and Promising Avenues*. Ostfildern: Thorbecke 199-225.

Pattee, Aaron (2016): *„Integrative Recording Methods of Historic Architecture—Burg Hohenecken from Southwest Germany"* Master's Thesis, University of Nebraska-Lincoln.

Spieß, Karl-Heinz (1975): *"Vom reichsministerilen Inwärtseigen zur eigenständigen Herrschaft—Untersuchungen zur Besitzgeschichte der Herrschaft Hohenecken vom 13. bis zum 17. Jahrhundert"* in: JbGKL 12/13: 84-106.

Untermann, Matthias (2007): *„Abbild, Symbol, Repräsentation—Funktionen mittelalterlicher Architektur?"* **in Wagener, O., Dinzelbacher, P., (Eds.)**: *Symbole der Macht? Aspekte mittelalterlicher und frühneuzeitlicher Architektur: Beihefte zur Mediaevistik—Monographen*, Editionen, Sammelbände.

Rooting through Direction – New and Old Approaches

Hoenen, Armin
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Deutschland

Introduction

Computational stemmatology is the discipline dealing with the reconstruction of the copy history of historical works, primarily transmitted by handwriting. The result is a directed acyclic graph or tree. Trees generated by computational means are often unrooted. While in biology, all lifeforms are connected in a huge tree of life (compare tolweb.org), for texts such a tree seems an unsuitable metaphor. For this reason, rooting a tree in biology is relatively easy while it is relatively hard in stemmatology. Biologists in the majority of cases use a so-called outgroup, that is a species remotely - but not too remotely - related to the group under investigation. Then they identify root that node which is closest to the

outgroup since the outgroup is considered to indicate where the group connects with the tree of life. There are some alternative rooting procedures none of which is applicable to stemmatology generally. Haigh (1970, 1971) proposes to assign root according to probabilities assuming a certain type of birth-process as a generating function for stemmata but immediately relativises his proposal saying that the process used is not historically realistic. Yet other approaches are pursued by Marmerola et al. (2016). We implement the approach of Haigh (1970), one of the approaches of Marmerola et al. (2016) (called minimum-cost heuristic) and present our own two approaches which we then test on 3 artificial stemmatological datasets.

Approach

Correctly classifying all directions of the edges in a tree would provide us with root. Such a classification would be one method to obtain a root, but would have a caveat. If only a few edges would be classified wrongly, the resulting contradiction could be resolvable in more than one equally probable way. Instead of looking at single edges or the complete tree, we focus on paths, namely all paths from one leaf to another leaf. The situation we start from is thus an unrooted tree (UT) with exactly one covertly underlying rooted tree (RT), the gold standard. What we can distinguish in the UT are only leafs and internodes. For any leaf, we suppose that there is at least one fellow leaf which is so remote that their latest common ancestor is root. Furthermore we do know that any shortest path from leaf to leaf in the UT must pass through the latest common ancestor of both leafs. We want to denote it by # l_i, l_j but we initially don't know which of the nodes on the path in the UT it is. # is pivotal for directionality in the RT: all directions of edges towards the two leafs point away from it. Thus, traversing the path by edges from one leaf to the other, direction changes in the RT when passing #. If, as we supposed, for each leaf there is at least one other leaf for which # is root, then, comparing all leaf-leaf paths starting in a leaf l_1, the # most distant from l_1 must coincide with root since there can be no later common ancestor in the RT. Moreover, all leafs should converge on the same node as the most distant #. Thus, if one could detect that node on any leaf-leaf path which is pivotal for direction, one would know for each such path the direction change point # and in consequence root. Detecting the most probable point of direction change on a path may be easier and lead to less contradictions than classifying each edge for direction since the path itself provides context. However, there is a case we have so far not taken into account: the case that RT is a planted tree. In that case, the highest direction change convergence point (HDCCP) would be the first node in the RT that has more than one child. Also, there would be one leaf, on the path to which all other leaves do not detect any direction change, which our algorithm tracks. For a graphical explanation in case root is internal, see Figure 1.

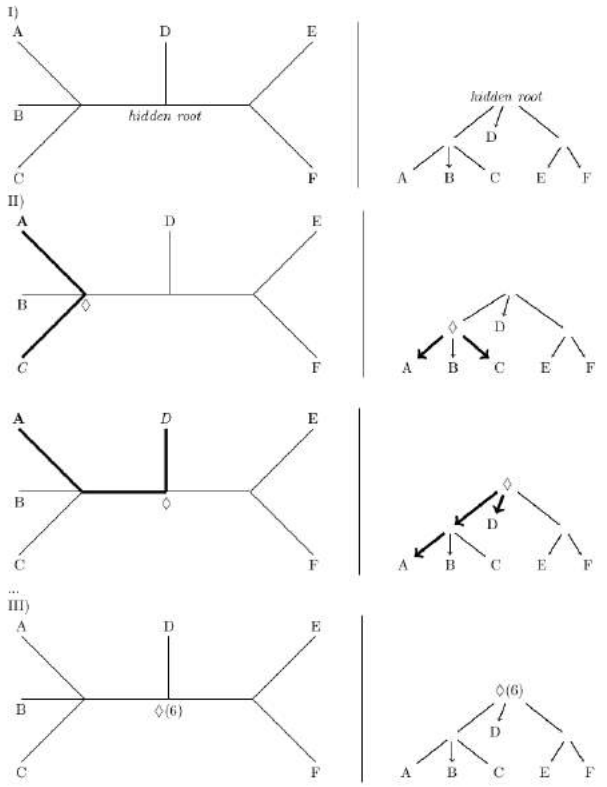


Figure 1. Detection of root from an unrooted tree (I, left) with an underlying rooted one (right). For each leaf on each path to another leaf, the node of direction change is detected and marked,here with # (II, **A**-C and **A**-D exemplarily). For each leaf separately, the # furthest away from it is chosen and if direction change detection works 100 %, all nodes converge on true root (III).

Direction

In practice, assigning direction is utterly complicated by the fact that only very few features allow an unambiguous detection of direction. In biology, Gogarten et al. (1989), Iwabe et al. (1989) have used gene duplications to root groups of bacteria at the root to the tree of life itself for which outgroups are not applicable. The logic behind this is that gene copies are relatively specific and that once present they are unlikely to be lost. The best parallel in textual criticism are probably certain types of skips such as line skips. The problem for a copyist having a model with a lineskip is that they cannot restitute the text unless knowing it by heart or having another exemplar. Now, lineskips appear to be no mass phenomenon and other indications for direction are rather rare and relatively unstable. For instance, one could try to classify variants for (linguistic or relative) age as done in the so-called CBGM method (Mink 2004) and then assign direction change to the node on a leaf-leaf path which has relatively most oldest variants.[1] But, the classification could be tricky and require masses of input data. Or one could take a heuristic explained by West (1973:32) who elaborates on groups of manuscripts sharing variants but this criterion may be too weak in our case and in addition lead to some overweighting of vulgate variants. We choose psycholinguistically obtained asymmetric letter confusion probabilities instead. Those are largely constrained

to one-to-one aligned letter pairs and in a nutshell give a probability of whether the letter sequence <tap> is more likely to be copied from or to <top>. Hoenen (2018) used these to determine distances (MMD) to compute stemmata and found that they are only interpretable with roughly a third of the cases of variation in the three artificial datasets Parzival (Spencer et al. 2004), Notre Besoin (Baret et al. 2004) and Heinrichi (Roos & Heikkilä 2009). Furthermore due to orthographic depth (Katz & Frost 1992) Finish (Heinrichi) was more processable than English (Parzival) or French (Notre Besoin). We detect that node as most probable DC point on a leaf-leaf path which has a minimal sum of MMD on the way to each leaf. Marmerola et al. (2016) use a heuristic where they take each node in UT as tentative root, sum the weights on all paths to all leafs and finally, comparing all possible roots take the overall minimum cost root. Weights come from one of their similarity functions, but in our implementation, we simply take the Hamming distance (Hamming 1950) and the MMD. Additionally, we implement a similar approach using the MMD distance only. We take each UT node as tentative root, then sum weights not on all paths root-leaf *in* but simply on all edges *of* the actual tree, since our distance matrix is asymmetric and will thus lead to a different sum for each tentative root. We choose the minimum cost root (LCM mincost). Finally, we implement the above outlined approach of the HDCCP root where we use again the letter confusion matrices, this time to determine the most likely DC nodes (#) on the leaf-leaf paths (HDCCP LCM). We used the matrix of Geyer (1977)[2] for lowercase andPaap et al. (1982) for uppercase.

Results

The results can be seen in Table 1.

Tradition	True root	Haigh MML	Marmerola et al. (2016), Hamming/MMD	LCM mincost, MMD	HDCCP LCM, MMD
Parzival	LM	LM	LM	JW(2)	SD(2)
Notre Besoin	n10	n9 (2)	n9(2)	n9(2);n11(1)	n9(2); planted
Heinrichi	h1	h1	h1	h16(4)	h1

Table 1. Results for different approaches to rooting. In brackets length of path to true root.

Discussion

While in 1970/1971 the artificial traditions were not yet present, Marmerola et al. (2016) likewise do not report their results on the complete artificial datasets but on their estimated stemmata. Both approaches yielded very good results and identified root twice. LCM mincost performed clearly worse, surprisingly also worse than Marmerola et al. (2016) with MMD, probably since the magnitude of differences when summing the paths is larger. Moreover, HDCCP LCM on the same distance (MMD) performs better than LCM mincost and identified root in one case. For the second tradition for which the Gold Standard is a planted tree, all methods had problems. HDCCP LCM detected that root could be a leaf and gave as candidates for that case n8, n10 (true root), n4 and n6. Hoenen (2018) had reported that for Finish due to orthographic depth, the psycholinguistic matrices would perform best which is consistent with our results. The artificial traditions are hardly representative of historical traditions in size or depth and thus allow only a rough approximation of the implications of these preliminary results. Also, MMD is only one way to approach the detection of directionality. The artificial traditions have been produced by recent scribes, their time-depth is not comparable to historical traditions and their sizes and source languages are not representative of historical data, which is why a more in depth investigation and a closer look onto the results is a priority for future elaboration of the method.

Conclusion

We have demonstrated the first implementation of a new rooting algorithm which is applicable not only for stemmatological but also biological (molecular and character data, substitution matrices) and historical linguistic data (lexico-statistics, sound shifts). The approach yielded encouraging results.

Fußnoten

1. Here, one could also use variants classified as older through the principles of lectio difficilior or lectio brevior. Some first Machine Learning based classifications suggest that this principle, at least for the artificial datasets mentioned below is a too low frequency phenomenon for directionality classification.
2. Boles & Clifford (1989) data yielded the same results on HDCCP LCM.

Bibliographie

Baret, Philippe / Macé, Caroline / Robinson, Peter (2004): *"Testing methods on an artificially created textual tradition"* in: **Baret, Philippe / Macé, Caroline / Bozzi, Andrea / Cignoni, Laura**: *The evolution of texts: confronting stemmatological and genetical methods* Linguistica Computationale XXIV-XXV. Pisa/Rom: Instituti Editoriali e Poligrafici Internationali, 255–281.

Boles, David B. / Clifford, John E. (1989): *"An upper- and lower case alphabetic similarity matrix, with derived generation similarity values"*. Behav. Res. Methods Instrum. Comput. 21, 597–586.

Geyer, L. H. (1977): *"Recognition and confusion of the lowercase alphabet"*. Percept. Psychophys. 22, 487–490.

Gogarten, Johann P. / Kibak, Henrik / Dittrich, Peter / Taiz, Lincoln / Bowman, Emma J. / Bowman, Barry J. / Manolson, Morris F. / Poole, Ronald J. / Date, Takayasu / Oshima, Tairo (1989): *"Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes."* Proc. Natl. Acad. Sci. 86, 6661–6665.

Haigh, John (1970): *"The recovery of the root of a tree"* J. Appl. Probab. 7, 79–88.

Haigh, John (1971): *"Mathematics in the Archaeological and Historical Sciences"* Edinburgh University Press, Scotland, UK, 396–400.

Hamming, Richard W. (1950): *"Error detecting and error correcting codes."* Bell System technical journal, 29(2), 147-160.

Hoenen, Armin (2018): *"Multi Modal Distance - An Approach to Stemma Generation with Weighting"* in: Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018. Miyazaki (Japan).

Iwabe, Naoyuki / Kuma, Kei-ichi / Hasegawa, Masami / Osawa, Syozo / Miyata, Takashi (1989): *"Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes"* Proc. Natl. Acad. Sci. 86, 9355–9359.

Katz, Leonard / Frost, Ram (1992): *"The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis"* Haskins Lab. Status Rep. Speech Res. SR-111, 147–160.

Marmerola, Guilherme D. / Oikawa, Marina A. / Dias, Zanoni / Goldenstein, Siome / Rocha, Anderson (2016): *"On the Reconstruction of Text Phylogeny Trees: Evaluation and Analysis of Textual Relationships"* PloS One 11, e0167822.

Mink, Gerd (2004): *"Problems of a highly contaminated tradition: the New Testament: Stemmata of variants as a source of a genealogy for witnesses"* in: **van Reenen, P., den Hollander, A., van Mulken, M. (Eds.)**: *Studies in Stemmatology II. John Benjamins*, pp. 13–86.

Paap, Kenneth R. / Newsome, Sandra L. / McDonald, James E. / Schvaneveldt, Roger W. (1982): *"An activation-verification model for letter and word recognition: the word-superiority effect"* Psychol. Rev. 89, 573–594.

Roos, Teemu / Heikkilä, Tuomas (2009): *"Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets"* Lit. Linguist. Comput. 24, 417–433.

Spencer, Matthew / Davidson, Elizabeth A. / Barbrook, Adrian C. / Howe, Christopher J. (2004): *"Phylogenetics of artificial manuscripts"* J. Theor. Biol. 227, 503–511.

West, Martin L. (1973): *"Textual Criticism and Editorial Technique: Applicable to Greek and Latin texts"* Teubner, Stuttgart.

Sachthematische Zugänge im Archivportal-D am Beispiel Weimarer Republik

Meyer, Nils
nils.meyer@la-bw.de
Landesarchiv Baden-Württemberg, Deutschland

Ein neuer Rechercheweg

Das Archivportal-D (www.archivportal-d.de) ist ein Subportal der Deutschen Digitalen Bibliothek (DDB, https://www.deutsche-digitale-bibliothek.de/) , das im Rahmen eines DFG-Projekts speziell für die Recherche nach Archiven und Archivgut entwickelt wurde. Derzeit sind im Archivportal zwei Nutzungsszenarien möglich: Nutzende können entweder nach einzelnen Archiven suchen und davon ausgehend durch die Bestände des jeweils ausgewählten Archivs browsen oder sie können mit der Freitextsuche nach einzelnen Objekten suchen. In beiden Fällen stehen verschiedene Facetten zur Einschränkung der Ergebnisse bereit.

In dem aktuellen, ebenfalls von der DFG geförderten Projekt sollen diese Nutzungsmöglichkeiten um einen dritten Zugang erweitert werden. Am Beispiel zweier Digitalisierungsprojekte des Landesarchivs Baden-Württemberg und des Bundesarchivs zum Thema „Weimarer Republik“ soll ein neuer, themenbasierter Zugang zu den Objekten im Archivportal entwickelt werden. Im Zentrum stehen dabei die Anreicherung der Erschließungsdaten des Archivguts mit sachthematischen Begriffen und die Nutzbarmachung der angereicherten Informationen für ein intuitives, forschungsgerechtes Angebot. Das Projekt läuft über zwei Jahre seit dem 1. Juni 2018. Projektpartner sind neben Bundesarchiv und Landesarchiv Baden-Württemberg die Deutsche Nationalbibliothek als Betreiberin der DDB sowie FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur.

Entwicklung von Systematik und Vokabularen

In einem ersten Schritt soll gemeinsam mit Vertreterinnen und Vertretern der Archivcommunity, der Geschichtswissenschaft und der Digital Humanities eine an den Bedürfnissen der Forschung ausgerichtete sachthematische Systematik zur Weimarer Republik erstellt werden, die sowohl als Sucheinstieg als auch als hierarchische Facette bei der Filterung der Suchergebnisse dienen kann. Den einzelnen Objekten werden dabei kontrollierte Indexbegriffe aus dem Themenbereich „Weimarer Republik“ zugewiesen, welche wiederum mit den Klassen der Systematik verknüpft sind. Für den Aufbau einer nutzerorientierten Systematik werden verschiedene wissenschaftliche und nicht-wissenschaftliche Systematiken als Grundlage genommen. Das kontrollierte Vokabular der Indexbegriffe wird aus einer umfangreichen Wortgutsammlung ausgehend von den vorhandenen Erschließungsdaten sowie externen Quellen (z.B. Normdateien, Wikidata) gespeist. Daneben soll ein kontrolliertes Vokabular zu den einzelnen Staaten der Weimarer Republik eine grundlegende geografische Suchmöglichkeit schaffen. Indexbegriffe und Systematik sollen dabei mit Wikidata-Datenobjekten und Normdateien verknüpft werden. Gleichzeitig soll auch die normdatenbasierte Indexierung bekannter Personen bei den einzelnen Archivobjekten verbessert werden.

Tools und algorithmische Zuordnung

Ein Schwerpunkt des Projekts und zentrale Forschungsaufgabe neben der Entwicklung von Systematik und Vokabularen ist die Entwicklung von Tools zur Zuordnung der einzelnen Archivobjekte zu den verschiedenen Vokabularen. Ziel der Arbeit soll ein komfortables Zuordnungswerkzeug zur einfachen Verknüpfung von ganzen Beständen sowie auch einzelnen Archivalien mit den verschiedenen sachthematischen, geografischen und

Personen-Indexbegriffen sein. Dieses Tool soll zunächst anhand der Weimar-Bestände des Bundesarchivs und des Landesarchivs Baden-Württemberg entwickelt und getestet werden. In einem späteren Schritt soll das Tool auch dem Personal in anderen Archiven sowie den Nutzenden zur Verfügung gestellt werden. Auf diese Weise können die Nutzenden kollaborativ an der Anreicherung der Erschließungsdaten und der Erweiterung des Themenzugangs mitwirken. Zugleich wird die bisher nur wenig genutzte Möglichkeit der Sachindexierung in den Archiven gestärkt.

Dem Zuordnungstool zur Seite gestellt werden soll ein Werkzeug zur algorithmischen Zuordnung von Archivobjekten zu den verschiedenen Indexbegriffen. Über Maschinenlernprozesse sollen die einzelnen Objekte beziehungsweise die Objektmetadaten und ihr jeweiliger Kontext analysiert werden. Dabei werden verschiedene externe Linked-Data-Ressourcen abgefragt und in die Analyse mit einbezogen. Daraus werden im Ergebnis Vorschläge für die Zuordnung der einzelnen Archivobjekte zu den verschiedenen Vokabularen generiert. Die Vorschläge dienen der Unterstützung des Fachpersonals und der Nutzer bei der Auswahl der passenden Indexbegriffe und können von diesen mit dem Zuordnungstool umgesetzt werden. Der hohe Anteil unstrukturierter Daten aufgrund der Erschließung in verschiedenen Freitextfeldern in den Archiven stellt die Arbeit in diesem Arbeitspaket vor besondere Herausforderungen. Unterschiedliche Ansätze (Improved Fisher Vector, Convolutional Neural Networks) sollen daher auf ihre Eignung für archivische Erschließungsdaten erprobt werden.

Grundlagenentwicklung für Archivportal und DDB

Um die sachthematische Systematik adäquat abbilden zu können, sollen die bisherigen Facetten des Archivportals und der DDB um eine hierarchische Dimension erweitert werden. Daneben werden die neuen sachthematischen, geografischen und personenbezogenen Metadaten der einzelnen Archivobjekte in einem eigenen, neu zu entwickelnden Repositorium unabhängig von den Ursprungsdaten des jeweiligen Archivs gesichert und sollen über eine Exportfunktion auch an das datengebende Archiv zurückgespielt werden können. Hierfür sollen umfangreichere Entwicklungsarbeiten im gemeinsamen Backend von DDB und Archivportal angestoßen werden.

Mit der Grundlagenarbeit aus dem Projekt sollen später mit geringem Aufwand auch Themenzugänge zu anderen Bereichen (Nationalsozialismus, Erster Weltkrieg, verschiedene Archivaliengattungen) möglich sein. Über das Frontend des Archivportals und der DDB sind diese unmittelbar nutzbar, über das API der DDB können die angereicherten Metadaten auch in anderen Kontexten weitergenutzt werden.

Bibliographie

Becker, Irmgard Christa / Maier, Gerald / Uhde, Karsten / Wolf, Christina (eds.) (2016): *Netz werken*. Das Archivportal-D und andere Portale als Chance für Archive und Nutzung. Beiträge zum 19. Archivwissenschaftlichen Kolloquium (= Veröffentlichungen der Archivschule Marburg 61), Marburg: Archivschule Marburg.

Fähle, Daniel / Maier, Gerald / Schröter-Karin, Tobias / Wolf, Christina (2015): *Archivportal-D*. Funktionalität, Entwicklungsperspektiven und Beteiligungsmöglichkeiten, in: Archivar 68: 10-19.

Hentschel, Christian / Wiradarma, Timur Pratama / Sack, Harald (2015): *If we did not have imagenet*. Comparison of fisher encodings and convolutional neural networks on limited training data, in: *Advances in Visual Computing* (= Lecture Notes in Computer Science 9475) 400-409.

Herrmann, Tobias / Zahnhausen, Vera (2018): *Auf dem Weg zum Digitalen Lesesaal*. Das Projekt "Weimar – Die erste deutsche Demokratie", in: *Kompetent!* Archive in der Wissensgesellschaft. 86. Deutscher Archivtag 2016 in Koblenz (= Tagungsdokumentationen zum Deutschen Archivtag 21), Fulda: Verband deutscher Archivarinnen und Archivare [Veröffentlichung Herbst 2018].

Hollmann, Michael (2016): *Deutschland in zwei Nachkriegszeiten*. Der Einstieg in das Online-Archiv des Bundesarchivs, in: Archivar 69, 6-9.

Laux, Susanne / Wolf, Christina (2016): *Forschungsprojekt "Von der Monarchie zur Republik" gestartet*. Digitalisierung von Quellen zur Demokratiegeschichte im deutschen Südwesten 1918-1923, in: Archivnachrichten 52, 28-29.

Lindenthal, Jutta (2016): *Datenqualität und Retrieval*. Vorschläge zur Verbesserung der Suche n der Deutschen Digitalen Bibliothek http://jl.balilabs.de/DDB/DQ/DDB_Datenqualit%C3%A4t_Retrieval_1.0.pdf [letzter Zugriff 24. September 2018].

Reisacher, Martin / Krauth, Wolfgang (2015): *Vernetzen als Herausforderung*. Die Deutsche Digitale Bibliothek, in: Archivnachrichten 51, 36 f.

Rühle, Stefanie / Schulze, Francesca / Büchner, Michael (2014): *Applying a Linked Data Compliant Model*. The Usage of the Europeana Data Model by the Deutsche Digitale Bibliothek http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/231/225 [letzter Zugriff 24. September 2018].

Sack, Harald (2014): *Linked Data Technologien*. Ein Überblick, in: **Pellegrini, Tassilo / Sack, Harald / Auer, Sören (eds.)**: *Linked Enterprise Data*. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien. Berlin: Springer 21-62.

Wolf, Christina (2016): *Eines für alle: Das Archivportal-D*. Neue Zugangswege zu Archivgut, in: Neue Wege ins Archiv. Nutzer, Nutzen, Nutzung. 84. Deutscher Archivtag 2014 in Magdeburg (= Tagungsdokumentationen zum Deutschen Archivtag 19). Fulda: Verband deutscher Archivarinnen und Archivare 47-63.

Semantisch angereicherte Präsentationsschichten für geisteswissenschaftliche Webanwendungen Methodenvergleich und Referenzimplementierung

Toschka, Patrick
patrick.toschka@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Hintergrund

Ausgangspunkt der folgenden Betrachtungen ist eine Master-Thesis, die sich mit einem Vergleich von Methoden zur Anreicherung von Präsentationsschichten geisteswissenschaftlicher Webanwendungen mit semantisch strukturierten Metadaten befasst. Viele aktuelle Forschungsapplikationen lagern Metadaten in eine eigene Datenschicht aus. Meist können diese Daten über spezifische Schnittstellen bezogen werden. In diesem Szenario bietet die Präsentationsschicht meist nur eine optische Darstellung dieser Daten ohne eine tiefergehende semantische Strukturierung. Alternativ können Metadaten aber auch direkt im HTML-Quelltext eingebettet werden. Eine externe, maschinelle Verarbeitung der Metadaten dort scheint allerdings noch aufwendig, da angeblich wenige bis keine Standards für die Auswertung von Daten direkt aus der Präsentationsschicht vorhanden sind.

Das Web hat sich weiterentwickelt. Große Konzerne wie Suchmaschinenbetreiber haben Methoden und Standards vorangetrieben, die im privatwirtschaftlichen Sektor neue Ansätze ermöglichen. Über Linked (Open) Data Technologien können Geräte selbstständig Inhalte fremder Quellen anfragen, interpretieren und aufbereiten. Hierfür sind Technologien erforderlich, die semantische Metadaten auf einer Vielzahl von Plattformen standardisiert zur Verfügung stellen und Inhalte miteinander verknüpfen.

Potentiale von JSON-LD und Schema.org

JSON-LD hat sich durch die Verwendung der JSON-Syntax, der Übersichtlichkeit des Quellcodes und der Auslagerung aus dem eigentlichen Dokument heraus als de facto Standard herausgestellt. Bereits bei der Evaluation des Schema.org-Vokabulars wird klar, dass dieses gut geeignet ist, grundsätzliche semantische Aussagen abzubilden. Beispielsweise ist die Auszeichnung von Beziehungen zwischen Personen wie *x kennt y, x ist dieselbe Person wie y* oder *x ist verwandt mit y* out-of-the box mit dem Schema.org-Vokabular möglich. Bei komplexeren semantischen Relationen wie *x hat schon von y gehört, x empfiehlt y* oder *x spottet über y* fehlen in Schema.org zuweilen noch bestimmte Aussagemuster. Dieser Situation kann aber durch die Einbindung zusätzlicher LOD-Vokabulare gut begegnet werden.

Die modulare Ausbaufähigkeit von JSON-LD birgt somit für eine semantische Anreicherung der Präsentationsschichten von Digital Humanities Anwendungen viele Möglichkeiten. Im Bereich Digitaler Editionen beispielsweise fügt sich eine Einbeziehung der Metadaten in die Präsentationsschicht gut in das Konzept der Edition als Interface ein.[1] Auch für die Interoperabilität webbasierter geisteswissenschaftlicher Ressourcen in Einklang mit den FAIR-Prinzipien spielt eine semantische Strukturierung der Präsentationsschicht eine wichtige, momentan aber noch unterschätzte Rolle.[2] So könnten Browser-Plugins zukünftig automatisch Vorschläge für verwandte Einträge auf anderen Webressourcen anzeigen, ohne dass dabei ein Mehraufwand bei der Datenpflege betrieben oder gar eine spezifische Schnittstelle implementiert werden muss.

Methodik

Es gilt also herauszufinden, inwiefern das Schema.org Vokabular für den geisteswissenschaftlichen Anwendungsfall zum einen qualitativ exakt genug ist und ob, zum anderen, alle nötigen Aussagen mit diesem Vokabular und eventuellen Erweiterungen getroffen werden können. Mittels Referenzimplementierungen auf Basis bestehender Webapplikationen, die bisher keine der genannten Technologien in der Präsentationsschicht nutzen, können die Potentiale genauer eingeschätzt werden. Dazu müssen Sets projektspezifischer Metadaten zunächst aufgearbeitet und dann unter Verwendung verschiedener Einbettungsmethoden in HTML verankert werden. Ziel ist es, möglichst viele verschiedene Anwendungsfälle auf diese Art und Weise vergleichend zu erfassen und dabei die möglichen Probleme mit den Technologien und dem Schema.org Vokabular herauszuarbeiten. Daraus leitet sich ab, ob sich einzelne Einbettungsverfahren gut oder weniger gut für geisteswissenschaftliche Webanwendungen eignen, welche Probleme auftreten und worauf bei zukünftigen Implementierungen geachtet werden sollte.

Fußnoten

1. Siehe zum Begriff der Edition als Interface auch https://www.i-d-e.de/publikationen/schriften/bd-12-interfaces/ (letzter Zugriff: 12. Januar 2019).
2. Siehe zu den FAIR-Prinzipen (Findable, Accessible, Interoperable, und Re-usable) auch http://www.forschungsdaten.org/index.php/FAIR_data_principles (letzter Zugriff: 12. Januar 2019).

Bibliographie

Carroll, Jeremy J. / Bizer, Christian / Hayes, Patrick / Stickler, Patrick (2004): *"Named Graphs, Provenance and Trust"*, pdf Format, URL: http://wifo5-03.informatik.uni-mannheim.de/bizer/SWTSGuide/carroll-ISWC2004.pdf (letzter Zugriff: 08. Januar 2019).

Google Inc. (2018): *"Understand how structured data works"*, URL: https://developers.google.com/search/docs/guides/intro-structured-data , text/html Format, (letzter Zugriff: 04. Januar 2019).

Guha, R.V. / Brickley, Dan / MacBeth, Steve (2015): *"Schema.org: Evolution of Structured Data on the Web"* in: Queue - Structured Data, Vol. 13 No. 9, text/html Format, URL: https://queue.acm.org/detail.cfm?id=2857276 (letzter Zugriff: 08. Januar 2019).

Halpin, Harry / Herman, Ivan / Hayes, Patrick J. (2009): *"When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web"*, pdf Format, URL: https://www.w3.org/2009/12/rdf-ws/papers/ws21 (letzter Zugriff: 08. Januar 2019).

Lahntaler, Markus (2012): *"Third Generation Web APIs. Bridging the Gap Between REST and Linked Data"*, pdf Format, URL: http://www.markus-lanthaler.com/research/third-generation-web-apis-bridging-the-gap-between-rest-and-linked-data.pdf (letzter Zugriff: 04. Januar 2019).

W3C World Wide Web Consortium (2014): *"JSON-LD 1.0. A JSON-based Serialization for Linked Data"*, text/html Format, URL: http://www.w3.org/TR/2014/REC-json-ld-20140116/ (letzter Zugriff: 04. Januar 2019).

W3C World Wide Web Consortium (2015): *"RDFa Core 1.1 – Third Edition. Syntax and processing rules for embedding RDF through attributes"*, text/html Format, URL: http://www.w3.org/TR/2015/REC-rdfa-core-20150317/ (letzter Zugriff: 04. Januar 2019).

Wettlaufer, Jörg (2018): *"Der nächste Schritt? Semantic Web und digitale Editionen"* in: Digitale Metamorphose: Digital Humanities und Editionswissenschaft, Hg. von Roland S. Kamzelak / Timo Steyer. 2018 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2), text/html Format, URL: http://dx.doi.org/10.17175/sb002_007 (letzter Zugriff: 08. Januar 2019).

Semantische Minimal-Retrodigitalisierung von Brief-Editionen

Rettinghaus, Klaus
klaus.rettinghaus@gmail.com
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland

Situation

Brief-Editionen sind in vielerlei Hinsicht ein unerlässliches Hilfsmittel für die (historische) Forschung. Gerade weil Briefe ein höchst subjektives Ausdrucksmedium sind, enthalten sie oftmals wertvolle Informationen, die für das Verständnis größerer Zusammenhänge essentiell sein können. Doch um ihr volles Potential entfalten zu können, müssten sie leicht zugänglich sein. Dabei liegen die allermeisten Editionen ausschließlich in gedruckter Form vor, nur neuere Vorhaben publizieren wenigstens teilweise digital.

Der Forscherin oder dem Forscher entgehen im Zweifelsfall wichtige Schriftzeugnisse, weil sie sie schlichtweg nicht finden. Oder um es mit Goethe zu sagen: „Man erblickt nur, was man schon weiß und versteht." Es stellt sich also die Frage, wie man diese zwischen Buchdeckel gepressten Schätze heben und der Forschung zugänglich machen kann.

OCR-gestützte Retrodigitalisierung von Volltexten kritischer Editionen ist hier qualitativ zumeist noch nicht ausreichend. Erschwerend kommt hinzu, dass diese nicht-semantisch ist – und bei urheberrechtlich geschützten Werken in aller Regel völlig unmöglich.

Lösungsansatz

Als möglicher Lösungsansatz wäre eine sogenannte semantische Minimal-Retrodigitalisierung vorzuschlagen, um gedruckte Editionen zu erschließen und online auffindbar zu machen. Hierbei werden ausschließlich die Metadaten der Briefe digitalisiert, das heißt die Namen des Schreibers und des Empfängers sowie das Datum des Briefes. Weitere Informationen, beispielsweise zu den Orten, sind wünschenswert, jedoch nicht notwendig. Versieht man diese Metadaten mit Normdaten und stellt sie im TEI-Austauschformat CMIF (*Correspondence Metadata Interchange-Format*) bereit, könnten sie auch im Kontext von vielen tausend weiteren Briefen im Webservice *correspSearch.net* sichtbar werden, der von der Berlin-Brandenburgischen Akademie der Wissenschaften bereitgestellt wird.

Der große Vorteil hierbei ist, dass auch jüngste Publikationen erschlossen werden können, da auf den notwendigen Metadaten kein Urheberrechtsschutz liegt. Nicht selten enthalten solche Editionen sogar ein tabellarische Korrespondenzregister, das relativ einfach gescannt und aufbereitet werden kann. Generell scheint es allerdings nicht ratsam die Brief-Metadaten direkt in XML zu kodieren, da ein solches Vorgehen relativ fehleranfällig und zeitaufwändig ist. Aufgrund des hohen Grades an Strukturiertheit ist es sinnvoller sie gleichsam in Tabellen zu erfassen; dies erhöht die Übersichtlichkeit und reduziert die Fehlerquote. Wünschenswert ist natürlich eine Anreicherung mit Normdaten. Diese kann manuell oder auch halbautomatisch mit entsprechenden Tools erfolgen. Die kompilierten Daten können dann mit dem Programm *CSV2CMI* automatisch in das entsprechende TEI-XML umgewandelt werden. Die resultierende Datei muss dann nur noch im World Wide Web bereitgestellt werden.

Ein weiterer, charmanter Vorteil hierbei ist, dass in der Tabelle beliebig viele zusätzliche Informationen untergebracht werden können, die im CMIF nicht oder noch nicht kodiert werden können. So besteht folglich sogar die

Möglichkeit zwei Dateien bereitzustellen, eine für den für den automatisierten und standardisierten Austausch (*machine-readable*) und eine für Endbenutzer leicht lesbare Übersicht (*human-readable*).

Ausblick

Gleichwohl das hier vorgestellte Verfahren vergleichsweise unkompliziert ist, macht sich die Arbeit dennoch natürlich nicht von allein. Für wen erscheint es also sinnvoll diese Daten aufzubereiten und bereitzustellen?

Zunächst kommen natürlich Bibliotheken in den Sinn, die – als Informations-Anbieter des 21. Jahrhunderts – ein höchst eigenes Interesse daran haben (sollten), Wissen zu und aus Ihren Beständen anzubieten. Des Weiteren kommen Forscher in den Sinn, die sowieso intensiv mit einem Korrespondenzkorpus arbeitet. Das Eigeninteresse an verbesserter Auffindbarkeit und Zugänglichkeit dort sollte ausreichen, die notwendigen Daten zu digitalisieren und vorrätig zu halten. Hinzu kommt in letzterem Falle natürlich die Möglichkeit, die so aufbereiteten Daten als Forschungsdaten zu publizieren.

Zusammenfassung

Das Poster soll die hier skizzierte Idee der Semantischen Minimal-Retrodigitalisierung von Brief-Editionen vorstellen und den vorgeschlagenen Workflow sowie die bereit stehenden Software-Tools präsentieren.

Bibliographie

Woesler, Winfried (1988): *Vorschläge für eine Normierung von Briefeditionen*, in: Editio 2, S. 8–18. doi: 10.1515/9783110241938.8

Ball, Rafael (2014): *Bibliotheken im 21. Jahrhundert. Vom Leser zum Kunden*, in: **Ceynowa, Klaus / Hermann, Martin:** *Bibliotheken: Innovation aus Tradition*, Berlin: De Gruyter Saur 226–231, doi: 10.1515/9783110310511.226

Stadler, Peter (2014): *Interoperabilität von digitalen Briefeditionen*, in: **Wolzogen, Hanna Delf von / Falk, Rainer:** *Fontanes Briefe ediert*, Würzburg: Königshausen & Neumann 278–287.

Dumont, Stefan (2016): *corrospSearch – Connecting Scholarly Editions of Letters*, in: Journal of the Text Encoding Initiative 10, https://journals.openedition.org/jtei/1511 [letzter Zugriff 10. Oktober 2018].

Rettinghaus, Klaus (2018): *saw-leipzig/csv2cmi* (Version v1.6.2), Zenodo, doi: 10.5281/zenodo.1461642 [letzter Zugriff 13. Oktober 2018].

text2ddc meets Literature - Ein Verfahren für die Analyse und Visualisierung thematischer Makrostrukturen

Mehler, Alexander
mehler@em.uni-frankfurt.de
Goethe University of Frankfurt, Deutschland

Uslu, Tolga
uslu@em.uni-frankfurt.de
Goethe University of Frankfurt, Deutschland

Gleim, Rüdiger
gleim@em.uni-frankfurt.de
Goethe University of Frankfurt, Deutschland

Baumartz, Daniel
baumartz@stud.uni-frankfurt.de
Goethe University of Frankfurt, Deutschland

In diesem Poster geht es um die thematische Analyse und Visualisierung literarischer Werke mithilfe automatisierter Klassifikationsalgorithmen. Hierfür wird ein bereits entwickelter Algorithmus namens text2ddc (Uslu et. al. 2018) verwendet, um die Themenverteilungen literarischer Werke zu identifizieren. Darüber hinaus thematisiert der Beitrag, wie diese Verteilungen von Themen und deren Abhängigkeiten untereinander visualisiert werden können.

Bei text2ddc handelt es sich um einen Klassifikator auf Basis neuronaler Netze, der Texte einer bestimmten Anzahl von Sprachen nach der Dewey-Dezimalklassifikation (DDC) kategorisiert. Die DDC ist ein internationaler Standard für die Themenklassifikation im Bereich von (digitalen) Bibliotheken. Um text2ddc zu trainieren, wurde die Wikipedia verwendet. Da viele Artikel der Wikipedia mit der Gemeinsamen Normdatei (GND) verlinkt sind und die GND Informationen zu den entsprechenden DDC-Kategorien hinterlegt, war es möglich, ein vergleichsweise großes und zugleich breites DDC-orientiertes Trainingskorpora für das Deutsche aufzubauen. Am Beispiel dieses Korpus erreicht unser Algorithmus einen F-Score von 87,4%. Da die Artikel der Wikipedia auch über Sprachgrenzen hinweg untereinander verlinkt sind, war es zudem möglich, text2ddc für über 40 Sprachen zu trainieren.

text2ddc wurde auf Korpora verschiedener Genres angewandt, um deren Themenverteilungen zu analysieren. Zum einen betrifft dies die Wikipedia selbst, aber auch Korpora basierend auf StadtWikis, anhand derer bestimmt wurde, welche Themen dominant sind und wie diese zusammenhängen. Ein drittes Beispiel betrifft literarische Texte bzw. historische Texte der Wissenschaft. Abbildung 1 zeigt etwa die Themenverteilung von *Die Geburt der Tragödie aus dem Geiste der Musik* von Friedrich Nietzsche.

In dieser Abbildung repräsentieren die Knoten die DDC-Kategorien, wobei die Knotenfarbe dazu dient, die jeweilige DDC-Hauptkategorie zu identifizieren. Kanten zwischen den Knoten repräsentieren den Zusammenhang der jeweiligen Themen. Hierfür wurde die semantische Ähnlichkeit von Sektionen, Paragraphen und Sätzen ausgewertet. Ein alternatives Beispiel bildet *Massenpsychologie und Ich-Analyse* von Sigmund Freud (siehe Abbildung 2). Die Beispiele verdeutlichen nicht nur die erwarteten Unterschiede beider Werke, sondern zeigen zugleich makrostrukturelle thematische Zusammenhänge auf.

Der zugrundeliegende Algorithmus basiert auf folgendem Prozedere: Zunächst wird der Inputtext in Sektionen untergliedert, wofür die jeweilige logische Dokumentstruktur ausgewertet wird. Anschließend werden verschiedene NLP-Methoden angewendet, um Informationen über Lemmata und *Named Entities* zu gewinnen, was wiederum auf einer automatischen Disambiguierung basiert. Mittels dieser Methoden erreichen wir eine höhere Genauigkeit bei der Klassifikation mit text2ddc. Im nächsten Schritt werden die Sektionen unter Verwendung der DDC als Zielklassifikation kategorisiert. Je mehr Sektionen auf dasselbe DDC-Thema abgebildet werden, desto höher ist das Gewicht des entsprechenden Zielknotens und desto größer kann dieser dargestellt werden. Da bei dieser Vorgehensweise nicht auf Linkstrukturen zurückgegriffen werden kann, erfolgt die Induktion von Themenkanten nach der inhaltlichen Ähnlichkeit der den Themen zugeordneten Sektionen. Hierfür werden Texteinbettungsalgorithmen aus dem Bereich neuronaler Netze angewandt.

Das Poster untersucht anhand der Werke einer Reihe von deutschsprachigen Autoren (u.a. Karl Marx, Sigmund Freud, Franz Kafka, Friedrich Nietzsche, Thomas Mann und Martin Heidegger) die Möglichkeiten und Grenzen von Themenkarten zur Erfassung makrostruktureller Themenzusammenhänge von Texten, wie sie unser Algorithmus erfasst. Auf diese Weise soll eine Alternative zu den in den DH omnipräsenten *topic models* aufgezeigt werden. Zu diesem Zweck experimentiert der Beitrag mit alternativen Visualisierungstechniken basierend auf interaktiven konzentrischen Netzwerken (PolyViz) und alternativ basierend auf klassischen Netzwerkdarstellungen.

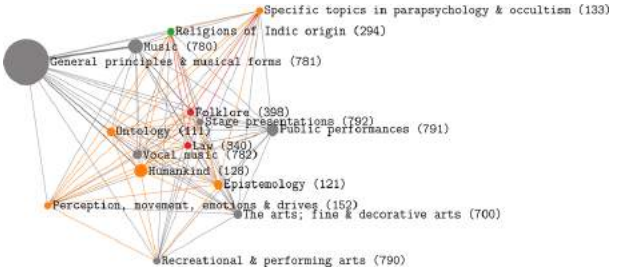


Abbildung 1: Friedrich Nietzsche (Die Geburt der Tragödie aus dem Geiste der Musik):

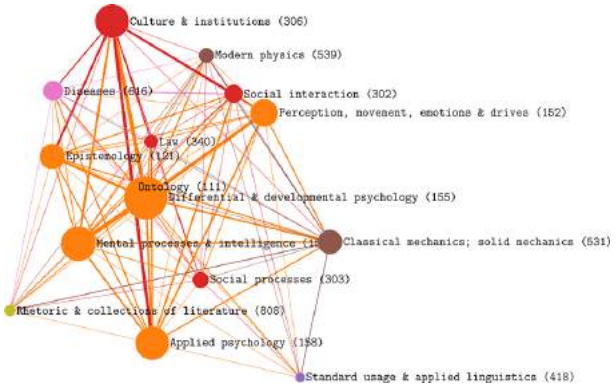


Abbildung 2: Siegmund Freud (Massenpsychologie und Ich-Analyse).

Bibliographie

T. Uslu, A. Mehler, A. Niekler, and W. Hemati (2018): *"Towards a DDC-based Topic Network Model of Wikipedia"*, in Proceedings of 2nd International Workshop on Modeling, Analysis, and Management of Social Networks and their Applications (SOCNET 2018), February 28, 2018, 2018.

T. Uslu and A. Mehler (2018): *"PolyViz: a Visualization System for a Special Kind of Multipartite Graphs"*, in Proceedings of the IEEE VIS 2018.

D. Baumartz, T. Uslu, and A. Mehler (2018): *"LTV: Labeled Topic Vector"*, in Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: System Demonstrations, August 20-26, Santa Fe, New Mexico, USA.

Umfrage zu Forschungsdaten in den Geistes- und Humanwissenschaften an der Universität zu Köln

Metzmacher, Katja

katja.metzmacher@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Helling, Patrick

patrick.helling@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Blumtritt, Jonathan

jonathan.blumtritt@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Mathiak, Brigitte
bmathiak@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Einleitung

Forschungsdaten spielen in den Geisteswissenschaften eine immer größere Rolle. Die für 2019 erwartete Ausschreibung zum Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI) durch das BMBF, die auf die Empfehlungen und Diskussionsimpulse des RfII zurückgehen (RfII 2016), zeigen den Stellenwert, den Forschungsdatenmanagement mittlerweile im wissenschaftspolitischen Diskurs erhalten hat. Für die Bildung von Konsortien in der NFDI wird explizit eine Ausrichtung der Strukturen an den Bedürfnissen der Community gefordert (RfII 2017). Eine Evaluation der eigenen Angebote sind für Forschungsinfrastrukturen, Hochschulen und Forschungsdatenzentren ein wichtiges Instrument, um die Planungen an die sich entwickelnden Bedürfnisse der Wissenschaftler*innen anzupassen.

Das Data Center for the Humanities (DCH) an der Philosophischen Fakultät der Universität zu Köln ist ein Forschungsdatenzentrum, das speziell die Fragen der geisteswissenschaftlichen Forschung im Blick hat und das sich durch eine große Nähe zum Forschungsalltag ausweist. Das Feedback aus dem Kreis der Wissenschaftler*innen der Fakultät trägt großes Gewicht. Das DCH hat eine Umfrage konzipiert, bei deren Gestaltung wir uns an verschiedenen FDM-Umfragen orientiert haben, insbesondere an unserer ersten Umfrage 2016 (Mathiak, Kronenwett 2017), welche unter anderem den Bedarf an Beratung, Speicherplatz und anderen FDM-Services fokussierte. Ergänzend haben wir 2018 die Kenntnisse und Gewohnheiten im Bereich FDM erfragt und die Umfrage auf die Mitglieder der Humanwissenschaftlichen Fakultät ausgeweitet, die vermehrt das Beratungsangebot des DCH in Anspruch nehmen.

Beschreibung der Umfrage

Die Umfrage an der Philosophischen und Humanwissenschaftlichen Fakultät der Universität wurde zwischen dem 06.06. und 08.07.2018 vom DCH in Kooperation mit den beiden Dekanaten sowie dem Kompetenzzentrum Forschungsdaten der Universität durchgeführt.[1] Der Online-Fragebogen bestand aus 36 geschlossenen Fragen, bei kategorialen Antworten gab es immer die Möglichkeit, ergänzende Kategorien zu benennen.

An der Humanwissenschaftlichen Fakultät nahmen insgesamt 115 Personen an der Umfrage teil, 67 haben alle Fragen beantwortet. Da teilweise nur die allerletzten Fragen unbeantwortet blieben, beinhaltet der Teildatensatz N=89 Fälle.

An der Philosophischen Fakultät nahmen insgesamt 215 Personen an der Umfrage teil, 128 beantworteten alle Fragen. Auch hier wurden einige nicht ganz vollständige Fälle in die Analyse miteinbezogen, der Teildatensatz beinhaltet N=179 Fälle. Diese können direkt mit der letzten Umfrage verglichen werden.

Insgesamt kann so auf eine Gesamtheit von 268 Datensätze zurückgegriffen werden. Alle Fächergruppen der beiden Fakultäten sowie alle Statusgruppen der Wissenschaftler*innen sind vertreten.

Methode

Bei der Umfrage handelt es sich um eine Online-Umfrage mit definierten Adressatenkreis, der aber nicht durch personalisierte Einladungen überprüft worden ist. Die Daten wurden bereinigt und statistisch in SPSS ausgewertet.

Ergebnisse

Die hier vorgestellten Ergebnisse beziehen sich auf die Erfahrung der Wissenschaftler*innen mit der Nutzung von Datenarchiven.

Von den befragten Personen haben deutlich weniger als die Hälfte Erfahrungen mit der Datenablage in Archiven. Lediglich 34,1% der Befragten haben bereits Daten in einem Datenarchiv abgelegt. Auch die Erfahrung mit sekundärer Datennutzung ist vergleichsweise gering. Dennoch ist das Interesse, insbesondere an unveröffentlichten Daten anderer Forscher*innen groß, sogar größer als die Bereitschaft selbst Daten zu veröffentlichen (siehe Abb. 1).

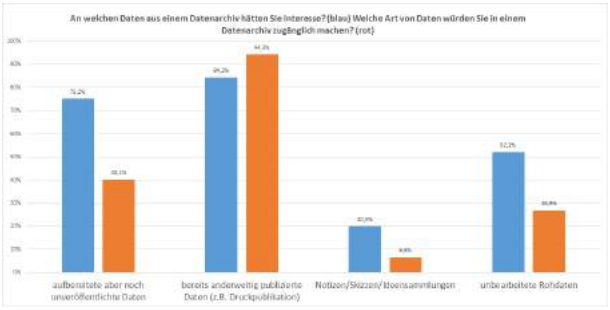


Abbildung 1. Gegenüberstellung von Interesse an Daten (N=165, Mehrfachantworten zulässig, blau) und Bereitschaft diese zu teilen (N=212, Mehrfachantworten zulässig, rot).

Grundsätzlich können sich aber nahezu alle Befragten vorstellen, zukünftig Forschungsdaten in einem Datenarchiv abzulegen. Mehrere Faktoren spielen dabei eine Rolle: so erfahren Empfehlungen von Datenarchiven durch Kolleg*inn*en eine hohe Relevanz und auch Möglichkeiten den Datenzugriff beschränken zu können bzw. Informationen darüber zu erhalten, wofür die eigenen Daten genutzt werden können, sind für die Befragten entscheidend. Gleichzeitig sind ein moderater Aufwand in der Datenaufbereitung und die Kostendeckung durch Forschungsförderung wichtige Faktoren (siehe Abb. 2).

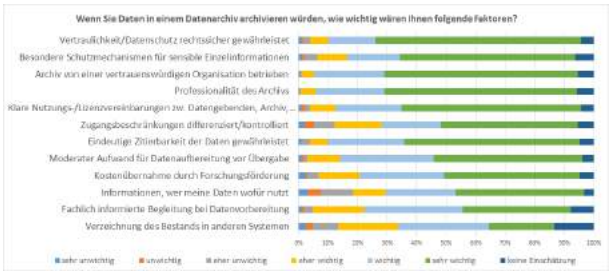


Abbildung 2. Entscheidungsfindung bei der Wahl eines Datenarchivs (N=268).

Einfluss auf die strategische Ausrichtung des DCH

In der vorherigen Umfrage 2016 war vor allem das Bedürfnis nach Beratungen deutlich geworden. In den darauffolgenden zwei Jahren haben wir stark daran gearbeitet, die Angebote auszuweiten und uns mit verschiedenen Organisationen zu vernetzen, um Forschungsdatenmanagementbedürfnisse in der Breite bedienen zu können. Beispielsweise bieten wir nun mit der Universitätsbibliothek Köln zusammen rechtliche Beratungen an.

Darüber hinaus haben wir als eines der zentralen Probleme identifiziert, dass Onlineressourcen, wie Webseiten und Onlinedatenbanken oft nicht besonders lange erhalten werden und schnell verschwinden. Um dieses Problem anzugehen, haben wir das Projekt SustainLife (Barzen et al. 2018) eingeworben, in dem es um den Erhalt von genau solchen Systemen geht. Ähnlich hilfreiche Anstöße erhoffen wir uns auch aus den Ergebnissen der Umfrage diesen Jahres.

Vorschau Poster

Im Rahmen der Postersession werden wir noch mehr zu den Ergebnissen auch der anderen Umfrageabschnitte präsentieren, diese mit den Ergebnissen von 2016 und anderen Erhebungen vergleichen, sowie eine Vorschau auf die sich daraus resultierenden strategischen Ziele des DCH geben.

Fußnoten

1. Dekanat der Philosophischen Fakultät der Universität zu Köln, https://phil-fak.uni-koeln.de/dekanat.html , Stand: 12.10.2018; Dekanat der Humanwissenschaftlichen Fakultät der Universität zu Köln, https://www.hf.uni-koeln.de/2008 , Stand: 12.10.2018; Kompetenzzentrum Forschungsdaten an der Universität zu Köln, https://fdm.uni-koeln.de/ , Stand: 12.10.2018.

Bibliographie

RfII - Rat für Informationsinfrastrukturen (2016): *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*, Göttingen.

RfII - Rat für Informationsinfrastrukturen (2017): *Schritt für Schritt - oder: Was bringt wer mit? Ein Diskussionsimpuls für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)*, Göttingen.

Mathiak, Brigitte / Kronenwett, Simone (2017): *A Survey on Research Data at the Faculty of Arts and Humanities of the University of Cologne*, in: Digital Humanities Conference 2017, Montreal 08.-11.08.2017, https://dh2017.adho.org/abstracts/164/164.pdf [Letzter Zugriff 12. Oktober 2018].

Barzen, Johanna / Blumtritt, Jonathan / Breitenbücher, Uwe / Kronenwett, Simone / Leymann, Frank / Mathiak, Brigitte / Neuefeind, Claes (2018): *SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften*, in: Book of Abstracts der 5. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2018), Köln 26.02.-02.03.2018: 471-474.

UPB-Annotate: Ein maßgeschneidertes Toolkit für historische Texte

Seemann, Nina
nina.seemann@upb.de
Universität Paderborn, Deutschland

Merten, Marie-Luis
mlmerten@mail.upb.de
Universität Paderborn, Deutschland

Einleitung

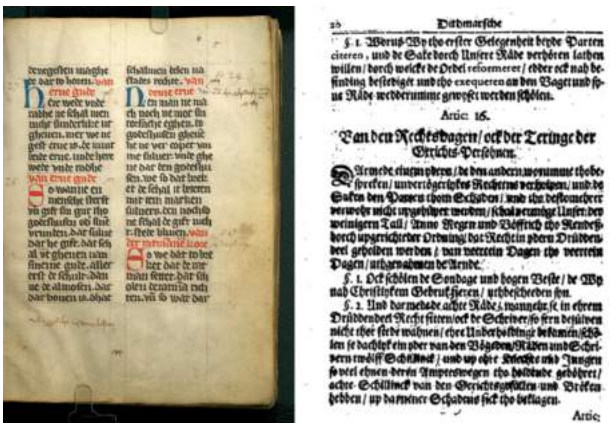
Das interdisziplinäre Projekt InterGramm geht dem literaten Ausbau des Mittelniederdeutschen bis hin zum Schriftsprachenwechsel zum Frühneuhochdeutschen nach. Im Zuge von Sprachausbauprozessen wandeln sich (häufig) bereits existierende Konstruktionen und neue literate Konstruktionen entstehen (Traugott / Trousdale 2013). Unser Ziel ist es, diese Wandelphänomene durch die Analyse historischer Rechtstexte und Arzneibücher zu erfassen. Insbesondere interessiert uns die Entwicklung komplexer Satztypen (subordinierende Konstruktionen), attributiver Techniken, die in komplexen Nominalphrasen resultieren, und textstrukturierender Elemente etc. Voraussetzung zur Erfassung dieser Elemente ist es, unser Korpus mit *Part-of-Speech*-Tags sowie mit Konstruktion-Tags (eine Art semantisch-syntaktische Annotation auf Phrasenebene) zu versehen. Naheliegenderweise steht für die linguistische Annotation von Daten bereits eine Vielzahl an Tools zur Verfügung, z.B. WebAnno (Yimam et al. 2014) oder CorA (Bollmann et al. 2014). Jedoch benötigen wir im Hinblick auf unsere graphematisch stark variierenden Texte und die herausfordernde Grammatikanalyse ein maßgeschneidertes Toolkit, das das Layout der historischen Texte 1:1 abbildet, das Editieren von Primärdaten erlaubt sowie morphologische

und konstruktionale Annotationen unterstützt. Das im Projekt erstellte Korpus wird der Forschergemeinde voraussichtlich durch CLARIN-D zur Verfügung gestellt.

Vom historischen Dokument zum Toolkit

1.) Korpus und Transkription

Unser Korpus besteht aus Rechtstexten und Arzneibüchern aus dem 13. bis 17. Jahrhundert, aufgeteilt in (I.) eine Zusammenstellung mittelniederdeutscher Texte von 1227 bis 1650 (1,2 Mio. Token) und (II.) eine Dokumentensammlung frühneuhochdeutscher Texte, die im ehemals mittelniederdeutschen Sprachraum verfasst wurden (400.000 Token). Nach Möglichkeit versuchen wir, nur Primärquellen zur Transkription zu benutzen. In den meisten Editionen wurden bereits – von uns unerwünschte – Normalisierungen bezüglich des Layouts vorgenommen. Unsere Transkriptionen sollen diplomatisch sein, d.h. textstrukturierende Elemente des historischen Dokuments, etwa Rubrizierungen, Leerzeilen etc., sollen 1:1 in das Transkript übernommen werden. Diese Elemente liefern wichtige Informationen im Hinblick auf strukturelle Änderungen und entscheidende Hinweise zum historischen Gebrauch dieser Texte. In Abbildung 1 zeigen wir zwei Auszüge aus unseren Primärquellen. In der linken, älteren Quelle wurden zur Strukturierung des Textes Rubrizierungen und Majuskeln genutzt. Rubrizierte Textstellen übernehmen dabei u. a. die Funktion einer Überschrift, während Majuskeln den Anfang eines Paragraphen kennzeichnen. Die rechte, jüngere Quelle hingegen nimmt Zentrierungen, Einrückungen und Absätze zur Hilfe.



Links: Stadtrecht von Lübeck (1294); rechts: Landrecht von Dithmarschen (1667).

2.) *XML*-Format

Jedes Transkript wird in unser *XML*-Format transformiert, das aus den drei Hauptkomponenten (i) *metadata* , (ii) *layoutinfo* und (iii) *token* besteht. Im *metadata* werden der Name des Dokuments, Entstehungsort, Entstehungsjahr und Texttyp (Rechtstext oder Arzneibuch) gespeichert. Diese Informationen sind für geplante temporale und lokale Visualisierungen des Sprachausbaus nötig. In *layoutinfo* werden all die Informationen bezüglich des Layouts gespeichert, die für eine 1:1-Abbildung im Tool benötigt werden. Jedes Wort des Textes wird als ein *token*-Element gespeichert, welches wiederum aufgeteilt wird in (i) diplomatische Ebene und (ii) moderne Ebene. Eine Illustration und Erläuterung folgen am Ende von 3.).

3.) upb::annotate

Das graphische Nutzer- *Interface* stellt die *token* der *XML*-Datei entsprechend der *layoutinfo* dar, d.h. dem Layout der Primärquelle entsprechend. Dies hat den Vorteil, dass die Annotatorinnen und Annotatoren sehr einfach erkennen können, was z.B. eine Überschrift ist oder ob am Zeilenende eine nicht-markierte Worttrennung stattfand. Weiterhin kommt dem Annotationsfluss zugute, dass die Daten in Leserichtung annotiert werden und der Kontext unmittelbar sichtbar ist. Das Toolkit ermöglicht, beide Annotationen (*POS*/ Konstruktionen) im gleichen Nutzer-Interface auszuführen, der Wechsel zur anderen Ebene erfolgt durch Klick auf den entsprechenden Trigger. Mit Blick auf die Nutzerpraxis ist es einfacher, zunächst Phrasen mit Konstruktion-Tags auszuzeichnen und dann den der Phrase zugehörigen Token *POS*-Tags entsprechend ihres Kontextes zuzuweisen. Wir zeigen einen Screenshot in Abbildung 2.

POS-Tag-Annotation über den Token; Konstruktion-Tag-Annotation unter den Token.

Eine wichtige Funktion des Tools ist das Editieren von Token. Aufgrund der historischen Schreibvariation ist es teilweise nötig, zwei Token zusammenzuziehen oder ein Token zu trennen, um konsistente *POS*-Tag-Annotationen vornehmen zu können. Für beide Operationen wird in der zugrundeliegenden *XML*-Datei jeweils ein eindeutiger Marker gesetzt, der die manuellen Editierungen nachvollziehen lässt. So sagt uns die diplomatische Ebene für „in|dhere“ (*t290* in Abbildung 3), dass es im Originaldokument als ein Wort geschrieben wurde. Jedoch besteht es aus zwei Wörtern, die jeweils ein eigenes *POS*-Tag auf moderner Ebene erhalten. Für „screi#mannen“ (*t302* in Abbildung 3) sagt uns die diplomatische Ebene, dass es als zwei Wörter im Originaldokument geschrieben wurde. Jedoch ist es ein Wort mit einem *POS*-Tag auf moderner Ebene.

```
<token id="t290" trans="in|dhere">
 <dipl id="t290_d1" trans="in|dhere" utf="indhere"/>
 <mod id="t290_m1" trans="in|" utf="in">
  <pos tag="APPR"/>
 </mod>
 <mod id="t290_m2" trans="dhere" utf="dhere">
  <pos tag="DDART"/>
 </mod>
</token>

<token id="t302" trans="screi#mannen">
 <dipl id="t302_d1" trans="screi#" utf="screi"/>
 <dipl id="t302_d2" trans="mannen" utf="mannen"/>
 <mod id="t302_m1" trans="screi#mannen" utf="screimannen">
  <pos tag="NA"/>
 </mod>
</token>
```

t290: Manuelle Trennung durch den Marker ‚|'. *t302:* Manuelle Zusammenfügung durch den Marker ‚#'.

Posterpräsentation

Auf dem Poster präsentieren wir einen Überblick über unser Projekt und beleuchten den obigen Prozess genauer. Zudem werden wir auf unsere Tagsets eingehen sowie auf die Generierung automatischer Vorschläge für POS- und Konstruktion-Tags durch *Machine Learning* und *Pattern Matching*.

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 (LaTeCH)* 86—90.

Traugott, Elizabeth / Trousdale, Graeme (2013): *Constructionalization and Constructional Change.* Oxford: Oxford University Press.

Yimam, Seid Muhie / Biemann, Chris / de Castilho, Richard / Gurevych, Iryna (2014): *"Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno"*, in: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 91—96.

VAnnotatoR: Ein Werkzeug zur Annotation multimodaler Netzwerke in dreidimensionalen virtuellen Umgebungen

Abrami, Giuseppe
abrami@em.uni-frankfurt.de
Goethe Universität Frankfurt, Deutschland

Spiekermann, Christian
s2717197@stud.uni-frankfurt.de
Goethe Universität Frankfurt, Deutschland

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

Abstract

Die Verarbeitung, Erzeugung und Visualisierung von multimodalen Netzwerken, in denen etwa symbolische und ikonographische Inhalte miteinander vernetzt werden, um sie multimodal rezipierbar zu machen, bilden neuartige Herausforderungen für die digitalen Geisteswissenschaften, da sie weit über textbasierte Medien hinausreichen. Digitalisate von Schrifttexten, Bildern und Illustrationen können in diesem Zusammenhang als Beispiele für knotenbildende Informationsobjekte ebenso genannt werden wie dreidimensionale Objektrepräsentationen. Die Anordnung dieser Objekte zu *semiotischen Netzwerken* hat den Zweck, deren Zusammenhangsstruktur zu visualisieren und rezipierbar zu machen. Der Begriff des Netzwerks erlaubt es dabei, Mikrostrukturen auf der Ebene von Knotenpaaren oder -triaden ebenso zu unterscheiden, wie Mesostrukturen von so genannten Netzwerkmotiven oder gar Makrostrukturen, welche die Gesamtorganisation des jeweiligen Informationsraums thematisieren (etwa im Sinne einer Zentrums-Peripherie-Struktur). Multimodale Netzwerke waren bereits Gegenstand einer Reihe von Projekten (siehe z.B. Li et al. 2007; Kane, Alavi 2008; Nasser, Gleich 2017). Allerdings befindet sich darunter kein Projekt, das eine generische Plattform zur Visualisierung und Annotation solcher Netzwerke im Rahmen virtueller Umgebungen zur Verfügung stellt. Um diese Lücke zu schließen, haben wir **VAnnotatoR** (Spiekermann, Abrami, Mehler 2018) entwickelt. **VAnnotatoR** ist ein Framework zur Erstellung und Visualisierung multimodaler Netzwerke in virtuellen Umgebungen. Das System erlaubt die Bearbeitung und Visualisierung dieser Netzwerke als *multimodale Hypertexte* (Mehler et al. 2018) in der *virtuellen Realität* (VR) und der *augmentierten Realität* (AR). **VAnnotatoR** wurde mithilfe von Unity3D[1] für VR implementiert und unterstützt gängige Desktop-VR-Brillen[2] (HTC Vive und Oculus Rift). Die AR-Implementierung von **VAnnotatoR** für Smartphones basiert wiederum auf AR-Core[3]. Mittels **VAnnotatoR** sind Inhalte visualisierbar, suchbar, relationierbar, attribuierbar und weit darüber hinaus annotierbar. Die derzeitige Implementierung dieses Systems erlaubt die Modellierung von Netzwerken auf der Ebene so genannter Hypergraphen (Berge 1989) und hierarchischer Graphen, in denen Knoten vollständige Graphen enthalten können. Abbildung 1 illustriert ein solches Netzwerk unter anderem durch Rekurs auf Texte, Bilder, Photographien, dreidimenionale Modelle und ein begehbares Haus. Um eine natürliche Interaktion mit den Komponenten solcher Graphen zu ermöglichen, beinhaltet **VAnnotatoR** eine Gesten- und Bewegungssteuerung mit dazugehörigen VR-Controllern. Da klassische VR-Eingabegeräte die Gestaltungsfreiheit

der Gestensteuerung limitieren, beinhaltet **VAnnotatoR** darüber hinaus eine Gestenerkennung unter Verwendung von Leap-Motion[4]. Gleichzeitig wurde eine Gestenerkennung basierend auf Datenhandschuhen[5] für den **VAnnotatoR** entwickelt (Kühn 2018). Da komplexe Annotationsprozesse i.d.R. von mehreren Annotatoren durchgeführt werden müssen, ermöglicht **VAnnotatoR** zudem die simultane und kollaborative Bearbeitung von multimodalen Netzwerken. Die Annotationen werden in UIMA-Strukturen, unter Verwendung des *UIMA-Database Inferface* (Abrami, Mehler 2018), abgespeichert und nutzbar gemacht.

Mit **VAnnotatoR** ist ein Annotationswerkzeug entstanden, welches die Bearbeitung und Visualisierung komplexer Netzwerke gestengesteuert ermöglicht. Der Beitrag demonstriert dies am Beispiel einer *Public History of the Holocaust*, in welcher, basierend auf dem Konzept der *Stolperwege* (Mehler et al. 2017), **VAnnotatoR** dazu genutzt wird, dreidimensionale Modelle nicht mehr existierender bzw. unzugänglicher Gebäude begehbar zu machen und zu dokumentieren. Auf diese Weise entsteht ein begehbarer und erweiterbarer Informationsraum, der insbesondere dazu dient, historische Prozesse zu dokumentieren. **VAnnotatoR** ist daher als ein Werkzeug zu betrachten, das verschiedene geisteswissenschaftliche Disziplinen verbindet. Inwiefern **VAnnotatoR** daher als interdisziplinäres Werkzeug der digitalen Geisteswissenschaften aufzufassen ist, soll Diskussionsgegenstand der Präsentation dieses Systems sein.

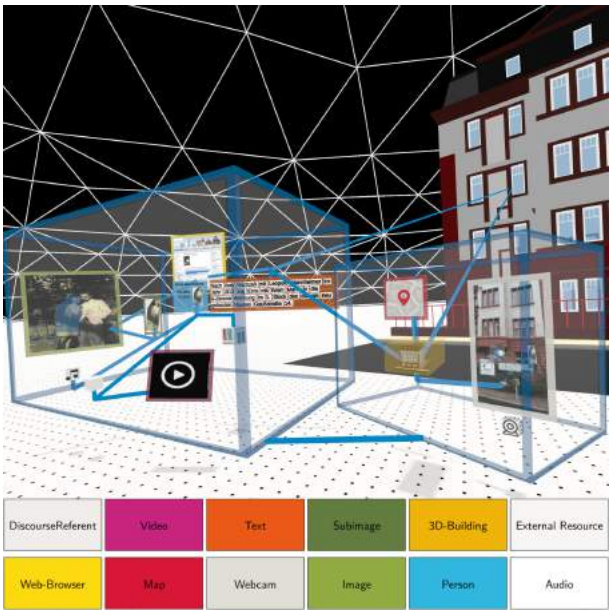


Abbildung 1: Ein Bild und ein Text wurden aus einem Browser extrahiert (linker Kubus). Beide wurden segmentiert und mit anderen Entitäten relationiert. Das Netzwerk zeigt in diesem Beispiel unter anderem Relationen zwischen einem 3D Modell eines Gebäudes (welches im extrahierten Text thematisiert wird), einer Postion (Karte), einer Videoübertragung aus der realen Welt, einer Personenrepräsentation, einem Audio-Dokument und einem Wikidata-Eintrag. Subnetzwerke können außerdem zu Knoten aggregiert werden (visualisiert als Kubus), welche untereinander wiederum relationiert werden können. Auf diese Weise entstehen virtuelle Netzwerke von annotierten Szenen oder Situationen, die von anderen Personen weiterverarbeitet oder rezipiert werden können. Die Legende zeigt die der jeweiligen Modalität zugewiesene Farbe.

Fußnoten

1. https://unity3d.com/de
2. Getestet mit HTC Vive (https://www.vive.com) und Oculus Rift (https://www.oculus.com/rift)
3. https://developers.google.com/ar/discover
4. https://www.leapmotion.com
5. Hi5 VR Glove (https://hi5vrglove.com)

Bibliographie

Abrami, Giuseppe / Mehler, Alexander (2018): *A UIMA Database Interface for Managing NLP-related Text Annotations*, in: Proceedings of the 11th edition of the Language Resources and Evaluation Conference. Miyazaki / Japan

Berge, Claus (1989): *Hypergraphs: Combinatorics of Finite Sets*. North Holland, Amsterdam.

Kane, Gerald C. / Alavi, Maryam (2008): *Casting the net: A multimodal network perspective on user-system interactions*, in: Information Systems Research, 19(3), pages 253–272.

Kühn, Vincent Roy (2018): *A gesture-based interface to VR*. Bachelor's thesis, Goethe-Universität Frankfurt.

Li, Zhi-Chun / Huang, Hai-Jun / Lam, William HK. / Wong, Sze Chun (2007): *A model for evaluation of transport policies in multimodal networks with road and parking capacity constraints* in: Journal of Mathematical Modelling and Algorithms 6(2), pages: 239–257.

Mehler, Alexander / Abrami, Giuseppe / Bruendel, Steffen / Felder, Lisa / Ostertag, Thomas / Spiekermann, Christian (2017): *Stolperwege: An App for a digital public history of the holocaust* in: Proceedings of the 28th ACM Conference on Hypertext and Social Media. Pages 319–320. New York, NY, USA. ACM.

Mehler, Alexander / Abrami, Giuseppe / Spiekermann, Christian / Jostock, Matthias (2018): *VAnnotatoR: A framework for generating multimodal hypertexts*, in: Proceedings of the 29th ACM Conference on Hypertext and Social Media. New York, NY, USA. ACM.

Nassar, Huda / Gleich, David F. (2017): *Multimodal network alignment* in: CoRR, abs/1703.10511.

Spiekermann, Christian / Abrami, Giuseppe / Mehler, Alexander (2018): *VAnnotatoR: a gesture-driven annotation framework for linguistic and multimodal annotation*, in: Proceedings of the Annotation, Recognition and Evaluation of Actions (AREA) Workshop. Japan.

Weltkulturerbe international digital: Erweiterung der Wittgenstein Advanced Search Tools durch Semantisierung und neuronale maschinelle Übersetzung

Röhrer, Ines
i.roehrer@campus.lmu.de
LMU München, Deutschland

Ullrich, Sabine
sabine.ullrich@campus.lmu.de
LMU München, Deutschland

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
LMU München, Deutschland

Einleitung

Mit der Aufnahme des Nachlasses von Ludwig Wittgenstein ins Internationale UNESCO-Weltdokumentenregister im Jahr 2017, gewinnen die Forschung an den Werken des Philosophen, sowie die Texte selbst an großer Bedeutung (Trötzmüller 2017). Durch langjährige fachübergreifende Kooperation mit dem Wittgenstein Archiv der Universität Bergen (WAB) (Pichler 2010, 2014) kann das Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwigs-Maximilians-Universität München einen umfassenden Zugang zum Nachlass Ludwig Wittgensteins anbieten. Der Zugang zum Nachlass wird mit einer Suchmaschine und integriertem Faksimile Reader über das Portal WiTTFind (http://wittfind.cis.uni-muenchen.de) ermöglicht. Zur Forschung an der textgenetischen Entwicklung des Prototraktatus wurde als neueste Applikation der Odyssee Reader entwickelt (Still 2018). Durch die intensive Zusammenarbeit mit den Philosophen konnte die Suchmaschine durch die Wittgenstein Advanced Search Tools (WAST) bereits umfangreich erweitert werden (Hadersbeck et al. 2012, 2014) und ermöglicht eine schnelle Suche von konkreten Textstellen im Nachlass, semantischen Ähnlichkeiten in den Themenbereichen Farbe (Krey 2014) und Musik (Röhrer 2017), einen Faksimile Reader (Lindinger 2015), sowie einen Geheimschrift-Übersetzer. Eine Übersicht inklusive der hier vorgestellten Erweiterungen ist in Abbildung 1 zu sehen.

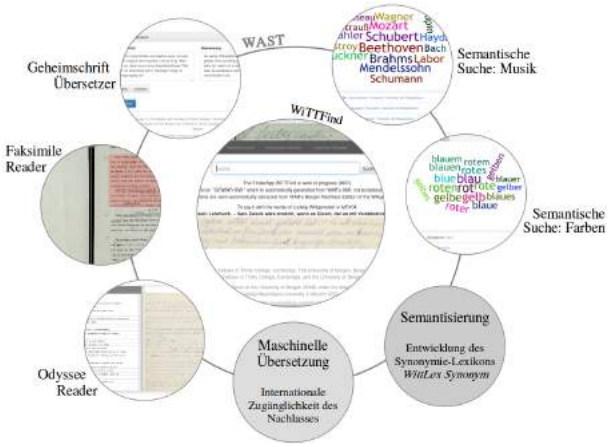


Abbildung 1. Übersicht der Wittgenstein Advanced Tools (WAST) in WiTTFind. Die grau hinterlegten Komponenten werden hier vorgestellt.

Integration der Semantisierung

Auf unserem Poster sollen nun zwei neue Komponenten der WAST vorgestellt werden. Zum einen wird das seit Jahren bestehende digitale Lexikon WiTTLex erweitert. Dieses Lexikon ist im DELA-Format verfasst und ermöglicht unserer FinderApp WiTTFind Suchbegriffe lemmatisiert und zerlegt nach Wortstamm und Partikel im Nachlass von Ludwig Wittgenstein zu finden. Die Besonderheit von WiTTLex besteht darin, dass es auf Wittgensteins Sprache zugeschnitten ist und nur Wörter enthält, die in seinem Nachlass vorkommen. Aufgrund dieser Eigenschaften bietet das Lexikon eine einzigartige Grundlage für sprachliche Untersuchungen in Ludwig Wittgensteins Werken. Um eine detailliertere Textforschung für Fragestellungen semantischer Natur zu ermöglichen wird derzeit im Rahmen eines studentischen Forschungsprojektes ein Synonymie-Speziallexikon, WiTTLex Synonym, entwickelt. Die Grundlage für dieses Lexikon ist einerseits die durch WiTTLex geschaffene Wortdatenbank, sowie andererseits die aus GermaNet (Hamp et al 1997, Henrich et al. 2010) und WordNet (Miller 1995, Fellbaum 1998) extrahierten Synonyme. Equivalent zu WorNet ist GermaNet ein lexikalisch-semantisches Wortnetzsystem, welches an der Universität Tübingen entwickelt wird. Anschließend wird diese Basis in eine dem DELA-System ähnliche Struktur formatiert, sowie manuell getestet und ergänzt. Das entstandene Lexikon kann die Suche von WiTTFind für die Nutzer anreichern, da ähnliche Textstellen gefunden werden können. Der Suchraum wird einerseits durch die Synonyme selbst, andererseits mit Wörtern erweitert, die eine Synonymverlinkung zum Suchwort haben. Ein derartiger schrittweiser Aufbau und Erweiterung eines Synonymielexikons ermöglicht eine Evaluation von GermaNet im sprachlichen Kontext der Philosophie und kann zeigen, für wie viele Wörter Synonyme automatisch gefunden werden konnten, und von welcher Güte die gefundenen Synonyme sind. In einem zweiten Evaluationsverfahren wird verglichen, ob unser finales WiTTLex Synonym eine Verbesserung gegenüber einem rein automatischen, auf GermaNet und WordNet basierenden Systems, bei unserer Ähnlichkeitssuche WiTTSim auf Ludwig Wittgensteins Nachlass zeigt.

Abbildung 2. Beispiel für eine Synonymverlinkung

Neuronale maschinelle Übersetzung

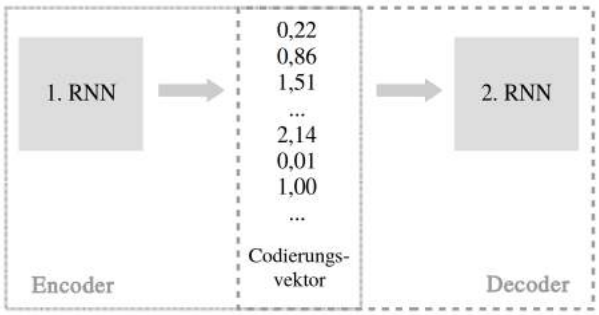


Abbildung 3. Übersicht des neuronalen maschinellen Übersetzungssystems.

Die zweite Erweiterung der WAST betrifft den internationalen Zugang der Suchmaschine für nicht deutschsprachige Wissenschaftler. Die gefundenen Faksimile und deren Transkription aus Bergen können derzeit nur in Originalsprache angezeigt werden. Nur ein sehr geringer Anteil dieser Originaltexte wurde auf Englisch verfasst, während der Großteil in deutscher Sprache geschrieben wurde. Daher wird derzeit im Rahmen eines weiteren studentischen Forschungsprojekts ein maschinelles Übersetzungssystem integriert, um Philosophen und anderen Interessierten aus aller Welt einen möglichst objektiven Zugang zum Nachlass zu ermöglichen. Das Übersetzungssystem wird als Sequence to Sequence Modell (Luong et al. 2015, 2017, Sutskever et al. 2014) implementiert. Dafür werden zwei rekurrente neuronale Netze (RNNs) trainiert, bestehend aus einem Encoder und einem Decoder. Der Encoder berechnet einen Codierungsvektor für den deutschen Textabschnitt, während der Decoder den entstandenen Vektor ins Englische transformiert (Abbildung 3). Werden diese zwei Netze aneinandergeschaltet, erhält man ein neuronales maschinelles Übersetzungssystem, welches den Nachlass vom Deutschen ins Englische übersetzt.

Es muss jedoch angemerkt werden, dass die automatische Übersetzung keinesfalls eine philosophisch-interpretatorische Übersetzung ersetzen kann. Sie kann lediglich eine Grundlage bilden, welche dann in intensiver Zusammenarbeit mit den Philosophen geprüft und optimiert werden kann.

Für zukünftige Arbeiten sollen die übersetzen Texte in die Ähnlichkeitssuche WiTTSim einfließen (Ullrich et al. 2018), wo sie der Aufdeckung von sprachübergreifende Ähnlichkeiten dienen kann (siehe Abbildung 4).

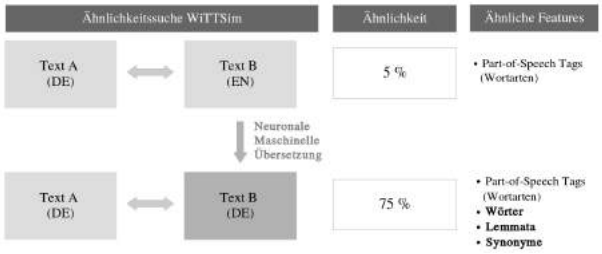


Abbildung 4. Ähnlichkeitsberechnung mit WiTTSim und Vergleich der Ergebnisse mit und ohne maschineller Übersetzung am Beispiel von zwei Texten A und B.

Bibliographie

Fellbaum, Christiane (1998): *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Seebauer, Patrick / Strutynska, Olga (2012): *New (re)search possibilities for Wittgenstein's Nachlass.* 35th International Wittgenstein Symposium 2012, Kirchberg am Wechsel, Contributions, pp. 102-105. Kirchberg am Wechsel: ALWS.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Øyvind (2014): *Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST)*. Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014), pp. 91-96. Madrid.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Bruder, Daniel / Arends, Ina / Baiter, Johannes (2015): *Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind*. 2. Tagung Digital Humanities im deutschsprachigen Raum 23.-27.2 (Graz).

Hamp, Birgit / Helmut Feldweg (1997): *GermaNet - a Lexical-Semantic Net for German*. Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid.

Henrich, Verena / Erhard Hinrichs (2010): *GernEdiT - The GermaNet Editing Tool.* Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta, pp. 2228-2235.

Krey, Angela (2014): *Semantische Annotation von Adjektiven im Big Typescript von Ludwig Wittgenstein*. Bachelorarbeit, CIS.

Lindinger, Matthias (2015): *Entwicklung eines WEB-basierten Faksimileviewers mit Highlighting von Suchmaschinen-Treffern und Anzeige der zugehörigen Texte in unterschiedlichen Editionsformaten*. Masterthesis, CIS.

Luong, Minh-Thang / Hieu Pham / Christopher D Manning (2015): *Effective approaches to attention-based neural machine translation* EMNLP.

Luong, Minh-Thang / Eugene Brevdo / Rui Zhao (2017): *Neural Machine Translation (seq2seq) Tutorial*, https// github.com/tensorflow/nmt, zugegriffen am 9.10.2018.

Miller, George A. (1995): *WordNet: A Lexical Database for English*, in: Communications of the ACM Vol. 38, No. 11: 39-41.

Pichler, Alois (2010): *Towards the New Bergen Electronic Edition*, in: Wittgenstein After His Nachlass. Ed. Nuno Venturinha, pp. 157-172. Houndmills: Palgrave Macmillan.

Pichler, Alois / Bruvik, Tone Merete (2014): *Digital Critical Editing: Separating Encoding from Presentation*, in: Digital Critical Editions. Ed. Daniel Apollon, Claire Bélisle,

Philippe Régnier, pp. 179-202. Urbana Champaign: University of Illinois Press.

Röhrer, Ines (2017): *Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass*, Bachelorarbeit, CIS.

Still, Sebastian (2018): *Ludwig Wittgenstein: 100 Jahre Traktatus. Der Odyssee-Reader, ein web-basiertes Tool zur text-genetischen Suche im Traktatus*, Masterthesis, Ludwig-Maximilians-Universität München.

Sutskever, Ilya / Oriol Vinyals, / Quoc V. Le. (2014): *Sequence to sequence learning with neural networks*, NIPS.

Trötzmüller, Eva (2017): *Unesco-Weltdokumentenerbe - Zwei Neuaufnahmen*, https://www.unesco.at/presse/artikel/article/unesco-weltdokumentenerbe-zwei-neuaufnahmen/, zugegriffen am 12.10.2018

Ullrich, Sabine /Bruder, Daniel / Hadersbeck, Maximilian (2018): *Aufdecken von "versteckten" Einflüssen: Teil-Automatisierte Textgenetische Prozesse mit Methoden der Computerlinguistik und des Machine Learning*, 5. Tagung Digital Humanities im deutschsprachigen Raum 26.2.-2.3. (Köln)

Wie sich die Bilder gleichen. Bildähnlichkeitssuche in Drucken des 16. Jahrhunderts

Götzelmann, Germaine
germaine.goetzelmann@kit.edu
Karlsruher Institut für Technologie, Deutschland

Die Anzahl von Buchdigitalisaten weltweit hat längst eine Datenmenge erreicht, die sich rein manueller Auswertung entzieht (HathiTrust verzeichnet beispielsweise tagesaktuell 5.860.937.950 digitalisierte Buchseiten). Aus diesem Grund gilt es, zusätzliche Auswertungsmethoden und Workflows zu entwickeln, die solche Digitalisatsdaten analysierbar und durchsuchbar machen und damit über explorative Ansätze und Zufallsfunde hinausgehen. Ein Beispiel für eine solche Auswertungsmethode stellt die Suche nach Bildähnlichkeiten zwischen Buchillustrationen dar, die wichtige Aufschlüsse über Zusammenhänge zwischen einzelnen Buchexemplaren, aber auch über die Signifikanz von Bildformeln sowie den Zusammenhang von Bild- und Texttradition liefern kann.

Anwendungsfall: Drucke des 16. Jahrhunderts

In deutschen (Wiegen-)Drucke des 15. und 16. Jahrhunderts insb. volkssprachlicher Texte werden dem geschriebenen Wort zahlreiche Illustrationen zur Seite gestellt. Während diese zu Beginn noch große Textnähe aufweisen, wird für sie zusehends konstitutiv, dass sie bereits von vornherein polyvalent und multifunktional angelegt sind (Vgl. Ott, 1999), um leicht in verschiedene Textzusammenhänge gestellt werden zu können. Dies spiegelt unter anderem ökonomische Interessen wider. Holzschnittvorlagen werden für immer neue Bücher wiederverwendet und somit rekontextualisiert. Die Einschätzung dieser Bildübernahmen reicht in der Forschung in den Einzelfällen von ‚reflektiert' bis ‚sorglos' (Speth 2017, 305).

Im VD16 stehen von den ca. 106000 verzeichneten Drucken des 16. Jahrhunderts etwa 68500 mit Link zu einem Komplettdigitalisat zur Verfügung, die aus den verschiedensten Digitalisierungsprojekten stammen. Dies entspricht auch bei vorsichtigster Schätzung mehreren Millionen von Einzelseiten. Die Posterpräsentation stellt einen Workflow vor, der ausgehend von diesen Digitalisaten diejenigen Buchseiten mit Illustrationen einer Bildähnlichkeitssuche unterzieht und so die Verwendung gleicher und ähnlicher Bilder in verschiedenen Buchexemplaren sichtbar macht.

Workflow zur Bildähnlichkeitssuche

Bildähnlichkeitssuchen erschließen zusehends digitale Bestände, dabei lassen sich zwei Vorgehensweisen unterschieden: die explorative Suche, wie sie die proprietäre Bildähnlichkeitssuche der Bayerischen Staatsbibliothek (BSB) bietet und die strukturierte Suche, wie sie im Projekt 15cBOOKTRADE mittels *VGG Image Search Engine* (VISE)) umgesetzt ist. VISE kommt auch im vorgestellten Workflow zum Einsatz. In einem ersten Schritt dieses Workflows werden aus Digitalisaten die Einzelseiten als Bilddaten extrahiert. In einem zweiten Schritt wird dann auf Methoden der Layoutanalyse zurückgegriffen, um Illustrationen als Bildregionen zu segmentieren. Für den Anwendungsfall der Drucke des 16. Jahrhunderts bietet sich das Tool *LAREX* an, das im Workflow vollautomatisch angewendet wird. Mit diesem Workflowschritt geht eine Reduktion des Suchraums für die Bildähnlichkeitssuche von allen Einzelseiten auf Seiten mit erkannten Bildregionen einher – in der Praxis bedeutet dies eine Verringerung der Seitenmenge um über 75%. Anhand der Segmentierungsergebnisse wird in einem dritten Schritt die Bildähnlichkeitssuche als vollständige Suche durchgeführt: Jede segmentierte Bildregion wird mit allen anderen Seiten mit erkannten Bildregionen verglichen. Aus den Ergebnissen der Bildsuche werden in Schritt vier Ergebnisgraphen erzeugt, die die Treffer ähnlicher Bildregionen abbilden. Dabei erfolgt die Graphgenerierung so, dass nur reziproke Ergebnisse der Bildsuche aufgenommen werden. Zwei Bildregionen im Graph werden nur verbunden, wenn Bildregion A Bildregion B findet und umgekehrt (Abbildung 1). Die einzelnen Bildregionen können anschließend nach Buchzugehörigkeit zusammengefasst werden, woraus sich ein Netzwerk von Buchexemplaren ergibt, dessen gewichtete Verbindungen die Anzahl wiederverwendeter Illustrationen darstellen. Mit diesem Ergebnis eröffnen sich die eigentlichen Auswertungsmöglichkeiten unter Einbeziehung der vorliegenden Metadaten.

Workflowergebnisse als Annotationen

Im automatischen Workflow sind die einzelnen Schritte sowohl parametrisierbar als auch gegen andere Werkzeuge mit gleicher Funktion (beispielsweise andere Layoutanalyse-Tools) austauschbar. Daher sind zum einen flexible Neudurchläufe mit anderen Werten wünschenswert, zum anderen müssen die Ergebnisse der einzelnen Schritte gut vergleichbar sein. Mit diesem Ziel werden die Ergebnisse der Einzelschritte 2-5 in einem einheitlichen, standardisierten Format gemäß der W3C-Empfehlung des *Web Annotation Data Model* in einem Annotationsserver gespeichert. Diese Annotationsform ermöglicht perspektivisch auch die Verbindung mit Metadaten in der Linked Open Data Cloud sowie Veröffentlichung und Rückbindung an die öffentlich zugänglichen Digitalisatsdaten der Bibliotheken zur Nachnutzung für weitere Forschungsfragen.

Herausforderungen

Methodisch kann mit dem beschriebenen Verfahren nicht klar zwischen identischen und ähnlichen Bildern unterschieden werden. Strukturelle Ähnlichkeiten wie großflächige gleichartige Muster in den Holzschnitten können zu systematischen Fehl-Erkennungen führen, sind jedoch relativ leicht herauszufiltern. Anders steht es um korrekte Treffer ohne Signifikanz für die Fragestellung, beispielsweise Druckermarken, Initialen oder Schmuckelemente. Diese müssen ggf. manuell bereinigt werden, was allerdings mit einer Benutzeroberfläche zur Darstellung der einzelnen Regionengraphen mit vergleichsweise geringem Aufwand auch für mehrere hundert Graphen durchführbar ist. Herausforderungen stellen auch die Schnittstellen an Beginn und Ende des Workflows dar. In den Onlinebibliographien fehlen häufig noch automatische Exportmöglichkeiten. Auch die Digitalisate sind oft nur manuell zugreifbar, sodass hier ein automatisches Anstoßen des Workflows basierend auf Rechercheergebnissen kaum möglich ist. Die Vielzahl verschiedener Varianten von Digitalisatspräsentationen der verschiedenen Bibliotheken machen einen Rückbezug der Ergebnisse auf die öffentlich verfügbaren Digitalisate mühsam – aber gerade hier eröffnet das Konzept des *Web Annotation Data Models* interessante Möglichkeiten.

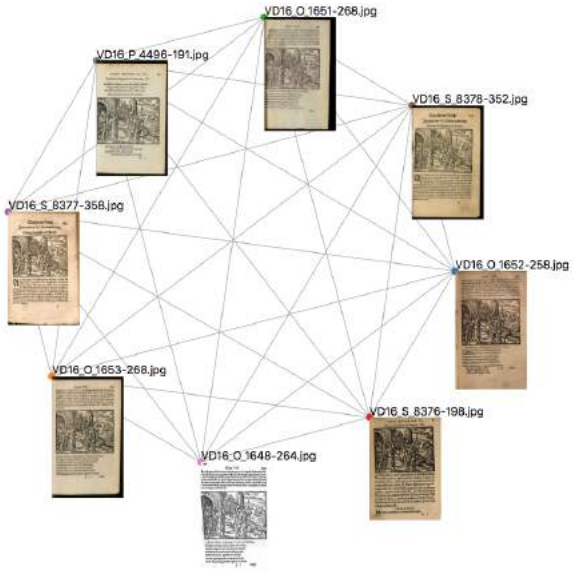


Abbildung 1. Graph von Seiten mit ähnlicher Bildregion

Bibliographie

Brantl, Markus, / Ceynowa, Klaus / Meiers, Thomas / Wolf, Thomas (2017): *Visuelle Suche in historischen Werken*, in: Datenbank Spektrum 17: 53–60.

Hathitrust Statistics and Visualizations: https://www.hathitrust.org/statistics_visualizations [letzter Zugriff 27.10.18].

Malaspina, Matilde / Zhong, Yujie (2017): *Image-matching technology applied to Fifteenth-century printed book illustration*, in: Lettera Matematica 5: 287–292.

Ott, Norbert (1999): *Leitmedium Holzschnitt*, in: Die Buchkultur im 15. und 16. Jahrhundert 2, Hamburg: Maximilian-Gesellschaft 163–252.

Reul, Christian / Springmann, Uwe / Puppe, Frank (2017): *LAREX – A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books*, in Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, ACM.

Speth, Sebastian (2017): *Dimensionen narrativer Sinnstiftung im frühneuhochdeutschen Prosaroman. Textgeschichtliche Interpretation von ‚Fortunatus' und ‚Herzog Ernst'*, Berlin/Boston: De Gruyter.

The 15cBOOKTRADE Project: http://15cbooktrade.ox.ac.uk/ [letzter Zugriff 27.10.18].

Web Annotation Data Model: https://www.w3.org/TR/annotation-model/ [letzter Zugriff 27.10.18].

Zedlers fehlende Seiten: Digitale Quellenkritik und Analoge Erkenntnisse

Müller, Andreas
andreas.mueller@geschichte.uni-halle.de
MLU Halle/Wittenberg, Deutschland

Das Zedlersche *Universal-Lexicon* (17-1754) stellt mit seinen 68 Bänden eine wichtige Quelle für die Erforschung des 18. Jahrhunderts dar. Durch das Portal „Zedler-Lexikon.de“ wurde dieses „bedeutendste Monument des enzyklopädischen Schreibens im Zeitalter der Aufklärung“ (Schneider 2013, S. 9) in einem Kooperationsprojekt zwischen der Herzog August Bibliothek Wolfenbüttel und der Bayerischen Staatsbibliothek München von 2004 bis 2007 digital erschlossen (siehe: Dorn et al. 2008). Entsprechend wurde dieser Webauftritt zu einer wichtigen digitalen Quelle, was sich in hohen Zugriffszahlen widerspiegelt (Dorn et al. 2008, S. 100). Daher erschien es grundlegend angebracht, diesen digitalen Zugriff auf die analoge Quelle einer gründlichen Kritik zu unterziehen.

Das vorgeschlagene Poster präsentiert die Ergebnisse dieser Quellenkritik. Bei einem Abgleich der Anzahl der digitalen Scanbilder mit der gedruckten Seitenzählung zeigte sich, dass teilweise bis zu 180 Seiten pro Band „fehlten“. Die fehlende Seitenzahl geht nach Überprüfung am Original jedoch auf Fehler im Buchdruck zurück und bildet ein deutliches Muster ab, dass die Tätigkeit der verschiedenen Redakteure (Jakob August Franckenstein 1731-32, Paul Daniel Longolius 1733-35 und Carl Günther Ludovici 1738-54) abbildet.

So zeigen deutliche Variationen in der Seitenzählung (Auslassung und Doppelzählung von Seiten) die anfänglichen Schwierigkeiten im verteilten Druck zwischen Halle, Leipzig und Delitzsch (1731-36) und die deutliche Professionalisierung des Drucks nach der Übernahme der Redaktion durch Carl Günther Ludovici ab 1739.

Dies ist für die Forschung interessant, da die Entstehungsbedingungen trotz bereits beachtlicher Erfolge (Calov 2007, Haug 2007, Löffler 2007, Prodöhl 2005, Quedenbaum 1977) immer noch viele Fragen aufwerfen. Zusätzlich zeigt das Poster Ergebnisse einer Analyse von Metadaten, die mittels eines Python Scripts von Zedler-Lexikon.de abgerufen wurden:

Kamen beispielsweise im ersten Band auf jeden Artikel nur etwa 0,3 Verweise, so stieg diese bis zum letzten Band auf rund 2 Verweise pro Artikel an. Darin zeigt sich ein zunehmender Anspruch an die Nutzerfreundlichkeit des Universal-Lexicon von Seiten der Redakteure. Weiters zeigt eine Untersuchung der Länge der einzelnen Artikel einen deutlichen Wandel vom knappen Konversationslexikon zur umfangreichen Enzyklopädie. Der Begriff »Enzyklopädie« bezeichnet in diesem Kontext enzyklopädische Lexika, eine Bedeutung, die sich ab der zweiten Hälfte des 18. Jahrhunderts durch die Vorbildwirkung der Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers herausbildete. Die ältere Bedeutung von »Enzyklopädie« als systematische Wissensordnung trifft auf das UL nicht zu. (Vergleiche hierzu ausführlich Dierse 1977).

Klar erkennbar wird hier jedoch, dass die durchschnittliche Länge der Artikel pro Band von 1732 bis 1754 stetig zunimmt und um 1754 durchschnittlich die 8-fache Länge gegenüber 1732 erreicht.

Hier zeigt sich vor allem in den sehr langen Artikeln der letzten Bände (wie Wien mit 134 Seiten, Wolfische Philosophie mit 174 Seiten oder Zunft mit 54 Seiten) nicht nur der Anspruch, Themen nun in größerem Detailgrad darzustellen, sondern auch ein finanzielles Motiv der Herausgeber. So erschienen von 1747-1750 noch acht Bände zum Buchstaben „W“ und nochmal fünf Bände zum Buchstaben „Z“, wohl um die profitablen Zahlungen der Subskribenten noch möglichst lange zu nutzen.

Die Zusammenführung dieser und weiterer erhobenen Metadaten mit der wechselvollen Geschichte des *Universal-Lexicon*, wie sie die bisherige Forschungsliteratur rekonstruieren konnte, ermöglichte es, viele Ergebnisse der historischen Forschung weitgehend zu bestätigen. Gleichzeitig mahnen die Erkenntnisse damit zu einer kritischen Verwendung des *Universal-Lexicons*, da dieses in den 23 Jahren seiner Entstehung großen inhaltlichen Wandlungsprozessen unterzogen war und daher Artikel aus dem Band A-Am (1732) anders bewertet werden müssen als Artikel aus dem Band Zm-Zz (1750) oder gar aus den vier Supplementbänden (1751-54).

Der Einsatz digitaler Quellenkritik an Zedler-Lexikon.de führte in diesem Fall nicht nur zu einer positiven Bestätigung der digitalen Repräsentation des Werkes, sondern lieferte auch Erkenntnisse zum analogen Werk, die ohne die vorhergehende Digitalisierung und den Einsatz entsprechender Methoden nicht realisierbar gewesen wäre.

Bibliographie

Calov, Carla (2007): *Quellen zu Johann Heinrich Zedler und seinem Lexikon im Stadtarchiv Leipzig*. In: Leipziger Jahrbuch zur Buchgeschichte 16, S. 203–244.

Dierse, Ulrich: *Enzyklopädie. Zur Geschichte eines philosophischen und wissenschaftstheoretischen Begriffs*. Band. 2. Bonn 1977.

Dorn, Nico / Oetjens, Lena / Schneider, Ulrich Johannes (2008): *Die sachliche Erschließung von Zedlers "Universal-Lexicon"*. Einblicke in die Lexicographie des 18. Jahrhunderts. In: Das achtzehnte Jahrhundert 32 (1), S. 96–125.

Haug, Christine (2007): *Das "Universal-Lexicon" des Leipziger Verlegers Johann Heinrich Zedler im politischen Konfliktfeld zwischen Sachsen und Preußen*. In: Leipziger Jahrbuch zur Buchgeschichte 16, S. 301–331.

Jürgens, Hanco / Lüsebrink, Hans-Jürgen (2017): *Enzyklopädismus und Ökonomie im Aufklärungszeitalter. Zur Einführung.* In: Das achtzehnte Jahrhundert 41 (2), S. 197–202.

Löffler, Katrin (2007): *Wer schrieb den Zedler? Eine Spurensuche*. In: Leipziger Jahrbuch zur Buchgeschichte 16, S. 265–284.

Lohsträter, Kai / Schock, Flemming (Hg.) (2013): *Die gesammelte Welt : Studien zu Zedlers "Universal-Lexicon"* ; [Ergebnisse einer internationalen Tagung, die im November 2010 in der Herzog-August-Bibliothek Wolfenbüttel stattgefunden hat]. Wiesbaden: Harrassowitz (Schriften und Zeugnisse zur Buchgeschichte).

Prodöhl, Ines (2005): *"Aus denen besten Scribenten." Zedlers "Universal Lexicon" im Spannungsfeld zeitgenössischer Lexikonproduktion*. In: Das achtzehnte Jahrhundert 29 (1), S. 82–94.

Quedenbaum, Gerd (1977): *Der Verleger und Buchhändler Johann Heinrich Zedler 1706-1751. Ein Buchunternehmer in den Zwängen seiner Zeit ; ein Beitrag zur Geschichte des deutschen Buchhandels im 18. Jahrhundert*. Hildesheim u.a.: Olms.

Schneider, Ulrich Johannes (2013): *Die Erfindung des allgemeinen Wissens : enzyklopädisches Schreiben im Zeitalter der Aufklärung*. Berlin: Akademie-Verlag.

Index der Autorinnen und Autoren

Partner:

Unterstützer: