DHd2020

SPIELRÄUME

DIGITAL HUMANITIES ZWISCHEN MODELLIERUNG UND INTERPRETATION

7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum

digital humanities im
deutschsprachigen raum

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

DHd 2020

Spielräume
Digital Humanities zwischen Modellierung und Interpretation

Konferenzabstracts

Universität Paderborn
02. bis 06. März 2020

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: Christof Schöch
Redaktion und Korrektur der Auszeichnungen:
Nina Seemann, Benjamin Bellgrau
Konvertierung TEI nach PDF: Nina Seemann
https://github.com/NinaSeemann/DHd2020-BoA
Historie der Autorinnen und Autoren sowie Versionen der Konversionsskripte:
Attila Klett (2019)
https://github.com/texttechnologylab/DHd2019BoA
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: Benjamin Bellgrau

Online verfügbar: https://doi.org/10.5281/zenodo.3666690

ISBN 978-3-945437-07-0

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

DHd2020

SPIELRÄUME

DIGITAL HUMANITIES ZWISCHEN MODELLIERUNG UND INTERPRETATION

Vorwort

Bereits zum siebten Mal kommen in diesem Jahr die digitalen Geisteswissenschaftler*innen aus dem deutschsprachigen Raum für die Jahrestagung des DHd-Verbands zusammen, diesmal in Paderborn. Mit dem Book of Abstracts dokumentieren wir dieses "Schaufenster" der aktuellen Forschung in den digitalen Geisteswissenschaften.

Auch in diesem Jahr haben wir weitere Schritte unternommen, um die im Book of Abstracts publizierten Beiträge in Richtung vollwertiger, wissenschaftlicher Publikationen zu entwickeln. Dies bezieht sich diesmal insbesondere auf die Vortragsvorschläge, die als Ergebnis des Reviewing-Verfahrens grundsätzlich für den Vortrag während der Konferenz angenommen wurden. Deren Autor*innen haben wir dieses Jahr aufgefordert, auf die Rückmeldungen in den Gutachten zu reagieren und eine überarbeitete Fassung des Abstracts für das Book of Abstracts einzureichen. Diese Fassung durfte dabei nicht nur deutlich länger sein als der ursprüngliche Beitrag, sondern wurde auch noch einmal von einer oder einem der ursprünglichen Gutachter*innen daraufhin überprüft, ob die Überarbeitung konstruktiv und sinnvoll ausgefallen ist. Diese inhaltliche ebenso wie die formale Aufwertung des Book of Abstract werden wir, so hoffe ich, auch in den nächsten Jahren weiter entwickeln können.

Dass das Konferenzprogramm und das Book of Abstract in dieser Form zustande kommen konnte, verdanken wir vielen tatkräftigen Menschen. An aller erster Stelle möchte ich hier allen Autor*innen von Beiträgen für unsere Jahreskonferenz danken, auch denjenigen, deren Beiträge wir schweren Herzens ablehnen mussten. Sie alle ermöglichen es uns als Community, bei der Jahrestagung die thematische Vielfalt und fachliche Qualität unserer Forschung aufzuzeigen. Ein bisschen Statistik sei erlaubt: wir konnten rund zwei Drittel der Einreichungen als Vortrag und Panel annehmen, bei den Postern und Workshops lag dieser Wert bei rund drei Vierteln.

Dank gebührt außerdem natürlich den mehr als 120 Gutachter*innen, die in diesem Jahr für rund 175 Einreichungen etwa 600 Gutachten verfasst haben. Ohne diese enorme kollektive Anstrengung wäre die Jahrestagung in dieser Form nicht möglich. Den Rahmen für die Begutachtung ebenso wie die Auswertung und Entscheidung über Annahmen und Ablehnung obliegt dem Programmkomittee, dessen Mitgliedern ich an dieser Stelle ebenfalls danken möchte: Stefanie Acquavella-Rauch, Kai-Christian Bruhn (bis Juli 2019 Vorsitzender), Alexander Czmiel, Lisa Dieckmann, Michaela Geierhos (Vertreterin des lokalen Organisationsteams), Katrin Glinka, Andreas Henrich, Patrick Sahle, Stefan Schmunk, Caroline Sporleder, Georg Vogeler und Lars Wieneke. Es war mir eine Freude, mit dieser diskussionsfreudigen, engagierten Gruppe von Kolleg*innen das wissenschaftliche Programm der DHd-Tagung zu erarbeiten.

Ein ganz besonderer Dank gebührt außerdem den lokalen Organisator*innen, allen voran Michaela Geierhos, mit der wir immer in engem Kontakt standen und die uns jederzeit tatkräftig und vorausschauend unterstützte. Sehr dankbar bin ich auch Bejamin Bellgrau, der als Mitglied des lokalen Organisationsteams bei Fragen rund um ConfTool immer schnell eine Lösung gefunden hat. Ein besonderes Dankeschön gebührt schließlich Nina Seemann, die aus den nicht immer perfekten XML-Dateien das vorliegende, wunderschöne Book of Abstracts generiert hat.

Trier, im Februar 2020
Christof Schöch
für das Programmkomitee der DHd2020

Inhaltsverzeichnis

Keynotes

From Modeling to Interpretation
Flanders, Julia ... 13
Humans in the Loop: Humanities Hermeneutics and Machine Learning
Liu, Alan ... 13

Workshops

Annotieren, Analysieren, Visualisieren – Einführung in CATMA 6
Horstmann, Jan; Meister, Jan Christoph; Petris, Marco; Schumacher, Mareike; Flüh, Marie ... 15
Barcamp data literacy: Datenkompetenzen in den digitalen Geisteswissenschaften vermitteln
Wuttke, Ulrike; Lemaire, Marina; Stefan, Schulte; Helling, Patrick; Blumtritt, Jonathan; Schmunk, Stefan ... 18
Bias in Datensätzen und ML-Modellen: Erkennung und Umgang in den DH
Lassner, David; Brandl, Stephanie; Guy, Louisa; Baillot, Anne ... 21
Deep Learning für visuelle Medien: Annotation, Training, Analyse
Howanitz, Gernot; Radisch, Erik ... 24
Digital Humanities from Scratch
Roeder, Torsten; Cremer, Fabian; Dogunke, Swantje; Elwert, Frederik; Lordick, Harald; Ott, Katrin; Söring, Sibylle; Wübbena, Thorsten ... 27
Einführung in TEI-ODD
Stadler, Peter; Bohl, Benjamin W.; Viglianti, Raffaele ... 29
Hackathon „Sortir de la guerre“
Schwandt, Silke; Baillot, Anne; Gervais, Ludovic; Braud, Camille; Thomas, Clement; Bonsergent, Lou-Ann; Strothotte, Adrian; Niewöhner, Laura Maria ... 31
Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse
Kremer, Gerhard; Jung, Kerstin ... 33
Modellierung und Verwaltung von DH-Anwendungen in TOSCA
Schildkamp, Philip; Neuefeind, Claes; Mathiak, Brigitte; Harzenetter, Lukas; Breitenbücher, Uwe; Leymann, Frank 36
Nachlass Ludwig Wittgenstein: Softwaretechnologien und computerlinguistische Methoden der Software-Infrastruktur um die FinderApp WiTTFind
Hadersbeck, Maximilian; Babl, Florian; Eisterhues, Marcel; Röhrer, Ines; Still, Sebastian; Ullrich, Sabine; Landes, Florian; Lindinger, Matthias ... 39
OCR4all – Eine semi-automatische Open-Source-Software für die OCR historischer Drucke
Wehner, Maximilian; Dahnke, Michael; Landes, Florian; Nasarek, Robert; Reul, Christian ... 43
Showtime – sehen und gesehen werden! Erzeugung semantischer (Spiel-)Räume für kollaboratives Arbeiten mit multimedialen Annotationen im Mehrdimensionalen
Wieners, Jan Gerrit; Schubert, Zoe; Türkoğlu, Enes; Niebes, Kai Michael; Eide, Øyvind ... 46
Spielplätze der Theoriebildung in den Digital Humanities
Geiger, Jonathan; Pfeiffer, Jasmin ... 48
Vom Phänomen zur Analyse – ein CRETA-Workshop zur reflektierten Operationalisierung in den DH
Ketschik, Nora; Krautter, Benjamin; Murr, Sandra; Pagel, Janis; Reiter, Nils ... 52

Panels

Altbausanierung mit Niveau – die Digitalisierung gedruckter Editionen ... 57
Datamodelling Drama and (Musical)theater ... 59
Events: Modellierungen und Schnittstellen ... 62
Intertextualität in literarischen Texten und darüber hinaus ... 65
Maschinelles Lernen in den Geisteswissenschaften. Systemische und epistemologische Konsequenzen einer neuen Technologie ... 68
What’s in the news? (Erfolgs-)Rezepte für das wissenschaftliche Arbeiten mit digitalisierten Zeitungen ... 70

Vorträge

Anwendungen von DH-Methoden in der Erschließung und Digitalisierung von Kulturerbe. Ein Vorschlag zur Systematisierung
Franken, Lina ... 74

„As a Hobby at First“ Künstlerische Produktion als Modellierung
Bernhart, Toni ... 77

Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane
Lüschow, Andreas ... 80

Best-practices zur Erkennung alter Drucke und Handschriften. Die Nutzung von Transkribus large- und small-scale
Hodel, Tobias ... 84

Bildrepositorien und Forschung mit digitalen Bildern im Bereich der Kunstgeschichte
Kröber, Cindy; Münster, Sander; Messemer, Heike ... 87

Computationelle Textanalyse als fünfdimensionales Problem
Gius, Evelyn ... 90

Confounding variables in Sub-Genre classification: instructive problems
Jannidis, Fotis; Konle, Leonard; Leinen, Peter ... 94

Critical Machine Vision. Eine Perspektive für die Digital Humanities
Bell, Peter; Offert, Fabian ... 98

Das Werk bildender Künstler*innen im Kontext – Digitale Werkverzeichnisse im semantischen Netz
Effinger, Maria; Sobriel, Nicole ... 101

Der Spielraum zwischen „zu wenig“ und „zu viel“
Du, Keli ... 104

DH’s Next Top-Model? Digitale Editionsentwicklung zwischen Best Practice und Innovation am Beispiel des „Corpus Masoreticum“
Liedtke, Clemens ... 107

Die Digitale Edition der Protokolle des Bayerischen Ministerrats – ein Erfahrungsbericht
Schrott, Maximilian; Reinert, Matthias ... 111

Die Falte: Ein Denkraum für interaktive und kritische Datenvisualisierungen
Brüggemann, Viktoria; Bludau, Mark-Jan; Dörk, Marian ... 114

Die Kanonfrage 2.0
Dziudzia, Corinna; Hall, Mark ... 116

Differenz und Ähnlichkeit in der computergestützten Filiation von Renaissancemusik. Zur datenbasierten Evaluation von Substitutionsmodellen mithilfe von Surrogatdaten
Plaksin, Anna ... 119

3D-Rekonstruktion als Werkzeug der Quellenreflexion
Messemer, Heike; Clados, Christiane ... 123

Ein Schritt zurück: Distinktive Eigenschaften im deutschsprachigen Drama
Krautter, Benjamin ... 127

Erzählerische Spielräume. Medienübergreifende Erforschung von Narrativen im Mittelalter mit ONAMA
Nicka, Isabella; Hinkelmanns, Peter; Landkammer, Miriam; Schwembacher, Manuel; Zeppezauer-Wachauer, Katharina ... 131

Friends with Benefits: Wie Deep-Learning basierte Bildanalyse und kulturhistorische Heraldik voneinander profitieren
Hiltmann, Torsten; Thiele, Sebastian; Risse, Benjamin ... 135

Game On! Digitale Archäologie und Edition zu(m) Spielen
Roeder, Torsten; Rettinghaus, Klaus ... 138

Geschichte aus erster Hand – Der Aufbau eines nationalen Zeitungsportals unter Berücksichtigung der Bedürfnisse verschiedener Nutzergruppen
Landes, Lisa; Dinger, Patrick ... 141

... hungere schon nach dem nächsten Band. Eine Untersuchung von Metaphern für Leseerfahrungen in Web 2.0 Literaturrezensionen
Herrmann, J. Berenike; Messerli, Thomas ... 144

Ikonizität als Erkenntnismittel – Vollständigkeit, Verständlichkeit und Kontextualisierung als Grundprinzipien der Visualisierung
Freyberg, Linda ... 148

Integrating user-specified Knowledge for semi-automatic Coreference Resolution
Schmidt, David; Krug, Markus; Puppe, Frank ... 151

Interpretations- spielräume. Undogmatisches Annotieren literarischer Texte in CATMA 6
Horstmann, Jan; Jacke, Janina ... *154*
Multimodaler Bedeutungstransfer vom Text zum Bild. Granulare Bildklassifikation durch Verteilungssemantik.
Donig, Simon; Maria, Christoforaki; Bernhard, Bermeitinger; Handschuh, Siegfried ... *158*
m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts
Schumacher, Mareike; Flüh, Marie ... *162*
Netzwerkanalyse spielerisch vermitteln mit DraCor und forTEXT: Zur nicht-digitalen Dissemination einer digitalen Methode in Form des Kartenspiels „Dramenquartett“
Horstmann, Jan; Flüh, Marie; Schumacher, Mareike; Fischer, Frank; Trilcke, Peer; Meister, Jan Christoph ... *167*
OMMR4all - ein semiautomatischer Online-Editor für mittelalterliche Musiknotationen
Wick, Christoph; Hartelt, Alexander; Puppe, Frank ... *171*
Partizipatives Design in Digital Humanities Projekten: Checklist, Maßnahmenkatalog und Use-Case
Dogunke, Swantje ... *174*
Passive Präsenz tragischer Hauptfiguren im Drama
Willand, Marcus; Krautter, Benjamin; Pagel, Janis; Reiter, Nils ... *177*
Positivismus der geistigen Gegenstände: Carnap und die Digital Humanities
Heßbrüggen-Walter, Stefan ... *182*
Public Humanities Tools: Der Bedarf an niederschwelligen Services
Hermes, Jürgen; Klinke, Harald; Demmer, Dennis ... *184*
(Re-)Collecting Theatre History: Wissensdinge, Biographien, Wirkungsräume
Mertgens, Andreas; Türkoğlu, Enes; Probst, Nora ... *187*
Redewiedergabe in Heftromanen und Hochliteratur
Brunner, Annelen; Jannidis, Fotis; Tu, Ngoc Duyen Tanja; Weimer, Lukas ... *190*
Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren
Wiedmer, Nathalie; Pagel, Janis; Reiter, Nils ... *194*
Spiele im Spiel – Datenbankbasiertes Arbeiten zur interaktionale Sprache im Dramenwerk von Andreas Gryphius
Eggert, Lisa; Müller, Melissa ... *200*
Spielräume bei der retroperspektivischen Analyse der Wittgenstein-Edition und die Herausforderungen für das Semantic Clustering
Hadersbeck, Maximilian; Ullrich, Sabine; Still, Sebastian; Pichler, Alois ... *202*
Spielräume definieren: Cooking Recipes of the Middle Ages
Steiner, Christian; Klug, Helmut W. ... *205*
Spielräume modellieren. Eine digitale Edition von Giovanni Domenico Tiepolos Bildzyklus Divertimento per li Regazzi auf der Grundlage von CIDOC CRM
Tumanov, Rostislav; Viehhauser, Gabriel; Feldmann, Alina; Koller, Barbara ... *207*
Sprachvarietäten- abhängige Terminologie in der neuronalen maschinellen Übersetzung: Eine Analyse in der Sprachrichtung Englisch-Deutsch mit Schwerpunkt auf der österreichischen Varietät der deutschen Sprache
Heinisch, Barbara ... *211*
SubRosa – Multi-Feature-Ähnlichkeitsvergleiche von Untertiteln
Luhmann, Jan; Burghardt, Manuel; Tiepmar, Jochen ... *215*
Syntaktische Profile für Interpretationen jenseits der Textoberfläche
Andresen, Melanie; Begerow, Anke; Franken, Lina; Gaidys, Uta; Koch, Gertraud; Zinsmeister, Heike ... *219*
Textanalyse mit kombinierten Methoden – ein konzeptioneller Rahmen für reflektierte Arbeitspraktiken
Kuhn, Jonas; Pichler, Axel; Reiter, Nils; Viehhauser, Gabriel ... *223*
Theatre-Tool: Erschließung, Verknüpfung und Web-Präsentation von Theater- und Musikbeständen mit unterschiedlichen Quellentypen
Capelle, Irmlind; Richts, Kristina; Schilke, Elena ... *227*
The rapid rise of Fraktur
Weichselbaumer, Nikolaus; Seuret, Mathias; Limbach, Saskia; Hinrichsen, Lena; Maier, Andreas; Christlein, Vincent ... *229*
„The Vectorian“ – Eine parametrisierbare Suchmaschine für intertextuelle Referenzen
Burghardt, Manuel; Liebl, Bernhard ... *232*
Typisierte Varianz-Analyse von Texten
Balbach, Nico; Reul, Christian; Puppe, Frank ... *235*
Unsichtbares sichtbar machen - semantische Modellierung interpretativer Vorgänge am Beispiel der historischen Bestandsaufnahme der Brandenburgisch-Preußischen Kunstkammern
Wagner, Sarah ... *238*

Varianz, Ambiguität, Unsicherheit. Methodische Schlaglichter zur mittelniederdeutschen Grammatikographie
Ihden, Sarah ... *240*
Volltexttransformation frühneuzeitlicher Drucke – Ergebnisse und Perspektiven des OCR-D-Projekts
Boenig, Matthias; Engl, Elisabeth; Baierer, Konstantin; Hartmann, Volker; Neudecker, Clemens ... *244*
Wege bereiten, vermitteln und Denkräume schaffen! Digital Humanities als community-induziertes Phänomen
Wuttke, Ulrike ... *247*
Welche Beziehungen steuern das Briefkorrespondenz- netzwerk der Reformatoren? Eine Netzwerkanalyse
Roller, Ramona; Schweitzer, Frank ... *250*
Wie wir lesen könnten. StreamreaderPS 0.1
Sahle, Patrick ... *255*
Würgegriff oder Rettungsanker? – Interpretationsspielräume handschriftlicher (Musik-)Quellen im digitalen Kontext
Veit, Joachim ... *258*
Zu den Anforderungen einer musikalischen Stilometrie
Kepper, Johannes ... *260*

Doctoral Consortium

Annotation of Non-Standard Varieties
Seltmann, Melanie E.-H. ... *265*
„Ein lebendiges psychologisches Parlament". Lazarus' und Steinthals Zeitschrift für Völkerpsychologie und Sprachwissenschaft.
Reiners, Stefan ... *266*
Raise your voice! - Über den Zusammenhang zwischen Lautstärkemerkmalen in literarischen Prosatexten und der Emanzipation der Frau von 1848 bis 1920
Guhr, Svenja ... *267*

Posterpräsentationen

Abstract Enhancement. Potentiale der DHd-Konferenzabstracts als Daten/Publikation
Steyer, Timo; Andorfer, Peter; Cremer, Fabian ... *271*
A Linked Open Data Platform for Historical Geographic Data
Görz, Günther; Seidl, Chiara; Thiering, Martin ... *272*
Anforderungen an das Forschungsdaten- management an einer mittelgroßen Universität und Konzeption einer prototypischen Lösung
Jegan, Robin; Gradl, Tobias; Henrich, Andreas ... *274*
Aufbau und Erfahrungen aus dem Digital Humanities Lab der Universität Erlangen-Nürnberg
Scholz, Martin; Klusik-Eckert, Jacqueline ... *276*
Becoming Urban: Ein Spiel mit Räumen
Bürgermeister, Martina; Holzer, Matthias; Nussmüller, Antonia; Scheuermann, Leif; Sonnberger, Jakob ... *277*
Besuch im »Marstheater« – Eine Netzwerkmodellierung von Karl Kraus' Riesendrama »Die letzten Tage der Menschheit«
Fischer, Frank; Busch, Anna; Hechtl, Angelika; Trilcke, Peer; Vogel, Andreas ... *278*
BeyondTheNotes: Ein Tool zur quantitativen Analyse in den digitalen Musikwissenschaften
Ortloff, Anna-Marie; Windl, Maximiliane; Güntner, Lydia; Schmidt, Thomas ... *280*
CASOTEX – Ein Projekt, das sozialwissenschaftliche, interpretative Methoden mit maschinellen Lernverfahren verschränkt
Albrecht, Jens; Lehmann, Robert ... *284*
Cooking Recipes of the Middle Ages: Nachnutzbare Ressourcen eines internationalen Forschungsprojekts
Steiner, Christian; Klug, Helmut W.; Böhm, Astrid; Raunig, Elisabeth; Laurioux, Bruno; Ardesi, Denise; Poirier, Corentin ... *286*
Das Erkenntnispotenzial Digitaler Musikedition
Iffland, Joachim ... *288*
Das gute Glas – Glasgestaltung im Zeitalter der guten Form
Kraft, Anneli ... *289*

Das Theater mit dem Theater: Thementransfer in den Spectators
Fuchs, Alexandra; Geiger, Bernhard; Hobisch, Elisabeth; Koncar, Philipp; More, Jacqueline; Saric, Sanja; Scholger, Martina ... 291
Der Datenpool eines frühneuzeitlichen Self-Trackers, oder: Johann Christian Senckenbergs „Observationes". Ein Distant Reading-Zugang
Faßhauer, Vera ... 292
Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte
Schmidt, Thomas; Bauer, Marlene; Habler, Florian; Heuberger, Hannes; Pilsl, Florian; Wolff, Christian ... 296
Der Event Crawl als Ansatz für den Aufbau von Webarchiven am Beispiel von politischen Wahlkämpfen
Eckl, Markus; Gassner, Sebastian ... 300
Der Haken von Frazier, der Ali 1971 zu Boden schickte, hat jetzt eine URI: Zur Modellierung, Transkription und Visualisierung performativer kultureller Objekte
Geißler, Nils ... 301
Die „Hans Kelsen Werke" (HKW) – eine rechtswissenschaftliche Hybridedition
Reinthal, Angela; Tscheu, Amelie; Trautmann, Marjam ... 303
Digitale Editionen im Spannungsfeld zwischen Formalisierung und Interpretation: Rezensionen der Online-Zeitschrift RIDE als Gradmesser für die Zukunft
Resch, Claudia; Rastinger, Nina ... 304
Digitales Publizieren im Spiegel der Zeitschrift für digitale Geisteswissenschaften: Eine Standortbestimmung
Fricke-Steyer, Henrike; Klaffki, Lisa ... 306
Digitalisierung und Erschließung arkaner Quellen im Virtuellen Archiv „Sachsen und das östliche Europa"
Kunze, Kristina ... 307
Discovery-Service TRIPLE
Schulte, Judith ... 309
Diskursive Strukturen als öffentliche Spielräume in Graphenstrukturen. Konzeption, Modellierung und Auswertung ideengeschichtlicher Netzwerke am Beispiel der Spätaufklärung im Fürstentum Lippe
Schneider, Philipp ... 310
Early Stage Digital Medievalist Subcommittee. Vernetzen, entgrenzen, Spielräume schaffen
Busch, Hannah; Gengnagel, Tessa; Schulz, Daniela ... 313
Eine Programmierschnittstelle für Metadaten zu DH Lehraktivitäten: Die DH Course Registry API
Schmeer, Hendrik; Wissik, Tanja ... 314
Ein Spielraum der Digital Humanities: Die Europäische Sommeruniversität „Kulturen & Technologien"
Annisius, Marie; Burr, Elisabeth; Fußbahn, Ulrike ... 315
Erweiterung eines Forschungsdaten- repositoriums um ein Modul für die Nachnutzbarkeit und Analyse von Textressourcen
Schneider, Gerlinde; Vasold, Gunter ... 316
Geisteswissenschaftliches Forschungsdatenmanage- ment in der Lehre – Konzepte, Methoden, Erfahrungen
Blumtritt, Jonathan; Helling, Patrick; Mathiak, Brigitte; Neuefeind, Claes; Rau, Felix; Schildkamp, Philip; Wieners, Jan Gerrit ... 318
Historische Schulbücher als Spielräume für Digital Humanities? Mapping von unterschiedlichen Metadatenformaten für Bibliotheken und linguistische Analysen
De Luca, Ernesto William; Fallucchi, Francesca; Hertling, Anke; Klaes, Jan Sebastian; Schmitz, Claudia; Towara, Nadine ... 321
Interdisziplinäres Streitgespräch – Nutzerkommentar- analysen aus ethisch-rechtlicher Perspektive
Brokering, Annalena; Guhr, Svenja ... 323
Intermediation der Forschungsinfrastruktur. Ein Rollenmodell für den Umgang mit einer komplexen Infrastrukturlandschaft
Wübbena, Thorsten; Neumann, Katrin; Cremer, Fabian ... 325
Keine Panik! Manifest für Softwareentwicklung in Studierendenprojekten
Eschweiler, Mark; Evers, Anna-Maria; Kruhl, Dominik; Reuhl, Elisabeth; Türkoğlu, Enes ... 327
Kein Spiel(raum): Rechtliche und ethische Rahmenbedingungen geisteswissenschaftlicher Forschung
Scholger, Walter; Hannesschläger, Vanessa ... 328
Keter Shem Ṭov - Prozessualisierung eines Editionsprojekts mit 100 Textzeugen
Molitor, Paul; Necker, Gerold; Pöckelmann, Marcus; Rebiger, Bill; Ritter, Jörg ... 330
Korpusbereinigung für größere Textmengen. Eine (kurze) Problematisierung und ein Lösungsansatz für Duplikate
Adelmann, Benedikt; Gius, Evelyn ... 331

Linked Ogham Stones – Semantische Modellierung und prototypische Analyse irischer Ogham-Inschriften
Homburg, Timo; Thiery, Florian ... 334
Merkmale registrieren oder textuelle Phänomene identifizieren? Zur Vereinbarkeit von automatischer und manueller Textsortenanalyse
Thielert, Frauke; Haaf, Susanne; Schuster, Britt-Marie; Georgi, Christopher ... 337
Metadaten-basierte Visualisierungen im Stilometrie-Paket „Stylo“
Pielström, Steffen; Maciej, Eder ... 340
Modeling disciplinary structure with uniform manifold approximation and projection
Noichl, Maximilian ... 341
Modellierung von Annahmen als Basis für Rekonstruktionen von Architektur
Albers, Laura; Große, Peggy ... 342
Normdaten der Faktenanker für Qualität im semantischen Retrieval. Der Ausbau der Gemeinsamen Normdatei (GND) im Projekt GND für Kulturdaten (GND4C).
Rosenkötter, Martha; Fischer, Barbara ... 344
Opaque – digitale Arbeitsumgebung für die Humanities
Schlicht, Helene; Jentsch, Patrick; Porada, Stephan ... 345
Orte in narrativen biographischen Interviews: automatische Methoden und manuelle Analysen
Ruppenhofer, Josef; Flinz, Carolina; Schmidt, Thomas ... 347
Prosopographische Interoperabilität – Stand der Dinge
Vogeler, Georg; Schlögl, Matthias; Vasold, Gunter ... 348
Requirements on the Punctuation Reconstruction for the Translation of Post-modern Poetry
Meyer-Sickendiek, Burkhard; Baumann, Timo; Hussein, Hussein ... 350
„Romantik“ im aktuellen parteipolitischen Diskurs auf Twitter
Duan, Tinghui; Buechel, Sven; Hahn, Udo ... 352
Routinen, Ressourcen und Tools der digitalen Texterforschung. Ein einfacher Einstieg
Horstmann, Jan; Flüh, Marie; Petris, Marco ... 354
Schmankerl Time Machine. Rechnerisch-explorative Zugänge zur Gastronomie in München
Schulz, Julian; Schneider, Stefanie; Cakir, Osman; Kohl, Linus; Reißer, Alexandra ... 356
Science Data Center für Literatur
Ulrich, Mona; Hess, Jan; Kamzelak, Roland; Kramski, Heinz Werner; Jung, Kerstin; Kuhn, Jonas; Schlesinger, Claus-Michael; Viehhauser, Gabriel; Schembera, Björn; Bönisch, Thomas; Kaminski, Andreas ... 358
SoNAR (IDH): Datenschnittstellen für historische Netzwerkanalyse
Bludau, Mark-Jan; Dörk, Marian; Fangerau, Heiner; Halling, Thorsten; Leitner, Elena; Menzel, Sina; Müller, Gerhard; Petras, Vivien; Rehm, Georg; Neudecker, Clemens; Zellhoefer, David; Moreno Schneider, Julian ... 360
Spielräume des digitalen Publizierens nutzen: Das Online Journal „Entangled Religions“ als ‚Research Hub’
Heinig, Julia; Elwert, Frederik ... 362
Spielräume zwischen Yakshis und Dibias: Vladimir Propps Morphologie des Märchens im ontologiegestützten interkulturellen Vergleich
Pannach, Franziska; Krishnan, Aravind ... 364
Stilometrische Untersuchung von Figurenreden in realistischen Erzähltexten
Weimer, Lukas ... 365
StreamReaderSD 0.2 – Eine prototypische Webanwendung für das Lesen von Texten als Zeichenstrom
Drach, Sviatoslav ... 367
Topic Modeling der Hugo-Schuchardt-Korrespondenz – Möglichkeiten und Grenzen
Saric, Sanja; Scholger, Martina ... 369
Wiener Ballette <Tanz Musik=“mei“ Bild=“jpg“ Text=“tei“ Bewegung=“?“/>
Vera, Grund; Henner, Drewes ... 371
Zwischen geisteswissenschaftlicher Offenheit und informatischer Explikation: Motivsuche als Herausforderung bei der Arbeit mit digitalen Ressourcen
Rastinger, Nina Claudia; Resch, Claudia ... 372

Anhang

Index der Autorinnen und Autoren ... 375

Keynotes

From Modeling to Interpretation

Flanders, Julia
College of Social Sciences and Humanities at the Northeastern University, USA

Abstract

Scholarly modeling and interpretation are complementary elements of a shared social geometry. In shaping corpora, editions, archives, and data sets, the work of modeling is directed at producing convergence and legibility: the preconditions of interpretation. The authority of such modeling work – the consensus it mobilizes and formalizes – is founded in the shared literacies that also animate even the most contrarian interpretive acts. The interpretive agency of the scholarly individual draws its power from the same sources, and moves along the same intellectual vectors, as the shared agency of the standards organization, the committee, the disciplinary imaginary. Modeling in this curation-based mode is a world-making tool whose products are not only models but also guidelines and specifications, constraint systems and conversion pathways, all operating to make a world whose interpretive gestures have been anticipated and accommodated in advance.

The value of such curatorial work within academic digital humanities is considerable, but in the widening and socially urgent space of community-led archiving and public humanities research, these forms of power and agency need renewed scrutiny. The sponsoring, authoritative "we" of the information standard elides the very publics who most need recognition. "Our" models do not yet account for the forms of knowledge and interpretive work arising in those publics. The processes by which digital models are created and applied are hermetic and enmeshed in technical interdependencies. Can we imagine instead techniques, processes, and literacies that can support community-oriented and community-led modeling and interpretation for a new public digital humanities?

Humans in the Loop: Humanities Hermeneutics and Machine Learning

Liu, Alan
English Department at the University of California, USA

Abstract

As indicated by the emergent research fields of computational “interpretability” and “explainability,” machine learning creates fundamental hermeneutical problems. One of the least understood aspects of machine learning is how humans learn from machine learning. How does an individual, team, organization, or society “read” computational “distant reading” when it is performed by complex algorithms on immense datasets? Can methods of interpretation familiar to the humanities (e.g., traditional or poststructuralist ways of relating the general and the specific, the abstract and the concrete, the structure and the event, or the same and the different) be applied to machine learning? Further, can such traditions be applied with the explicitness, standardization, and reproducibility needed to engage meaningfully with the different Spielraum – scope for “play” (as in the “play of a rope,” “wiggle room”, or machine-part “tolerance”) – of computation? If so, how might that change the hermeneutics of the humanities themselves?

In his keynote lecture, Alan Liu uses the example of the formalized “interpretation protocol” for topic models he is developing for the Mellon Foundation funded WhatEvery1Says project (which is text-analyzing millions of newspaper articles mentioning the humanities) to reflect on how humanistic traditions of interpretation can contribute to machine learning. But he also suggests how machine learning changes humanistic interpretation through fresh ideas about wholes and parts, mimetic representation and probabilistic modeling, and similarity and difference (or identity and culture).

Workshops

Annotieren, Analysieren, Visualisieren – Einführung in CATMA 6

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

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

Der ohne technische Vorkenntnisse besuchbare hands-on Workshop bildet eine Einführung in die Möglichkeiten der für Geisteswissenschaftlerïnnen entwickelten, webbasierten Annotations- und Analyseplattform CATMA, deren sechste Version im Oktober 2019 veröffentlicht wurde. Im Zentrum stehen theoretische und praktische Aspekte der digitalen Annotation von (literarischen) Texten sowie die Analyse und Visualisierung dieser Texte und der erstellten Annotationen.

CATMA (*Computer Assisted Text Markup and Analysis*; https://catma.de) ist ein webbasiertes open-source-Tool, das seit 2008 an der Universität Hamburg entwickelt und derzeit von über 60 Forschungsprojekten und ca. 12.000 Nutzerïnnen weltweit genutzt wird. Die im Zuge des DFG-Projektes forTEXT (https://fortext.net) entwickelte sechste Version bietet neben erweiterten technischen Möglichkeiten (wie beispielsweise der Datenversionierung und der Organisation kollaborativer Arbeit in einer Projektstruktur), ein völlig überarbeitetes, intuitiver nutzbares User Interface, das einen leichten Einstieg in die digitale Textannotation und -analyse ermöglicht, ohne dass umfangreiche technische Kenntnisse vonnöten wären, und ohne dass die Nutzerïnnen mit zu vielen (Experten-)Funktionen gleichzeitig konfrontiert würden. Das gesamte Repertoire an Funktionen (wie beispielsweise kollaborative Annotation oder automatische Annotation von Textkorpora) kann von erfahreneren Nutzerïnnen bei Bedarf genutzt werden.

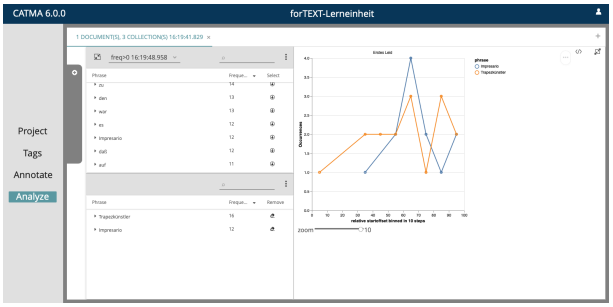


Abbildung 1: CATMA

CATMA unterstützt...

- private und teambasierte Texterforschung durch individuelle wie kollaborative Annotation, Analyse und Visualisierung;
- explorative, non-deterministische Praktiken der Textannotation – CATMA liegt ein diskursiver, diskussionsorientierter Ansatz zur Textannotation zugrunde, der auf die Forschungspraktiken hermeneutischer Disziplinen zugeschnitten ist;
- die nahtlose Verknüpfung von Textannotation, Analyse und Visualisierung 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 (durch die Nutzung von Standoff-Markup) daher möglich, mehrere und sogar widersprechende Annotationen zu vergeben und so der Bedeutungsvielfalt der Texte Rechnung zu tragen. Mit der *Build-Query*-Funktion lassen sich zudem ganz ohne Kenntnisse der Query Language Schritt für Schritt Abfragen kreieren und Textanalysen durchführen. Die Ergebnisse der Analyse können in verschiedenen Varianten visualisiert und für die literaturwissenschaftliche Interpretation und Argumentation genutzt werden. Die sechste Version des Tools integriert gemäß der im Projekt 3DH (http://threedh.net) formulierten Kriterien einer *Dynamic Data Visualisation and Exploration for Digital Humanities Research* ein geisteswissenschaftlich orientiertes Visualisierungskonzept. Dieses Konzept nutzt die *Vega Visualization Grammar*, die auf die von Wilkinson (2005) formulierte generische *Grammar of Graphics* zurückgeht. Schließlich bietet CATMA 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 seit Jahrhunderten zu den textwissenschaftlichen Kernpraktiken gehörende Annotation (vgl. Moulin 2010) lässt sich in sog. Highlights, Freitextkommentare sowie taxonomiebasierte Annotation und Textauszeichnung aufteilen, wobei die Übergänge häufig fließend sind (vgl. Jacke 2018, § 9). Während CATMA 6 auch die Möglichkeit für Highlights und 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 für CATMA spezifische Dreigliederung bietet mehrere Vorteile:

- 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. *Standoff Markup* gespeichert und können damit wahlweise angezeigt oder ausgeblendet werden. Der eigentliche Text wird hierbei nicht verändert. Arbeitet eine Gruppe von Annotatorïnnen 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, Relationalität 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 die sog. *BuildQuery*-Funktion, ein Wizzard, 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, im Erscheinen).

Ablauf

Im Workshop werden wir uns in abwechselnden Präsentations- und Hands-on-Phasen der textanalytischen Arbeit in CATMA 6 nähern. Nach einer generellen Einführung in das Tool werden die Teilnehmerïnnen 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 Teilnehmerïnnen 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 kritische Reflektionen, Diskussionen sowie individuelle Rückfragen (theoretischer, praktischer und technischer Art) auf jedem Niveau und in Bezug auf die Projekte der Teilnehmerïnnen wird ausreichend Möglichkeit bestehen.

Zeitplan

Im Workshop werden wir anhand von Kafkas Erzählung *Erstes Leid* und narratologischen Kategorien der erzählerischen Distanz beispielhaft den Arbeitsablauf der digitalen Texterforschung praktisch kennenlernen (die Teilnehmenden können jedoch auch gerne mit selbst mitgebrachten Texten und Annotationskategorien arbeiten):

- 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 (https://fortext.net), in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die traditionelleren 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.

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.

Marie Flüh, M.Ed.

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

Marie ist Master of Education und interessiert sich besonders für Christoph Martin Wieland und seine Zeitgenossen. Außerdem liegt ihr Forschungsschwerpunkt auf der Wertung von Literatur. Sie studierte in Kiel und Hamburg und ist nun wissenschaftliche Mitarbeiterin an der Universität Hamburg.

Zahl der möglichen Teilnehmerïnnen

Bis zu 30 Personen.

Benötigte technische Ausstattung

Teilnehmerïnnen 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: *DH-Commons Journal* 1.

Gius, Evelyn (im Erscheinen): „Digitale Hermeneutik: Computergestütztes close reading als literaturwissenschaftliches Forschungsparadigma?", in: Jannidis, Fotis / Winko, Simone / Rapp, Andrea / Meister, Jan Christoph / Stäcker, Thomas (eds.): *Digitale Literaturwissenschaft. DFG-Symposium Villa Vigoni, 2017*. Berlin, New York: de Gruyter.

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 (letzter Zugriff 20. Dezember 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". *Digital Humanities Conference* 2010, London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html (letzter Zugriff 20. Dezember 2018).

Wilkinson, Leland (2005): *The grammar of graphics* . 2. Aufl. New York: Springer.

Barcamp data literacy: Datenkompetenzen in den digitalen Geisteswissenschaften vermitteln

Wuttke, Ulrike
ulrike.wuttke@gmx.net
Fachhochschule Potsdam, Fachbereich Informationswissenschaften, RDMO, Deutschland

Lemaire, Marina
marina.lemaire@uni-trier.de
Universität Trier, Servicezentrum eSciences, Deutschland

Stefan, Schulte
stefan.schulte@uni-marburg.de
Philipps-Universität Marburg, Marburg Center for Digital Culture & Infrastructure (MCDCI), 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

Schmunk, Stefan
stefan.schmunk@h-da.de
Hochschule Darmstadt, Fachbereich Media, HeFDI, Deutschland

Beschreibung des Themas: Vermittlung von data literacy in den Geisteswissenschaften

Nachdem beim Thema Forschungsdatenmanagement (FDM) auf politischer Ebene lange die institutionelle Verankerung, z. B. über FDM-Policies (vgl. Forschungsdaten.org (o.J.), Helbig et al. 2018) sowie der Infrastrukturaufbau im Vordergrund stand, findet mittlerweile eine Fokussierung auf die Vermittlung von Kompetenzen im Umgang mit Forschungsdaten – data literacy (vgl. RfII 2019, Schüller et al. 2019) – statt, auch im Prozess zur Errichtung einer Nationalen Forschungsdateninfrastruktur (NFDI). In praktisch allen geisteswissenschaftlich geprägten Konsortien finden sich Aussagen zur Kompetenzvermittlung und mit der fachübergreifenden Konsortiumsinitiative CompeNDI gibt es sogar einen Antrag, der Datenkompetenzen in den Mittelpunkt stellt.[1] Auch in der Neuauflage der Empfehlungen zur Sicherung der guten wissenschaftlichen Praxis (DFG 2019) wird wiederholt der verantwortungsvolle und möglichst offene Umgang mit Forschungsdaten thematisiert und es werden FDM-Kompetenzen für die Sicherung der Forschungsqualität und -exzellenz als unabdingbare wissenschaftliche Schlüsselkompetenzen gezählt (vgl. Wuttke & Klar 2019).

Diese zunehmende Fokussierung auf data literacy steht durchaus im Kontrast zur Situation in den geisteswissenschaftlichen Studiengängen und im Forschungsalltag, in denen die explizite Vermittlung von Datenkompetenzen keine Selbstverständlichkeit ist. Dies liegt u. a. an einer grundsätzlichen Akzeptanzproblematik von FDM in diesen Disziplinen. So werden beispielsweise der Nutzen des FDM und der FAIR-Prinzipien[2] angezweifelt, da der zu erwartende zeitliche und finanzielle Mehraufwand den Nutzen (im Sinne wissenschaftlicher Reputation) nicht rechtfertige. Vielen Wissenschaftler*innen fehlt die Vorstellungskraft, dass ihre Daten für nachfolgende Forschungsprojekte nützlich sein könnten. Zum Teil werden Forderungen zur Offenlegung der Daten und der zur Erstellung und Analyse verwendeten Methoden und Werkzeuge als Angriff auf die Wissenschaftsfreiheit verstanden beziehungsweise als indirekter Vorwurf bislang nicht nach den Regeln der guten wissenschaftlichen Praxis gearbeitet zu haben. Auch wird in Frage gestellt, ob sich die qualitativen, insbesondere hermeneutischen Methoden und Erkenntnisprozesse der Geisteswissenschaften, mittels digitaler Daten und Systeme abbilden lassen. Zudem schätzen viele Geisteswissenschaftler*innen ihre FDM-Kompetenzen als unzureichend ein und fühlen sich von den Anforderungen (zu recht?) überfordert (vgl. Lemaire 2018, 238).

Die beschriebenen Hemmnisse für die Akzeptanz und Implementierung des FDM in der geisteswissenschaftlichen Praxis sind inzwischen bekannt und Hochschulen und Forschungseinrichtungen sind dazu aufgefordert, verstärkt Maßnahmen und Angebote zur Vermittlung von Datenkompetenzen, als Oberbegriff unter den hier FDM-Kompetenzen subsumiert werden sollen, zu etablieren (vgl. RfII 2016, 50). Hierfür liegen bereits gute allgemeine Konzepte vor (vgl. FDMentor & DINI/nestor-AG Forschungsdaten 2018, Dolzycka et al. 2019, Wiljes & Cimiano 2019), es gibt jedoch noch wenig disziplinspezifische Erfahrungen im Bereich der Geisteswissenschaften. Es stellen sich Fragen nach dem “Wann” und “Wie” des Erwerbs von FDM-Kompetenzen, nach geeigneten didaktischen Formaten oder der Abgrenzung bezüglich tiefergehender DH-Kompetenzen, wie sie in spezialisierten Studiengängen vermittelt werden. Diese und weitere Fragen sollen unter Einbeziehung unterschiedlicher Perspektiven im Rahmen eines Barcamps diskutiert werden, weil dieses offene, partizipatorische Format, das stark vom Input aller Teilnehmer*innen lebt, besonders geeignet scheint für eine explorative Diskussion komplexer Themenbereiche.

Das geplante Barcamp ist Teil der Bemühungen der DHd-AG Datenzentren, weiterführende Anforderungen und Aufgaben des langfristigen digitalen Kulturwandels in den Geisteswissenschaften zu eruieren. Hierfür ist bei der DHd 2020 auch ein Panel zur Datenqualität vorgesehen. Ziel beider Aktivitäten ist langfristig die Schaffung positiver Anreize für FDM und Forschungsdatenpublikationen aus den Bedürfnissen der wissen-

schaftlichen Praxis heraus. Speziell mit dem Barcamp möchte die DHd-AG Datenzentren:

- einen Erfahrungsaustausch über konkrete Vermittlungsformen und didaktische Konzepte initiieren,
- Geisteswissenschaftler*innen, FDM-Expert*innen etc., die sich mit der Kompetenzvermittlung befassen, vernetzen,
- einen Beitrag zur Diskussion über den Stellenwert von Datenkompetenzen in den Geisteswissenschaften leisten sowie
- Impulse für die zukünftige Arbeit der AG Datenzentren ableiten.

Workshopformat: Barcamp

Die Einreichenden möchten im Rahmen eines eintägigen Barcamps gemeinsam mit interessierten Wissenschaftler*innen, Forschungsdatenmanager*innen etc. die oben skizzierten Aspekte der Vermittlung von Datenkompetenzen in den Geisteswissenschaften, sowie weiterführende Themen und Fragen, diskutieren. Die für dieses Format typische dynamische, interaktive Entwicklung der Tagesordnung scheint für eine agile "Szene", wie die des FDM, als großer Vorteil und das Format hat sich schon in ähnlichen Kontexten bewährt auf deren Erfahrungswerte die Organisator*innen zurückgreifen können (vgl. Budd et al. 2015, Dogunke et al. 2018, Tóth-Czifra & Wuttke 2019, Muuß-Merholz 2019).

Das wichtigste Merkmal eines Barcamps ist die gemeinsame Programmgestaltung durch Organisator*innen und Teilnehmer*innen, d.h. zu einem Barcamp können alle Beteiligten hierarchieunabhängig aus ihrer Erfahrungswelt beitragen und gemeinsam zu neuen Lösungsansätzen gelangen.

Potentielle Themen und Fragen

Die folgende Sammlung potentieller für das Barcamp zentraler Themen und Fragen aus dem Bereich der Lehre und Vermittlung von data literacy, insbesondere FDM, in den Geisteswissenschaften, dient einem ersten Eindruck. Sie beruht auf den Erfahrungen der Einreichenden und der aktuellen Forschung und erhebt keinen Vollständigkeitsanspruch:

- Welche didaktischen Konzepte eignen sich für welche Zielgruppen?
- Welche Formate eignen sich für die (weiterbildende) Sensibilisierung und Qualifikation (z. B. allgemeine Workshops, Coffee Lectures, Learning by Doing etc.)?
- Welche Strategien eignen sich, um data literacy in geisteswissenschaftliche Curricula zu integrieren?
- Über welche Kompetenzen müssen FDM-Lehrende und -Trainer*innen verfügen und wie kann man diese vermitteln?
- Welche Datenkompetenzen benötigen alle Geisteswissenschaftler*innen und welche sollten spezifisch in DH-Studiengänge integriert werden?
- Welche Aspekte von data literacy sind spezifisch für die Geisteswissenschaften, welche generisch?
- Wie lassen sich Forschende für FDM gewinnen: Top-Down oder Bottom-Up, Push oder Pull?
- Welche Akteure spielen für die Etablierung von Vermittlungsangeboten eine Rolle?
- Welche Maßnahmen gibt es zur Verbesserung der Zugänglichkeit zu FDM-Anlaufstellen bzw. -Institutionen und Serviceangeboten?
- Welche Beratungskompetenzen und -strategien sind zur Vermittlung von bedarfsorientierten FDM-Kompetenzen und -Lösungen innerhalb von Beratungsgesprächen nötig?

Durchführung / Ablauf

Das Barcamp-Format sieht vor, dass zum Veranstaltungsbeginn alle Themenvorschläge gesammelt werden und auf einem „Marktplatz" verhandelt wird, welche Diskussionsgruppen mit welchen Formaten (z. B. Gruppendiskussion, Fishbowl, Knowledge Café, vgl. DCC 2019) entstehen und wer an welcher Diskussionsgruppe teilnimmt. Zur Themenfindung für das Barcamp werden daher die Organisator*innen im Vorfeld zum einen über verschiedene Kanäle aus dem Bereich der DH- und FDM-Communities für Vorschläge werben, zum anderen werden die Organisator*innen aus der eigenen Praxis Themen vorschlagen. Alle Themenvorschläge werden zentral online gesammelt, damit sich alle Interessierten über die eingegangenen Vorschläge informieren können. Alle Vorschläge werden zusammen mit tagesaktuellen Vorschlägen zu Beginn des Barcamps auf dem "Marktplatz" anhand des Feedbacks der Teilnehmer*innen gruppiert, priorisiert und darauf basierend die endgültige Tagesordnung festgelegt. Zusätzlich wird es zur allgemeinen inhaltlichen Unterfütterung bzw. zur Einstimmung auf die Barcamp-Sessions kurze "Teaser-Talks" (ca. 2-3 Min.) geben (u. a. von Mitgliedern des Organisationskomitees bzw. Teilnehmer*innen, im Vorfeld wird auf diese Möglichkeit hingewiesen).

Der zeitliche Rahmen ist so gestaltet, dass es möglich sein wird, grundlegende Fragen zu thematisieren und spezifische Aspekte zu vertiefen. Wir sehen für die thematische Diskussion 45-minütige Sessions für kleinere Gruppendiskussionen (ggf. andere Formate) vor, die je nach Anzahl der Themenvorschläge, Interesse und Teilnehmerzahl parallel stattfinden können. Für die Dokumentation, Moderation und Durchführung des Barcamps sind insbesondere die Einreichenden verantwortlich. Zusätzlich werden alle thematischen Sessions durch die Gruppen selbst dokumentiert (z. B. Flipcharts, Etherpad) und anschließend die Ergebnisse dem Plenum präsentiert. Hierfür wird jeweils ein/e Verantwortliche/r benannt (Dokumentator*in & Präsentator*in). Für die Ergebnispräsentation der thematischen Sessions sind am Ende des Nachmittags gesonderte Slots vorgesehen. Die Dokumentationsmaterialien dienen als Grundlage für weitere Formate der Ergebnissicherung und -verbreitung (siehe unten).

Organisatorisches

Ziele & Sicherung der Ergebnisse

Es ist ein ausführlicher Blogpost zu den Ergebnissen geplant. Abhängig vom Feedback der Teilnehmenden und der breiteren Community sind weitere Formate (White Paper, Artikel) möglich.

Beitragende

Das Barcamp wird von Mitgliedern der DHd-AG Datenzentren und ausgewiesenen Expert*innen organisiert und durchgeführt:

- Ulrike Wuttke (Potsdam) ist stellvertretende Sprecherin der AG Datenzentren und Mitarbeiterin im DFG-Projekt RDMO. Sie verfügt über Expertise im Bereich Forschungsdatenmanagement unter besonderer Berücksichtigung nationaler und internationaler Infrastrukturen und Community-Anforderungen und lehrt in diesem Bereich.
- Marina Lemaire (Trier) ist Referentin für Projektmanagement im Bereich digitaler Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften am Servicezentrum eSciences und Mitglied in der AG Datenzentren. Ihre Expertise beruht auf mehr als 10-jähriger FDM-Beratungspraxis in interdisziplinären Forschungskontexten, der Forschung zur FDM-Implementierung an Forschungseinrichtungen und der Durchführung von FDM-Workshops, Einzelschulungen und Informationsveranstaltung.
- Patrick Helling und Jonathan Blumtritt (Köln) vertreten das DCH in der AG Datenzentren. Sie greifen auf eine über 6-jährige Beratungserfahrung im geisteswissenschaftlichen FDM zurück, betreiben aktives FDM an der Universität zu Köln (UzK) und sind im Bereich der universitären FDM-Lehre tätig. Jonathan Blumtritt ist außerdem technischer Koordinator im BMBF-Verbundprojekt KA3.
- Stefan Schmunk (Darmstadt) ist Professor für Informationswissenschaft / Digital Libraries an der Hochschule Darmstadt (h-da) und beschäftigt sich in Forschung und Lehre seit zehn Jahren mit Forschungsdaten, FDM und digitalen Forschungsinfrastrukturen in den Geistes- und Kulturwissenschaften. Er leitet seit April 2019 das HDA-Teilprojekt der Hessischen Forschungsdateninfrastruktur (HeFDI).
- Stefan Schulte (Marburg) ist Koordinator des Marburg Centers for Digital Culture & Infrastructure (MCDCI, in Gründung) und arbeitet seit mehreren Jahren zum Forschungsdatenmanagement in den Geisteswissenschaften. Er ist Mitherausgeber der Open Access-Zeitschrift "Bausteine Forschungsdatenmanagement" und hat 2018 das Projekt „TRUST - Training zum Umgang mit sensiblen Forschungsdaten" durchgeführt.

Zusätzlich liegen bereits Interessenbekundungen zur Teilnahme aus der Community vor (u. a. seitens FDMentor und der DINI-Nestor UAG Schulungen/Fortbildungen).

Zahl der möglichen Teilnehmerinnen und Teilnehmer

Ca. 30-40 Teilnehmer*innen zuzüglich des Organisationskommitees.

Benötigte technische Ausstattung

- Aufgrund der vorgesehenen Gruppendiskussionen wäre ein gut unterteilbarer, großer Raum bzw. mehrere Räume sinnvoll
- Moderationsmaterialien, insbesondere Flipcharts und eine große Pinnwand für die Sessionplanung (inkl. Moderationskarten, Flipchartstiften und Pins) und kleinere Pinnwände für die Gruppenarbeit
- Beamer für Begrüßung und einführende Teaser-Talks

Fußnoten

1. Siehe NFDI-Absichtserklärungen: https://www.dfg.de/foerderung/programme/nfdi/absichtserklaerungen/index.html [letzter Zugriff 10.09.2019].
2. Siehe https://www.go-fair.org/fair-principles/ [letzter Zugriff 10.09.2019].

Bibliographie

Budd, A. / Dinkel, H. / Corpas, M. / Fuller, J.C. / Rubinat, L. / Devos, D.P. / Khoueiry, P.H. / Förstner, K.U. / Georgatos, F. / Rowland, F. / Sharan, M. / Binder, J.X. / Grace, T. / Traphagen, K. / Gristwood, A. / Wood, N.T. (2015): "Ten simple rules for organizing an unconference", in: *PLoS Comput Biol.* 11, e1003905, DOI: 10.1371/journal.pcbi.1003905.

DCC (2019): "Unconference", Blogpost: http://www.dcc.ac.uk/events/idcc19/unconference [letzter Zugriff 10.09.2019].

DFG (2019): *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*, Bonn, https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf [letzter Zugriff 10.09.2019].

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 10.09.2019].

Dolzycka, Dominika / Biernacka, Katarzyna / Helbig, Kerstin / Buchholz, Petra (2019): *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement (Version 2.0)*, DOI: http://doi.org/10.5281/zenodo.2581292.

FDMentor & DINI/nestor-AG Forschungsdaten (2018): "Materialkatalog zum Forschungsdatenmanagement (Version 1.0) [Data set]". DOI: http://doi.org/10.5281/zenodo.1209284.

Forschungsdaten.org (o. J.): "Data Policies", Webseite, https://www.forschungsdaten.org/ [letzter Zugriff 10.09.2019].

Helbig, Kerstin / Hahn, Uli / Jagusch, Gerald / Rex, Jessica (2018): "Erstellung und Realisierung einer institutionellen Forschungsdaten-Policy" in: *Bausteine Forschungsdatenmanagement* 1: 17-23, DOI: https://doi.org/10.17192/bfdm.2018.1.7945.

Lemaire, Marina (2018): "Vereinbarkeit von Forschungsprozess und Datenmanagement. Forschungsdatenmanagement nüchtern betrachtet" in: *o-bib. Das offene Bibliotheksjournal* 5(4): 237–247, DOI: https://doi.org/10.5282/o-bib/2018H4S237-247.

Muuß-Merholz, Jöran (2019): *Barcamps & Co: Peer to Peer-Methoden für Fortbildungen*, Wein-

heim: Beltz http://www.content-select.com/index.php?id=bib_view&ean=9783407367082 [letzter Zugriff 11.12.2019].

RfII (2016): *Leistung aus Vielfalt – Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*, Göttingen, http://www.rfii.de/?wpdmdl=1998 [letzter Zugriff 10.09.2019].

RfII (2019): *Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft*, Göttingen, http://www.rfii.de/?wpdmdl=3883 [letzter Zugriff 10.09.2019].

Schüller, Katharina / Busch, Paulina / Hindinger, Carina (2019): *Future Skills: Ein Framework für Data Literacy*. Hochschulforum Digitalisierung. Arbeitspapier Nr. 47. https://hochschulforumdigitalisierung.de/sites/default/files/
dateien/HFD_AP_Nr_47_DALI_Kompetenzrahmen_WEB.pdf [letzter Zugriff 10.09.2019].

Tóth-Czifra, Erzsébet / Wuttke, Ulrike (2019): "Loners, Pathfinders, or Explorers? How are the Humanities Progressing in Open Science?" (Bericht vom Open Science Barcamp Berlin, 2019), Blogpost Generation R, 20.04.2019, DOI: https://doi.org/10.25815/x516-wf23.

Wiljes, Cord / Cimiano, Philipp (2019): "Teaching Research Data Management for Students" in: *Data Science Journal* 18(1), DOI: http://doi.org/10.5334/dsj-2019-038.

Wuttke, Ulrike / Klar, Jochen (2019): "How FAR is FAIR? Der öffentliche Zugang zu geisteswissenschaftlichen Forschungsdaten als gute wissenschaftliche Praxis und die Rolle des Forschungsdatenmanagements", Vortragsfolien, DOI: http://doi.org/10.5281/zenodo.3365979.

Bias in Datensätzen und ML-Modellen: Erkennung und Umgang in den DH

Lassner, David
lassner@tu-berlin.de
TU Berlin

Brandl, Stephanie
stephanie.brandl@tu-berlin.de
TU Berlin

Guy, Louisa
louisa.guy.etu@univ-lemans.fr
Le Mans Université

Baillot, Anne
anne.baillot@univ-lemans.fr
Le Mans Université

Vortragende

David Lassner

Master Informatik David Lassner, Doktorand an der TU Berlin im Bereich Maschinelles Lernen für Digital Humanities, insbesondere für quantitative Literaturanalyse.
lassner@tu-berlin.de

Stephanie Brandl

Dipl. Math. Stephanie Brandl, Technische Universität Berlin. Forschungsschwerpunkte: Maschinelles Lernen, Natural Language Processing.
stephanie.brandl@tu-berlin.de

Louisa Guy

Louisa Guy, Doktorandin, Le Mans Université. Forschungsinteressen: Digitale Textanalyse, Anwendung von Methoden der Computerlinguistik auf sozialwissenschaftliche Kontexte.
louisa.guy.etu@univ-lemans.fr

Anne Baillot

Prof. Dr. Anne Baillot, Le Mans Université. Forschungsschwerpunkte: Digitale Philologie, Digital Humanities, Translation Studies.
anne.baillot@univ-lemans.fr

Anforderungen

Maximalanzahl Teilnehmender: 25

Räumliche Anforderungen:

- Beamer
- Whiteboard/Tafel
- Stromversorgung für Laptops der Teilnehmenden
- Wifi

Anforderungen an die Teilnehmenden:

Wir erwarten, dass die Teilnehmenden ihre eigenen Laptops mitbringen, die bestenfalls schon die nötige Software vorinstalliert haben. Wir werden kurz vor der Konferenz eine Willkommens-E-Mail mit den Softwareanforderungen verschicken. Die praktischen Sitzungen werden mithilfe von Jupyter Notebooks (Python3, Jupyter) abgehalten. Wir planen zusätzlich als Absicherung einen Online-Zugang zu einem Jupyter-Hub Server mit vorinstallierten Paketen für Teilnehmende, bei denen die Installation Schwierigkeiten macht. Die praktischen Sitzungen sind so konzipiert, dass nur sehr geringe, bis gar keine Programmierkenntnisse notwendig sind. Im Wesentlichen sollen die Teilnehmenden die Parameter und Eingabedaten der vorgegebenen Programme modifizieren, Teil-

nehmende mit mehr Programmierkenntnissen ermutigen wir natürlich tiefer in die Programme einzusteigen und auch diese zu modifizieren.

Beschreibung

Der Workshop besteht aus einem allgemeineren Teil zu Bias im Maschinellen Lernen, in dem grundlegend in die Thematik eingeführt wird, und einem spezifischeren Teil, in dem ML-Biases im Kontext von DH behandelt werden. Beide Teile beinhalten Vortrags- sowie Mitmachsessions. Ziel des Workshops ist es, dass die Teilnehmenden sich des Problems von Bias in Machine Learning Modellen bewusst werden und die grundlegenden Techniken zur Erkennung und zur Unterdrückung von Biases kennenlernen. Es soll außerdem gemeinsam erarbeitet werden, auf welche Weise DH-ForscherInnen mit den Biases umgehen können - denn in vielen Anwendungen sind diese nicht gewünscht: Ein System zur Vorauswahl von Bewerbern sollte Männer nicht bevorzugen,[1] ein Modell zur Gesichtserkennung sollte keinen Unterschied in der Genauigkeit haben, weil sich die Hautfarbe der Personen auf den Bildern ändert (Buolamwini et al. 2018), und ein Modell zur Erkennung von Hate-Speech im Internet sollte nicht kontextfrei bspw. Begriffe wie "homosexuell" als toxisch einstufen.[2]

Gleichzeitig können Biases in ML Modellen erwünscht sein, wenn man beispielsweise die Veränderung von Biases in Sprache analysiert.

Teilnehmende werden im Vorfeld ermutigt eigene Daten mitzubringen, mit denen sie im zweiten praktischen Teil experimentieren können.

Das Workshopprogramm wird online unter bias-ml-dh.davidlassner.com öffentlich zur Verfügung gestellt und die Kursmaterialien auf Github unter github.com/millawell/bias-ml-dh veröffentlicht. Dort sollen die Teilnehmenden auch schon im Vorfeld einen Eindruck bekommen, welche ihrer eigenen Daten möglicherweise zum Workshop mitgebracht werden könnten.

Zeittafel

Zeit	Titel	Vortragende
Halbtag 1.1	Einleitung, Motivation	Anne Baillot, David Lassner
Halbtag 1.2	Erkennung von Biases in ML	David Lassner
Kaffepause		
Halbtag 1.3	Verhinderung von Biases in ML	Stephanie Brandl
Halbtag 1.4	Praktische Sitzung 1	
Halbtag 2.1	Autorinnen um 1800	Anne Baillot
Halbtag 2.2	Revolte auf Twitter	Louisa Guy
Kaffepause		
Halbtag 2.3	Praktische Sitzung 2	
Halbtag 2.4	Abschlussdiskussion	

Erkennung von Biases

Zu Beginn steht die Begriffsklärung (Datenbias, Modellbias, etc.) und konkrete Beispiele zur Erkundung verschiedener Biases in verschiedenen Datensätzen, sowie Modellarchitekturen. Beispielsweise anhand konkreter Architekturen neuronaler Netze zur Textklassifikation, deren erster Layer ein Embedding-Layer auf Word2Vec-Basis ist (Mikolov et al. 2013).

Es werden verschiedene Methoden vorgestellt, wie Biases in Modellen und Daten erkannt werden können (Caliskan et al. 2017, May et al. 2019, Garg et al. 2018, Bolukbasi et al. 2016, Swinger et al. 2019).

Wie lassen sich Biases verhindern?

Innerhalb der letzten 3 Jahre wurden zahlreiche Methoden veröffentlicht, die darauf abzielen Biases in Word Embeddings und anderen NLP Anwendungen zu verringern. In diesem Teil wollen wir einen Überblick über die wichtigsten Methoden verschaffen, ihre Stärken und Schwächen aufzeigen und diskutieren.

Aktuell können diese Methoden in 3 Kategorien eingeteilt werden:

Manipulation von Datensätzen

Datensätze werden so verändert, beispielsweise durch Datenanreicherung, dass Biases im Datensatz nicht mehr zu finden sind und so auch nicht mitgelernt werden. Zum Beispiel schlagen Zhao et al (2018) vor, jeden Satz in einem Datensatz zu kopieren, sodass dieser in mehreren Varianten vorkommt: eine für jedes grammatikalische Geschlecht. So wird eine balancierte Repräsentation zwischen den (binären) Geschlechtern garantiert. Bestehende ML-Methoden die ansonsten biased Ergebnisse erzeugen, können so faire Modelle lernen.

Anpassung der Methode

Zhang et al (2018) schlagen vor den Einfluss geschützter demografischer Informationen wie Geschlecht oder Postleitzahl auf das Klassifikationsergebnis mit Adversarial Learning zu verringern. Drei verschiedene Definitionen von „equality" und Parität werden analysiert und für jeden Definition wird eine entsprechende Strategie vorgestellt um demografische Parität zu sichern.

Zusätzlicher Analyseschritt

Bolukbasi et al (2016) zeigen, dass mit Hilfe von Wortlisten ein Unterraum errechnet werden kann, der die geschlechtsbezogene Information in Word Embeddings beinhaltet. Wörter werden mit Hilfe dieser Wortlisten in geschlechtsneutral (z.B. doctor) und geschlechtsspezifisch (z.B. grandmother) eingeteilt. In dem entsprechenden Unterraum werden dann alle Wörter, die grammatikalisch geschlechtsneutral sind, auch neutralisiert, so dass beispielsweise *doctor* zentriert zwischen den Word Embeddings für "Mann" und "Frau" liegt.

Allerdings zeigen auch einige dieser Methoden Schwächen und es wurde bereits gezeigt, dass in vielen Fällen Biases weiterhin rekonstruiert werden können (Gonen & Goldberg, 2019).

Praktische Sitzung 1

Im ersten praktischen Teil sollen dann ML Modelle selbst ausprobiert und werden und, anhand von verschiedenen Analysemethoden, Biases explorativ erkundet werden.

Wir stellen eine fertige ML-Pipeline zur Textklassifizierung zur Verfügung, die mit vortrainierten Word Embeddings arbeitet. Die Klassifizierung soll dahingehend analysiert werden, welche Biases sie enthält. Dann sollen die vortrainierten Word Embeddings mithilfe von Tensorflow Projector erkundet werden und es sollen Richtungen identifiziert werden, die für die Biases in den Ergebnissen verantwortlich sein könnten. Die Teilnehmenden sollen die vortrainierten Word Embeddings auf Grundlage ihrer Erkenntnisse modifizieren und untersuchen, wie sich das Klassifikationsergebnis dadurch ändert.

Des Weiteren sollen die Biases dieser Pipeline mithilfe von standardisierten Wort-list Tests (SEAT, May et al. 2019 / WEAT, Caliskan et al. 2017) analysiert werden.

Zuletzt soll den Teilnehmenden auch die Möglichkeit gegeben werden, die Korpuszusammensetzung für das Training der Word Embeddings zu modifizieren und selber trainierte Word Embeddings anstelle der vortrainierten zu verwenden, beispielsweise mithilfe von Sampling, Vereinigung, Mitteln.

Erkenntnisgewinn für DH durch Untersuchung von Biases

Biases in historischen Textdatensätzen können auf Biases in den Gesellschaften ihrer Entstehung sowie in ihrer Aufbewahrungs- und Tradierungsgeschichte aufdecken. Mit Blick auf die wachsende Wichtigkeit von Cultural Heritage Studies in den Digital Humanities sind diese Art von Biases ein hochaktuelles Forschungsfeld (Garg et al 2018). Der Korpuskonstruktion muss in diesen Fällen allerdings besondere Sorgfalt beigemessen werden, da nur bei einem für die jeweilige Forschungsfrage möglichst ausgewogenen Korpus auch tatsächlich durch die Biases im Korpus auch auf die Biases in der Gesellschaft Rückschlüsse gezogen werden können (Underwood 2019, Bode 2020). Kurz gesagt birgt jeder Schritt in der Geschichte der zu untersuchenden Objekte die Gefahr eines unbewusst und ungewollt induzierten Bias, die der bewussten und gewollten Analyse von Biases im Wege stehen können.

Autorinnen um 1800

Digitale Methoden machen es möglich, das traditionelle Narrativ der Literaturgeschichte zu überdenken und damit Literatur in den Vordergrund zu rücken, die etwa aus Gendergründen im Kanon als zweitrangig überliefert worden war. Zumindest machen sie es theoretisch möglich: Es soll nämlich gezeigt werden, dass digitale Korpora und Methoden die Biases der traditionellen Historiographie auch im literarischen Bereich nur zu leicht reproduzieren und dass die Korpusbildung und der Trainingsprozess einer besonderen Zuspitzung brauchen, um z.B. die Rolle von schreibenden Frauen deutlich machen zu können. Argumentiert wird hier am Beispiel von Autorinnen aus der Zeit um 1800 – der Phase nämlich, wo der (wohl männliche) Autor sich als literarischer, wirtschaftlich tragfähiger Wert etabliert.

Tweetanalyse von #aufschrei und #blacklivesmatter

Auf dem sozialen Netzwerk Twitter führten die Hashtags „aufschrei“ und „blacklivesmatter“ 2013 zu kollektiven Revolten, die online begannen, sich dann aber auch auf den Alltagsdiskurs ausweiteten. Unter #aufschrei berichteten Frauen über ihre Erfahrungen mit Sexismus und unter #blacklivematters ging es um Erlebnisse mit Rassismus. An diesem Beispiel werden Methoden zur Quellenanalyse vorgestellt. Ziel ist es, die Dynamik der digitalen Bewegungen von #aufschrei und #blacklivesmatter anschaulich zu machen.

Praktische Sitzung 2 und Abschlussdiskussion

Im zweiten praktischen Teil sollen die Teilnehmenden ihre eigene Expertise einbringen und in Gruppen individuelle Fragestellungen formulieren, die mithilfe der zuvor kennengelernten Modelle untersucht werden können. Wenn möglich, sollen sofort erste Prototypen entwickelt werden.

Falls Teilnehmende keine eigenen Korpora bzw. Fragestellungen mitbringen, stellen wir eine ML-Pipeline zur Verfügung, die existierende Systeme zur Erkennung von Hatespeech im Internet auf Tweets mit dem Hashtag #aufschrei bzw. #blacklivesmatter sowie einer Kontrollgruppe aus zufälligen anderen Tweets anwendet. Mithilfe dieser Pipeline sollen Teilnehmende untersuchen, wie Sprache einer neu entstehenden Bewegung, die nicht dem Mainstream entspricht, möglicherweise automatisch als Hatespeech erkannt wird.

Fußnoten

1. Was tatsächlich bei Amazon passiert ist: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
2. Beispiel von https://twitter.com/jessamyn/status/900867154412699649 bezüglich des www.perspectiveapi.com Interfaces, außerdem Davidson et al. (2019)

Bibliographie

Bode, Katherine (Forthcoming 2020): Why You Can’t Model Away Bias, Modern Language Quarterly 81.1. preprint: katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf [letzter Zugriff 27. September 2019].

Bolukbasi, Tolga / Kai-Wei Chang / James Y Zou / Venkatesh Saligrama / Adam T Kalai (2016): Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Conference of NIPS.

Buolamwini, Joy / Timnit Gebru (2018): Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency.

Caliskan, Aylin / Joanna J. Bryson / Arvind Narayanan. (2017): Semantics derived automatically from language corpora contain human-like biases. Science 356.

Davidson, Thomas / Debasmita Bhattacharya / Ingmar Weber (2019): Racial Bias in Hate Speech and Abusive Language Detection Datasets. arXiv preprint arXiv:1905.12516.

Garg, Nikhil / Londa Schiebinger / Dan Jurafsky / James Zou (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences.

Gonen, H. / Yoav Goldberg (2019): Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. Conference of the NAACL.

May, Chandler / Alex Wang / Shikha Bordia / Samuel R. Bowman / Rachel Rudinger (2019): On Measuring Social Biases in Sentence Encoders. Conference of the NAACL.

Mikolov, T. / Chen, K. / Corrado, G. / Dean, J. (2013): Efficient estimation of word representations in vector space. In ICLR.

Sap, Maarten / Dallas Card / Saadia Gabriel / Yejin Choi / Noah A. Smith (2019): The Risk of Racial Bias in Hate Speech Detection. Conference of the ACL.

Swinger, Nathaniel / Maria De-Arteaga / Neil Heffernan IV / Mark Leiserson / Adam Kalai (2019): What are the biases in my word embedding?. Conference on Artificial Intelligence, Ethics, and Society (AIES).

Underwood, Ted (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Zhang, B. H. / Lemoine, B. / Mitchell, M. (2018): Mitigating unwanted biases with adversarial learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Zhao, J. / Wang, T. / Yatskar, M. / Ordonez, V. / Chang, K. W. (2018): Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876.

Deep Learning für visuelle Medien: Annotation, Training, Analyse

Howanitz, Gernot

gernot.howanitz@uni-passau.de
Universität Passau, Deutschland

Radisch, Erik

e.radisch@gmx.at
Sächsische Akademie der Wissenschaften zu Leipzig

Workshop-Konzept

Zahlreiche Projekte der eHumanities fokussieren auf die Verarbeitung von Information, die in Textform codiert sind. Andere Modalitäten der Informationsübermittlung und -übertragung, wie beispielsweise visuelle Medien, bleiben in den DH häufig außen vor. Ein Grund dafür ist das breite Spektrum an etablierten Verfahren, das für solche Fragestellungen zur Verfügung steht. Ist man bei der Analyse von Texten in der Zwischenzeit so weit, sich den Inhalten und Kontexten durch automatisierte computergestützte Analyseverfahren zu nähern, verweilt man bei anderen Modalitäten wie Bildern allzu oft auf einer Ebene, wo verglichen mit der Textanalyse eher Buchstaben gezählt werden. Aus der Perspektive der Kulturwissenschaften ergibt sich hier ein *desideratum*; schließlich widmen sich diese der (menschlichen) Kultur in ihrer ganzen Bandbreite und decken kulturelle Äußerungen im weitesten Sinne ab, die unterschiedlichste Modalitäten, wie beispielsweise physische Artefakte und performative Handlungen, mit einschließen. Zwar ist es eingeschränkt möglich, kulturelle Phänomene zu transkribieren, also in textuelle Form zu bringen, was aber kaum automatisierbar ist und Informationsverluste birgt. Native Ansätze, welche auf jeweils spezifische Eigenschaften der zu untersuchenden Modalität eingehen, erscheinen deshalb vielversprechend. Neueste Ansätze der Computer Vision bieten hier ein großes Potential, denn sie ermöglichen es, sich auch Inhalten jenseits des Textes automatisiert zu nähern, was Möglichkeiten für computergestützte multimodale Analysen eröffnet.

In jüngster Zeit halten nun Methoden der Computer Vision langsam Einzug in die Digital Humanities (Tilton/Arnold 2018). Dabei wird allerdings das volle Potential des Deep Learning nicht ausgeschöpft. Zwar finden Neuronale Netze zur Bilderkennung Anwendung in den Digital Humanities, diese beschränken sich aber oft auf die Nutzung vortrainierter Netze. Hier ergeben sich potentielle Probleme, werden diese Netze doch in der Regel auf einige wenige etablierte Bildkorpora wie etwa Microsofts COCO-Dataset (Common Objects in Context, http://cocodataset.org) trainiert. Dabei handelt es sich um Bildmaterial, das vorwiegend aus dem Nordamerika des 21. Jahrhunderts stammt. Für viele kulturwissenschaftliche Fragestellungen, die auf andere Zeiträume und/oder andere Kulturkreise abzielen, ergibt sich daraus ein Bias, der die Ergebnisse verfälschen kann. In solchen Fällen ist dann ein selbst durchgeführtes, zielgerichtetes Training notwendig, das auf die spezifische Fragestellung abgestimmt ist. Ein solches Training neuronaler Netze ist jedoch keineswegs trivial, sondern erfordert eine Menge Vorarbeit, Wissen um grundlegende Trainingsstrategien und vor allem auch Erfahrung im Tweaken der Parameter.

Der Workshop soll grundlegende Kenntnisse zur Anwendung von State-of-the-Art-Algorithmen der Computer Vision in den Digital Humanities vermitteln. Er baut auf den Erfahrungen auf, die die beiden Workshopleiter im Rahmen ihrer Tätigkeit am Passau Center for eHumanities (PACE) sammeln konnten. Die Ergebnisse wurden auch in den letzten zwei DHd-Konferenzen in Vorträgen vorgestellt (Bermeitinger et al. 2017; Decker et al. 2018). In der dreijährigen Projektlaufzeit wurde ein reicher Schatz an Erfahrungen im Umgang mit Neuronalen Netzen gesammelt und eine Reihe von einfachen und klar strukturierten Workflows entwickelt, die nun mit einer interessierten Öffentlichkeit geteilt werden sollen. Eine Hoffnung ist, dass der Workshop Anstoß für Projekte gibt, die visuelle Medien quantitativ erfassen wollen und gleichzeitig die vorgestellten Methoden einer kritischen Evaluation unterziehen und weiterentwickeln.

Der Workshop *Deep Learning für visuelle Medien* intendiert in mehrere der in PACE erarbeiteten Workflows einzuführen , die es erlauben, Neuronale Netze für visuelle Medien auf bestimmte Fragestellungen der Geisteswissenschaften anzuwenden, für eigene Fragestellungen zu adaptieren bzw. zu trainieren und die Ergebnisse zu analysieren. Damit soll die Grundlage gelegt werden, Forschern selbst das Training und die Anwendung Neuronaler Netze sowie die Analyse deren Ergebnisse zu ermöglichen. Im Zentrum des Workshops stehen drei Neuronale Netze, die über verschiedene Features verfügen.

Das von Facebook Artificial Intelligence Research Group (FAIR) entwickelte Framework *Detectron* (Girshick et al. 2018) kombiniert verschiedene neuronale Netze und ermög-

licht ein breites Nutzungsspektrum. Dieses leistungsstarke Framework erlaubt nicht nur das Trainieren der Objekterkennung, sondern kann ebenfalls eine Reihe wichtiger Keypoints des menschlichen Körpers (z.B.: Kopf, Schultern, Ellenbogen, Knie, usw.) erkennen, die wiederum wichtige Rückschlüsse auf die Haltung der Personen zulassen. OpenPose (Zhe et al. 2017), das ebenfalls im Rahmen des Workshops vorgestellt wird, befasst sich ebenfalls mit diesen Keypoints. Im Gegensatz zu Detectron kann OpenPose auch einzelne Finger erkennen. Anders ausgedrückt liefert dieses Netz deutlich mehr Informationen zurück. Das dritte Neuronale Netz, auf das eingegangen werden wird, ist OpenFace von Tadas Baltrusaitiš (Baltrusaitiš et al. 2018). Dieses mächtige Neuronale Netz kann nicht nur Gesichter erkennen, sondern auch deren dreidimensionale Ausrichtung errechnen und eine ganze Reihe von Keypoints im Gesicht erkennen. Diese Keypoints lassen ebenfalls Rückschlüsse auf sogenannte Facial Expression Units (Decker et al. 2019), welche genutzt werden können, um Aussagen über die Emotionen machen zu können, die eine Person zeigt.

Im Rahmen des Workshops werden sowohl Installation als auch Setup und erste Schritte mit diesen Frameworks thematisiert. Darüber hinaus geht es im Rahmen dieses Workshops auch darum, Netze zielgerichtet für die eigene Fragestellung zu trainieren. Wie ein Netz erfolgreich mit verhältnismäßig kleinen Korpora trainiert werden kann, wird im Kurs vermittelt werden. Auch die Evaluierung von Trainingsergebnissen wird diskutiert. In einem letzten Schritt soll auch in die Arbeit mit den extrahierten Features eingeführt und verschiedene Analysemöglichkeiten vermittelt werden.

Programm

Vor dem Workshop:
Vernetzung der Teilnehmerinnen und Teilnehmer über Github (https://github.com/passau-centre-for-ehumanities/visual_media), Identifizierung von gemeinsamen Forschungsinteressen, Gruppenbildung, falls notwending Hilfestellung zur Installation grundlegender Software (Jupyter Notebooks), um beim Workshop selbst möglichst wenig Zeit zu verlieren.

9:00-9:30 *Kick-Off und Kennenlernrunde, Abfragen der Erwartungen*

9:30-10:30 *Allgemeine Einführung in Deep Learning*
Zum Auftakt des Workshops wird eine allgemeine kurz gehaltene Einführung zu künstlichen neuronalen Netzen und Deep Learning-Algorithmen im Allgemeinen gegeben, um ein Verständnis der Funktionsweise von Detectron zu entwickeln.

10:30-11:00 *Kaffeepause*

11:00-12:30 *Einführung in Detectron und OpenPose, Praktische Erfahrungen*
Im Anschluss daran wird der generelle Aufbau von Detectron in die Funktionsweise der wichtigsten Bestandteile des Frameworks vorgestellt. Neben dem Trainingsaufbau wird hier aufgezeigt, wie man Standardmodelle lädt und auf visuelle Medien anwenden kann. Insbesondere geht es darum, Detectron und OpenPose zu nutzen, um Personen und deren Haltung in Bildern erkennen (Abb. 1)

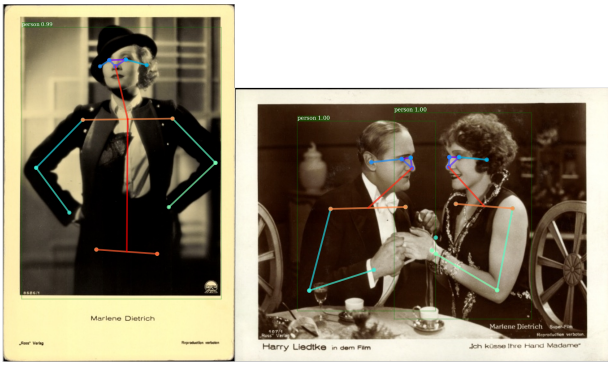


Abbildung 1: Posenerkennung mit Detectron

Als drittes Standbein soll im Kurs des Weiteren in die Anwendung von OpenFace eingeführt werden, mit dessen Hilfe es möglich ist, Keypoints von Gesichtern auszulesen, die dafür verwendet werden, um sogenannte Action Units abschätzen zu können. Action Units sind Basisbestandteile von menschlichen Emotionsausdrücken (Abb. 2)

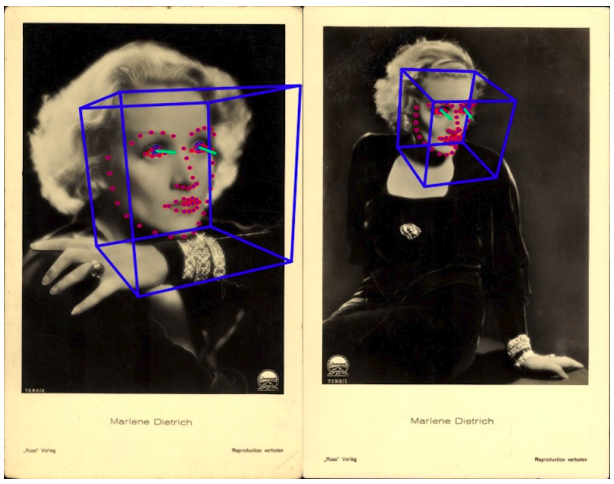


Abbildung 2: OpenFace blickt Marlene Dietrich ins Gesicht: Lage im Raum (blau), Keypoints (rot), berechnete Blickrichtung (grün)

12:30-14:00 *Mittagspause*

14:00-16:00 *Detectron Trainieren*
Ein wichtiger Bestandteil des Workshops wird es sein, in ein im Rahmen des Passau Center for eHumanities entwickelten Workflow zum Trainieren von Detectron einzuführen.

Es wird konkret an einfachen Beispielen vermittelt, wie man die einzelnen Bestandteile des Workflows installieren und auf die individuelle Forschungsfrage hin anwenden kann. Es wird neben der Vermittlung des Workflows ebenfalls großen Wert darauf gelegt, den Kursteilnehmern zu vermitteln, welche typischen Fehler beim Trainingsaufbau zu vermeiden sind. Die Teilnehmer sollen am Ende des Workshops dazu in der Lage sein:

- selbstständig Trainingskorpora für Detectron erzeugen zu können
- ein grundlegendes Verständnis dafür entwickelt haben, auf welche Parameter zu achten sind, um auch mit kleinen Bildkorpora trainieren zu können

- den Trainingsprozess mit den eigenen Daten initiiren zu können

Abbildung 3: Beispiel der Annotation. Als Werkzeug wird Labelme genutzt (https://github.com/wkentaro/labelme)

Abbildung 4: Beispiel der Anwendung von Detectron, trainiert auf ukrainische Symbole

16:00-16:30 *Kaffeepause*

16:30-18:00 *Vorstellung von Analysetechniken für die produzierten Ergebnisse*

Im letzten Teil des Kurses werden Analysetechniken vorgestellt, die es ermöglichen, die Produzierten daten nach interessanten Mustern zu explorieren bzw. deren Inhalte zu analysieren (Abb. 5, 6).

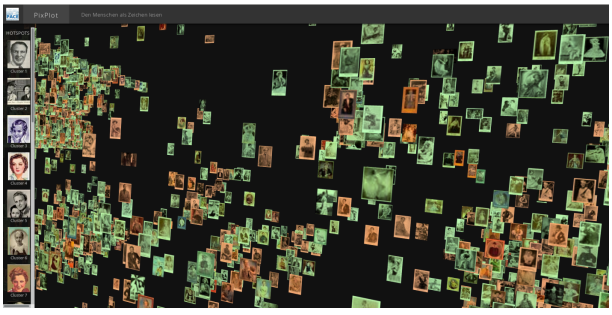


Abbildung 5: Beispiel einer skalierbaren, dreidimensionalen Visualisierung eines Clusterings mittels einer modifizierten Version von Pixplot. Die Bilder werden anhand von Metadateninformationen zusätzlich eingefärbt.

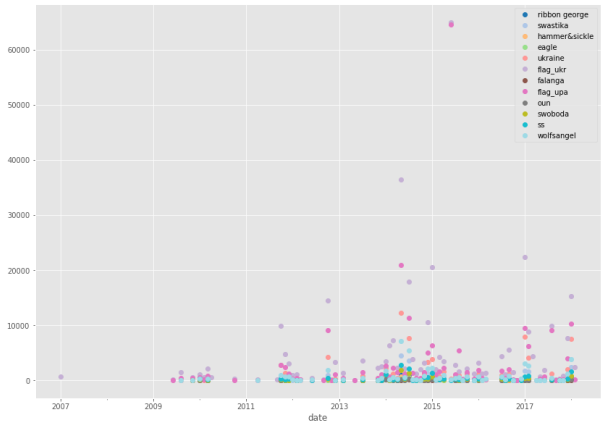


Abbildung 6: Beispiel für eine Analyse der Ergebnisse. Hier konkret: die Verteilung von Symbolen in Youtube-Videos über die Zeit.

18:00-18:30 *Schlussrunde, Workshop-Evaluation*

Dieses Programm ist hoffentlich geeignet, als Inspiration und erste Einführung in das Thema Deep Learning für visuelle Medien zu dienen. Eine Nachbereitung und weitere Vernetzung über Github ist ausdrücklich erwünscht, um eine weitere Begleitung der Projekte zu garantieren.

Zusätzliche Angaben

- Benötigte technische Ausstattung: Beamer, WLAN-Zugang, ausreichend Steckdosen für die Laptops der Teilnehmerinnen und Teilnehmer
- Zahl der möglichen Teilnehmer: 20

Bibliographie

Bermeitinger, B. / Howanitz, G. / Radisch, E. (2018): Contextualizing Bandera: Ein Distant Watching Ansatz, in Kritik der digitalen Vernunft Konferenzabstracts. Köln, 26.02-02.03.2018. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [abgerufen am 01.05.2018].

Baltrusaitiš, T. / Zadeh, A. / Chong Lim, Y., Morency, L.-P. (2018): "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on. IEEE, 2018, S. 59-66.

Decker, J.-O. / Howanitz, G. / Radisch, E. / Rehbein, M. (2019): Den Menschen als Zeichen lesen. Quantitative Lesarten körperlicher Zeichenhaftigkeit in visuellen Medien. In: Dhd 2019, Digital Humanities: multimodal & multimedial, Konferenzabstracts, Frankfurt am Main 2019, S. 106-109. https://zenodo.org/record/2596095#.XLBUNUNS-V4

Girshick, R. / Radosavovic, I. / Gkioxari, G. / Dollár, P. / He, K. (2018): Detectron. https://github.com/facebookresearch/detectron. [Letzter Zugriff 25. 09. 2018]

Zhe C. / Tomas S. / Shih-En W. / Yaser Sh. (2017): Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR.

Digital Humanities from Scratch

Roeder, Torsten
torsten.roeder@leopoldina.org
Leopoldina, Halle (Saale)

Cremer, Fabian
cremer@maxweberstiftung.de
Max Weber Stiftung, Bonn

Dogunke, Swantje
swantje.dogunke@htwk-leipzig.de
HTWK Leipzig

Elwert, Frederik
frederik.elwert@rub.de
Ruhr-Universität Bochum

Lordick, Harald
lor@steinheim-institut.org
Steinheim-Institut

Ott, Katrin
katrin.ott@uni-erfurt.de
Universität Erfurt

Söring, Sibylle
ssoering@cedis.fu-berlin.de
Freie Universität Berlin

Wübbena, Thorsten
wuebbena@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte, Mainz

Thema

Thema des Workshops „Digital Humanities from Scratch" sind Koordinationsaufgaben im Bereich von DH-Aktivitäten, die an wissenschaftlichen Institutionen aller Größenordnungen stetig intensiver und bedeutender zum Tragen kommen. Der Workshop bietet dafür ein offenes und moderiertes Forum, das mit Impulsbeiträgen und Diskussionsrunden von der DH-Community ausgefüllt wird.

Während einige Einrichtungen auf jahrzehntelange Erfahrungen zurückgreifen können, beginnen andere erst heute und nur langsam damit, die oft zahlreichen DH-Aktivitäten am eigenen Haus zu koordinieren. Solche praxisorientierten Organisations- und Koordinationsaufgaben sind aber nur selten Teil der wissenschaftlichen DH-Ausbildung (Cremer 2019) und reichen weit in angrenzende Gebiete – z.B. Informationstechnologie, Wissenschaftsmanagement, Forschungsdatenmanagement – hinein. Auch die Entwicklung und Umsetzung eines erfolgreichen DH-Gesamtkonzepts, das solide in institutionelle, lokale und regionale Kontexte eingebettet ist, stellt oft eine organisatorische und politische Herausforderung dar.

Angesichts der voranschreitenden Digitalisierung aller Wissenschaftszweige ist DH-Koordination als grundlegende Aufgabe aller Wissenschaftseinrichtungen anzusehen. Hinzu kommt die allgemeine Forderung nach Interdisziplinarität und Methodenvielfalt, die durch Vernetzung verschiedener Fachbereiche, in den Geisteswissenschaften aber speziell durch die Öffnung für digitale Arbeits-, Forschungs- und Publikationsverfahren erwidert werden kann. Allerdings kann die Vielfalt an den Möglichkeiten, die sich durch Digitalisierung und speziell im Feld der DH bieten, sowie die Komplexität der Anforderungen, die dies an den Wissenschaftsbetrieb stellt, für Forschende und Institutionen eine Überforderung bedeuten und Unsicherheit erzeugen. Ein zentrales Anliegen der DH-Koordination liegt deshalb darin, Prozesse des Umdenkens und Neugestaltens zu begleiten, (Denk-)Räume zu schaffen, zu öffnen und darin Handlungsoptionen zu vermitteln („Change Management").

Versteht man die DH als Teil eines umfassenden Digitalisierungsprozesses, sind sie als Vielzahl von Schnittstellen zwischen Forschenden, Lehrenden, Studierenden, Bibliothek, IT, Administration, Leitung, Forschungsförderung etc. anzusehen. Diese Schnittstellen können, wenn nicht als (Kompetenz-)zentren, in Form von Arbeitsgruppen, Schulungen, Veranstaltungen etc. ausgestaltet werden, benötigen aber Organisation und Moderation. Dies bringt aktuell mehrere Problemstellungen mit sich. Zuvorderst konfligieren hier typische DH-Stellenprofile (mit vornehmlich informationstechnischen Qualifikationen) mit dem tatsächlichen Aufgabenfeld der DH-Koordination: Kommunikation, Vermittlung, Projektmanagement, Outreach, vielfach auch institutionelle Strategie (Wuttke 2019). Desweiteren sind DH-Koordinationsaufgaben an Universitäten oft fächer- oder fakultätsübergreifend angelegt, Forschungsverbünde schaffen Querschnittsstellen, und Akademien pflegen zum Teil eigene Referate; auch die Verortung der Stellen (Bibliothek, Administration, Forschungsabteilung) wird sehr unterschiedlich gehandhabt.

Ein Diskurs über die verschiedenen Profile, Konfigurationen und Aufgabenfelder der DH-Koordination sowie ein Austausch über konkrete Erfahrungen und Strategien kann nicht nur den Koordinator*innen helfen, sondern vor allem auch das Profil der DH an wissenschaftlichen Institutionen insgesamt schärfen, Herausforderungen benennen und Lösungsmodelle bündeln.

Bisherige Aktivitäten

Die Herausforderungen auf institutioneller, organisatorischer, disziplinärer, personeller und technischer Ebene wurden bereits in dem vielbeachteten Panel „Digital Humanities from Scratch" auf der DHd 2019 diskutiert (Roeder et al. 2019a und 2019b). Ein deutlicher Bedarf an Austausch, Vernetzung und Bündelung von Initiativen und Aktivitäten offenbarte sich sowohl bei quereinsteigenden DH-Koordinator*innen als auch beim wissenschaftlichen Nachwuchs aus DH-Studiengängen, der zukünftig mit Aufgaben der DH-Koordination konfrontiert werden wird. Auch im internationalen Kontext ist die Relevanz der Thematik erkannt worden. So wurde 2017 die Digital Scholarship Working Group im Kontext

der Digital Library Federation gegründet (DLF 2019). Ebenso wurde das Thema in den Workshops „Getting Things Done“ und „Libraries as Research Partners in Digital Humanities“ auf der DH 2019 in Utrecht behandelt (Keegan et al. 2019 sowie Wilms et al. 2019).

Ziele und Ergebnisverwertung

Der Workshop möchte insbesondere einen Beitrag dazu leisten, dem aktuellen Austauschbedarf eine offene Plattform zu bieten, DH-Koordination als Aufgabenfeld Feld in den DH zu thematisieren sowie das Netzwerk von DH-Koordinator*innen langfristig zu stärken und auszubauen. Die bisherigen, regen Diskussionen werden fortgesetzt, indem gezielt Kontroversen und Interessenschwerpunkte aufgegriffen werden. Sowohl gegenüber den jeweiligen Institutionen als auch im Kontext von Forschungsverbünden, fachspezifischen und übergeordneten Verbänden und Gremien kann dadurch den Aufgabenfeldern der DH-Koordination eine deutlich verbesserte Sichtbarkeit und Stimme verliehen werden, die bislang fehlt.

Der Workshop dient darüber hinaus zur Abstimmung zukünftiger Aktivitäten, beispielsweise:

- Kartierung der DH-Koordinations-Stellen (Ausprägungen, Stellenprofile, Standorte)
- Vergleich der Ausschreibungen mit tatsächlichen Tätigkeiten
- Checkliste für DH-Koordinationsaufgaben (Handlungsempfehlungen, Erfahrungsaustausch)
- Bildung einer DHd-Arbeitsgruppe „DH-Koordination“ zur Produktivierung des Netzwerks (Abstimmung lokaler Schwerpunkte, Ausarbeitung von Working Papers)

Um die Ergebnisse für die Community öffentlich sichtbar und verfügbar zu halten, werden diese anschließend (in einer noch zu bestimmenden Form) publiziert.

Organisation

Veranstaltungsformat

Der auf 3,5 Stunden angesetzte Workshop ist in zwei Hauptteile gegliedert, um zweierlei Bedürfnisse abzudecken: Erstens Themenkomplexe zu definieren und Fragestellungen aufzuwerfen, und zweitens den Raum für Diskussionen sowohl in kleinen Kreisen als auch in großer Runde zu öffnen. Zu dem Workshop werden maximal 40 Personen zugelassen. Aktive Teilnahme ist ausdrücklich erwünscht.

Der erste Hauptteil wird durch freie Beiträge bestritten, die über einen Call eingeworben werden (s.u.). Das gewünschte Vortragsformat ist Pecha Kucha (sprich: pe'tscha-k-tscha). Dabei handelt es sich um ein alternatives Präsentationsformat, das den Vortrag durch ausgewählte Bilder anstelle von textlastigen Folien begleitet (PechaKucha 2019). Das Format sieht vor, dass während des Vortrags 20 Bilder für jeweils 20 Sekunden eingeblendet werden. Die Gesamtdauer ist dadurch auf 6:40 min definiert. Die Verbindung von Vortrag und Bildern kann frei gestaltet werden.

Der zweite Hauptteil wird als World Café mit parallelen Thementischen gestaltet (The World Cafe 2019). In moderierten Gruppen wird über vorher umrissene Themengebiete (s.u.) diskutiert und gearbeitet. Zur Unterstützung und Dokumentation der Diskussion werden Pinnwände und Schreibmaterial (z.B. für „graphic recording“) zur Verfügung gestellt.

Im abschließenden Wrap-Up werden mögliche weiterführende Aktionen eruiert und die Ergebnisverwertung beschlossen.

Der Workshop versteht sich als offene und inklusive Initiative, die dediziert „Collegiality and Connectedness“ in den DH (Spiro 2012: 26–28) befördern und mit kreativen Formaten das Konferenzmotto „Spielräume“ adressieren möchte.

Zeitplan

Nach Begrüßung und organisatorischen Hinweisen (15 Minuten) werden im ersten Hauptteil acht Pecha Kuchas präsentiert. Einschließlich der Zeit für Kurzvorstellung und Übergang (2–3 Minuten) sind dafür 75 Minuten zu veranschlagen. Nach einer 20-minütigen Pause werden im zweiten Hauptteil drei World Cafés mit jeweils fünf parallelen Thementischen angesetzt. Nach jeweils 20 Minuten können die Teilnehmer*innen an andere Thementische wechseln. Einschließlich des Wechsels (3–4 Minuten) sind dafür 70 Minuten zu veranschlagen. Das Workshop-Team übernimmt die Moderation sowie das Wrap-Up (30 Minuten).

Begrüßung und Organisatorisches	15 Minuten
Hauptteil 1 Pecha Kucha (8x 6:40 Minuten)	75 Minuten
Pause	20 Minuten
Hauptteil 2 World Café (3x 20:00 Minuten)	70 Minuten
Wrap-Up	30 Minuten
Gesamtdauer	210 Minuten

Call for Pecha Kuchas

Digital Humanities from Scratch: Wie geht man Koordinationsaufgaben im Bereich der DH an, wenn sich diese in einem frühen Entwicklungsstadium befinden? Die Anforderungen sind vielfältig und gehen oft über reines Expert*innenwissen (sei es im Programmierbereich oder in einem geisteswissenschaftlichen Fach) hinaus, denn DH ist nicht nur interdisziplinär, sondern erstreckt sich auch auf organisatorische, institutionelle und soziale Handlungsfelder. Der Workshop „Digital Humanities from Scratch“ auf der DHd 2020 „Spielräume“ (2. bis 6. März 2020) vertieft diese Thematik in kreativen Vortrags- und Diskussionsformaten.

Wir laden hiermit dazu ein, Kurzvorträge einzureichen, die Erfahrungen typischer Herausforderungen oder Lösungsansätze im Bereich der DH-Koordination vorstellen. Die Vorträge sollen dem Format „Pecha Kucha“ folgen: Die Dauer des mündlichen Vortrags beträgt exakt 6:40 Minuten. Für die Präsentation sind 20 Bilder auszuwählen (möglichst ohne Text), die während des Vortrags für jeweils 20 Sekunden eingespielt werden. Die Pecha Kuchas geben Impulse für ein anschließendes World Café, wo in kleinen Gruppen intensiv diskutiert werden kann.

World Café Thementische

- **DH und GLAM** (Galleries, Libraries, Archives, Museums). DH-Koordination wird häufig in oder nahe an Bibliotheken angelegt. Wie verhalten sich die Aufgabengebiete der DH-Koordination zu bibliothekarischen Anforderungen? Wo kann dies helfen? Und wie ließe sich das Verhältnis zu anderen Kultur- und Gedächtnisorganisationen gestalten? Hier spielen Vermittlungsaufgaben auf der einen Seite und konservatorische Aspekte auf der anderen Seite hinein.
- **DH und IT.** Die Informationstechnologie ist als Rückgrat der DH nicht wegzudenken. Dennoch treten immer wieder Abgrenzungstendenzen auf. In der Tat will die fließende Grenze zwischen „harter IT“ und „softer IT“ praktikabel organisiert werden. Wie kommt man an Hosting und Webdesign? Wie lässt sich die Entwicklung von Forschungssoftware organisieren? Welche Risiken sind mit der Beauftragung externer Dienstleister verbunden? Und müssen DH-Koordinator*innen selbst mit anpacken?
- **DH und Wissenschaftsmanagement.** Bei der Drittmitteleinwerbung stehen DH-Koordinator*innen oft vor der Aufgabe, digitale Methoden und Arbeitsabläufe in Stellen-, Kosten- und Arbeitsplänen einzuflechten. Auch während der Durchführungsphase suchen Projekte immer wieder Beratung und Unterstützung in DH-Fragen. Aber nicht immer: Wie ist mit Projekten umzugehen, die den DH ablehnend gegenüber stehen oder schlicht noch keine Kompetenz aufgebaut haben?
- **DH als Vermittlung.** Den digitalen Wandel in der Wissenschaft voranzutreiben ist nicht zuletzt eine psychologische Aufgabe. Wie lassen sich Räume schaffen und Brücken bauen für grundlegende strukturelle Veränderungen? Erfolg versprechen hier informelle Formate, beispielsweise Stammtische, Coffee Lectures, Barcamps. Wie aber lässt sich nachhaltig Kompetenz aufbauen sowie geeignetes Personal akquirieren und entwickeln, so dass DH zu einer Gemeinschaftsaufgabe wird?
- **DH-Institutionalisierung.** Jede Institution definiert selbst, ob und wie sie DH in ihre Organisationsstruktur integriert. Wird dies mit Koordinationsstellen oder mit wissenschaftlichen IT-Stellen ausgefüllt? Liegen die Aufgaben schwerpunktmäßig im Schulungs- und Vermittlungsbereich, im Projektmanagement oder im technischen Feld? Ist DH einer bestehenden Abteilung zugeordnet, als eigene Abteilung organisiert oder wird es als Querschnittsaufgabe implementiert? Welche Möglichkeiten gibt es, bestehende Strukturen weiterzuentwickeln?

Bibliographie

Cremer, Fabian (2019): „Gottes Werk und Teufels Beitrag: Ein Essay zu Digital Humanities und Projektmanagement“ (Blogpost). Website: *DHd-Blog*, 19. März 2019. https://dhd-blog.org/?p=11283 (archiviert am 13.09.2019).

DLF (2019): „Digital Scholarship Working Group“. Website: *Digital Library Federation Wiki*, 10. Januar 2019. https://wiki.diglib.org/Digital_Scholarship_Working_Group (archiviert am 25.09.2019).

Keegan, Tom / Gehlsen Morlan, Leah / Leonard, Peter / DeRose, Catherine (2019): „Getting Things Done: Administrative Tips, Tricks, Helps, And Hindrances In Digital Scholarship“ (Workshop). *Digital Humanities Conference*, Utrecht. URL: https://dev.clariah.nl/files/dh2019/boa/1107.html (archiviert am 13.09.2019).

PechaKucha (2019) ohne Autor: „Frequently Asked Questions“. Website: *PechaKucha. 20 Images x 20 Seconds*, ohne Datum. URL: https://www.pechakucha.com/faq (archiviert am 24.09.2019).

Roeder, Torsten / Söring, Sibylle / Dogunke, Swantje / Elwert, Frederik / Wübbena, Thorsten / Lordick, Harald / Cremer, Fabian / Klammt, Anne (2019a): „Digital Humanities ‚from Scratch‘. Herausforderungen der DH-Koordination zwischen Querschnittsaufgaben und ‚one-(wo)man-show‘“ (Panel). In: *DHd 2019, Digital Humanities: multimedial & multimodal. Konferenzabstracts*, S. 68–72, DOI: 10.5281/zenodo.2596094 (21.03.2019), als Auszug: 10.5281/zenodo.3244179 (12.06.2019).

Roeder, Torsten / Söring, Sibylle / Dogunke, Swantje / Elwert, Frederik / Wübbena, Thorsten / Lordick, Harald / Cremer, Fabian / Klammt, Anne (2019b): „Digital Humanities ‚from Scratch‘. Ein Panel-Bericht zur DHd 2019“ (Blogpost). Website: *DHd-Blog*, 3. Juli 2019. URL: https://dhd-blog.org/?p=11804 (archiviert am 13.09.2019).

Spiro, Lisa (2012): „‚This Is Why We Fight‘: Defining the Values of the Digital Humanities“. In: Matthew K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis/London: University of Minnesota Press. S. 16–35.

The World Cafe (2019) ohne Autor: „World Cafe Method“. Website: *The World Cafe*, 2019. URL: http://www.theworldcafe.com/key-concepts-resources/world-cafe-method/ (archiviert am 24.09.2019).

Wilms, Lotte / Klaassen, Martine / Claeyssens, Steven / Lefferts, Marian (2019): „Libraries as Research Partners in Digital Humanities“ (Preconference Workshop). *Digital Humanities Conference*, The Hague: National Library of the Netherlands, 8. Juli 2019. URL: https://adholibdh.github.io/dh2019-preconference/assets/pdfs/LibDH2019_Programme.pdf

Wuttke, Ulrike (2019): „Infrastrukturelle Erfolgsfaktoren für einen Digital Humanities-Schwerpunkt an deutschen Universitäten“ (Masterarbeit TH Köln, 28.06.2019). URL: https://publiscologne.th-koeln.de/frontdoor/index/index/docId/1396

Einführung in TEI-ODD

Stadler, Peter

stadler@weber-gesamtausgabe.de
Universität Paderborn, Deutschland

Bohl, Benjamin W.

bohl@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Viglianti, Raffaele

rviglian@umd.edu
MITH, University of Maryland, College Park, MD, USA

ODD (One Document does it all) ist eine Metasprache, entwickelt im Kontext der Text Encoding Initiative (TEI), zur (formalen) Beschreibung und Dokumentation von XML Schemata. ODD bildet die Grundlage (d.h. den Quellcode) der Richtlinien und Schemata sowohl der Text Encoding Initiative (TEI) als

auch der Music Encoding Initiative (MEI). Aber ODD ist nicht auf diese bestehenden Codierungsrichtlinien beschränkt; so lässt es sich auch zur Beschreibung anderer bestehender XML-Dialekte benutzen (beispielsweise HTML [4]), oder zur Entwicklung eigenständiger Codierungsrichtlinien in ganz anderen Kontexten, wie etwa im Falle von Music Performance Markup (MPM) [1].

Das grundlegende Designprinzip von ODD folgt einem Literate Programming Ansatz, d.h. in einem ODD-Dokument sind sowohl die Code-Bestandteile zur Beschreibung der Grammatik eines Schemas als auch die menschenlesbare Beschreibung – und Exemplifizierung – dieser Regeln miteinander verwoben [6]. Im Kontext der TEI und MEI Communities wird dies insbesondere von digitalen Editionen genutzt. Der besondere Anreiz liegt für diese Unternehmungen hierbei darin, dass die Dokumentation (im ODD-Format) der jeweils spezifischen Nutzung der TEI bzw. der MEI-Richtlinien gewissermaßen das digitale Pendant zu herkömmlichen Editionsrichtlinien darstellt.

Aus „Endnutzersicht" wird ODD meist zur Maßschneiderung von TEI oder MEI Schemata genutzt. Die aktive Weiterentwicklung von ODD hat jedoch interessante neue Möglichkeiten eröffnet; so ist es neuerdings mittels so genanntem ODD-Chaining [2] auch möglich geworden, eigene Editionsrichtlinien nicht unmittelbar von den TEI- oder MEI-Richtlinien abzuleiten, sondern vermittelt von anderen ODD-Anpassungen. Dies wird beispielsweise im Deutschen Textarchiv genutzt, um die Richtlinien zur Auszeichnung von Drucken bzw. von Manuskripten von einem gemeinsamen DTA Basisformat abzuleiten [3].

Solchen Anwendungsfällen von ODD werden jedoch meist nur wenige „Eingeweihte" gewahr, da die häufig eine Verständnisbarriere für die Mechanismen von ODD vorliegt. Diesem soll der vorliegende Workshop entgegenwirken. Deshalb ist der ganztägige Workshop speziell auf die Bedürfnisse von Einsteigern ausgerichtet. Es werden sowohl die notwendigen theoretischen Hintergründe vermittelt, als auch praktische Hilfestellungen zum Erstellen eines ersten eigenen Schemas bzw. dessen Dokumentation vermittelt. Hierfür wird im ersten Teil des Workshops zunächst als niederschwelliger Einstieg der von Raffaele Viglianti neu entwickelte webbasierte ODD-Editor „Roma" [5] vorgestellt und damit die Erstellung einfacher Anpassungen des TEI-Schemas verwendet; dazu gehören etwa Operationen wie das Hinzufügen oder Entfernen von ganzen teilbereichen des Schemas (Modulen), von einzelnen Elementen oder Attributen, sowie das Einschränken von Attributwerten auf geschlossene Listen.

Im zweiten Teil des Workshops sollen dann durch direktes Bearbeiten des ODD-Quelldokuments erweiterte Funktionen wie Modularisierung, ODD-Chaining oder das Generieren von Schemata mit mehreren Namespaces erläutert werden.

Der Workshop ist als Hands-On-Session konzipiert, in der die Teilnehmer an Ihrem eigenen Laptop direkt Erfahrungen sammeln sollen. Das Tutoren-Team steht ihnen dabei stets mit Rat und Tat zur Seite.

Formalia

- Maximale Zahl der möglichen Teilnehmerinnen und Teilnehmer: 25
- benötigte technische Ausstattung: Beamer; die Teilnehmer*innen sollen eigene Laptops mit vorinstalliertem oXygen-XML-Editor mitbringen. Eine oXygen-Lizenz zur Nutzung im Workshop wird gestellt

Beiträger

Peter Stadler

Wissenschaftlicher Mitarbeiter an der Carl-Maria-von-Weber-Gesamtausgabe an der Universität Paderborn und Mitglied des TEI Councils. Zu seinen Forschungsgebieten zählen digitale Musik- und Briefeditionen.

Raffaele Viglianti

Research Programmer am Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland und Mitglied des TEI Councils. Seine Forschung kreist um Textwissenschaft und digitale Editionen mit einem Fokus auf Musik.

Benjamin W. Bohl

Research Software Engineer der Bernd Alois Zimmermann-Gesamtausgabe und in seiner Mitgliedschaft im MEI Board als Co-chair des MEI Technical Teams eingesetzt. In seiner Forschung befasst er sich mit Datenmodellierung und -haltung in digitalen Editionsprojekten mit plurimedialen Gegenständen.

Bibliographie

[1] Berndt, Axel / Bohl, Benjamin W. (2018): Music Performance Markup: Formale Beschreibung musikalischer Interpretationen. In: Editio Bd. 32 (2018), Nr. 1, S. 185–204.

[2] Burnard, Lou: ODD chaining for beginners, http://teic.github.io/PDF/howtoChain.pdf

[3] Deutsches Textarchiv: Schema und Dokumentation der DTABf Schema, http://www.deutschestextarchiv.de/doku/basisformat/schema.html

[4] Holmes, Martin 2018: Using ODD for HTML, in Proceedings of of the Text Encoding Initiative Conference and Members Meeting The Markup Conference, Tokyo, Japan, September 9–13 2018. Pages 240–241, https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf

[5] TEI Consortium: Roma ODD Editor, https://romabeta.tei-c.org

[6] Viglianti, Raffaele: One Document Does-it-all (ODD): a language for documentation, schema generation, and customization from the Text Encoding Initiative, https://www.balisage.net/Proceedings/vol24/html/Viglianti01/BalisageVol24-Viglianti01.html

Hackathon „Sortir de la guerre“

Schwandt, Silke
silke.schwandt@uni-bielefeld.de
Universität Bielefeld, Deutschland

Baillot, Anne
anne.baillot@univ-lemans.fr
Université Le Mans, Frankreich

Gervais, Ludovic
ludovic.gervais.etu@univ-lemans.fr
Université Le Mans, Frankreich

Braud, Camille
camille.braud.etu@univ-lemans.fr
Université Le Mans, Frankreich

Thomas, Clement
clement.thomas@univ-lemans.fr
Université Le Mans, Frankreich

Bonsergent, Lou-Ann
lou-ann.bonsergent.etu@univ-lemans.fr
Université Le Mans, Frankreich

Strothotte, Adrian
adrian.strothotte@uni-bielefeld.de
Universität Bielefeld, Deutschland

Niewöhner, Laura Maria
l.niewoehner@uni-bielefeld.de
Universität Bielefeld, Deutschland

Leitung

Prof. Dr. Anne Baillot

Forschungsschwerpunkte: Digitale Philologie, Digital Humanities, Translation Studies.

Dr. Silke Schwandt

Forschungsschwerpunkte: Digitale Geschichtswissenschaft, Digital Humanities, Geschichte des Mittelalters.

Beitragende

Studierende der Universität Le Mans (Studierende der Germanistik, Digital Humanities und Europäischen Studien): Lou-Ann Bonsergent, Camille Braud, Clément Thomas (Le Mans); Ludovic Gervais (2019/2020 Erasmus in Köln)

Teilnehmerzahl

Optimal zwischen 12 und 24 Personen, möglich ab 6.

Ausstattung

Beamer; gute Internetverbindung; großer Raum, in dem sich die TeilnehmerInnen in 3er/4er-Gruppen für die praktischen Teile zusammentun und besprechen können, ohne die anderen Gruppen zu stören; ein Laptop pro TeilnehmerInnengruppe (e.g. max. 8 Laptops).

Pädagogischer Ansatz

Die digitalen Geisteswissenschaften eröffnen vielfältige Spielräume in der Forschung durch die Anwendung neuer Methoden sowie durch die Interaktion mit anderen Disziplinen. Die gleichen Spielräume bieten sich auch in der Vermittlung. Jenseits von der Entwicklung von E-Learning-Angeboten, MOOCs und dergleichen, fehlt eine Exploration vergleichbarer Spielräume für die Lehre in geisteswissenschaftlichen Fächern (vgl. Schön et al. 2017). Eine solche Exploration kann zweierlei adressieren: *digitale Inhalte*, Daten, Wissen und Informationen, oder *digitale Methoden* wie den Umgang mit etwa Datenbanken, Foren, oder Anwendungen.

Auf der Ebene der Inhalte spielt die Auseinandersetzung mit Fragen der Authentizität und Verlässlichkeit von digital verfügbaren Informationen eine zentrale Rolle. Gerade die Geisteswissenschaften beschäftigen sich in ihrem Kern mit der Produktion von verlässlichen Wissensbeständen für die Gesellschaft.

So lernen Studierende, wie sie Informationen und Wissen extrahieren, bewerten und einordnen. Sie lernen den kritischen Umgang mit scheinbar Gegebenem und werden darin geschult, fremde Perspektiven zu erkennen, einzunehmen und zu reflektieren. Diese zentralen Kompetenzen gilt es in der universitären Lehre mit Blick auf digitale Objekte, auf moderne Informationsmedien zu erweitern und anzupassen (vgl. Büttner 2019). Digitale Medien zu kennen bedeutet dabei, diese auch zu verstehen, in ihrer Entstehung und ihrem Stellenwert beurteilen zu können oder ihnen Informationen zu entnehmen.

Verschiedene Institutionen wie auch die Universität Bielefeld verfolgen Data Literacy Education Initiativen, um diese Herausforderungen zu adressieren. Dabei geht es vor allem darum, ein Bewusstsein für die Relevanz von Daten und die damit verbundene Notwendigkeit zu wecken, umsichtig mit Daten umzugehen. An der Universität Le Mans wird ein ähnliches Ziel im Rahmen des Germanistik-Curriculums verfolgt, das Grundlagen der Datenbankverwaltung, der Entwicklung von Web-Interfaces, des Datenmanagements, aber auch der Vermittlung in die Gesellschaft hinein vermittelt.

Mit diesem Workshop kommen beide pädagogischen Ansätze zum Tragen und kehren dabei auch die traditionelle Lehrperspektive um insofern als die Studierenden selbst den Workshop leiten sollen.

Studentische Kompetenzen stärken

Dieser Workshop versteht sich als spielerische Exploration der im Rahmen des Projekts „Sortir de la guerre (1919-1930)" gesammelten Daten. Diese werden den Workshopteilnehmenden sowohl in virtueller als auch in physischer Form zur Verfügung gestellt (die Ausstellung soll im März 2020 in Workshop-Nähe gezeigt werden). Die Studierenden, die an der Konzeption und Realisierung der Ausstellung mitgearbeitet haben, werden an der Konzeption und Durchführung des Workshops aktiv beteiligt: Die französischen Studierenden bekommen im Herbst 2019 eine professionelle Einweisung in die Führung durch die Ausstellung und sie werden zwischen November 2019 und Februar 2020 Schulklassen aus Le Mans die physische und die virtuelle Ausstellung nahebringen, wobei der Akzent auf die zweisprachige Vermittlung liegt. Auf diese Weise sollen sie neben dem Wissen um die Ausstellungsthematik auch ihre Datenkenntnisse (Zusammenspiel von digitalem Bild-, Text- und Tonmaterial und Identifizierung der einschlägigen Formate; Arbeit mit digitalen Umgebungen; Orientierung in Metadaten) einbringen und Erfahrungen in der Wissensvermittlung sammeln. Diese Kompetenzen werden im Rahmen des Workshops eingebracht.

(Daten-)Exploration als zentrales Moment der DH steht im Mittelpunkt des Workshops. Kombiniert werden Elemente aus dem Bereich der DH, der Data Literacy und der spielerischen Vermittlung von Wissen. Spielerische Kultur- und Geschichtsvermittlung sind seit der Entstehung der Museumspädagogik etablierter Teil der Arbeit im Museum. Museen und andere kulturelle Institutionen gehören zudem zu den bevorzugten Arbeitsfeldern von Studierenden aus den Geisteswissenschaften.

Datengrundlage

Die Datengrundlage des Workshops ist das Ergebnis einer Lehrkooperation, die an den Universitäten Le Mans und Paderborn durchgeführt wurde. Bei der Ausstellung « Sortir de la guerre (1919-1930) » handelt es sich um ein gemeinsames wissenschaftliches und pädagogisches Projekt der Faculté des lettres, langues et sciences humaines der Université du Mans und der kulturwissenschaftlichen Fakultät der Universität Paderborn. Aufbauend auf das Potential der Städtepartnerschaft haben DozentInnen, Studierende und die jeweiligen Stadtarchive eine Ausstellung konzipiert, die unter dem Titel „Sortir de la guerre" die Nachkriegsjahre 1919-1930 in beiden Städten dokumentiert. Die Ausstellung gliedert sich in vier thematische Schwerpunkte: Demobilisierung, Neuaufbau, Erinnerungskultur, Zurück ins Leben. Die Ausstellung wird im November 2019 in beiden Städten zeitgleich eröffnet und zirkuliert dann in öffentlichen Einrichtungen. In Le Mans werden Führungen organisiert, die von Studierenden der Geschichte und der Germanistik angeboten werden.

Die Ausstellung präsentiert sich sowohl in physischer Form (23 Tafeln, die Archivbestände abbilden und in beiden Sprachen kommentieren) als auch virtuell. Studierende aus dem 2. Jahr der Germanistik/DH aus Le Mans erarbeiteten eine virtuelle Ausstellung in einer Omeka-Umgebung. Eine interaktive Karte der Erinnerungsorte in Le Mans wurde ebenfalls entwickelt (http://umap.openstreetmap.fr/fr/map/les-commemorations-de-la-ville-du-mans_323901#13/48.0074/0.2123). Das einschlägige Blog (https://sortir1919.hypotheses.org/) informiert über die Fortschritte der Ausstellung. Die Daten der virtuellen Ausstellung werden im Rahmen dieses Workshops bearbeitet um das Projekt durch „Neulektüren" zu bereichern, die im Anschluss in das Ausstellungskonzept eingebaut werden können.

Bei den Daten handelt es sich um archivarische Metadaten, Digitalisate, Tonaufnahmen und Kommentare in Textform. Diese werden den Workshop-TeilnehmerInnen auf GitHub zur Verfügung gestellt. Die Daten wurden im Rahmen der Lehrkooperation erhoben, analysiert und in Form der Ausstellung präsentiert. Nicht alle gesammelten Rohdaten wurden aber ausgewertet: Auch die noch unbearbeiteten Daten werden im Rahmen des Workshops zur Verfügung gestellt.

Programm des Workshops

Der Workshop hat Hackathon-Charakter (vgl. Meyer 2019 und Knoll 2017): Die TeilnehmerInnen werden in 3 bis 4 Teams gegliedert, die jeweils 3 Aufgaben zu bewältigen haben. Für jede Aufgabe stehen 45 Minuten zur Verfügung.

Alle 3 Aufgaben sollen auf der Grundlage der Daten bewältigt werden, die zur Verfügung gestellt werden. Es dürfen aber auch ergänzend andere Daten herangezogen werden (etwa aus den Sammlungen der Europeana zum Ersten Weltkrieg).

Der Workshop ist für ein Zielpublikum konzipiert, dem die Arbeit mit Daten geläufig ist. Sollten sich überwiegend Studierende oder digital nicht rüstige TeilnehmerInnen melden (etwa Studierende), würde jedes Team insgesamt nur eine der drei Aufgaben zuteil bekommen.

Aufgabe 1 hat „so viel wie möglich" zum Motto: Da sollen sich die Teilnehmenden eine Auswertung der Daten ausdenken, die es ermöglicht, soviel Daten wie möglich zu bearbeiten. (möglich wären Visualisierungen oder Sonification)

Aufgabe 2 lautet „so gut wie möglich": Da sollen sich die Teilnehmenden eine Auswertung ausdenken, die so qualitativ wie möglich ist. Es steht ihnen frei, unter dem zur Verfügung gestellten das Material auszusuchen (Bild, Metadaten, Text), was sie auswerten wollen. Die Qualität wird vorrangig nach den FAIR-Prinzipien evaluiert.

Aufgabe 3 ist dann „so weit wie möglich": Die TeilnehmerInnen sollen hier Public History-Tools einsetzen, um die Daten einem größtmöglichen Publikum schmackhaft zu machen.

Die Studierenden sollen im Vorfeld die Daten aufbereiten, den Workshop leiten, die Teams aufteilen helfen, die Daten präsentieren, die Teams unterstützen und sich an der Evaluation der Ergebnisse aktiv beteiligen.

Zeitplan(3,5 Std. Workshop):

- 15 Min. Warming-Up (Präsentation des Konzepts, der Daten, Konstitution der Teams, Raumgestaltung für die Teams)
- 45 Min. Aufgabe 1

- 15 Min. Präsentation und Auszeichnung
- 45 Min. Aufgabe 2
- 15 Min. Präsentation und Auszeichnung
- 45 Min. Aufgabe 3
- 15 Min. Präsentation und Auszeichnung
- 15 Min. Auszeichnung und Ausblick (mögliche Auswertungen der Ergebnisse)

Bibliographie

Büttner, Stephan, ed. (2019): *Die Digitale Transformation in Institutionen Des Kulturellen Gedächtnisses: Antworten Aus Der Informationswissenschaft*. Berlin: Simon Verlag für Bibliothekswissen.

Knoll, Nico (2017): "„HackHPI": How to Organize a Hackathon." In *Veranstaltungen 4.0*. Edited by Thorsten Knoll, 155–69. Wiesbaden: Springer Fachmedien Wiesbaden.

Meyer, Francine / Monika Taddicken (2019): "Hackdays Als Alternatives Lehrformat? Eine Empirische Betrachtung Eines Beispiellehrformats in Bezug Auf Mediale Und Technologische Bildung." In *Teaching Trends 2018: Die Präsenzhochschule Und Die Digitale Transformation*. Edited by Susanne Robra-Bissantz et al. 1. Auflage. Digitale Medien in der Hochschullehre 7. Münster: Waxmann.

Schön, Sandra / Veronika Hornung-Prahauser, Patricia Schedifka, and Markus Alsleben, eds. (2017): *Innovation Durch Exploration: Innovationsanstöße Zum Internet Der Dinge (Internet of Things, IoT) Durch Offenes Explorieren Und Experimentieren in Technologielaboren, Kreativ- Und Innovationsräumen*. Salzburg: Books ON DEMAND, 2017.

Trilcke, Peer / Frank Fischer (2018): "Literaturwissenschaft Als Hackathon: Zur Praxeologie Der Digital Literary Studies Und Ihren Epistemischen Dingen." In *Wie Digitalität Die Geisteswissenschaften Verändert: Neue Forschungsgegenstände Und Methoden*. Edited by Martin Huber and Sibylle Krämer. Sonderband der Zeitschrift für digitale Geisteswissenschaften 3. DOI: http://www.zfdg.de/sb003_003.

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

Kremer, Gerhard
gerhard.kremer@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Das Ziel dieses Tutorials ist es, den Teilnehmenden 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] 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 (s. a. Willand et al., 2019 zu dieser Form in den DH), 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.Zusätzlich werden bei jedem Testlauf Informationen über die Entscheidungen protokolliert, um die Erklärbarkeit der Ergebnisse zu unterstützen.

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.

Abgrenzung zur Einreichung „Vom Phänomen zur Analyse – ein CRETA-Workshop zur reflektierten Operationalisierung in den DH“

Neben diesem CRETA-Hackatorial befindet sich noch ein weiterer Workshop des Stuttgarter DH-Zentrums CRETA in Begutachtung. Auch wenn es eine gewisse Schnittmenge zwischen den Workshops gibt (Textgrundlagen, Anwendungsfälle), ist die jeweilige Zielsetzung grundsätzlich verschieden: Während es beim hier vorgestellten CRETA-Hackatorial um Verfahren des Maschinellen Lernens geht, konzentriert sich der parallel ausgearbeitete CRETA-Workshop auf den grundlegenderen Schritt der Operationalisierung – es geht also darum, Ansätze aufzuzeigen, wie ein Untersuchungsvorhaben oder theoretisches Konzept überhaupt für die computergestützte Analyse “vor- bzw. aufbereitet” werden kann. Beide Workshops ergänzen einander sinnvoll, was die Teilnahme an beiden oder an nur einem der Workshops möglich macht.

Anhang

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 und ausreichend Stromanschlüssen für die Laptops der Teilnehmenden keine besondere technische Ausstattung benötigt. Es sollte sich um einen Raum handeln, in dem genügend Platz ist, durch die Reihen zu gehen und den Teilnehmenden über die Schulter zu blicken.

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

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.

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

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

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.

Willand, Marcus / Gius, Evelyn / Reiter, Nils (2019): "Ein neues Format für die Digital Humanities: Shared Tasks. Zur Annotation narrativer Ebenen." in *Abstracts of DHd: multimedial und multimodal,* Frankfurt.

Modellierung und Verwaltung von DH-Anwendungen in TOSCA

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH)

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH)

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH)

Harzenetter, Lukas
lukas.harzenetter@iaas.uni-stuttgart.de
Universität Stuttgart, Institut für Architektur von Anwendungssystemen (IAAS)

Breitenbücher, Uwe
uwe.breitenbuecher@iaas.uni-stuttgart.de
Universität Stuttgart, Institut für Architektur von Anwendungssystemen (IAAS)

Leymann, Frank
frank.leymann@iaas.uni-stuttgart.de
Universität Stuttgart, Institut für Architektur von Anwendungssystemen (IAAS)

Abstract

Das aktuell vom Institut für Architektur von Anwendungssystemen (IAAS) der Universität Stuttgart und vom Data Center for the Humanities (DCH) der Universität zu Köln bearbeitete Projekt *SustainLife – Erhalt lebender, digitaler Systeme für die Geisteswissenschaften* befasst sich mit der Konservierung von Forschungssoftware im Bereich der Digital Humanities (DH). Dabei wird der Topology Orchestration Specification for Cloud Applications (TOSCA) Standard verwendet, um das Deployment von DH-Anwendungen vollständig zu automatisieren und diese langfristig verfügbar zu halten. Um der DH Community unseren Ansatz interaktiv zu demonstrieren, möchten wir im Vorfeld der DHd 2020 einen Workshop zur *Modellierung und Verwaltung von DH-Anwendungen in TOSCA* durchführen. Dabei sollen Kernkompetenzen bezüglich der Modellierung von Softwaresystemen mit TOSCA sowie Erfahrungen und Best Practices im Umgang mit OpenTOSCA, einer open-source Implementierung des TOSCA Standards, vermittelt werden.

Problemstellung

Die zunehmende Etablierung der DH als ein eigenes Forschungsfeld sowie der damit einhergehend vermehrte Einsatz von digitalen Methoden im Forschungsprozess erfordern daran angepasste Mittel der Ergebnissicherung. Zur Langzeitarchivierung von Forschungsprimärdaten gibt es bereits etablierte Strategien, bspw. die Nutzung standardisierter Datenformate und die Übermittlung relevanter Daten an einschlägige Repositorien. Weitestgehend unberücksichtigt bleibt dabei, dass viele der in DH-orientierten Forschungsprozessen erzeugten digitalen Artefakte nicht in Form von Primärdaten, sondern in Form von Forschungssoftware vorliegen. Die Vielfalt der in den DH erzeugten Software beinhaltet auch sog. *lebende Systeme* (Sahle, Patrick / Kronenwett, Sabine: 2013), deren Laufzeitumgebung unerlässliche Daten enthält und die somit nicht statisch abbildbar sind. Da solche lebenden Systeme im Gegensatz zu klassischen Erkenntnisträgern wie bspw. Monographien oder Lexika nicht ohne kontinuierliche Wartung auskommen, stellen Erhalt, Betreuung und dauerhafte Bereitstellung große technische, organisatorische und finanzielle Hürden dar. Weiterhin erfordert die Heterogenität der in den DH erzeugten Forschungssoftware eine höchst flexible Methodologie bzw. Technologie, die Standardisierung, Nachnutzbarkeit und Archivierung von möglichst vielen digitalen Artefakten gewährleisten kann (Barzen, Johanna et al.: 2018). Neben den genannten Herausforderungen (He-

terogenität, Unterfinanzierung und Überalterung digitaler Artefakte) fordert die wissenschaftliche Praxis dauerhafte Interoperabilität und Nachvollziehbarkeit aller Erkenntnisträger. Bezogen auf digitale Systeme sind diese Forderungen (1) konstante Zugänglichkeit, (2) die Möglichkeit eines fehlerfreien Betriebs und (3) die Möglichkeit jeden Entwicklungsstand einer Forschungsanwendung zu jedem Zeitpunkt und ohne große strukturelle Hürden nachzuvollziehen bzw. wiederherzustellen zu können.

Lösungsansatz

Da der TOSCA-Standard (OASIS: 2013, 2019) es erlaubt, Anwendungen standardisiert und anbieterunabhängig zu modellieren, zu provisionieren und zu deployen, eignet er sich auch zum langfristigen Archivieren und Betreiben von in den DH erzeugter Forschungssoftware (vgl. Neuefeind et al.: 2018). Hierbei werden Anwendungen mithilfe von wiederverwendbaren Komponententypen, sog. *Node Types*, modelliert. Um Abhängigkeiten zwischen diesen unterschiedlichen Komponenten einer Anwendung darzustellen, werden verschiedene Beziehungstypen, sog. *Relationship Types*, verwendet. So kann bspw. eine einfache PHP Webanwendung, die auf eine Datenbank zugreift, als eine Instanz des Node Types *PHP Anwendung* modelliert werden, welche sich zu einer Instanz des *MySQL Datenbank* Node Types verbindet. Die Verbindung der beiden Komponenten zueinander wird durch den Relationship Type *connectsTo* dargestellt. Zusätzlich kann bspw. angegeben werden, dass beide Komponenten auf einer Ubuntu virtuellen Maschine (VM) installiert werden müssen, welche wiederum eine Instanz des Node Types *Ubuntu VM* ist (vgl. Neuefeind et al.: 2019).

Solch eine Beschreibung der Anwendungskomponenten und deren Beziehungen untereinander wird *Anwendungstopologie* genannt. Weiterhin ermöglicht TOSCA durch sein Typensystem die Modellierung von wiederverwendbaren Komponententypen, sodass bspw. der *PHP Anwendung* Node Type in mehreren unterschiedlichen Anwendungen verwendet werden kann. Dadurch kommen Synergieeffekte zum Tragen, da bereits existierende Node Types von anderen Modellen wiederverwendet werden können, womit die Modellierung neuer Anwendungen deutlich schneller und einfacher wird. Darüber hinaus bietet die open-source TOSCA Implementierung OpenTOSCA die Möglichkeit Anwendungen grafisch per drag-and-drop zu modellieren, wodurch die Modellierung nochmals vereinfacht wird.

Inhalte des Workshops

Neben einem Einblick in verschiedene Lösungsansätze wird den Teilnehmenden zunächst der konzeptuelle Rahmen des TOSCA-Standards vermittelt. Auf Basis dieser theoretischen Vorarbeit sollen praxisorientierte Arbeitseinheiten in den Umgang mit OpenTOSCA einführen. Durch die Vermittlung sowohl der theoretischen Grundlagen als auch der praktischen Anwendung des TOSCA-Standards werden die Teilnehmenden in die Lage versetzt, (Forschungs-) Software standardkonform zu modellieren und mit Hilfe von OpenTOSCA bereitzustellen.

Die praxisorientierten Arbeitseinheiten werden wie folgt strukturiert: Ausgehend von der Identifikation aller Komponenten eines Softwaresystems soll dieses im Hinblick auf den TOSCA-Standard als Anwendungstopologie erfasst und abgebildet werden. Dabei werden auch theoretische Konzepte wie sog. *Software-Stacks* praxisnah eingebunden. Daraufhin soll die erarbeitete Anwendungstopologie mittels OpenTOSCA und dem darin enthaltenen graphischen Editor *Winery* (vgl. Kopp et al.: 2013) modelliert werden, um die modellierte Anwendung letztendlich mit OpenTOSCA automatisiert bereitzustellen. Des Weiteren werden unsere Erfahrungen und Best-Practices im Umgang mit TOSCA und der Modellierung von Anwendungen in OpenTOSCA an die Community weitergegeben.

Zielgruppe des Workshops

Der Workshop richtet sich in erster Linie an Mitarbeiter von Datenzentren, Bibliotheken und sonstigen Institutionen mit Ausrichtung auf Infrastrukturen für Langzeitarchivierung und -betrieb heterogener lebender Systeme. Vorerfahrungen im Umgang mit Linux und mit den Themen *Shell-Scripting*, *Software-Stacks* und *Service-Orchestrierung* sind hilfreich, aber nicht notwendig zur erfolgreichen Teilnahme. Um einen produktiven Kontext zur Vermittlung der aufgezeigten Inhalte zu schaffen und individuelle Beratung und Betreuung zu ermöglichen, streben wir ein Ideal von 20 bis maximal 30 Teilnehmenden an.

Technische Vorbedingungen

Zur erfolgreichen Teilnahme am Workshop ist es notwendig, dass jeder Teilnehmer ein eigenes Arbeitsgerät mitbringt. Weiterhin ist es wünschenswert, dass alle Teilnehmer im Vorfeld des Workshops eine OpenTOSCA-Instanz auf ihren Geräten aufsetzen, um eigene Modellierungen durchzuführen und zu sichern. Zwar wird eine zentral erreichbare Instanz bereitgestellt, jedoch kann keine Garantie für den langfristigen Erhalt dieser Instanz und damit auch der dort hinterlegten Ergebnisse übernommen werden (es ist jedoch problemlos möglich diese Ergebnisse zum Abschluss des Workshops aus der zentralen Instanz zu exportieren und somit weiterhin nutzen zu können). Darüber hinaus sind eine stabile Internetverbindung sowie eine umfassende Versorgung der Teilnehmer mit Netzstrom unabdingbar.

Für den erfolgreichen Ablauf des Workshops werden alle angemeldeten Teilnehmer im Vorfeld des Workshops mit allen notwendigen Informationen zur Inbetriebnahme von OpenTOSCA ausgestattet. Weiterhin werden einschlägige Dokumentationen, Publikationen und Anleitungen sowohl vorab als auch im Kontext des Workshops bereitgestellt.

Forschungsgebiete der Referenten

Brigitte Mathiak

Brigitte Mathiak ist Vorsitzende Sprecherin des Data Center for the Humanities und insbesondere an den Themen Datenmanagement und Text Mining interessiert. Die Idee zum *SustainLife* LIS-Projekt entstand, nachdem sie immer wieder erlebt hat wie lebende Systeme aufgegeben oder vernachlässigt werden müssen. Sie ist Juniorprofessorin für Digital Humanities an der Universität zu Köln und darüber hinaus Se-

nior Scientist am Leibniz-Institut für die Sozialwissenschaften (GESIS).

Claes Neuefeind

Claes Neuefeind ist Postdoc am Cologne Center for eHumanities (CCeH) der Universität zu Köln. Bis Oktober 2019 bearbeitete er gemeinsam mit Philip Schildkamp das DFG-LIS-Projekt *SustainLife* für das DCH und ist seither am CCeH verantwortlich für die Geschäftsführung der Koordinierungsstelle Digital Humanities der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste.

Frank Leymann

Frank Leymann ist Professor für Informatik und Direktor des Institute of Architecture of Application Systems (IAAS) an der Universität Stuttgart. Seine Forschungsinteressen umfassen serviceorientierte Architekturen und zugehörige Middleware, Workflow- und Geschäftsprozessmanagement, Cloud Computing und damit verbundene Aspekte des Systemmanagements sowie Design Patterns. Frank ist Mitautor von mehr als 400 peer-reviewed Papers, etwa 70 Patenten und mehreren Industriestandards. Er ist ein gewähltes Mitglied der Europäischen Akademie.

Lukas Harzenetter

Lukas Harzenetter ist wissenschaftlicher Mitarbeiter am Institut für Architektur von Anwendungssystemen (IAAS) an der Universität Stuttgart. Seinen Master of Science Abschluss erhielt er von der Universität Stuttgart im Studiengang Software Engineering im Jahr 2018. Seine Forschungsinteressen liegen im Bereich Cloud-Deployment und Management. Er beschäftigt sich vor allem damit, wie sich Deploymentmodelle entwickeln. Lukas ist Teil des DFG-LIS-Projekts *SustainLife* und arbeitet an nachhaltigen Anwendungsimplementierungen im Bereich der digitalen Geisteswissenschaften.

Philip Schildkamp

Philip Schildkamp forscht seit 2015 und lehrt seit 2017 an der Universität zu Köln. Er studierte Soziologie, Psychologie und Informationsverarbeitung. Schwerpunktthemen seiner Beschäftigung sind technische Infrastrukturmaßnahmen im Bereich der (digitalen) Geisteswissenschaften und die Orchestrierung von verteilten Softwaresystemen. Seit März 2018 bearbeitet er am DCH das DFG-LIS-Projekt *SustainLife*.

Uwe Breitenbücher

Uwe Breitenbücher ist wissenschaftlicher Mitarbeiter und Postdoc am Institut für Architektur von Anwendungssystemen (IAAS) der Universität Stuttgart. Seine Forschungsvision ist die Verbesserung der Bereitstellung von Cloud-Anwendungen und des Anwendungsmanagements durch die Automatisierung der Anwendung mithilfe von Managementmustern. Uwe war Teil des CloudCycle-Projekts, in dem das OpenTOSCA Ecosystem entwickelt wurde. Seine aktuellen Forschungsinteressen umfassen cyber-physikalische Systeme, Blockchains und Microservices.

Acknowledgments

Dieser Workshop wird teilweise durch das DFG-LIS Projekt *SustainLife* (379522012) finanziert.

Bibliographie

Barzen, Johanna / Blumtritt, Jonathan / Breitenbücher, Uwe / Kronenwett, Sabine / 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), S. 471-474.

Kopp, Oliver / Binz, Tobias / Breitenbücher, Uwe / Leymann, Frank (2013): Winery – A Modeling Tool for TOSCA-based Cloud Applications. In: Proceedings of the 11th International Conference on Service-Oriented Computing (ICSOC 2013), S. 700-704.

Neuefeind, Claes / Harzenetter, Lukas / Schildkamp, Philip / Breitenbücher, Uwe / Mathiak, Brigitte / Barzen, Johanna / Leymann, Frank (2018): The SustainLife Project – Living Systems in Digital Humanities. In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing (SummerSoC 2018) (IBM Research Report RC25681), S. 101-112.

Neuefeind, Claes / Schildkamp, Philip / Mathiak, Brigitte / Marčić, Aleksander / Hentschel, Frank / Harzenetter, Lukas / Breitenbücher, Uwe / Barzen, Johanna / Leymann, Frank (2019): Sustaining the Musical Competitions Database. A TOSCA-based Approach to Application Preservation in the Digital Humanities. In: Book of Abstracts der 29. Digital Humanities Conference (DH 2019), https://dev.clariah.nl/files/dh2019/boa/0574.html (Stand: 10.09.2019).

OASIS (2013): Topology and Orchestration Specification for Cloud Applications Version 1.0, http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html (Stand: 10.09.2019).

OASIS (2019): TOSCA Simple Profile in YAML Version 1.2, http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.2/TOSCA-Simple-Profile-YAML-v1.2.html (Stand: 10.09.2019).

Sahle, Patrick / Kronenwett, Sabine (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.

Nachlass Ludwig Wittgenstein: Softwaretechnologien und computerlinguistische Methoden der Software-Infrastruktur um die FinderApp WiTTFind

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
Ludwig-Maximilians Universität München

Babl, Florian
Florian.Babl@campus.lmu.de
Ludwig-Maximilians Universität München

Eisterhues, Marcel
Marcel.Eisterhues@campus.lmu.de
Ludwig-Maximilians Universität München

Röhrer, Ines
I.Roehrer@campus.lmu.de
Ludwig-Maximilians Universität München

Still, Sebastian
Sebastian.Still@campus.lmu.de
Ludwig-Maximilians Universität München

Ullrich, Sabine
sabine.ullrich@campus.lmu.de
Ludwig-Maximilians Universität München

Landes, Florian
florian.landes@kbl.badw.de
Bayerische Akademie der Wissenschaften, München

Lindinger, Matthias
matthias.lindinger@bsb-muenchen.de
Bayerische Staatsbilbliothek, München

Die Infrastuktur und das Projekt

Seit 2010 kooperieren das Wittgenstein Archiv der Universität Bergen und das Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians Universität München in der Forschungsgruppe „Wittgenstein Advanced Search Group“ (WAST). Die Forschungsgruppe entwickelt Web-Frontends (FinderApps) und spezielle Suchwerkzeuge, die sich gut für die Forschung und Lehre im Bereich der Digital Humanities eignen. Ihre erste Suchmaschine, die FinderApp WiTTFind (wittfind.cis.lmu.de, siehe Abb. 1), die den von der UNESCO zum Weltkulturerbe (im Jahr 2017) erhobenen (Schmidt 2018) Nachlass von Ludwig Wittgenstein durchsucht, gewann im Jahre 2014 der EU-Open-Humanity Award. Der Preis zeichnet Gruppen aus, die herausragende Technologie im Bereich der Humanities entwickelt haben. Die in der Forschergruppe programmierte FinderApp WiTTFind erlaubt es, mit hochqualifizierten, computerlinguistisch orientierten Suchwerkzeugen Nachlasstrans-kriptionen zu durchsuchen. Die Transkriptionen entstammen der *Bergen Normalized Edition*, die die Grundlage der Wittgenstein Edition bildet. Neben den gefundenen Treffern der Suchmaschine, werden in den Suchergebnissen von WiTTFind die Faksimile-Extrakte aus den Originaldokumenten angezeigt. So kann der Nutzer die „Aura“ der gefundenen Textstelle im Original studieren und nicht nur den transkribierten Text sehen.

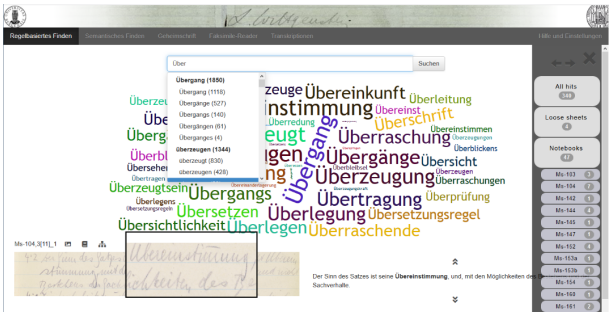


Abbildung 1: WiTTFind (http://wittfind.cis.lmu.de)

Damit derNutzer auch den seitenweisen Kontext des Suchtreffers im Original studieren kann, wurde am CIS eine weitere WEB-Applikation entwickelt, der doppelseitige Reader. Dieser Reader ermöglicht es, vom Suchtreffer direkt an die entsprechende Stelle im entsprechenden Dokument des Originals zu springen. Im doppelseitigen Lesemodus kann der Nutzer in den Faksimile des originalen Dokuments blättern. Eine symmetrische Autovervollständigung gibt während der Suchanfrage einen statistischen und lexikalischen Zugang zu den Wörtern, die in der Edition vorkommen. Im Zentrum der Suche steht die selbstprogrammierte C++ Suchmaschine wf, die mit Hilfe von Vollformlexika (WiTTlex), verbessertem POS-Tagging und weiteren Metainformationen regelbasiertes Suchen erlaubt. Zum Aufspüren semantisch ähnlicher Textpassagen in der Edition gibt es das NLP-Tool WiTTSim.

Die thematisch getrennten Aufgaben innerhalb der Infrastruktur der WAST-Tools (siehe Abb. 2) werden über REST-API’s von einzelnen Microservices realisiert, deren zentrale Datenhaltung über eine mongo Datenbank realisiert wird. Die Oberflächen der FinderApps werden mit HTML5, Javascript

und Bootstraptechniken für WEB-Browser programmiert und möglichst browserunabhängig gehalten.

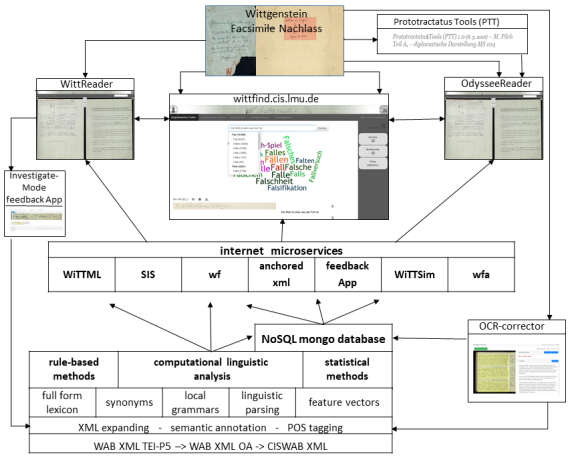


Abbildung 2: Infrastruktur der WAST-Tools (http://gitlab.cis.lmu.de)

Alle Programme, Schnittstellen und Entwicklungen werden dokumentiert (siehe Abb. 3) und Tutorials für Anschlussprojekte entwickelt. So ist gewährleistet, dass die Tools und Suchmaschinen nachhaltig verwendet und auch für die Forschung und Lehre eingesetzt werden können. Als Versionskontrollsystem wird git verwendet.

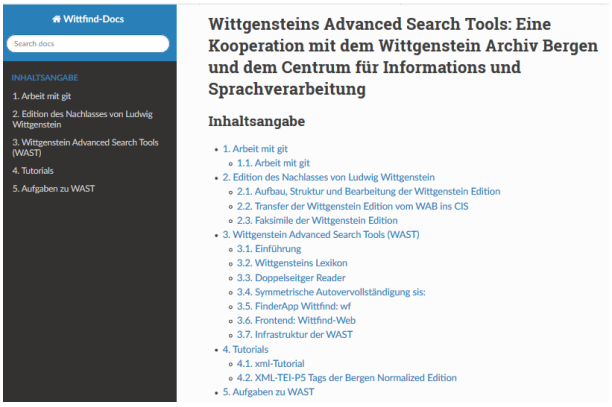


Abbildung 3: Dokumentation der WAST-Tools: http://wittfind.cis.uni-muenchen.de/wast/infrastruktur/index.html

Bei der Entwicklung der Infrastruktur der WAST-Tools wurden die strengen Vorgaben des EU-Open-Humanity Awards eingehalten: Forderungen nach Open-Source, interdisziplinäre Öffnung und Nachhaltigkeit. Diese Offenheit ermöglichte es weitere FinderApps für andere Wissenschaftsbereiche zu implementieren: GoetheFind (Faust-I und Faust-II Edition, Deutsches Textarchiv Berlin (XML-TEIP5, DTA Basis Format)), HistoFind (Briefwechsel Erzherzog Leopold Wilhelms an Kaiser Ferdinand III. aus dem Reichsarchiv Stockholm; Kooperation mit Historikern) und den OdysseeReader (Schreibprozess der zur Logisch-Philosophischen-Abhandlung führte; Kooperation mit Philosophen).

In diesem Workshop werden die verwendeten Softwaretechnologien und computerlinguistischen Methoden im konkreten Einsatz vorgestellt. Den Teilnehmer*innen wird ein Debian-10 Container mit allen notwendigen Programmen, Tools und Dokumentation der gesamten Softwareinfrastruktur zur Verfügung gestellt. Innerhalb dieses Containers können die Teilnehmer*innen die einzelnen Tools der WAST-Projektgruppe kennenlernen und bekommen von den Projektmitarbeiter*innen kleine Aufgaben gestellt, die sie dann mit ihnen bearbeiten. So können sie die Arbeitsweise der WAST Infrastruktur konkret kennenlernen.

Im Workshop werden folgende Datenformate, Tools und Programmierkonzepte vorgestellt und geübt

Gitlab Projektmanagement und Continuous Integration, XML TEI-P5 Edition CISWAB, Faksimilestrukturierung und Texterkennung, lexikalische Arbeit, WEB-Oberfläche der FinderApps und Einsatz mit Micorservices, doppelseitiger Faksimilereader mit MongoDB, NLP-Tools zur semantischen Ähnlichkeitssuche, Vorstellung und Programmierung einer regelbasierten Suchmaschine und die Erstellung eines Dokumentationssystems mit Sphinx.

Voraussetzungen an die Kursteilnehmer*innen

Programmierkenntnisse (Grundkenntnisse): LINUX (Arbeit mit der UNIX-Shell), Python, XML, HTML, git, javascript, POS-Tagging.

Da beim Workshop einige Entwickler der WAST-Tools anwesend sein werden, gibt es die Möglichkeit auch vertieft in die jeweilige Thematik einzusteigen.

Gitlab Projektmanagement und Continuous Integration (Hadersbeck, Still)

Im gesamten Projekt wird als Versionierungssystem git verwendet. Die Projektrepositories werden auf zwei unterschiedlichen Rechnern ausgerollt: Dem preview-Server für Tests und einem Projektserver für die offizielle Onlineversion. Es wird das in der Praxis bewährte „git branching model“ kombiniert mit einer „continuous integration“ Technik eingesetzt. Mit einer Feedbackapp können Nutzer Fehler melden oder Implementierungswünsche äußern, die in Issues innerhalb der Projektrepositories bearbeitet werden.

XML TEI-P5 Edition CISWAB (Hadersbeck)

Als Datenbasis für das WiTTFind Projekt wird die „Bergen Nachlass Edition“ (BNE) verwendet, die sich an den Richtlinien der Text Encoding Initiative (TEI-P5) orientiert. Im Workshop werden die wichtigen TEI-XML-Elemente der BNE vorgestellt.

Faksimilestrukturierung und Erkennung (Eisterhues, Landes)

Da in den FinderApps neben den gefunden Textstellen auch die zugehörigen Faksimileextrakte aus der Edition dargestellt werden, sind Kenntnisse der Bildkoordinaten der Textstellen nötig. Diese Koordinaten werden mit Hilfe einer Kette von Bildverarbeitungstools ermittelt. Da bei Manuskripten und bei manuellen Änderungen in Dokumenten die automatische Zeichenerkennung unbrauchbare Ergebnisse liefert, wurden eigene Strategien entwickelt, die die Informationen aus der BNE nutzen. Im Workshop werden die eingesetzten Tools und Optimierungsstrategien vorgestellt.

Lexikalische Arbeit (Lokale Grammatiken, Semantik) (Röhrer)

Zur lemmatisierten Suche, Partikelverberkennung und semantischen Wortfeldern wurden spezielle Projektlexika entwickelt (Röhrer 2017). Die Lexika enthalten alle Wörter der zu durchsuchenden Edition und sind mit grammatischen Angaben und zum Teil mit zusätzlichen semantischen Informationen versehen. Diese Lexika und ein nachgestelltes optimiertes Part-of-Speech Tagging ist die Grundlage für die computerlinguistischen Methoden, die bei der regelbasierten Suche im Nachlass von Ludwig Wittgenstein eingesetzt werden.

Regelbasierte Suchmaschine (Babl)

Im Zentrum der FinderApps steht die Suchmaschine wf, ein multithreaded C++ Programm, das viele Anfragemöglichkeiten zur Suche implementiert: Einwort und Mehrwortsuche (mit internem Rankingverfahren) und reguläre Ausdrücke kombiniert mit linguistischen Anfragen (Morphologische Eigenschaften, POS-Tags, semantische und syntaktische Tags). Für das Rankingverfahren wird für jeden Suchtreffer die Relevanz zur Suchanfrage berechnet. Die Qualität für jeden Suchtreffer, die Distanz zwischen den einzelnen Wörtern und unterschiedlichen Belohnungs- und Bestrafungsparametern, gehen in die Berechnung der Relevanz ein. Die Treffer werden dann nach dieser sortiert und auf der Website ausgegeben. Durch dieses neuartige Ranking kann nun auch nach verschiedenen Wörtern gesucht werden, die im Text nicht direkt hintereinander stehen müssen.

NLP-Tool Semantische Ähnlichkeitssuche (Ullrich)

Zur Extraktion von semantisch ähnlichen Bemerkungen wurde das Analysetool WiTTSim (Ullrich 2018) entwickelt, welches anhand von semantischen und syntaktischen Features ähnliche Texte identifiziert. Da die enorm hohe Anzahl von etwa 100.000 Features in Kombination mit den zu vergleichenden 54.000 Bemerkungen eine effiziente Suche unmöglich macht, wurde ein semantisches Clustering-Verfahren vorgeschaltet (Ullrich 2019), welches durch Dimensionsreduktion und Gruppierung der Texte die Rechenzeit der Ähnlichkeitssuche um den Faktor 100 beschleunigt.

WEB-Oberfläche der FinderApps und Micorservices (Hadersbeck, Still)

Zur Arbeit mit WiTTFind wird dem User eine WEB-basierte FinderApp zur Verfügung gestellt, die über REST-APIs und „internet microservices“ mit den WAST-Tools kommuniziert. HTML5, Javascript und Bootstrap-css erlauben den Aufbau der WEB-page, die nahezu browserunabhängig die Schnittstelle zum Anwender darstellt.

Doppelseitiger Faksimilereader und MongoDB (Lindinger)

Der doppelseitiger Faksimilereader ist eine komplett eigenständige Anwendung mit Suchschlitz und Investigate Mode zur gleichzeitigen Betrachtung von Faksimile und Transkription. Außerdem gibt es zahlreiche weitere Features, die es den Nutzern sehr bequem erlauben, die gefunden Treffer der Suchmaschine im Kontext einer doppelseitigen Darstellung der Faksimile zu sehen und gleichzeitig durch die Dokumente der Forschungsdomäne zu blättern. Sämtliche Informationen bzgl. Edition und Faksimile sind in einer MongoDB gespeichert und werden über HTTP-Schnittstellen abgefragt.

Dokumentationssystem Sphinx (Babl) (siehe Abb.2)

Für jedes Teilprojekt der Wittgenstein Advanced Search Tools (WAST) wird im entsprechenden Gitlab Ordner eine README.md Datei erstellt, das in einer Dokumentation, die alle Projekte umspannt mithilfe der Software Sphinx zusammengefasst und online auf ansprechende Art und Weise darstellt. Die Dokumentation hilft, neuen Studierenden einen schnelleren Einstieg in das Projekt zu finden und ermöglicht es, das gesamte WAST-Projekt schnell nach bestimmten Fachbegriffen zu durchsuchen.

Programm des Workshops (ganztages Workshop)

Überblick/Einführung/Vorstellungsrunde

Digitaler Zugang zum Nachlass von Ludwig Wittgenstein, das Projekt WAST (Dr. Max Hadersbeck)

Fragen/ Diskussion/ gewünschte Schwerpunkte der Teilnehmer*innen des Workshops

WAST-Spezialthemen (jeweils ca. 15 Min. Theorie / 20 Min. Praxis)

- Gitlab Projektmanagement und Continuous Integration mit git production / testing server (Hadersbeck, Still)
- XML TEI-P5 Edition CISWAB (Hadersbeck): Bergen Normalized Edition und xslt-Transformationen und Investigate-Mode von WiTTFind

- Faksimilestrukturierung und OCR Erkennung (Eisterhues, Landes)
- Lexikalische Arbeit (Röhrer): Lemmatisierte Suche, Lexika, Lokale Grammatiken, Query Beispiele
- WEB-Oberfläche der FinderApps und Microservices (Hadersbeck, Still): Flask server, Javascript
- Doppelseitiger Faksimilereader und mongodb (Lindinger)
- NLP-Tool Semantische Ähnlichkeitssuche (Ullrich): NLP-Python Libraries, Funktionalitäten
- Regelbasierte Suchmaschine (Babl): Programmierung C++, make/cmake, client-server Programmierung mit C++
- Dokumentationssystem Sphinx (Babl): Markdown, Sphinx Installation, 2HTML, 2PDF

Arbeitsgruppen: Diskussionen/Spezialfragen

Je nach Interesse der Teilnehmer*innen unter der Leitung der einzelnen Dozent*innen.

Kurzbiographie der Dozent*innen

Florian Babl (CIS)

Bachelorarbeit: Entwicklung eines Rankingverfahrens der Suchtreffer für die FinderApp WiTTfind im Nachlass Ludwig Wittgensteins

Forschungsschwerpunkte: verschiedene Rankingalgorithmen und ihre Funktionalität mit dem Ziel der Rankingverbesserung.

Marcel Eisterhues (CIS)

Forschungsschwerpunkte: Der momentane Forschungsschwerpunkt ist die automatische Seitensegmentierung von handgeschriebenen Texten.

Max Hadersbeck (CIS)

Projektleiter und Dozent am CIS

Forschungsschwerpunkte: Digitaler Zugang zum Nachlass von Ludwig Wittgenstein, FinderApp WiTTFind, Wittgenstein Advanced Search Tools, Programmierung: C++, Python, XML

Florian Landes (Kommission für bayerische Landesgeschichte bei der Bayerischen Akademie der Wissenschaften)

Bachelorarbeit: Optical Character Recognition (OCR) – Optische Zeichenerkennung (OZE) Ein Werkzeug zur Verknüpfung von digitaler Edition und Faksimile? Semiautomatische Ermittlung von Bildkoordinaten für WiTTFind

Forschungsschwerpunkte: OCR, OZE, Bavarikonprojekt Ortsnamen des Regierungsbezirks Schwaben

Ines Röhrer (CIS)

Masterarbeit: Lexikon, Syntax und Semantik - computerlinguistische Untersuchungen zum Nachlass Ludwig Wittgensteins

Forschungsschwerpunkte: Digitales Speziallexikon WiTTLex für den Nachlass von Ludwig Wittgenstein

Sebastian Still (CIS)

Masterarbeit: Ludwig Wittgenstein: 100 Jahre Traktatus. Der Odyssee-Reader, ein web-basiertes Tool zur textgenetischen Suche im Traktatus

Forschungsschwerpunkte: moderne Frontend Programmierung, NLP (Backend)

Sabine Ullrich (CIS)

Masterarbeit: Clustering zur Verbesserung der Performanz einer Ähnlichkeitssuche

Forschungsschwerpunkte: Natural Language Processing, Data Mining, semantische Ähnlichkeitserkennung im Nachlass von Ludwig Wittgenstein

Bibliographie

Babl, Florian (2019): *Entwicklung eines Rankingverfahrens der Suchtreffer für die FinderApp WiTTFind im Nachlass Ludwig Wittgensteins*. Bachelor‘s thesis. LMU.

Landes, Florian (2019): *Optical Character Recognition (OCR) – Optische Zeichenerkennung (OZE). Ein Werkzeug zur Verknüpfung von digitaler Edition und Faksimile? Semiautomatische Ermittlung von Bildkoordinaten für WiTTFind*, Bachelorarbeit, LMU.

Lindinger, Matthias (2013): *Highlighting von Treffern des Suchmaschinentools WiTTFind im zugehörigen Faksimile.* Bachelor‘s thesis, LMU.

Lindinger, Matthias (2015): *Entwicklung eines WEB-basierten Faksimileviewers mit Highlighting von Suchmaschinen-Treffern und Anzeige der zugehörigen Texte in unterschiedlichen Editionsformaten*. Master's thesis, LMU.

Pichler, Alois (2017): *Wittgenstein Archives at the University of Bergen (WAB): Open Access to Wittgenstein's Nachlass. XML based Interactive Dynamic Presentation (IDP) of WAB's Nachlass transcriptions.* 16. Mai 2017. http://wab.uib.no/transform/wab.php?modus=opsjoner [letzter Zugriff 20.09.2019].

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Øyvind L. (2014): „Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST)“ , in: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 91-96. ACM.

Hadersbeck , Maximilian / Pichler, Alois / Bruder, Daniel / Schweter, Stefan (2016): *New (re)search possibilities for Wittgenstein's Nachlass II: Advanced Search, Navigation and Feedback with the FinderApp WiTTFind*. http://wab.uib.no/alois/Hadersbeck_Pichler%20Kirchberg2016.pdf [letzter Zugriff 20.09.2019].

Röhrer, Ines / Ullrich, Sabine / Hadersbeck, Maximilian (2019): *Weltkulturerbe international digital: Erweiterung der Wittgenstein Advanced Search Tools durch Semanti-*

sierung und neuronale maschinelle Übersetzung. multimedial multimodal. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 25. - 29.03.2019 an den Universitäten zu Mainz und Frankfurt.

Röhrer, Ines (2017): *Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass.* Bachelor's thesis, LMU.

Schmidt, Alfred (2018): „Ludwig Wittgenstein's Nachlass in the UNESCO Memory of the World register.", in: *Nordic Wittgenstein Review* 7(2):209–213.

Ullrich, Sabine / Bruder, Daniel / Hadersbeck, Maximilian (2018): Aufdecken von „versteckten" Einflüssen: Teil-Automatisierte Textgenetische Prozesse mit Methoden der Computerlinguistik und des Machine Learning. Kritik der digitalen Vernunft. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 26.02.-02.03. 2018 an der Universi tät zu Köln, veranstaltet vom Cologne Center for eHumanities (CceH).

Ullrich, Sabine (2019): *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*. Master's thesis. LMU.

OCR4all – Eine semi-automatische Open-Source-Software für die OCR historischer Drucke

Wehner, Maximilian
maximilian.wehner@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Dahnke, Michael
michael.dahnke@uni-siegen.de
Universität Siegen, Deutschland

Landes, Florian
florian.landes@kbl.badw.de
Bayerische Akademie der Wissenschaften, Deutschland

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

Reul, Christian
christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Das Problemfeld der OCR früher Drucke

Lange galt die automatisierte Texterkennung oder sog. Optical Character Recognition (OCR) historischer Drucke des späten Mittelalters und der Frühen Neuzeit, das heißt die Überführung des gedruckten Textes in eine maschinenverarbeitbare Form, als sehr problematisch (Rydberg-Cox 2009). Die OCR moderner Texte wird dagegen auch aufgrund technischer Innovationen wie des zeilen- statt zeichenbasierten OCR-Ansatzes (Breuel et al. 2013) weithin als informatisch gelöstes Problem angesehen. Die teils höchst komplexen Layoutstrukturen von Inkunabeln und der bis zum Ende des 18. Jahrhunderts gedruckten Werke, ihr oft schlechter Erhaltungs- und Druckzustand sowie die Vielfalt und Varianz der in ihnen verwendeten Drucktypen stellen dagegen bis heute sogar den kommerziellen State of the Art der Texterkennungssoftware wie beispielsweise ABBYY FineReader[1] vor erhebliche Probleme. Auch die vermeintlich einfach gedruckten Frakturromane des 19. Jahrhunderts bereiten bei ihrer Überführung in eine E-Text-Variante immer wieder große Schwierigkeiten. Trotz der durch Bibliotheken und andere öffentliche Einrichtungen bereit gestellten, wachsenden Bestände bilddigitalisierter Vorlagen dieser Epochen ist darum der Umfang digitalisierter Texte nicht annähernd im selben Maß gewachsen, obwohl in den vergangenen Jahren bereits deutliche Fortschritte für die OCR vormoderner Drucke aufgezeigt werden konnten (Springmann / Lüdeling 2017).

Vor allem für die geistes- und kulturwissenschaftliche Editionsphilologie eröffnet sich auf diese Weise ein erhebliches Problemfeld, ist diese vor dem Hintergrund der Entwicklung hin zu immer mehr digitalen Editionen doch auf meist große Textmengen in digitaler Form angewiesen, die im besten Fall neben ihrer hohen Zeichengenauigkeit bereits Metainformationen über das gedruckte Ursprungsmedium aufweisen – zu denken wäre hier besonders an die Typisierung unterschiedlicher Layoutregionen (Überschriften, Marginalien, Bildbeischriften etc.) oder die Lesereihenfolge der einzelnen Layoutelemente des originalen Textes. Und auch mit Blick auf neuere Forschungsfelder innerhalb der Geisteswissenschaften und Digital Humanities (Text Mining, Sentiment Analysis usw.) sowie deren Bedarf an großen Textmengen zur Anwendung quantitativer Analyseverfahren stellt sich zunehmend die Frage nach Möglichkeiten einer OCR früher und vormoderner Drucke, die sowohl hohen Qualitätsansprüchen als auch einem entsprechenden Automatisierungsgrad genügt.

Werkzeuge, die diese Anforderungen erfüllen, sollten zudem frei verfügbar sein, sich einfach und selbstständig von einem informatisch nicht vorgeschulten Nutzerkreis auf einer einheitlichen Benutzeroberfläche bedienen lassen und die unterschiedlichen Submodule wie beispielsweise die Vorverarbeitung von Bilddateien, Möglichkeiten der Layouttypisierung sowie die eigentliche Zeichenerkennung integrativ zu einem kohärenten OCR-Workflow zusammenführen.

Am Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik der Julius-Maximilians-Universität Würzburg wurde deshalb die OCR-Software OCR4all[2] entwickelt, welche die genannten Notwendigkeiten in sich vereint und sich als erstes Programm überhaupt mit Blick auf die besonders herausfordernden Textgruppen direkt an Geisteswissenschaftler*innen richtet.

OCR-Workflow

Typischerweise gliedert sich ein OCR-Workflow in vier Hauptkomponenten (s. Abbildung 1). Im sog. **Preprocessing** werden die Originalbilder in Vorbereitung späterer Arbeits-

schritte binarisiert (Konvertierung des Ausgangsbildes in ein Schwarzweißbild) und gerade gestellt, um die nachfolgenden Arbeitsschritte zu erleichtern.

Abbildung 1: Hauptkomponenten eines typischen OCR-Workflows. Von links nach rechts: Originalbild, Preprocessing, Segmentierung, OCR, Nachkorrektur.

Während der **Segmentierung** erfolgt die Erkennung und Typisierung der Layoutbestandteile. Dazu werden zuerst die Text- und Nicht-Textregionen (Bilder, Bordüren etc.) unterschieden, optional die Textregionen anschließend als Haupttext, Überschriften, Marginalien etc. semantisch ausgezeichnet. Abschließend werden die Textregionen zur Vorbereitung der OCR in einzelne Zeilenbilder zerschnitten.

In einem dritten Schritt, der **OCR**, werden die identifizierten Bildzeilen durch die Anwendung von sog. Modellen in maschinenverarbeitbaren Text umgewandelt. Je nach Material können dazu entweder sog. gemischte Modelle verwendet werden, die mithilfe einer Vielzahl ganz unterschiedlicher, jedoch epochentypischer Werke erstellt wurden. Handelt es sich bei den zu bearbeitenden Werken hinsichtlich der Vielfalt und Varianz der in ihnen verwendeten Drucktypen sowie deren Erhaltungszustand jedoch um sehr spezifische Drucke, können sog. werkspezifische Modelle für die Erkennung erstellt und verwendet werden.

In der **Nachkorrektur** können die generierten maschinenverarbeitbaren Texte und Daten abschließend nachbearbeitet und korrigiert werden.

OCR4all orientiert sich in seinem Aufbau an den beschriebenen Hauptkomponenten eines OCR-Workflows, gliedert diese jedoch noch einmal in unterschiedliche Teilmodule. Der modulare Aufbau erlaubt dabei eine Einbindung und Verwendung bereits bestehender Softwarelösungen, die gemäß ihrer Stärken zu einem kohärenten OCR-Workflow kombiniert werden.

Grundsätzlich kann der Workflow vollautomatisch durchlaufen werden. Dennoch hat der Nutzer immer die Möglichkeit, korrigierend in jeden Teilschritt einzugreifen, um ein optimales Ergebnis zu garantieren, welches als Startpunkt des dann folgenden Teilschritts fungiert. Dafür können die für jedes Teilmodul vorgegebenen Einstellungen durch den Nutzer individuell angepasst werden.

Das Preprocessing erfolgt in OCR4all wie oben beschrieben. Dabei werden alle gängigen Eingabeformate für Bilddateien unterstützt. Dem schließt sich die Layouttypisierung mithilfe des Segmentierungstools LAREX[3] (s. Abbildung 2) an. Hier können werkspezifische Parameter zur Text- und Bildtypisierung festgelegt sowie zu erkennende Layoutregionen (Haupttext, Überschriften, Marginalien, Seitenzahlen etc.) definiert werden. Je nach Komplexität des vorliegenden Seitenlayouts ist nach einer automatischen Layouterkennung ein Eingriff in das vorliegende Ergebnis mittels unterschiedlicher Korrekturwerkzeuge möglich. Weiterhin kann in LAREX die Lesereihenfolge der Layoutbestandteile markiert werden, um den Lesefluss des Originals vorlagengetreu nachbilden zu können. Vor allem für die Verwendung des maschinenverarbeitbaren Textes in digitalen Editionen sind diese Funktionen unverzichtbar.

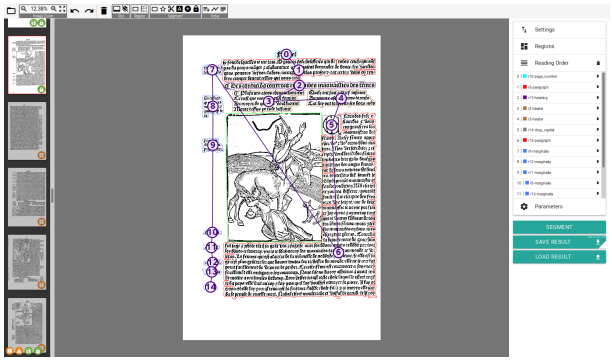


Abbildung 2: Im Teilmodul der Segmentierung erfolgen die Typisierung der Layoutelemente sowie die Festlegung der Lesereihenfolge.

Der Layouttypisierung folgt die Zeilensegmentierung. In dieser werden die Text beinhaltenden Layoutbestandteile in einzelne Zeilenbilder zerteilt (OCRopus[4]), um die eigentliche OCR vorzubereiten.

Anschließend wird im Erkennungsschritt aus den vorliegenden Einzelzeilen (mittels Calamari[5]) maschinenverarbeitbarer Text generiert. Dazu können in OCR4all bereits standardmäßig integrierte gemischte Modelle für Fraktur- und Antiquaschriften unterschiedlicher Epochen genutzt werden. Es besteht die Möglichkeit, die entstandenen Texte anschließend in einem Editor komfortabel zu korrigieren (s. Abbildung 3).

Abbildung 3: Im Editor kann generierter Text mithilfe eines sog. Virtual Keyboard (rechts) zeichengetreu korrigiert werden.

Für die Feststellung der Fehlerrate der Zeichenerkennung kann im Evaluationsmodul der ursprünglich erkannte Text mit der durch den Nutzer vorgenommenen Korrektur verglichen werden.

Darüber hinaus bietet OCR4all die Möglichkeit, die oben angesprochenen werkspezifischen Modelle unter Verwendung vorgenommener Textkorrekturen selbst zu trainieren, stetig zu verfeinern und anzuwenden. Besonders bei Werken mit erheblicher Typenvielfalt und -varianz, bei denen ein bestehendes gemischtes Modell keine hinreichenden Erkennungsergebnisse erzielt, können auf diese Weise dennoch sehr hohe Zeichenerkennungsraten erreicht werden.

In der abschließenden Nachkorrektur können die generierten Texte editionsreif korrigiert und als Plain Text oder PageXML[6] ausgegeben werden. Letzteres Format beinhaltet neben dem eigentlichen Text auch dessen Verankerung in semantischen Positionen auf den Druckseiten in Form von Koordinaten.

In Abhängigkeit des Ausgangsmaterials variiert der zum Erreichen einer sehr hohen Genauigkeit benötigte Arbeitsaufwand zwischen wenigen Minuten bei Werken mit einfachen Layoutstrukturen, für die ein passendes Modell vorliegt, und

einigen Stunden bei sehr komplexen, frühen Drucken, für die werkspezifische Modelle trainiert werden müssen (Reul et al. 2019).

Workshopkonzeption

Der ganztägige Workshop soll einem informatisch und technisch nicht spezifisch vorgeschulten Nutzerkreis einen nachvollziehbaren und verständlichen Einstieg in das Themen- und Problemfeld der OCR historischer Drucke bieten. Er wird dazu befähigen, mithilfe der vorgestellten Software eigenständig qualitativ hochwertige Texte aus ganz unterschiedlich anspruchsvollen Ausgangsdaten zu generieren – und dies mit zeitlich vertretbarem Aufwand. Die Konzeption erfolgt aus diesem Grund sehr praxisbezogen. Konkret bedeutet dies einen angeleiteten und individuell betreuten Durchgang durch den oben vorgestellten OCR-Workflow anhand verschiedener, nach Layoutkomplexität, Typographie, Erhaltungszustand und Entstehungszeitraum geclusterter Drucke. Dabei sollen anwendungsbezogen wichtige Grundfragen der OCR beantwortet werden:

- Wie verändert sich entsprechend des Ausgangsmaterials die Anwendung der OCR-Workflows und der in ihm enthaltenen Submodule?
- Mit welchem Aufwand ist in unterschiedlichen Bearbeitungsphasen des Materials zu rechnen?
- Wie stark lässt sich der Workflow in Abhängigkeit des vorliegenden Materials automatisieren?
- Wie schnell sind bei einem werkspezifischen Training welche Erkennungsraten erreichbar?
- Welcher Aufwand ist mit Blick auf die spätere Verwendung der produzierten Texte überhaupt sinnvoll?
- …

Da sich neben den oben beschriebenen, meist vormodernen Textspezifika auch eine grundlegende technische Expertise der Benutzer*innen im Bereich der OCR als eine wichtige Bedingung für die Produktion hochwertiger digitaler Texte herausgestellt hat, strebt der Workshop neben einer besonders praktischen Handlungsanleitung auch die Vermittlung der wichtigsten Funktionskonzepte der in OCR4all integrierten Submodule an.

Der Workshop umfasst neben den oben beschriebenen Inhalten auch Fragen der Einrichtung und Installation der Software. Zusätzlich wird eine Serverversion der Software zur Verfügung gestellt, die einen reibungslosen Ablauf gewährleistet und Trainingsprozesse werkspezifischer Modelle effizient durchführbar macht. Die max. 25 Teilnehmer*innen benötigen einen Laptop und Internetzugang. Die Verwendung einer Maus wird empfohlen.

Forschungsinteressen der Beitragenden

Maximilian Wehner ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik sowie am Zentrum für Philologie und Digitalität „Kallimachos“ der Julius-Maximilians-Universität Würzburg. Forschungsinteressen sind die Literatur der Frühen Neuzeit, die OCR früher Drucke sowie die Entwicklung entsprechender Vermittlungskonzepte.

Dr. Michael Dahnke arbeitet als Wissenschaftlicher Mitarbeiter am Zentrum für Informations- und Medientechnologie der Universität Siegen. Seine Forschungsschwerpunkte bewegen sich in den Bereichen digitaler Editionsphilologie, Datenmodellierung im Rahmen von TEI sowie der OCR und der Modellierung gewonnener Textdaten.

Florian Landes ist als Wissenschaftlicher Mitarbeiter bei der Bayerischen Akademie der Wissenschaften beschäftigt. Seine Forschungsinteressen liegen in den Bereichen der OCR sowie der digitalen Rekonstruktion.

Robert Nasarek ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschafts- und Sozialgeschichte der Martin-Luther-Universität Halle-Wittenberg sowie des Zentrums für Wissenschaftsforschung der Nationalen Akademie der Wissenschaften Leopoldina. Seine Arbeit bewegt sich im Bereich der Wirtschafts- und Sozialgeschichte, OCR und Digital Humanities.

Christian Reul ist Kommissarischer Leiter der Digitalisierungseinheit des Zentrums für Philologie und Digitalität „Kallimachos“ der Julius-Maximilians-Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR auf historischem Material sowie die Entwicklung von OCR-Software.

Fußnoten

1. https://www.abbyy.com/de-de/finereader/
2. https://www.uni-wuerzburg.de/zpd/ocr4all
3. https://github.com/OCR4all/LAREX
4. https://github.com/tmbdev/ocropy
5. https://github.com/Calamari-OCR/calamari
6. https://www.primaresearch.org/tools/PAGELibraries

Bibliographie

Breuel, Thomas M. / Ul-Hasan, Adnan / Al-Azawi, Mayce Ali / Shafait, Faisal (2013): High-Performance OCR for Printed English and Fraktur Using LSTM Networks, in: 12th International Conference on Document Analysis and Recognition: 683-687.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, in: *ArXiv Preprints (submitted to MDPI – Applied Sciences)* https://arxiv.org/abs/1909.04032.

Rydberg-Cox, Jeffrey A. (2009): Digitizing Latin Incunabula: Callenges, Methods, and Possibilities, in: *Digital Humanities Quarterly* 3, 1 http://digitalhumanities.org:8081/dhq/vol/3/1/000027/000027.html.

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.

Showtime – sehen und gesehen werden! Erzeugung semantischer (Spiel-)Räume für kollaboratives Arbeiten mit multimedialen Annotationen im Mehrdimensionalen

Wieners, Jan Gerrit
jan.wieners@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland

Schubert, Zoe
zoe.schubert@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland

Türkoğlu, Enes
enes.tuerkoglu@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland

Niebes, Kai Michael
kai.niebes@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland

Eide, Øyvind
oeide@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland

Zusammenfassung

Dieser Workshop soll seinen Teilnehmerinnen und Teilnehmern aufzeigen, wie sie für ihre Forschung relevante multimediale Objekte (3D-Modelle, Bilder, Sounds und Videos) für kollaboratives interdisziplinäres Arbeiten online bereitstellen, durch Annotationen mit Informationen anreichern und untereinander verknüpfen können. Einerseits steht dabei das Präsentieren der Objekte, andererseits das Sammeln von Informationen zu diesen im Vordergrund. In der praktischen Umsetzung wird hierfür das browserbasierte open-source Tool Kompakkt (https://kompakkt.de) eingesetzt, welches an der Universität zu Köln entwickelt und Mitte des Jahres 2019 veröffentlicht wurde. Es bietet seinen Nutzerinnen und Nutzern einen beinahe spielerischen Interaktionsraum, in dem die Objekte über einen modernen Webbrowser bereitgestellt, kollaborativ exploriert, erforscht und annotiert werden können. Kompakkt nutzt die 3D-Darstellung eines Objekts als Ausgangspunkt für das Sammeln heterogener Informationen, die durch den Einsatz von multimedialen Annotationen entstehen. Annotationen dienen dabei als flexible (Meta-)Daten, die die klassische Erfassung der Informationen erweitern.

Durch individuelle und geteilte Sammlungen von multimedialen Objekten ermöglicht die Software eine neuartige Lösung zum kollaborativen Sammeln und Erzeugen von Informationen. Mittels Annotationen können sowohl textuelle Beschreibungen als auch Objekte als Annotationsinhalt angefügt werden. Anhand dieser können entsprechend Verbindungen zwischen Objekten aufzeigt und Netzwerkstrukturen erstellt werden. Darüber hinaus hebt die Erstellung von Annotationen im dreidimensionalen Raum die zu einer Annotation zugehörige Perspektive auf eine neue Weise hervor: Annotationen sind nicht nur mit einer bestimmten Position im Raum in Relation zu einem Objekt, sondern auch mit der vom Nutzer oder von der Nutzerin gewählten Perspektive verknüpft. Das Festlegen der Reihenfolge von Annotationen eines Objekts wird in Kompakkt dazu genutzt, dass man sich von einer Annotation und der entsprechenden Perspektive zur einer anderen bewegen kann. Die resultierenden interaktiven Kamerafahrten implizieren dann die geführte Bewegung in der Zeit durch den Raum. Dies ermöglicht neue Wege der Präsentation bis hin zum annotationsbasierten Storytelling. Die beschriebene Funktionalität ermöglicht zudem das Erstellen von Bewegungspfaden in VR- und AR-Betrachtungen. Ein 3D-Objekt und dazugehörige Annotationen in der virtuellen oder erweiterten Realität betrachten zu können, bringt eine besondere Qualität in den Interaktionsraum, die es noch zu erforschen gilt. Unter anderem soll dies als eine der zentralen Ausgangsfragen des Workshops diskutiert werden. Das Ziel ist es hierbei, die Möglichkeiten der webbasierten 3D-Anwendungen zu evaluieren, während neue Erkenntnisse und Anwendungsbereiche aus interdisziplinärer Sicht generiert werden.

Finger weg! Damit spielt man nicht.

Physische Objekte können einen zentralen Anknüpfungspunkt für den Austausch und das Erzeugen von Wissen unterschiedlichster Art darstellen. Die Betrachtung eines Gegenstandes aus divergenten Perspektiven eröffnet dabei neue Blickwinkel und kann somit zum Erkenntnisgewinn beitragen. Dies gilt in Bezug auf die Darstellung räumlicher Verhältnisse, aber auch im bildungssprachlichen Sinn, und meint hier konkret multiple wissenschaftliche Perspektiven.

Eine physische Interaktion mit relevanten Objekten ist in vielen Fällen – insbesondere mit wertvollen historischen Artefakten – nicht möglich, da sie dabei beschädigt werden können. Für viele Arten von Objekten birgt dieser Umstand die Gefahr, dass sie ihren interaktiven Kontext verlieren, und zwar auch dann, wenn er ein wichtiger Bestandteil ihrer kulturellen Energien war. Dies trifft besonders zu, wenn der Untersuchungsgegenstand ein zur interaktiven Nutzung geschaffenes Objekt ist. Die Materialität des Objekts kann hiermit dazu beitragen, dass die Objekte zum einen de-kontextualisiert werden, zum anderen aus physikalischen Gründen nur

eingeschränkt erreichbar sind. Manuelle Rekonstruktionen kulturhistorischer Objekte sind oft zu teuer, und solche Rekonstruktionen sind – genauso wie die ursprünglichen Objekte –, nicht unbedingt jenseits ihres festgelegten eingenommenen physischen Standorts off-site erreichbar.

Nicht immer ist aber die Haptik und physische Präsenz eines Objektes für die Forschung und den Erkenntnisgewinn ausschlaggebend, sodass ein 3D Modell, welches das Objekt repräsentiert und abbildet, sowohl einen anderen Zugang als auch eine ähnliche Annäherung zu dem Objekt gewährleisten kann. Die Voraussetzung dafür ist, neben einer adäquaten Abbildung, die Fähigkeit zur effizienten Exploration eines solchen. Die Erstellung dreidimensionaler Objekte, die durch den Modellierungsprozess selbst zum Erkenntnisgewinn beitragen kann, ist dank der technischen Möglichkeiten mit immer geringerem finanziellem, arbeitsintensiven und zeitlichen Aufwand möglich; zeitgemäße Webtechnologien und Programmierschnittstellen (API) wie WebGL und WebXR ermöglichen zudem eine Darstellung multimedialer Objekte im Webbrowser ohne Installation eines externen PlugIns. Digitale 3D-Modelle können durch ihre Verfügbarkeit als Objekte im Netz eine deutliche größere Verwendergruppe erreichen.

Hands-on: Kompakkt

Der Fokus dieses Workshops ist auf die praktische Arbeit mit Kompakkt gerichtet. So lernen die Teilnehmerinnen und Teilnehmer der Veranstaltung, 3D-Objekte mitsamt ihrer Metadaten in das Objektrepositorium einzupflegen und online bereitzustellen. Zum anderen führt der Workshop ein in die Erstellung multimedialer Annotationen und geführter Touren durch interessante Aspekte des mit Kompakkt annotierten Objektes; die Teilnehmerinnen und Teilnehmer sind dazu eingeladen, eigene 3D-Objekte (falls vorhanden) einzusetzen. Zudem ist bereits eine Vielzahl von 3D-Objekten, die für unterschiedliche Disziplinen von Interesse sind, frei online und im Kompakkt-Repository verfügbar, die ihre Verwendung im Workshop finden können. Auch Bilder (ebenfalls annotierbar), Sounds und Videos sollen zum Einsatz kommen.

Multimediale Objekte werden von Kompakkt im dreidimensionalen Raum dargestellt, die Interaktionsmöglichkeiten werden dabei individuell auf das Objektmedium angepasst: 3D-Modelle lassen sich wie gewohnt auf x-, y- und z-Achse im kartesischen Koordinatensystem bewegen und rotieren, Rastergraphiken lassen sich horizontal und vertikal verschieben, Audioströme werden über einen interaktiven Platzhalter und Steuerelemente zugänglich gemacht. Zahlreiche Objekte, die in Kompakkt bereitgestellt werden, verweisen auf ein physisches Objekt (das Original). Andere Objekte sind ausschließlich digitaler natur und können nicht einmal in die Welt außerhalb von Computern und Projektionen übersetzt werden – Objekte wie z.B. CGI-Elemente, die in Kino, Theater oder anderen Kunstformen verwendet werden. Kompakkt akkumuliert solch unterschiedlichste Objekte und referenziert sie nicht ausschließlich, somit wird eine unmittelbare Interaktion mit den Objekten ermöglicht. Neben der Angabe allgemeiner standardisierter Metadaten, die sich auf das Objekt als Ganzes beziehen und über eine Schnittstelle zur Eingabe von Metadaten während des Upload-Prozesses gesammelt werden, besteht mit der Annotationsfunktionalität eine weitere Möglichkeit, den Datensatz zu erweitern.

Für das Erstellen einer Annotation wird durch einen Doppelklick auf die Oberfläche eines Objektes zunächst eine Markierung gesetzt. Der ausgewählte Punkt im Raum, relativ zum Objekt wird anschließend gespeichert und die weitere Bearbeitung der Annotation initialisiert. Der erstellte Referenzpunkt der Annotation wird in Kompakkt durch einen kleinen 2D-Kreis visualisiert. Dieser zeigt eine Zahl, die zusätzlich den Rang in der geordneten Liste von objektbezogenen Annotationen wiedergibt. Außerdem wird die durch die Nutzerin oder den Nutzer eingenommene Kameraeinstellung und Perspektive zum Zeitpunkt der Erstellung der Annotation erfasst und. Die wiederherstellbare Benutzerperspektive wird somit Teil der Annotation selbst und ist grundlegend für die Funktionalität, eigene sogenannte Walk-Throughs durch den virtuellen Raum zu erstellen. Dafür werden zusätzliche Steuerelemente im Benutzerinterface bereitgestellt, wenn ein Objekt über mehr als eine Annotation verfügt. Die Walk-Through Funktionalität ermöglicht es den Anwenderinnen und Anwendern, mit einer animierten Kamerafahrt von einer Annotation zur anderen zu navigieren und neben den entsprechend angefügten Informationen auch die verschiedenen Perspektiven der unterschiedlichen Annotationen zu explorieren.

Eine Annotation verfügt über einen Annotationstitel und einen Annotationsinhalt, der textuelle und multimediale Objekte aufnehmen kann – multimediale Objekte wie Texte, Bilder, 3D-Modelle oder Audiodateien aus dem Kompakkt-Repositorium oder Hyperlinks zu externen Webressourcen. Verfügt die Annotation über keinen Inhalt, so ist sie dennoch gebunden an die Nutzerperspektive und ermöglicht geführte Touren durch das Objekt, wie sie zuvor referiert wurden. Der Inhalt einer Annotation wird in einem HTML-Element dargestellt, das dynamisch neben der Markierung der Annotation positioniert wird. Selbst wenn sich die Kamera um das Objekt herum bewegt, werden der ausgewählte Punkt und der Körper der Annotation korrekt positioniert.

Als ein weiteres zentrales Feature der Webanwendung Kompakkt stellt sich die Möglichkeit dar, kollaborativ an Objekten zu arbeiten. So können Nutzerinnen und Nutzer andere Nutzer einladen, gemeinsam ein Objekt zu annotieren. Die Änderungen können von allen mitarbeitenden Benutzerinnen und Benutzern beobachtet werden. Derzeit wird an einem Feature gearbeitet, dass die kollaborierenden Nutzerinnen und Nutzer in Echtzeit visuell über Annotationen informiert, die von anderen online erstellt, bearbeitet oder entfernt werden. Registrierte Benutzer sind eingeladen, annotierbare Sammlungen von Objekten aus dem Repository zu erstellen. Eine Sammlung enthält alle relevanten Materialien, um mit der Annotation eines Objektes zu beginnen: Das Objekt selbst, darüber hinaus jedoch auch Objekte, die als Teil einer Annotation verwendet werden sollen. Der Zugriff auf eine Sammlung lässt sich individuell gestalten. So können Sammlungen privat, eingeschränkt sichtbar oder für alle Benutzer des Systems zugänglich sein. Neben bereits im Repository vorhandenen Objekten ist es Benutzerinnen und Benutzern möglich, eigene Objekte bereitzustellen, die in Annotationen verwendet werden können. Sowohl einzelne Objekte als auch Sammlungen lassen sich mit ihren Annotationen per Iframe auf externen Webseiten einbetten.

Mit seiner Kernfunktionalität bietet Kompakkt den Teilnehmerinnen und Teilnehmern des Workshops eine leicht zugängliche und leistungsstarke Anwendung, um multimediale Objekte bereitzustellen, kollaborativ zu annotieren und mit der Walk-Through Komponente eigene Narrationen zu realisieren – Narrationen, die den Fokus auf Aspekte des betrachteten Objektes lenken. Narrationen, die interessant und relevant in Forschung und Lehre sind.

Veranstaltungsdetails

- Dauer: ein halber Tag
- Maximale Teilnehmerzahl: 25
- Benötigte technische Ausstattung: Computer-Lab oder eigener Computer (Laptop) mit einem darauf installierten modernen Webbrowser (aktuelle Version von Chrome oder Firefox) und ein ständiger Internetzugang.

Kontaktdaten & Forschungsinteressen der Beitragenden

Institut für Digital Humanities
Universität zu Köln info@dh.uni-koeln.de
Albertus-Magnus-Platz
D-50931 Köln
+49 221 470-4430
https://idh.uni-koeln.de

Prof. Dr. Øyvind **Eide** (oeide@uni-koeln.de) ist Professor für Digitale Geisteswissenschaften an der Universität zu Köln. Er wurde am King's College London (2013) in Digital Humanities promoviert. Von 1995 bis 2013 war er als Mitarbeiter in verschiedenen Positionen an der Universität Oslo tätig und beschäftigte sich mit digitalen Geisteswissenschaften und der Informatik im Kontext des kulturellen Erbes. Von 2013 bis 2015 war er Dozent und wissenschaftlicher Mitarbeiter an der Universität Passau. Er war 2016–19 Vorsitzender der European Association for Digital Humanities (EADH) und engagiert sich zudem aktiv in mehreren internationalen Organisationen wie ICOM's International Committee for Documentation (CIDOC). Seine Forschungsinteressen konzentrieren sich auf transformative digitale Intermedia-Studien, wobei er die kritische schrittweise Formalisierung als Methode zur konzeptionellen Modellierung von Informationen über das kulturelle Erbe verwendet. Dies wird als Werkzeug für die kritische Auseinandersetzung mit medialen Unterschieden eingesetzt, insbesondere mit den Beziehungen zwischen Texten und Karten als Kommunikationsmedien. Er beschäftigt sich auch mit theoretischen Studien zur Modellierung in den Geisteswissenschaften und darüber hinaus.

Kai Michael **Niebes** (kai.niebes@uni-koeln.de) ist Software-Entwickler am Institut für Digital Humanities der Universität zu Köln. Seine Forschungsinteressen bestehen aus Machine Learning, moderner Webentwicklung und Datenbanktechnologien.

MA Zoe **Schubert** (zoe.schubert@uni-koeln.de) ist wissenschaftliche Mitarbeiterin, Software-Entwicklerin und Dozentin für Medieninformatik und Informationsverarbeitung am Institut für Digital Humanities der Universität zu Köln. Sie leitet dort das Projekt "Lehre in 3D" an, in dessen Kontext auch Kompakkt entstanden ist. Sie hat einen Master-Abschluss in Medienkulturwissenschaften und Medieninformatik (2013) und schreibt ihre Dissertation über "Virtuelle Realität als transformative Technologie in den Geisteswissenschaften - Theater in der virtuellen Realität". Ihre Forschungsinteressen umfassen mediale Transformation, Virtual and Augmented Reality, Visualisierung, Annotation, Modellierung in digitalen Geisteswissenschaften und Webtechnologien, sowie die Entwicklung von Anwendungen in diesen Bereichen.

BA Enes **Türkoğlu** (enes.tuerkoglu@uni-koeln.de) ist am Cologne Center for eHumanities, Institut für Digital Humanities und an der Theaterwissenschaftlichen Sammlung der Universität zu Köln tätig. Er hat an der Universität Istanbul Radio, TV und Kino studiert, 2009 kam er nach Deutschland, wo er in Köln seinen Bachelorabschluss in Medieninformatik absolviert hat. In seiner Forschung beschäftigt er sich mit der Digitalisierung heterogener Objektarten und der Relevanz ihres kulturellen Kontextes.

Dr. Jan Gerrit **Wieners** (jan.wieners@uni-koeln.de) ist wissenschaftlicher Mitarbeiter, Software-Entwickler und Dozent für Medieninformatik und Informationsverarbeitung am Institut für Digital Humanities der Universität zu Köln. Jan G. Wieners hat einen Magister Artium in Historisch-Kulturwissenschaftlicher Informationsverarbeitung (HKI), Germanistik und Philosophie und wurde an der Universität zu Köln über spielübergreifende künstliche Intelligenz in klassischen Brettspielen promoviert. Seine Forschungsinteressen umfassen virtuelle und augmentierte Realität, mediale Transformationen, Modellierung, Theorie und Praxis der digitalen Geisteswissenschaften, Künstliche Intelligenz, Computer Vision und Game Studies.

Spielplätze der Theoriebildung in den Digital Humanities

Geiger, Jonathan

jonathan.geiger@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Pfeiffer, Jasmin

jasmin.pfeiffer@uni-saarland.de
Universität des Saarlandes, Lehrstuhl für Neuere deutsche Literaturwissenschaft | Medienwissenschaft, Saarbrücken, Deutschland

Die Digital Humanities (DH) existieren als Forschungsfeld (wenn auch nicht unter diesem Label) bereits seit den 1940er Jahren, so man Roberto Busas Projekt des Index Thomisticus als Grundstein ansieht. Wechselt man von der wissenschaftshistorischen zu einer wissenschaftspolitischen Perspektive und betrachtet Faktoren wie z. B. Forschungsförderungen und Projekte, sind die DH erst seit gut 10–20 Jahren Teil der (deutschen) Wissenschaftslandschaft. Ein Ringen um Akzeptanz ist teilweise bis heute zu beobachten. Gleichwohl lässt sich sagen, dass die DH mit Kuhn gesprochen durchaus mittlerweile den Status einer Normalwissenschaft erlangt haben und sich im Produktivbetrieb befinden.

Typischerweise lässt sich in der wissenschaftlich-disziplinären Ontogenese nach der Etablierung einer Disziplin bzw. eines Forschungsfeldes der Übergang in eine neue Phase verzeichnen. Einerseits, weil sich durch den Produktivbetrieb

und das vermehrte Einbringen vergleichbarer wissenschaftlicher Erkenntnisse Fragen nach dem epistemischen Status, der Validität, der Verwertbarkeit und der weiterführenden Fragenentwicklung stellen, andererseits, weil der wissenschaftspolitische Legitimationsdruck abnimmt und dadurch Ressourcen frei werden und sich neue Handlungsspielräume eröffnen. In dieser Phase der Reife und Entwicklung von Selbst-bewusstsein befinden sich die DH derzeit.

Vereinzelte Beiträge zur theoretischen Reflexion der DH als solche, ihrer Objekte und Methoden, Diskussionen auf Twitter und Blogs, sowie Konferenzthemen (exemplarisch: der Titel der DHd-Konferenz 2018 "Kritik der digitalen Vernunft") stellen eindeutige Marker für diesen Befund dar. Es ist allerdings auch zu konstatieren, dass sich an der von Thiel in der FAZ 2012 vorgebrachten Kritik an der Theorielosigkeit der DH (https://www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html, zuletzt abgerufen am 27.09.2019) bislang wenig verändert hat. Denn obwohl die Relevanz der Theoriebildung für die DH schon verschiedentlich betont wurde und immer wieder auch Theoriebeiträge vorgelegt werden, sind derlei Überlegungen bislang eher nebenbei und wenig zentralisiert in einzelnen, unabhängigen Projekten oder projektlosen Einzelarbeiten angestellt worden. Ähnlich institutionalisierte Diskurse wie z. B. im angelsächsischen Raum die "Debates in the Digital Humanities" sucht man vergeblich. Im deutschsprachigen Raum entstanden so in der (bisweilen naiven) Akzentuierung der Entwicklung digitaler Werkzeuge klaffende Lücken in der Theoretisierung der Aktivitäten und Gegenstände, welche hinter der vordergründigen Methoden-Orientierung nicht sofort ins Auge springen. Da die Werkzeuge und Methoden ja einfach scheinbar "funktionieren", werden Fragen nach ihrem epistemologischen Status verhängnisvollerweise allzu leicht in die zweite Reihe gestellt.

Das Gebot der Stunde ist also die systematische wissenschaftlich-disziplinäre Selbstreflexion, Theoriebildung und epistemologische Positionierung der DH. Nachdrücklich sollte daher ein tiefgreifendes, akademisches Aufspüren jener Besonderheiten des Digitalen gefordert werden, die unter dem Signum des fundamentalen und allumfassenden Wandels einen Wissenschaftsbereich wie die DH nun schon seit einigen Jahrzehnten glaubwürdig rechtfertigen. Jene Suche sollte sich über die spezifischen Methoden der digitalen Transformation spannen. Was offenkundig fehlt sind z. B. informationstheoretische, kulturwissenschaftliche und philosophische Grundlagen und ein theoretisches Fundament, welches mit Hilfe einer flächendeckenden, systematischen Untersuchung die isolierten und verstreuten Ansätze sinnbringend und letztlich auch den einzelnen digitalen GeisteswissenschaftlerInnen Souveränität stiftend miteinander verknüpfen könnte. Insbesondere die Geisteswissenschaften, die sich durch ihre epistemische Sensibilität auszeichnen, besitzen das theoretische und methodische Rüstzeug um die unreflektierte Anwendung digitaler Werkzeuge und den naiven Glauben in digital konstruierte wissenschaftliche Erkenntnisse vermeiden bzw. überwinden zu können.

Der Workshop "Spielplätze der Theoriebildung in den Digital Humanities" möchte an genau dieser Stelle ansetzen. Es soll ein Impuls gesetzt werden, der sich auf mehreren Ebenen erstreckt: Mit der dezidierten Thematisierung der Theoretisierung der DH wird die Community für die Relevanz des Themas sensibilisiert und gleichermaßen wird der Status Quo bestimmt, inwiefern Interesse und Kapazitäten von Seiten der einzelnen ForscherInnen für dieses Thema bereits vorhanden sind. Dies dient auch als Grundlage für etwaige Verstetigungsansätze dieser Forschungsrichtung auf mittelfristiger Perspektive hin, z. B. in Form einer eigenen Zeitschrift oder einer AG in der DHd. Inhaltlich wird eine Kartographierung der Objekte, Perspektiven und Methoden als Teil einer kritischen Refraktion der Digital Humanities unternommen, sowie Ansätze zur wissenschaftlichen Selbstdeutung der DH (als Disziplin, Feld oder Hilfswissenschaft) gesammelt. Hierzu loten die Teilnehmenden gemeinsam die Spielräume wissenschaftstheoretischer Grundlagen und Arbeitsfelder aus und schaffen damit eine Basis für systematische Deutungen.

Der Workshop hat die Struktur eines World Cafés: Nach einer kurzen Begrüßung und Vorstellung des Programms im Plenum rotieren die Teilnehmende zwischen einer Reihe unterschiedlicher Themenfelder bzw. Thementische. So erhalten sie die Möglichkeit, sich innerhalb stetig aktualisierter Gruppenkonstellationen in Diskussionen über Perspektiven, Thesen und Themen der DH einzubringen. Insbesondere kontroverse sachliche Diskussionen sollen provoziert werden, um eine möglichst differenzierte und breite Grundlage für ein Theorienfundament zu schaffen. Die Diversität der beteiligten wissenschaftlichen Disziplinen auf dem Gebiet der DH, die möglicherweise in der Vergangenheit einer holistischen Theoriebildung der DH im Wege stand, wird dabei als Trumpfkarte gespielt. Für die Einigung auf eine gemeinsame Sprache und die Integration der Perspektiven werden im Rahmen des Workshops erste Ansätze konturiert und protokolliert. Moderierende an den Thementischen leiten die Gespräche, geben Denkimpulse, dokumentieren die Ergebnisse analog und digital und stellen diese abschließend dem gesamten Plenum vor. Eine zusammenfassende Reflexion der Ergebnisse und die Entwicklung eines thematischen Ausblicks runden den Workshop ab. Für die Moderation stehen zunächst die Einreichenden zur Verfügung. Sie werden ergänzt von verschiedenen einschlägigen KollegInnen, die an der Entwicklung des Workshops beteiligt waren (u. a. Patrick Sahle, Enes Türkoğlu und Rabea Kleymann). Nach der Bewilligung und Veröffentlichung des Workshops können sich außerdem noch weitere Interessenten für die Moderation einzelner Thementische melden und werden dann von den Organisatoren ausgewählt. Das Format des World Cafés ist für die Zielsetzungen des Workshops optimal, da die materialen Beiträge von Seiten der Teilnehmenden kommen, zudem kann ihre Heterogenität nicht nur aufgefangen, sondern produktiv genutzt werden. Die Unterteilung in Thementische gibt nur eine lockere Strukturierung vor und dient auch der Feststellung von Interessensprioritäten der Community. Weiterhin findet "am Rande" eine wechselseitige Identifikation und Vernetzung der Teilnehmenden statt.

Die Themeninseln sollen folgende Schwerpunkte haben:

- **Objekte der DH**: Aus geisteswissenschaftlicher Sicht stellen sich die Gegenstände der Informatik alles andere als selbstverständlich dar: Daten sind nicht "neutral", sondern bereits Interpretationen und Produkte von Forschungsprozessen und -methoden. Dasselbe gilt für Datenmodelle und letztlich auch für Algorithmen, deren Einfluss auf die Transformation von Daten für den geisteswissenschaftlichen Forschungsprozess selbstverständlich mitreflektiert werden muss. Aus informatischer Sicht sollte außerdem die Frage nach digitalen bzw. digitalisierten Objekten neu gestellt werden, welche durch Verfahren der technischen Reproduzierbarkeit notwendigerweise einen neuen ontologischen Status aufweisen.

- **Methoden der DH**: Die Forschungsgegenstände werden maßgeblich durch die Forschungsmethoden geprägt. Man könnte auch sagen, dass sie durch Forschungsmethoden erst als Gegenstände hervorgebracht werden. Eine Reflexion der Methoden ist daher unerlässlich und stellt sich nicht nur aus wissenschaftssoziologischer und wissenschaftspolitischer Perspektive im Hinblick auf Forschungsgelder, sondern auch aufgrund des neuen Zuganges, den die DH zu Forschungsgegenständen ermöglichen, z. B. in Form einer digitalen Hermeneutik und des Distant Readings.
- **Werkzeuge der DH**: Forschungsmethoden und Werkzeuge stehen in einem dialektischen Verhältnis zueinander. Software wird geformt von den Daten und den Datentransformationsaufgaben, die Daten hingegen werden strukturiert nach der verarbeitenden Software. Es ergeben sich Sachzwänge, deren Ausläufer sich bis hinein in das Research Software Engineering, die Prototypenkonzeption und Usability-Testing bemerkbar machen.
- **Medialität und Digitalität der DH**: Die Digitalität ist kein Phänomen der Geisteswissenschaften, sondern muss in einem größeren gesellschaftlichen Rahmen gedacht werden – die Digitalität der Kultur ist der Kontext einer Digitalisierung von Kultur. Logiken der Algorithmizität, Hyperreferentialität und technischer Performativität werden in die Forschung eingeschrieben und müssen bei einer Theorie der DH mitberücksichtigt werden (Beispiele: Informationstheorie geisteswissenschaftlicher Forschungsdaten, Transmedialisierung, Materialität des Digitalen).
- **Wissenschaftstheorie der DH**: Dies ist der bis dato wohl am prominentesten diskutierte Punkt, der sich auf das Verhältnis der DH zu den "klassischen" Geisteswissenschaften und der Informatik bezieht, sowie auf den Status der DH als eigenständige Disziplin, als Feld oder als Hilfswissenschaft. Es stellt sich die Frage, ob die DH eine eigene Wissenschaftstheorie brauchen oder befriedigend über etablierte Wissenschaftstheorien (z. B. von Fleck, Kuhn, Popper) beschrieben werden können. Der Theorienpluralismus und neue epistemische Forschungsdarstellungen sind hier ebenso zu diskutieren, wie das Problem der Inkommensurabilität, dass sich mit dem Semantic Web für Forschungsdaten neu stellt.
- **DH und Öffentlichkeit**: An das wissenschaftlich-disziplinäre Selbstverständnis der DH als Forschungsfeld schließen sich auch die Untersuchung des Verhältnisses zwischen DH und Öffentlichkeit an. Dies umfasst Fragen nach der Positionierung der DH im öffentlichen Diskurs rund um (geisteswissenschaftliche) Forschung, Fragen der Forschungsethik und Forschungsförderung, Open Access und Bürgerbeteiligung ("citizen science").

Der Workshop bietet damit Raum für verschiedene "Spielplätze" im Bereich der Theoriebildung der Digital Humanities: Spielräume des Theoretischen durchsetzen dann die Spielräume der Forschungspraxis und machen diese wissenschaftstheoretisch greifbar. Es lässt sich argumentieren, dass die DH gewappnet dafür sind, ein neues Kapitel ihrer jungen Wissenschaftsgeschichte zu schreiben. In diesem Sinne besteht die Hoffnung, dass der Workshop "Spielplätze der Theoriebildung in den Digital Humanities" durch die Zusammenführung interessierter WissenschaftlerInnen zur Initialzündung wird, nach der ForscherInnen gemeinsam und engagiert die Fundamentbildung der DH vorantreiben.

Interessierte ForscherInnen haben auch nach dem Workshop die Möglichkeit in Kontakt zu bleiben, nicht nur, weil die Ergebnisse des Workshops digital zur Verfügung gestellt werden, sondern auch weil die Organisatoren die Möglichkeit für weitere Kollaboration anbieten wollen. Glückt das Vorhaben des Workshops als Inkubator, ist eine Verstetigung und Institutionalisierung des Forschungsinteresses geplant, um einerseits eine offene Plattform des Austausches und der Diskussion zu bieten und andererseits um Forschungsergebnisse wieder in die Community (und auch die Öffentlichkeit) zurückzuspielen.

Einreichende:

- Jonathan D. Geiger, M. A.
  - Akademie der Wissenschaften und der Literatur | Mainz, Digitale Akademie
  - Geschwister-Scholl-Str. 2 in 55131 Mainz
  - Forschungsinteressen: (Sozial)Epistemologie, Wissenssoziologie, Philosophie der Digitalität, Digital Humanities, Theorie von Informatik, Informations- und Dokumentationswissenschaft
- Jasmin Pfeiffer, M. A.
  - Universität des Saarlandes, Lehrstuhl für Neuere deutsche Literaturwissenschaft | Medienwissenschaft
  - Campus, Gebäude A22, Raum 0.20, 66123 Saarbrücken
  - Forschungsinteressen: Theorie und Analyse des Computerspiels, Fiktionstheorien, Virtuelle Realitäten, Medialität und Materialität, Digitalität, Theorie der Algorithmen

Teilnehmende: max. 40
Anforderungen an die Raumausstattung:

- Beamer: Ja
- Tafel/Whiteboard: Nein
- Flipchart: Nein (aber Flipchart-Papierbögen, die dann an die Pinnwände geheftet werden können)
- Moderationskoffer: 6 Stück
- Pinnwand: 6 Stück
- Steckdosenleisten: 3 Stück
- weitere Anmerkungen:
  - Ein Laptop für den Beamer im Plenum wäre gut.
  - Ideal wäre die Möglichkeit neben dem Raum für das Plenum auch Zugang zu 1–3 weiteren (kleineren) Räumlichkeiten zu haben bzw. überhaupt Raum zum Ausweichen zu haben, sodass sich die Arbeitsgruppen etwas verteilen können.
  - Ein Bonus (wenn auch keine Notwendigkeit) wäre ein kleiner Stehtisch für jede Arbeitsgruppe, also vor jede Pinnwand (insgesamt 6 Stück).

Bibliographie

Bauer, J. (2011): 'Who are you Calling Untheoretical?', *Journal of Digital Humanities*, 1(1). Available at: http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/.

Brügger, N. (2016): 'Digital Humanities in the 21st Century: Digital Material as a Driving Force', *digital humanities quarterly*, 10(2). Available at: http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html.

Capurro, R. (1978): *Information: ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs*. München: Saur.

Capurro, R. (2017): *Homo Digitalis: Beiträge zur Ontologie, Anthropologie und Ethik der digitalen Technik*. Wiesbaden: Springer.

Castells, M. (2003): *Das Informationszeitalter: Wirtschaft, Gesellschaft, Kultur.* Wiesbaden: Springer.

Cecire, N. (2011): 'Theory and the Virtues of Digital Humanites', *Journal of Digital Humanities*, 1(1). Available at: http://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire/.

Ciula, A. / Eide, Ø. (2017): 'Modelling in digital humanities. Signs in context', *Digital Scholarship in the Humanities*, 32(1), pp. 33–46. DOI: https://doi.org/10.1093/llc/fqw045.

Dahlström, M. (2011): 'Critical Editing And Critical Digitisation', *Text Comparison and Digital Creativity*, (The Production of Presence and Meaning in Digital Text Scholarship). DOI: https://doi.org/10.1163/ej.9789004188655.i-328.29.

Deck, K.-G. (2018): 'Digital Humanities – Eine Herausforderung an die Informatik und an die Geisteswissenschaften', *Sonderband der Zeitschrift für digitale Geisteswissenschaften*, 3. DOI: 10.17175/sb003_002.

Flanders, J. / Jannidis, F. (2019): *The shape of data in the digital humanities. Modeling texts and text-based resources.* London, New York: Routhledge (Digital research in the arts and humanities).

Floridi, L. (2013): *The philosophy of information*. Oxford: Oxford Univ. Press.

Frabetti, F. (2015): *Software theory: a cultural and philosophical study*. London; New York: Rowman & Littlefield International (Media philosophy).

Friedewald, M. / Leimbach, T. (2011): 'Computersoftware als digitales Erbe: Probleme aus Sicht der Technikgeschichte', in *Neues Erbe. Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung.* KIT Scientific Publishing.

Gius, E. / Jacke, J. (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.

Gnadt, T. et al. (2017): 'Faktoren und Kriterien für den Impact von DH-Tools und Infrastrukturen'. DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek. Available at: http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2017-21.pdf.

Hall, G. (2012): "*Blog Post: Has Critical Theory Run Out of Time for Data-Driven Scholarship?*, *Debates in the Digital Humanities*. Available at: https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/1a9b138c-eb51-4f48-bcb8-039505f88ff8#p2b2 (Zugriff: 19. August 2019).

Heßbrüggen-Walter, S. (2018): 'Philosophie als digitale Geisteswissenschaft', *Sonderband der Zeitschrift für digitale Geisteswissenschaften*, 3. DOI: 10.17175/sb003_006.

Hui, Y. (2016): *On the existence of digital objects*. London: University of Minnesota Press.

Kaden, B. (2016): 'Zur Epistemologie digitaler Methoden in den Geisteswissenschaften', *Berliner Beiträge zu Digital Humanities*.

Koch, G. (ed.) (2017): *Digitalisierung. Theorien und Konzepte für die empirische Kulturforschung.* Köln: Herbert von Halem Verlag.

Matzner, T. (2016): 'Beyond data as representation. The performativity of Big Data in surveillance.', *Surveillance & Society*, 14(2), pp. 197–204.

McCarty, W. (2014): 'Getting there from here. Remembering the future of digital humanities: Roberto Busa Award lecture 2013', *Literary and Linguistic Computing*, Volume 29(Issue 3), pp. 283–306. DOI: 10.1093/llc/fqu022.

Mohabbat Kar, R. / Parycek, P. (2018): 'Berechnen, ermöglichen, verhindern: Algorithmen als Ordnungs- und Steuerungsinstrumente in der digitalen Gesellschaft', in *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft*. Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT), pp. 7–39. Available at: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-57562-7.

Nassehi, A. (2019): *Muster. Theorie der digitalen Gesellschaft.* München: Beck Verlag.

Nerbonne, J. (2015): 'Die Informatik als Geisteswissenschaft', *Sonderband der Zeitschrift für digitale Geisteswissenschaften*, 1(1). DOI: 10.17175/sb001_003.

Porter, D. (no date): *The Uncanny Valley and the Ghost in the Machine: a discussion of analogies for thinking about digitized medieval manuscripts*, *Dot Porter Digital*. Available at: http://www.dotporterdigital.org/the-uncanny-valley-and-the-ghost-in-the-machine-a-discussion-of-analogies-for-thinking-about-digitized-medieval-manuscripts/.

Reiche, R. et al. (2014): 'Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften'. (DARIAH-DE Working Papers), (4).

Sahle, P. (2015): 'Digital Humanities? Gibt's doch gar nicht!', *Grenzen und Möglichkeiten der Digital Humanities*. DOI: 10.17175/sb001_004.

Scheinfeldt, T. (2012): "*Blog Post: Where's the Beef? Does Digital Humanities Have to Answer Questions?*, *Debates in the Digital Humanities*. Available at: https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/3c03ecdb-2dcf-4597-8fc4-e42f8dcc21e1 (Zugriff: 19 August 2019).

Schröter, J. / Böhnke, A. (eds) (2004): *Analog/Digital – Opposition oder Kontinuum? Zur Theorie und Geschichte einer Unterscheidung.* Bielefeld: transcript Verlag.

Stalder, F. (2016): *Kultur der Digitalität*. Berlin: Suhrkamp.

Türkoglu, E. (2019): 'Vom Digitalisat zum Kontextualisat – einige Gedanken zu digitalen Objekten', in. *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, Frankfurt am Main. DOI: 10.5281/zenodo.2600812.

Wettlaufer, J. (2016): 'Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern', *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2016_011.

Vom Phänomen zur Analyse – ein CRETA-Workshop zur reflektierten Operationalisierung in den DH

Ketschik, Nora
nora.ketschik@ilw.uni-stuttgart.de
Universität Stuttgart

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

Murr, Sandra
sandra.murr@ts.uni-stuttgart.de
Universität Stuttgart

Pagel, Janis
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart

Reiter, Nils
nils.reiter@uni-koeln.de
Universität zu Köln; Universität Stuttgart

Einleitung

Der Workshop adressiert eine der großen Herausforderungen für Arbeiten in den Digital Humanities – die Operationalisierung geisteswissenschaftlicher Konzepte und Fragestellungen für computergestützte Methoden (vgl. Jannidis 2010, 109–132; Moretti 2013; Flanders, Jannidis 2015; Jacke 2014, 118–139). Während Geisteswissenschaftler vor allem mit komplexen, häufig textübergreifenden Phänomenen arbeiten und als relevant erachtete Kontexte der behandelten Themen heranziehen, ist die computergestützte Arbeit an identifizierbare Phänomene auf der Textoberfläche gebunden. Die hieraus erwachsende Diskrepanz zwischen Erwartungen und Ergebnissen gilt es über eine adäquate Operationalisierung, **also eine Messbarmachung theoretischer Konzepte**, zu überbrücken. Mit unserem Workshop wollen wir genau diese Schnittstelle in den Fokus rücken. Anhand dreier Anwendungsfälle zeigen wir auf, welche Herausforderungen sich aus dem Einsatz computergestützter Methoden für geisteswissenschaftliche Zwecke ergeben und wie mit ihnen umgegangen werden kann. In einem praktischen Teil haben die Teilnehmenden die Möglichkeit, selbst an der Operationalisierung eines Phänomens zu arbeiten; hierfür stellen wir Anwendungsfälle mit geeigneten Tools und Technik-„Baukästen“ zur Verfügung. Programmierkenntnisse werden dabei nicht vorausgesetzt. Ziel des Workshops ist es, das Bewusstsein für die Differenzen zwischen geisteswissenschaftlicher und computergestützter Arbeitsweise zu schärfen, typische Herausforderungen zu adressieren und Herangehensweisen zur Operationalisierung geisteswissenschaftlicher Phänomene aufzuzeigen. Denn nur durch die reflektierte Auseinandersetzung mit den Operationalisierungsannahmen kann ein angemessener (also reflektierter) Umgang mit den Ergebnissen gewährleistet werden.

Use Cases

Als Anwendungsfälle stellen wir drei unterschiedliche literatur- und sozialwissenschaftliche Phänomene vor, zu denen wir im Rahmen des Stuttgarter „Center for Reflected Text Analytics“ (CRETA)[1] umfangreiche Erfahrungen gesammelt haben. Die gewählten Beispiele decken verschiedene Aufgabentypen ab: Wir behandeln erstens die Extraktion bestimmter Instanzen aus einem Text, zweitens die Segmentierung eines Textes und drittens ein holistisches Textphänomen.

Entitäten und Entitätenreferenzen

Zum einen befassen wir uns mit dem Konzept der Entität und ihrer Referenz in literatur- und sozialwissenschaftlichen Texten (vgl. Reiter u.a. 2017, 19–22; Blessing u.a. 2017). Als Entitätenreferenzen gelten alle Ausdrücke, die auf eine Entität der realen oder fiktiven Welt referieren. Dazu zählen Personen/Figuren, Orte, Organisationen sowie Ereignisse, so dass das Konzept der Entität bewusst weit gefasst und für verschiedene Forschungsfragen anschlussfähig ist. Auf Entitäten kann auf verschiedene Weise referiert werden, u.a. über Eigen- und Gattungsnamen (z.B. “Angela Merkel”, “die Kanzlerin”). Um Entitäten in einem Text zu extrahieren, müssen folglich die Entitätenreferenzen annotiert und kookkurrente Ausdrücke aufgelöst werden. Die Herausforderungen bestehen vor allem in der Festlegung der Referenzausdrücke (welche Ausdrücke werden berücksichtigt?), in der Abgrenzung von Entitätenreferenzen gegenüber Generika sowie im Umgang mit Verschachtelungen, Metonymien und textspezifischen Besonderheiten. Am Beispiel zweier Textsorten (mhd. Artusroman und Bundestagsdebatten) stellen wir das Phänomen und Möglichkeiten der Umsetzung vor.

Erzählebenen

Des Weiteren beschäftigen wir uns mit der Annotation von Erzählebenen.[2] Hierbei geht es formal darum, einen Text in sinnvolle Segmente zu zerlegen, die seriell aneinandergereiht oder ineinander verschachtelt sein können. Auch wenn das narratologische Konzept ‘Erzählebene’ recht klar definiert erscheint, wird das Phänomen je nach theoretischer Grundlage unterschiedlich aufgefasst und analysiert (vgl. Genette 1988 [1983]; Ryan 1991). Um eine intersubjektive Annotation von Erzählebenen zu erreichen, gilt es deshalb zunächst, einen gemeinsamen Konsens zu theoretischen Grundannahmen zu finden. Ferner macht es die Operationalisierung von Erzählebenen notwendig, das vage Konzept akkurat zu formalisieren

und distinktive Merkmale zu bestimmen, die das Phänomen sinnvoll abgrenzen können.

"Wertherness"

Als dritten Anwendungsfall stellen wir die sog. "Wertherness" vor, womit eine Sammlung von Texteigenschaften gemeint ist, die Texte als "Wertheriaden" identifizieren können. Die Veröffentlichung von Goethes "Die Leiden des jungen Werthers" 1774 zog eine Reihe an literarischen Adaptationen nach sich, die sich durch verschiedene Bezugnahmen auf den Originaltext als sog. Wertheriaden ausweisen. Die Referenzen können dabei sowohl formaler (z.B. Briefroman, Dreiecksbeziehung) als auch inhaltlicher (z.B. Rolle der Natur, Verhältnis Subjekt-Gesellschaft) Art sein. Für eine computergestützte Analyse solcher Referenztexte müssen einerseits die einzelnen formalen und semantischen Kategorien operationalisiert und in den Texten identifiziert werden, andererseits ist zu untersuchen, welche Kriterien in bekannten Wertheriaden in Kombination miteinander auftreten.

Ansätze zur Operationalisierung

Im Workshop stellen wir zwei Ansätze zur Operationalisierung vor, die sich – in verschiedenen Phasen des Forschungsprozesses – sehr gut gegenseitig ergänzen. Der erste Ansatz besteht dabei in der Schärfung von **Konzeptdefinitionen durch Annotationen** und richtet sich an Menschen. Die Ergebnisse sind also keine Skripte oder Funktionen, sondern klare(re) Definitionen der fraglichen Konzepte, die von Menschen mit größerer intersubjektiver Übereinstimmung umgesetzt werden können, aber auch die theoretische Diskussion bereichern (vgl. Gius/Jacke, 2017; Pagel et al., 2018; Reiter et al., im Erscheinen). Daneben führt der Annotationsprozess auch zu einer intensiven und kritischen Beschäftigung mit dem Material und den textuellen Instanzen des Konzeptes und liefert damit auch Ideen für eine computergestützte Operationalisierung.

Als zweiten Ansatz stellen wir die Idee vor, Zielphänomene **indirekt zu operationalisieren**. Hierbei werden pro Phänomen mehrere messbare Eigenschaften in den Blick genommen, die mit dem Zielkonzept verwandt, aber nicht deckungsgleich sind. Aufschlussreich ist dabei in erster Linie nicht die Inspektion einzelner Eigenschaften, sondern die Gesamtschau der verschiedenen Einflussfaktoren (vgl. "instrumental variables" in Sack, 2011; "indirekte Operationalisierung" in Reiter/Willand, 2018). Bei textbasierten Phänomenen können so insbesondere linguistische und strukturelle Eigenschaften betrachtet werden, die größtenteils mit großer Reliabilität automatisch extrahierbar sind.

Ablauf

In einem Theorieteil führen wir in die Problematik der Operationalisierung von geisteswissenschaftlichen Phänomenen für die computergestützte Analyse ein. Anhand der drei oben genannten Beispiele aus der CRETA-Praxis thematisieren wir die Problematik und stellen die Ansätze der Operationalisierung im Detail vor. Je nach Interesse kann anschließend einer dieser Anwendungsfälle ausgewählt und bearbeitet werden.

Im praktischen Teil des Workshops haben die Teilnehmenden die Möglichkeit, beide Operationalisierungsansätze an ihrem gewählten Anwendungsfall zu erproben. Hierfür befassen sie sich zunächst mit dem Phänomen, indem sie es anhand eines Textauszugs manuell annotieren und parallel stichpunktartig die Richtlinien schärfen. In einer ersten Diskussionsrunde werden die verschiedenen Ergebnisse gesammelt und diskutiert. Zur Erprobung des zweiten Ansatzes stellen wir für jeden Anwendungsfall einen Operationalisierungs-„Baukasten" vor. Dieser besteht aus einer Sammlung von Python-Skripten in einem Jupyter-Notebook[3], die auf das jeweilige Untersuchungsvorhaben zugeschnitten ist und den Teilnehmenden die Möglichkeit gibt, sich dem zu untersuchenden Phänomen über computergestützte Verfahren anzunähern. Die Teilnehmenden können in Kleingruppen in diesem Baukasten verschiedene Parameter einstellen sowie manuell Eigenschaften an- oder abwählen, wobei sie auf ihr Vorwissen über den Untersuchungsgegenstand aus der ersten Praxisrunde zurückgreifen (können). Nachdem die Teilnehmenden die Eigenschaften ausgewählt und ggf. parametrisiert haben, können sie die Ergebnisse visualisieren und mit den Texten abgleichen. Damit erhalten die Teilnehmenden ein direktes Feedback zu den ausgewählten Parametern und können prüfen, ob das Untersuchungsvorhaben mit den festgelegten Einstellungen angemessen umgesetzt wird. Der Baukasten ist zur iterativen Nutzung vorgesehen, so dass der Einfluss verschiedener verwandter Eigenschaften auf die Ausgaben sichtbar wird und die Teilnehmenden sich einer geeigneten technischen Umsetzung sukzessiv annähern können. In einer abschließenden Diskussion werden die Ergebnisse gesammelt und es wird ausgewertet, wie adäquat sich die jeweiligen Zielphänomene mittels der gewählten Annahmen haben abbilden lassen.

Lernziele

Ziel unseres Workshops ist es, die Teilnehmenden für die Wichtigkeit der Operationalisierung in den Digital Humanities zu sensibilisieren und ihnen Lösungsangebote vorzustellen. Durch die interdisziplinäre Ausrichtung von DH-Arbeiten kommt der Operationalisierung eine Schlüsselposition zu, indem diese eine Brücke zwischen geisteswissenschaftlichem Phänomen und computergestützter Umsetzung schlägt. Mit den gewählten Anwendungsfällen wollen wir den Teilnehmenden ein "Repertoire" für die Operationalisierung verschiedener Aufgabentypen mitgeben. Wir zeigen zum einen, dass die Annotation eines Phänomens als Methode seiner Operationalisierung dienen kann (vgl. Gius, Jacke 2017, 233–254); zum anderen führen wir für textbasierte Phänomene eine approximative Operationalisierung ein (vgl. Reiter/Willand, 2018). Beide Verfahrensweisen sind auf andere Anwendungsfälle übertragbar. Gleichzeitig möchten wir deutlich machen, dass es für jedes Untersuchungsvorhaben nicht nur eine, sondern verschiedene Wege der Operationalisierung gibt. Die Spielräume, die bei der Operationalisierung geisteswissenschaftlicher Fragestellungen entstehen, machen es notwendig, Entscheidungen reflektiert zu treffen, sie offenzulegen und ihren Einfluss auf die Ergebnisse als Voraussetzung für eine angemessene Interpretation zu bedenken.

Abgrenzung zum CRETA-Hackatorial "Maschinelles Lernen lernen"

Neben diesem Workshop zur Operationalisierung wird noch ein weiterer Workshop des Stuttgarter DH-Zentrums CRETA während der diesjährigen DHd-Konferenz stattfinden (Gerhard Kremer, Kerstin Jung: "Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse"). Auch wenn es eine gewisse Schnittmenge zwischen den Workshops gibt (Textgrundlagen, Anwendungsfälle), ist die jeweilige Zielsetzung grundsätzlich verschieden: Während es beim CRETA-Hackatorial um Verfahren des Maschinellen Lernens geht, konzentriert sich der hier vorgestellte Workshop auf den grundsätzlicheren Schritt der Operationalisierung. Es geht also darum, Ansätze aufzuzeigen, wie ein Untersuchungsvorhaben oder theoretisches Konzept überhaupt für die computergestützte Analyse "vor- bzw. aufbereitet" werden kann. Beide Workshops ergänzen einander sinnvoll, was die Teilnahme an beiden oder an nur einem der Workshops möglich macht.

Anhang

Zeitplan

(insgesamt 3 Stunden + 30 Min. Pause)

1. Einführung und Ablauf (10 Min.)
2. Theoretischer Teil (insgesamt 40 Min.)
   - Erläuterung der Problemstellung
   - Vorstellung der drei Anwendungsfälle
3. Praktischer Teil
   - Einführung in die Primärtexte und Tools, Ausgabe der skizzierten Guidelines (10 Min.)
   - Erste Praxisrunde (Kleingruppen): Manuelle Annotation eines Phänomens, parallele Erweiterung/Überarbeitung der Guidelines, iterativ (30-40 Min.)

     - Kaffeepause (30 Min.) -

   - Sammeln der Ergebnisse und Diskussion der Herangehensweisen (20 Min.)
   - Zweite Praxisrunde (Kleingruppen): Arbeit am Operationalisierungsbaukasten, Feedback über Ausgabedatei, iterativ (30-40 Min.)
4. Abschlussdiskussion: Sammeln der „Ergebnisse", Diskussion der Erfahrungen und Lernziele (30 Min.)

Zahl der möglichen Teilnehmer

Zwischen 15 und 25.

Angaben zur technischen Ausstattung

Abgesehen von Beamer und ausreichend Steckdosen ist keine besondere technische Ausstattung erforderlich. Die Teilnehmenden arbeiten im praktischen Teil an ihrem eigenen PC. Informationen zu eventuellen Vorab-Installationen werden rechtzeitig mitgeteilt.

Beitragende

Der Workshop wird von Mitarbeitenden des "Center for Reflected Text Analytics" (CRETA) der Universität Stuttgart veranstaltet, die bereits erfahrene Workshop-Leiter/-innen im DH-Bereich sind (DHd 2017, DH 2017, DHd 2018, ESU 2018, DHd 2019, HCH 2019).

Das BMBF-geförderte eHumanities-Zentrum CRETA ist auf die interdisziplinäre Zusammenarbeit von Literaturwissenschaft, Linguistik, Philosophie und Sozialwissenschaft mit Maschineller Sprachverarbeitung und Visualisierung ausgerichtet. Die übergreifende Zielsetzung besteht in der Erarbeitung systematischer und transparenter Workflows, in denen die Entwicklung komputationeller Modelle und Methoden kritisch reflektiert und adäquat auf die unterschiedlichen geistes- und sozialwissenschaftlichen Forschungsfragen angepasst wird.

Nora Ketschik
nora.ketschik@ilw.uni-stuttgart.de
Universität Stuttgart
Institut für Literaturwissenschaft, Abt. für Germ. Mediävistik
Keplerstraße 17
70174 Stuttgart

Nora Ketschik ist Promotionsstudentin in der Abteilung für Germanistische Mediävistik. Im Rahmen von CRETA führt sie Netzwerkanalysen zu ausgewählten mittelhochdeutschen Romanen durch und setzt sich dabei kritisch mit der Verwendung computergestützter Methoden für literaturwissenschaftliche Analysezwecke auseinander.

Benjamin Krautter
Benjamin.Krautter@ilw.uni-stuttgart.de
Keplerstraße 17
70174 Stuttgart

Benjamin Krautter ist Promotionsstudent in der Abteilung für Neuere Deutsche Literatur II und Mitarbeiter im Projekt QuaDramA - Quantitative Drama Analytics. Dort arbeitet er an der Operationalisierung Aristotelischer Kategorien für die quantitative Dramenanalyse. Er beschäftigt sich zudem mit der Integration quantitativer Methoden in literaturwissenschaftliche Fragestellungen (*scalable reading*).

Sandra Murr
sandra.murr@ts.uni-stuttgart.de
Universität Stuttgart
Institut für Literaturwissenschaft, Abt. für Neuere Deutsche Literatur I
Keplerstraße 17
70174 Stuttgart

Sandra Murr ist Promotionsstudentin in der Abteilung für Neuere Deutsche Literatur I. In CRETA arbeitet sie an der digitalen Analyse des "Wertheriaden-Korpus", Texte, die in der Folge von Goethes "Werther" seit 1774 erschienen sind. Mittels computergestützter Verfahren wird sich mit der Frage auseinandergesetzt, anhand welcher charakteristischer Kriterien eine "Wertheriade" als solche definiert wird und wie sich entsprechende strukturelle und inhaltliche Kriterien operationalisieren, in den Texten automatisch identifizieren und reflektiert vergleichen lassen.

Janis Pagel
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Janis Pagel ist Promotionsstudent am Institut für Maschinelle Sprachverarbeitung und Mitarbeiter im QuaDramA-Projekt. Er forscht zu Anwendungen von computerlinguistischen Methoden auf literaturwissenschaftliche Fragestellungen und innerhalb von CRETA hauptsächlich zu Koreferenzresolution für literarische Texte.

Nils Reiter
nils.reiter@uni-koeln.de
Institut für Digital Humanities
Universität zu Köln
Albertus-Magnus-Platz
50931 Köln

Nils Reiter hat Computerlinguistik/Informatik an der Universität des Saarlandes studiert, wurde 2013 an der Uni Heidelberg promoviert und ist seit 2014 Post-Doc am Institut für Maschinelle Sprachverarbeitung. Seit seiner Promotion ist er im Bereich Digital Humanities unterwegs, mit einem besonderen Interesse an Fragen der Operationalisierung, und zwar sowohl im Hinblick auf Automatisierung wie auch auf manuelle Annotation. Er arbeitet dabei auch an praktischen Fragen der Kooperation zwischen Geistes- und Computerwissenschaftler*innen, und organisiert einen shared task zur Erkennung von Erzählebenen. Derzeit ist er Vertretungsprofessor für Sprachliche Informationsverarbeitung/Digital Humanities an der Universität zu Köln.

Fußnoten

1. www.creta.uni-stuttgart.de
2. Bei der Umsetzung des Konzepts wurde auf Vorarbeiten des Shared Tasks "SANTA" (Systematic Analysis of Narrative Texts through Annotation) zurückgegriffen, https://sharedtasksinthedh.github.io/. Das Material ist veröffentlicht in Reiter u.a. (2019).
3. https://jupyter.org

Bibliographie

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.

Julia Flanders / Fotis Jannidis (2015): Knowledge Organization and Data Modeling in the Humanities, .

Gérard Genette (1988 [1983]: Narrative Discourse Revisited. (Translated by Jane E. Lewin), Ithaca.

Marie-Laure Ryan (1991): Possible Worlds, Artificial Intelligence and Narrative Theory, Bloomington, Indianapolis.

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

Janina Jacke (2014): Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology, in: Journal of Literary Theory 8, S. 118–39.

Fotis Jannids (2010): Methoden der computergestützten Textanalyse, in: Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen, hg. v. Vera Nünning, Ansgar Nünning und Irina Bauder-Begerow, Stuttgart, Weimar, S. 109–132.

Franco Moretti (2013): "Operationalizing": or, the function of measurement in modern literary theory, in: Literary Lab 6, S. 1–13.

Janis Pagel / Nils Reiter / Ina Rösiger / Sarah Schulz (2018): A Unified Text Annotation Workflow for Diverse Goals, in: Proceedings of the Workshop for Annotation in Digital Humanities (annDH), hg. v. Sandra Kübler und Heike Zinsmeister, Sofia, Bulgaria, August 2018, S. 31–36.

Nils Reiter / Evelyn Gius / Marcus Willand (Hrsg.) (2019): A Shared Task for the Digital Humanities. Special issue of Cultural Analytics. November 2019.

Nils Reiter / Marcus Willand (2018): Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse, in: Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven, hg. v. Toni Bernhart, Marcus Willand, Sandra Richter und Andrea Albrecht, Stuttgart, S. 45–76.

Nils Reiter / André Blessing / Nora Echelmeyer / Gerhard Kremer / Steffen Koch / Sandra Murr / Maximilian Overbeck / Axel Pichler (2017): CUTE: CRETA Unshared Task zu Entitätenreferenzen, in: DHd 2017 Bern, Conference Abstracts, S. 19–22.

Graham Alexander Sack (2011): Simulating Plot: Towards a Generative Model of Narrative Structure, in: Papers from the AAAI Fall Symposium (FS-11-03).

Panels

Altbausanierung mit Niveau – die Digitalisierung gedruckter Editionen

Kontext

Während das Buch immer noch den höchsten Stellenwert in der geisteswissenschaftlichen Forschung im Allgemeinen besitzt, sind Editionen, die als Buch erscheinen, seit Jahren rückläufig (Eggert 2009). Bestehende Druckeditionen wirken mittlerweile neben ihren digitalen Nachfolgerinnen wie Relikte aus einer anderen Zeit. Ihr wissenschaftlicher Wert bleibt weitestgehend in den Grenzen des Buches verhaftet, während der digitale Editionskosmos wächst und perspektivisch zu einem dichten Wissensnetz wird. Um Druckeditionen besser verfügbar zu machen, sie mit anderen Editionen zu vernetzen, oder einen neuen Blick auf die Quellen zu ermöglichen, häufen sich in den letzten Jahren Unternehmungen zur Digitalisierung von Druckeditionen.[1]

Die mit der Digitalisierung von Editionen verbundenen, generalisierbaren Anforderungen und Implikationen sind, trotz ihrer unmittelbaren Relevanz für den Bereich der Digitalen Editionen, bisher noch nicht systematisch und projektübergreifend untersucht worden. Da bis dato zudem kaum auf die zahlreichen Erfahrungen bestehender Digitalisierungsprojekte zurückgegriffen werden kann, existiert sowohl bei laufenden als auch neuen Projekten stets die Gefahr, dass die organisatorischen, konzeptionellen und technischen Herausforderungen unterschätzt oder gar nicht erst erkannt werden. So entpuppen sich bspw. Projekte, die zunächst mit geringem Aufwand umsetzbar scheinen, nicht selten als Mammutaufgaben, die in Bezug auf Komplexität und Ressourcenbedarf die Anforderungen vergleichbarer *born digital*-Editionen teils deutlich übersteigen können.

Aus wissenschaftstheoretischer Perspektive stellt sich die Frage, welchen Stellenwert digitalisierte Editionen im Kosmos digitaler Editionstypen einnehmen können, wenn sie, wie Sahle formuliert, gar keine digitalen Editionen sind (Sahle 2013: 58ff.). In diesem Spannungsfeld gilt zu diskutieren, wie gedruckte editorische Leistungen der Vergangenheit unter den neuen medialen Bedingungen methodisch angemessen transformiert und für die Zukunft gesichert werden können.

Konzeption des Panels

Das Panel richtet sich als Forum für den Erfahrungsaustausch und die Diskussion über theoretische und praktische Implikationen bei der Digitalisierung von Editionen sowohl an SoftwareentwicklerInnen aus den digitalen Geisteswissenschaften als auch an FachwissenschaftlerInnen. Vier Fragefelder sollen aus der Perspektive verschiedener Akteure im Panel diskutiert werden:

- Wie lassen sich Typen von digitalisierten Editionen im Spektrum der digitalen Editionen kartieren? Können sie sich an *born digital*-Editionen annähern oder bleiben sie im Paradigma des Drucks verhaftet?
- Welche strukturellen, technischen und wissenschaftlichen Hürden können von der Planung bis zum Abschluss einer digitalisierten Edition auftreten?
- Welche Komponenten und Verfahren erfolgreicher Digitalisierungs *workflows* lassen sich erkennen?
- Welche Handlungsempfehlungen und *Best Practices* können auf Grundlage der vorhergehenden Fragen formuliert werden?

Das Panel beginnt mit einer Einleitung durch die Moderatoren, der kurze Statements der Beitragenden mit Schwerpunkt auf bestimmte Fragefelder folgen und die mit einer These oder Fragestellung enden. Sie dienen als Problemaufriss und zur Identifizierung unterschiedlicher Positionierungen im Kontext der (Retro)Digitalisierung, über die im Anschluss debattiert wird. Es folgt eine Diskussion im Plenum. Darauf aufbauend werden die Beitragenden (sowie weitere Interessierte) im Nachgang der DHd2020 die Arbeit an einem Leitfaden aufnehmen, der sowohl technisch-praktische als auch methodische Fragen der Digitalisierung von Druckeditionen berücksichtigt und als Ausgangspunkt für einen weiterführenden Diskurs dient. Der Entwurf des Leitfadens soll online vorab veröffentlicht werden. Die diskutierte und finalisierte Fassung (in englischer und deutscher Sprache) wird dauerhaft zugänglich gemacht werden.

Leitfragen der Statements

‚Born', ‚reborn', ‚retro': Kartierung von Editionstypen

Frederike Neuber

Im editionswissenschaftlichen Diskurs unterscheidet man im Spektrum der digitalen Editionstypen meist zwischen „*born digital*" und „Retrodigitalisierungen". Letzterem Typus wird dabei abgesprochen, eine „digitale Edition" im engeren Sinne zu sein; laut Sahle etwa überschreiten „*retrokonvertierte gedruckte Editionen oder vertiefende Digitalisierungs- und Erschließungsprojekte [...] oft nicht die Schwelle zu ‚digitalen Editionen'*" (Sahle 2014). Sind digitalisierte Editionen also dazu verdammt, als 'digitalisierte Bücher' im Paradigma der Druckkultur verhaftet zu bleiben oder konstituieren sie einen weiteren, und neu zu definierenden Editionstyp? Bei der Beantwortung dieser Frage spielt zum einen der Grad ihrer „*Verdatung*" (Krämer/Huber 2018) eine zentrale Rolle. Zum anderen rückt der doppelte Rückbezug auf eine historische Quelle/Dokument einerseits und die Druckedition andererseits die digitalisierte Edition in ein Spannungsfeld von Tradition und Wandel.

„Das bisschen Edition macht sich doch von selbst": Herausforderungen bei der Retrodigitalisierung von Editionen in der Praxis

Torsten Schaßan/Timo Steyer

In der Umsetzung der Retrodigitalisierung können vor allem zwei paradigmatische Schwierigkeiten ausgemacht werden: Zum einen wird der mit dieser Transformation verbundene Aufwand unterschätzt. Zum anderen wird der gedruckten Vorlage allzu häufig ein sakrosankter Status zugeschrieben. Damit verbunden sind zahlreiche Fragen, die einer Klärung im jeweiligen Projektkontext bedürfen. Häufig unklar ist bspw. ob und wenn ja, in welcher Form in den Text eingegriffen werden darf; sei dies aus Gründen der Fehlerkorrektur oder der Angleichung an den aktuellen Forschungsstand. Zentraler Diskussionspunkt wird im Statement die Frage nach dem Einfluss des Layouts der Druckedition auf die digitale Präsentation sein. Ebenso wird in die Debatte der Aspekt eingebracht, dass die Retrodigitalisierung häufig als rein technischer Prozess ohne philologischen Anspruch und wissenschaftlichen Mehrwert bewertet wird (Ball et. al 2016; Sahle 2012) und die beteiligten digital affinen WissenschaftlerInnen zum Dienstleister marginalisiert werden. Dies wird auch durch Missverständnisse bedingt, die mit dem Eingang neuer Terminologie in das Editionsprojekt aufgrund der *Datafication* einhergehen können.

Vier auf einen Streich – Zum Verhältnis von Workflow und Mindsets (nicht nur) im PROPYLÄEN-Projekt

Dominik Kasper

Auf dem Weg vom Druck zur digitalisierten oder gar digitalen Edition können unterschiedliche *Workflows* und Werkzeuge zum Einsatz kommen. Dabei lassen sich grundsätzlich zwei Verfahrensweisen nach ihrem Schwerpunkt unterscheiden: manuelle und automatische Erfassung. Im Statement werden häufig wiederkehrende Komponenten möglicher *Workflows* benannt und aufgezeigt, wann welches Verfahren - und ggf. auch deren Kombination - sinnvoll erscheint.

Welches Vorgehen Anwendung findet, wird unter anderem dadurch bestimmt, welche Erwartungen und Mentalitäten das Projekt prägen. Unterschiedliche *Mindsets* und Perspektiven können bspw. unterschiedliche Priorisierungen von Arbeitspaketen nach sich ziehen. Aber auch das Anhaften am Buch-Medium und divergente Sichtweisen auf den Charakter von Text-Modellierung oder den Stellenwert von Automatisierung einerseits und händischem Arbeiten/Annotieren andererseits können sich hier auswirken.

Dienstleistung als ein Baustein digitalisierter Editionen

Martina Gödel

Ein Projekt *workflow* muss sich nicht allein auf Leistungen der Projektpartner beschränken. Aufträge an externe Dienstleister können eine Option sein, um auf zügigem Wege eine solide Datenbasis zu erhalten, von der die fachwissenschaftliche Arbeit aus beginnen kann. Die genaue Definition von Umfang und Art der Leistungen, die je nach Projektbeschaffenheit und Personaldecke flexibel ausgeschrieben werden können, stellt eine Herausforderung dar, die zugleich das Bewusstsein für die konkreten Projektanforderungen erhöhen kann. Das Spektrum geht von reiner Texterkennung (entsprechend der gewünschten Fehlerfreiheit) bis zur Entwicklung und Anwendung von TEI-Datenmodellen, die auf die Spezifika der Druckedition eingehen und die Weiterarbeit möglichst weit vorbereiten und unterstützen.

Projekten, die in der Planungs- oder *Controlling*phase sind, soll anhand von erfolgreichen Projektbeispielen eine Entscheidungshilfe angeboten werden, wann und in welchen Bereichen Dienstleistung sinnvoll sein kann.

Bewahrung des kulturellen Erbes durch Transformation oder die Edition der Edition. Das Spannungsfeld von digitalisierter zur digitalen Edition aus Sicht der Bibliotheken

Thomas Stäcker

Bibliotheken bewahren gedruckte Editionen. Mit der Durchsetzung des digitalen Paradigmas werden diese selbst Gegenstand editorischer Prozesse. Nicht nur die Edition, sondern auch der digitale Transformationsprozess stellt eine neue erschließende Dimension dar: *„Throughout history, the act of editing stands out as the conscious effort, anonymous or non-anonymous, of making existing texts available in a new form"* (Haugen 2016: 206). Die erschließende ‚Übersetzung' im Sinne der Herstellung von Maschinenlesbarkeit bzw. *Datafication* ist eine wichtige Aufgabe von Bibliotheken. Dabei geht es weniger um eine hermeneutische Neuaneignung als um die Remediatisierung des Druckes (Bolter 2001). Sie dient - mit einem erweiterten Editionsbegriff (Price 2009) - der Sicherung der Zugänglichkeit nach Maßgabe der FAIR-Prinzipien und ist als umfassende, auf Dauer angelegte Aufgabe zu verstehen. Dabei ergeben sich u.a. Aufgaben und Fragen der Remodellierung, der Metadatenerfassung, Standardisierung sowie Entwicklung geeigneter Schnittstellen und Suchmöglichkeiten.

Teilnehmende

Frederike Neuber ist wissenschaftliche Mitarbeiterin bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften. Sie ist Mitherausgeberin von „Jean Paul - Sämtliche Briefe digital" und im Institut für Dokumentologie und Editorik u. a. als *Managing Editor* der Zeitschrift RIDE aktiv.

Torsten Schaßan ist wissenschaftlicher Mitarbeiter an der Herzog August Bibliothek Wolfenbüttel. Er betreut dort den Bereich Digitale Editionen. An der HAB wurden mehrere Retrodigitalisierungsvorhaben umgesetzt, darunter die Briefe der Fruchtbringenden Gesellschaft und "Controversia et Confessio".

Dominik Kasper ist wissenschaftlicher Mitarbeiter an der Akademie der Wissenschaften und der Literatur Mainz. Erfahrungen mit Retrodigitalisierung konnte er in den Projekten "Deutsche Inschriften Online" und "PROPYLÄEN – Goethes Biographica" (Leiter der Frankfurter Arbeitsstelle) sammeln.

Martina Gödel ist seit 2011 freiberuflich unter dem Namen *textloop* im Bereich Texterkennung, -korrektur und TEI-Auszeichnung tätig. Erfahrungen mit der Digitalisierung von gedruckten Editionen konnte sie unter anderem in der Arbeit für

die Projekte Dingler-Online, Blumenbach-online, Schule von Salamanca oder der Leibniz-Edition sammeln. Sie ist Mitglied in der DTABf-Steuerungsgruppe.

Thomas Stäcker (https://orcid.org/0000-0002-1509-6960) ist Direktor der ULB Darmstadt und nebenamtlicher Professor für *Digital Humanities* an der FH Potsdam. Zu seinen zahlreichen initiierten oder begleiteten DH-Projekten gehören zum Bereich der digital(isierten)en Editionen bspw. die Briefe Athanasius Kirchers an Herzog August, Lessings Übersetzungen, Lipsius' *De Bibliothecis*, die Werke Andreas Bodensteins gen. Karlstadt, Christoph Heidmanns *Oratio de Bibliotheca Julia* oder „Europäische Religionsfrieden".

Max Grüntgens (Moderation) ist wissenschaftlicher Mitarbeiter an der Akademie der Wissenschaften und der Literatur Mainz. Erfahrungen mit Retrodigitalisierung konnte er in den Projekten "Deutsche Inschriften Online" (Leiter der Mainzer Arbeitsstelle) und "PROPYLÄEN – Goethes Biographica" sammeln.

Martin Prell (Moderation) ist DH-Koordinator der PROPYLÄEN-Edition (Goethe- und Schiller-Archiv Weimar) und des "Editionenportal Thüringen" (Universität Jena). Er gibt unter anderem die Briefe Erdmuthe Benignas von Reuß-Ebersdorf heraus.

Fußnoten

1. Unter Digitalisierung von Editionen verstehen wir die Überführung bereits gedruckter Publikationen in ein elektronisches Format zum Zwecke der digitalen Verarbeitung und Bereitstellung. Beispiele dafür sind u.a. die Teilvorhaben von *PROPYLÄEN - Goethes Biographica*, http://www.goethe-biographica.de/; *Jean Paul - Sämtliche Briefe digital, http://jean-paul-edition.de* oder *Briefe der Fruchtbringenden Gesellschaft und Beilagen. Die Zeit Fürst Ludwigs von Anhalt-Köthen 1617–1650*, http://diglib.hab.de/edoc/ed000213/start.htm.

Bibliographie

Ball et al. (2016): „Der gedruckten Edition eine digitale Schwester. Das AEDit-Projekt und die digitale Edition der Fruchtbringenden Gesellschaft", in: *Denkströme. Journal der Sächsischen Akademie der Wissenschaften zu Leipzig* 16: 69-81, http://www.denkstroeme.de/heft-16/s_69-81_ball-dickel-herz-steyer [letzter Zugriff: 25.09.2019].

Bolter, Jay David (2001): *Writing Space: Computers, Hypertext, and the Remediation of Print*. 2. Aufl. Mahwah, NJ u.a.: Erlbaum.

Eggert, Paul (2009): "The book, the E-text and the 'Worksite'", in: Deegan, Marilyn (ed.): *Text editing, print and the digital world*. Farnham u.a.: Ashgate 63-82.

Haugen, Odd Einar (2014): "The Making of an Edition", in: Apollon, Daniel / Bélisle, Claire / Régnier, Philippe (eds.): *Digital Critical Editions. Topics in the Digital Humanities.* Urbana: University of Illinois Press 203-245.

Krämer, Sybille / Huber, Martin (2018): „Dimensionen Digitaler Geisteswissenschaften", in: *Zeitschrift für digitale Geisteswissenschaften*, http://dx.doi.org/10.17175/sb003_013.

Price, Kenneth M. (2009): „Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?", in: *Digital Humanities Quarterly* 3, Nr. 3, http://digitalhumanities.org:8081/dhq/vol/3/3/000053/000053.html [letzter Zugriff: 25.09.2019].

Sahle, Patrick (2012): „Mal wieder und immer noch: Digitized vs. Digital", in: *DHd-Blog* (27. November 2012) https://dhd-blog.org/?p=1122 [letzter Zugriff: 25.09.2019].

Sahle, Patrick (2013): *Digitale Editionsformen: zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik* (= Schriften des Instituts für Dokumentologie und Editorik 8). Norderstedt: Books on Demand.

Sahle, Patrick (unter Mitarbeit von Georg Vogeler und den Mitgliedern des IDE) (2014): „Kriterienkatalog für die Besprechung digitaler Editionen" (Version 1.1) https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ [letzter Zugriff: 25.09.2019].

Datamodelling Drama and (Musical)theater

Die Katalogisierung von Sammlungs- und Bibliotheksbeständen und zahlreiche Datenbankprojekte aus der Dramen-, Theater- und Musikforschung haben in den letzten Jahrzehnten eine bisher kaum berücksichtigte Fülle von Material zum (Musik-)theater in gedruckter und handschriftlicher Form zu Tage gefördert und recherchierbar gemacht. Die ebenfalls rasch voranschreitende Bild- und Metadatendigitalisierung macht dieses Material der Forschung einfach zugänglich. Ein Portal, das die vielen Einzelprojekte zu Aufführungsdaten, Texten, Noten und Werken gemeinsam recherchierbar machen würde, so dass Strukturen und Zusammenhänge wie Werk- und Aufführungsserien, Gattungszusammenhänge oder Popularität und Wirkung sichtbar werden, gibt es derzeit jedoch nicht.[1]

Stattdessen wurden in den letzten Jahren inhaltlich und technisch sehr heterogene Datenbanken zur Erfassung von Dramentexten, Libretti, Noten, Aufführungen und Theatertruppen angelegt.[2] Je nach Disziplin und Fragestellung bieten sie verschiedene Zugänge, indem sie etwa primär Aufführungen verzeichnen[3] oder Aufführungen und handschriftliche oder gedruckte Theatertexte und Theaternoten bzw. weitere Text und Bilddokumente die damit in Verbindung stehen.[4] Ein anderer Zugang geht von Gattungen, Werken, deren Fassungen und von Werkreihen aus.[5] Von dieser textuellen Seite lassen sich Materialzusammenhänge dann erschließen, wenn in Portalen die Suche nach Gattungen und Subgattungen möglich ist und die Verlinkung von Werktiteln auf den Eintrag in der Gemeinsamen Normdatei (GND) verwendet wurde. Dies ist in einigen Portalen wie etwa dem Karlsruher Virtuellen Katalog oder dem Verzeichnis der Drucke des 17. Jahrhunderts[6] und dem Verzeichnis der Drucke des 18. Jahrhunderts[7] möglich. Ein Normdatensatz enthält eine auf Grundlage eines Regelwerks definierte Ansetzungsform (bevorzugte Namensform), beliebig viele von der Ansetzungsform abweichende Namensformen (Verweisungsformen), Erläuterungen zur Ansetzungsform (z.B. Quellenangabe), weitere normdatenspezifische Informationen und ggf. redaktionelle Hinweise (vgl. Plassmann et a. 2011, Gantert/Hacker 2008). Normdatensätze werden in strukturierter Form in überregionalen, regionalen und lokalen Normdateien erfasst bzw. gespeichert. Im deutschspra-

chigen Raum gibt es mit der gemeinsamen Normdatei (GND) eine überregionale von der Deutschen Nationalbibliothek gehostete Normdatei.[8] Hier werden derzeit Personen, Körperschaften, Orte, Werke und Veranstaltungen erfasst. Die Einträge sind freilich von sehr unterschiedlicher Detailliertheit und Vollständigkeit, werden aber laufend ergänzt. Im Gegensatz dazu ist die Verzeichnung von Materialien und deren Einbettung in Werkzusammenhänge in der Musikwissenschaft bereits gut strukturiert und etabliert und über den Katalog Répertoire International des Sources Musicales (RISM) recherchierbar.[9]

Um diesen Teil des kulturellen Erbes strukturiert zugänglich zu machen, wäre es nötig Material aus Bibliothekskatalogen, Archivbeständen, Findbüchern und bereits bestehenden Datenbankprojekten zu verknüpfen. Grundlage dafür wären eine umfassende Ontologie des (Musik-)Theaterbereichs, das die Relationen der Objekte und Metadaten abbildete und die Abfrage der Zusammenhänge möglich machte. Zu erfassen wären einerseits die Materialien wie handschriftliche Soufflierbücher, Noten, Theaterzettel, Ariendrucke, Kupferstiche, Videomaterial, Zeitungsberichte usw., aber auch die sie verknüpfenden Praktiken des Produzierens, Bearbeitens, Übersetzens, Aufführens, Kombinierens, Druckens und Distribuierens. Einen zentralen Ansatzpunkt scheint Swiss Performing Arts Datamodel zu bieten.[10] Es schließt an das European Collected Library of Artistic Performance, Performing Arts Vocabulary (ECLAP) an und entwickelt dieses weiter.[11] Es handelt sich um ein äußerst komplexes Datenmodell, das sechs verschiedene Typen von Klassen, 17 Attribute, 23 Relationen und fünf Arten von Qualifiern vorsieht und so Materialien, Aufführungen Akteure verbindet.

Das Panel geht von den jüngsten Bestrebungen zur Entwicklung einer Ontologie für die Darstellenden Künste und ihrer Implementierung in einer Triple-Store-Datenbank aus (Birk Weiberg). Ein Vorteil dieses Datenbankformats ist es, dass es auch die Koexistenz von widersprüchlichen Informationen ermöglicht, was insbesondere bei offenen Forschungsfragen etwa bei der Datierung von Aufführungen oder der Zuordnung zu Truppen etc. ein wichtiger Fortschritt ist (Rusher 2002-2004). In einem weiteren Schritt sollen Datenbank-Projekte aus der Theater- (Klaus Illmayer), Musiktheater- (Gesa zur Nieden) und Bibliothekswissenschaft (Katrin Bicher) vorgestellt werden. Wie die Recherche nach Werkzusammenhängen basierend auf neuen Gegenstandsontologien für eine aufführungsorientierte und materialbasierte Dramengeschichte genutzt werden kann, wird Katrin Dennerlein aufzeigen. Dabei wird auch reflektiert, wo die Bibliotheken und Archive ihre Kooperationsmöglichkeiten bisher noch nicht ausschöpfen, wie jüngst erst wieder angemerkt wurde (vgl. Knoche 2017). Die einzelnen Projekte sollen ihre Modellierung im Abgleich mit SPA-Datenmodell darstellen. Am Ende des Panels soll diskutiert werden, um welche Objekte, Attribute und Relationen eine Ontologie zu ergänzen wäre, die die Komplexität des literarischen, aufführungbezogenen und musikalischen Materials abdecken könnte.

Die SPA-Ontologie und ihre Implementierung

Dr. Birk Weiberg (Luzern)

Das Schweizer Archive der Darstellenden Künste (SAPA) ist vor zwei Jahren aus der Fusion zweier Tanz- und Theaterarchive entstanden. Ein wesentlicher Teil der Fusion ist die Zusammenführungen sehr unterschiedlicher Datenbanken. Dafür wurde in einem ersten Schritt ein auf CIDOC-CRM, FRBRoo und dem noch in Entwicklung befindlichen Archivstandard Records in Context (RiC) basierendes Datenmodel entwickelt, welches versucht, die Darstellenden Künste sowohl mittels archivarischer als auch dokumentarischer Ordnungsstrukturen abzubilden. Zur Zeit wird das Datenmodel implementiert und die Bestandsdatenbanken sukzessive in eine Graphdatenbank migriert. Dabei stellt sich immer wieder die Frage, welche Elemente des komplexen Datenmodels sich in welcher Weise praktikabel implementieren lassen und wie die Daten im Triplestore in Zukunft editiert werden können. Ein wesentlicher Aspekt des SAPA-Projekts ist die Zusammenarbeit mit Wikidata, wo es das "WikiProject Performings Arts" gibt, in dem die Schweizer mit internationalen Daten zusammengeführt werden. Eine nachhaltige Anbindung an den Fachinformationsdienst Darstellende Künste ist ebenfalls geplant.

Modellierung von Inszenierungsdaten am Beispiel von theadok.at

Dr. Klaus Illmayer

In der Aufführungsdatenbank „Theadok“,[12] einem Gemeinschaftsprojekt des *Instituts für Theater-, Film- und Medienwissenschaft der Universität Wien* mit der *Österreichischen Akademie der Wissenschaften – Austrian Centre for Digital Humanities* , werden Daten von Theaterinszenierungen aller Sparten aus Österreich gesammelt und für die wissenschaftliche Forschung aufbereitet. Bei der Modellierung konzentriert man sich auf Inszenierungen, Vorlagen, Personen, Bühnen, Ensembles und Festivals, die als Entities konzipiert sind, zwischen denen Relationen bestehen. Bei Personen findet sich der Verweis auf den jeweiligen Eintrag in der Gemeinamen Normdatei (GND) der Deutschen Nationalbibliothek. Die Verweise auf die GND für Werke sind in Arbeit und auch Archivmaterial wird nach und nach ergänzt. Der Datenbestand umfasst derzeit ~30.000 Inszenierungen aus den Jahren 1945-2001, die entweder produziert oder aufgeführt wurden in Österreich.

Normdaten für Bühnenwerke - Chancen und Herausforderungen einer kollaborativen und nachhaltigen Erschließung

Dominik Stoltz/Katrin Bircher

An der Staats- und Universitätsbibliothek Dresden (SLUB) soll der RISM zu einem zentralen Nachweisinstrument für Musikdrucke des 16. bis 18. Jahrhunderts ausgebaut werden.

Dabei spielen die Metadatenauszeichnung des RISM, Werk- und Fassungsklassifikationen nach FRBR und der Verweis auf Personen- und Werk-Normdatensätze der Gemeinsamen Normdateit (GND) eine wichtige Rolle. Im Rahmen des Fachinformationsdienstes Musik *musiconn* wird an der SLUB eine Datenbank zur Erfassung von Aufführungsdokumenten entstehen, sowie ein Fachrepositorium, in dem musikwissenschaftliche Fachliteratur open access zur Verfügung gestellt wird. Die Zusammenführung aller Informationen in einer Datenbank, die eine gemeinsame Abfrage erlaubt, ist eine besondere Herausforderung, die ebenfalls Gegenstand des Vortrages sein wird.

Quellenklassifikation und Datenbank im Projekt „Pasticcio“

Prof. Dr. Gesa zur Nieden

Das frühneuzeitliche Opernpasticcio, für das zu überregional bekannten und bereits mehrmals vertonten Libretti Arien verschiedener Komponisten zusammengestellt wurden, zeichnet sich durch die Vielzahl der am Produktionsprozess beteiligten Akteure aus. Für die Produktion von Opernpasticcios war das zugrundeliegende Libretto und seine Anpassung an lokale Gegebenheiten wie auch die Wünsche der Impresari sowie der Sängerinnen und Sänger bei der Auswahl der Einzelarien in gleicher Weise maßgeblich. Im deutsch-polnischen Projekt "Pasticcio. Ways of Arranging Attractive Operas" sollen die damit einhergehenden Transferprozesse durch die Kombination einer digitalen Edition der Notentexte mit einer Datenbank zu den Karrieren der beteiligten Sängerinnen und Sänger abgebildet werden. Bei der Strukturierung der Daten auf der Grundlage verbreiteter Modelle wie FRBR stellt sich jedoch die Frage, welchen Werkstatus das Opernpasticcio besitzt, vor allem im Hinblick auf die Abbildung der Querverbindungen mit weiteren Opern über die im Pasticcio jeweils enthaltenen Einzelarien. Im Vortrag sollen unterschiedliche Möglichkeiten einer Klassifizierung diskutiert werden, die vom Libretto als Werk ausgehen, eine "work cloud" aus Libretto und Notentext vorsehen oder den musikalischen Text des Pasticcios als Werk ansetzen kann. Eine solche Alternative muss nicht zuletzt anschlussfähig an den Umgang mit FRBR in weiteren Disziplinen wie der Literatur- und der Theaterwissenschaft sein, um die digitale Darstellung der Forschungsergebnisse zum Pasticcio über Normdaten in größere Kontexte einbinden zu können.

Rewriting History of Drama

PD Dr. Katrin Dennerlein

Dramengeschichte wird in der Literaturwissenschaft bisher zumeist als Geschichte einzelner Sprechtheaterwerke nach dem Perlenschnurprinzip erzählt, die kaum einmal in Aufführungskontexte eingebunden sind oder gar die Fassung spezifizieren, von der sie ausgehen. Will man das umfangreiche und heterogene gedruckte und handschriftliche Material in seinen zahlreichen medialen Formen und die Praktiken des Produzierens, Aufführens, Druckens, Distribuierens, Rezensierens, Kritisierens berücksichtigen, benötigt man eine Ontologie, um diese Materialien und Praktiken aufeinander beziehen zu können. Gegenstand des Vortrages sollen die Anforderungen an eine solche Ontologie, sowie die neuen multimedialen und narrativen Darstellungsmöglichkeiten sein, die durch sie ermöglicht werden würden.

Timetable

- 5 Min. Katrin Dennerlein: Einführung
- 10 Min. Dr. Birk Weiberg: Die SPA-Ontologie und ihre Implementierung
- 10 Min. Dr. Klaus Illmayer: Modellierung von Inszenierungsdaten am Beispiel von theadok.at
- 10 Min. Katrin Bircher: „Normdaten für Bühnenwerke - Chancen und Herausforderungen einer kollaborativen und nachhaltigen Erschließung“
- 10 Min. Prof. Dr. Gesa zur Nieden: Quellenklassifikation und Datenbank im Projekt „Pasticcio“
- 10 Min. PD Dr. Katrin Dennerlein: Rewriting History of Drama
- 35 Min. Diskussion des Plenums

Fußnoten

1. Man vergleiche etwa die sehr unterschiedlichen Informationszusammenstellungen, die die Recherche nach populären Musiktheaterwerken um 1800 in Datenbanken wie dem Karlsruher Virtuellen Katalog (KVK), dem Fachinformationsdienst Performing Arts (https://performing-arts.eu), der Europeana (https://www.europeana.eu/portal/de) oder dem Répertoire International des Sources Musicales (RISM, https://opac.rism.info/metaopac/start.do?View=rism).
2. Vgl. die informative Zusammenstellung von Datenbanken aus diesem Bereich von Klaus Illmayer: https://www.zotero.org/groups/494335/digital_humanities_in_theatre_film_and_media_studies/items
3. Zum Beispiel die Aufführungsdatenbank „Theadok“ (https://theadok.at), ein Gemeinschaftsprojekt des Instituts für Theater-, Film- und Medienwissenschaft der Universität Wien mit der Österreichischen Akademie der Wissenschaften - Austrian Centre for Digital Humanities.
4. Die 2017 fertig gestellte Datenbank zum *Hamburger Stadttheater* (http://www.stadttheater.uni-hamburg.de/) verzeichnet Aufführungen, Theaterzettel, handschriftliche Theatermaterialien wie Soufflierbücher und Inspektionsbücher, letztere mit händisch eigegebenen Signaturen. Rollenauszügen, Partituren, Stimmen oder Drucke aus dem lokalen oder aus überregionalen Bibliothekskatalogen sind nicht angegeben oder verlinkt. Ähnlich *Berliner Klassik. Das vollständige Repertoire von Ifflands Direktion, Dezember 1796 bis 1814*, http://berlinerklassik.bbaw.de/BK/theater/Theaterzettel.html. Ebenfalls eine Aufführungsdatenbank ist die Weimarer Theaterzetteldatenbank, weil sie die Erfassung der Aufführungen im *Weimarer Hoftheater* und im *Deutschen Nationaltheater* von 1754–1990 zum Ziel hat. Die extrahierten Aufführungs-, Werk- und Personen-Datensätze verweisen auf die Normdatensätze, die Digitalisate der Theaterzettel sind verlinkt.
5. Vgl. z.B. „Die Oper in Italien und Deutschland zwischen 1770 und 1830“ (http://www.musikwissenschaft.uni-mainz.de/musikwissenschaft/projekte/operberlin.html). In diesem Projekt kooperieren mehrerer Bibliotheken, so dass über Normdatensätze tatsächlich Fassungen und Aufführungsserien von Opernwerken recherchiert werden können (Vgl. auch die Dokumentation hier: https://de.wikipedia.org/wiki/DFG-Opernprojekt). Erfasst werden Opernwerke, Komponisten, Manuskripte, Fassungen, Libretti, Librettisten, Aufführungsserien und Aufführungsdaten. Verlinkt werden hier digital verfügbare Manuskripte der Noten, aber keine Texte und Drucke.
6. http://www.vd17.de/.
7. https://gso.gbv.de/DB=1.65/.
8. http://www.d-nb.de/standardisierung/normdateien/gnd.htm.
9. https://opac.rism.info/metaopac/start.do?View=rism.
10. https://datahub.ckan.io/dataset/360d-b967-078c-4eca-bcd4-58bb5870753f/resource/9b-b9d231-6b39-4e06-a44a-d0d9d939d45f/download/spa_data_model_v0-51_20170926.pdf

11. http://www.eclap.eu/schema/eclap/. Eine Auflistung weiterer Datenmodelle aus diesem Bereich findet sich hier: https://www.wikidata.org/wiki/Wikidata:WikiProject_Performing_arts/Data_structure unter „Existing Data Models / Ontologies outside Wikidata“
12. https://theadok.at

Bibliographie

Jonathan Bollen (2016) : Data Models for Theatre Research: People, Places, and Performance. In: Theatre Journal 68,4 2016, S. 615-632.

Estermann, Beat / Christian Schneeberger (2017): Data Model for the Swiss Performing Arts Platform. Draft Version 0.51 September 2017 https://datahub.ckan.io/dataset/360db967-078c-4eca-bcd4-58bb5870753f/resource/9bb9d231-6b39-4e06-a44a-d0d9d939d45f/download/spa_data_model_v0-51_20170926.pdf

Gantert, Klaus / Rupert Hacker (2008): (2008): Bibliothekarisches Grundwissen. München.

Knoche, Michael (2017): *Die Idee der Bibliothek und ihre Zukunft.*

Engelbert Plassmann (et. al.) (2011): Bibliotheken und Informationsgesellschaft in Deutschland. Eine Einführung. Wiesbaden.

Events: Modellierungen und Schnittstellen

Kurzzusammenfassung

Einerseits sind Ereignisse im menschlichen Leben, und damit auch in dessen Dokumentation, allgegenwärtig. Andererseits gibt es in den Digital Humanities gerade für diesen zentralen Bereich noch deutlichen Nachholbedarf sowohl a) betreffend die Modellierung und darauf aufbauend Kodierung von Ereignissen, als auch b) betreffend die Austauschbarkeit von Daten über Ereignisse, für die es weder Standards noch – in anderen Bereichen bestehende – umfassende Normdatenquellen etwa aus dem Bibliotheksbereich (GND, VIAF, GeoNames etc.) in breit akzeptierter Form gibt.

Die EinreicherInnen haben angesichts dessen zur TEI-Konferenz 2019 eine Initiative unternommen, das Datenmodell von tei:event zu erweitern, ihm in Analogie zu anderen Named Entities aus dem TEI-Modul ‘namesdates’ ein tei:event-Name zwecks Referenzierung im tei:body zur Seite zu stellen und die Möglichkeiten des tei:listEvent-Wrapperelements zu ergänzen. Während wir – deren Domäne die Digitale Edition ist – dieses Ziel in Abstimmung mit dem TEI-Konsortium verfolgen, streben wir außerdem einen breiteren Austauschprozess mit anderen Bereichen innerhalb der Digital Humanities an, welche sich ebenfalls mit Ereignissen beschäftigen. Das vorgeschlagene Panel versammelt ausgewiesene Expertinnen und Experten zu einer offen angelegten Diskussion zur Modellierung von Ereignissen sowie zu möglichen Einsatzszenarien einer ‘eventSearch API’.

Hintergrund

Die Projektidee – in Anlehnung an CorrespSearch (Korrespondenzdatennetzwerk) ein Ereignisdatennetzwerk aufzubauen –, Ereignisdaten homogenisiert zu sammeln, diese zusammenzuführen und sie als historischen Background datumsbezogen zur Verfügung zu stellen, entstand aus den individuellen Interessen dreier unterschiedlicher Editionsprojekte an drei unterschiedlichen Institutionen, die von WissenschaftlerInnen verschiedener akademischer Disziplinen betrieben werden: Christiane Fritze und Christoph Steindl, Österreichische Nationalbibliothek; Infrastruktur für digitale Editionen an der Österreichischen Nationalbibliothek mit einer Tagebuchedition in TEI // Helmut W. Klug vom Zentrum für Informationsmodellierung der Uni Graz; mehrere Editionsprojekte mittelalterlicher Texte im GAMS // Stephan Kurz, Österreichische Akademie der Wissenschaften; hybride Edition Ministerratsprotokolle der Habsburgermonarchie in TEI.

> Wegen der Vielfalt der Anwendungsfälle und der edierten Materien (im einreichenden Team sind das etwa: Tagebuch, Itinerar, Kalender, Protokoll) verzichten wir auf eine normative Definition des Ereignisbegriffs abseits von „Ein Ereignis ist eine in Zeit, Frequenz und Raum verortbare Zustandsänderung eines oder mehrerer Objekte oder ihrer wechselseitigen Bezüge, die durch eine oder mehrere Quellen belegbar und mit einem oder mehreren Bezeichnungen benannt sein kann.“ In diesem Rahmen lassen sich in unseren Editionen ganz unterschiedliche Qualitäten von Ereignissen beschreiben, wie etwa: a) Frau X schreibt am 14.6.1932 am Ort Y in ihr Tagebuch, welches sie 3 Monate später an Herrn Z in A sendet; b) In dem Tagebuch berichtet sie vom Traum der letzten Nacht, in dem sie von der Geburt einer Tochter geträumt hat, c) Der Benediktinerpater F. langt am Michaelistag 1340 nach mehrtägiger Reise in Modriach an, d) Im Protokoll mit Signatur 777 steht: Die Minister A, B, D und N lehnen in der Sitzung vom 15.7.1876 das Gnadengesuch des Mörders M. ab, e) Kaiser F. lässt die Kriegserklärung übermitteln. Die Verantwortung für die Granularität, Reziprozität (nesting events) und Auswahl der für einen Datenbestand als Ereignisse erwähnenswerten Zustandsänderungen liegt bei der jeweiligen Herausgeberinnen.

Zielsetzung

Ausgehend von dem Bedürfnis, die Daten zu Ereignissen aus (den oben skizzierten, aber auch allgemein) TEI-Editionen bzw. anderen Datenquellen vergleichbar und distribuierbar zu machen, verfolgen wir mit dem Panel das übergeordnete Ziel, den wissenschaftlichen Diskurs zur Modellierung, Sammlung und Darstellung von Ereignissen nach der überaus erfolgreichen Diskussion mit der TEI-Community auch in einem breiteren DH-Kontext voranzutreiben. Eine erfolgreiche Umsetzung eines derartigen Service kann nur erreicht werden, wenn Vorschläge und Feedback eines möglichst heterogenen InteressentInnenkreises vorliegen und die Möglichkeiten anderer Dokumentationsstandards (z.B. CIDOC CRM) sowie die Interoperabilität der Ansätze in Betracht gezogen werden. Darüberhinaus müssen Fragen zur Vernetzung von editionsgetriebenen und fremddatengestützten Ereignisdaten mit weiteren Daten z.B. aus prosopographischen Forschungs-

unternehmungen bzw. aus dem Semantic Web (Linked Open Data) diskutiert werden.

Konkret heißt das:

- Vorstellung der Ideen zu eventSearch mit kritischer Rückmeldung
- Sammlung von Zugängen zur Ereignismodellierung aus LOD/Semantic Web
- Verbreiterung der Basis an Überlegungen durch Öffentlichkeit dieser Diskussion
- Vernetzung von Event-Interessierten (Konsortialbildung? Infrastrukturschaffung?)
- Einladung zur Mitarbeit

Methoden

Panel. Um statt der im CFP definierten 30 Minuten für die Diskussion mit dem Publikum zumindest 45 Minuten aufwenden zu können und damit notwendige Rückkoppelung und Feedback zu erhalten, sind nach einer Themeneinführung sehr kurze Einleitungsstatements (5 Minuten pro beteiligter Initiative/Person) geplant. Nach einer Diskussionsrunde innerhalb der Panel-TeilnehmerInnen zu dem Thema Modellierung von Ereignissen sowie zum Thema Schnittstelle werden Fragen und Anregungen gemeinsam mit dem Publikum diskutiert.

Stand der Diskussion zu Ereignissen im Kontext TEI-gestützter digitaler Editionen

Ereignisse sind allgegenwärtige Merkmale menschlichen Lebens zu allen Zeiten an allen Orten. Die deutschsprachige Wikipedia listet für den 27. September 65 Ereignisse in Politik und Wirtschaft ('events'), TV-Sender und Zeitungen bieten ähnliche Formate, indem sie in einer Art Panoptikum historische Ereignisse Revue passieren lassen, die EU plant ein monumentales Time Machine Projekt. Allgemein gesprochen: Es besteht immer Interesse daran, unterschiedlichste Ereignisse in Relation zueinander zu setzen. Besonders spannend wird das, wenn man dies nicht nur als eine Art Rückschau betreibt, sondern versucht, dieses Nebeneinander von Ereignissen für einen Tag, eine Periode in der Vergangenheit darzustellen. Entsprechend unausweichlich trifft man auf Erwähnung und Dokumentation von Ereignissen in historischen wie zeitgenössischen Quellen, die Grundlage geisteswissenschaftlicher Forschung sind. Aber ein Nebeneinander von Ereignissen kann auch für fiktionale Texte festgestellt werden. Historische Zeugen können Lebensdokumente (Tagebücher, Reisenotizen) genauso sein wie amtliche Dokumente (Urkunden, Protokolle, Kalender) oder Kulturüberlieferungen materieller Natur (Inschriften, Teppiche, tei:object u.a.). Während Ontologien wie CIDOC CRM Ereignisse als Zustandsänderungen beschriebener Objekte modellieren, gehen textbasierte Schemata wie die TEI bislang eher implizit davon aus, dass einem Überlieferungsträger auch ein (z.B. Schreib-)Ereignis ursächlich zugrunde liegt; im Text beschriebene Ereignisse sind bestenfalls im Sinne einer Referenzierung in Analogie zu anderen Klassen von Named Entities erfasst. Auch die Linked Open Data-Welt beschreibt/referenziert Ereignisse. Den Hiatus zwischen diesen Ansätzen zu harmonisieren fällt aus unserer Sicht am ehesten konkreten Projektanwendungen zu, die bspw. im Zusammenhang mit prosopographischen Auxiliardaten für digitale Editionen stehen – deren Aufmerksamkeitsspanne endet allerdings an den Enden des konkreten Anwendungsfalles. Das vorgeschlagene Panel richtet seine Aufmerksamkeit auf diesen neuralgischen Punkt in der DH-Forschungslandschaft.

Besonderes Augenmerk der Panel-Diskussion gilt auch der Frage der Granularität von Ereignissen in den verschiedenen Ansätzen: In CIDOC CRM etwa ist prinzipiell jede distinkte kleinteilige Zustandsänderung als Ereignis konzeptionalisiert, wogegen für das Ziel eines Discovery- und Disseminationservices eher ein alltagssprachliches Verständnis von Ereignissen Nutzen verspricht (im Editionskontext mit der Lösung in TEI, dass all jenes zum Ereignis wird, das von der Editorin/dem Editor als tei:event ausgezeichnet ist).

Alternativen zu CIDOC CRM und/oder TEI, stehen zur Verfügung und sollten ebenfalls diskutiert werden:

- The Simple Event Model Ontology, https://semanticweb.cs.vu.nl/2009/11/sem/
- Normdaten zu historischen Einzelereignissen z.B. aus der GND-Ontologie, die die bibliothekarischen Schlagwortketten in super/sub-Relationen modelliert
- CMIF (nicht erst in der Version 2, siehe Dumont et al. 2019) als ausmodellierter Spezialfall: correspActions als Subklasse von Ereignissen
- ...

Ereignisse im Namensraum TEI:

Werden Quellen und Texte unabhängig von ihrem Fiktionalitätsgrad digital ediert, hat sich zwar die Kodierung nach den Richtlinien der TEI inklusive der Erfassung von Named Entities durchgesetzt, jedoch fehlt eine gängige Praxis für die semantisch eindeutige Erfassung und Auszeichnung von Ereignissen. In den TEI-Guidelines werden zum Zeitpunkt der Einreichung Ereignisse (tei:event) geradezu nachlässig behandelt und es wird ihnen keine Eigenständigkeit zugesprochen: Das tei:event-Element ist derzeit nur recht eingeschränkt verfügbar und kann ausschließlich als Kindelement anderer Elemente aus dem Names-Dates-Modul verwendet werden: Im Allgemeinen sind Ereignisse anderen Konzepten (Date, Person, Organisation usw.) untergeordnet. Der Eindruck, dass 'Events' nicht besonders im Fokus der TEI-Community stünden, täuscht allerdings, denn seit 2010 gibt es die unterschiedlichsten Diskussionen im Rahmen der TEI-Mailingliste und im TEI-C GitHub. Dabei geht es ganz allgemein um die Auszeichnung von Ereignissen, aber es werden auch vielversprechende Themen wie die Einführung eines <eventName>-Elements besprochen und natürlich, dass es notwendig wäre, das Event-Element zu stärken und es z.B. den vergleichbaren Elementen aus dem Names-Dates-Modul anzugleichen. Mit der TEI P5 Version 3.6.0 wurden tei:ptr und tei:idno als Kindelemente von tei:event erlaubt. Dieses Event untermauert den Bedarf an einer tiefergehenden und viele AkteurInnen einschließenden Diskussion zum Ereignis. Hinzu kommt, dass zum Ereignis gehörige grundlegende Informationen wie ein zeitlicher Hinweis, eine Lokalisierung oder die Angabe zu beteiligten Personen nicht ohne Weiteres möglich ist. Im entsprechenden Vorschlag zur Änderung/Erweiterung

der TEI Guidelines wird die Kodierung von Ereignissen im Rahmen der TEI angepasst und werden dem TEI-Konsortium umfassend ausgearbeitete Änderungsvorschläge für eine universelle Auszeichnung von Events vorgelegt. Dabei wird eine möglichst flache Auszeichnung angestrebt, um die Datenmigration für Ereignisdaten so niedrigschwellig wie möglich zu halten. Das vorgeschlagene Modell wird auf bestehende Editionsprojekte der InitiatorInnen angewendet, um so ein exemplarisches Korpus an eventSearch-Daten bereitzustellen.

Elemente einer Verschaltung von Ereignissen

Ein Vorschlag, den die InitiatorInnen des Panels verfolgen, ist die Erstellung einer Programmierschnittstelle (API) basierend auf dem erarbeiteten und mit dem TEI-Konsortium abgestimmten und tw. noch abzustimmenden Datenmodell für Ereignisse sowie einer webbasierten grafischen NutzerInnenschnittstelle für in digitalen Editionen kodierte Ereignissen nach dem Vorbild von correspSearch . Events aus externen Quellen werden ebenso eingebunden und ermöglichen eine Betrachtung einzelner Ereignisse im historischen Kontext. So bietet eine kalendarische Übersicht raschen Überblick über singuläre Ereignisse, die möglicherweise von mehreren Zeugen und Zeugnissen betrachtet worden sind oder die koinzidieren. Die Übersicht lädt so zur Interpretation und Kontextualisierung ein. Im Gegensatz zu der bloßen Auflistung von Ereignissen zu einem bestimmten Zeitpunkt, wie sie leicht etwa über Wikipedia (historische Jahrestage) einzusehen ist, erfahren die NutzerInnen hier über die Verlinkung in die jeweiligen Quellen individuelle Wahrnehmung und Bewertung: Aus welcher Quelle/Edition stammen die Informationen, wessen Sicht auf ein bestimmtes Ereignis wird wiedergegeben, wer stellt das Ereignis/den chronologischen Kontext des Ereignisses wie dar?

Die von den PanelinitiatorInnen angedachte EventSearch-Infrastruktur ermöglicht es, die aufgezeichneten Ereignisse einzelner digitaler Editionen sichtbarer und leichter auffindbar zu machen. Eine Rückverlinkung von der Webschnittstelle/API zu digitalen Editionen gewährleistet, dass Ereignisbezüge einfach in der originalen Fassung angezeigt und analysiert werden können. Eine niederschwellige Möglichkeit zu partizipieren wird angeboten, sodass Editionsprojekte unabhängig von Größe und Budget nicht daran gehindert werden, ihre Ergebnisse zu teilen.

Bestätigte TeilnehmerInnen

- eine Person aus dem Kreis der Initiatorinnen:
  Wir stellen die Ergebnisse der Event-Extraction aus unseren eigenen Editionsdaten bis dato vor:
  - 5 Editionsprojekte, zu denen wir in TEI modellierte listEvents bereits vorliegen haben (CF/ÖNB: Okopenko-Tagebuch; SK/ÖAW-IHB: Ministerratsprotokolle Habsburgermonarchie, Mächtekongresse, HWK/ZIM-ACDH: Itinerar Santonino, Wirtschaftskalender, Heiligenkalender)
  - Mockups von Timelines, Kalendern, Widgets als Vorentwurf
- Matthias Schlögl, Österreichische Akademie der Wissenschaften; Projekt Austrian Prosopographical Information System (APIS):
  - Mapping von APIS-Daten zu CIDOC CRM und TEI
- Sascha Grabsch, Berlin-Brandenburgische Akademie der Wissenschaften, correspSearch
  - Ereignisse im Rahmen von Briefeditionen; Correspondence Metadata Interchange Format (CMIF) als Modell für Überschneidungen bei der Entwicklung von übergreifenden Infrastrukturen zur Aggregierung, Speicherung und Dissemination von ([correspAction]event-)Daten,

Bibliographie

Anderson, C. / Eide, Ø. / Orlowska, A. / Pindl, K. / Tomasek, K. / Vogeler, G. (2016): Modeling semantically Enhanced Digital Edition of Accounts (MEDEA) for Discovery and Comparison on the Semantic Web, 2016. https://hcommons.org/deposits/item/hc:24347

CIDOC Conceptual Reference Model Version 6.2, Mai 2015, http://www.cidoc-crm.org/Version/version-6.2

Dumont, Stefan (2016): correspSearch – Connecting Scholarly Editions of Letters, in: Journal of the Text Encoding Initiative, Issue 10, 2016, http://journals.openedition.org/jtei/1742.

Dumont, Stefan et al. (2019): Correspondence Metadata Interchange Format (CMIF). In: Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf. Edited by Stefan Dumont, Susanne Haaf and Sabine Seifert. Berlin 2019. https://encoding-correspondence.bbaw.de/v1/CMIF.html

Grishman, Ralph (2015): Information Extraction, in: The Oxford Handbook of Computational Linguistics, hrsg. v. Ruslan Mitkov, 2015 https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199573691.001.0001/oxfordhb-9780199573691-e-009.

Piotrowski, Michael (2012): Natural Language Processing for Historical Texts, in: Synthesis Lectures on Human Language Technologies, Bd. 5, San Rafael 2012, S. 1–157.

Mersch, Dieter (2002): Was sich zeigt. Materialität, Präsenz, Ereignis, München, Fink

Johnson, Uwe (1970ff): Jahrestage. Frankfurt: Suhrkamp.

Kritsotaki, Athina / Doerr, Martin (2006): Documenting Events in Metadata. In: The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST 2006. http://www.cidoc-crm.org/sites/default/files/Documenting%20Events%20in%20Metadata.pdf

Van Hage et al. (2009): The Simple Event Model Ontology. http://semanticweb.cs.vu.nl/2009/11/sem/

Vogeler, Georg (2019): The 'assertive edition', in: International Journal of Digital Humanities, 1,2, 2019, S. 309–322. http://dx.doi.org/10.17613/M6JS9H76P

Buckland, M. / Ramos, M.R. (2010): Events as a structuring device in biographical mark-up and metadata. Bulletin of the American Society for Information Science and Technology 36, 26–29. https://doi.org/10.1002/bult.2010.1720360209

Fokkens, A. / ter Braake, S. / Ockeloen, N. / Vossen, P. / Legêne S. / Schreiber G. / de Boer, V., (2018): BiographyNet: Extracting Relations Between People and Events. arXiv preprint .

Peroni, S. / Tomasi, F. / Vitali, F., (2013): Reflecting on the Europeana Data Model. https://doi.org/10.1007/978-3-642-35834-0_23

Intertextualität in literarischen Texten und darüber hinaus

Einleitung

Die Analyse der Formen und Funktionen von Intertextualität ist ein Forschungsbereich, dessen heuristischer Anspruch seit dem erstmaligen Auftreten des Terminus ‚Intertextualität' in Julia Kristevas Aufsatz *Bachtin, das Wort, der Dialog und der Roman* (1972 [1967]) von einer Spannung zwischen modellhafter Strenge und interpretativen Spielräumen geprägt ist. So betonte Roland Barthes im Anschluss an Kristeva die interpretative Freiheit der Rezipierenden bei der Herstellung intertextueller Relationen (vgl. z. B. Barthes 1974: 53f.). Im Kontrast dazu entwarfen Gérard Genette, Manfred Pfister, Susanne Holthuis und andere umfassende Modelle intertextueller Beziehungen, die das hier meist chronologisch gedachte Verhältnis von Prä- und Posttext systematisieren sollten (Genette 1993, Pfister 1985, Holthuis 1993).

Diese divergierenden Tendenzen innerhalb der Intertextualitätsforschung können nicht zuletzt auf die Tatsache zurückgeführt werden, dass sich literarische Intertextualität selbst – wie bereits das Wort ‚An*spiel*ung' nahelegt – durch einen gewissen Spielcharakter auszeichnet. Dabei beziehen sich intertextuelle Relationen aber stets auf bestimmte Texteigenschaften zweier oder mehrerer Texte, die in einer Relation von Übereinstimmung und Abweichung zueinander stehen: sprachliche Merkmale, Charaktere, Plotstrukturen etc. Dies verweist auf ein systematisches Funktionieren intertextueller Beziehungen. Intertextualität basiert damit konzeptuell auf der Doppelgesichtigkeit des Konzepts ‚Spielraum': „Öffnung und Schließung, Freiheit und Vorschrift können nicht getrennt voneinander betrachtet werden, sondern bedingen sich gegenseitig" (Dettke/Heyne 2016: 11f.).

Gerade dieses Changieren zwischen Regelhaftigkeit und Dynamik bereitete bisherigen Untersuchungen literarischer Intertextualität mit den Mitteln analoger Textarbeit stets große Probleme: Klassische Textanalysen und abstrakte Modelle erweisen sich gleichermaßen als defizitär, indem für eine nachvollziehbare Erfassung der bestehenden Vielfalt intertextueller Relationen gerade das Ineinandergreifen von Modellierung und Interpretation entscheidend ist (vgl. Nantke/Schlupkothen 2018, 2019). Die formale Modellierung bietet hier gesteigerte Möglichkeiten der systematischen Erfassung und der induktiven Kategorienbildung sowie der unmittelbaren Visualisierung. Auf diese Weise können Modelle entstehen, welche flexibel genug sind, um unterschiedlichste Formen von Intertextualität adäquat zu erfassen, und dabei gleichzeitig eine formale Strenge aufweisen, die einer maschinellen Abfrage sowie der Kombination mit (teil-)automatisiert erzeugten Analyseergebnissen offensteht. Bislang finden sich Beispiele für den Einsatz computergestützter Verfahren zur Intertextualitätsdetektion vor allem im Bereich der *digital classic studies*[1]. Das Panel zielt auf eine kritische Reflexion und Erweiterung der bestehenden text reuse-Studien in der digital arbeitenden Altphilologie und demonstriert anhand von Anwendungsbeispielen aus Literaturwissenschaft, Philosophie und Wissenschaftsgeschichte das Potenzial einer computergestützten Intertextualitätsforschung in weiteren Teilbereichen der Digital Humanities.

Konkret soll im Panel anhand verschiedener Beispiele aufgezeigt und diskutiert werden, wie und wo sich digitale Ansätze zur Erfassung und Modellierung intertextueller Beziehungen zwischen den Polen ‚Formalisierung' und ‚interpretative Freiheit' verorten lassen. Dabei verstehen wir eine intertextuelle Referenz als eine von einer Leserin/einem Leser wahrgenommene "Wiederholung" aus einem anderen Text, wobei die Wiederholung im Regelfall nicht (nur) die Textoberfläche betrifft, sondern Ideen, Gedanken, Formulierungen, Syntax- oder Plotstrukturen. Im Rahmen des Panels werden Verbindungsmöglichkeiten von quantitativen und qualitativen Verfahren zur Erschließung von Intertextualität evaluiert. Ebenfalls wird dabei erörtert, wie im Zuge der Modellierung interpretative Spielräume immer wieder zur Herausforderung für die Formalisierungsbestrebungen werden und wie derartige Situationen positiv gewendet spezifische Funktionsweisen von Intertextualität sichtbar machen können.

Panelvorträge

Mehrstufige Annotation literarischer Intertextualität jenseits der Textoberfläche

Julia Nantke (Universität Hamburg) & Ben Sulzbacher (Bergische Universität Wuppertal)

Für die systematisierende Erfassung intertextueller Relationen wurde im Projekt *FormIt* eine *Linkbase* entworfen, welche durch ihre mehrstufige Anlage verschiedene Möglichkeiten zur Verknüpfung und Annotation von intertextuellen Phänomenen eröffnet sowie eine unmittelbare Visualisierung der Ergebnisse leistet.

Mithilfe der *Linkbase* können in mehreren Texten parallel intertextuelle Bezüge als stabile Links annotiert werden. Das resultierende Modell bezieht sich neben sprachlichen Übereinstimmungen ebenso auf literaturwissenschaftlich relevante Kategorien wie Figurengestaltung, Perspektive, Erzählerstimme etc. Die Beziehung der annotierten Textstellen wird dabei hinsichtlich der beteiligten literaturwissenschaftlichen Kategorien sowie der Art der Relation (Hinzufügung/Auslassung, semantische Verschiebung, Kanalisierung, Relativierung etc.) bestimmt.

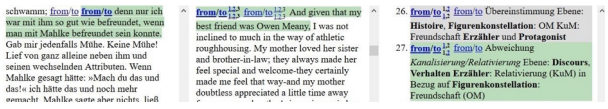


Abbildung 1: Ausschnitt *Linkbase* mit mehrstufiger Annotation einer intertextuellen Beziehung zweier Texte

Daraus abgeleitet erfolgt eine Modellierung der relevanten Kategorien in drei ‚Bäumen', welche die Phänomene nach den Ebenen *Textoberfläche* (Verortung im Text), *Discours* (Ausge-

staltung der Darstellung) und *Histoire* (Elemente des Dargestellten) gliedern. Eine kollaborative und rekursive Kategorienbildung verhindert zusammen mit der Anbindung an konkrete Texte und Textstellen das „blackboxing" (vgl. Latour 2000: 373) der verschiedenen Modellierungsschritte. Der Vortrag soll zeigen, wie verschiedene Möglichkeiten, Annotationen an Texte anzuknüpfen und Textstellen und Annotationen jeweils untereinander zu clustern, literarischen Strukturen angemessene Repräsentationen intertextueller Beziehungen ermöglichen.

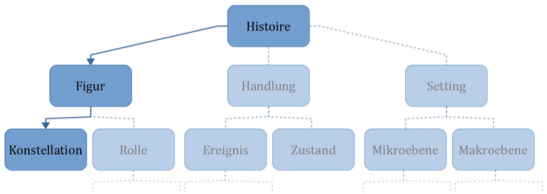


Abbildung 2: Ausschnitt des Kategorienbaums auf der Histoire-Ebene zur oben gezeigten Annotation (Übereinstimmung)

Die Struktur der *Linkbase* bildet eine Brücke zwischen der formalen Strenge, die für eine abstrahierende Modellierung literarischer Formen notwendig ist, und interpretativen Spielräumen, die bei der Repräsentation von Intertextualität häufig einbezogen werden müssen. Durch die unmittelbare Zusammenschau und interaktive Exploration der *Linkbase*-Ebenen können neue Erkenntnisse über die Funktionsweise intertextueller Beziehungen und literarischer Strukturen im Allgemeinen entstehen. Die verschiedenen Modellierungsstufen sorgen dafür, dass Anknüpfungspunkte für die Integration quantitativer Verfahren generiert werden. So gilt es bspw. zu evaluieren, wie Übereinstimmungen auf der Basis impliziter Informationen mit Methoden des Machine Learning (teil-)erfasst werden könnten.

Computergestützte Ansätze zur Detektion von Shakespeare-Referenzen in postmoderner Fiktion

Manuel Burghardt (Universität Leipzig) & Johannes Molz (Ludwig-Maximilians-Universität München)

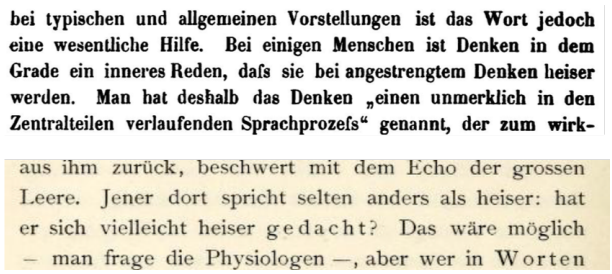
Als der einflussreichste Autor der westlichen Kulturhemisphäre wird Shakespeare bis heute in vielen literarischen Genres referenziert (vgl. etwa Taylor, 1989; Maxwell & Rumbold, 2018) und eignet sich damit wie kein anderer zu Untersuchungen des literaturwissenschaftlichen Phänomens der Intertextualität. Um die vielfältigen intertextuellen Bezüge auf Shakespeares Werk systematisch zu identifizieren verwenden wir computergestützte Methoden zur Erkennung von Textähnlichkeit (*text similarity*) und Textwiederverwendung (*text reuse*).

Wir präsentieren erste Ergebnisse aus einer Pilotstudie zur Identifikation von Shakespeare-Referenzen in Romanen aus den Bereichen Fantasy, Magischer Realismus und postmoderne Fiktion, da erste Voruntersuchungen zeigten, dass diese Genres in besonderem Maße dazu neigen, Shakespeare zu zitieren. Im Rahmen dieser Pilotstudie wurden unterschiedliche computergestützte Ansätze, wie bspw. *local alignments* (Burghardt et al., 2019) sowie Verfahren aus dem Bereich des maschinellen Lernens (bspw. *sentence embeddings*) erprobt, die jeweils ganz eigene Herausforderungen in Hinblick auf die eng miteinander verzahnte Modellierung von Hyper- und Hypotexten (vgl. Genette, 1993) und die Interpretation automatisch generierter Ergebnislisten mit sich bringen.

Annotation und Erkennung semi-literarischer Interferenz am Beispiel Nietzsche

Nils Reiter (Universität Stuttgart/Universität zu Köln) & Axel Pichler (Universität Stuttgart)

Intertextuelle Referenzen spielen neben der Literatur auch in der Philosophie eine große Rolle. So werden etwa in der Zeitschrift *Nietzsche-Studien* seit 1972 Nachweise intertextueller Verweise durch Nietzsche gesammelt. Abbildung 3 zeigt ein Beispiel, demzufolge Nietzsche die Vorstellung, dass denken heiser machen kann, von Höffding übernommen hat. Diese Daten können als Referenzdaten dienen, wobei sie natürlich nicht exhaustiv sind, obwohl Nietzsche zu denjenigen Autoren zählt, dessen Quellen am umfangreichsten erforscht sind.[2] Im dritten Panel-Beitrag werden zwei Ansätze und erste Arbeiten diskutiert, die sich an den Nietzsche-Nachweisen orientieren.


bei typischen und allgemeinen Vorstellungen ist das Wort jedoch eine wesentliche Hilfe. Bei einigen Menschen ist Denken in dem Grade ein inneres Reden, daſs sie bei angestrengtem Denken heiser werden. Man hat deshalb das Denken „einen unmerklich in den Zentralteilen verlaufenden Sprachprozeſs" genannt, der zum wirk-

aus ihm zurück, beschwert mit dem Echo der grossen Leere. Jener dort spricht selten anders als heiser: hat er sich vielleicht heiser gedacht? Das wäre möglich – man frage die Physiologen —, aber wer in Worten

Abbildung 3: Nietzsche-Nachweis aus Höffding, Harald: Psychologie in Umrissen, dokumentiert von Brobjer, Thomas (erschienen 2001 in Nietzsche-Studien (30))

Zunächst stellen wir ein Kategoriensystem vor, dass in einem Bottom-Up-Verfahren etabliert wurde. Dazu wurden die Nietzsche-Nachweise als existierende Annotationen aufgefasst und eine „Meta-Annotation" zugefügt, die die Art der Referenz charakterisiert (z.B. „semantisch äquivalente Paraphrase" oder „syntaktische Ähnlichkeit"). Mit den üblichen Methoden aus der reflektierenden Annotationspraxis (Übereinstimmung) können Definitionen für diese Charakterisierungen geschärft werden, so dass ein robuster Überblick über verschiedene Arten der Referenzen vorliegt. Im Gegensatz zu Ansätzen, die vollständig „from scratch" annotieren, bewahrt der Rückgriff auf existierende Referenzen davor, eine subjektiv motivierte Teilmenge an Referenzen in Betracht zu ziehen. Anknüpfungspunkte und Gemeinsamkeiten mit den im ersten Beitrag vorgestellten Kategorien zu eruieren ist eines der Ziele des Panels.

Daneben diskutieren wir Möglichkeiten, Referenzen automatisch zu erkennen. Klar ist, dass exhaustive Referenzdaten auf absehbare Zeit nicht zur Verfügung stehen werden, da die Menge an Referenzzielen tendenziell steigt und zu großen Teilen auch unbekannt ist. Auch Negativbeispiele lassen sich nur unter stark einschränkenden Annahmen sicher feststellen. Damit können überwachte maschinelle Lernverfahren nur noch bedingt eingesetzt werden. Unser Ansatz orientiert sich daher an den zuvor etablierten Annotationskategorien, und besteht

aus einer Sammlung von Erkennern, die die Kategorien operationalisieren. Ziel ist, potentiellen Benutzer_innen Vorschläge in verschiedenen Kategorien machen zu können, die dann individuell gewichtet und ausgewählt werden können.

WordWeb/IDEM: Datenbasierte Erfassung von Intertextualität durch eine Graphdatenbank zum frühneuzeitlichen englischen Theater

Regula Hohl-Trillini (Universität Basel)

Anstelle eines neuen Modells implementiert WordWeb/IDEM[3] ein 50jähriges, ikonoklastisches Nicht-Modell. Wie Ende Sechzigerjahre postuliert, realisiert die Datenbank[4] ein Netzwerk ohne Mitte, ein Universum von Texten ohne Bezug auf ein zentrales Werk. Die verbalen, motivischen und onomastischen Beziehungen zwischen Dramen der Shakespearezeit werden durch tausende Textausschnitte abgebildet, die dieselben Phrasen oder Namen enthalten. Diese verbindenden "Lexias" (Barthes 1973) repräsentieren "wahrgenommene Wiederholung aus einem anderen Text": Zitierende sind Leser, die schreiben und so den zitierten Text mitbestimmen. Die Weiterentwicklung der Hypertextdatenbank HyperHamlet[5] vollzieht dies nach: in WordWeb wird statt der Fixierung auf einen Autor ein radikales Konzept von Intertextualität umgesetzt, die "Intersubjektivität" ablöst (Kristeva 1967).

Das englische Drama um 1600 ist der ideale Testfall für WordWeb, weil poststrukturalistische Konzepte der Realität des frühneuzeitlichen Theaters vollkommen entsprechen. Im Londoner “Hollywood” arbeiteten Dramatiker zusammen, hörten und lernten die Werke der Kollegen (als Schauspieler), schrieben um, verfassten sequels, improvisierten und zitierten, meistens ohne "korrekte" Signalisierung. Wie Drehbuchschreiber amüsieren sie durchs Recycling von «memes», wie eine Bühnenfigur klarmacht: "My horse, my horse my kingdom for a horse - look, I speak play scraps!" (Marston 1601). Diese wettbewerbsorientierte, kommerzielle Theaterszene war tatsächlich ein " ‘tissue of quotations drawn from innumerable centres of culture" (Barthes 1968), eine Echokammer (Barthes 1975), die "likes" in der Form von Zitaten enthält.

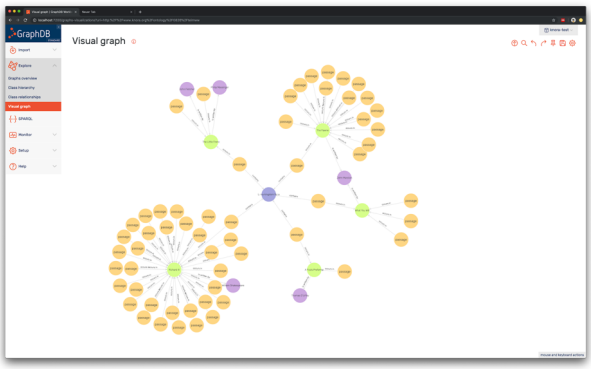


Abbildung 4: Graphenvisualisierung der Lexia “A horse, a horse, my kingdom for a horse”.

So kann WordWeb auch Shakespeares Beitrag zum "web of words" seiner Zeit klären, da es seine scheinbare Dominanz im Kontext der verbalen Landschaft zwischen 1550 und 1688 neu liest.

Struktur

Es ist geplant, dass das Panel der folgenden Struktur folgt:

- Kurze Einführung in die Thematik / größerer thematischer Rahmen
- Impulse durch vier Einzelvorträge
- Moderierte Abschlussdiskussion mit dem Publikum

Frau Prof. Dr. Evelyn Gius, TU Darmstadt, hat zugesagt, die Moderation des Panels zu übernehmen.

Fußnoten

1. für eine umfassende Bibliographie zu diesem Bereich vgl. https://wiki.digitalclassicist.org/Text_Reuse#References.
2. https://gepris.dfg.de/gepris/projekt/281581212
3. vgl. http://p3.snf.ch/project-183259
4. www.wordweb.unibas.ch
5. www.hyperhamlet.unibas.ch

Bibliographie

Barthes, Roland (1973). “Textual Analysis of Poe’s ‘Valdemar’” in Lodge, David (ed.): *Modern Criticism and Theory*. London: Longman, 1993, 172–195.

Barthes, Roland (1974): *Die Lust am Text*. Frankfurt a.M.: Suhrkamp.

Barthes, Roland (1968): “The Death of the Author” in: Lodge, David (eds.): *Modern Criticism and Theory*. London: Longman, 167–172.

Barthes, Roland (1994): *Roland Barthes by Roland Barthes*. Berkeley: University of California Press.

Burghardt, Manuel / Meyer, Selina / Schmidtbauer, Stephanie / Molz, Johannes (2019): “The Bard meets the Doctor” – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who”, in: *Book of Abstracts, DHd 2019* 222-225.

Dettke, Julia / Heyne, Elisabeth (2016): “Zugänge zum Spielraum der Literatur” in: Dettke, Julia / Heyne, Elisabeth (eds.): *Spielräume und Raumspiele in der Literatur*. Würzburg: Könighausen & Neumann, 11–45.

Genette, Gerard (1993): *Palimpseste. Die Literatur auf zweiter Stufe*. Frankfurt am Main: Suhrkamp. Translation of the revised second edition. [Genette, Gerard (1982): Palimpsestes. *La littérature au second degré*. Paris: Éditions de Seuil. Revised 2nd edition 1983.]

Hohl-Trillini, Regula / Sixta Quassdorf (2010): “A 'Key to all Quotations'? A corpus-based parameter model of intertextuality.” *Literary and Linguistic Computing* 10.1093/llc/fqq003.

Holthuis, Susanne (1993): *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption*. Tübingen: Stauffenburg.

Kristeva, Julia (1972): “Bachtin, das Wort, der Dialog und der Roman” in: Jens Ihwe (ed.): *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen*

Basis der Literaturwissenschaft II. Frankfurt a.M.: Athenäum, 345–375.

Kristeva, Julia (1986): "Word, Dialogue and the Novel" in: Moi, Toril (ed.): *The Kristeva Reader*. New York: Columbia University Press 35–59.

Latour, Bruno (2000): *Die Hoffnung der Pandora. Untersuchungen zur Wirklichkeit der Wissenschaft*. Frankfurt a.M.: Suhrkamp.

Nantke, Julia / Schlupkothen, Frederik (2018): "Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches' Semantic Web", in: *Book of Abstracts, DHd 2018* 345–349.

Nantke, Julia / Schlupkothen, Frederik (2019): "FormIt: Eine multimodale Arbeitsumgebung zur systematischen Erfassung literarischer Intertextualität", in: *Book of Abstracts, DHd 2019* 289–291.

Marston, John (1601): "What You Will" in Woord, Harvey (ed.): *The Plays of John Marston*. Edinburgh and London: Oliver and Boyd, 1934-1939. 2:227-295.

Maxwell, Julie / Rumbold, Kate (eds.) (2018): Shakespeare and Quotation. Cambridge: Cambridge University Press.

Pfister, Manfred (1985): „Konzepte der Intertextualität" in: Broich, Ulrich / Pfister, Manfred (eds.): *Intertextualität. Formen, Funktionen, anglistische Fallstudien*. Tübingen: Niemeyer 1-30.

Taylor, Gary (1989): *Reinventing Shakespeare: A Cultural History from the Restoration to the Present*. Oxford: Oxford University Press.

Maschinelles Lernen in den Geisteswissenschaften. Systemische und epistemologische Konsequenzen einer neuen Technologie

Seit einigen Jahren machen maschinelles Lernen und Überlegungen zu den Konsequenzen der dadurch entstehenden Artificial Intelligence Schlagzeilen. Von Spracherkennung über selbstfahrende Autos bis hin zu komplexen Spielen, maschinelles Lernen macht Computer in einzelnen Handlungsfeldern leistungsfähiger als Menschen.

In der Theorie werden drei Formen (*supervised*, *unsupervised* und *reinforcing*) des maschinellen Lernens unterschieden. Während die erste Form (*supervised*) auf Training basiert, also dem Versuch vorgegebene Resultate zu imitieren, ist das Ziel des zweiten (*unsupervised*) in einer Gesamtmasse Muster zu erkennen und zu *clustern*. Die dritte Form schliesslich (*reinforcing*) ist eine Mischung der beiden ersten Ansätze, der Lösungswege aufgrund von positiven oder negativen Rückmeldungen in eine gewünschte Richtung lenkt. Unabhängig von der Form des maschinellen Lernens stellen die Algorithmen im Handumdrehen komplexe Programme in den Schatten, die Spezialisten über Jahrzehnte hinweg entwickelt haben.

Ein Ansatz, das sogenannte *deep learning* basiert auf neuronalen Netzen, die dem menschlichen Gehirn nachempfunden sind. Sogenannte Neuronen (eigentlich Speicherbereiche) werden über mehrere Schichten vernetzt, mit Eingangs- sowie den gewünschten Ausgangsdaten konfrontiert und auf dieser Grundlage trainiert. Der Algorithmus „lernt" oder „imitiert" erwartetes Verhalten (Leifert et al., 2016).

Ebenso werden andere unüberwachte und überwachte Verfahren des maschinellen Lernens eingesetzt, um Strukturen in großen Datenmengen zu finden und die Zusammenhänge zwischen den Daten und ihnen zugeordneten Kategorien zu erkennen (z.B. Verfahren zur Dimensionalitätsreduktion, Clustering, Klassifikation, siehe einführend Alpaydin, 2014).

Die Technologien, die auf die 1980er Jahre zurückgehen, wurden lange nur testweise eingesetzt, weil die Leistungsfähigkeit der Computersysteme nicht ausreichend war (Fausett, 1993). Inzwischen lernen Maschinen mit den Methoden erfolgreich auf Gebieten, die schwer formalisierbar sind. Kommerzielle Anbieter wie Google, Amazon, Apple und Facebook implementieren *machine learning* heute schon in fast all ihren Produkten. Mit jeder Suchanfrage bei Google nutzen Menschen diese Technologie, ohne sich dessen bewusst zu sein, mit teils problematischen Folgen (Noble, 2018).

Unterschiedliche Perspektiven auf maschinelles Lernen

Das Panel hat zum Ziel, die Entwicklung und Anwendung des maschinellen Lernens mit einer Reflexion zu verbinden, die die Konsequenzen des Einsatzes aufzeigt. Dabei soll weder der häufig mit euphorischen Erwartungen verbundene Nutzen, noch unberechtigte Fundamentalabwehr befeuert werden. Vielmehr ist die differenzierte Beurteilung aus unterschiedlichen Blickwinkeln das Ziel.

Im Panel zentral gesetzt werden epistemologische Fragen, die gerade aufgrund der imitierenden Natur des maschinellen Lernens entscheidend sind für die Aufbereitung von Trainingsmaterial oder die Implementierung in Entscheidungsprozesse. Gleichzeitig ähneln die Prozesse, die die Algorithmen übernehmen Vorgehensweisen geisteswissenschaftlicher Verstehensprozesse, die unter dem Begriff der „Hermeneutik" versammelt werden. Maschinelles Lernen hat entsprechend das Potential, als Methode unsere Zugänge und den Blick auf unser Material fundamental zu erweitern.

Im Rahmen des Panels werden vier Protagonist*innen ihre Perspektive auf die Konsequenzen der Nutzung des maschinellen Lernens werfen:

DH Segment: Generischer Ansatz für historische Dokumente

Sofia Ares Oliveira (Lausanne)

Der Einsatz des maschinellen Lernens erfordert insbesondere bei der Erstellung neuer Algorithmen Fertigkeiten aus den Computerwissenschaften. Genau dieser Aufgabe stellt sich Sofia Ares Oliveira täglich, wenn sie als Ingenieurin selbständig neuronale Netze für dhlab der Eidgenössisch Technischen Hochschule in Lausanne (EPFL) erstellt. Im Rahmen des

Panels wird sie verantwortlich sein für eine kurze Einführung in maschinelles Lernen.

Aufgrund jahrelanger Beschäftigung mit der visuellen Analyse digitalisierter Dokumente, ist Ares Oliveira Spezialistin für den Aufbau und die Umsetzung neuronaler Netze zur semantischen Aufbereitung von Dokumenten (Segmentierung und Annotation). „DH segment" (Ares Oliveira et al., 2018) eine Applikation, die für die Analyse und Identifikation von Dokumententeilen genutzt werden kann, beruht auf einem eigens erstellten, trainierbaren neuronalen Netz und dient als Ausgangspunkt zu Überlegungen zum Aufbau von *machine learning* Algorithmen.

Die zwei Teilbeiträge von Sofia Ares Oliveira werden auf Englisch vorgetragen.

Machine Learning-Algorithmen für die Digitalen Literaturwissenschaften

Christof Schöch (Trier)

Anhand von Beispielen aus der jüngeren Forschung in den Computational Literary Studies (u.a. Underwood 2019 und So 2019) möchte der Beitrag aufzeigen, dass Verfahren des überwachten *machine learning* auch gewinnbringend für Fragestellungen aus den Geisteswissenschaften eingesetzt werden können, die sich nicht auf die eindeutige Zuordnung von Items zu Klassen reduzieren lassen. Als hierfür nützlich erweisen sich Zugänge zu den beim *machine learning* entstehenden Daten, die es beispielsweise erlauben, Grade der Unsicherheit zu modellieren, die Interpretierbarkeit von Algorithmen zu erhöhen oder statt der Kategorisierung das Verständnis des untersuchten Gegenstandsbereichs in den Vordergrund zu rücken. Das impliziert, dass nicht die Fragestellungen an die vorhandenen Verfahren des Machine Learning angepasst werden müssen, sondern umgekehrt, die Verfahren so eingesetzt oder modifiziert werden können, dass sie sich bestmöglich für den Erkenntnisgewinn in den Geisteswissenschaften eignen.

Gattungsstilistik und maschinelles Lernen

Ulrike Henny-Krahmer (Würzburg)

In dem Beitrag werden verschiedene Möglichkeiten vorgestellt, maschinelle Lernverfahren für die Erforschung historischer Gattungen anhand des Textstils einzusetzen, insbesondere Clustering, Klassifikation und Topic Modeling (Henny-Krahmer, 2018; Schöch et al., 2016). Dabei wird diskutiert, welche neuen Möglichkeiten sich durch die Verfahren für die Gattungsforschung ergeben (u.a. automatische Gattungsbestimmung, Untersuchung umfangreicher Textkorpora, umfassende und systematische Untersuchung von Textmerkmalen), aber auch, welche Konzepte von Gattung und Textstil durch maschinelle Lernverfahren in den Vordergrund rücken, wodurch der Anschluss an neuere gattungstheoretische Diskussionen (z.B. Gattungen als literarisch-soziale Institutionen, Familienähnlichkeitsbeziehungen in Gattungen, siehe dazu Hempfer, 2010; Voßkamp, 1977) nicht immer gegeben ist. Am Beispiel der Gattungsstilistik soll so aufgezeigt werden, wie maschinelles Lernen die Möglichkeiten empirischer Untersuchungen in den Geisteswissenschaften erweitern kann, aber auch wie sich der Erkenntnisgewinn auf bestimmte sprachlich-formale textuelle Aspekte konzentriert.

Ground-Truth und Fragen der geisteswissenschaftlichen Datenaufbereitung

Tobias Hodel (Bern)

Im Rahmen von Projekt READ wurde mit der Einführung von maschinellen Lernverfahren die Erkennung von Handschriften und alten Drucken markant verbessert. Da die neuronalen Netze auf Trainingsmaterial basieren (also *supervised* sind), müssen Fragen nach der Aufbereitung gestellt und eine Verständigung epistemologischer Grundannahmen, insbesondere nach dem Konzept der „Ground-Truth" untersucht werden. Solche Diskussionen bilden einerseits eine Aussensicht auf die verwendeten Algorithmen, andererseits lassen sich Vorstellungen aus den Disziplinen (Germanistik, Geschichte, Editionswissenschaften) kritisch in den Blick nehmen.

Die Panelisten werden kurz und thesenhaft ihre Perspektive auf die Technologie darlegen, dabei sollen sie u.a. zu drei Komplexen Stellung nehmen:

Chancen und Grenzen der Technologie

Wo wird der Einsatz der Technologie in den Geisteswissenschaften neue Erkenntnisse bringen, welche Dokumente/Materialien/Daten eignen sich nicht für die Behandlung mit *machine learning* Algorithmen? Inwiefern ähnelt oder unterscheidet sich der Einsatz der Technologie von hermeneutischen Prozessen?

Epistemologische Konsequenzen (für die DH/geisteswissenschaftliche Disziplinen)

Fragen nach Erkenntnismöglichkeiten werden in diversen geisteswissenschaftlichen Disziplinen seit Jahrzehnten diskutiert. Die Nutzung von Algorithmen des maschinellen Lernens erfordern jedoch klare Aussagen zur untersuchten Materie, unabhängig davon, ob es sich um *supervised* oder *unsupervised* Zugäng handelt (was ist Text, was soll identifiziert werden, welche Einheiten sind sinntragend etc.). Der Einsatz des maschinellen Lernens zwingt entsprechend zur Offenlegung von Konzepten und Vorstellungen.

Regeln zur Nutzung von *machine learning* Algorithmen

Neben der Angst vor dem Kontrollverlust und etwaigem Rückgang von Arbeitsplätzen oder der Überwachung von Menschenmassen, sind es nicht zuletzt Skandale zur Verletzung der Privatsphäre, die in den vergangenen Monaten zum Ruf nach der Regelung des Einsatzes der Technologie führten (Lobo, 2019). Neuste Forschungen zeigen, dass die Vorstellungen einer ethischen AI stark divergieren (Jobin et al., 2019). Ethische Regelungen sind in den Geisteswissenschaften unüblich, gerade deshalb sind die Diskussionen im Umfeld der Technologie fruchtbar. Inwiefern besteht ein Zusammenhang zwischen geforderter Diversifizierung der DH mit der Anwendung der Technologie?

Maschinelles Lernen in den Digital Humanities

Im wissenschaftlichen Bereich sind es zurzeit vor allem die angewandte Informatik und Mathematik sowie die Computerlinguistik, die maschinelles Lernen in ihre Forschungen integrieren. In den Digital Humanities spielt die Technologie bislang von wenigen Zentren abgesehen eine untergeordnete Rolle. In absehbarer Zeit dürfte sie ein wichtiger Teil der Disziplin werden – nicht nur im Recherche –, sondern auch im Auswertungs- und Schreibprozess. Insbesondere im Umgang mit digitalisierten Dokumenten, großen Datenmengen und Bildquellen können neuronale Netze ein wichtiges Mittel sein, um Daten zu finden, zu sortieren und auszuwerten.

Die digitalen Geisteswissenschaften umfassen mit ihrem Methodenapparat sowohl komplexe Softwareentwicklung, als auch die Anwendung statistischer Modelle und das Erklären mit hermeneutischen Verfahren. Daher ist die Disziplin prädestiniert in den Diskussionen dieser gesellschaftsverändernden Technologie eine Vorreiterrolle einzunehmen.

What's in the news? (Erfolgs-)Rezepte für das wissenschaftliche Arbeiten mit digitalisierten Zeitungen

Einleitung

Why newspapers? Wie geht man methoden-kritisch mit digitalisierten Zeitungskorpora um? Welcher technische Aufwand ist nötig? Wie gestaltet sich die Zusammenarbeit mit Bibliotheken zum Thema Verfügbarkeit und Lizenzierung? Die zahlreichen Panels, Präsentationen und Poster auf der DH2019 haben gezeigt, dass ein großes Interesse seitens der Wissenschaft und auch der Bibliotheken besteht umfangreiche Zeitungs- und Zeitschriftensammlungen für die Öffentlichkeit und Wissenschaft digital verfügbar zu stellen (vgl. ADHO). Dabei konzentrierten sich die Debatten aber überwiegend auf die Provider-Perspektive, d.h. welche Herausforderungen sind im Rahmen des Digitalisierungsprozesses und der Zusammenarbeit mit wissenschaftlichen Institutionen zu überwinden. Im Spielraum der Wissenschaft bedeutet diese Erzeugung der Forschungsdaten, Zeitungen als Primärtexte aufgrund ihrer Menge, Zeitspannen und textuellen und thematischen Vielfalt nutzen zu können, um temporale, grenzübergreifende, multilinguale und -modale Querbezüge zu erstellen. Diese Forschung bietet die Möglichkeit operative, methodische und organisatorische Herausforderungen anzugehen, um innovative computergestützte Modelle, Tools, Codes, Daten und Infrastrukturen zu entwickeln. Die Panelvorträge geben Einblicke in Forschungsprojekte zu digitalisierten Zeitungen, die mit unterschiedlichen Korpora und Fragestellungen, aber teils ähnlichen Verfahren und Forschungsdesigns an umfangreichen historischen Zeitungssammlungen arbeiten. Dabei sollen die folgenden Kategorien angesprochen werden: Herausforderungen, Forschungsfragen, Methoden (inklusive Tools), Teamkomposition, Korpusbeschreibung, Projektformat und Projektdauer. Das Ziel dieses Panels ist es gemeinsam das Generalisierungspotential der Methoden und den wissenschaftlichen Output zu diskutieren, um eine Zutaten- und Werkzeugliste für das wissenschaftliche Arbeiten mit digitalisierten Zeitungen zu generieren. Entsprechend der vorgestellten Beiträge sollen gemeinsame Fragen diskutiert werden wie:

- Wie ist das Verhältnis von Korpuskomposition und Fragestellung in den jeweiligen Projekten?
- Welche forschungsspezifischen Hürden oder Fallstricke gab es in den einzelnen Projekten?
- Welche geistes- und kulturwissenschaftlichen Fragen können in Anforderungen in digitale Methoden übersetzt werden und wie werden diese im digitalen Umfeld operationalisiert?
- Wird die Fragestellung ggf. durch ein Interface beeinflusst, oder andersherum: lässt sich das Interface der Fragestellung anpassen?
- Wie können wir sicherstellen, dass wir nicht nur projektspezifische Tools bauen und Methoden entwickeln, sondern diese auch für weitere Anwendungen in der Wissenschaft und Öffentlichkeit nachnutzbar machen?

Panelvorträge

NewsEye Case Study: Rückkehrmigration in österreichischen Tageszeitungen zwischen 1850 und 1950

Sarah Oberbichler

Ein nicht unbedeutender Teil jener Menschen, die freiwillig oder unfreiwillig zwischen 1850 und 1950 ihre Heimat verlassen hatten, kehrten in ihr Ursprungsland zurück. In österreichischen Tageszeitungen wurde regelmäßig über die Rückkehr von freiwilligen Auswander*innen, in Gefangenschaft geratenen Soldaten oder Flüchtlingen berichtet, weshalb dieses Medium eine geeignete Quelle für die Erforschung des bis dato vernachlässigten Themas der Rückkehrmigration darstellt. Im Rahmen des NewsEye Projektes wurden deshalb folgende Forschungsfragen aufgegriffen: Wie und in welchem Kontext wurde in österreichischen Tageszeitungen über Heimkehrer*innen berichtet und wie hat sich die Berichterstattung im Laufe der Zeit verändert? Das zur Beantwortung der Fragestellung notwendige Korpus wird mithilfe des Online Zeitungsarchives der Österreichischen Nationalbibliothek ("ANNO") erstellt und für die weitere Anwendung von Text-Mining Methoden und qualitativen, diskursanalytischen Analysen aufbereitet. Gerade aber dieser erste Schritt – die Bildung des Korpus – bringt eine Reihe von Herausforderungen mit sich. Nicht alle Suchbegriffe führen zu eindeutigen Ergebnissen und fehlende Speicher- und Download-Optionen führen zu langen und aufwendigen "Copy und Paste"-Verfahren. Vor diesem Hintergrund stellt sich die Frage, welche Methoden und Ansätze (beispielsweise Article Separation, Topic Modeling oder Wort Embeddings) für die Bildung eines Korpus herangezogen werden können, wenn manuelle Vorgehen ei-

nen zu großen Zeitaufwand darstellen. Ebenfalls stellt sich die Frage, wie viel Spielraum zwischen vorgegeben Funktionen und individuellen Einstellungen zielführend ist.

More than a Feeling: Media Sentiment as a Mirror of Investors' Expectations at the Berlin Stock Exchange, 1872-1930

Lino Wehrheim / Bernhard Liebl

Das Verhalten von Finanzinvestoren wird nicht nur durch Fundamentalwerte wie etwa künftige Zahlungsströme, sondern auch durch „weiche" Faktoren wie Stimmungen, Launen und Gefühle beeinflusst. Entsprechend hat sich in der Finanzmarktforschung das Konzept des „Investor Sentiment" etabliert, was als eine (individuelle oder kollektive) Einstellung in Bezug auf künftige Marktentwicklungen verstanden werden kann, die nicht auf rationaler Abwägung basiert. Das Ziel des Projekts ist es, die Bedeutung von Sentiment für die Berliner Börse zwischen 1872 und 1930 zu erfassen, dem bedeutendsten deutschen Finanzplatz dieser Zeit. Inwieweit beeinflussten historische Erfahrungen wie Kriege oder politische Ereignisse die Stimmung von Finanzinvestoren, und welchen Einfluss übte Sentiment auf die Entwicklung historischer Börsenkurse aus? Hat sich dieser Einfluss im Zeitverlauf verändert? Um die Stimmung an der Berliner Börse zu quantifizieren, wird auf Basis historischer Zeitungsartikel ein Sentiment-Index erstellt, ein Ansatz, der seit der Arbeit von Tetlock (2007) verbreitet Anwendung findet, so etwa bei Ferguson et al. (2015), García (2013) und Hanna et al. (2017). Konkret werden wörterbuchbasierte Verfahren sowie Ansätze des maschinellen Lernens herangezogen. Besondere Bedeutung erhält die Generierung eines domain-spezifischen Sentiments-Wörterbuchs (Finanzmarktdeutsch des 19. Jahrhunderts). Um den Sentiment-Index um die Komponente medialer Narrative zu ergänzen, werden die Zeitungsartikel zusätzlich mit Topics Models ausgewertet. Das zugrundeliegende Korpus besteht aus Artikeln der Berliner Börsen-Zeitung, die in täglichen Marktberichten über das Geschehen und die Stimmung am Finanzplatz Berlin berichtete.

"Horizontales Lesen" als digitale Analysemethode von Zeitungskritiken

Torsten Roeder

Zeitungen und Zeitschriften sind spätestens seit dem 19. Jahrhundert - bis in die Jetztzeit - ein Medium für öffentliche Debatten über Kunst und Kultur. Während historische Untersuchungen sich vor nicht allzu langer Zeit noch aufgrund der diffusen Quellenlage auf Einzelfalluntersuchungen beschränken mussten, macht aktuell die immense Menge an Material, das durch die Digitalisierung verfügbar ist und wird, die Entwicklung und Anwendung neuer Verfahren notwendig. Ein Ansatz besteht in dem Verfahren des "horizontalen" Lesens, mit dem sich thematisch zusammenhängende Texte zu einem (historischen) Meinungsspektrum anordnen lassen. Den "Named Entities" fällt dabei eine Schlüsselrolle zu, da diese die zentralen Vergleichspunkte liefern. Anhand einer ausgewählten Entity (z.B. der Titel eines musikalischen Werkes) und aller Textausschnitte, die sich direkt darauf beziehen, kann manuell und ggf. mithilfe übertragener Anwendung von Sentimentanalyse oder ggf. Topic Modeling die gesamte Bandbreite zu einem historischen Zeitpunkt oder an einem historischen Publikationsort abgebildet werden. Voraussetzungen dafür sind jedoch eine hochwertige Texterschließung und semantische Annotationen, bestenfalls mit Normdaten versehene Eigennamen von Personen, Orten, Werken etc. p.p. Während diese Methode durchaus verwertbare Ergebnisse produziert, bleibt für alle Anwendungsfälle, an welcher Stelle für die notwendige Qualität der Daten Sorge zu tragen ist: Kann dies bereits durch Provider geschehen, oder muss dies notwendigerweise - ggf. auch abgestuft - im jeweiligen Forschungsprojekt geschehen?

Oceanic Exchanges: Transnationale Textmigration

Jana Keck

Im 19. Jahrhundert entstand die Massenpresse. Die technischen Innovationen der Druckpressen, fehlende Regulierungen der Gesetzgebung und Durchsetzung des Urheberrechts und das wachsende Interesse der Bevölkerung an Informationen weltlicher, sensationeller und politischer Natur schuf eine globale Kultur reichhaltiger, schnell zirkulierender Informationsquellen. Auch wenn dies auf den grenzübergreifenden Charakter der Presse hinweist, wurde die Zeitungsforschung bisher weitgehend in Metropol- und nationalen Räumen definiert. Im Rahmen des internationalen Forschungsprojektes "Oceanic Exchanges" wurden über 100 Millionen digitalisierte Zeitungsseiten aus mehr als 7 Nationen gesammelt, um den transnationalen Charakter der Presse im 19. Jahrhundert zu untersuchen. Welche Texte und Ideen – literarisch, politisch, wissenschaftlich, wirtschaftlich, religiös – zirkulierten im öffentlichen Raum? Wie wurden diese Texte in dem jeweiligen Raum und Sprache übersetzt und modifiziert? Text Reuse Detection hat sich schon trotz „noisy" OCR als erfolgreiche Text-Mining Methode erwiesen, um ähnliche Textpassagen in umfangreichen Datenbanken digitalisierter englischsprachiger Zeitungen zu erkennen und zu modellieren (Viral Texts Project). Das Ziel dieses Beitrages ist zu zeigen, welche methodischen Möglichkeiten, aber auch Herausforderungen beim Erkennen und Modellieren von Reprints in multi-lingualen Sammlungen und der zeitlichen Klassifizierung entstehen. Damals wie heute, nutzen diese Reprinting-Praktiken die Paradoxien eines jeden internationalen Mediensystems: scheinbare Verbundenheit und doch beständige Distanzen.

Digitale Ideengeschichte: der antimoderne Diskurs über Europa in der schweizer Presse (1900-1945) (Estelle Bunout / Marten Düring)

Estelle Bunout / Marten Düring

Die Diskussionen über Europa haben sich in der Schweiz in der ersten Hälfte des 20. Jahrhunderts auf mehreren Ebenen entwickelt: über die Initiativen zum Aufbau eines institutionellen Europas oder noch über die Rolle, die die Schweiz dabei spielen sollte. Ein Blick auf die Presse gibt eine breitere Perspektive auf diese Diskussionen. Vertreter aus unterschiedlichen Bereichen wie der Mathematik, mit Sophie Piccard (1904–1990), Verfechterin der Paneuropa-Union und der Literatur, mit Gonzague de Reynold (1880–1970) verteidigen jeweils eine eigene Vision für Europa. De Reynold, ein bekannter Antimodernist, versuchte ein Europa der Aristokratie und des Korporatismus gegenüber dem liberalen, demokrati-

schen Europa zu verteidigen und war in den Zeitungsredaktionen gut eingebunden. In diesem Zusammenhang gehen wir der Frage nach ob es besondere Bemühungen von Antimodernisten gab, die Diskussion über Europa wieder anzueignen und die Assoziation Europas mit Frieden, Progressivismus und Aufklärung im Kontext ihrer Weltanschauung neu zu gestalten. Diese Frage wird anhand der digitalisierten Schweizer Presse bearbeitet, die im Rahmen des impresso Projekts mit NLP bereichert wurde und durch eine forschungsorientierte Oberfläche zugänglich gemacht wurde. Diese Infrastruktur und weitere NLP-Methoden helfen bei der Erstellung einer Sammlung von Artikeln, die diesen antimodernen Diskurs zu Europa beinhalten, was durch gewöhnlichen Stichwortsuche nicht möglich wäre.

Bibliographie

ADHO: https://dh2019.adho.org/conftool/ [letzter Zugriff 31. August 2019].

ANNO (AustriaN Newpapers Online): http://anno.onb.ac.at/ [letzter Zugriff 31. August 2019].

Crymble, Adam (2016): "Digital Library Search Preferences amongst Historians and Genealogists: British History Online User Survey", in: *Digital Humanities Quarterly* 10(4) http://www.digitalhumanities.org/dhq/vol/10/4/000270/000270.html [letzter Zugriff 31. August 2019].

Ehrmann, Maud / Bunout, Estelle / Düring, Marten (2019): "Historical Newspaper User Interfaces: A Review", in: *IFLA* http://library.ifla.org/2578/1/085-ehrmann-en.pdf [letzter Zugriff 31. August 2019].

Ferguson, NJ et al. (2015): "Media Content and Stock Returns: The Predictive Power of Press", in: *Multinational Finance Journal* 19(1): 1–31.

García, D. (2013): "Sentiment during Recessions", in: *The Journal of Finance* 68(3): 1267–1300.

Hanna, Alan J. / Turner, John D. / Walker, Clive B. (2017): "News Media and Investor Sentiment Over the Long Run", in: QUCEH Working Paper Series, Working Paper No. 2017-06.

impresso: https://impresso-project.ch/ [letzter Zugriff 31. August 2019].

Koenen, Erik (2018): "Digitale Perspektiven in der Kommunikations- und Mediengeschichte", in: *Publizistik* 63(4): 535–56 https://doi.org/10.1007/s11616-018-0459-4 [letzter Zugriff 31. August 2019].

NewsEye: https://www.newseye.eu/ [letzter Zugriff 31. August 2019].

Oceanic Exchanges: http://oceanicexchanges.org/ [letzter Zugriff 31. August 2019].

Tetlock, PC (2007): "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", in: *The Journal of Finance* 62 (3): 1139–1168.

Viral Texts: https://viraltexts.org./ [letzter Zugriff 31. August 2019].

Wijfjes, Huub (2017): "Digital Humanities and Media History. A Challenge for Historical Newspaper Research", in: *Tijdschrift Voor Mediageschiedenis* 20(1): 4-24–24 https://doi.org/10.18146/tmg20277 [letzter Zugriff 31. August 2019].

Vorträge

Anwendungen von DH-Methoden in der Erschließung und Digitalisierung von Kulturerbe. Ein Vorschlag zur Systematisierung

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

Bevor Kulturerbe digital verfügbar und nachnutzbar ist, stehen komplexe Prozesse an, die als spezifische „invisible work“ (Star/Strauss 1999) bezeichnet werden können. Diese werden kaum außerhalb der engeren Community problematisiert: Wo finde ich in den analogen Katalogen, Zettelkästen, Findbüchern u.ä. die notwendigen Informationen zur Erfassung und Erschließung? Welche Sammlungsbestände sind besonders wichtig und deshalb zu digitalisieren? Ist eine Massen-Digitalisierung mit geringer Detailtiefe der Erfassung oder eine detaillierte Erschließung einiger Teilbestände sinnvoll? Wie können die Vorgehensweisen von unterschiedlichen Personen und Institutionen vereinheitlicht werden? Und wie können Informationen, die bereits digital vorliegen – etwa in Tabellen oder älteren Erfassungssystemen – an internationale Regelwerke und Standards zur Metadatenhaltung angepasst werden?

Kernanliegen des vorgeschlagenen Beitrags ist es, für die Sammlungserschließung und Digitalisierung – gerade in Museen, aber auch in Archiven und Bibliotheken – Methoden, Tools und Analyseperspektiven der Digital Humanities stärker als bisher zu nutzen. Dafür werden mögliche Synergien und exemplarische Anwendungen in einer ersten Exploration aufgezeigt. Dabei wird auf eigene Projekterfahrungen zurückgegriffen, die im Kontext des Digitalisierungsvorhabens „Digitales Portal Alltagskulturen im Rheinland“ (2013-2017) gesammelt wurden.[1] Diese sind bereichert um Perspektiven des DH-Forschungsverbundes an der Universität Hamburg, „Automatisierte Modellierung hermeneutischer Prozesse (hermA)“ (2017-2020).[2]

What (not) to do with 100.000 Pictures. Sammlungserschließung als Sisyphos-Arbeit?

Im LVR-Institut für Landeskunde lagern umfangreiche Bestände, die nach Sammlungsgeschichte gegliedert und nur rudimentär erschlossen sind. Vor allem die Fotodokumentationen sind für Erschließung und Digitalisierung prädestiniert, zeigen sie doch plurale Facetten immateriellen Kulturerbes seit etwa 1900[3]. Dazu kommen umfangreiche Materialien zu 31 schriftlichen Befragungen aus den 1970er bis 2000er Jahren sowie kleinere Einzelsammlungen. Die Fotobestände liegen in unterschiedlichen Negativformaten sowie als Dias chronologisch sortiert vor. Dazu sind Fotoabzüge auf Karteikarten geklebt, mit Metadaten versehen und thematisch abgelegt in einer über Jahrzehnte gewachsenen Systematik. Negative und Abzüge sind also unterschiedlich archiviert und nur in Einzelfällen einander zuzuordnen. Es bestehen auf Einzelplatzrechnern gepflegte Bestandslisten sowie eine Datenbank[4] mit Teilinventarisierung. Seit 2013 werden im Rahmen der LVR-weiten Digitalisierungsstrategie Digitalisate erstellt, mit angereichert und Metadaten abgelegt, um sie mittelfristig öffentlich zugänglich zu machen.[5] Die Systeme werden aktuell in die Datenbank digiCULT.web[6] übertragen, wodurch mit dem LIDO-Format größere Datenmengen standardkonform publiziert werden können.

Vergleichbare gewachsene Systeme gibt es in vielen Museen, Archiven und Bibliotheken. Gerade in kleineren Museen und Forschungseinrichtungen, deren Hauptaugenmerk auf der Ausstellung bzw. Forschung und weniger auf der Sammlungserschließung liegt, sind die Erfassungen unvollständig und in der Institutionsgeschichte von verschiedenen Akteuren mit spezifischen Wissenshintergründen und Systematiken erfolgt. Diese Systeme funktionieren vor allem aufgrund des Wissens von Archivar*innen und Forschenden (vgl. zu Wissenskonzepten Koch 2006): Sie wissen, welche Materialien wo liegen, welche Strukturen was auffindbar machen, was in welchen Kontexten verwendet wurde. Wenn Stellenwechsel oder Verrentungen anstehen, geht dieses Wissen verloren oder wird bruchstückhaft weitergegeben.

Wenn analoge Bestände digitalisiert werden sollen, geht es in der Vorbereitung (z.B. der Antragsstellung für Drittmittel) um hohe technische Qualität, Fragen des Workflows sowie der Bereitstellung, einzuhaltende Standards oder eigene Präsentationsoberflächen. Die konkreten Arbeitsschritte zur Umsetzung dieser Zielvorgaben werden oft erst in der Projektrealisierung entwickelt. Gerade die Museumsdatenbanken sind geprägt von einer gewachsenen Pluralität, die in dieser Form nicht als veröffentlichungswürdig gelten und strukturelle Nachbearbeitungen erfordern.

Für die hier exemplarisch stehenden Fotobestände des LVR-Instituts fiel die Entscheidung, sich nicht an den bestehenden (Ablage-)Systematiken zu orientieren. Neben anderen Argumenten war dabei die unvollständige und heterogene, thematisch abgelegte Erschließung per Karteikarte ausschlaggebend. Um eine Auswahl zur inhaltlichen Erschließung zu ermöglichen und analoge Vorarbeiten gering zu halten, wurden alle Negative und Dias digitalisiert (zur Problematik von Original und Kopie, die sich für Foto-Abzüge spezifisch stellt, vgl. Schönholz 2017). Im Zuge der Auftragsvergabe wurden Bestände erstmals gezählt und liegen als vollständige Digitalisate vor. Ein Meilenstein dieser Fleißarbeit waren die im Vergleich und Überblick erstmals digital verfügbaren Bildbestände.

Doch was macht man mit 100.000 Fotos, die außer einer Dateibenennung keinerlei Informationen mit sich bringen? Hier werden die eingangs aufgeworfenen Fragen konkret. Vergleichbar mit den Diskussionen um Distant Reading (Moretti 2007 und 2016; Crane 2006) stellt sich ob der Verfügbarkeit der Digitalisate die Frage, wie diese zielführend erschlossen werden können. Wie findet man relevante Bildbestände für

eine Tiefenerschließung? Wo helfen die bestehenden Metadaten zielführend?

Maschinelle Unterstützung nutzen, aber wie?

Für diese Arbeitsschritte sind Methoden der Digital Humanities vielversprechend. Zwar wird zunehmend die Frage nach der Nutzung von Digitalisaten als Open Data für die Wissensproduktion[7] diskutiert, oft jedoch erst nach Abschluss der Erschließungsarbeiten. Nicht erst die publizierten Daten digitalen Kulturerbes können mit DH-Verfahren erforscht werden – diese sind bereits in der Erschließung enorm hilfreich. Zwei Verfahren aus dem Bereich des Machine Learning scheinen besonders erfolgsversprechend: maschinelle Bilderkennung sowie die Analyse bestehender Metadaten mittels Text Mining.

Die großen Mengen unerschlossener Fotos können mit maschineller Bilderkennung hinsichtlich ihrer Ähnlichkeit gruppiert werden, wie es etwa PixPlot realisiert:[8] Bildanordnungen im Vektorraum machen Schwerpunkte des Bestandes deutlich, außerdem lassen sich Subkorpora bilden, die mit Massenbearbeitung formal erschlossen werden können. In der Arbeitspraxis ist daneben das Identifizieren von Dubletten relevant: Wenn beispielsweise Abzüge des Archivs in der Vergangenheit abfotografiert wurden, existiert das Foto in unterschiedlichen Kopien im digitalisierten Bestand – manuell eine Suche nach der Nadel im Heuhaufen, automatisiert mit Bildvergleich gut zu identifizieren (vgl. die vielversprechenden Ansätze bei Schneider 2019). Auch ähnliche Aufnahmen, z.B. aus einer Bildserie, sind so zuzuordnen. Die inhaltliche Erschließung wird durch eine Ähnlichkeitssuche ebenfalls deutlich vereinfacht: Hat man etwa eine gute Aufnahme eines Gegenstandes, so lassen sich relativ eindeutig andere Abbildungen dessen im Bestand finden. Hier wäre jedoch eine menschliche Intervention (zumindest zu Beginn eines möglichen Active Learning-Verfahrens) aufgrund der feinen Unterschiede notwendig. Zudem sind zweifelsfrei die Trainingsdaten von großer Relevanz und sollten wo möglich aus bereits erschlossenen, vergleichbaren Kulturerbe-Datensätzen bestehen.

Text Mining-Verfahren würden in GLAM-Datenbanken nicht nur aus dem Museumsbereich präzisiere Suchabfragen und gerade in bestehenden Erfassungen systematische Funde ermöglichen. Ansätze wie facettierte Suchmasken mit Linked Open Data, wie sie in Plattformen zur Datenpräsentation zunehmend realisiert werden,[9] wären im Backend enorme Arbeitserleichterungen. Schon banale Automatisierungen wie Rechtschreibkorrektur und Vereinheitlichungen der Formalerschließung sind momentan in der Regel nicht vorgesehen. Sie kosten viel Zeit und Konzentration, sobald nicht simple Ersetzungen vorzunehmen sind. Gleichzeitig fehlt den Entwickler*innen der eingesetzten Datenbanken die zielgenaue Kollaboration mit entsprechender DH-Forschung zu konkreten Tools und Verfahren sowie der Testung von verschiedenen Funktionen für eine hohe Qualität der Ergebnisse, die vor der Übernahme in die Infrastruktur erfolgen muss. LOD wird zudem aktuell nur in Ausnahmen direkt in die Erfassungssysteme eingebunden – erst so könnte die Arbeit an Ontologien und konkreten Datensätzen gezielt verbunden werden. Dazu kommt die Notwendigkeit, die Erfassungen besser zu vernetzen; auch in Fällen, in denen dies erst nach der Veröffentlichung möglich oder notwendig wird. Hier sollte eine Öffnung für fortlaufende kollaborative Ergänzung und Korrektur von Daten in Verbindung mit der Erfassung von Paradaten (McIlvain 2013) geschaffen werden.

Viel Zeit wird mit der Erzeugung von metacrap (Doctorow 2001) verbracht. Die messy Metadaten, die für viele Erfassungen – oft im Backend, aber auch publiziert – bestehen, sind durch Tools einfach zu identifizieren und automatisch zu beheben: Museum Analytics[10] beispielsweise ermöglicht es, große Mengen von Museumsdaten zu analysieren, ist allerdings für publizierte Metadaten vorgesehen. Gerade in der Migration zwischen Datenbanksystemen sowie für den internen Gebrauch zwecks Qualitätskontrolle, Bestandssichtung und Entscheidung über Nachbearbeitungen vor der Veröffentlichung erlaubt dieses einen anderen Blick auf die Bestände. Das Tool Breve[11], das Tabellen visualisiert, könnte ergänzende Funktionen übernehmen. Die Entwicklungsvorhaben des Verbundprojektes GND4C[12] oder des Projekts Qrator[13] sind richtungsweisend, leider aber noch nicht verfügbar, und entsprechende Konferenzworkshops zu DH für Gedächtnisinstitutionen (Döhl/Voges 2019) erfreulich.

Erste Ansätze zur Nutzung von DH-Analyseverfahren werden auch von Museumsseite diskutiert, etwa hinsichtlich Netzwerkanalyse der Sammlungsbestände (Werner 2019) oder Möglichkeiten der Visualisierung (Mayr/Windhager 2019), bilden dort jedoch (noch) die absolute Ausnahme. Dies spiegelt sich auch in den Programmen der entsprechenden Tagungen wie „Museums and the Internet“[14] oder denen der Fachgruppe Dokumentation des Deutschen Museumsbundes[15]. Falls entsprechende Ansätze bereits genutzt werden, so geschieht dies wiederum weitestgehend als „invisible work“ (s.o.) ohne Darstellung in der (Forschungs)Öffentlichkeit. In weiten Teilen der entsprechenden GLAM-Community wird gerade von Museumsseite die DH noch zu wenig als möglicher Kooperationspartner wahrgenommen, um entsprechende Workflows und Implementierungen zu konzipieren.

Die vorgestellten Zugänge könnten im Ergebnis der Implementierung nicht nur aufzeigen, welche Daten in einem Bestand enthalten sind, sondern auch, welche Leerstellen in der Erfassung noch geschlossen werden sollten. Von einer linearen Durchsicht, Überarbeitung und Freigabe der Datenbank-Einträge kann mit entsprechender Tool-Unterstützung – und einhergehender Interoperabilität! – zu einer gezielten Nachbearbeitung von Teilbeständen oder Vereinheitlichung einzelner Metadatenfelder übergegangen werden. So bleibt mehr Zeit für eine inhaltliche Erschließung und Analyse sowie die dringend notwendige epistemologischen Reflexionen dieser Prozesse.

Fazit: DH-Verfahren in Sammlungsdatenbanken

Was fehlt in der Gesamtschau momentan? Vor allem die Öffnung von Erschließungssystemen für die dargestellten Methoden sowie die Öffnung der entsprechenden Communities zueinander.

Erforderlich ist dabei der frühzeitige Einbezug von bestehenden Tools und Analyseverfahren – lange vor der Veröffentlichung der Datensätze. Gerade in der Exploration weitestgehend nicht erfasster Bestände zur Vorbereitung der Erschließung und in der Qualitätskontrolle von Metadaten lie-

gen große Potentiale, die noch zu wenig genutzt werden. Wenn gleichzeitig bereits tiefenerschlossene Bestände als Trainingsdaten genutzt werden, können die Ansätze auch unabhängig von der Nutzung konkreter Datenbanken gegenseitigen Mehrwert sowohl in der Methodenentwicklung als auch in der Erschließung bringen.

Eine öffentliche Finanzierung und Weiterentwicklung von den entsprechenden Tools ist dabei dringend notwendig. Statt weiter in gewinnorientierte Software zu investieren, sollten Genossenschaften und Vereine gegründet und ausgebaut werden. Eine Unterstützung der Toolentwicklung durch entsprechend kompetente DHler*innen ist vielversprechend. So wäre etwa ein Hackathon zur Erweiterung von Erschließungssystemen eine Möglichkeit, um über das Tagesgeschäft hinausgehende Innovationen umzusetzen. Entsprechend erweiterte Datenbanken sollten viel häufiger auch in der Forschung verwendet werden, die aktuell noch viel zu oft in Form von Excel-Sheets (weiter-)arbeitet, obwohl Datenbanken mit erweiterten Funktionen existieren. Mit einfachen Import/Exportfunktionen innerhalb der Tools und Verbindungen zu Analyseverfahren könnten Forschungsumgebungen geschaffen werden, in denen kollaboratives Arbeiten gleichzeitig die Datensätze anreichert und Forschungsfragen beantwortet.

Fußnoten

1. Projektergebnisse unter https://alltagskulturen.lvr.de/. Das DFG-geförderte Projekt wurde durch die Autorin koordiniert, von Dagmar Hänel geleitet und maßgeblich auch durch den wissenschaftlichen Dokumentar im Projekt, Christian Baisch, vorangetrieben.
2. Das von Gertraud Koch und Heike Zinsmeister geleitete Verbundprojekt befragt DH-Perspektiven auf Möglichkeiten zur Modellierung hermeneutischer Prozesse. Vgl. https://www.herma.uni-hamburg.de/.
3. Vgl. Sammlungsbeschreibungen unter https://alltagskulturen.lvr.de/de/sammlungen .
4. Faust Software, vgl. https://www.land-software.de/.
5. Genutzt wird eine Weiterentwicklung von MediaFiler, vgl. https://mediafiler.com/en .
6. digiCULT.web als entitätsbasierte Online-Datenbank ist vorrangig für Museumsbestände entwickelt und baut auf CIDOC-CRM auf. Vgl. https://www.digicult-verbund.de/de/digicultweb. Zu LIDO vgl. http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/.
7. Vgl. etwa die Schwerpunktsetzung „Open Data – now what?“ der Sharing is Caring-Konferenz 2019. http://sharecare.nu/programme/. Danke an Samantha Lutz für den Hinweis.
8. Vgl. https://github.com/YaleDHLab/pix-plot . Vgl. auch Leonard 2019.
9. Vgl. etwa Suchfacetten der DDB, https://www.deutsche-digitale-bibliothek.de/, mittlerweile auch der Europeana, https://www.europeana.eu/.
10. https://www.max.gwi.uni-muenchen.de/.
11. http://hdlab.stanford.edu/breve/.
12. Ziel 3 des Projektes ist die „Bereitstellung von Schnittstellen und Werkzeugen zur Unterstützung nicht-bibliothekarischer Anwendungskontexte.“ Vgl. https://wiki.dnb.de/pages/viewpage.action?pageId=134055796 . Danke an Axel Vitzthum für den Hinweis.
13. Vgl. https://qurator.ai/. GLAM-Institutionen mit digitalem Kulturerbe sind hier ein Anwendungsfall.
14. Bisher waren entsprechende Beiträge bei der Tagung absolute Ausnahme. Vgl. das Archiv unter https://mai-tagung.lvr.de/de/startseite.html.
15. Das Archiv der Tagungsprogramme und Vortragsfolien lässt nicht auf entsprechende Diskussionen schließen. Vgl. https://www.museumsdokumentation.de/?lan=de&q=Who%20is%20who/FG%20Dokumentation%20im%20DMB/Tagungsarchiv.

Bibliographie

Crane, Gregory (2006): “What Do You Do with a Million Books?” In: D -*Lib Magazine* 12/3. http://www.dlib.org/dlib/march06/crane/03crane.html.

Doctorow, Cory (2001): “Metacrap. Putting the Torch to Seven Straw-Men of the Meta-Utopia.” https://people.well.com/user/doctorow/metacrap.htm.

Döhl, Frédéric / Voges, Ramon (2019): „Erklärt und ausprobiert – Digital Humanities für Gedächtnisinstitutionen.“ Workshop im Rahmen der Tagung *Zugang gestalten 2019*. https://zugang-gestalten.org/dokumentation-2019/.

Koch, Gertraud (2006): „Die Neuerfindung als Wissensgesellschaft. Inklusionen und Exklusionen eines kollektiven Selbstbildes.“ In: Hengartner, Thomas; Moser, Johannes (Hg.): *Grenzen & Differenzen. Zur Macht sozialer und kultureller Grenzziehungen*. Leipzig, S. 545–559.

Leonard, Peter (2019): “Large images dataset overtime: PixPlot new features.” In: *Culture Analytics Workshop: Time Series, Digital Humanities 2019*. https://dev.clariah.nl/files/dh2019/boa/1079.html und https://github.com/Culture-Analytics/DH2019.

Mayr, Eva / Windhager, Florian (2019): „Vor welchem Hintergrund und mit Bezug auf was? Zur polykontexturalen Visualisierung kultureller Sammlungen“. Vortrag im Rahmen der Tagung *Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Zeitalter*. Folien unter http://objekte-im-netz.fau.de/projekt/sites/default/files/2019-11/Mayr%26Windhager_PolyContext.pdf.

McIlvain, Eileen (2013): “Paradata. What is Paradata?” In: *NSDL Documentation Wiki*. https://wiki.ucar.edu/display/nsdldocs/Paradata.

Moretti, Franco (2007): “Graphs, Maps, Trees. Abstract Models for Literary History.” London, New York.

Moretti, Franco (2016): “Distant Reading.” Göttingen.

Schneider, Stefanie (2019): „Über die Ungleichheit im Gleichen. Erkennung unterschiedlicher Reproduktionen desselben Objekts in kunsthistorischen Bildbeständen.“ In: *DHd 2019. Digital Humanities im deutschsprachigen Raum 2019. Konferenzabstracts*, S. 92–94. https://doi.org/10.5281/zenodo.2596095.

Schönholz, Christian (2017): „Jede Kopie ein Original!. Aspekte eines kulturellen Größenverhältnisses.“ In: Koch, Gertraud (Hg.): *Digitalisierung. Theorien und Konzepte für die empirische Kulturforschung*. Konstanz/München 2017, S. 157–182.

Star, Susan Leigh / Strauss, Anselm L. (1999): “Layers of Silence, Arenas of Voice. The Ecology of Visible and Invisible Work.” In: *Computer Supported Cooperative Work* 8/1-2, S. 9–30.

Werner, Claus (2019): „Die Sammlung als Graph. Gephi als Tool der Sammlungsevaluation. Deutsches Bergbau-Museum Bochum.“ Vortrag im Rahmen der Tagung *Museums and the*

Internet (Mai-Tagung) 2019. https://mai-tagung.lvr.de/media/mai_tagung/pdf/2019/MAI-2019-Werner.pdf.

„As a Hobby at First“ Künstlerische Produktion als Modellierung

Bernhart, Toni
toni.bernhart@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Plädoyer für eine Fachgeschichte digitaler Geisteswissenschaften

Die Geschichte digitaler Geisteswissenschaften wurde bislang kaum erforscht. Schlaglichtartige Beiträge liegen vor (Hoover 2007, Kelih 2008, Cortelazzo/Tuzzi 2008, Viehhauser 2015, Weitin 2015, Jannidis 2015, Twellmann 2016, Thaller 2017, Schöch 2017, Lauer und Pacyna 2017, Bernhart 2018), doch umfassende Studien, die nicht nur die letzten Jahrzehnte, sondern auch die zahlreichen Vorläufe seit dem späten 18. Jahrhundert in systematischer und historischer Perspektive in den Blick nehmen, gibt es noch nicht.[1]

Der Vortrag möchte für historisches Bewusstsein digitaler Geisteswissenschaften sensibilisieren und für eine Fundierung der Wissenschaftsgeschichte des Faches werben. Denn historisch informierte digitale Geisteswissenschaften sind in der Lage, auf Erfahrungen und Experimente aus mindestens zwei Jahrhunderten zurückzugreifen und diese für die Erkenntnisgewinnung zu nutzen und kritisch zu reflektieren. In ihrer genuinen Koppelung von Informatik, die tendenziell eher gegenwartsbezogen operiert, und geisteswissenschaftlichen Disziplinen, die tendenziell größere Aufmerksamkeit auf die Betrachtung historischer Wissensbestände legen, sind die digitalen Geisteswissenschaften dazu berufen, historische Perspektivierungen bei ihrer Arbeit an Modellierung und Interpretation kultureller Artefakte zu integrieren und zu systematisieren. Historische Perspektivierung macht die gesellschaftliche Relevanz digitaler Geisteswissenschaften transparent und dient der didaktischen Vermittlung des Faches, indem Zeitverlauf und Erkenntnisgewinn in Korrelation miteinander erzählt werden können. Unter den zahlreichen Strängen einer Wissenschaftsgeschichte digitaler Geisteswissenschaften greift der Vortrag den Aspekt künstlerischer Produktion als Modellierung heraus und demonstriert diese am Beispiel der Arbeiten von Theo Lutz und Wilhelm Fucks.

Künstlerische Produktion als Modellierung

Das Thema der Jahrestagung stellt Modellierung und Interpretation als zentrale Arbeitsfelder der Digital Humanities in den Vordergrund. Modellierung wird dabei jedoch vorwiegend theoretisch als Mittel der Erkenntnisgewinnung verstanden. Ein anderer Aspekt der Modellierung ist die künstlerische Produktion, die aber nicht zu den originären Arbeitsgebieten der Digital Humanities zählt. Während im internationalen Feld der Digital Studies die Grenze zwischen den Künsten und den Wissenschaften sehr viel stärker aufgehoben scheint und auch akademische Forschung sich an künstlerischer Produktion beteiligt, verharren die Digital Humanities, insbesondere jene des deutschsprachigen Raums, in eher beobachtendem Status. Sie sehen Analyse und Interpretation als ihre primären Zuständigkeitsbereiche und halten weiterhin die Dichotomie zwischen Medienkunst und Medienwissenschaft aufrecht. Digitale Kunst wird außerhalb der Digital Humanities produziert; für künstlerische Produktion werden Digital Humanities kaum genutzt. Dabei verweist gerade die Metapher der Spielräume auf den transgressiven Charakter experimenteller Laboratorien, die in den Digital Humanities bereitstehen.

Die Frühzeit digitaler Geisteswissenschaften und insbesondere die Kybernetik der späten 1950er und 1960er Jahre waren in dieser Hinsicht sehr viel verspielter und experimenteller. Vertreter akademischer Disziplinen wie etwa der Mathematik und Physik verstanden Modellierung ganz selbstverständlich auch im Sinne künstlerischer Produktion. Beispiele dafür sind der Mathematiker Theo Lutz und der Physiker Wilhelm Fucks. Wissenschaftsgeschichtlich bezeichnend ist ferner, dass sich in dieser Zeit aus Mathematik und Elektrotechnik ein neues Fach zu emanzipieren beginnt, das in den 1970er Jahren unter dem Namen Informatik sehr rasch an internationaler Bedeutung gewinnt. Dabei war in dieser frühen Zeit noch nicht ausverhandelt, für welche technischen, angewandten, theoretischen und geisteswissenschaftlichen Problemlösungen die Informationsverarbeitung zuständig sein soll; vielmehr waren Explorationen in sehr unterschiedliche Richtungen charakteristisch (Gunzenhäuser 1968).

„Stochastische Texte“ von Theo Lutz

Theo Lutz (1932–2010), Mathematikstudent an der vormaligen Technischen Hochschule, heute Universität Stuttgart, schrieb im Frühjahr 1959 an der hochschuleigenen Zuse Z 22 seine Diplomarbeit über elektrotechnische Netzwerke. In seiner Freizeit entwickelte er die Idee zu einem Umkehrschub: Wenn es mithilfe computergestützter und statistischer Verfahren möglich ist, Texte zu analysieren und zu interpretieren, muss es auch möglich sein, mithilfe derselben Verfahren Texte zu produzieren. Sein Lehrer Max Bense, Philosoph der rationalen Avantgarde, und sein Studienfreund Rul Gunzenhäuser, der später gemeinsam mit Helmut Kreuzer das Grundlagenwerk „Mathematik und Dichtung“ herausgeben (Kreuzer/Gunzenhäuser 1965) und die „Zeitschrift für Literaturwissenschaft und Linguistik (LiLi)“ begründen und in Stuttgart zu einem Pionier der Informatik avancieren wird, waren begeistert von Lutz’ Idee und unterstützten das Vorhaben. Das Ergebnis wa-

ren die „Stochastischen Texte“, die Wortmaterial aus Franz Kafkas Roman „Das Schloss“ (1926) wahrscheinlichkeitsmathematisch zu grammatikalisch sinnvollen Sätzen kombinierten (Lutz 1959).[2] Programmiert wurde die Z 22 im sogenannten Freiburger Code. Eine besondere Herausforderung stellte dabei die maschinelle Generierung der für die Textherstellung erforderlichen Zufallszahlen dar. Mit seinen „Stochastischen Texten“ schrieb Lutz Literaturgeschichte: Sie waren nach den „Love Letters“ von Christopher Strachey die ersten mithilfe einer programmierten Rechenmaschine generierten Texte in deutscher Sprache (Strachey 1954). Doch der ursprüngliche Zweck der „Stochastischen Texte“ war ein anderer: Sie sollten als Vergleichstexte zur Untersuchung natürlichsprachlicher Texte dienen (Bernhart 2019: 329–331, Bernhart/Richter 2019, Reiter/Bernhart eingereicht). Die Rekonstruktion der Genese der „Stochastischen Texte“ wird im Vortrag flankiert von der Berücksichtigung der poetologischen, philosophischen, politischen und ästhetischen Voraussetzungen, die für maschinelle und programmgesteuerte Generierung von Kunst in den späten 1950er und 1960er Jahren stil- und programmbildend waren.[3]

Wilhelm Fucks und Neue Musik

Wilhelm Fucks (1902–1990) war Physiker an der RWTH Aachen. In den Geisteswissenschaften ist Fucks vor allem für seine sehr zahlreichen kybernetischen Forschungen zu Literatur, Musik und bildender Kunst und für die beiden Monographien „Formeln zur Macht“ (Fucks 1965) und „Nach allen Regeln der Kunst“ (Fucks 1968) bekannt. Bislang kaum beachtet wurden dagegen seine kompositorischen Versuche.

Erstaunlich ist dabei der Echoraum, den sich Fucks für seine Kompositionen verschaffen konnte. Dank seines kommunikativen Talents und seiner wissenschaftlichen Adaptionsfähigkeit war er in der Lage, sich innerhalb weniger Jahre als zeitgenössischer Komponist zu etablieren und sich Ende der 1960er Jahre neben internationalen Vertretern der Neuen Musik wie Iannis Xenakis und John Cage zu behaupten, obwohl er auf dem Gebiet des musikalischen Schaffens Amateur war. Seine ersten Versuche reichen in die Zeit der letzten Monate des Zweiten Weltkriegs zurück, wie er im Diskussionsprotokoll des Bandes zum Symposium „Information Theory“ an der Royal Institution 1955 in London festhält. Zunächst habe er sich hobbymaßig, „as a hobby at first“ (Fucks 1956: 169), mathematisch mit Fragen literarischer Stilistik beschäftigt. Erst später habe er sich mit den Theorien und Ansätzen etwa von Benoit Mandelbrot, Gustav Herdan, Claude E. Shannon oder Norbert Wiener vertraut gemacht und seine Studien auf den Bereich der Musik ausgedehnt (ebd.).

Unter dem Eindruck der probabilistischen Logik von John von Neumann (Neumann 1956) intensivierte Fucks seine Beschäftigung mit stochastischer Musik, suchte in Paris den Austausch mit Abraham Moles und den Kontakt zu Iannis Xenakis, die ihn in den Kreis um den einflussreichen Experimentator und Theoretiker der Neuen Musik Hermann Scherchen einführten. Auf Scherchens legendärer Tagung in Gravesano im Schweizer Tessin stellte Fucks schließlich 1962 seine umfangreichen harmonischen Entropieforschungen sowie eigene Kompositionen vor, die – ähnlich wie Lutz’ Generierung der „Stochastischen Texte“ – aus der „umgekehrten“ Anwendung empirischer Verteilungen hervorgingen (Fucks 1962). Für den Vortrag seiner Stücke konnte Fucks die namhafte Pianistin Margot Pinter (1915–1982) gewinnen, Professorin für Klavier am Innsbrucker Konservatorium und Spezialistin für Neue Musik. Eine bislang unbekannte Tonbandaufzeichnung davon konnte ich kürzlich im Archiv der Akademie der Künste, Berlin, ausfindig machen (Akademie der Künste, Archiv, Signatur AVM-31 6332, Band 14 und Band 15). Die aufgezeichnete Musik und vor allem die ebenfalls aufgezeichnete Plenumsdiskussion im Anschluss an den Vortrag sind aufschlussreiche, bislang unbekannte Quellen, die im Vortrag vorgestellt werden.[4]

Fucks’ weitere Stationen führten nach Berlin und London. Auf Einladung des Architekten und Präsidenten der Berliner Akademie der Künste Hans Scharoun stellte Fucks 1965 auf einer prominent besetzten Tagung zum Thema „Kybernetik“ in Berlin seine Musiken vor. In London war Fucks als Komponist auf der von Jasia Reichardt kuratierten Ausstellung „Cybernetic Serendipity“ 1968 vertreten (Reichardt 1968), die als eine der ersten internationalen Ausstellungen kybernetischer Künste gilt. Auf der gleichnamigen Langspielplatte, die im Rahmen der Ausstellung erschien, ist Fucks’ Stück „Quatro due [sic]“, gespielt von Margot Pinter, zu hören. Auf der Platte vertreten sind unter anderem John Cage, Iannis Xenakis und James K. Randall (Cybernetic Serendipity Music 1968).

Spielräume künstlerischer Produktion in Kybernetik und digitalen Geisteswissenschaften

Im Fazit des Vortrags wird danach gefragt, inwiefern sich Kybernetik und Digital Humanities hinsichtlich künstlerischer Produktion unterscheiden. Produktiv dafür kann die schwierige Vergleichbarkeit der beiden Bewegungen sein: Die Kybernetik operierte vorwiegend quantitativ und statistisch und delegierte die Interpretation an eine imaginäre Zukunft; die Digital Humanities dagegen integrierten quantitative und qualitative Verfahren von Anfang an und gleichermaßen in die Modellierung und Interpretation der Untersuchungsgegenstände. Die Dichotomie zwischen Medienkunst und Medienwissenschaft bleibt dabei tendenziell aufrecht, während die Kybernetik ihre Spielräume sehr viel transgressiver auch für künstlerische Produktionen nutzte. Spielerischer scheinen derzeit digitale Medienkünste zu agieren, die in und neben ihrer künstlerischen Produktion oft auch forschend tätig sind. Auch die zahlreichen KI-Labore (etwa von Google oder OpenAI) pflegen neben ihrer angewandten Forschung mitunter spielerischen Umgang mit Künstlicher Intelligenz, der bisweilen an die Experimente von Lutz und Fucks erinnert. Der Geist der Kybernetik entsprang der jungen und technikbegeisterten Aufbruchstimmung der Nachkriegszeit, während digitale Geisteswissenschaften und Künste mittlerweile auf kollaborative Projekterfahrungen, Tools und Formate eines halben Jahrhunderts digitaler Kompetenz zurückblicken und auch kritischere und differenziertere Positionen vertreten als die historische Kybernetik. Hinsichtlich der Einmischung in künstlerische Produktion liegen in den Digital Humanities Spielräume verborgen, über deren (Nicht-)Nutzung nachzudenken lohnen kann.

Fußnoten

1. Der Beitrag entstand im Rahmen des Forschungsprojekts „Quantitative Literaturwissenschaft“, gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 259167649.
2. Für die Genehmigung der Verwendung von bislang unveröffentlichtem Text- und Bildmaterial von Theo Lutz danke ich Hannelore und Heike Lutz sowie dem Deutschen Literaturarchiv (DLA) Marbach.
3. Anhand des Nachlasses von Theo Lutz, der seit 2019 im Deutschen Literaturarchiv (DLA) Marbach liegt, lässt sich die komplexe und voraussetzungsreiche Genese der „Stochastischen Texte“ sehr genau und detailreich rekonstruieren. Vertiefende Ausführungen dazu bieten der Beitrag von Reiter/Bernhart (eingereicht) und zwei Vorträge von Toni Bernhart: „Beiwerk als Werk. ‚Stochastische Texte‘ von Theo Lutz“ bei der 18. internationalen Tagung der Arbeitsgemeinschaft für germanistische Edition „Werk und Beiwerk. Zur Edition von Paratexten“ vom 12. bis 15. Februar 2020 im Deutschen Literaturarchiv (DLA) Marbach und „Theo Lutz auf Zuse Z 22: ‚Stochastische Texte‘ (1959). Präliminarien einer Edition“ beim XIV. Kongress der Internationalen Vereinigung für Germanistik „Wege der Germanistik in transkulturellen Perspektiven“ vom 26. Juli bis 2. August 2020 in Palermo.
4. Für die Genehmigung der Verwendung von bislang unveröffentlichtem Text-, Bild- und Tonmaterial von Wilhelm Fucks danke ich Thomas Fucks, Anton Voigt und dem Archiv der Akademie der Künste, Berlin.

Bibliographie

Bernhart, Toni (2018): „Quantitative Literaturwissenschaft: Ein Fach mit langer Tradition?“ in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (Hg.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven.* Berlin, Boston: Walter de Gruyter 207–219 https://doi.org/10.1515/9783110523300-009 [6.1.2020].

Bernhart, Toni (2019): „Rul Gunzenhäuser und die Stuttgarter Schule der mathematischen Literaturwissenschaften“, in: Albrecht, Andrea / Bonitz, Masetto / Skowronski, Alexandra (Hg.): *Max Bense. Werk – Kontext – Wirkung.* Stuttgart: Metzler 323–335.

Bernhart, Toni / Richter, Sandra (2019): „Maschinen können Gedichte schreiben“, in: Süddeutsche Zeitung, Nr. 244, vom 22. Oktober 2019: 12.

Cortelazzo, Manlio / Tuzzi, Arjuna (2008): *Metodi statistici applicati all'italiano.* Bologna: Zanichelli.

Cybernetic Serendipity Music (1968). ICA [LP]. https://cyberneticserendipity.net/ [6.1.2020].

Fucks, Wilhelm (1956): „Mathematical Theory of Word Formation“, in: Cherry, Colin (ed.): *Information Theory. Papers read at a Symposium on 'Information Theory' held at the Royal Institution, London, September 12th to 16th 1955.* London: Butterworth 154–170.

Fucks, Wilhelm (1962): „Mathematische Musikanalyse und Randomfolgen. Musik und Zufall. Musical Analysis by Mathematics. Random Sequences“, in: *Gravesaner Blätter* 6/23–24: 132–155. https://archiv.adk.de/objekt/2971402 [6.1.2020].

Fucks, Wilhelm (1965): *Formeln zur Macht. Prognosen über Völker, Wirtschaft, Potentiale.* Stuttgart: Deutsche Verlagsanstalt.

Fucks, Wilhelm (1968): *Nach allen Regeln der Kunst. Diagnosen über Literatur, Musik, bildende Kunst – die Werke, ihre Autoren und Schöpfer.* Stuttgart: Deutsche Verlagsanstalt.

Gunzenhäuser, Rul (Hg.) (1968): *Nicht-numerische Informationsverarbeitung. Beiträge zur Behandlung nicht-numerischer Probleme mit Hilfe von Digitalrechenanlagen*. Wien / New York: Springer.

Hoover, David L. (2007): „Quantitative Analysis and Literary Studies“, in: Siemens, Ray / Schreibman, Susan (eds.): *A Companion to Digital Literary Studies* (= Blackwell companions to literature and culture 50). Malden, MA: Blackwell: 517–533 http://www.digitalhumanities.org/companion/DLS/ [6.1.2020].

Jannidis, Fotis (2015): „Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft. Ein Essay“, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte (DVJs)* 89/4: 657-661.

Kelih, Emmerich (2008): *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft* (= Studien zur Slawistik 19). Hamburg: Kovač.

Kreuzer, Helmut / Gunzenhäuser, Rul (Hg.) (1965): *Mathematik und Dichtung. Versuche zur Frage einer exakten Literaturwissenschaft* (= Sammlung Dialog 3) . München: Nymphenburger Verlagshandlung.

Lauer, Claudia / Pacyna, Jana (2017): „Zählen und Erzählen. Mittelalterliche Literatur- und Geschichtswissenschaft im methodischen Dialog“, in: Schweiker, Marcel / Hass, Joachim / Novokhatko, Anna / Halbleib, Roxana (Hg.): *Messen und Verstehen in der Wissenschaft. Interdisziplinäre Ansätze.* Wiesbaden: Springer 23–41.

Lutz, Theo (1959): „Stochastische Texte“, in: *augenblick. zeitschrift für tendenz und experiment* 4/1: 3–9.

Neumann, John von (1956): „Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components“, in: Shannon, C. E. / McCarthy, J. (eds.): *Automata Studies* (= Annals of Mathematics Studies 34). Princeton, NJ: Princeton University Press 43–98.

Reichardt, Jasia (Hg.) (1968): *Cybernetic Serendipity. The Computer and the Arts. A Studio International special issue,* 2nd ed. London: Studio International. https://cyberneticserendipity.net/ [6.1.2020].

Reiter, Nils / Bernhart, Toni (eingereicht): „Theo Lutz: Poetry Generation in 1959 on Zuse Z 22“, in: *Digital Humanities Quarterly. Special Issue on Minimal Computing.*

Schöch, Christof (2017): „Quantitative Analyse“, in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (Hg.): *Digital Humanities. Eine Einführung. Mit Abbildungen und Grafiken.* Stuttgart: Metzler 279–298.

Strachey, Christopher (1954): „The ‘Thinking’ Machine“, in: *Encounter. A monthly review of literature, the arts, and politics* 13: 25–31. http://www.unz.com/print/Encounter-1954oct-00025/ [6.1.2019].

Thaller, Manfred (2017): „Geschichte der Digital Humanities; Digital Humanities als Wissen-schaft“, in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (Hg.): *Digital Humanities. Eine Einführung. Mit Abbildungen und Grafiken.* Stuttgart: Metzler 3–18.

Twellmann, Marcus (2016): „Gedankenstatistik. Proto-digitale Wissenschaften vom ‚objektiven Geist‘ und ihre Archivverfahren“, in: Gretz, Daniela / Pethes, Nicolas (Hg.): *Archiv / Fiktionen. Verfahren des Archivierens in Literatur und Kultur*

des langen 19. Jahrhunderts. Freiburg i. Br. / Berlin / Wien: Rombach 409–431.

Viehhauser, Gabriel (2015): „Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte“, in: Baum, Constanze / Stäcker, Thomas (Hg.): *Grenzen und Möglichkeiten der Digital Humanities* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften 1) http://dx.doi.org/10.17175/sb01 [6.1.2020].

Weitin, Thomas (2015): „Digitale Literaturwissenschaft“, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte (DVJs)* 89/4: 651–656.

Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane

Lüschow, Andreas
andreas.lueschow@gmx.de
Universität Trier, Deutschland

Einleitung

Unter dem Begriff des *Semantic Web* (Berners-Lee, Hendler, Lassila 2001) werden Techniken, Standards und Methoden zusammengefasst, mit deren Hilfe im Internet verfügbare Daten der semantischen Verarbeitung durch Maschinen zugänglich gemacht werden können. Durch die Einführung und Nutzung von offenen Standards wie z. B. RDF (Schreiber & Raimond 2014) soll hierbei die Interoperabilität unterschiedlicher Datenquellen sichergestellt werden. Diese Standards beziehen sich auf die Art, wie Informationen repräsentiert werden und wie Verknüpfungen mit anderen Informationen hergestellt werden können. Daher wird oftmals auch der Begriff der *Linked Data* verwendet (Bizer, Heath, Berners-Lee 2009). In einer Visualisierung der Linked-Data-Cloud von 2017 (Freyberg 2017: 29) sind die Geisteswissenschaften als eigener Bereich nicht explizit aufgeführt, was die geringe Veröffentlichung geisteswissenschaftlicher semantischer Daten widerspiegelt bzw. vermuten lässt, wenngleich z. B. im Bereich der Graphentechnologien durchaus einige Projekte existieren (Kuczera 2017).

Metadaten als Basis literaturwissenschaftlicher Forschung

Dabei sind solche Daten Basis vieler (literatur-)wissenschaftlicher Fragestellungen: Soll bspw. eine quantitative Textanalyse einer großen Anzahl von Romanen durchgeführt werden, müssen zunächst einmal die in Frage kommenden Werke ermittelt und ausgewählt werden. Die Erstellung solcher möglichst repräsentativen Samples ist allerdings ohne eine Kenntnis der gesamten Romanproduktion einer Epoche, der dort behandelten Themen und Motive und weiterer Angaben über die inhaltliche Ausgestaltung der zu betrachtenden Textproduktion nicht ohne Weiteres möglich.

Hierbei helfen können Nachschlagewerke wie z. B. Fachbibliographien, in denen bibliographische Metadaten verzeichnet sind. Teilweise liegen solche Metadaten bereits als Linked Data vor, da Bibliothekskataloge (retro-)digitalisiert wurden. Diese Metadaten sind als Basis literaturhistorischer Arbeit jedoch häufig nicht ausreichend, da für eine zielgerichtete Auswahl relevanter Literatur oftmals mehr als die üblicherweise erschlossenen bibliographischen Angaben notwendig sind.

Einen weiteren, großen Anteil an der prinzipiell verfügbaren Literatur haben jedoch auch Werke, die nicht digitalisiert, sondern nur in gedruckter Form vorliegen. Die *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne, Frautschi 1977) fasst alle von den Autoren auffindbaren französischsprachigen Romane aus der zweiten Hälfte des 18. Jahrhunderts zusammen. Neben bibliographischen Daten zu Autoren, Werktiteln, Verlegern u. a. sind, soweit möglich, auch Angaben zu weiteren Auflagen (Reeditionen) und zum Inhalt der Werke zusammengetragen worden. Die Bibliographie enthält somit inhaltliche Informationen zu den einzelnen Romanen, die weit über eine Auflistung bibliographischer Metadaten hinausgehen. Solche Informationen sind wie o. g. notwendige Voraussetzung für die Erstellung repräsentativer Samples, u. a. zur weiteren literaturhistorischen Untersuchung der Textproduktion einer Sprache bzw. Epoche.

Zielsetzung

Im Rahmen des hier präsentierten Vorhabens – einer Masterarbeit im Studiengang Digital Humanities an der Universität Trier – wurde die o. g. Bibliographie eingescannt und mittels *Optical Character Recognition* (OCR) in maschinenlesbaren Text umgewandelt. Auf dieser Grundlage wurden mithilfe eines Verfahrens des überwachten maschinellen Lernens die einzelnen Einträge extrahiert, in ein selbst entwickeltes semantisches Modell überführt und mit externen Daten verknüpft, sodass die Bibliographie nunmehr als RDF-Datensatz vorliegt und weiterverwendet werden kann.[1] Zielsetzung der Arbeit war es, die in der Bibliographie enthaltenen Informationen unter Nutzung bibliographischer Standards und aktueller, verbreiteter Datenmodelle auf eine Art und Weise zu repräsentieren, die zukünftig weitere Verarbeitungen und Anreicherungen ermöglicht. Die so entstandene digitale Bibliographie kann darüber hinaus als Basis für buchwissenschaftliche, literaturhistorische und verwandte Forschungen dienen, da in ihr sowohl formale als auch inhaltliche Metadaten zur französischsprachigen Romanproduktion eines definierten Zeitraums enthalten sind.

Metadatenextraktion

Ablauf

Der Ablauf der Metadatenextraktion ist in Abbildung 1 dargestellt.

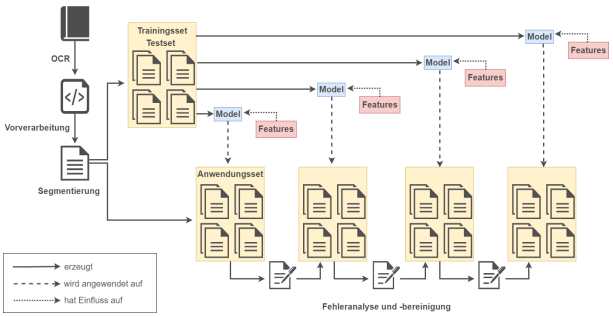


Abbildung 1: Ablauf der Metadatenextraktion

Nach dem Einscannen der gedruckten Vorlage, der OCR, der Vorverarbeitung (Korrektur von Fehlern, Entfernen von Vorwort und Abbildungen, einheitliche Zeichenkodierung etc.) wurden die einzelnen Jahreslisten der Bibliographie und innerhalb dieser die einzelnen Einträge/Romane durch XML-Markup voneinander getrennt (Segmentierung).

Anschließend wurde ein Trainingsset erstellt, mit welchem der verwendete Algorithmus trainiert werden konnte. Für die Trainingsdaten wurde aus jedem Jahrzehnt ein Jahr ausgewählt und die Metadaten der in diesem Jahr erschienenen Romane wurden manuell mit XML-Markup ausgezeichnet. Zur Evaluation der Modelle wurde ein Teil der Daten als Testset zurückgehalten.

Das maschinelle Lernen verlief iterativ, sodass jeweils Modelle für unterschiedlich „tiefe“ Metadatenebenen gelernt wurden, da eine mehrstufige Anwendung mehrerer Modelle oftmals bessere Ergebnisse als die Verwendung eines einzigen Modells für die gesamten Daten erzielt (Kovacevic et al. 2011: 388) und simpler strukturierte Modelle weniger Trainingsdaten benötigen (Candeias 2011: 28). Ein erstes Modell wurde bspw. zur Bestimmung der Makrostruktur der Metadaten verwendet (Titel, Autor, Publikationsdetails etc.), weitere Modelle verfeinerten jeweils die Auszeichnung innerhalb einer dieser Gruppen (z. B. Differenzierung der Publikationsdetails: Ort, Verleger, Jahr, Format, Seitenangabe). Insgesamt wurden sechs Modelle trainiert, die durch stichprobenartige Analyse der erzeugten Daten sukzessive angepasst wurden, bis keine Verbesserungen mehr möglich waren. Das jeweils beste Modell einer Iteration wurde dann auf die restlichen, noch nicht im Trainings- bzw. Testset enthaltenen Jahreslisten angewendet.

Algorithmus und Features

Zur Modellbildung wurden *Conditional Random Fields* (CRF), ein Verfahren des überwachten maschinellen Lernens, verwendet (Lafferty, McCallum, Pereira 2001), das sich in den letzten Jahren zu einem wesentlichen Verfahren im Rahmen der Informationsextraktion entwickelt hat (vgl. z. B. Groza, Grimnes, Handschuh 2012). CRF kombinieren die Vorteile von *Hidden-Markov-Modellen* (HMM) und *Support Vector Machines* (SVM), zwei weiteren gut untersuchten Verfahren (Peng, McCallum 2004: 329).

Die in den Algorithmus eingespeisten Daten (hier: Wörter bzw. Token) werden als Sequenzen von Zuständen modelliert und auf Grundlage dieser beobachteten Zustände werden Label für die einzelnen Elemente vergeben. Im Gegensatz zu HMM berücksichtigen CRF jedoch mögliche Beziehungen der Elemente untereinander – im vorliegenden Fall also der Metadatenfelder bzw. der berücksichtigten Features. Da die Einträge der Bibliographie einem definierten Schema folgen (z. B. steht immer zuerst die Autorenangabe, dann folgt der Titel), ist dieser Algorithmus zur Modellierung der vorliegenden Daten besonders geeignet.

Tabelle 1: In den Modellen berücksichtigte Features

Feature	Erklärung
word	Einzelnes Wort wie es im Text vorkommt
word.lower	Wort in Kleinbuchstaben
word[-3:], word[-2:], word[-1:]	Die letzten Zeichen des Wortes
word[-1:].isalpha	Endet das Wort mit einem Buchstaben?
word[:3], word[:2], word[:1]	Die ersten Zeichen des Wortes
word[:1].isalpha	Ist das erste Zeichen ein Buchstabe?
word.isupper	Besteht das Wort nur aus Großbuchstaben?
word.istitle	Beginnt das Wort mit einem Großbuchstaben?
word.isdigit	Besteht das Wort nur aus Zahlen?
word.isalpha	Besteht das Wort nur aus Buchstaben?

Damit ein CRF-Modell trainiert werden kann, müssen Features erhoben werden, die den Inhalt der einzelnen Metadatenfelder repräsentieren. Tabelle 1 gibt die genutzten Features wieder. Diese Features wurden nicht nur für das jeweilige Wort, sondern auch für das vorherige und das nachfolgende Wort erhoben. So kann im Modell bspw. gelernt werden, dass auf ein bestimmtes Wort stets eine Zahl folgt.

Die genutzten Features wurden ausgehend von einer manuellen Analyse der Einträge in der Bibliographie und basierend auf den ausführlichen Erläuterungen der Autoren zur Sammlung und Strukturierung der Daten im Vorwort der Bibliographie ausgewählt. In der gedruckten Vorlage wurde Großschreibung bspw. zur Hervorhebung von Familiennamen verwendet und Angaben zum Inhalt eines Romans folgten fest definierten einleitenden Begriffen.

Eine ausführliche Evaluation unterschiedlicher Feature-Kombinationen fand im Rahmen der Arbeit nicht statt, da bereits die o. g. simplen Features zu ausreichend hoher Genauigkeit der Metadatenextraktion führten. Weitere Optimierungen hätten überdies vom eigentlichen Ziel der Arbeit weggeführt. Die zur Unterscheidung der einzelnen Metadatenfelder günstigsten Features wurden jedoch erhoben, um die Wirksamkeit und innere Struktur der gelernten Modelle zu überprüfen. Hierbei zeigte sich z. B., dass die einleitenden Wendungen zur inhaltlichen Beschreibung der Romane auch vom Algorithmus als solche gelernt und zur Auszeichung neuer Daten verwendet wurden.

Um auch weniger strukturierte Datengrundlagen als Bibliographien mit dem entwickelten Workflow verarbeiten zu können, bestünde hier ein möglicher, näher zu untersuchender Ansatzpunkt für eine genauere Analyse hilfreicher Features und die eventuelle Einführung weiterer Features.

Evaluation

Das maschinelle Lernen wurde mithilfe der Programmiersprache *Python* und der dort verfügbaren Bibliothek *sklearn-crfsuite*[2] implementiert. Die Evaluation der Modelle geschah mit der zu *sklearn-crfsuite* kompatiblen Bibliothek für wissenschaftliche Programmierung *scikit-learn*[3]. In der folgenden Tabelle sind die gängigen Maße Precision, Recall und der F1-Score für die sechs gelernten Modelle angegeben.

Tabelle 2: Evaluation der einzelnen Modelle

Modell	Precision	Recall	F1
entry (Makrostruktur des Eintrags)	0,954	0,953	0,951
det (Publikationsdetails)	0,987	0,986	0,986
res (Schlagwörter)	0,919	0,907	0,908
au (Autorennamen)	0,975	0,986	0,980
ae (Makrostruktur weiterer Editionen)	0,997	0,997	0,997
ae_se (einzelne Einträge weiterer Editionen)	0,961	0,960	0,960

Für alle Metadatenfelder konnte eine sehr hohe Genauigkeit erreicht werden. Der so erzeugte Datensatz mit allen Einträgen aus der Bibliographie ist somit nahezu vollständig korrekt ausgezeichnet.

Semantische Modellierung

Zurzeit existiert kein einheitlicher, akzeptierter Standard, der in der Bibliothekswelt für die semantische Repräsentation bibliographischer Daten verwendet wird. Stattdessen orientieren sich diejenigen Bibliotheken, die bereits Linked Data zur Verfügung stellen, an unterschiedlichen Datenmodellen, Schemas und Ontologien. Es existieren jedoch Versuche, die bereits entwickelten Modelle in ein möglichst generisches und von vielen Bibliotheken nachnutzbares Modell zu integrieren (Suominen, Hyvönen 2017).

Vorhandene Ontologien

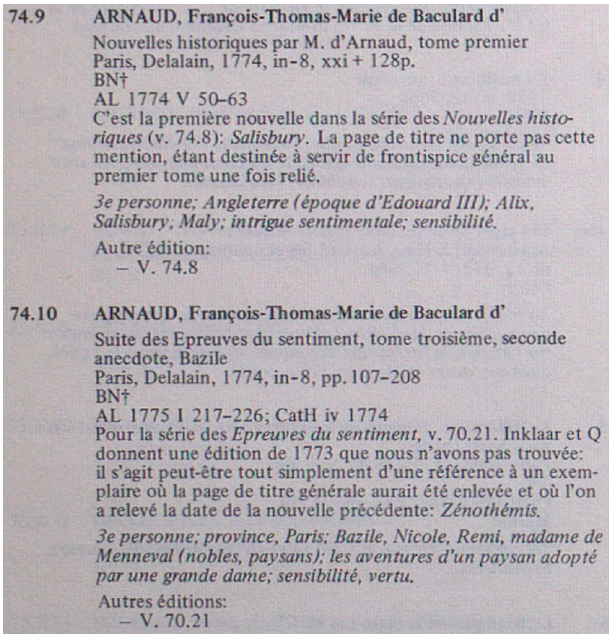
Vor allem die folgenden Datenmodelle sind für die semantische Modellierung der Metadaten aus der Bibliographie relevant, da sie entweder bereits weit verbreitet sind oder spezifische Elemente enthalten, die nachgenutzt werden können.

- *FRBR: Functional Requirements for Bibliographic Records* und *RDA: Resource Description and Access* (IFLA 2009)
- *DCTerms: Dublin Core Metadata Terms* (Dublin Core Metadata Initiative 2012)
- *PRISM: Publishing Requirements for Industry Standard Metadata* (IDEAlliance 2008)
- *SPAR Ontologies* (Peroni, Shotton 2018)

Die Entwicklung der SPAR-Ontologien wird von den Autoren u. a. damit begründet, dass bisherige Systeme uneinheitlich seien und deutliche Schwächen aufwiesen. PRISM und FRBR seien bspw. „top-level vocabularies rather than something specifically developed to characterise specific aspects of scholarly publishing“ (Peroni, Shotton 2018). Gleichzeitig benutzen die SPAR-Ontologien jedoch Elemente aus den anderen o. g. Vokabularen, um Redundanzen und doppelte Element-Definitionen zu vermeiden. In der hier beschriebenen Arbeit wurde daher ebenfalls versucht, aus den o. g. Datenmodellen vorrangig diejenigen Elemente zu verwenden, die bereits im Bibliothekswesen etabliert und nicht zu spezifisch, gleichzeitig aber ausreichend detailliert sind.

Modellentwicklung

Nach einer eingehenden Analyse der in der Bibliographie vorhandenen Metadaten wurden aus den o. g. Ontologien diejenigen Elemente zur weiteren Berücksichtigung ausgewählt, die zur möglichst genauen und eindeutigen Modellierung der einzelnen Einträge der Bibliographie (siehe Abbildung 2) benötigt werden. Hierbei wurde darauf geachtet, nicht bloß die einzelnen Romane mit ihren Metadaten abzubilden, sondern auch den Aufbau und die Struktur der Bibliographie an sich. Dadurch konnte das gesamte zu erzeugende Modell an den bereits im Linked-Data-Service der *Bibliothèque nationale de France* (BnF) vorhandenen Eintrag für die *Bibliographie du genre romanesque français* angebunden werden.[4]



74.9 **ARNAUD, François-Thomas-Marie de Baculard d'**
Nouvelles historiques par M. d'Arnaud, tome premier
Paris, Delalain, 1774, in-8, xxi + 128p.
BN†
AL 1774 V 50–63
C'est la première nouvelle dans la série des *Nouvelles historiques* (v. 74.8): *Salisbury*. La page de titre ne porte pas cette mention, étant destinée à servir de frontispice général au premier tome une fois relié.
3e personne; Angleterre (époque d'Edouard III); Alix, Salisbury; Maly; intrigue sentimentale; sensibilité.
Autre édition:
– V. 74.8

74.10 **ARNAUD, François-Thomas-Marie de Baculard d'**
Suite des Epreuves du sentiment, tome troisième, seconde anecdote, Bazile
Paris, Delalain, 1774, in-8, pp. 107–208
BN†
AL 1775 I 217–226; CatH iv 1774
Pour la série des *Epreuves du sentiment*, v. 70.21. Inklaar et Q donnent une édition de 1773 que nous n'avons pas trouvée: il s'agit peut-être tout simplement d'une référence à un exemplaire où la page de titre générale aurait été enlevée et où l'on a relevé la date de la nouvelle précédente: *Zénothémis.*
3e personne; province, Paris; Bazile, Nicole, Remi, madame de Menneval (nobles, paysans); les aventures d'un paysan adopté par une grande dame; sensibilité, vertu.
Autres éditions:
– V. 70.21

Abbildung 2: Beispieleinträge in der gedruckten Bibliographie

Durch die im Vorfeld bereits erfolgte Extraktion der einzelnen Metadatenfelder aus den OCR-Daten konnten diese schließlich direkt auf die entsprechenden Elemente in dem erstellten RDF-Modell abgebildet werden. Dies geschah überwiegend mithilfe der Programmiersprache *Java* und der dort verfügbaren Bibliothek *Apache Jena*[5].

Verknüpfung mit anderen Ressourcen

Um die Möglichkeit der Anreicherung der Daten mit Informationen aus externen Ressourcen beispielhaft darzustellen, wurden die Namen der Autoren der einzelnen Romane aus dem RDF-Modell extrahiert und mithilfe von Apache Jena an die API der *Virtual International Authority File* (VIAF)[6] gesendet. Von dort wurden – sofern vorhanden – die VIAF-IDs extrahiert und dem RDF-Modell hinzugefügt. Weitere externe

Ressourcen könnten auf ähnliche Weise integriert werden. Voraussetzung für die erfolgreiche Nutzung der API ist, dass die Einträge im RDF-Modell keine Schreibfehler oder OCR-Fehler aufweisen. Dies kommt allerdings relativ häufig vor (Gründe sind u. a.: kleine Schrift in der Vorlage, viele Eigennamen, kurze Wörter mit wenig Kontext) und ist eines der wesentliche Probleme des Datensatzes.

Fazit

Sowohl die Extraktion der einzelnen Metadaten aus den OCR-Texten als auch die Erstellung und anschließende Überführung in ein RDF-Modell ließen sich mit gutem Erfolg umsetzen. Die Erkennungsgenauigkeit des CRF-Algorithmus war mit einem F1-Score von durchschnittlich 0,964 (0,908–0,997) außerordentlich hoch. Grund hierfür war sicherlich vor allem die bereits stark strukturierte Datengrundlage. Fehlende einheitliche Standards zur Repräsentation bibliographischer Metadaten und Fehler in den Textdaten sind jedoch Schwachstellen, die eine genauere Analyse und evtl. umfangreiche Bereinigung/Korrektur der zu repräsentierenden Daten nötig machen.

Das vorgestellte Projekt hat durch die Kombination von modernen Verfahren zur Informationsextraktion und die Zusammenstellung von aktuellen Ontologien zur Repräsentation bibliographischer Metadaten einen für die Datengrundlage passenden Ansatz entwickelt, der als Standard-Workflow für ähnliche Projekte verwendet werden könnte und in solchen überprüft und verfeinert werden sollte. Denkbar wären z. B. die Digitalisierung und Metadatenextraktion weiterer Bibliographien, um den erzeugten Datenbestand zu ergänzen, zu erweitern oder anzureichern. Auch die Überprüfung des hier beschriebenen Vorgehens in verwandten Kontexten (andere Nachschlagewerke, andere Sprachen, andere Epochen) unter Nutzung weiterer oder anderer Features wäre sinnvoll.

Der Workflow und die Daten werden daher am *Trier Center for Digital Humanities* im Rahmen des von der Forschungsinitiative Rheinland-Pfalz geförderten Projektes „MiMoText – Mining and Modeling Text“ weiterverwendet und erweitert. Ziel ist hier der Aufbau eines „aus unterschiedlichen Quellen gespeisten Informationsnetzwerks für die Geisteswissenschaften, das durch die Bereitstellung als Linked Open Data nicht nur frei verfügbar und mit anderen Wissensressourcen des Semantic Web verknüpfbar ist, sondern auch neuartige und effiziente Zugriffsmöglichkeiten auf fachwissenschaftliche Informationen bietet“.[7] Die beschriebene Arbeit liefert hierfür eine geeignete Grundlage.

Fußnoten

1. Der Datensatz ist verfügbar unter https://zenodo.org/record/3401429 (Lizenz: CC-BY).
2. https://github.com/TeamHG-Memex/sklearn-crfsuite
3. https://scikit-learn.org/stable/index.html
4. http://data.bnf.fr/ark:/12148/cb34586696d
5. https://jena.apache.org/
6. https://platform.worldcat.org/api-explorer/apis/VIAF
7. https://kompetenzzentrum.uni-trier.de/de/projekte/projekte/m/

Bibliographie

Berners-Lee, Tim / Hendler, James / Lassila, Ora (2001): *"The Semantic Web",* in: *Scientific American* 284.5: 29–37.

Bizer, Christian / Heath, Tom / Berners-Lee, Tim (2009): "Linked Data – The Story So Far", in: *International Journal on Semantic Web and Information Systems* 5.3: 1–22 http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf [letzter Zugriff 03. Januar 2020].

Candeias, Ricardo Pereira (2011): *Metadata Extraction from Scholarly Articles*. Master Thesis, Universidade Técnica de Lisboa. https://fenix.tecnico.ulisboa.pt/downloadFile/395143160947/dissertacao.pdf [letzter Zugriff 03. Januar 2020].

Dublin Core Metadata Initiative (2012): *DCMI Metadata Terms*. DCMI Recommendation. http://dublincore.org/documents/2012/06/14/dcmi-terms/ [letzter Zugriff 03. Januar 2020],

Freyberg, Linda (2017): "Density of Knowledge Organization Systems", in: *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)* 25–30.

Groza, T. / Grimnes, A. / Handschuh, S. (2012): "Reference Information Extraction and Processing Using Conditional Random Fields", in: *Information Technology and Libraries* 31.2: 6–20.

IDEAlliance – International Digital Enterprise Alliance (2008): *The PRISM Namespace – Final* http://www.prismstandard.org/specifications/2.0/PRISM_prism_namespace_2.0.pdf [letzter Zugriff 03. Januar 2020].

IFLA Study Group on the Functional Requirements for Bibliographic Records (2009): *Functional Requirements for Bibliographic Records – Final Report*. (IFLA Series on Bibliographic Control, Vol. 19) https://www.ifla.org/publications/functional-requirements-for-bibliographic-records [letzter Zugriff 03. Januar 2020].

Kovacevic, Aleksandar / Ivanovic, Dragan / Milosavljevic, Branko / Konjovic, Zora / Surla, Dusan (2011): "Automatic extraction of metadata from scientific publications for CRIS systems“, in: *Program* 45.4: 376–396.

Kuczera, A. (2017): "Graphentechnologien in den Digitalen Geisteswissenschaften", in: *ABI Technik*, 37.3: 179–196.

Lafferty, John D. / McCallum, Andrew / Pereira, Fernando C. N (2001): "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", in: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)* 282–289.

Martin, Angus / Mylne, Vivienne / Frautschi, Richard (1977): *Bibliographie du genre romanesque français 1751-1800*. London, Paris: Mansell, France expansion.

Peng, Fuchun / McCallum, Andrew (2004): "Accurate Information Extraction from Research Papers using Conditional Random Fields", in: *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLTNAACL)* 329–336 https://www.cs.umass.edu/~mccallum/papers/hlt2004.pdf[letzter Zugriff 03. Januar 2020].

Peroni, Silvio / Shotton, David (2018): "The SPAR Ontologies", in: Vrandečić D. et al. (eds.): *The Semantic Web –*

ISWC 2018. Lecture Notes in Computer Science. Cham: Springer 119–136.

Schreiber, Guus / Raimond, Yves (2014): *RDF 1.1 Primer*. W3C Note. http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624 [letzter Zugriff 03. Januar 2020].

Suominen, Osma / Hyvönen, Nina (2017): "From MARC silos to Linked Data silos?", in: *o-bib. Das offene Bibliotheksjournal* 4.2: 1–13.

Best-practices zur Erkennung alter Drucke und Handschriften. Die Nutzung von Transkribus large- und small-scale

Hodel, Tobias
tobias.hodel@uzh.ch
Universität Bern, Schweiz

In den vergangenen Jahren konnte die automatisierte Erkennung sowohl handschriftlicher als auch alter Druckschriften stark verbessert werden. Sowohl für Handschriften als auch für alte Drucke hat sich der Einsatz von Handwritten Text Recognition (HTR) bewährt, die auf dem Einsatz neuronaler Netze beruht.[1] Führend in der Implementierung der Technologie ist die Plattform Transkribus, die im Rahmen von Projekt READ zwischen 2016 und 2019 stetig weiter entwickelt wurde (Muehlberger u. a. 2019) und auf der mittlerweile (Stand Ende 2019) mehr als 1‘000‘000 Dokumentenseiten bearbeitet wurden.[2] Die Verbesserung der Technologie führte gleichzeitig zur Einführung von unterstützenden Tools und Methoden, die reine Texte entweder mit höherer Genauigkeit suchbar machen oder strukturierende Maßnahmen ermöglichen. Es spielt also eine Rolle, welche Resultate erreicht werden sollen und welche Ziele der vorgesehene Zugriff hat. Im Rahmen des Papers werden drei Herangehensweisen skizziert, die ausgehend von unterschiedlichen Zielvorstellungen andere Aufbereitungsschritte und die Allokation von Ressourcen an verschiedenen Stellen nötig machen. Das Paper fokussiert auf die Nutzung von Transkribus, da die Software frei nutzbar und aufgrund des GUI innerhalb von zwei Arbeitstagen ohne Vorkenntnisse erlernbar ist (siehe dazu auch die How-To Guides: READ 2019). Darin eingeschlossen ist die Anfertigung eigener Modelle zur Erkennung von Handschriften bzw. alter Drucke.

Verbesserung der Handschriftenerkennung

Handschriftenerkennung ist seit den 1990er Jahren und dem Aufkommen der OCR (Optical Character Recognition) ein Forschungsfeld der Computerwissenschaften. Nach einer frühen Euphorie folgte bis vor fünf Jahren eine Ernüchterung, da die erzielten Resultate, die häufig auf statistischen Modellen (insbesondere Hidden Markov Models) basierten, für den Einsatz in der Praxis ungenügend waren. Zeichenfehlerquoten von bestenfalls 16% CER (Character Error Rate) zeigten zwar die Chancen auf, der erkannte Text war aber weder lesbar noch sinnvoll durch Postkorrektion aufzubereiten (Sánchez et al. 2013). Erst der Einsatz neuronaler Netze (erst rekursive, später konvolutionale) führte dazu, dass die Fehlerquote auf unter 12% CER gedrückt werden konnte (Leifert et al. 2016). Ab der Schwelle um 12% wird die Korrektur von erkanntem Text gegenüber von händisch erstellten Transkriptionen ökonomisch sinnvoll. Gleichzeitig sind die Resultate ab 12% für Menschen insofern nützlich, da die Navigation im Text, insbesondere für Personen mit Kenntnissen der Dokumente, rasch und zielsicher möglich ist.

Zugriffsformen

Obwohl geisteswissenschaftliche Forschung häufig „Text“ im Fokus hat, ist die Diskussion, was darunter verstanden wird, bereits mehrfach in Bezug zu digitalen Editionsformen in den Digital Humanities breitgetreten worden (Sahle 2013, zu Texterkennung Hodel 2018). Aus der Perspektive von Forschenden, die Text als Grundlage nutzen, werden gleichzeitig unterschiedliche Fragen an das aufbereitete Material gestellt. Ob etwa ein Dokument nur durchsucht oder aber mit *text-mining* Methoden ausgewertet werden soll, führt zu unterschiedlichen Anforderungen an die Texterkennung. Zwischen den beiden Polen besteht auch die Möglichkeit nur visuell abgegrenzte Textteile auszuwerten. Für alle diese Zielvorstellungen unterscheidet sich der Aufwand für die Aufbereitung der Materialien.

Um eine Vergleichbarkeit herzustellen, lohnt sich eine systematisierte Sicht auf den Workflow mit Angaben, wo der Grossteil der Arbeit anfällt. Es werden daher im Folgenden Fragen nach Ablauf und Umfang der Arbeit sowie nach Schwierigkeiten/Probleme beschrieben.

Nicht thematisiert werden unterschiedliche Möglichkeiten bei Upload sowie Exportformate und insgesamt Fragen der Nachbereitung.

Hohe Textgüte als Ziel

Der klassische Zugriff zielt darauf ab, mit möglichst wenig Aufwand eine möglichst gute Texterkennung zu erzielen. Das Training von passgenauen Modellen steht dabei im Vordergrund. Aufbauend auf der Erkennung können *text-mining* Technologien ebenso eingesetzt werden, wie die Weiterverarbeitung als digitale Edition, etwa mit textkritischer Kommentierung oder durch die Annotation von *named entities*.

Ablauf

Primär muss möglichst viel Trainingsmaterial bereitgestellt werden, das bereits früh im Arbeitsprozess in Modelle umgesetzt (= trainiert) wird. Generische Modelle spielen eine untergeordnete Rolle, da diese die Qualität der passgenauen Modelle nicht erreichen. Die Arbeitsweise ist iterativ, das heisst nach Aufbereitung von bereits 3'000 Wörtern lohnt sich die

Herstellung eines Modells. Darauf aufbauend werden weitere Seiten erkannt und korrigiert. Der Prozess wird gestoppt, sobald die Verbesserung des Modells sich im Zehntelprozent Bereich bewegt (siehe dazu auch unten: Evaluation von trainierten Modellen).

Spezifische Typen von Layoutanalysen spielen keine Rolle.

Umfang der Arbeit

Gute Resultate (Zeichenfehler unter 5%) werden bei Dokumenten von derselben Hand mit 10'000 Wörter erreicht. Trainings können selbständig gestartet und überprüft werden.[3]

Schwierigkeiten/Probleme

Sobald unterschiedliche Hände in den Dokumenten erkannt werden sollen, ist eine Erhöhung der Trainingsmaterialien notwendig.

Semantische Textsegmentierung als Ziel

Dokumententypen können aufgrund von schematischem Aufbau nur in Teilen interessant sein, d.h. nur ein visuell abgesetzter Teil soll ausgewertet werden. Einzelne Textteile, bspw. Marginalien mit inhaltlichen Zusammenfassungen oder Fussnoten, werden für spezifische Forschungszwecke identifiziert.

Ablauf

Ein spezifischer Layouttyp wird trainiert. Danach wird unabhängig davon ein Textmodell entwickelt (analog zu „Hohe Textgüte“). Extraktion von Textregionen bzw. Einbindung in externe (webbasierte) Applikationen ist möglich.

Umfang der Arbeit

Mindestens 100, besser 200-300 Vorkommen der zu identifizierenden Teile (bspw. Textregionen). Zusätzlich müssen Aufwände für die Aufbereitung von Modellen veranschlagt werden.

Schwierigkeiten/Probleme

Das Training der semantischen Layouterkennung ist erst experimentell in Transkribus implementiert und schlecht dokumentiert (das Training kann auch extern über ein eigene Tool erfolgen [Ares Oliveira u. a. 2018; Quirós 2017]). Die Technologie ist insgesamt noch experimentell. Die Extraktion spezifischer Textregionen erfordert zudem Erfahrung im Umgang mit der REST API von Transkribus.

Suche in grossen Textbeständen als Ziel

Als regelmässige Anforderung wird die Suche in großen Dokumentenkorpora vorgegeben. Dazu scheint zwar ein probates Mittel die Erkennung mit hoher Textgüte zu sein, da für Handschriften und alte Drucke die Genauigkeit von OCR nicht gegeben ist, drängt sich ein alternativer Zugriff auf. Mittels Suche in allen vom Algorithmus erkannten Varianten, kann auch mit nicht passgenauen Modellen eine hohe Trefferquote (hoher *recall*) erreicht werden.

Ablauf

Primär wird ein möglichst passendes Model identifiziert. Einige generische Modelle (etwa für mittelalterliche Buchschriften oder lateinische Schriften aus den Niederlanden sowie deutsche Kurrent) sind bereits publiziert und in Zukunft werden noch weitere Modelle veröffentlicht. In einem zweiten Schritt werden einige wenige typische Seiten als Validierungsseiten aufbereitet bzw. sog. Samples (zufällig ausgewählte Zeilen, die eine statistisch valide Aussage ermöglichen) erstellt. Aufgrund der eruierten Zeichenfehlerquote in den Validierungssets, ist es möglich Aussagen über die erwartete Trefferquote zu machen. Ab Fehlerquoten von 15% CER und weniger, wird der Recall (Umfang aller gefundenen, möglichen Treffer) 99% betragen und somit werden nur wenige Treffer verpasst.

Umfang der Arbeit

Beschränkt sich vorwiegend auf die Identifikation passender Modelle und der Herstellung von Validierungssets (Validierungsseiten oder Samples bestehend aus einzelnen Zeilen) zur Validierung der Ergebnisse. Die Auswertung der Treffer mit Identifikation von *false-positive* Treffern am Ende des Prozesses bedarf manueller Arbeit.

Schwierigkeiten/Probleme

Die Auswertenden der Trefferliste müssen die Handschrift/den Druck selbst lesen können, um korrekte Treffer zu identifizieren. Die Suche erfolgt in den Erkenntabellen (sog. confidence matrices), da diese nicht durch etablierte Datenbanksysteme etabliert werden, ist die Performanz niedrig und Suchen in mehr als 1'000 Seiten dauern mehrere Minuten (50'000 Seiten werden in knapp 3 Stunden durchsucht). Die Abfragen müssen in Transkribus erfolgen bzw. über die REST API.

Visualisierung

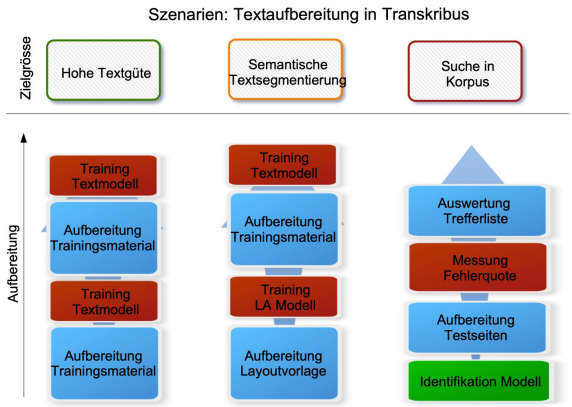


Abbildung 1: Visualisierung der Szenarien zur Textaufbereitung in Transkribus. Abbildung des Autors, CC-BY.

Trainingskurven: Evaluation von trainierten Modellen

Eine Aufgabe, die auch von Geisteswissenschaftlern mit nur bedingten technischen Vorkenntnissen übernommen werden kann, ist das Training von Textmodellen. Dazu muss ein Trainingsset (ca. 90% aller aufbereiteten Seiten) und ein Validierungsset (ca. 10% der Seiten) definiert werden. Das Training erfolgt danach auf den Servern in Innsbruck, es gibt nur zwei Optionen, die angepasst werden können. Erstens kann die Anzahl der Epochen (siehe unten) angepasst werden und zweitens können bereits vorhandene Modelle als Basismodelle gewählt werden.

Im Trainingsmodus erstellt ein Fehlertool Kurven, die anzeigen wie gut das Training ablief. Anhand dieser Trainingskurven lässt sich abschätzen, inwiefern ein Modell noch verbessert werden kann bzw. gar ein Re-training notwendig ist, da sich das Netz nicht wunschgemäss verbesserte.

Drei Begriffe müssen zum Verständnis vorgängig geklärt werden: **Neuronale Netze** stammen aus dem Bereich des maschinellen Lernens und versuchen aufgrund von Trainingsmaterial (Input und gewünschter Output) einen wertenden Algorithmus (ein Netz basierend auf je nach Input unterschiedlich reagierenden Speicherzellen) zu entwickeln (*trainieren*), der den gewünschten Ausgaben möglichst nahe kommt (Schöch 2017). **Epochen** meint die Anzahl an Wiederholungen, mit denen ein Netz mit denselben Trainingsdaten zwecks Verbesserung gefüttert wird. Am Ende jeder Epoche wird das Validierungsset durch das Netz erkannt und eruiert, welche Resultate erreicht worden wären. Dadurch entstehen zwei **Kurven** (in den Abbildungen rot für das Validierungsset und blau für das Trainingsset), die Aussagen über die Fähigkeit eines Netzes machen.

Trainings- und Validierungskurve divergiert stark

Wenn sich die Trainingskurve während dem gesamten Training stetig verbessert, das Validierungsset jedoch auf einer (weit) schlechteren Fehlerquote stehen bleibt, spricht man von *overfitting*. Das bedeutet, das Modell lernt die Trainingsseiten auswendig, ohne dass die Fähigkeit zur Erkennung der Zeichen wirklich eingelernt wird. Probates Mittel um den Effekt zu reduzieren, ist das Bereitstellen von mehr Trainingsmaterial.

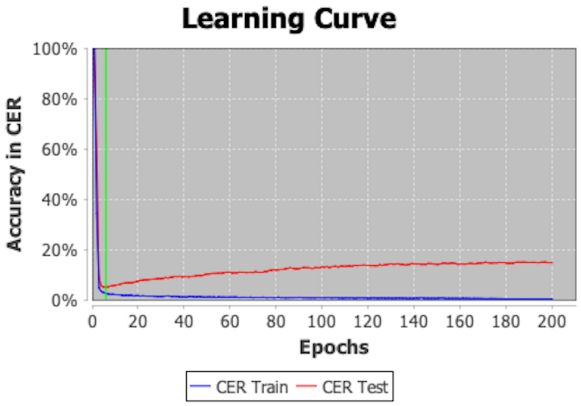


Abbildung 2: Lernkurve mit zuwenig Trainingsmaterial, das zum „overfitting“ führt.

Validierungs- und Trainingskurve verbessern sich bis am Ende des Trainings

Wenn beide Kurven bis zu den letzten Epochen (Trainingszyklen) leichte Verbesserungen feststellen lassen, ist das Netz noch nicht „austrainiert“. Optimal verbessert sich das Netz in den letzten 10-15 Epochen nur noch minimal bzw. gar nicht mehr. Wenn der Effekt der Verbesserung bis zum Ende anhält, sollte das Training nochmals mit mehr Epochen gestartet werden. Erfahrungsgemäss ist die Erkennung von austrainierten Netzen besser und zuverlässiger.

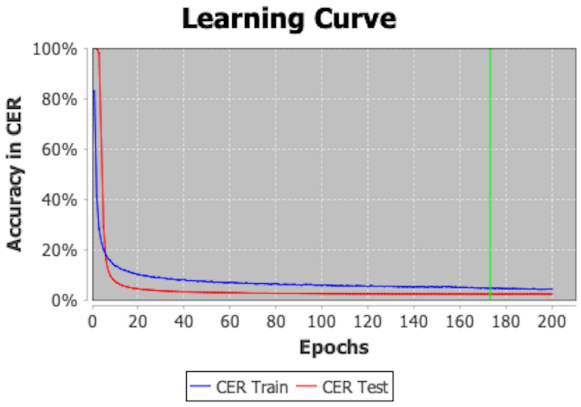


Abbildung 3: Lernkurve eines Netzes, das noch weiter austrainiert werden könnte.

Die Anwendung von Texterkennung, etwa mit Transkribus, ist ohne technische Vorkenntnisse problemlos erlernbar. Mit Rücksicht auf einige wenige Kniffe und mit basalen Kenntnis-

sen der angewandten Algorithmen lassen sich grössere Dokumentenmengen sinnvoll und effizient aufbereiten.

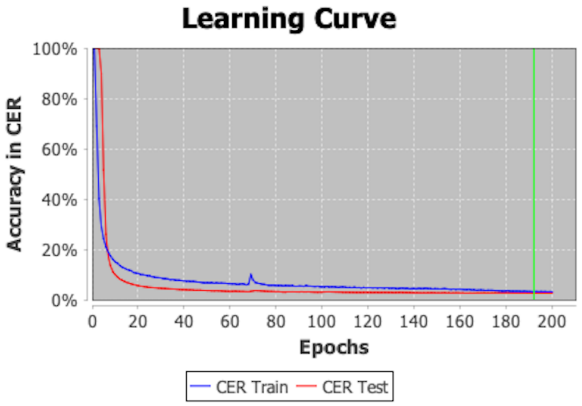


Abbildung 4: Lernkurve eines austrainierten Netzes mit genügend Epochen und ausreichend Trainingsmaterial.

Fußnoten

1. Im Rahmen eines Wettbewerbs an der ICFHR 2018 wurden nur noch Algorithmen basierend *machine learning* zur Erkennung der Handschriften eingesetzt, siehe: https://scriptnet.iit.demokritos.gr/competitions/10/viewresults/.
2. Eine Alternative zu Transkribus ist die Open Source Software Kraken, die ebenfalls auf der Basis neuronaler Netze Trainings individueller Handschriftenmodelle erlaubt. Für die Erkennung alter Drucke (gedruckt vor 1830) eignen sich auch Open Source Tools wie Tesseract. Siehe dazu den aktuellen Stand der Förderinitiative OCR-D, Neudecker u. a. 2019.
3. Aktuell ist der Zugang zur Trainingsfunktionalität nicht standardmässig gegeben. Per Mail (an email@transkribus.eu) wird die Möglichkeit aber rasch und unkompliziert gewährt.

Bibliographie

Ares Oliveira, Sofia / Seguin, Benoît / Kaplan, Frédéric (2018): dhSegment: A generic deep-learning approach for document segmentation. In: *Frontiers in Handwriting Recognition (ICFHR), 2018, 16th International Conference.* 7–12.

Hodel, Tobias (2018): Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning. In: Vogeler, Georg (Hg.). *DHd 2018. Kritik der digitalen Vernunft Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018* Köln, 249–251. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

Leifert, Gundram u. a. (2016): Cells in Multidimensional Recurrent Neural Networks. *Journal of Machine Learning. Res.* 17/97:1–97:37.

Muehlberger, Guenter u. a. (2019): Super Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation 75/5* , 954-976. https://doi.org/10.1108/JD-07-2018-0114

Neudecker, Clemens u. a. (2019): OCR-D: An end-to-end open source OCR framework for historical documents. *EuropeanaTech Insight* 13.

READ: Transkribus . https://read.transkribus.eu/transkribus/ [12.9.2019].

Quirós, Lorenzo (2017): *P2PaLA: page to PAGE layout analysis tookit.* https://github.com/lquirosd/P2PaLA.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. *Schriften des IDE 8* Bd. 2, Norderstedt: BoD. http://kups.ub.uni-koeln.de/5352/ [25.7.2014].

Sánchez, J.A. u. a. (2013): TranScriptorium: a European Project on Handwritten Text Recognition. In: Marinai, Simone / Marriott, Kim (Hg.). *ACM Symposium on Document Engineering DOCENG.* ACM, 227–228.

Schöch, Christoph (2017): Quantitative Analyse, in: Jannidis, Fotis / Kohle, Hubertus, Rehbein, Malte (Hg.). *Digital Humanities: Eine Einführung*. J.B. Metzler, Stuttgart, 279–298. https://doi.org/10.1007/978-3-476-05446-3_20

Bildrepositorien und Forschung mit digitalen Bildern im Bereich der Kunstgeschichte

Kröber, Cindy
cindy.kroeber@tu-dresden.de
TU Dresden, Deutschland

Münster, Sander
sander.muenster@tu-dresden.de
Friedrich-Schiller-Universität Jena, Deutschland

Messemer, Heike
heike.messemer@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Das Verständnis zu Forschungspraktiken und Bedürfnissen von WissenschaftlerInnen und angehenden ForscherInnen im Bereich der Kunstgeschichte sowie ihren Herausforderungen und Ansprüchen beim Zugang und bei der Nutzung digitaler Ressourcen, wie z.B. Bildrepositorien, ist essentiell für den Aufbau geeigneter digitaler Infrastrukturen, die den wissenschaftlichen Arbeitsablauf effektiv erleichtern und den Wert der Bestände steigern. Vor diesem Hintergrund geht es uns darum, zu untersuchen, wie KunsthistorikerInnen mit digitalen Informationen und insbesondere mit Bildbibliotheken umgehen und welche Qualitätskriterien sie dabei heranziehen.

Dieser Beitrag zielt darauf ab, Implikationen für eine Nutzung und nutzerzentrierte Gestaltung von Ressourcen und insbesondere von Bildrepositorien abzuleiten. Um forschungsbezogene Aufgaben adäquat zu unterstützen, haben verschiedene Ansätze versucht, den Forschungsprozess zu

formalisieren, indem sie sogenannte Forschungsprimitive als grundlegende Aufgaben (z.B. Entdecken, Kommentieren, Vergleichen oder Referenzieren) innerhalb der Forschung identifiziert haben (Palmer/Teffeau/Pirmann 2009, Unsworth 2000, Ross 2010) oder das Zwischenspiel von Aufgaben während der Forschungstätigkeit modellieren (z.B. Benardou et al. 2010, Pertsas/Constantopoulos 2017). Da sich solche Beschreibungen leicht in Software-Designs übertragen lassen, besteht ein Widerspruch zu den vielschichtigen Methoden und komplexen Vorgehen, mit denen Wissenschaft tatsächlich praktiziert wird.[1]

Nutzerstudie mit Studierenden der Kunstgeschichte

Die Daten für diese Umfrage stammen aus einer Fokusgruppen-Diskussion vom September 2016 mit 15 Studierenden des Studiengangs Kunst- und Architekturgeschichte der Universität Würzburg und werden ergänzt durch Interviews mit 5 Studierenden der Kunstgeschichte an der TU Dresden vom Juni 2019. Es handelte sich um Bachelor- und Masterstudierende im Alter zwischen 20 und 70 Jahren.

Die Studierenden aus Würzburg sollten für eine Exkursion Informationen über bestimmte Gebäude der Stadt Dresden vorbereiten und präsentieren. Im Rahmen einer Fokusgruppe hat der Moderator Fragen nach dem Rechercheverhalten für die Vorbereitung der Exkursionsaufgabe gestellt und wollte, dass die Studierenden ihren allgemeinen Forschungsprozess beschreiben. Der schriftliche Bericht der Fokusgruppen lieferte die Daten für eine qualitative Auswertung. Eine geeignete Methode ist die von Philipp Mayring eingeführte qualitative Inhaltsanalyse (Mayring 2008). Die Umfrage ist durch mehrere Aspekte eingeschränkt, wie z.B. die geringe Zahl der studentischen TeilnehmerInnen, die alle von der gleichen Universität kamen oder ihre sehr spezifische Aufgabe im Bereich Architekturgeschichte. Die Anwesenheit des Lehrenden und der KommilitonInnen kann ebenfalls zu einer Verzerrung der Antworten führen. Daher wurde im Juni 2019 mit einer weiteren Interviewphase begonnen. Diese Interviews sind losgelöst von einer spezifischen Aufgabe. Die TeilnehmerInnen werden insbesondere zu ihrem Forschungsprozess und ihren Qualitätskriterien im Umgang mit digitalen Bildrepositorien befragt. Von diesen Interviews wurden Audioaufnahmen gemacht, die transkribiert und dann ebenfalls nach Mayrings qualitativer Inhaltsanalyse ausgewertet wurden.

Ergebnisse der Studie

Phasen im Forschungsprozess

Im Gespräch berichteten die Studierenden, dass sie sich zu Beginn der Bildersuche einen ersten Eindruck verschaffen und die Aufgabe erfassen wollen. Meist wird auch vorab online und bei der Bibliothek geschaut, ob genug Material zum Thema zur Verfügung steht. Die initiale Bildersuche diente der Inspiration und Recherche nach verwandten Themen sowie der Hypothesenbildung (Frage: Wie sind Sie an Ihren Rechercheprozess herangegangen?). Im Anschluss änderte sich ihr Suchverhalten und Bedarf, um alle relevanten Informationen zu sammeln und zu strukturieren, um schlussendlich die Präsentation vorzubereiten. Die gesamte Recherche wurde von den Informationen beeinflusst, auf die sie währenddessen stießen.

Die Studierenden waren sich einige, dass Bilder hauptsächlich online gesucht werden. Bei der Frage, wo sie nach den Bildern gesucht haben (Frage: Welche digitalen Datenbanken und Plattformen haben Sie für die Bildsuche gewählt?), ergab sich folgende Listung: 1) Google, 2) der Bibliothekskatalog, 3) verschiedene Literatur und digitale Texte mit Bildern wie Bücher und PDF-Dokumente und 4) andere Datenbanken, z.B. Wikipedia, (Flickr), Instagram, Pinterest und andere ausländische Plattformen. Die Studierenden sind sich bewusst, dass sie auch eigene Fotografien nutzen können. In den Interviews gaben die Studierenden der TU Dresden an, spezifische Kunstwerke vorrangig bei Prometheus zu suchen, da es ihnen so im Seminar gezeigt wurde. Alternativen wie das Bildarchiv Foto Marburg und die Deutsche Fotothek waren teilweise sogar unbekannt, was aber auch am unterschiedlichen Fokus der Sammlungen liegen kann. Die Deutsche Digitale Bibliothek, die Zugang zu Sammlungen verschiedener Institutionen bietet und somit eine größere Trefferquote verspricht, wurde gar nicht genannt.

In Bezug auf die Bedeutung von Bildern für die Gestaltung der Forschung wurde bereits die starke Beeinflussung der WissenschaftlerInnen durch die Primärquellen im Forschungsprozess festgestellt (Long/Schonfeld 2014: 18). Empirisch gesehen suchen NutzerInnen benötigte Bilder mit vertrauten Abläufen (Beaudoin 2009: 286) und verwenden für die Suche in (unbekannten) Repositorien Fachbegriffe – „concept-based for theme, and object-based for thing" (Beaudoin 2009: 297).

Kriterien für die Bildersuche

Die Fokusgruppe zeigte, dass für die Studierenden eine präzise Suche am wichtigsten ist. Eine Suchanfrage enthält zwei Probleme: Irrelevante Punkte müssen weggelassen werden, um eine wesentlich schnellere Überprüfung der Ergebnisse zu ermöglichen, und relevante Punkte müssen hervorgehoben werden, um eine unverzerrte Wahrnehmung der Daten zu gewährleisten (Datta et al. 2008). Auch wenn das „Browsen" von großen Bildmengen als gründlicher Weg beim Zugriff auf eine Bilddatenbank (Besser 1990) anerkannt ist und von Menschen gerne als Inspiration genutzt wird (Hastings 1999), ist es angesichts der Vielzahl der heute angebotenen digitalisierten Bilder sehr zeitaufwändig. In der Regel greifen BenutzerInnen über eine Stichwortsuche auf Bildrepositorien zu. Dies erfordert die Übersetzung von visuellen Bedarfen in Text (Pisciotta 2001).

In der Kunstgeschichte fungieren Bilder als Digitalisat eines Kunstobjekts. Der Fokus liegt auf dem abgebildeten Objekt, das typischerweise ein bekanntes Objekt wie ein Gemälde oder eine Skulptur ist und gezielt nach z.B. Titel oder KünstlerIn gesucht wird (Hastings 1999). Fotografien einer Stadtansicht oder eines Gebäudes, die für die Architekturgeschichte von Bedeutung sind, benötigen eine spezielle Beschreibung für den Suchvorgang (Matusiak 2006). In den meisten Datenbanken müssen Schlüsselwörter für die Suche mit den Metadaten eines Bildes übereinstimmen.

Die Suche nach Bildern oder die Produktion von Texten sind in der Kunstgeschichte nach wie vor grundlegende Vorgänge auch wenn im Bereich der Digitalen Kunstgeschichte viel über neuartige Forschungsansätze diskutiert wird (Heusinger 1989, Kohle 2013, Drucker 2013, Bentkowska-Kafel 2015). Die digitalen Technologien zielen nicht unbedingt darauf ab die Methoden der ForscherInnen zu verändern (Long/Schon-

feld 2014: 42), sondern wollen neuartige Forschungsfragen beantworten, neue Analysetechniken anwenden oder die Nutzung von Technologie als Medium für neue Forschungspraktiken etablieren.

Wo wird gesucht?

Die Studierenden gaben an, gern Google zu nutzen, weil die Suchmaschine Schlagworte sehr gut handhabt und neben den zufriedenstellenden Suchergebnissen auch den weiteren Rechercheprozess befördert. Die vorzugsweise Verwendung dieser generischen Suchmaschine wurde bereits in anderen Studien festgestellt (Kemman/Kleppe/Scagliola 2014, Gregory 2007), auch wenn es verschiedene Repositorien gibt, die sich speziell an WissenschaftlerInnen aus den Bereichen Kunst und Architektur richten (Chen 2009). Die Volltextsuche von Google kann sehr gut mit der Stichwortsuche umgehen: Rechtschreibung, Sprach- und Namensvariationen oder lokale Namen sind für Google keine Herausforderung. Es wurde festgestellt, dass sich Google als hilfreich erwies, indem es geeignete Keywords vorschlug, die für andere Plattformen verwendet werden können (Gibbs/Owens 2012). Verwandte Suchergebnisse oder weitere Vorschläge sind z.B. auf Ähnlichkeitsanalysen mit z.B. Künstlicher Intelligenz oder die Analyse von Präferenzen anderer Nutzer zurückzuführen.

Die Handhabung von Google wird als Standard akzeptiert. Plattformen, die sich stark von Google unterscheiden und nicht den Usability-Standards entsprechen, sind im Nachteil (Kemman/Kleppe/Scagliola 2014). Die Studierenden sind sich bewusst, dass Bilder von Google aus urheberrechtlichen Gründen nicht unbedingt für Veröffentlichungen verwendet werden können und dass Google nicht die einzige verwendete Quelle sein sollte. Dies könnte damit zusammenhängen, dass den Studierenden die Qualität von Google für die Wissenschaft nicht genügt oder sie ein Bewusstsein für den Einfluss der Algorithmen auf die von Google angezeigten Ergebnisse entwickelt haben (Kemman/Kleppe/Scagliola 2014).

Veranstaltungen zum Wissenschaftlichen Arbeiten in der Kunstgeschichte sind Teil des Curriculums. In der Regel wird dort auch der Bibliothekskatalog oder eine Plattform wie *prometheus. Das verteilte digitale Bildarchiv für Forschung und Lehre* vorgestellt. Dadurch haben die Studierenden ein höheres Maß an Vertrauen in diese Angebote (Kemman/Kleppe/Scagliola 2014) und schauen nicht nach weiteren Quellen zur Verifizierung ihrer Ergebnisse. Urheberrechtsstatus und Bildqualität sind in der Regel zufriedenstellend. Die Studierenden nutzen auch andere Quellen, um einen umfassenderen Eindruck vom Forschungsobjekt zu erhalten. Allerdings werden Bilder in Textdokumenten wie digitalisierten Schulbüchern in der Regel nicht extra indiziert und können daher nicht direkt in einer Datenbank abgerufen werden. Sie sind aber sehr wertvoll, da die schriftlichen Ausführungen weitere Informationen liefern. Darüber hinaus ist es für die Studierenden sehr hilfreich, über Links Zugang zu verwandten Themen und weiteren Informationen zu haben. Ein eingeschränkter Zugang durch eine Registrierung stellt eine Barriere dar, die die Studierenden nur zögerlich angehen, da sie nicht wissen, ob sich die darüber zugänglichen Inhalte lohnen.

Empfehlungen

- Verbesserung der Metadatenqualität

Falsche oder unvollständige Metadaten sind auch heute noch ein Problem, welche den Erfolg bei der Suche nach Bildern beeinflussen (Beaudoin 2009: 298). Zum einen, weil die einschlägigen Datenbanken ihre Suchfunktion immer noch fast ausschließlich darauf ausrichten und zum anderen, weil die Nutzer sich damit auch immer noch arrangieren.

Crowdsourcing als Ansatz besitzt das Potential diesen Zustand der Metadaten zu verbessern (Nowak/Rüger 2010). Sofern Bilddatenbanken Kommentarfunktionen aufweisen, wird den KommentarverfasserInnen allerdings kaum geantwortet, weil kein Personal dafür zur Verfügung steht. Benötigt werden daher partizipative Plattformen. Insbesondere bei der Suche nach Fotos sind fehlende oder unzureichende Beschreibungen zu Bildinhalt und Kontext (Fotograf, Erstellungsdatum etc.) ebenfalls ein Hindernis.

Ein weiterer Ansatz stützt sich auf Künstliche Intelligenz und deren Einsatz für z.B. eine intelligente Verschlagwortung[2] sowie u.a. Objekt- und Kontexterkennung[3] .

Für eine Suche überflüssig werden Metadaten, wenn Fotografien in einem digitalen 3D-Stadtmodell verortet werden. So können auf den Fotos abgebildete Gebäude mit den virtuellen Architekturmodellen direkt verlinkt und darüber gefunden werden (Maiwald et al. 2019: 9).

- Verbesserung der Benutzerfreundlichkeit

Digitale Repositorien sind heute nicht nur Werkzeuge für ExpertInnen aus der Informatik. Daher ist es umso wichtiger, Feedback zu Funktionalitäten und Interfacedesign von den tatsächlichen NutzerInnen zu erhalten und bei der Entwicklung digitaler Lösungen zu berücksichtigen. Somit kann auch eine Nachhaltigkeit der Anwendungen gestärkt werden.

- Sensibilisierung für digitale Prozesse

WissenschaftlerInnen als potentielle NutzerInnen sollten schon in den frühen Phasen der Digitalisierungsprozesse einbezogen werden, in dem ihre Bedarfe identifiziert und berücksichtigt werden. Ebenso wichtig ist die Transparenz hinsichtlich der Entscheidungen, die beim Aufbau einer digitalen Ressource getroffen werden.

Ausblick

Der in diesem Beitrag vollzogene Überblick über die Verwendung von Bilddatenbanken durch KunsthistorikerInnen und deren Nutzungsbedarfen soll Ansätze zur Umsetzung und Weiterentwicklung der formulierten Empfehlungen liefern. Darüber hinaus werden wichtige Handlungsbedarfe in den Bereichen Verbesserung der digitalen Zugänglichkeit von Bildern (z.B. hinsichtlich Verknüpfung von Bildern und Informationen, beständige technologische Aktualisierung von Anwendungen) und Langzeitarchivierung digitaler Daten gesehen.

Forschungsförderung

Die diesem Beitrag zugrundeliegende Forschung ist Teil der Aktivitäten der Nachwuchsforschungsgruppe HistStadt4D, die vom Bundesministerium für Bildung und Forschung im Rahmen der Fördervereinbarung Nr. 01UG1630 gefördert wird.

Fußnoten

1. Vgl. dazu Artikel „Scientific method“ in der englisch-sprachigen Wikipedia, online verfügbar unter: https://en.wikipedia.org/wiki/Scientific_method (16.09.2019).
2. Exploring art with open access and AI: What's next?, in The Met Museum. Online verfügbar unter: https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-exploring-art-open-access-ai-whats-next. (09.09.2019)
3. Artificial intelligence as a bridge for art and reality, von J. H. Dobrzynski, in The New York Time, 25.10.2016

Bibliographie

Beaudoin, Joan E. (2009): *An Investigation of Image Users across Professions: A Framework of Their Image Needs, Retrieval and Use*. PhD, Drexel University.

Benardou, Agiatis / Constantopoulos, Panos / Dallas, Costis / Gavrilis, Dimitris (2010): "Understanding the Information Requirements of Arts and Humanities Scholarship", in: *International Journal of Digital Curation* 1 (5): 18-33.

Bentkowska-Kafel, Anna (2015): "Debating Digital Art History", in: *DAH-Journal* 1: 50-64.

Besser, Howard (1990): "Visual access to visual images: the UC Berkeley Image Database Project", in: *Library trends* 38 (4): 787-798.

Chen, Ching-Jung (2009): "Art history: a guide to basic research resources", in: *Collection Building* 28 (3): 122-125.

Datta, Ritendra / Joshi, Dhiraj / Li, Jia / Wang, James Z. (2008): "Image retrieval: Ideas, influences, and trends of the new age", in: *ACM Comput. Surv.* 40 (2): 1-60.

Drucker, Johanna (2013): "Is There a 'Digital' Art History?", in: *Visual Resources* 29 (1-2): 5-13.

Heusinger, Lutz (1989): "Applications of Computers in the History of Art", in: Hamber, Anthony / Miles, Jean / Vaughan, William (Hrsg.): *Computers and the History of Art*. London and New York: Mansell Pub. 1-22.

Gibbs, Fred / Owens, Trevor (2012): "Building better digital humanities tools", in: *DH Quarterly*, 6 (2): o.S.

Gregory, Tori R. (2007): "Under-Served or Under-Surveyed: The Information Needs of Studio Art Faculty in the Southwestern United States", in: *Art Documentation: Journal of the Art Libraries Society of North America* 26 (2): 57-66.

Hastings, Samantha Kelly (1999): "Evaluation of image retrieval systems: Role of user feedback", in: *Library trends* 48 (2): 438.

Kemman, Max / Kleppe, Martijn / Scagliola, Stef (2014): 'Just Google It'. in: Mills, Clare /Pidd, Michael /Ward, Esther (Hrsg.): *Proceedings of the Digital Humanities Congress 2012 . Studies in the Digital Humanities*. Sheffield: HRI Online Publications o.S.

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

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

Maiwald, Ferdinand / Bruschke, Jonas / Lehmann, Christoph / Niebling, Ferdinand (2019): "A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR", in: *Virtual Archaeology Review* 10 (21): 1-13.

Matusiak, Krystyna K. (2006): "Towards user-centered indexing in digital image collections", in: *OCLC Systems & Services: International digital library perspectives* 22 (4): 283-298.

Mayring, Philipp (2008): *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Beltz Deutscher Studien Verlag.

Nowak, Stefanie / Rüger, Stefan (2010): "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-labelimage annotation", in: *Proceedings of the international conference on Multimedia information* 557-566.

Palmer, Carole L. / Teffeau, Lauren C. / Pirmann, Carrie M. (2009): *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*, Report commissioned by OCLC Research, online verfügbar unter: http://www.oclc.org/programs/publications/reports/2009-02.pdf (letzter Zugriff 16. September 2019).

Pertsas, Vayianos / Constantopoulos, Panos (2017): "Scholarly Ontology: modelling scholarly practices", in: *International Journal on Digital Libraries* 18 (3): 173-190.

Pisciotta, Henry / Brisson, Roger / Ferrin, Eric / Dooris, Michael / Spink, Amanda (2001): "Penn State visual image user study", in: *D-Lib Magazine* 7(7/8): 169-196.

Ross, Seamus (2010): *Expert Forum on Scholarly Activity and Information Process*, 10.-11. Juni 2010 in Athen.

Unsworth, John (2000): "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?", in: *Symposium on Humanities Computing: formal methods, experimental practice*.

Computationelle Textanalyse als fünfdimensionales Problem

Gius, Evelyn
gius@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Die Einschätzung und Kritik computationeller Textanalyse

In diesem Beitrag wird ein Modell vorgestellt, das zu einer Einschätzung der Komplexität von Forschungsansätzen dient, die sich Texten mit computationellen Analysen nähern. Das Modell wurde vor dem Hintergrund der (literaturwissenschaftlichen) Analyse von literarischen Texten entwickelt, es ist jedoch – ggf. mit leichten Anpassungen – für Textanalysen generell geeignet.

Die Komplexität von Digital Humanities-Projekten ist bestimmt von der Aushandlung von Vorannahmen, Methoden, der Passung zum Gegenstand, der konkreten interdisziplinären Zusammenarbeit, die fachlich, persönlich und oft auch

karrierestrategisch eine große Herausforderung für die Beteiligten sein kann, bis hin zur Darstellung von Ergebnissen für eine oder mehrere Forschungscommunities. Neben Fragen der Projektplanung und -steuerung, wissenschaftspolitischen und wissenschaftskommunikativen Aspekten geht es auch um Fragen, die das eigentliche Forschungsgeschehen betreffen. Dieses wird aktuell in Bezug auf seine Relevanz und Ausrichtung diskutiert: Eine harsche Kritik von Nan Z. Da (2019a) an den Verfahren der DH initiierte eine mit dem etwas überzogenen Begriff „Digital Humanities War" bezeichnete Auseinandersetzung.[1] Diese Debatte wird z.T. als Auseinandersetzung zwischen angeblichen Strukturalist*innen und Poststrukturalist*innen dargestellt. Zumindest von letzteren, die den Strukturalismus als solchen benennen und eine Kluft zwischen diesem und den eigenen Zugängen diagnostizieren (vgl. z.B. Dobson 2019 und Bode im Erscheinen). Hinzu kommt, dass in der Auseinandersetzung förderpolitische Aspekte zumindest als Hintergrund eine große Rolle spielen.[2]

Ein methodenunabhängiges Modell

Diese Auseinandersetzungen gehen zum Großteil an den eigentlichen Forschungszugängen vorbei. Dabei wäre es aus Sicht der Digital Humanities *und* der Literaturwissenschaft erhellend, die diskutierten Verfahren oder gar Methodenlinien detaillierter zu beschreiben und ihre Bedeutung zu reflektieren. Deshalb möchte ich ein Modell vorschlagen, das eine solche Betrachtung von computationellen Textanalyseansätzen ermöglicht und eine Grundlage bildet, auf der Textanalyse-Zugänge unabhängig von ihrer literaturtheoretischen Fundierung beschrieben, kritisiert und zu verglichen werden können.

Ausgangspunkt des Modells sind die drei Aspekte, die für jede computergestützte Textanalyse wesentlich sind: Die Phänomene, denen das Interesse gilt, die Texte, die untersucht werden, und die Art, wie Erkenntnis erzeugt wird.[3] Aus diesen Aspekten lassen sich insgesamt fünf Dimensionen ableiten, die für die Einschätzung der Komplexität eines Zugangs genutzt werden können: die Kontextualisierung von Phänomenen, die Zusammengesetztheit von Phänomenen, die Heterogenität von Texten, der Analysemodus und der Erkenntnisbeitrag computationeller Analysen.

Zusammengesetztheit von Phänomenen

Eine Einschätzung der Phänomene, die in einer computationellen Textanalyse untersucht werden, kann anhand der Phänomenbeschreibung stattfinden. Für diese kann man fragen: Wird das Phänomen als einfach, nicht weiter unterteilt, oder als aus mehreren Phänomenen zusammengesetzt betrachtet? Dabei geht es wohlgemerkt nicht um eine allgemein gültige Definition des entsprechenden Phänomens, sondern um die von den Forscher*innen genutzte Beschreibung.

Beschreibungen für dasselbe Phänomen können in unterschiedlichen Forschungsprojekten entsprechend unterschiedlich ausfallen. In Bezug auf ein aktuelles Forschungsprojekt zu Gender und Krankheit in literarischen Prosatexten[4] sind zum Beispiel folgende Unterschiede denkbar: Man könnte das Phänomen „Krankheit einer literarischen Figur" ausschließlich daran festmachen, ob diese ärztlich behandelt wird. Man kann aber ebenso eine Reihe von Phänomenen wie körperliche Reaktionen, Aussagen der Figur etc. nutzen, um Krankheit zu bestimmen.

Kontextualisierung von Phänomenen

Neben der Bestimmung der Teile, aus denen eine Phänomenbeschreibung zusammengesetzt ist, geht es auch um die Frage, welches Wissen zur Bestimmung des Phänomens herangezogen werden muss. Dies kann zum einen Wissen sein, das der Text vermittelt. Aber es kann auch weiteres Wissen nötig werden, wie etwa spezielles Domänenwissen, zusätzliches (innerfiktionales oder außerfiktionales) Weltwissen u.ä. Die Kernfrage ist entsprechend: Braucht man über das Textwissen hinausgehendes weiteres Wissen, um ein Phänomen zu identifizieren?

Auch hier gilt: Die Einstufung der Komplexität gilt für den betrachteten Anwendungsfall, andere Fälle haben ggf. für dieselben Phänomene andere Komplexitätsgrade. Im Projekt Gender und Krankheit wurde etwa mit Koreferenz-Auflösung experimentiert, die überwiegend auf Textphänomenen basiert. Das Krankheitskonzept wiederum wurde unter Rückgriff auf Wissen für zeitgenössische Krankheiten und Krankheitsbezeichnungen bearbeitet (etwa „Phthise" als Bezeichnung für Tuberkulose).

Abbildung 1 stellt beispielhaft die beiden Dimensionen der Komplexität einiger Phänomene dar, die im Projekt Gender und Krankheit eine Rolle spielen.

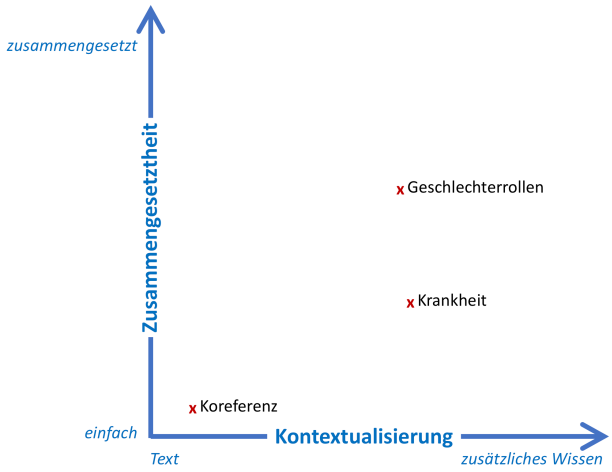


Abbildung 1: Komplexitätsdimensionen für Phänomene

Textheterogenität

Oft wird vorschnell angenommen, dass für die textorientierten Digital Humanities die nun wesentlich größere Menge an untersuchten Texten distinktiv ist. Dabei ist die Frage, ob es sich um – vermeintliche – Big Data handelt oder nicht, aus Sicht der computationellen Textanalyse nur insofern interessant, als damit die Frage zusammenhängt, ob man die Texte, die man analysiert, kennt bzw. kennen kann oder nicht. In Bezug auf die Komplexität der genutzten Texte relevanter ist hingegen die umfassendere Frage: Wie viele (wie) verschiedene Texte werden analysiert? Dabei fällt unter Heterogenität von Texten die Anzahl der Texte selbst, aber auch die Anzahl von

verschiedenen Texteigenschaften, die für die Fragestellung relevant sind bzw. sein könnten. Im Fall literarischer Texte sind das typischerweise Eigenschaften wie Gattung, Genre, Epoche, Autorgender, Erscheinungsort etc.

Die Textheterogenität reicht von einem Text bis zu sehr vielen, sehr heterogenen Texten reicht[5] und ist v.a. im Vergleich zu anderen Vorhaben beurteilbar. Im Projekt Gender und Krankheit liegt eine vergleichsweise hohe Textheterogenität vor, da das Korpus aus über 2.000 deutschsprachigen Texten besteht, die verschiedene Genres, Autor*innen und Epochen zuzuordnen sind.

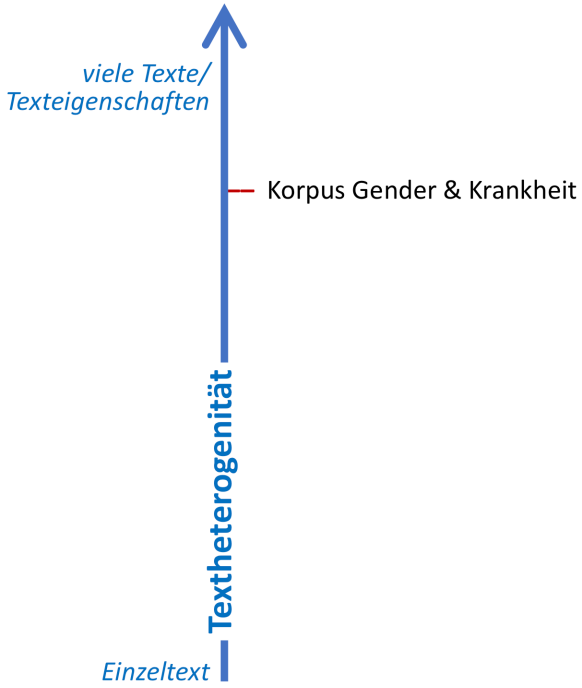


Abbildung 2: Komplexitätsdimension Textheterogenität

Analysemodus

In der Komplexitätsdimension des Analysemodus geht es darum, wer die Erkenntnisse produziert. Hier sind die beiden Möglichkeiten recht offensichtlich: Auf der einen Seite steht (menschliches) Lesen, auf der anderen Seite maschinelles Erschließen. Die Hauptfrage ist also: Wird die Textbasis durch Menschen oder durch Computer erschlossen? Dabei wird für alle Zugänge als gegeben vorausgesetzt, dass der Computer genutzt wird. Während das Lesen in Annotationen von Textstellen oder zumindest in die Ergänzung der Texte um Metainformationen resultiert, wird beim maschinellen Erschließen im Normalfall Textmining betrieben. Beide Textzugangsarten können weiter differenziert werden nach der Interpretationstheorie (etwa in text-, leser- oder autororientierte Zugänge) bzw. dem angewendeten maschinellen Verfahren (etwa in regelbasierte und Lernverfahren).

In konkreten Forschungsprojekten kommen fast immer beide Modi vor. So werden im Projekt Gender und Krankheit manuelle Annotationen von Textpassagen und halb-automatische Verfahren zur Wortfeldgenerierung für die weitere Verarbeitung oder die Methodenentwicklung mit automatischen Verfahren zur Figurenerkennung, Segmentierung und Sentimentanalyse kombiniert. Da die Zwischenschritte in der Analyse zumeist manuell überprüft und teilweise ergänzt werden, handelt es sich hier um ein Verfahren zwischen Lesen und automatischem Erschließen und damit um eine eher geringe Komplexität.

Erkenntnisbeitrag

Schließlich geht es bei der Betrachtung von computationellen Textanalysen auch darum, wie der Computer eingesetzt wird, um Erkenntnisse zu generieren. Wenn man von der literaturwissenschaftlichen Praxis der Textanalyse ausgeht, ist die komplexeste Aufgabe jene, die Textbasis insgesamt im Hinblick auf die gewählte Fragestellung zu interpretieren. Interpretation ist jedoch bislang nicht der Fokus computationeller Zugänge zu literarischen Texten. Trotzdem lohnt es sich, Interpretation als ein Extrem der Dimension der Erkenntnis zu denken. In Anlehnung an die literaturwissenschaftliche Praxis kann man die Komplexitätsdimension des Erkenntnisbeitrags computationeller Analysen als von der Analyse des Textes für ein erstes Textverständnis bis hin zur Interpretation der Textbasis als Ganzes ausgedehnt sehen.[6] Alternativ kann auch die sozialwissenschaftliche Kategorisierung von Forschungslogiken[7] in Anlehnung an Peirce (1935) in Deduktion, Induktion und Abduktion als Skala für die Erkenntnisdimension genutzt werden.

Unabhängig von der Frage, welche Systematik man für die Tätigkeiten verwendet, die mit Textverstehen befasst sind, ist die zentrale Frage in der letzten Komplexitätsdimension: Wie weit geht der Erkenntnisbeitrag der computationellen Methode? Es geht also um die Frage nach der Neuheit des computationell Erforschten. Grob kann man die Komplexitätsstufen des Erkenntnisbeitrags wie folgt erfassen: Werden in einer deduktiven bzw. einfachen Textanalyse aufgrund von bestehenden Hypothesen bzw. Regeln (also bestehenden Analysekategorien und -verfahren) durchgeführt, werden aus der Betrachtung von Texten neue Analysekategorien oder auch Taxonomien entwickelt oder handelt es sich um Hypothesen über größere Zusammenhänge in den Texten, also um ihre Interpretation?[8]

Bei der Auseinandersetzung mit der Komplexitätsdimension des Erkenntnisbeitrags ist zu beachten, dass in einer typischen literaturwissenschaftlichen Textanalyse meist alle Modi vorliegen und fließend ineinander übergehen. Für die Komplexitätseinschätzung ist relevant, welche Modi davon computationell unterstützt werden sollen. Im Fall des Projekts zu Gender und Krankheit soll etwa deduktiv die Veränderung der Figurenkonstellation anhand der Figurennennungen analysiert werden. Ein induktives Verfahren liegt vor, wenn Genderkategorien durch Clustering von Figurenrede herausgearbeitet werden (die dann wieder deduktiv in der Analyse genutzt werden). Und schließlich liegt ein abduktiver Zugang vor, wenn durch eine Gesamtbetrachtung ein neues Element entdeckt würde, das Figurenkrankheit beeinflusst.

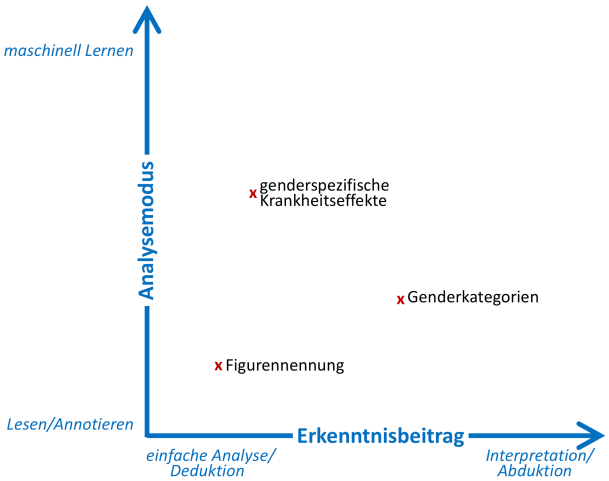


Abbildung 3: Komplexitätsdimensionen für Erkenntnis: Analysemodus und Erkenntnisbeitrag

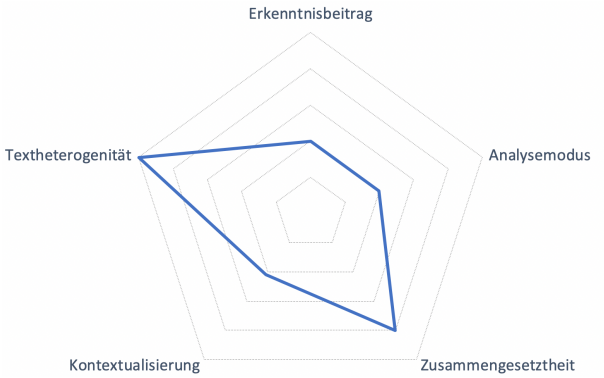


Abbildung 4: Übersichtsdarstellung: Komplexität im Projekt Gender und Krankheit

Zur Nutzung des Modells

Wie bereits dargelegt, betrifft die Bestimmung der Komplexität in den fünf Dimensionen primär die normativen Setzungen durch die Forscher*innen. Ausschlaggebend ist weniger, wie Texte, Phänomene und Erkenntnis an sich modelliert werden *sollten*, sondern vielmehr, wie die Modellierung konkret umgesetzt wird. Die vorgeschlagenen Dimensionen sind außerdem von der mit einem Zugang verbundenen Interpretationstheorie unabhängig. Damit ist das Modell für alle literaturwissenschaftlichen Textanalyseverfahren geeignet, für jene, die in einer strukturalistischen Tradition gesehen werden können, genauso wie für solche, die eher postmoderne Zugangsweisen umsetzen – oder andere Zugänge.

Für die Betrachtung und Kritik eines Zugangs sollten alle fünf Dimensionen berücksichtigt werden. Damit vermeidet man auch vorschnelle Kritik, die sich auf eine einfache Modellierung einer Dimension beschränkt und den Zugang insgesamt als unterkomplex betrachtet, obwohl er in einer oder mehreren anderen Dimensionen Erhebliches leistet.

Darüber hinaus eignet sich das Modell als Instrument für den Entwurf eines Zugangs. Es kann in allen Phasen computationeller Textanalyse genutzt werden – vom Design des Forschungszugangs zu Beginn der Forschungsarbeit über die wiederholten Bestandsaufnahme oder Nachjustierung im Projektverlauf bis hin zur Einordnung der erzielten Ergebnisse am Ende und der Reflektion des gesamten Prozesses.

Abschließend seien noch einmal die fünf Dimensionen mit ihren Kernfragen dargestellt:

1. Komplexitätsdimension 1: die Zusammengesetztheit von Phänomenen
   - Frage *: Wird das Phänomen als einfach, nicht weiter unterteilt, oder als aus mehreren Phänomenen zusammengesetzt betrachtet?*
   - Komplexität: von einfach bis hin zu vielfach zusammengesetzten Phänomenen
2. Komplexitätsdimension 2: die Kontextualisierung von Phänomenen
   - Frage: *Braucht man über das Textwissen hinausgehendes weiteres Wissen, um ein Phänomen zu identifizieren?*
   - Komplexität: von Textwissen bis hin zu verschiedenen Arten von umfangreichem weiterem Wissen
3. Komplexitätsdimension 3: Textheterogenität
   - Frage *: Wie viele (wie) verschiedene Texte werden analysiert?*
   - Komplexität: von einem Text mit homogenen Eigenschaften bis hin zu vielen, in sich und zueinander heterogenen Texten
4. Komplexitätsdimension 4: Analysemodus
   - Frage *: Wird die Textbasis durch Menschen oder durch Computer erschlossen*?
   - Komplexität: von von Menschen annotiert bis hin zu von Maschinen durch Lernen analysiert
5. Komplexitätsdimension 5: der Erkenntnisbeitrag computationeller Analysen
   - Frage: *Wie weit geht der Erkenntnisbeitrag der computationellen Methode?*
   - Komplexität: von der Anwendung simpler Regeln auf einzelne Textelemente bis zur Interpretation der gesamten Textbasis.

Fußnoten

1. Vgl. dazu den Artikel „The Digital Humanities Debacle. Computational methods repeatedly come up short“ von Da (2019b) und als „Digital Humanities War“ zusammengefasste Reaktionen auf https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986, sowie das „Special Forum on Responses to Nan Z. Da“ in *Cultural Analytics* auf https://culturalanalytics.org/2019/09/special-forum-on-re-

sponses-to-nan-z-da/, sowie Jannidis (2019) und Krajewski (2019).
2. Nan Z. Da (2019b) könnte diesbezüglich so zusammengefasst werden, dass sie vorschlägt, keine Mittel mehr in die Computational Literary Studies zu stecken und damit zu verschwenden.
3. Damit ist der Zugang wesentlich spezifischer als die umfassende „TaDiRAH - Taxonomy of Digital Research Activities in the Humanities“, die alle digitalen Forschungsaktivitäten zu erfassen versucht (vgl. http://tadirah.dariah.eu/vocab/, gesehen am 21.12.2019).
4. Vgl. dazu z.B. Gius et al. (2019), Andresen et al. (2019) und https://www.herma.uni-hamburg.de/subprojects.html , gesehen am 21.12.2019).
5. Genau genommen werden hier zwei Gegensatzpaare abgebildet: Anzahl (von Texten) und Heterogenität (von Texteigenschaften). Diese Eigenschaften werden zu einer Dimension zusammengefasst, da sie die Komplexität von Texten vergleichbar steigern.
6. „Textanalyse“ ist literaturwissenschaftlich mehrdeutig, da der Begriff sowohl eine Textanalyse meint, die das Textverständnis im Fokus hat und der anschließenden Interpretation als Voraussetzung dient, als auch den Prozess der Analyse und Interpretation insgesamt, vgl. dazu Winko (2003).
7. Vgl. dazu auch die Arbeit im Projekt hermA zu den verschiedenen Forschungslogiken im Kontext von Annotationen (Gaidys et al. 2017 bzw. www.herma.uni-hamburg.de).
8. Vgl. dazu auch Eco (1987): „[D]er Text ist ein Objekt, das die Interpretation im Verlauf ihrer zirkulären Anstrengungen um die eigene Schlüssigkeit bildet auf der Basis dessen, was sie als ihr Resultat erschafft. Ich schäme mich nicht, daß ich auf diese Weise den alten und immer noch gültigen hermeneutischen Zirkel definiere. Die Logik der Interpretation ist die Peircesche Logik der ‚Abduktion‘.“

Bibliographie

Adelmann, Benedikt / Melanie Andresen / Anke Begerow / Lina Franken / Evelyn Gius / Michael Vauth (2019): „Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics“. In *DH2019 Book of Abstracts*. Utrecht.

Bode, Katherine. Im Erscheinen. „Why you can’t model away bias“. Preprint: Modern Language Quarterly 80.3.

Da, Nan Z. (2019a): „The Computational Case against Computational Literary Studies“. *Critical Inquiry* 45 (3): 601–39. https://doi.org/10.1086/702594.

Da, Nan Z. (2019b): „The Digital Humanities Debacle“. *The Chronicle of Higher Education*, 27. März 2019. https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986.

Dobson, James E. (2019): *Critical Digital Humanities: The Search for a Methodology*. Topics in the digital humanities. Urbana, Illinois: University of Illinois Press.

Eco, Umberto (1987): *Lector in Fabula. Die Mitarbeit der Interpretation in erzählenden Texten*. München: Hanser.

Gaidys, Uta / Evelyn Gius / Margarete Jarchow / Gertraud Koch / Wolfgang Menzel / Dominik Orth / Heike Zinsmeister (2017) : „Project description – hermA: Automated modelling of hermeneutic processes“. *Hamburger Journal für Kulturanthropologie*. https://journals.sub.uni-hamburg.de/hjk/article/view/1213.

Gius, Evelyn / Katharina Krüger / Carla Sökefeld (2019): „Korpuserstellung als literaturwissenschaftliche Aufgabe“. In *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts*, 164–166. Frankfurt & Mainz.

Jannidis, Fotis. (2019): „Digitale Geisteswissenschaften: Offene Fragen - schöne Aussichten“. Herausgegeben von Lorenz Engell und Bernhard Siegert. *Zeitschrift für Medien- und Kulturforschung*. https://doi.org/DOI: 10.28937/ZMK-10-1.

Krajewski, Markus (2019): „Hilfe für die digitale Hilfswissenschaft. Eine Positionsbestimmung“. Herausgegeben von Lorenz Engell und Bernhard Siegert. *Zeitschrift für Medien- und Kulturforschung*. https://doi.org/DOI: 10.28937/ZMK-10-1.

Peirce, Charles S (1935): *Collected Papers of Charles Sanders Peirce, Volumes V and VI: Pragmatism and Pragmaticism and Scientific Metaphysics*. Herausgegeben von Charles Hartshorne und Paul Weiss. Cambridge, Mass: Belknap Press of Harvard Univ. Press.

Winko, Simone (2003): „Textanalyse“. In *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, herausgegeben von Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, und Klaus Weimar, 3., neubearb. Aufl. Berlin: De Gruyter: 597–601.

Versch. Autoren (2019): „Special Forum on Responses to Nan Z. Da“. Journal of Cultural Analytics. 17. September 2019. https://culturalanalytics.org/2019/09/special-forum-on-responses-to-nan-z-da/.

Confounding variables in Sub-Genre classification: instructive problems

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

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Würzburg

Leinen, Peter
P.Leinen@dnb.de
Deutsche Nationalbibliothek

This paper started out as a report on the state of the art in text classification, but over time it became much more a reflection on the pitfalls in modeling genre using classification. The start of our research was motivated by developments in text classification: Recent years have seen new approaches like gradient boosting and deep neural networks. Our initial goal was to inform about these approaches, which are seldom used yet in the digital humanities. But this proved to be only a starting point for a deeper exploration of genre structures of our collection of dime novels (‘Heftromane’, ‘Groschenromane’).

Most research on genre classification has been looking into what you could call 'high level classes' like newspaper genres (news, editorials etc.; e.g. Frank and Bouckaert, 2006) or web genres (blog, personal website etc.; e.g. Eissen and Stein, 2004). Under this perspective all texts we are looking at belong to one genre: the novel. The subgenres are types of love stories like the doctor novel (‚Arztroman') or the country novel (‚Heimatroman') and types of adventure novels, mainly distinguished by the setting: the war novel (‚Kriegsroman') or the science fiction novel. These novels are cheap ('dime novels') and published in a booklet format and are usually distributed via magazine kiosks and not book shops (Stockinger 2018). From the very beginning it was clear to us, that they don't contain a random collection of each genre. On the contrary, the crime novels for example are just a small and very specific subsection of crime novels in general. But nevertheless we assumed that genre is the main aspect to group novels - for publishers and readers.

Our dataset consists of 11,600 dime novels from 12 different genres (see Fig.1). The genre label come from the four publishers who divide the market among themselves. (Bastei, Martin Kelter, Pabel Moewig and Cora). The corpus has been documented in previous studies such as Jannidis et. al. (2019a) and Jannidis et. al (2019b).

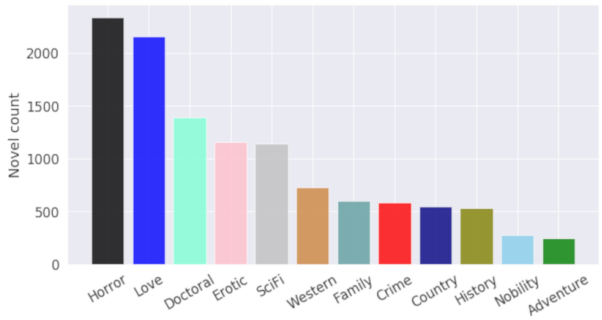


Figure 1: Novels per Genre

We have employed three groups of methods: traditional feature-based classifiers (Group A), modern feature-based classifiers (Group B) and deep learning (Group C). While Group A and B are based on document-term-matrix (20,000 most-frequent-words, tf-idf-weighted, stopwords removed, dimensionality reduction with LSI to 1000 features) as input, Group C works with unprocessed text. Named entities are removed completely. Hyperparameter optimization was done by sampling from the space of values recommended by the documentation of the libraries and (Olson et al. 2017) using Optuna (Akiba et al. 2019): In table 1 we report the best performance. We evaluated the performance of the deep learning approaches in advance on a smaller dataset, so that later only the best architecture had to be extensively tested (table 1). To increase speed initialized with pre-trained (wikipedia.de+30.000 novels) fasttext embeddings (Bojanowski et al. 2016). As a compromise between performance and speed we used the BiRNN architecture for all following experiments.

Table 1: Prestudy of deep learning architectures (4 subgenres, 800 novels)

	Fasttext	Flair	CNN	CNN+BiRNN	BiRNN	HATTN
f1-score	.886	.931	.925	.935	.923	.926
Time per epoch (seconds)	<5	288	210	190	90	215
Time to converge (minutes)	3	48	28	25	6	21

Table 2: Results of subgenre classification[1]

	Multi. NaiveBayes[2]	Logistic Regression	SVM (svc)	K-Nearest Neighbors
f1-score	.932	0.940	0.948	0.915
	XGBoost	LightGBM	CatBoost	BiGRU
f1-score	*	.878	*	.907

As was to be expected from the experience of previous studies on genre classification, the results were initially very good (Jannidis et. al. 2019a). They decreased slightly (~ 2 %) when we added novels from the publisher "CORA". With this addition our collection contains almost all dime novels published in recent years. Table 2 shows the classification results for this collection.

The decrease of our F1-score alone wasn't a great surprise, as the addition of new data is expected to increase diversity within groups and complicates classification. But two observations were irritating: First, we noticed classification results were improved when we included stopwords. Usually removing stopwords improves classification performance (Toman et al. 2006; Gonzales and Quaresma 2014). As most stopwords are typical function words which are used in stylometric research, this indicated that authorship information was used in the classification. Secondly, we noticed strong fluctuations between cross-validation folds, which seemed to indicate a very uneven class distribution.

To understand the first phenomenon better, we plotted the distribution of the authors across the genres (see Fig. 2): Many authors write exclusively within a genre. The greatest overlap can be found in the genres *Love* and *Family*.

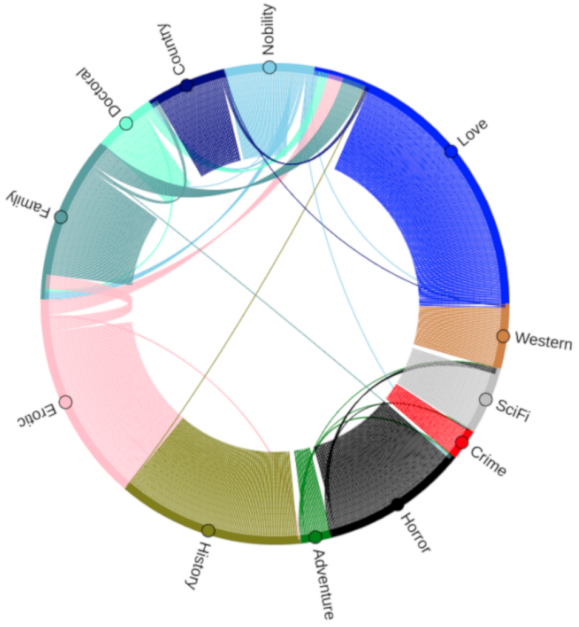


Figure 2: Inter-genre authorship

So, indeed, the authorship information could be used to identify the genre of text, but not in all genres equally.

In order to gain an insight into the influence of genre and publisher on the text form, we use Ivis (Szubert 2019) for unsupervised dimensionality reduction. The coloring of the data points according to publisher (figure 3) and genre (figure 4) shows the strong influence of these variables on the texts. It is also clear that Cora Verlag allows less variance among genres and thus becomes the most discriminatory factor. Figure 5 shows a detail of the previous plot, but focuses on microstructures. Theses structures indicate, that on this level genre and publisher are not enough to explain the distribution and that something else – author or series – comes into play.

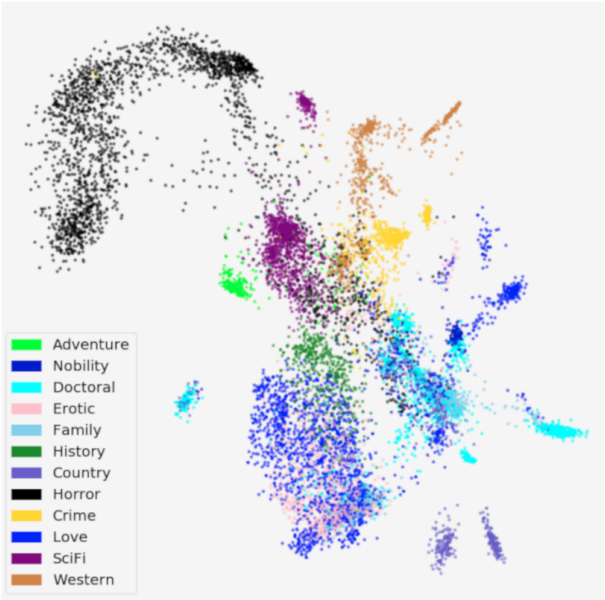


Figure 3: Ivis dimension reduction based on 20.000 mfw. Colors indicate genre.

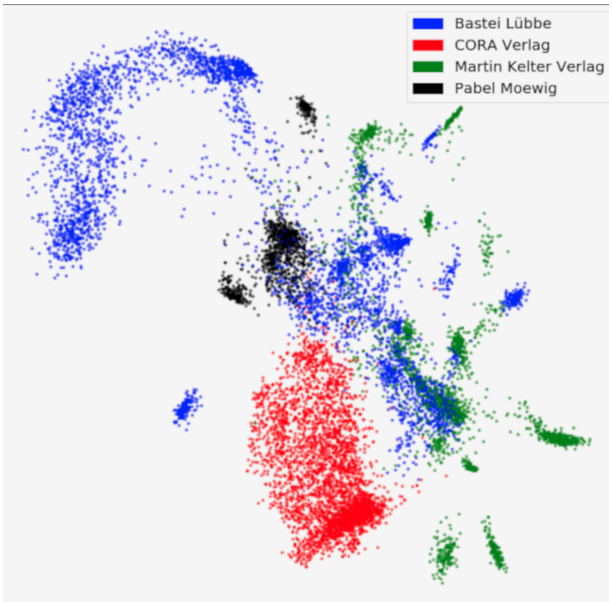


Figure 4: Ivis dimension reduction based on 20.000 mfw. Colors indicate publishers.

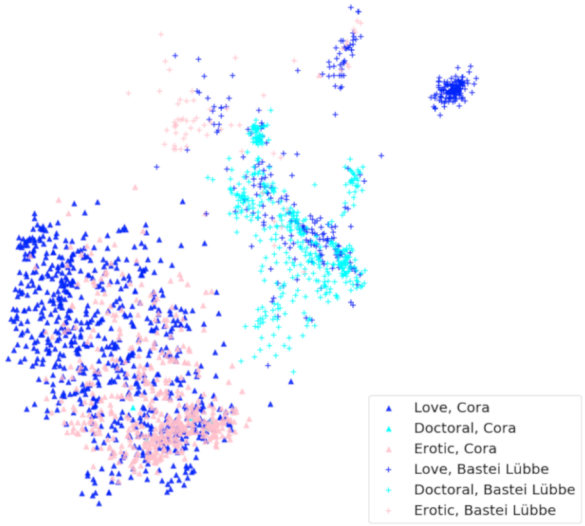


Figure 5: Detail of fig. 4, showing genres from publishing houses Bastei and Cora.

Obviously the variables publisher, author and series are influencing the distributions of our features, the words of the texts, and the variable, we want to predict, the genre. In a classical scientific model publisher, author and series would be called *confounding variables*, but in text classification the role of confounding has been mostly overlooked, probably because usually the main goal is prediction and not causal inference (Landeiro / Culotta 2016). Confounding variables are those factors in statistical models, that lead to false correlations or bias. For example, in an experiment that investigates the relationship between age of a person and the tendency to drive fast, the car would be a confounding variable. Because older people have probably a higher income and own faster cars. Something very similar is happening here. In the next section, we will apply a standard measure to control for confounding variables (restriction), while keeping the machine learning setup.

We created a restricted setup with a clear separation of authors, series and publishers between training and test data (i.e. authors which were in the training data, were not included in the test data etc.), and tested the subgenres in an one-vs-rest scheme. Figure 6 shows the results of this setup with at least 30 different combinations of test and training data per genre and a sample size of 200 novels split in half for training and test data.

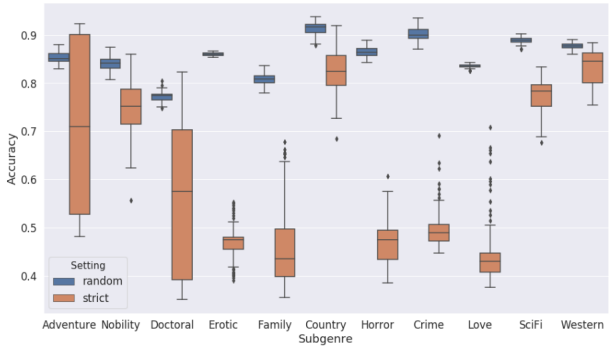


Figure 6: Binary Classification of Genres (Logistic Regression). Strict: No shared authors, series or publishers in training and testing dataset. Random: random sample to compare performances. Historic novels are excluded due to insufficient data.

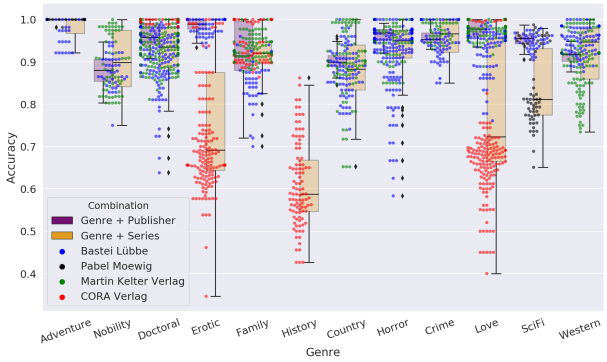


Figure 7: Classification of series and publisher within a genre (one-vs-rest scheme). Points of single observations are colored by the series publishers.

The performance of the 'strict setups' is lower, sometimes even below 50%. This behavior is the result of negative examples in the training data being more similar to the positive examples of the test data, for example in love and doctoral novels of Cora.

Though now we control for confounding variables, it is less clear, what it implies for the genre model. It is not unusual in genre theory to conceptualize genre in an ideal way as independent of other factors like authorship, time, publisher etc. which corresponds to the 'strict' version of splitting train and test data. But at the same time, these factors may be so intertwined with the genre features, that it is difficult, if not impossible to separate them at all (Hempfer 2010). Under this perspective our attempt to construct a 'clean' and strict model of genre, independent of publishers etc. is a misguided attempt.

Looking back we now see that we started our research with some assumptions which seem to be unfounded for this part of the literary market which is dominated by four publishers: We assumed that the genre labels have the same function as in the rest of the literary market. But the small number of publishers seems to create a different situation. We assume now, that at least in some instances combinations of genre names with publisher names (loves stories from Cora vs. love stories from Bastei-Lübbe) describe the clusters best. To start to evaluate this hypothesis, we trained the corpus on label combinations: 1) Genre and Publisher, e.g. 'Cora-Love', 2) Genre and Series. Figure 6 shows, that in many, but not all cases these combinations achieve very good results, which indicates that a clear-cut set of features corresponds these combinations. In some genres the same is true for series, for example doctoral or horror, while in others the series have no clear feature set (erotic, love).

To explore this in more detail, we looked at those genres where the values for a randomized and a strict setup in figure 4 are markedly different, which we see as a sign of a heterogeneity of the genre which was masked in the random setup. In this experiment we classified each of these six genres, using different setups for the separation of training and test set in order to control for the confounding variables. For the love novels this shows for example, that separating cleanly between the authors didn't reduce the performance, while doing the same with the series results in a drastic drop (figure 7), showing again, that in this genre, the genre cohesion is quite low, while the publishers and even the series have distinctive features.

Following up the indications for confounding variables we uncovered the complicated situation of genre in this subfield of the literary market. We succeeded to explore some of its substructures which haven't been described yet in literary studies, though it has been always one of its topics that this kind of literature is a commodity (Nusser 1973, Nusser 1991, Nutz 1999, Stockinger 2018). It is quite astonishing that almost every genre behaves differently, but this may be the result of a decades-old competition between this small number of publishers. Probably the different structures correspond to different strategies of each publisher. Bastei-Lübbe for example seems to follow a strategy where each series has a distinct profile, while Cora is focussing more on the publisher name as brand (Fig. 7 and Fig. 4) - though the clustering may also be influenced by the fact that Cora translates many novels from English. It would be an interesting follow-up-project, to find out, whether the readers of these genres know about these structures and how this knowledge directs their choices. Last but not least, we think that the strategies to control for known and unknown confounding variables in text classification, especially if it is done to understand existing structures and not so much to predict really new data, needs to be explored in more detail.

Acknowledgements

We like to thank Reviewer 2 for providing detailed and very informative feedback especially on the relation between data leakage and confounding variables as well as on the evaluation of dimension reduction techniques.

Fußnoten

1. Our code can be found: https://github.com/LeKonArD/info_leakage
2. For Multinomial Naive Bayes, Logistic Regression, SVM and K-NN we used the library Scikit-Learn (Pedregosa 2011). For the new gradient boosting approaches we used XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), CatBoost (Dorogush et al. 2017).

Bibliographie

Akiba, Takuya / Shotaro Sano / Toshihiko Yanase / Takeru Ohta / Masanori Koyama (2019): „Optuna: A Next-generation Hyperparameter Optimization Framework“. *CoRR* abs/1907.10902. http://arxiv.org/abs/1907.10902.

Bojanowski, Piotr / Edouard Grave / Armand Joulin / Tomas Mikolov (2016): „Enriching Word Vectors with Subword Information“. *CoRR* abs/1607.04606. http://arxiv.org/abs/1607.04606.

Chen, Tianqi / Carlos Guestrin (2016): „XGBoost: A Scalable Tree Boosting System“. *CoRR* abs/1603.02754. http://arxiv.org/abs/1603.02754.

Dorogush, Anna Veronika / Andrey Gulin / Gleb Gusev / Nikita Kazeev / Liudmila Ostroumova Prokhorenkova / Aleksandr Vorobev (2017): „Fighting biases with dynamic boosting“. *CoRR* abs/1706.09516. http://arxiv.org/abs/1706.09516.

Eissen, Sven Meyer zu / Stein, Benno (2008): „Retrieval models for genre classification“. *Scandinavian Journal of Information Systems.*

Frank, Eibe / Remco R. Bouckaert (2006): „Naive Bayes for Text Classification with Unbalanced Classes“. In *Knowledge Discovery in Databases: PKDD 2006* , herausgegeben von Johannes Fürnkranz, Tobias Scheffer, und Myra Spiliopoulou, 503–510. Berlin, Heidelberg: Springer Berlin Heidelberg.

Goncales, T. / Quaresma, P. (2014): Evaluating preprocessing techniques in a Text Classification problem. In Information Processing & Management 50. Jg., Nr. 1, S. 104-112.

Hempfer, Klaus W. (2010): „Zum begrifflichen Status der Gattungsbegriffe: Von ‘Klassen’ zu ‘Familienähnlichkeiten’ und ‘Prototypen’.“ *Zeitschrift Für Französische Sprache Und Literatur* 120, 1: 14-32. http://www.jstor.org/stable/40619075.

Jannidis, Fotis, / Konle, Leonard / Leinen, Peter (2019a): Thematic Complexity. DH 2019 in Utrecht. Conference Abstracts.

Jannidis, Fotis / Konle, Leonard / Leinen, Peter (2019b): Makroanalytische Untersuchung von Heftromanen. DHd 2019. Conference Abstracts.

Kaufman, Shachar / Saharon Rosset / Claudia Perlich / Stitelman (2012): „Leakage in Data Mining: Formulation, Detection, and Avoidance“. *ACM Trans. Knowl. Discov. Data* 6 (4): 15:1–15:21. https://doi.org/10.1145/2382577.2382579.

Ke, Guolin / Qi Meng / Thomas Finley / Taifeng Wang / Wei Chen / Weidong Ma / Qiwei Ye / Tie-Yan Liu (2017): „LightGBM: A Highly Efficient Gradient Boosting Decision Tree“. In *Advances in Neural Information Processing Systems 30* , herausgegeben von I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, und R. Garnett, 3146–3154. Curran Associates, Inc. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

Landeiro, V. / Culotta, A. (2016): „Robust Text Classification in the Presence of Confounding Bias“. AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 186–193.

Nusser, Peter (1973): *Romane für die Unterschicht. Groschenhefte und ihre Leser.* Stuttgart: Metzler.

Nusser, Peter (1991): *Trivialliteratur.* Stuttgart: Metzler.

Nutz, Walter (1999): *Trivialliteratur und Populärkultur.* Opladen: Wiesbaden.

Olson, Randal S. / William La Cava / Zairah Mustahsan / Akshay Varik / Jason H. Moore (2017): „Data-driven Advice for Applying Machine Learning to Bioinformatics Problems“. *arXiv:1708.05070 [cs, q-bio, stat],* August. http://arxiv.org/abs/1708.05070.

Pedregosa, F. / G. Varoquaux / A. Gramfort / V. Michel / B. Thirion / O. Grisel / M. Blondel u. a. (2011): „Scikit-learn: Machine Learning in Python“. *Journal of Machine Learning Research* 12: 2825–2830.

Ribeiro, Marco Tulio/ Sameer Singh / Carlos Guestrin (2016): „‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier“. *arXiv:1602.04938 [cs, stat],* Februar. http://arxiv.org/abs/1602.04938.

Stockinger, Claudia (2018): „Das All dort draußen zeigt uns, wer wir sind. Die Leseuniversen der Groschenhefte“. In Steffen Martus / Carlos Spoerhase (ed.): *Gelesene Literatur: Populäre Literatur im Medienwandel.* Text und Kritik. edition text und kritik.

Szubert, Benjamin, et al. (2019): “Structure-Preserving Visualisation of High Dimensional Single-Cell Datasets.” *Scientific Reports,* vol. 9, no. 1, June, p. 8914, doi: 10.1038/s41598-019-45301-0.

Toman, M. / Tesar, R. / Jezek, K. (2006): “Influence in Word Normalization on Text Classification." Proceedings of InSciT 4 (2006): 354-358.

Critical Machine Vision. Eine Perspektive für die Digital Humanities

Bell, Peter
peter.bell@fau.de
FAU Erlangen-Nürnberg, Deutschland

Offert, Fabian
fabian@zentralwerkstatt.org
FAU Erlangen-Nürnberg, Deutschland

Wir fragen nach neuen Spielräumen der Digital Humanities im Feld des maschinellen Lernens. Dazu dekonstruieren wir etablierte Computer-Vision-Modelle mit Methoden der Bildwissenschaft/Visual Studies.

Computer Vision, also der visuelle Zweig künstlicher Intelligenz, spielt eine immer wichtigere Rolle in Wirtschaft (z.B. Industrie 4.0, autonomes Fahren), Sozialem (Überwachung, Medizin) und Wissenschaft (vorrangig in den Natur-, Ingenieur-, und Lebenswissenschaften). Auch wenn erste Prototypen in den Digital Humanities, zum Beispiel in der digitalen Kunstgeschichte (Bell / Impett, 2019) und in den A/V-orientierten Di-

gital Humanities (Arnold / Tilton, 2019a), entwickelt werden, wird die rasante Entwicklung des visuellen maschinellen Lernens in den Geisteswissenschaften noch relativ wenig reflektiert. Dies liegt teilweise daran, dass dessen Erforschung und kritische Reflektion eine fundierte Kenntnis der technischen Prozesse erfordert.

In besonderen Maße gilt dies für die Interpretierbarkeit von Computer-Vision-Modellen aus dem Bereich des Deep Learning, also des maschinellen Sehens mit komplexen neuronalen Netzwerken wie Convolutional Neural Networks (LeCun et. al. 1989, Krizhevsky et. al, 2012). Obwohl Interpretation als Eckpfeiler humanistischer Methoden und Theoriemodelle gilt, und obwohl Interpretable Machine Learning gegenwärtig in den Technikwissenschaften mit großer Aufmerksamkeit bedacht wird (Lipton, 2016, Selbst / Barocas, 2018, Mittelstadt et. al., 2019, Doshi-Velez und Kim, 2017, Gilpin et. al. 2018), ist das Problem der Interpretation, also der sinnschaffenden Analyse und Kritik von Computer-Vision-Modellen und Arbeitsabläufen weder in den Geisteswissenschaften noch in den Digital Humanities ausführlicher gewürdigt worden. Erste Ansätze finden sich z.B. in (Underwood, 2019) oder (Arnold / Tilton, 2019b). Dies ist insofern überraschend, als die Interpretation von Computer-Vision-Modellen eine Reihe von Fragen aufwirft, die Strukturähnlichkeiten zu Problemen in den Geisteswissenschaften im Allgemeinen, und in den Bildwissenschaften im Besonderen aufweisen. Dazu gehören das Problem der visuellen Mehrdeutigkeit, das epistemologische Problem der Verortung von Wissen, und das Problem des Verhältnisses von Form und Bedeutung.

Obwohl diese Aspekte sich im Bereich der Computer Vision als technische Probleme mit technischen Lösungsansätzen manifestieren (Olah et. al. 2017, 2018, Hohman et. al. 2018), bleibt ihre kritische Sprengkraft erhalten und erfordert eine nicht-technische Aufarbeitung. Beispielhaft ist hier die Andersartigkeit der maschinellen Wahrnehmung mit Convolutional Neural Networks zu nennen, die nachweisbar sehr viel mehr auf das Erkennen von Oberflächenbeschaffenheit aufbaut als auf das Erkennen von Formen (Geirhos et. al., 2019), und generell mit kaum wahrnehmbaren Bildbestandteilen operiert (Ilyas et. al., 2019). Wie beeinflusst diese andersartige Weltsicht die Aussagekraft von maschinellen Analysen in den Digital Humanities? Interpretierbarkeit könnte daher als ein grundsätzlich interdisziplinäres Problem angesehen werden, welches das Potenzial hat, Anstrengungen in der Informatik und den Digital Humanities zu verbinden und zu festigen.

Unter dem Begriff “Critical Machine Vision” möchten wir in den Digital Humanities daher einen Bereich etablieren, in dem die Digital Humanities nicht nur digitale Methoden auf geisteswissenschaftliche Gegenstände anwenden, sondern umgekehrt die informatischen Werkzeuge mit Methoden der Digital Humanities und der Geisteswissenschaften analysieren. Critical Machine Vision stellt drei zentrale Fragen: (1) Was und wie wird von mit Hilfe von Computer Vision gelernt, (2) welche Stereotypen und Vorurteile werden in diesem Lernprozess affirmiert oder erzeugt, und (3) wie können diese Verzerrungen durch neuartige Formen von Bilddatensätzen und Annotationsmethoden gemindert werden, und so Ansätze aus dem Forschungsbereich Fairness, Accountability, and Transparency of Machine Learning, kurz FAT-ML, (vgl. Friedler et. al., 2019, Suresh / Guttag, 2019) für Bilddatensätze neu gedacht werden. Wir befassen uns insbesondere mit der kritischen Analyse der wichtigen Bilddatensätze ImageNet (Deng et. al., 2009) bzw. der ILSVRC2012-Auswahl von ImageNet (Russakovsky et. al., 2015) und COCO (Lin et. al., 2014), mit denen Convolutional Neural Networks trainiert und evaluiert werden.

ImageNet ist ein umfangreicher digitaler Bilddatensatz, der die automatische Klassifizierung von Bildern in Bezug auf die abgebildeten Objekte ermöglichen soll (Object Recognition). Er besteht aus über vierzehn Millionen Bildern in über 21.000 Kategorien. Wir konzentrieren uns in unserer Analyse weniger auf aus unserer Sicht eher unkritische Klassifizierungen (z. B. Hunderassen oder Fahrzeugtypen), sondern auf streitbare Zuschreibungen: die Kennzeichnung der Menschen, ihre Assoziation mit sozialen Gruppen und menschlichen Interaktionen. Diesem kritischen Bereich von ImageNet entsprechen ähnlichen Kategorien in der Bilddatenbank COCO (Common Objects in Context), die mit ihrem Fokus auf "Common Objects" den Alltag und dementsprechend auch viele Menschen und menschliche Interaktionen einbezieht. Im Gegensatz zu ImageNet hat COCO weniger Kategorien, aber mehr Instanzen pro Kategorie. Auf diese Weise können detaillierte Objektmodelle erlernt werden, die eine präzise 2D-Lokalisierung ermöglichen. Am relevantesten für den vorgeschlagenen Beitrag ist jedoch die Tatsache, dass die alltäglichen Szenen und Objekte in COCO hauptsächlich aus westlichen, bürgerlichen Kontexten des 21. Jahrhunderts stammen, also nur einen begrenzten Ausschnitt von Welt bieten, der wiederum von einer ebenfalls nicht repräsentativen Gruppe von Menschen annotiert wurde.

Beide Bilddatensätze werden mit Methoden der Informatik und der Bildwissenschaft untersucht, aber eben ganz bewusst als Teil der Digital Humanities. Unser Beitrag liegt also nicht nur im neuartigen Ansatz der granularen, technisch fundierten, Dekonstruktion und konstruktiven Umgestaltung von digitalen Bilddatensätzen, sondern auch in der Transdisziplinarität unter dem Dach der Digital Humanities Computer Vision/Informatik und Bildwissenschaften zu verbinden. Wir verändern damit die Blickrichtung der Digital Humanities. Sahen wir bisher mit den Werkzeugen der Computer Vision auf geisteswissenschaftliche Gegenstände, schauen wir jetzt mit geisteswissenschaftlichen Werkzeugen auf die Methoden der Computer Vision. Dabei ist dieses Verhältnis allerdings mehrfach gebrochen, denn wir nutzen dabei wiederum digitale Werkzeuge wie z.B. Convolutional Neural Networks und Generative Adversarial Networks (Goodfellow et. al., 2014), oder Werkzeuge aus dem Bereich der Visual Analytics wie Summit (Hohman et. al., 2019) und schauen auf geistes- und sozialwissenschaftliche Gegenstände (z.B. Gender, Race, Habitus und Diskurs). Die Öffnung der Black Box ist somit ein Ergebnis der konsequenten gegenseitigen Ergänzung von geisteswissenschaftlich-kritischen Werkzeugen und der Nutzbarmachung experimenteller informatischer Werkzeuge aus dem Bereich des maschinellen Lernens.

Eine der großen Herausforderungen der Computer Vision ist die Vielfalt und Heterogenität der realen Bildwelt, die sich mit technischen Mitteln nur schwer erfassen lässt. Während sich Computer Vision in der Vergangenheit auf ausgefeilte algorithmische Ansätze zur Erkennung von Merkmalen in Bildern konzentrierte, gelang es der jetzigen Generation des maschinellen Lernens diese weit zu übertreffen, indem komplexe (d.h. „tiefe“) Convolutional Neural Networks verwendet wurden, die auf großen Bilddatensätzen trainiert wurden. Mit der Einführung solcher Datensätze in den Computer-Vision-Prozess entsteht jedoch ein für die Schnittstelle von Computer Vision und maschinellem Lernen spezifisches Problem: Wie lässt sich die Vielfalt und Heterogenität der realen visuellen Welt in einer Reihe von Bildern – begrenzter Größe – darstellen? Histo-

risch gesehen hängt dieses Problem mit dem allgemeinen erkenntnistheoretischen Problem zusammen, Taxonomien des Bestehenden zu erschaffen. Symbolische Taxonomien und Kategorien dienen hier der Ordnung konkreter Repräsentanten in Form von annotierten Fotografien. Unsere kritische Analyse setzt somit an sämtlichen Punkten des Prozesses an: die Ordnung der Taxonomien, die Auswahl und Art der Abbildungen, der Vorgang der Annotation bis hin zu den algorithmischen Details des Trainings.

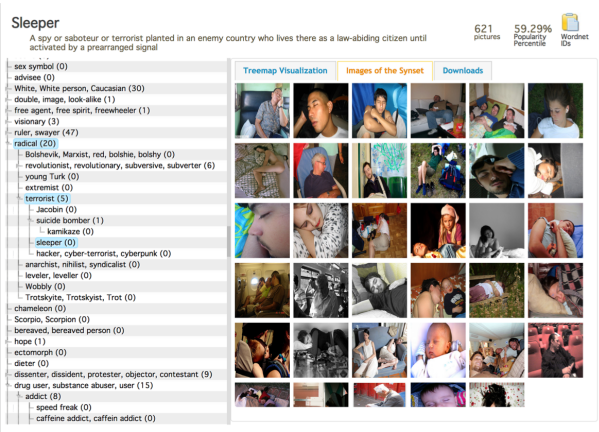


Abbildung 1: Die Kategorie "terrorist" in ImageNet enthält die Unterkategorie "sleeper" im Sinne verdeckter Terroristen (links). ImageNet illustriert den Begriff mit Bildern schlafender Menschen.

In jedem Schritt zeigen sich dabei Verzerrungen aufgrund von subjektiven Einschätzungen und (hegemonialen) Diskursen, die der demographischen Struktur der Akteur*innen (Fotograph*innen, Datenkurator*innen und Annotierenden) geschuldet sind. Diese Vorurteile lassen sich direkt an den Trainingsdaten, Kategorien und Strukturen ablesen, beispielsweise durch die von den Annotierenden erstellten Bildbeschreibungen (Captions) oder die vorgegebenen Objektklassen in COCO. Die in Teilen ungewollt komischen Bildbeschreibungen und US-amerikanisch geprägten Kategorien müssen immer vor dem Hintergrund betrachtet werden, dass Sie zum Training, Validieren und Testen von KI-Systemen verwendet werden. Eine anhand derart belasteter Datensätze trainierte KI wird zu einer Gefahr in jenem Moment, in dem die spezifische, eingeschränkte, und vorurteilsbelastete "Weltsicht" der KI auf Situationen der realen visuellen Welt trifft, und sich Mängel in den Datensätzen in undurchsichtige (da vielfach mediierte) und potenziell diskriminierende Fehlentscheidungen übersetzen.

Uns geht es jedoch nicht ausschließlich um eine Kritik der bestehenden Computer-Vision-Methoden, sondern um die Entwicklung und Erprobung neuer Verfahren in denen existierende Vorurteile durch bildwissenschaftliche Beschreibungs- und Ordnungsmodelle reduziert werden. Dabei stellt sich auch die Frage, wie sich eine größere Diversität der Bilddaten und letztlich des künstlichen Sehens nicht nur über eine räumlich größere Diversität, sondern auch eine zeitliche Diversität erreichen lässt. Zu welchem Maß spielen historische Bildwelten, das kulturelle Erbe, eine Rolle für unsere gegenwärtige Wahrnehmung, in welchem Maß muss sie Berücksichtigung finden?

Welche Veränderungen ergeben sich durch die Annotation von Expert*innen oder eine bessere Vorbereitung der Crowdworker? Die Analyse und die methodischen Ansätze zur Veränderung von Modellen und Prozessen zeigen wir anhand von wenigen Fallbeispielen (wie Abb. 1).

Unser Beitrag untersucht diese Fragen mit den kombinierten Mitteln der Computer Vision und der Bildwissenschaft, mit dem Ziel, diesen interdisziplinären Ansatz als neue Forschungsrichtung innerhalb der Digital Humanities vorzuschlagen, damit die Digital Humanities als kritischen Partner der Informatik neu zu etablieren, und ihre Spielräume somit signifikant zu erweitern.

Bibliographie

Arnold, T. / Tilton, L. (2019a): Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities.

Arnold, T. / Tilton, L. (2019b): Depth in Deep Learning: Knowledgeable, Layered, and Impenetrable.

Bell, P. / Impett, L. (2019): Ikonographie und Interaktion. Computergestützte Analyse von Posen in Bildern der Heilsgeschichte. Das Mittelalter 24, 31–53.

Deng, J. / Dong, W. / Socher, R. / Li, L. / Li, K. / Fei-Fei, L. (2009): Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference, 248–255.

Doshi-Velez, F. / Kim, B. (2017): Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Friedler, S.A / Scheidegger, C. / Venkatasubramanian, S. / Choudhary, S., Hamilton, E.P. / Roth, D. (2019): A comparative study of fairness-enhancing interventions in machine learning, in: ACM Conference on Fairness, Accountability, and Transparency (FAT*).

Geirhos, R. / Rubisch, P. / Michaelis, C. / Bethge, M. / Wichmann, F.A. / Brendel, W. (2019): ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.

Gilpin, L.H. / Bau, D. / Yuan, B.Z. / Bajwa, A. / Specter, M. / Kagal, L. (2018): Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE: 80–89.

Goodfellow, I. / Pouget-Abadie, J. / Mirza, M. / Xu, B. / Warde-Farley, D. / Ozair, S. / Courville, A. / Bengio, Y. (2014): Generative adversarial nets, in: Advances in Neural Information Processing Systems: 2672–2680.

Hohman, F.M. / Kahng, M. / Pienta, R. / Chau, D.H. (2018): Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. IEEE Transactions on Visualization and Computer Graphics.

Hohman, F. / Park, H. / Robinson, C. / Chau, D.H. (2019): Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. arXiv preprint arXiv:1904.02323.

Ilyas, A. / Santurkar, S. / Tsipras, D. / Engstrom, L. / Tran, B. / Madry, A. (2019): Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175.

Krizhevsky, A. / Sutskever, I. / Hinton, G.E. (2012): Imagenet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems: 1097–1105.

LeCun, Y. / Boser, B. / Denker, J.S. / Henderson, D. / Howard, R.E. / Hubbard, W. / Jackel, L.D. (1989): Backpro-

pagation applied to handwritten zip code recognition. Neural computation 1: 541–551.

Lin, T.-Y. / Maire, M. / Belongie, S. / Hays, J., Perona, P. / Ramanan, D. / Dollár, P. / Zitnick, C.L. (2014): Microsoft COCO: Common objects in context, in: European Conference on Computer Vision. Springer: 740–755.

Lipton, Z.C. (2016): The mythos of model interpretability, in: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

Mittelstadt, B. / Russel, C. / Wachter, S. (2019): Explaining Explanations in AI, in: ACM Conference on Fairness, Accountability, and Transparency (FAT*).

Olah, C. / Mordvintsev, A. / Schubert, L. (2017): Feature visualization. Distill. https://doi.org/10.23915/distill.00007 [letzter Zugriff 26 September 2019]

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A. (2018). The building blocks of interpretability. Distill. https://doi.org/10.23915/distill.00010 [letzter Zugriff 26 September 2019]

Russakovsky, O. / Deng, J. / Su, H. / Krause, J. / Satheesh, S. / Ma, S. / Huang, Z. / Karpathy, A. / Khosla, A. / Bernstein, M. et al. (2015): Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115: 211–252.

Selbst, A.D. / Barocas, S. (2018): The intuitive appeal of explainable machines. Fordham Law Review 87.

Suresh, H. / Guttag, J.V. (2019): A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002.

Underwood, T. (2019): Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Das Werk bildender Künstler*innen im Kontext – Digitale Werkverzeichnisse im semantischen Netz

Effinger, Maria

effinger@ub.uni-heidelberg.de
Universitätsbibliothek Heidelberg, Deutschland

Sobriel, Nicole

sobriel@ub.uni-heidelberg.de
Universitätsbibliothek Heidelberg, Deutschland

Ein Werkverzeichnis – auch Catalogue raisonné oder Œuvrekatalog genannt – hat den Anspruch, alle Werke eines bildenden Künstlers oder einer bildenden Künstlerin aufzulisten und zu beschreiben, auch wenn die Werke nicht mehr erhalten oder verschollen sind und ihre Existenz nur indirekt (z.B. durch Schriftquellen) nachweisbar ist. In der wissenschaftlichen Literatur oder Auktionskatalogen wird ein Werkverzeichnis in der Regel mit dem Autor*innennamen und der betreffenden Nummer des Werkes, der sogenannten „Werkverzeichnisnummer“ zitiert.

Œuvrekataloge in gedruckter Form haben eine sehr lange Tradition, bedeutet doch das Vorhandensein für den oder die Künstler*in gleichsam den „Ritterschlag“. Zugleich bezeugt es das Expertentum des Verfassers oder der Verfasserin über das erforschte Werk und dient überdies für den Kunstmarkt als verbindliches Nachweisinstrument.

Werkverzeichnisse sind genuin auf stetige Fortschreibung hin angelegt, damit kontinuierlich notwendige Ergänzungen zum jeweiligen Forschungsstand (wie Nachträge und Revisionen) möglich sind. Demzufolge ist eine zukünftig wachsende Bedeutung digitaler Werkverzeichnisse zu prognostizieren (Roettgen 2019, S. 376). Welchen Stellenwert die Erarbeitung von Œuvrekatalogen nach wie vor hat, zeigte die große Resonanz, auf den Aufruf einen „Arbeitskreis Werkverzeichnis“ (https://arbeitskreis-werkverzeichnis.de) ins Leben zu rufen. Bei der Gründungsveranstaltung in der Hamburger Kunsthalle am 03.11.2018 diskutierten rund 120 Teilnehmer und Teilnehmerinnen (https://arthist.net/archive/19049) über die vielfältigen Aspekte bei der Erstellung von Werkverzeichnissen. Die Chancen und Herausforderungen digitaler Angebote sind dabei wesentlicher Bestandteil der jährlich stattfindenden Workshops (und bildeten in der Novembertagung 2019 sogar den Schwerpunkt, vgl. https://www.belvedere.at/arbeitskreis-werkverzeichnis-3-tagung).

Im Rahmen des DFG-Programms „Fachinformationsdienste für die Wissenschaft“ (FID) entwickeln die Universitätsbibliothek Heidelberg und die SLUB Dresden mit dem Fachportal arthistoricum.net ein maßgeschneidertes Angebot für die kunsthistorische Fachcommunity. Die UB Heidelberg zeichnet dabei u.a. für den Bereich des elektronischen Publizierens im Open Access verantwortlich. Ihr aktueller Fokus liegt hierbei auf kollaborativen, mit Linked-Data-Technologien realisierten, dynamischen Publikationsmöglichkeiten und deren Weiterentwicklung. Nach dem Prinzip von „Enhanced e-Books“ sollen wissenschaftliche Texte unter anderem mit Bildern oder Multimediadateien verknüpft sowie gemeinsam mit weiteren Forschungsdaten weltweit zur Verfügung gestellt werden.

Obgleich es für die Forschung nur förderlich ist, haben Onlineressourcen immer noch ein Autoritätsproblem in der kunsthistorischen Forschung: Während in naturwissenschaftlichen Plattformen Kollektiv- und Mikropublikationen zunehmen, gibt es für Kunsthistoriker*innen verhältnismäßig wenige nachhaltige und damit zitierfähige Optionen.

Vor dem oben skizzierten Hintergrund lag es nahe, dass die UB Heidelberg die Gattung „Digitales Werkverzeichnis“ als einen ihrer Arbeitsschwerpunkte gewählt hat, da sich die Datenhaltung in einem Repositorium oder einer virtuellen Forschungsumgebung, die zugleich als Präsentationsplattform dient, besonders anbietet. Im Kontext von arthistoricum.net werden aktuell in Heidelberg mit der Objekt- und Multimediadatenbank heidICON sowie der „Wissenschaftlichen Kommunikations-Infrastruktur“ WissKI dafür unterschiedliche Lösungsansätze erarbeitet und z.T. auch schon bereitgestellt (https://www.arthistoricum.net/themen/wvz/). Ziel ist der Aufbau von Angeboten zu den Werken bildender Künstler*innen, die gemäß Linked-Data-Prinzipien und Semantic-Web-Standards die 2014 erarbeiteten Prämissen einer Digitalen Kunstgeschichte umsetzt (vgl. „Zürcher Erklärung zur digitalen Kunstgeschichte“, http://www.digitale-kunstgeschichte.de/w/images/6/6b/ZuercherErklaerungzurdigitalenKunstgeschichte2014.pdf).

Der erste, eher pragmatische Lösungsansatz – die Veröffentlichung eines Werkverzeichnisses in heidICON – kommt immer dann zur Anwendung, wenn nur wenige Eckdaten zu den Werken erhoben werden und es primär um die nachhaltige Veröffentlichung dieser Inhalte geht, d.h. weniger um darüber hinausgehende vertiefende Analysen wie z.B. Beziehungen der Werke untereinander oder zu Künstlernetzwerken. Auch stellen knappe Personalressourcen häufig einen limitierenden Faktor dar.

Die objektorientierte Datenbank heidICON (https://heidicon.ub.uni-heidelberg.de) ist das zentrale, nachhaltige Repositorium für die Universität Heidelberg und dient dort als Erschließungssystem sowie Präsentationsplattform für zahlreiche Institute, Sammlungen und Projekte der Universität Heidelberg. heidICON basiert auf der Software easydb der Firma programmfabrik (Berlin).[1] In dem durch das gemeinsam von Universitätsbibliothek und Universitätsrechenzentrum betriebenen Kompetenzzentrum Forschungsdaten (KFD) entwickelten OAIS-kompatiblen Langzeitarchivsystem heiARCHIVE werden alle in heidICON erfassten Daten in die Langzeitarchivierung überführt.

Hinsichtlich Technik und Erschließung auf internationale Standards setzend, hat sich heidICON zu einem unverzichtbaren Infrastrukturangebot entwickelt und leistet damit einen wichtigen Beitrag für die Etablierung der Digital Humanities an der Universität. Darüber hinaus wird heidiCON im Kontext des nationalen Auftrags der UB im Rahmen Fachinformationsdienste (FID) auch für die nachhaltige Bereitstellung von Bildsammlungen für die Fachcommunity der FIDs arthistoricum.net (Kunstgeschichte), Propylaeum (Altertumswissenschaften) und CrossAsia (Asienwissenschaften) bereitgestellt.

Besondere Aufmerksamkeit erfährt dieses Angebot aktuell bei der Erstellung digitaler Werkverzeichnisse. Naturgemäß stehen die Werke mit ihren digitalen Reproduktionen im Zentrum. So gilt es, umfangreiches, digital vorliegendes Abbildungsmaterial, Dokumentation von Restaurierungsmaßnahmen oder historische Quellen zu integrieren. Die in Heidelberg gewählte Lösung sieht vor, neben 2D- und 3D-Bildern auch vorhandenes Multimediamaterial nachhaltig zu speichern. Die Datenmaske zur Erfassung der Objekte ist nicht explizit auf Kunstwerke zugeschnitten, hat jedoch einen Schwerpunkt auf Gütern des kulturellen Erbes. Das eingesetzte Datenmodell bzw. die Erfassungskategorien der zu beschreibenden Werke sind am XML-Harvesting Schema LIDO (Lightweight Describing Objects, http://network.icom.museum/cidoc/working-groups/lido/what-is-lido) angelehnt. Der ereignisbasierte Aufbau deckt alle wesentlichen Aspekte für die Darstellung einer Objektgeschichte und den Anforderungen eines Werkverzeichnisses ab. Die Homogenisierung der Erschließungsdaten durch die Orientierung am LIDO-Schema erleichtert die spätere Nachnutzung der Daten, wenn die Datensätze exportiert und an Portale wie z.B. an die Deutsche digitale Bibliothek (DDB) geliefert werden.

Als Linked Data wurde kontrolliertes Vokabular und Normdaten, wie die Gemeinsame Normdatei (GND), GeoNames und der iDAI Gazzetteer implementiert, wodurch erste Schritte semantischer Verknüpfungen über die Datenwerte möglich sind. Dank dieser Referenzen zu Orten und Institutionen (Aufbewahrungs- und Standort, Entstehungsort oder dargestellter Ort), Personen (Künstler*in, ehemalige Eigentümer*innen [Provenienz], Darstellte oder Auftraggeber*innen), Klassifikation (Gattungen), Materialien, Technik und Werktitel ist eine eindeutige Identifizierung möglich (und soll zukünftig mit der Einbindung der Getty-Thesauri weiter ausgebaut werden).

Die einzelnen Datensätze sind zitierfähig durch persistente URIs oder die Vergabe von DOIs (Digital Object Identifiers). DOIs und die Einspeisung der Dateien in die Langzeitarchivierungssysteme der Universität Heidelberg garantieren eine nachhaltige Verfügbarkeit. Außerdem fördert die Ausgabe der Bilder und Metadaten via IIIF, die Referenzierbarkeit sowie die Bereitstellung der Inhalte als Linked Data im Semantic Web.

Neben dem Download der Multimediadateien ist der Export der Metadaten in CSV, XML oder JSON-Formaten möglich oder die Erstellung von Powerpoint-Präsentationen und deren Download. Sämtliche Funktionen in heidICON lassen sich über eine Programmierschnittstelle (API) steuern, was eine bessere Vernetzung mit bestehenden Systemen für die Datenpräsentation und die Archivierung zur Folge hat.

Die in heidICON erstellten Werkkataloge – wie aktuell beispielsweise das Online-Werkverzeichnis des Künstlers Georg Jakob Best (1903 – 2003, https://www.arthistoricum.net/themen/wvz/best/) oder des Bildhauers Bernhard Vogler (*1930, https://www.arthistoricum.net/themen/wvz/vogler/) – werden in arthistoricum.net in eine Themenseite über den Künstler oder die Künstlerin eingebettet, welche individuelle Angaben zu Leben und Werk, ggf. eine Bibliographie oder andere weiterführende Informationen beinhalten kann. Darüber hinaus können die digitalen Werkverzeichnisse aber auch optisch in institutionelle Webauftritte von Museen, Sammlungen oder anderen Kultureinrichtungen integriert werden. Hierfür werden eigens entwickelte Javascript-Module bereitgestellt, welche die Präsentation der Suche, verschiedene Browsingeinstiege sowie die Visualisierung der Trefferanzeigen außerhalb von heidICON im „Look and feel“ der das Werkverzeichnis verantwortenden Institution.

Der zweite Heidelberger Lösungsansatz zur Realisierung digitaler Werkverzeichnisse setzt für die Tiefenerschließung und Präsentation der Daten primär auf WissKI. Er kommt immer dann zur Anwendung, wenn komplexere Fragestellungen behandelt und vor allem auf der Grundlage semantischer Tiefenerschließung visualisiert werden sollen. Im Gegensatz zu heidICON ermöglicht die „Wissenschaftliche Forschungsinfrastruktur – WissKI“ (http://wiss-ki.eu/) eine individuellere und projektspezifischere Datenhaltung. Neben der Werkbeschreibung können auch andere Kategorien (Entitäten), zum Beispiel verknüpfte Personen oder Institutionen, ausführlicher beschrieben und vorhandene Kategorien untereinander verlinkt werden.

Das Alleinstellungsmerkmal von WissKI im Vergleich zu anderen Dokumentationssystemen ist die Fähigkeit, durch die ontologiebasierte wissenschaftliche Tiefenerschließung auf der Basis des ISO-Standards CIDOC-CRM (http://www.cidoc-crm.org) die Komplexität der kunstgeschichtlichen Dokumentation mit all ihren vielfältigen Bezügen abzubilden und als Linked Open Data bereitzustellen. Durch eine ontologiebasierte Datenhaltung in einem Triple Store stehen Forschungsergebnisse weltweit zur Verknüpfung mit anderen Datenrepositorien bereit. Kontrollierte Vokabulare und Normdaten (GND, Getty-Thesauri etc.) werden eingebunden. Besonders im Fokus steht hierbei die Nutzung von Werktitelnormdaten der Gemeinsamen Normdatei (GND) (http://www.arthistoricum.net/netzwerke/graphik-vernetzt/werktitelnormdaten/). Ausgehend von im Fach anerkannten Referenzwerken für Graphik oder Werkverzeichnisse erstellt die UB Heidelberg in enger Abstimmung mit dem „Arbeitskreis Graphik vernetzt“ und der

Deutschen Nationalbibliothek (DNB), systematisch Werktitelnormdaten mit wissenschaftlich gesicherten Inhalten.

Aber auch der zweite Lösungsansatz setzt für die nachhaltige Archivierung der Bild- und Multimediadateien auf heidICON. Für die aufwendiger angelegten digitalen Werkverzeichnisse erfolgt ebenfalls die Basiserschließung wie bereits oben beschrieben in heidICON. Das dort erfasste Bildmaterial kann inklusive ausgewählter Erschließungsdaten über einen eigens programmierten Picker direkt in das WissKI-basiertes Werkverzeichnis eingebunden werden.

Diese webbasierte und kollaborative Arbeits- und Publikationsweise sowie die multiplen Verbindungen von Bild und Text schaffen gegenüber bisher üblichen Printpublikationen neue Möglichkeiten der Visualisierung und Verbreitung stets aktueller Forschungsergebnisse. So können die zwischen den in der Datenbank erfassten Artefakten oder Personen bestehenden komplexen, geographischen, überlieferungskontextuellen, sprachlichen, inhaltlichen, ikonographischen oder editorischen Bezüge komfortabel recherchiert, visualisiert und dynamisch ausgebaut werden.

Individuelle Umsetzungen erfolgen u.a. in Kooperation mit dem Institut für Kunstgeschichte Regensburg (Prof. Christoph Wagner) für den Schweizer Maler, Kunsttheoretiker und Bauhausmeister Johannes Itten (1888-1967), des Weiteren wird in Zusammenarbeit mit der Kunsthistorikerin Prof. Dr. Steffi Roettgen (München) das bereits 1999 gedruckte Werkverzeichnis Anton Raphael Mengs in ein digitales Werkverzeichnis überführt und durch neue Werke bzw. Entitäten (wie etwa Neuzuschreibungen, Abschreibungen und Kopien) komplementiert. Darüber hinaus wird in einer Kooperation mit dem Albrecht-Dürer-Haus in Nürnberg mit einem digitalen Werkverzeichnis des Œuvres Albrecht Dürers begonnen.

Um diese ersten Prototypen zu einer nachhaltigen und vor allem nachnutzbaren Forschungsinfrastruktur ausbauen zu können, wird aktuell in dem DFG-Projekt „Semantics4Art&Architecture“ (http://www.ub.uni-heidelberg.de/wir/projekt_semantics4art.html) gemeinsam mit dem Herder-Institut für Ostmitteleuropaforschung in Marburg, eine auf CIDOC-CRM beruhende nachnutzbare Basis-Ontologie für digitale Werkverzeichnisse mit WissKI entwickelt und bereitgestellt. Die Erstellung anpassbarer Templates für Datenmodelle hilft einerseits der Homogenisierung der WissKI-Werkverzeichnisse, andererseits dient ein Datenmodell „von der Stange“ der Zeitersparnis und reduziert die benötigten informationstechnologischen Kenntnisse.

Bei der Konzeption digitaler Werkverzeichnisse muss zudem berücksichtigt werden, dass diese nicht nur der kunstwissenschaftlichen Dokumentation und Forschung dienen, sondern auch in anderen Bereichen, wie z.B. der Provenienzforschung, dem Kunsthandel oder aber auch dem Ausstellungswesen im Fokus des Interesses stehen. Die interdisziplinären Anforderungen für ein Werkverzeichnis werden deshalb in Workshops erarbeitet und in Datenmodell-Templates überführt, welche der Nachnutzung durch die Fachcommunity dienen. Wenngleich es das Ziel sein muss, ein sehr umfassendes Mustertemplate zu erstellen, um den unterschiedlichen Bedürfnisse verschiedener Gattungen Rechnung zu tragen, aus dem sich die Anwender – wie aus einem Bausteinkasten – bedienen können, ist die Modellierung und die Wahl der CRM-Klassen in manchen Fällen passender als in anderen.

Trotz der vielen Vorteile gibt es offene Fragen und Probleme, deren Klärung wichtig ist, wie zum Beispiel die Frage nach der Versionierung: Ab wann liegt eine neue Version vor und wie kann und sollte die technische Umsetzung erfolgen, wenn im besten Fall kenntlich gemacht werden soll, welche neuen Informationen im Vergleich zur Vorgängerversion vorliegen?

Rechtliche Hindernisse kommen insbesondere bei der von der Provenienzforschung gewünschten lückenlosen Erfassung ehemaliger Besitzer*innen auf. Diese können nicht immer veröffentlicht werden, was zum einen eine Herausforderung an die Modellierung und das WissKI-System setzt, wenn bestimmte Informationen eines Werkdatensatzes nicht in allen Fällen oder für eine gewisse Zeitspanne nicht als linked open Data zur Verfügung gestellt werden können.

Eine andere, im ersten Moment vielleicht weithergeholte Frage ist, ob Hackerangriffe eine mögliche Gefahr darstellen können? Eine illegale Platzierung eines Werkes in das System, wäre für Kunstfälscher von großem Vorteil, schließlich ist die Nennung eines Kunstwerkes in einem Werkverzeichnis für den Kunstmarkt von entscheidender Bedeutung.

Eine Folge digitaler Werkverzeichnisse unter der linked Data-Prämisse könnte die Auflösung der Monopolstellung der Autor*innen sein, wenn zum Beispiel durch kooperatives Fortschreiben der Inhalte, durch die Verknüpfung von Normdaten oder die Einbindung der Forschungsergebnisse von z.B. Restaurator*innen oder Provenienzforscher*innen das Wissen unterschiedlicher Forscher*innen zusammenfließt und die Grenzen der individuellen Autorenschaft verschwimmen.

Und welche Rollen und möglichen Probleme hat der Hoster – in unserem Fall also die Bibliothek – auszufüllen und zu berücksichtigen? Wenngleich sie nicht für den Inhalt zuständig ist, muss sie Lösungen finden, wenn die ursprünglich Hauptverantwortlichen nicht mehr zur Verfügung stehen, die Idee des dynamisch fortgeführten Werkverzeichnisses jedoch Bestand haben soll?

Das Heidelberger modulare Angebot für digitale Werkverzeichnisse soll den unterschiedlichen fachlichen Bedarfen hinsichtlich der geschilderten Komplexität und Erschließungstiefe sowie verschiedenartigen Nutzungsmöglichkeiten Rechnung tragen. Ebenfalls Berücksichtigung finden müssen dabei die hierfür jeweils zur Verfügung stehenden personellen und finanziellen Ressourcen. Stets gewahrt werden soll jedoch immer die Forderung, die digitalen Forschungsergebnisse nachhaltig, zitierfähig und interoperabel im Open Access zugänglich zu machen.

Fußnoten

1. Im Rahmen der Anfang 2019 erfolgtem Migration auf die neue Softwareversion easydb 5 wurde heidICON von der bisherigen bildorientierten Erfassung auf eine hierarchische und objektorientierte Datenhaltung auf der Basis des internationalen Standard LIDO (Lightweight Information Describing Objects) umgestellt.

Der Spielraum zwischen „zu wenig“ und „zu viel“

Du, Keli
keli.du@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Ausgangspunkt

Als eine quantitative textanalytische Methode wurde Topic Modeling[1] in den letzten Jahren in Digital Humanities häufig eingesetzt, um zahlreiche unstrukturierte Textdaten zu explorieren. Wenn Topic Modeling verwendet wird, muss man zuerst selbst entscheiden, wie viel Topics trainiert werden sollen. Es ist zwar bekannt, dass die Topic-Anzahl erhebliche Einfluss auf das Topic-Modell hat. Aber es ist nicht so ganz klar, wie groß der Unterschied zwischen den zwei Topic-Modellen ist, wenn man diese zwei Topic-Modelle mit unterschiedlicher Topic-Anzahl auf demselben Korpus trainiert.

In (Wallach et al., 2009) wurde Perplexität als interne Evaluationsmaß des Topic-Modells vorgeschlagen. Ein Topic-Modell wird als ein statistisches Sprachmodell betrachtet. Je niedriger die Perplexitätswerte ist, ist das Modell besser. In (Murphy, 2012, S. 954-955) wurde vorgestellt, dass die Perplexität von LDA-Topic-Modell mit der Erhöhung von Topic-Anzahl reduziert (Abbildung 1). In (Jurafsky & Martin, 2009, S. 43) wurde aber betont, dass die Korrelation zwischen Perplexität und Leistungsfähigkeit des Modells keine Kausalität ist. Deshalb kann eine interne Verbesserung in Perplexität nicht garantieren, dass das Modell bei den externen Aufgaben auf jeden Fall besser funktionieren kann. Eine End-to-End Evaluation (z. B. Dokument-Klassifikation) ist immer notwendig.

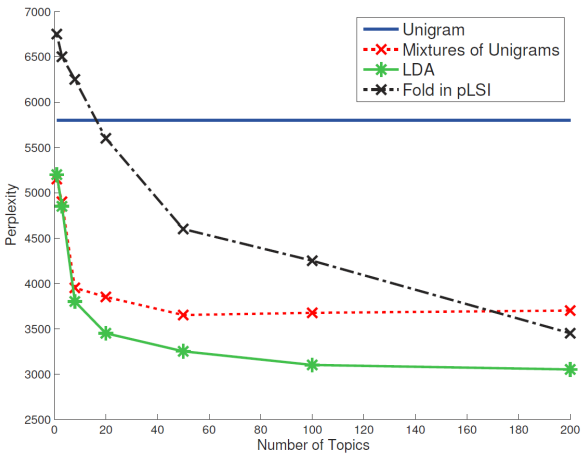


Abbildung 1: Perplexität vs. Topic-Anzahl auf TREC-AP-Korpus[2] (Murphy, 2012, S. 955)

Außerdem, wenn Topic Modeling für Forschung in Digital Humanities eingesetzt wird, interagieren die Benutzer normalerweise direkt mit Topics. Deshalb sind die standardmäßigen internen Evaluationsmethoden[3] für die Evaluation des Topic-Modells nicht ausreichend, weil sie die Qualität bzw. die Interpretierbarkeit der Topics nicht wiederspiegeln können. Über den Einfluss der Topic-Anzahl auf die Interpretierbarkeit der Topics wurde zum Beispiel von Matthew Jockers erklärt, wenn ein Topic-Modell zu viel Topics enthält, könnten die Topics ungenügend semantisch verwandte Wörter enthalten, um sinnvolle und interpretierbare Kontexte/Themen zu bilden. Im Gegensatz dazu, könnte ein Topic-Modell mit zu wenigen Topics dazu führen, dass die Topics zu allgemein sind und sie im ganzen Korpus vorkommen (Jockers, 2013, S. 128).

Aber was heißen eigentlich „zu wenig“ und „zu viel“? Um ein besseres Verständnis von dem Spielraum zwischen „zu wenig“ und „zu viel“ zu bekommen, wurden die vorliegende Untersuchungen durchgeführt. Diese Arbeit konzentriert sich nicht darauf, eine Methode zu finden, die die ideale Topic-Anzahl schätzen kann. Diese Arbeit möchte auch nicht, die Leistungsfähigkeit des Topic-Modells zu evaluieren. Das Ziel der Untersuchung in dieser Arbeit ist den Einfluss der Topic-Anzahl auf das Topic-Modell aus zwei Perspektiven zu verstehen: Topic Modeling basierte Dokument-Klassifikation und Topic-Kohärenz.

Korpus und Tools

Das Korpus der Untersuchung besteht aus 2000 deutschen Zeitungsartikeln zwischen 2001 und 2014[4]. Sie teilen sich in 10 thematische Klassen: „Digital“, „Gesellschaft“, „Karriere“, „Kultur“, „Lebensart“, „Politik“, „Reisen“, „Sport“, „Studium“ und „Wirtschaft“. Jede Klasse enthält 200 Dokumente. Das Korpus enthält insgesamt über 3,4 Millionen Tokens und die durchschnittliche Dokumentlänge ist ca. 1700. Alle Dokumente sind lemmatisiert. Abbildung 2 stellt die Verteilung der Dokumentlänge dar. Die meisten Dokumente enthalten 1400 bis 2000 Lemmata.

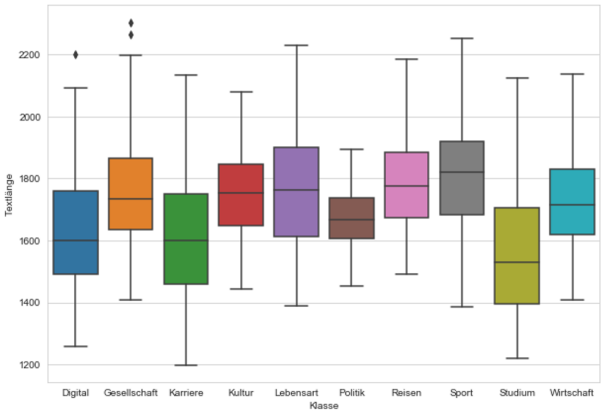


Abbildung 2: Verteilung der Dokumentlänge

Die Topic-Modelle wurden durch MALLET (McCallum, 2002) trainiert. Als Ergebnis bekommt man durch Topic Modeling eine Dokument-Topic-Verteilung und die Topics. In der Dokument-Topic-Verteilung wird jedes Dokument durch einen *n*-dimensionalen Vektor repräsentiert, während *n* die Topic-Anzahl des Topic-Modells ist. Aufgrund der Dokument-Topic-Verteilung wurde die Topic Modeling basierte Dokument-Klassifikation durchgeführt und die Klassifikation er-

folgte als 10-fache Kreuzvalidierung mit linearer SVM. Die Topic-Kohärenz wurde durch das Java-Programm Palmetto[5] automatisch berechnet und die erste 10 wichtigste Topic-Wörter wurden für die Berechnung genommen. Das Referenzkorpus für die Berechnung der Topic-Kohärenz ist die lemmatisierte deutschsprachige Wikipedia. In Palmetto wurden mehrere Topic-Kohärenz-Maße implementiert. Für diese Arbeit wurde das Normalised Pointwise Mutual Information (NPMI) basierte Kohärenz-Maß genommen, das in (Aletras & Stevenson, 2013) vorgeschlagen wurde.

Vor der Topic Modeling basierten Dokument-Klassifikation wurde Bag-of-Words (BoW) basierte Klassifikation zuerst durchgeführt, um eine Baseline der Klassifikation zu definieren. Die Tests erfolgten auch als 10-fache Kreuzvalidierung mit linearer SVM[6] , bei welchen der Accurary 0,765 und der F1(Makro)-Wert 0,758 betrug. Eine Baseline des NPMI-Wertes wurde auch definiert. Mit nur einer Iteration wurden zuerst 18 Topic-Modelle auf das Untersuchungskorpus trainiert, die jeweils 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500 Topics enthalten. Dadurch wurden 3150 „Topics ohne Topic Modeling“ erstellt und die NPMI-Werte dieser Topics wurden dann berechnet und der Durchschnittswert ist die NPMI-Baseline: -0,0619. Die Baseline wird durch eine schwarze Linie in den unteren Abbildungen dargestellt.

Die Untersuchungen

Das Ziel der folgenden Untersuchungen ist, den Einfluss der Topic-Anzahl zu überprüfen. Das Setting von Anzahl der Topics war **T** = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500. Für alle anderen Parameter-Einstellungen wurden die vorgegebenen Werte von MALLET genommen. Standard-Stoppwörter wurden vom Korpus entfernt. Aus technischen Gründen, nämlich die zufällige Initialisierung bei der Zuweisung von Topics und Gibbs Sampling, sind zwei Topic-Modelle von einem Korpus nicht völlig identisch, auch wenn das Setting beim Training gleich eingestellt ist. Deshalb können die Ergebnisse der Dokument-Klassifikation und die Topic-Kohärenz von den zwei Modellen unterschiedlich sein. Es wurden deshalb 10 Modelle für jedes Setting trainiert, um den Einfluss von der technischen Seite sichtbar zu machen.

Dokument-Klassifikation: Zuerst wurde der Einfluss von **T** auf die Dokument-Klassifikation mit LDA-Modell untersucht. Abbildung 3 stellt die Verteilungen der Klassifikationsergebnisse dar, die auf 10 Topic-Modelle von jeweiligen Settings basieren. Eine aufsteigende Tendenz ist deutlich erkennbar, wenn **T** von 10 auf 80 erhöht wurde. Eine signifikante weitere Verbesserung der Klassifikation kann in der Abbildung nicht mehr beobachtet werden, wenn **T** von 80 auf 500 erhöht wurde. Die meisten F1-Werte liegen zwischen 0,725 und 0,74. Am besten erzielte die Klassifikation das Accurarcy von 0,759 und den F1-Wert von 0,753. Eine Verbesserung gegenüber der Baseline konnte nicht festgestellt werden. Außerdem ist in der Abbildung zu beobachten, dass es größere Unterschiede unter den 10 Klassifikationsergebnissen gibt, wenn **T** = 10 ist. Diese große Abweichung zeigt, dass die zufällige Initialisierung und Gibbs Sampling eine größere Auswirkung auf das Training des Modells haben, wenn Topic-Modelle mit zu wenig Topics trainiert werden.

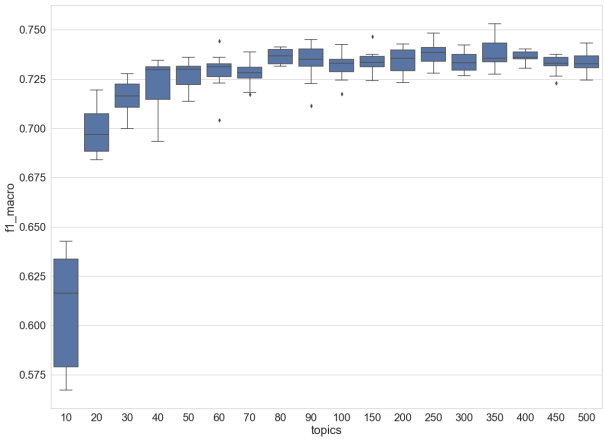


Abbildung 3: F1(Makro)-Werte der Topic Modeling basierten Dokument-Klassifikation im Verhältnis zu Anzahl der Topics

Topic-Kohärenz: Die zweite Untersuchung bezieht sich auf den Einfluss von **T** auf die Kohärenz der Topics. Die Verteilungsdichte der NPMI-Werte wird durch die Violin-Plots in der Abbildung 4 sichtbar dargestellt. Mit der Erhöhung von **T** geht der Median der NPMI-Werte (der weiße Punkt in die Mitte jedes Violin-Plots) unter. Der gesamte Wertbereich der NPMI-Werte ist außerdem breiter geworden, wenn **T** von 10 auf 60 steigt. Der Wertbereich der mittleren 50% der Daten geht mit der Erhöhung von **T** unter und ist hier besonders interessant. Der Bereich verbreitet sich zuerst, wenn **T** von 10 auf 100 steigt. Dann verengt der Bereich sich, wenn **T** von 150 auf 500 steigt.

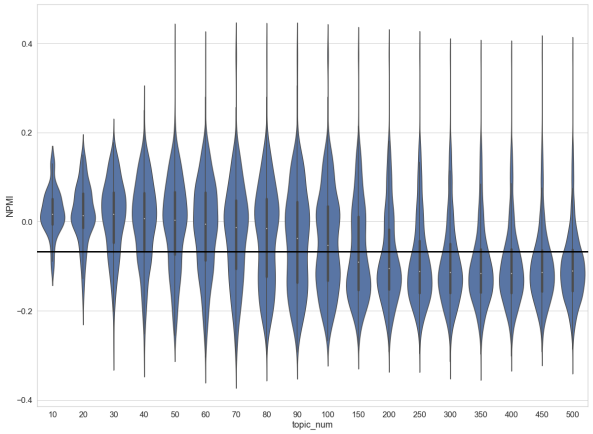


Abbildung 4: NPMI-Werte der Topics im Verhältnis zu Anzahl der Topics

Wenn man die NPMI-Werte der Topics mit der Baseline vergleicht, ist zu sehen, dass man mit der Erhöhung von **T** ständig durch Topic Modeling mehr Topics bekommen kann, deren NPMI-Wert größer als die Baseline ist (Abbildung 5, links). Aber wenn man diese absolute Anzahl normalisiert, also durch die gesamte Anzahl der Topics teilt, ist eine abnehmende Tendenz ganz deutlich erkennbar (Abbildung 5, rechts). Der Anteil von Topics, deren NPMI-Wert größer als die Baseline ist, sinkt von über 90% auf weniger als 30% ab. Das Ergebnis zeigt, dass man mit der Erhöhung von **T** durch Topic Modeling ständig viel mehr nicht kohärente Topics bekommen kann.

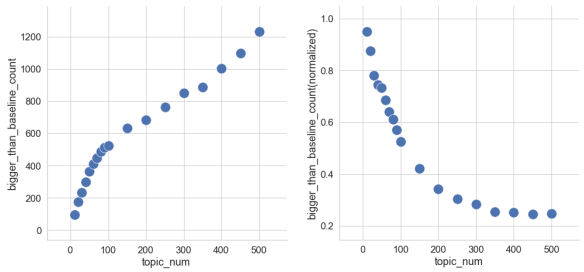


Abbildung 5: Topics, deren NPMI-Wert größer als die Baseline ist (links: absolute Anzahl; rechts: Prozentzahl)

In der Abbildung 4 werden von links nach rechts 10 * **T** , also 100 bis 5000 NPMI-Datenpunkte visualisiert. Um sicherzustellen, dass der Unterschied nicht auf eine ungleiche Anzahl von Datenpunkten zurückzuführen ist, wurde ein weiterer Test gemacht. Es wurden 500 Topic-Modelle à 10 Topics, 50 Topic-Modelle à 100 Topics und 10 Topic-Modelle à 500 Topics trainiert. Danach wurden die NPMI-Werte aller Topics berechnet und visualisiert. In der Abbildung 6 enthalten die drei Violin-Plots jeweils 5000 Datenpunkte. Hier wird eine ähnliche Verteilung wie in der Abbildung 4 beobachtet: Der gesamte Wertbereich verbreitet sich, der Median und der Wertbereich der mittleren 50% der Daten sinken, wenn **T** von 10 über 100 auf 500 erhöht wird.

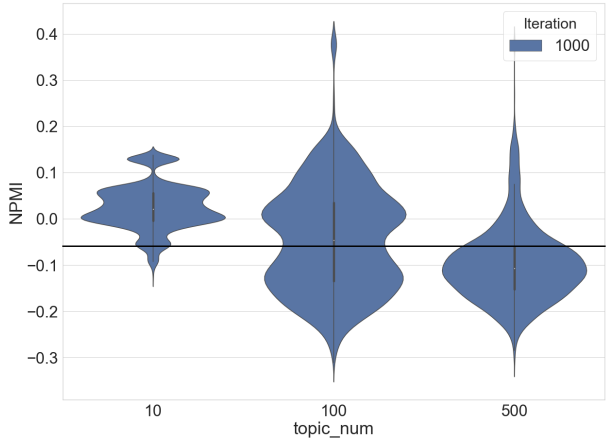


Abbildung 6: NPMI-Wert-Verteilungen von drei Topic-Modelle, die jeweils 5000 Topics enthält

Wenn man die Topics überprüft, sind 10 Topics für 2000 Zeitungsartikel sogar nicht „zu wenig“. In der Tabelle 1 sind die Topics aus einem Topic-Modell mit 10 Topics und sie sind keine allgemeinen Topics, die für Menschen nicht interpretierbar sind. Trotz bestimmter „Geräusche“ (wie z. B. „www“ und „geben“ im Topic 10), können 8 Topics zu den entsprechenden Klassen zugeordnet werden. Auch wenn Topic 6 und 7 zwei Topics sind, in denen vorwiegend Verben gruppiert werden, können sie wegen den Wörtern „kind“, „arbeiten“ und „haus“ mit der Klasse „Lebensart“ oder „Gesellschaft“ verbunden werden.

Nr.	Klasse	NPMI	Top 10 Wörter des Topics
1.	Sport	0,12959	spiel, fußball, spielen, spieler, trainer, tor, wm, fc, minute, verein
2.	Digital	0,06896	internet, facebook, datum, geben, netz, neu, google, online, nutzer, information
3.	Kultur	0.04433	welt, neu, schreiben, film, buch, musik, künstler, spielen, lassen, werk
4.	Wirtschaft	0.03412	euro, prozent, unternehmen, sagen, million, bank, deutsch, firma, neu, geben
5.	Politik	0.01903	sagen, deutsch, geben, krieg, regierung, russisch, neu, präsident, polizei, staat
6.	???	0.01083	sagen, geben, kind, online, wissen, mal, sehen, finden, einfach, arbeiten
7.	???	0.00799	sagen, stehen, haus, sehen, geben, mal, fahren, paar, sitzen, liegen
8.	Politik	0.00109	politisch, neu, politik, lassen, geben, kirche, stehen, politiker, spd, öffentlich
9.	Studium	-0.00441	universität, sagen, hochschule, uni, studium, schule, deutsch, geben, prozent, studieren
10.	Reisen	-0.06509	essen, lassen, restaurant, geben, www, hotel, tier, küche, kochen, liegen

Tabelle 1. 10 Beispieltopics aus einem 10-Topic Topic-Modell

In der Tabelle 2 sind drei Gruppen von Beispieltopics aus drei Topic-Modellen, die jeweils 10, 100, 500 Topics enthalten. Da die meisten Wörter in Topic 1a, 2a und 3a (auch 1b, 2b und 3b) gleich sind, kann festgestellt werden, dass die sinnvollen Topics mit der Erhöhung von **T** , statt sich in mehreren nicht kohärente Topics aufzulösen, kohärent bleiben können. Durch die Erhöhung von **T** werden eher weitere spezifische Topics produziert, wie z.B. „fußball, verein, fc, fan, stadion, bayern, bundesliga, spieler, trainer, liga“ oder „spielerin, birgit, frauenfußball, neid, tor, länderspiel, trabant, wm, sobiech, dfb“.[7]

Nr.	Topics-Anzahl (T)	NPMI	Top 10 Wörter des Topics
1a.	10	0,12959	spiel, fußball, spielen, spieler, trainer, tor, wm, fc, minute, verein
1b.	10	0,06896	internet, facebook, datum, geben, netz, neu, google, online, nutzer, information
2a.	100	0,10963	spieler, spielen, spiel, fußball, trainer, wm, ball, team, deutsch, tor
2b.	100	0,11812	facebook, internet, netz, nutzer, netzwerk, google, twitter, sozial, information, datum
3a.	500	0,10963	spielen, spieler, spiel, fußball, trainer, wm, ball, tor, deutsch, team
3b.	500	0.09401	facebook, netzwerk, internet, netz, nutzer, sozial, twitter, google, freund, information

Tabelle 2: Drei Gruppen von Beispieltopics aus drei Topic-Modellen

Fazit

Die vorliegenden Untersuchungen haben den Spielraum im Sinne von Topic-Anzahl bei Topic Modeling aus zwei Perspektiven eingegrenzt, nämlich Dokument-Klassifikation und Topic-Kohärenz. Angesichts der Untersuchung ist es festzustellen, dass man vermeiden sollte, ein Topic-Modell mit zu wenig Topics zu trainieren, wenn man eine bessere Topic Modeling basierte Dokument-Klassifikation sichern möchte. Ein To-

pic-Modell mit hoher Topic-Anzahl zu trainieren kann auch mehr kohärente Topics erzielen. Aber gleichzeitig muss man mit noch mehr nicht kohärenten Topics kämpfen. Deshalb ist es notwendig, die Topic-Kohärenz nach Topic Modeling zu berechnen, um die kohärenten und die nicht kohärenten Topics zu unterscheiden. Am Ende muss noch betont werden, dass das Ergebnis abhängig von Untersuchungskorpus sein könnte. Deshalb ist es geplant, die gleiche Untersuchung auf anderen Korpora (z. B. statt Zeitungsartikeln eine Sammlung von literarischen Texten für die Untersuchung zu nehmen) in der Zukunft durchzuführen.

Fußnoten

1. Topic Modeling ist eine Reihe von Algorithmen. Da das Latent Dirichlet allocation (LDA)-Modell am meisten verbreitetes Topic-Modell ist, bezieht „Topic Modeling" und „Topic-Modell" in dieser Arbeit sich nur auf LDA.
2. http://www.daviddlewis.com/resources/testcollections/trecap/
3. Hier beziehen die internen Evaluationsmethoden sich nicht nur auf die Perplexität, sondern auch auf die Methoden, die in (Deveaud, SanJuan, & Bellot, 2014); (Arun, Suresh, Veni Madhavan, & Narasimha Murthy, 2010); (Cao, Xia, Li, Zhang, & Tang, 2009) und (Griffiths & Steyvers, 2004) vorgeschlagen wurden
4. Das Korpus ist eine private Textsammlung, die leider nicht veröffentlicht werden kann. Mit MALLET wurde das Korpus importiert und in eine MALLET-Datei umwandelt, die hier verfügbar ist: https://www.dropbox.com/s/jpfhmtneu8q352z/Zeit_10_Klasse_lemma.mallet?dl=0
5. http://aksw.org/Projects/Palmetto.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
7. „birgit", „trabant" und „sobiech" sind drei Namen, die mit dem Thema Frauenfußball verbunden sind. Birgit Prinz und Anne Trabant sind zwei ehemalige deutsche Fußballspielerinnen. Gabriele Sobiech ist die Autorin vom Buch „ *Spielen Frauen ein anderes Spiel?:- Geschichte, Organisation, Repräsentationen und kulturelle Praxen im Frauenfußball*"

Bibliographie

Aletras, N. / Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22).

Arun, R. / Suresh, V., / Veni Madhavan, C. E. / Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Hrsg.), Advances in Knowledge Discovery and Data Mining (Bd. 6118, S. 391–402). https://doi.org/10.1007/978-3-642-13657-3_43

Cao, J. / Xia, T. / Li, J. / Zhang, Y. / Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Deveaud, R. / SanJuan, E. / Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique, 17(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Griffiths, T. L. / Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Jurafsky, D. / Martin, J. H. (2009): *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed* . Upper Saddle River, N.J., London: Pearson Prentice Hall (Prentice Hall series in artificial intelligence).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (13.12.2019).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective* . MIT press.

Wallach, H. M. / Murray, I. / Salakhutdinov, R. / Mimno, D. (2009, Juni): Evaluation methods for topic models. In: *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105-1112), ACM.

TREC-AP-Korpus: http://www.daviddlewis.com/resources/testcollections/trecap/ (14.12.2019)

DH's Next Top-Model? Digitale Editionsentwicklung zwischen Best Practice und Innovation am Beispiel des „Corpus Masoreticum"

Liedtke, Clemens
Clemens.Liedtke@hfjs.eu
Hochschule für Jüdische Studien Heidelberg, Deutschland

Einführung: Handschriftenforschung und Digitale Transformation

Wenn sich auch in den Geisteswissenschaften so etwas wie ein „Digital Turn" oder eine „Digitale Transformation" (Pousttchi 2017) beobachten lässt, wirft das unweigerlich die Frage auf, inwiefern die Einführung digitaler Prozesse in geisteswissenschaftliche Forschung Konsequenzen für den Aufbau bzw. das Design von konkreten Forschungsprojekten hat.

In folgendem Vortrag soll diese Frage an einem Fallbeispiel aus dem Bereich der Jüdischen Studien entwickelt werden. Hier wurde unlängst von Gerben Zaagsma herauspräpariert, dass durch die nunmehr in großem Umfang verfügbaren digitalen Ressourcen zu jüdischer Geschichte und Kultur sich

"die Sicherung, Bereitstellung und Analyse des in alle Welt verstreuten vielsprachigen und mehrschriftlichen Quellenmaterials“ als eine der zukünftigen Schlüsselaufgaben stellt (Zaagsma 2019:3ff.).

Dies lässt sich in besonderem Maße an dem Teilbereich der Handschriftenforschung exemplifizieren: Gerade durch die großen Digitalisierungsinitiativen der letzten Jahre an hebräischen Manuskripten, v.a. National Library of Israel, der Polonsky Foundation in Zusammenarbeit mit der British Library und der Bodleian Library stehen der wissenschaftlichen Community umfangreiche Quellenbestände zur Verfügung, die erst die Grundlage für weitere inhaltliche Tiefenerschließung, Edition und Corpusanalyse bilden.

Betrachtet man weiter das Teilgebiet der wissenschaftlichen Erforschung des hebräischen Bibeltextes, so lässt sich zeigen, dass forschungsgeschichtlich gesehen die Editionspraxis hebräischer Bibelhandschriften bereits teilweise auf digitale Prozesse zurückgreifen kann, etwa in computerlinguistischer Hinsicht durch die langjährigen Projekte des Eep Talstra Centre for Bible and Computer (ETCBC[1]; van Lit 2019) oder durch die elektronische Edition des Westminster Leningrad Codex (WLC[2]). Ob man dies bereits als Symptom eines digitalen Transformationsprozesses betrachten sollte, der jenseits der Einführung einzelner Tools und Verfahren übergreifende Veränderungen in Fragestellung, Methodologie und Methodenkritik sichtbar macht, ist damit noch nicht ausgemacht. Immerhin gilt nach wie vor die Print-Edition der "Biblia Hebraica Stuttgartensia" auf Grundlage der Handschrift Ms. Fircovitch B19a als eine der massgeblichen kritischen Textausgaben für Theologie und hebräische Bibeltextforschung und bildet damit noch den Stand und die Möglichkeiten der analogen Bibeltextkritik ab.

Hinsichtlich der Perspektive von primär digitalen wissenschaftlichen Editionen der Hebräischen Bibel stellt sich infolgedessen die Frage nach der Anwendbarkeit bereits etablierter Verfahren; während sowohl im WLC als auch im "Digital Mishnah Project"[3] XML-Textauszeichungen zum Einsatz kommen (entweder teilweise oder vollständig entlang der TEI P5 Spezifikationen implementiert), finden XML-basierte "best practices“ im Bereich linksläufiger Schriftsysteme wie dem Hebräischen bislang nur zögerlich Akzeptanz, zumal unter Einsatz von XML-Quelltext-Autorensystemen wie oXygen ganz basale handwerkliche Probleme das Schreiben von rechtsläufigen Tagsets und linksläufigen Schriften zur Herausforderung macht. So konstatiert auch noch das DARIAH Wiki: „Solange dieses Problem nicht grundsätzlich gelöst ist, wird die Akzeptanz von TEI und/oder XML in Hebraistik und Arabistik gering sein.“[4]. Gleichwohl berührt dies eher die Frage, inwiefern sich solche technischen Hürden durch geeignete grafische Benutzerschnittstellen nehmen lassen, die die systemischen Anforderungen bidirektionaler Unicode-Texte von anwendungsseitigen Annotationsebene wegabstrahieren.

Textcodierung: Modelle

Inhaltlich bedeutsamer scheint aber die mittlerweile ausführlich beschriebene Problematik zu sein, dass sich XML als semi-strukturierte, hierarchisch organisierte Markup-Sprache mit seinen strikten Regeln zur Wohlgeformtheit und Validität von Auszeichnungen nur bedingt dazu eignet, Phänomene zu annontieren, die nicht linear/hierarchisch, sondern mit Überlappungen oder sich überschneidenden Sequenzen strukturiert sind. Ebenso zwingt es die Bearbeitenden, die dokumentenzentrierte und die textzentrierte Perspektive einer zu edierenden Quelle durch zwei unterschiedliche Kodierungsstrategien zu lösen (Brüning/Henzel/Pravida 2013; Pierazzo 2017); gleichzeitig belastet die Verarbeitung von internen wie externen Verweisstrukturen im Dokument (Lesartvarianten, Zitate, Querbezüge) die Kodierung damit, die referentielle Integrität von Links zuverlässig verwalten zu können. Innovative Lösungsansätze werden für dieses Problem unter anderem entlang des Modells von Textvarianten-Graphen beschrieben (Schmidt 2008; Schmidt 2009; Schmidt/Colomb 2008:498) oder unter Verwendung von "Labelled Property Graph“-Systemen wie der Graphendatenbank Neo4J diskutiert (Kuczera 2016a, Kuczera 2016b).

An dieser beispielhaften Gegenüberstellung verschiedener Datenmodellierungsansätze einen Unterschied zwischen Standardverfahren/Best Practice und Innovation auszuloten, der bereits eine implizite Wertung von Innovation als „fortschrittlich“ mitmeint, griffe sicherlich zu kurz - gleichwohl spannt sich durch die in der Literatur diskutierten Anwendungsfälle durchaus ein Spannungsfeld auf: Einerseits belastet der Einsatz spezialisierter Datenbankmanagementsysteme[5] die Anforderungen an Offenheit und Langzeitverfügbarkeit von zu speichernden Forschungsdaten; auch die Neumodellierung von zu erhebenden Forschungsdaten schneidet im Sinne der FAIR-Prinzipien (Wilkinson / Dumontier / Aalbersberg, *et al.* 2016) zunächst von Anschlussmöglichkeiten ab, sind doch neuartige Datenmodelle, möglicherweise eben noch nicht „interoperable“ und „reusable“.

Andererseits bedeutet auch das Anwenden bestehender *best practices* eine interpretative Einschränkung: Mit der Umsetzung standardisierter Auszeichnungsschemata, sei es TEI-XML, eine bestimmte RDF-Ontologie oder Datenbankstruktur lässt sich am Quellenmaterial nur beobachten, was sich innerhalb der Unterscheidungsmöglichkeiten des jeweiligen Schemas bezeichnen lässt. Die Praxis der XML-basierten Quellenannotation zeigt hier, dass gerade bei steigenden Komplexitätsgraden am Material sich der Focus stärker in Richtung auf Einhaltung der Schema-Compliance und weg von der Beschreibung neuer Merkmalskategorien bewegt. Gute Indikatoren für dieses Phänomen sind beispielsweise vermehrter Einsatz von Standoff-Markup, individuelle, d.h. projektbezogene Schema-Erweiterungen, aber auch steigende Mehrdeutigkeiten im Markup bestimmter Phänomene wie etwa Marginalien in Handschriften (Estill 2016).

Dieses hier am Beispiel zweier Modellierungsstrategien angedeutete Spannungsverhältnis zwischen Standardisierung und Innovation lässt sich gerade im Rahmen des Projektdesigns produktiv nutzen, zwingt doch zum einen das Einführen digitaler Methoden in die Quellenerschliessung zu einer strengen Formalisierung des eigenen Forschungsprozesses, zum anderen gewinnt die Perspektive der Datenmodellierung (Owens 2011) eine größere Bedeutung. Beides hat nicht zuletzt auch entscheidenden Einfluss auf die Auswahl der zum Einsatz kommenden Technologie-Stacks.

Fallbeispiel: Corpus Masoreticum

Am folgenden Fallbeispiel soll entwickelt werden, wie die skizzierten Überlegungen in einem konkreten Projekt umgesetzt werden können: Das von der Deutschen Forschungsgemeinschaft geförderte Langzeit-Editionsvorhaben „Corpus Masoreticum“, das an der Hochschule für Jüdische Studien

Heidelberg angesiedelt ist, befasst sich mit dem sogenannten masoretischen Text in mittelalterlichen Bibelcodices. In der heutigen Forschung meint der Begriff der Masora alle metatextuellen Elemente zum Konsonantentext der Hebräischen Bibel. Dazu gehören Grapheme, grammatische, syntaktische und statistische Notizen, Referenzen und Verweise. Ab dem 12. Jh. entstehen im Kulturraum Aschkenas (Nord-Frankreich und Deutschland) hebräische Bibel-Kodizes, in denen die Masora mit mikrographischer Schrift in ornamentalen Formen auf der Seite platziert wurde und als Fabelwesen, vor allem aber als zoomorphe Gestalten (Hunde, Pferde, Hasen, Gazellen, Vögel, Fische) und sogar als anthropomorphe Darstellungen gestaltet werden - hierfür wurde der Begriff der Masora figurata zur Unterscheidung von linearer Masora magna geprägt. Sie kann darüber hinaus Zitate aus Kommentarliteratur enthalten, die weit über die üblichen quantitativen und referentiellen Annotationen zum hebräischen Konsonantentext hinausgingen (vgl. Ms Vat. ebr. 14). Als Beispiel für in diesen masoretischen Metatexten häufig enthaltenes Listenmaterial lassen sich die sogenannten „Okhla-Listen“ herausgreifen, in denen als bewahrenswert gedachte Textphänomene und Schreibungen in unterschiedlichen Strukturen und Layouts dem Bibeltext mitgegeben werden und ihrerseits auf verschiedene extern überlieferte Rezensionen dieser Listen referieren (als Überblick: Liss/Petzold 2016).

Bereits oberflächliche Untersuchungen an diesem sehr speziellen Quellenmaterial zeigen, dass hier besonders komplexe Anforderungen an das zu definierende Editionsdatenmodell gestellt werden: Zu dokumentieren ist nicht nur linearer Text, sondern hochgradig vernetzte interne und externe Verweisstrukturen nicht nur mit Bezug auf Lesartvarianten, sondern auch auf Kommentarliteratur und Listenmaterial mit spezifischen Listenmustern, die auf extern tradiertes Listenwissen verweisen. Darüber hinaus bedarf die doppelte Lesbarkeit von Masora figurata als Text und Bild gleichermaßen in ihrem Bezug zum Bibeltext eines besonderen Dokumentationsverfahrens.

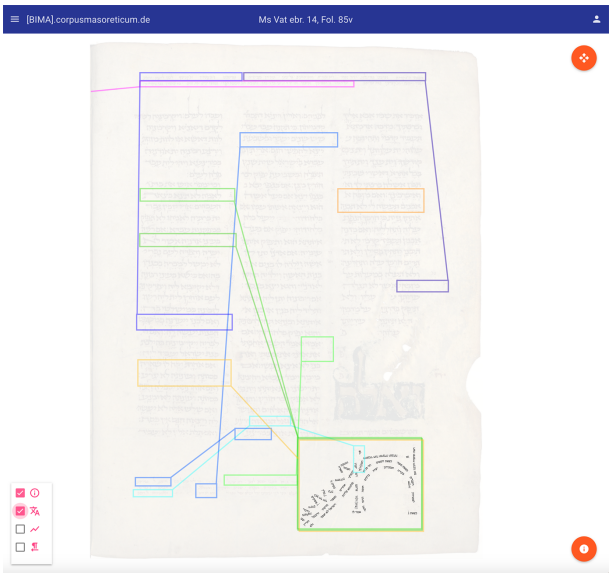


Abbildung 1: Überblick über das beispielhafte mise-en-page von Bibeltext, Masora parva, Masora Magna und Masora figurata in Ms Vat ebr. 14, Fol. 85v. Prototyp einer Visualisierungssoftware für digitale Erschliessung hebräischer Bibelcodices. Quelle: http://bima.corpusmasoreticum.de/figurata/tor

Implementierung von Modellen und Workflows

Durch die netzwerkartige Struktur der in die untersuchten Handschriften eingebetteten Metatexte lag es nahe, die Beschreibung von Text als Daten von vornherein als Graph zu modellieren; der „labelled-property“-Ansatz von Graphdatenbankensystemen wie Neo4J macht es durch seine sogenannte „whiteboard-friendliness“[6] möglich, einfache Modellskizzen rasch in lauffähige digitale Speichermodelle zu implementieren. Im Editionsworkflow wird zunächst der Import von Handschriftendigitalisaten samt Metadaten über IIIF-kompatible Archive realisiert und die Handschriftendaten im Importprozess in Graphendaten als Knoten und Kanten umgewandelt. Ab hier werden über eine grafische Benutzeroberfläche erzeugte Texttranskriptionen kontextbezogen als Datenknoten verlinkt, wobei der Bezug zum Digitalisat über die Kodierung von Text als SVG-Textpfaden erhalten bleibt. Transkriptionen werden bei Bedarf weiter tokenisiert, um weitere Metadaten oder Kontextrelationen in den Graphen integrieren zu können. Aus dem so generierten Text- bzw. Knowledge Graph lassen sich Subsets (Datenaggregate) generieren, die im späteren Prozess sowohl als TEI-XML, RDF-Graph oder auch als angereicherte IIIF-Manifeste (Text, Übersetzung, Kommentar) ausgeliefert werden können.

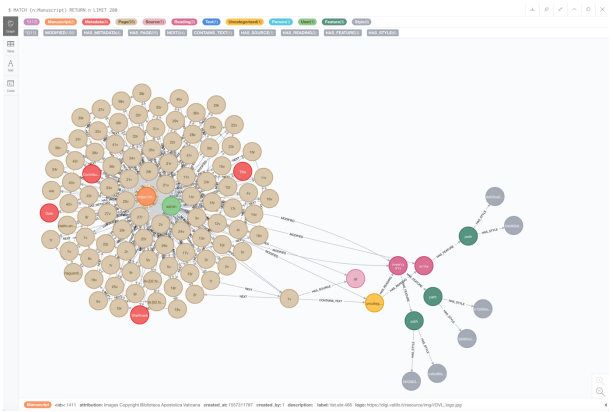


Abbildung 2: Beispielgraph anhand Cod. Vat. Ebr. 468, folio 1v, Darstellung in Neo4J. Quelle: Liedtke 2019 (in Vorbereitung)

Die technischen Komponenten sind stark modularisiert und, wo möglich, als Microservices implementiert, so dass die einzelnen Ressourcen der Edition mit REST-APIs ausgeliefert und abgefragt werden können. Die Softwarearchitektur ist als Docker-Container-Umgebung in einer skalierbaren Cloud-Computing Landschaft realisiert, die vom heiCloud-Service des Rechenzentrums der Universität Heidelberg bereitgestellt wird, wobei die Rahmenbedingungen von Langzeitarchivierung und Nachnutzbarkeit in Zusammenarbeit mit Fachdiensten der Universitätsbibliothek Heidelberg[7] gewährleistet werden.

Ausblick

Bezieht man das Spannungsverhältnis von Standardisierung und Innovation digitaler Prozesse in den architektonischen Aufbau eines geisteswissenschaftlichen Forschungsprojektes ein, lässt sich diese Dynamik produktiv für die Entwicklung eigener Workflows nutzen und öffnet Spielräume für das Modellieren der eigenen Forschungsdaten. Die formale Beschreibung eines digitalen Datenmodells kann dann methodenkritisch dazu verwendet werden, die im Prozess anstehenden Ergebnisse wieder an die Ausgangsfragestellung rückzubinden und Modelle auf ihre Plausibilität zu prüfen. Die Umsetzung in Technologie-Stacks oder digitale Frameworks führt im gezeigten Fallbeispiel zu der Konstruktion einer quasi „hybriden" Editionsumgebung und beschränkt den digitalen Anteil eines Projektes nicht nur auf das Ausliefern von „Tools", sondern betrachtet den Aspekt der digitalen Transformation als integralen Bestandteil des gesamten Forschungsprozesses.

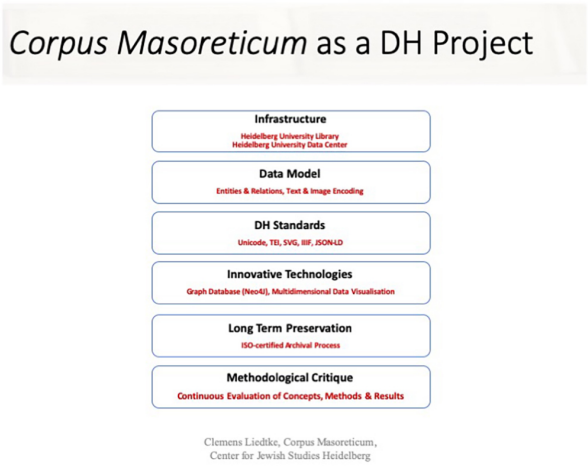


Abbildung 3: Corpus Masoreticum als DH-Projekt (Schema). Quelle: Liedtke 2019 (in Vorbereitung)

Fußnoten

1. http://etcbc.nl
2. http://tanach.us/
3. http://dev.digitalmishnah.umd.edu/
4. https://wiki.de.dariah.eu/display/publicde/3.5+Judaistik+und+Hebraistik
5. Die auch im nachfolgenden Fallbeispiel verwendete Community-Version von Neo4J ist unter einer GPLv.3 Lizenz als Open Source verfügbar, ist also im strengen Sinne keine proprietäre Software – die Entwicklung wird aber massgeblich von Neo4J Inc. als kommerziellem Unternehmen vorangetrieben.
6. https://neo4j.com/developer/guide-data-modeling/#whiteboard-friendly
7. https://heidata.uni-heidelberg.de/

Bibliographie

Estill, Laura (2016): "Encoding the Edge: Manuscript Marginalia and the TEI." *Digital Literary Studies* 1, no. 1. https://journals.psu.edu/dls/article/view/59715/59912.

Kuczera, Andreas (2016a): "Graphbasierte Digitale Editionen." Blog. *Mittelalter*. Interdisziplinäre Forschung Und Rezeptionsgeschichte (blog), April 19, 2016. https://mittelalter.hypotheses.org/7994.

Kuczera, Andreas (2016b): "Digital Editions beyond XML – Graph-Based Digital Editions.", in: *Proceedings of the 3rd HistoInformatics Workshop on Computational History* (HistoInformatics 2016), edited by Marten Düring, Adam Jatowt, Johannes Preiser-Kappeller, and Antal van Den Bosch. Krakow, 2016. http://ceur-ws.org/Vol-1632/paper_5.pdf.

Liedtke, Clemens (2019): 'How am I supposed to read this?' Challenges and Opportunities of Medieval Western Masorah as a Digital Scholarly Edition", in: J. Leipziger/H. Liss/K. J. Petzold (eds.), *Philology and Aesthetics:* Figurative Masorah in Western European Manuscripts (Judentum und Umwelt), Frankfurt am Main et al.: Peter Lang (in Vorbereitung)

Liss, Hanna / Petzold, Kay Joe (2016): "Die Erforschung der westeuropäischen Bibeltexttradition als Aufgabe der Jüdischen Studien. Ein halbes Jahrhundert Forschung und Lehre über das Judentum in Deutschland.", in: *Orchidee oder Mimose.* Versuch einer Standortbestimmung der Jüdischen Studien, edited by Andreas Lehnardt and Guiseppe Veltri. Berlin et al, 2016.

Owens, Trevor (2011): "Defining Data for Humanists: Text, Artifact, Information or Evidence?" *Journal of Digital Humanities* 1, no. 1. http://journalofdigitalhumanities.org/1-1/defining-data-forhumanists-by-trevor-owens/.

Pierazzo, Elena (2017): "Facsimile and Document-Centric Editing.", in: *Creating a Digital Scholarly Edition with the Text Encoding Initiative*, edited by Marjorie Burghart. https://www.digitalmanuscripts.eu/wp-content/uploads/sites/6/2017/09/05-Digital-Facsimiles-EP.pdf.

Pousttchi, Key (2017): "Digitale Transformation.", in: *Enzyklopädie der Wirtschaftsinformatik.* http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/technologien-methoden/Informatik--Grundlagen/digitalisierung/digitale-transformation/digitale-transformation/?searchterm=digitale%20transformation.

Schmidt, Desmond (2008): "What's a Multi-Version Document?" *Multi-Version Documents* (blog), May 3. http://multiversiondocs.blogspot.com/2008/03/whats-multi-version-document.html.

Schmidt, Desmond / Colomb, Robert (2009): "A Data Structure for Representing Multi-Version Texts Online." *International Journal of Human-Computer Studies* 67, no. 6 (June 2009): 497–514. https://doi.org/10.1016/j.ijhcs.2009.02.001.

Schmidt, Desmond (2009): "'Merging Multi-Version Texts: A Generic Solution to the Overlap Problem.' Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009," Proceedings of Balisage: The Markup Conference 2009, 2 (2009). https://doi.org/10.4242/BalisageVol3.Schmidt01.

Wilkinson, M. / Dumontier, M. / Aalbersberg, I. et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3 , 160018. doi:10.1038/sdata.2016.18

Zaagsma, Gerben (2018): "#DHJewish – Jewish Studies in the Digital Age." *Medaon.* Magazin für Jüdisches Leben in Forschung und Bildung, no. 12: 1–11.

Zundert, Joris J. van / Andrews, Tara L. (2016): "Apparatus vs. Graph: New Models and Interfaces for Text.", in: *Interface Critique*, edited by Florian Hadler and Joachim Haupt, 139:183–206. Berlin: Kulturverlag Kadmos.

Die Digitale Edition der Protokolle des Bayerischen Ministerrats – ein Erfahrungsbericht

Schrott, Maximilian

maximilian.schrott@ndb.badw-muenchen.de
Historische Kommission bei der Bayerischen Akademie der Wissenschaften, Deutschland

Reinert, Matthias

Matthias.Reinert@ndb.badw-muenchen.de
Historische Kommission bei der Bayerischen Akademie der Wissenschaften, Deutschland

Im Sommer 2019 konnte die Historische Kommission bei der Bayerischen Akademie der Wissenschaften ihr 2014 begonnenes Pilotprojekt für ein neues Konzept historisch-kritischer Editionsarbeit zu einem erfolgreichen Abschluss bringen. Mit dem neunten Band der Reihe „Protokolle des Bayerischen Ministerrats 1945–1962"[1], bearbeitet von Dr. Oliver Braun (München), entstand die erste Edition der Historischen Kommission auf XML-Basis. Das neue Konzept ermöglicht es, ohne großen Zusatzaufwand, einen vollwertigen gedruckten Band und eine digitale Version mit zahlreichen Such- und Verlinkungsfeatures herzustellen. Vor allem aber ist es darauf ausgelegt, dass die Bearbeiter und Bearbeiterinnen ihren Editionstext in vollwertigem TEI-XML[2] herstellen können, ohne technische Vorkenntnisse zu benötigen. Dank einer speziell eingerichteten Arbeitsumgebung im Programm *Oxygen XML Editor*[3] können sie ihre bisher gewohnte Arbeitsweise weitgehend beibehalten. Unser Vortrag möchte dieses neue Editionskonzept und die Arbeitsumgebung vorstellen sowie von den Erfahrungen berichten, die wir bei der Herstellung des Bands 9 der Bayerischen Ministerratsprotokolle gemacht haben.

Die Edition „Protokolle des Bayerischen Ministerrats 1945–1962"[4], die seit Anfang der 1990er Jahre , dank Förderung durch den Freistaat Bayern, bei der Historischen Kommission entsteht, ist eines der wichtigsten Forschungsprojekte zur bayerischen Zeitgeschichte. In bisher acht Bänden, die die Jahre 1945 bis 1951 abdecken, dokumentieren die ausführlichen Gesprächsprotokolle der Ministerratssitzungen das Handeln der Bayerischen Staatsregierung. Sie geben Einblick in die Fragen und Herausforderungen der Nachkriegszeit in Bayern, erst unter amerikanischer Kontrolle, dann in der noch jungen Bundesrepublik. Deutlich lassen sich aus Ihnen die Kontroversen und die teilweise heftig geführten Diskussionen zwischen den Kabinettsmitgliedern herauslesen. Nicht selten gleichen dabei die vor 70 Jahren geführten Debatten zu Themen wie Flüchtlingen, Verkehrspolitik und Bildungswesen verblüffend den aktuellen politischen Auseinandersetzungen in Bayern und Deutschland. Die ersten acht Bände der Edition (bis 1952) stehen im Volltext retrodigitalisiert auch online zur Verfügung.[5]

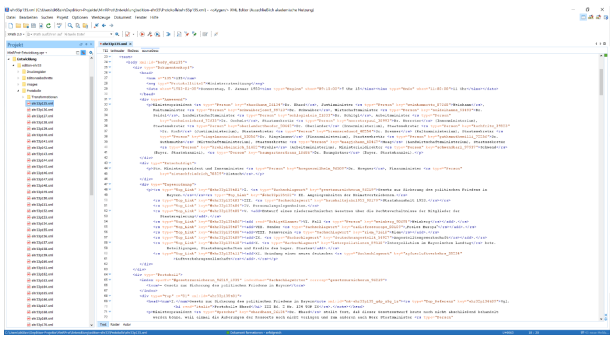


Abbildung 1: Ein Protokoll aus der Edition in seiner XML-Grundform, ...

Die Arbeiten an dem hier vorgestellten Editionskonzept begannen im Jahr 2014. Ausgangsbasis war damals das Projekt *ediarum*[6] der Berlin-Brandenburgischen Akademie der Wissenschaften (Dumont/Fechner 2015). Während es Inspiration und eine wertvolle technische Grundlage lieferte, ergaben sich rasch Anforderungen, die mit dem damals verfügbaren Werkzeugen von *ediarum* nicht vollständig zu erfüllen waren. Und so begannen Matthias Reinert und Maximilian Schrott (beide München), Mitarbeiter in der Abteilung „Digitale Publikationen" mit der Entwicklung eines eigenen Ansatzes. Dessen Kerngedanke ist es, dass der Bearbeiter möglichst gar keine Berührungspunkte mit der XML-Syntax der Dokumente hat. Stattdessen soll er sich voll auf die Arbeit am Editionstext konzentrieren können. Dennoch legt Herr Dr. Braun teils komplexe Strukturen auf Basis des TEI Lite Schemas in den Protokolltext an und baut während der gesamten Editionsarbeit interne und externe Verknüpfungen in den Text ein. So wird das Projekt mit Metainformationen angereicht, die vor allem bei der digitalen Veröffentlichung im Internet einen Mehrwert bringen. Entscheidend unterstützt wird er dabei durch den *Oxygen XML Editor*. Dieses leistungsstarke XML-Bearbeitungsprogramm, bietet die Möglichkeit eine stark individualisierte Arbeitsumgebung einzurichten.

Die wichtigsten Bestandteile dieser Arbeitsumgebung für unser Editionsprojekt sind Operationen und Stile. Die Operationen sind eine Reihe von vorbereiten Funktionen, die eine Eingabe des Bearbeiters entgegen nehmen, daraus XML-Fragmente generieren und in den Text einfügen. Sie werden entweder aus Standardoperationen, die der *Oxygen XML Editor* bietet, konfiguriert oder können selbst über eine Java-Schnittstelle programmiert werden. Herr Dr. Braun löst sie nach Bedarf über Buttons in der Menüleiste des Editors aus. Die Operationen ermöglichen es ihm mit wenigen Mausklicks und Eingaben in Textmasken komplexe Strukturelemente wie den Protokollkopf einzufügen, den Protokolltext in Tagesordnungspunkte und Unterpunkte zu gliedern oder Verweise auf andere Abschnitte und Fußnoten innerhalb des Editionskorpus zu setzen. Auf diese Weise kann Herr Dr. Braun das XML seiner Editionsdokumente bequem und ohne Angst vor lästi-

gen Flüchtigkeitsfehlern bearbeiten und das mit minimalem Einarbeitungsaufwand.

Die Stile werden über Cascading Style Sheets (CSS) eingerichtet. Mit diesen kann die Darstellung der XML-Dokumente in *Oxygen XML Editor* weitreichend angepasst werden. Für die Bayerischen Ministerratsprotokolle fiel der Entschluss, das XML-Markup, das oft als störend empfunden wird, größtenteils auszublenden und nur den reinen Editionstext darzustellen. Die semantischen Strukturen werden stattdessen über Formatierung, Textsetzung und farbliche Markierungen kenntlich gemacht. Herr Dr. Braun kann zwischen mehreren hinterlegten Stilen wechseln, die jeweils einen bestimmten Schritt des Workflows besonders unterstützen.

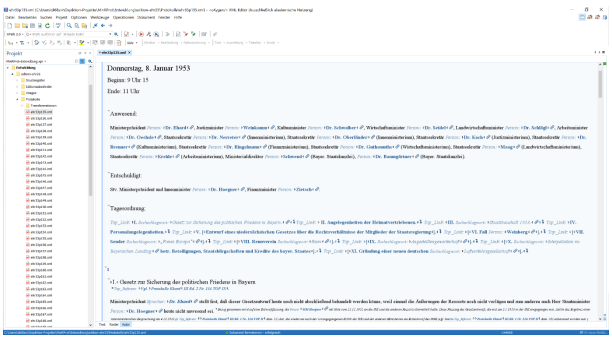


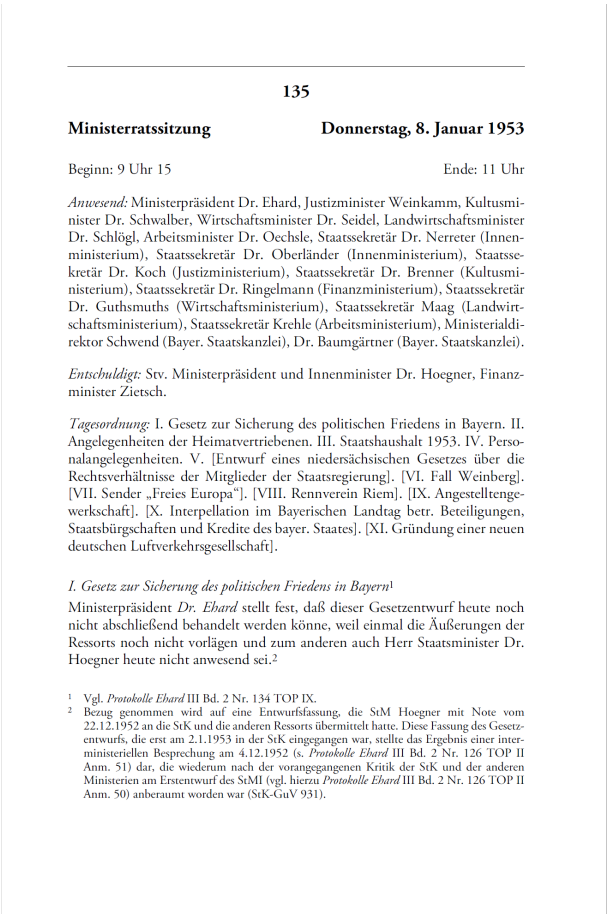
Abbildung 2: … in der Darstellung der Arbeitsumgebung …

Zusammengenommen ermöglichen Operationen und Stile ein bequemes und produktives Arbeiten in XML. Durch das vollständige Verlagern der Syntax in den Hintergrund, sinkt die Einstiegshürde für XML-unkundige Bearbeiter. Gleichzeitig ist die Arbeit schneller, bequemer und weniger fehleranfällig, als wenn das Markup von Hand eingegeben werden müsste. Und der Blick des Bearbeiters auf den Text wird nicht durch unübersichtliche XML-Element-Konstruktionen versperrt.

Eine neu hinzugekommene Aufgabe für Herrn Dr. Braun während der Editionsarbeit ist die Markierung der registerrelevanten Entitäten (Personen, Orte, Schlagwörter, Gesetze, Quellen und Literaturtitel) im Text. Die systematische Referenzierung dieser Textstellen erhöht den Grad der thematischen Verknüpfung innerhalb der Edition enorm und ermöglicht ein hohes Maß an Erschließ- und Durchsuchbarkeit für die digitale Präsentation. Kernstück dieser Verknüpfungsarbeit ist die Registerdatenbank (RDB). In diese trägt der Bearbeiter sämtliche registerrelevanten Entitäten im TEI-Format ein, ebenfalls unterstützt durch die Arbeitsumgebung. Da Herr Dr. Braun zwar an verschiedenen Orten aber alleine an seinen Editionstexten arbeitet, schien es zunächst unnötig die RDB als *eXist*-Datenbank[7] einzurichten. Stattdessen handelt es sich um eine Reihe von einfachen XML-Dateien, die jeweils einen einzelnen Entitätstypen beinhalten. Diese Lösung spart Wartungsaufwand und vermeidet einen Onlinezwang während der Arbeit, ohne dass sich dadurch ein merkbarer Nachteil ergab. Die Sicherung der RDB-Dateien, wie auch der Editionsdokumente, erfolgt stattdessen über *Sync+Share*[8], das Cloud-Speicherangebot des Leibniz-Rechenzentrums. Dieses sorgt gleichzeitig für den Austausch der Arbeitsdateien mit den technischen Betreuern.

Vom Text aus kann Herr Dr. Braun über entsprechende Operationen den jeweils benötigten RDB-Eintrag heraussuchen und eine Referenz auf diesen hinterlegen. Zusätzlich zu den Registerinformationen wie Personen- und Ortsname kann die RDB auch weiterführende Daten zu ihren Einträgen speichern. Zum Beispiel die geographischen Koordinaten eines Ortes, Normdaten-Identifier (GND)[9] oder biographische Informationen zu den Personen. Diese können während der Bearbeitung und später in der digitalen Präsentation genutzt werden, um zusätzlichen Funktionen zu realisieren.

Während sich Herr Dr. Braun auf seine Editionsarbeit konzentrieren kann, kümmert sich die Abteilung Digitale Publikationen der Historischen Kommission um die technische Umsetzung. Zu Projektbeginn prüfte sie, in Absprachen mit dem Bearbeiter, die Anforderungen des Editionsprojekts. Auf dieser Basis wurde dann entschieden, welche Phänomene des Editionstextes in XML kodiert werden mussten und wie. Aus diesen Überlegungen entstanden schließlich die digitalen Editionsrichtlinien. Anhand dieser richtete Maximilian Schrott dann die Editionsumgebung in *Oxygen XML Editor* ein, designte die Stile und programmierte die Operationen. Nach dieser Erstkonfiguration leistete er Herrn Dr. Braun während des gesamten Bearbeitungszeitraums fortlaufend Support. Er verbesserte Fehler in der Arbeitsumgebung, entwickelte diese auf Basis seines Feedbacks weiter und überwachte die korrekte Form der Editionsdokumente. Zum Ende der Editionsarbeit hin übernahm die Abteilung Digitale Publikation schließlich die Umwandlung des Editionstextes in die gedruckte Form.



135

Ministerratssitzung **Donnerstag, 8. Januar 1953**

Beginn: 9 Uhr 15 Ende: 11 Uhr

Anwesend: Ministerpräsident Dr. Ehard, Justizminister Weinkamm, Kultusminister Dr. Schwalber, Wirtschaftsminister Dr. Seidel, Landwirtschaftsminister Dr. Schlögl, Arbeitsminister Dr. Oechsle, Staatssekretär Dr. Nerreter (Innenministerium), Staatssekretär Dr. Oberländer (Innenministerium), Staatssekretär Dr. Koch (Justizministerium), Staatssekretär Dr. Brenner (Kultusministerium), Staatssekretär Dr. Ringelmann (Finanzministerium), Staatssekretär Dr. Guthsmuths (Wirtschaftsministerium), Staatssekretär Maag (Landwirtschaftsministerium), Staatssekretär Krehle (Arbeitsministerium), Ministerialdirektor Schwend (Bayer. Staatskanzlei), Dr. Baumgärtner (Bayer. Staatskanzlei).

Entschuldigt: Stv. Ministerpräsident und Innenminister Dr. Hoegner, Finanzminister Zietsch.

Tagesordnung: I. Gesetz zur Sicherung des politischen Friedens in Bayern. II. Angelegenheiten der Heimatvertriebenen. III. Staatshaushalt 1953. IV. Personalangelegenheiten. V. [Entwurf eines niedersächsischen Gesetzes über die Rechtsverhältnisse der Mitglieder der Staatsregierung]. [VI. Fall Weinberg]. [VII. Sender „Freies Europa"]. [VIII. Rennverein Riem]. [IX. Angestelltengewerkschaft]. [X. Interpellation im Bayerischen Landtag betr. Beteiligungen, Staatsbürgschaften und Kredite des bayer. Staates]. [XI. Gründung einer neuen deutschen Luftverkehrsgesellschaft].

I. Gesetz zur Sicherung des politischen Friedens in Bayern[1]

Ministerpräsident *Dr. Ehard* stellt fest, daß dieser Gesetzentwurf heute noch nicht abschließend behandelt werden könne, weil einmal die Äußerungen der Ressorts noch nicht vorlägen und zum anderen auch Herr Staatsminister Dr. Hoegner heute nicht anwesend sei.[2]

[1] Vgl. *Protokolle Ehard* III Bd. 2 Nr. 134 TOP IX.
[2] Bezug genommen wird auf eine Entwurfsfassung, die StM Hoegner mit Note vom 22.12.1952 an die StK und die anderen Ressorts übermittelt hatte. Diese Fassung des Gesetzentwurfs, die erst am 2.1.1953 in der StK eingegangen war, stellte das Ergebnis einer interministeriellen Besprechung am 4.12.1952 (s. *Protokolle Ehard* III Bd. 2 Nr. 126 TOP II Anm. 51) dar, die wiederum nach der vorangegangenen Kritik der StK und der anderen Ministerien am Erstentwurf des StMI (vgl. hierzu *Protokolle Ehard* III Bd. 2 Nr. 126 TOP II Anm. 50) anberaumt worden war (StK-GuV 931).

Abbildung 3: … und in der finalen PDF-Druckfassung.

Denn auch wenn die Veröffentlichung einer hochwertigen, digitalen Version im Fokus des Editionskonzepts steht, sollte auch der neunte Band der Bayerischen Ministerratsprotokolle noch gedruckt erscheinen. Die PDF-Vorlage für den Druck wird dabei direkt aus den erstellten XML-Dateien erzeugt. Als Satzprogramm dient das in den *Oxygen XML Editor* integrierte *Apache-FOP*[10]. In dieses werden die Editionsdateien mittels einer XSL-Transformation überführt. So lässt sich in weniger als einer Minute eine vollständige Druckversion des kompletten 1000seitigen Druckbandes im gewünschten Layout erstellen. Der finale Satz erfolgt somit komplett intern, ohne Zuarbeiten durch den Verlag. Der automatisierte Satz funktioniert insgesamt sehr gut, es bleiben aber vereinzelte Mängel. Vor allem Zeilen-, Spalten- und Seitenumbruch müssen an einigen Stellen manuell nachgearbeitet werden. Auch bei der Erstellung der gedruckten Version ergeben sich durch die Anreicherungsarbeiten in XML Vorteile. So kann durch die wortgenaue Kennzeichnung der registerrelevanten Entitäten und die Verknüpfung mit der RDB auch das Druckregister halbautomatisch erstellt und dem Bearbeiter so eine besonders mühsame Arbeit erspart werden.

Das abgeschlossen Pilotprojekt wird von Seiten der Historischen Kommission als sehr erfolgreich eingestuft. Die Arbeit im *Oxygen XML Editor* wurde von Herrn Dr. Braun als angenehm empfunden, die Drucklegung wurde pünktlich abgeschlossen und es konnten wertvolle Erfahrungen zur Verbesserung der Arbeitsumgebung und des Workflows gesammelt werden. Der Komfort für den Bearbeiter und die Qualität der Ergebnisse wird freilich durch eine vergleichsweise hohen Einrichtungs- und Supportaufwand erkauft. Diese Mehrarbeiten werden aber von der technischen Betreuung geleistet. Eine zusätzliche Belastung für den wissenschaftlichen Editor wird weitestgehend vermieden. Die Veränderungen in seinem Workflow sind zwar merkbar, aber nicht einschneidend. In der jetzigen Form funktioniert die Zusammenarbeit zwischen Bearbeiter und Technikern für beide Seiten sehr fruchtbringend. Außerdem sind die gewonnen Erfahrungen und Programmierarbeiten zum größten Teil für andere Projekte wiederverwendbar. Die Historische Kommission sieht sich deshalb in ihrem Entschluss bestätigt, zukünftig verstärkt auf das digitale, XML-basierte Editionskonzept zu setzen. Neben den noch ausstehenden Bänden der Bayerischen Ministerratsprotokolle sind bereits drei Briefeditionsprojekte in Arbeit, die auf diesem Ansatz beruhen.[11] Weitere Projekte befinden sich in der Vorbereitungs- und Planungsphase.

Der neunte Band der Bayerischen Ministerratsprotokolle ist im Oktober 2019 gedruckt erschienen. Die Veröffentlichung im Internet wird sich noch bis voraussichtlich 2022 verzögern. Dann endet mit dem Erscheinen des 10. Bandes, die mit dem Verlag vereinbarte Exklusivitätsperiode im Druck. Der Inhalt von Band 9 und der RDB können somit zum bestehenden Webangebot der Protokolle des Bayerischen Ministerrats[12] hinzugefügt werden. Ein größeres Funktionsupdate für die Website, das hiermit einhergehen soll, befindet sich derzeit noch in der Konzeptionsphase. Es soll das Webangebot um zahlreiche neue Features erweitern, die die Anreicherungen und Vernetzungen im Editionstext für die User zugänglich macht. Noch geprüft wird, welche Verlinkungsmöglichkeiten zu externen Webangeboten und Normdatenbanken umgesetzt werden können. Außerdem gibt es Überlegungen für spezialisierte Recherche- und Visualisierungsmöglichkeiten, zum Bespiel mittels der in der RDB erfassten Orte auf Karten.

Fußnoten

1. Das Kabinett Ehard III 18.Dezember 1950 bis 14. Dezember 1954, Band 3 8.1.1953–29.12.1953. bearb. von Oliver Braun, München [2019] (= Die Protokolle des Bayerischen Ministerrats. 1945-1962, Hg. von der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften durch Andreas Wirsching und von der Generaldirektion der Staatlichen Archive durch Margit Ksoll-Marcon, Band 9).
2. http://www.tei-c.org ; alle URLs in diesem Artikel wurden zuletzt am 25.09.2019 aufgerufen.
3. http://oxygenxml.com
4. Die Protokolle des Bayerischen Ministerrats 1945-1962 [ehemals 1945-1954], Hg. von der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften und der Generaldirektion der Staatlichen Archive Bayerns, 9 Bände, bearbeitet von Karl-Ulrich Gelberg [Bände 1-5] und Oliver Braun [Bände 6-9], 1995-2019.
5. www.bayerischer-ministerrat.de
6. http://www.bbaw.de/telota/software/ediarum
7. https://exist-db.org
8. https://syncandshare.lrz.de
9. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
10. https://xmlgraphics.apache.org/fop/
11. Die wissenschaftliche Korrespondenz des Historikers Karl Hegel (1813-1901), bearbeitet von Marion Kreis; Der Briefwechsel zwischen Adolf Harnack und Friedrich Althoff (1886-1908), bearbeitet von Claudia Kampmann; Zwischen Wissenschaft und Politik. Hans Delbrück – Ausgewählte Korrespondenz (1868-1929), bearbeitet von Andreas Rose und Jonas Klein.
12. Siehe Anm. 5.

Bibliographie

Dumont, Stefan / Fechner, Martin (2012): „Digitale Arbeitsumgebung für das Editionvorhaben ‚Schleiermacher in Berlin 1808–1834'" http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834 [letzter Zugriff 25.09.2019].

Dumont Stefan / Fechner Martin (2015): „Bridging the Gap: Greater Usability for TEI encoding, in: Journal of the Text Encoding Initiative 8" http://jtei.revues.org/1242 [letzter Zugriff 25.09.2019].

Die Falte: Ein Denkraum für interaktive und kritische Datenvisualisierungen

Brüggemann, Viktoria
viktoria.brueggemann@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Deutschland

Bludau, Mark-Jan
mark-jan.bludau@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Deutschland

Dörk, Marian
doerk@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Deutschland

Einführung

Obwohl die Geisteswissenschaften verstärkt auf Möglichkeiten der Datenvisualisierung zurückgreifen, bleibt es eine Herausforderung, interaktive visuelle Repräsentationen zu erzeugen, die erkenntnisreich und kritisch, aber auch zugänglich und ansprechend sind. Zu selten sind Beziehungen zwischen dem einzelnen Objekt und der ganzen Sammlung, sowie animierte Übergänge zwischen Ansichten bedacht und bewusst gestaltet (Chevalier et al. 2016). Obwohl sich die vielfältigen kulturellen Sammlungen zweifellos für digitale Realisierungen eignen, ist eine Umsetzung der ihnen zugeschriebenen Qualitäten in reichhaltige und kohärente Datenvisualisierungen sowohl theoretisch als auch praktisch oft noch nicht ausreichend.

Wir schlagen den vom französischen Philosophen Gilles Deleuze (Deleuze 1996) entwickelten Begriff der Falte als Denkraum für die Interpretation und Gestaltung interaktiver Visualisierungen vor. Die Falte bietet eine vielversprechende Perspektive auf Wissensräume und deren Repräsentation und wirft ein kritisches Licht auf die zugrunde liegenden Daten und ihre Komplexität. Anhand von Illustrationen stellen wir Operationen und Qualitäten der Falte vor und diskutieren, wie aus ihnen eine kritische Perspektive auf interaktive Datenvisualisierungen und Orientierungshilfe für deren geisteswissenschaftlichen Einsatz erwachsen kann.

Die Falte

In „Die Falte. Leibniz und der Barock“ (1988) bezieht sich Gilles Deleuze auf Leibniz' Monadologiebegriff, der eine dualistische Ontologie ablehnte und stattdessen die Monade als Grundbaustein der materiellen Welt etablierte (Leibniz 1898). Leibniz stellte sich die Seele als Monade vor, ein Haus ohne Türen und Fenster, in dem sich die Außenwelt nur als Innenbild widerspiegelt (Wagner 1995). Deleuze rekonstruiert Leibniz' Philosophie als barocke Metaphysik, die in seinem Sinne die Falte als Element der unendlichen Wiederholung enthält (Laerke 2010). Die Monade in Deleuzes Sinn ist auf zwei Ebenen mit Falten gefüllt: den „Faltungen der Materie“ und den „Falten der Seele“, die abgegrenzt und doch kontinuierlich miteinander verwoben sind. Diese Metapher bezieht sich auch auf den menschlichen Körper und die menschliche Seele und macht Prozesse der Informationsinterpretation und -akkumulation verständlich: Würde man mit einer Information auf der Ebene der Sinne konfrontiert, würden sich die Falten der Materie automatisch in Bewegung versetzen, woraufhin die resultierenden Verbindungen zu anderen Informationen sichtbar würden. Jede Information ist somit in den unendlichen (und „virtuellen“) Faltungen der Seele bereits angelegt, sie wird aber erst durch den Prozess des Faltens und Entfaltens sichtbar.

Die Operationen der Falte

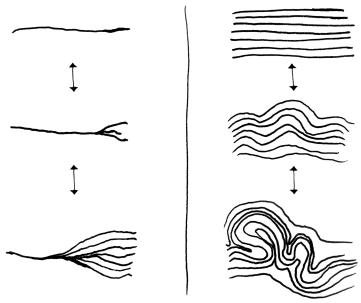


Abbildung 1. Die Operationen der Falte illustriert: Explikation und Implikation (links), Komplikation (rechts).

Für Deleuze bilden die drei Operationen „Explizieren-Implizieren-Komplizieren“ die Triade der Falte (Deleuze 1996, 45). Die ersten beiden Operationen müssen als ein Paar verstanden werden, bei dem das Eine das Andere umkehrt. Explizieren beschreibt den Prozess der Entfaltung, z.B. das Öffnen eines Buches oder das Ausfächern in Unterabschnitte (siehe Abb. 1, links). Im Gegenteil bezieht sich das Implizieren auf den Prozess des Faltens, der etwas in Größe und Detail reduziert. Wenn etwas in der Monade durch den Prozess der Implikation verborgen wird, umfasst das Gefaltete immer noch alles andere, auch wenn dies nicht immer wahrnehmbar ist. Durch die Explikation werden versteckte Verbindungen wieder sichtbar, wobei das Entfaltete seine Verbindung zum Ursprungspunkt nicht verliert und nahezu unendliche Verbindungen zu anderen Punkten besitzen kann.

Beim Ein- und Ausfalten können daher unvorhersehbare Ergebnisse und neue Verbindungen entstehen, da eine Faltung das plötzliche Nebeneinander von ehemals gegenüberliegenden Punkten erzeugen kann. An dieser Stelle führt Deleuze die Komplikation als dritte Operation der Falte ein (siehe Abb. 1, rechts). Die Funktionsweise der Komplikation erklärt den Prozess der Informationsakkumulation und Vernetzung von allem Wahrnehmbaren, während sie andererseits dessen Beliebigkeit anspricht. Da alles in der Monade zur Unendlichkeit gefaltet ist, liegt jede Möglichkeit bereits in ihr, auch wenn sie zu keinem Zeitpunkt in ihrer Gesamtheit zu erfassen ist. Überraschende Ereignisse sind also nicht mehr als ein „komplizierter“ Faltprozess, bei dem die Verbindungen neu geord-

net werden und jeweils nur ein Teil des verbundenen Universums sichtbar wird.

Die Qualitäten der Falte

Deleuzes Theorie der Falte lädt dazu ein, die dynamischen Eigenschaften digitaler Informationsräume näher zu betrachten. Im Folgenden beleuchten wir drei hervorstechende Qualitäten näher:

Kohärenz: Anstatt Informationspunkte als diskrete Objekte zu betrachten, drückt die Falte die kohärente Qualität monadologischer Strukturen aus, die durch Kontexte und Beziehungen definiert sind. Diese Qualität vervielfacht sich unendlich: Unabhängig davon, wie weit ein Informationsraum durchlaufen wird, besitzt jede Ausgabe oder jeder Informationspunkt einen Bezug zum Beginn der Suchanfrage oder zum gesamten Universum.

Elastizität: Diese Qualität erfasst die ständige Bewegung von Informationen, Gedanken oder Verbindungen, bei dem ein Impuls dem anderen folgt. Durch ihre Faltbarkeit können einzelne Elemente und ganze Anordnungen mehrere mögliche Erscheinungsformen annehmen und ihre Form und Richtung ändern, was zu flexiblen Prozessen des Dehnens und Verzerrens führt, ohne jedoch das erste Prinzip der Kohärenz aufzugeben.

Unendlichkeit: Da die Falte unendliche Möglichkeiten bietet, bleiben die Faltprozesse bis zu einem gewissen Grad unvorhersehbar und zufällig. Dies bedeutet nicht, dass Faltungen nicht auch wiederholbar oder rückverfolgbar sind, sondern dass sie nie als endgültig oder abgeschlossen gelten können. Die Qualität der Unendlichkeit ist stark mit der Operation der Komplikation verbunden und erinnert daran, dass Informationsräume auf den ersten Blick strukturiert und transparent erscheinen mögen, aber stark von der Perspektive der Betrachter*innen und einer Vielzahl von Datendimensionen abhängig sind.

Ein Denkraum für die Datenvisualisierung

Die Falte bietet eine einzigartige, kritische Perspektive auf die Form und Funktion von Informationsstrukturen. Die fortwährende Unvollständigkeit von Erkenntnisprozessen macht diese Theorie so relevant für die Datenvisualisierung, die darauf abzielt, komplexe Sachverhalte zu kommunizieren, aber ihre Auslassungen und Reduktionen transparent machen muss. Mit der zunehmenden Relevanz der Datenvisualisierung in den digitalen Geisteswissenschaften wächst ebenso die Notwendigkeit, sich mit der Interaktivität als einem ihrer wesentlichen Aspekte auseinanderzusetzen. Visualisierungen entlang der Falte zu entwerfen und analysieren bedeutet, Informationsräume als elastische, kohärente und potenziell unendliche Räume zu verstehen. Im Folgenden stellen wir anhand von beispielhaften Illustrationen dar, wie Operationen der Falte bereits in existierenden Visualisierungen umgesetzt werden und wie die formulierten Qualitäten den Gestaltungsprozess unterstützen können.

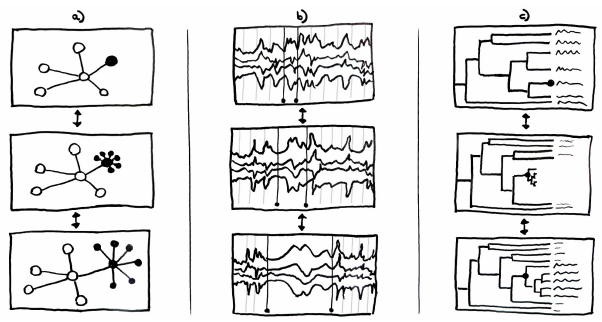


Abbildung 2. Beispiele für Explikation (von oben nach unten) und Implikation (umgekehrt) in Datenvisualisierungen: a) Das Entfalten eines Netzwerkgraphen zeigt detaillierte Nachbarschaftsbeziehungen des ausgewählten Knotens. b) Das Ausdehnen eines ausgewählten Bereichs eines Streamgraphen führt zur Detailerhöhung im ausgewählten und zur Kompression der nicht ausgewählten Bereiche. c) Die Auswahl eines Elements in einem Baumdiagramm öffnet zusätzliche Unterzweige, wobei durch Kompression der übrigen Zweige neuer Raum geschaffen wird.

Die **Kohärenz** der Falte manifestiert sich in der tiefen Kontextualisierung und Vernetzung aller Elemente. Um diesen hohen Grad an Kohärenz in interaktiven Visualisierungen zu realisieren, müssen die visuelle Kodierung und die interaktiven Funktionen konsequent über alle Ansichten hinweg gekoppelt werden. Abbildung 2 zeigt, wie sich einzelne Aktionen auf die gesamte Visualisierung auswirken können, beispielsweise wenn bestimmte Bereiche komprimiert werden, ohne dabei den Zusammenhang zum Rest zu verlieren. Die Gestaltung der jeweiligen Übergänge sollte sinnvoll und konsistent sein, ebenso wie konsistente Designentscheidungen über die gesamte Visualisierung, unabhängig vom dynamischen Zustand, getroffen werden sollten.

Ein hohes Maß an **Elastizität** bedeutet, dass Elemente flexibel in eine Anordnung eingebettet sind und dass sie in der Lage sind, ihre Form zu verändern oder ihre Ausgangsposition zu verlassen. Sie können sich so immer wieder neu zeigen - auch in unvorhergesehenen Darstellungen. Abbildung 3 a) zeigt eine Zeitachse, welche entgegen ihrer linearen Anordnung flexible Positionierungen zulässt, welche auf der veränderten Kodierung ihrer einzelnen Informationspunkte beruht. Während in einer elastischen Visualisierung nichts vollständig fixiert ist, sind Bewegungen jedoch auch nicht unbegrenzt dem Zufall überlassen. Das Spektrum der dynamischen Veränderungen einzelner Elemente und ganzer Anordnungen sollte daher sorgfältig abgewogen werden.

Die Qualität der **Unendlichkeit** bezieht sich auf eine Darstellungsvielfalt und -kontinuität, zum Beispiel durch verschiedene Kombinationen von visuellen Formen und interaktiven Funktionalitäten sowie kreisförmige oder offene Navigationsmechanismen. Dies kann neue Ausdrücke eines Datensatzes und eine Vielzahl von möglicherweise unerwarteten Entdeckungen hervorrufen, wie sie das Konzept der Serendipität hervorhebt (Leong et al. 2011, Thudt et al. 2012). Des Weiteren sollte eine kontinuierliche Navigation zwischen verschiedenen Visualisierungszuständen möglich sein, ohne eine Fokussierung auf bestimmte Teile vorzunehmen.

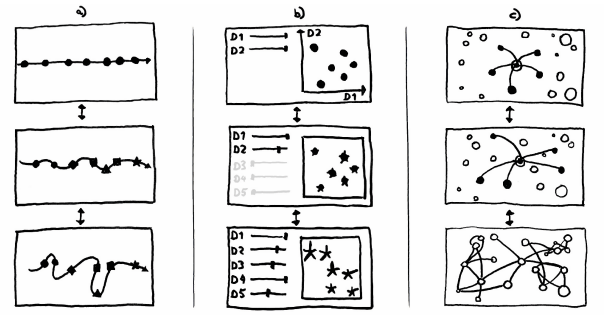


Abbildung 3. Komplikationsbeispiele: a) Faltung einer Zeitachse basierend auf der Ähnlichkeit zwischen Datenpunkten. b) Verwendung von mehrdimensionalen Reduktionstechniken in Kombination mit kodierten Glyphen zur schrittweisen Hinzufügung weiterer Dimensionen. c) Umschalten zwischen egozentrischen und nicht-zentrischen Gesamtzuständen eines Netzwerkdiagramms.

Mit der Falte schlagen wir einen Ansatz vor, der die Notwendigkeit einer gleichzeitigen und koordinierten Betrachtung von Interaktion und Repräsentation in der Datenvisualisierung betont. Während die Herausforderungen kultureller Daten und ihrer Umsetzung in Visualisierungen bereits kritische Aufmerksamkeit erlangt haben, werden die Chancen von Interaktionstechniken zu diesem Zweck bisher wenig diskutiert. Nicht nur die Konzeption einer visuellen Kodierung von Informationsräumen, sondern Interaktion und Übergänge zwischen verschiedenen Zuständen sind entscheidend für die Entwicklung überraschender Visualisierungen im Sinne der Falte.

Darüber hinaus können die zugrunde liegenden Daten selbst als Falten betrachtet werden, was uns daran erinnert, dass jede Perspektive nur eine mögliche Version der Realität darstellt, während sie unendlich viele andere Möglichkeiten in sich einschließt. Die häufige Ansicht von Daten als „gegeben“ (Drucker 2011), „objektiv“ oder „unveränderlich“, wird durch die Theorie der Falte in Frage gestellt. Diese Perspektive ist besonders relevant im Zusammenhang mit kritischen Sichtweisen auf die Macht und Rhetorik von Daten und ihrer Repräsentation (D'Ignazio et al. 2016, Dörk et al. 2013, Hullman und Diakopoulos 2011). Die Falte bietet hier einen Denkraum, welcher zu einer kritischen Analyse und Gestaltung von interaktiven Datenvisualisierungen anregt.

Bibliographie

Bredekamp, Horst (1988): „Leibniz' Gewebe: Strumpfband, Falte, Leinwand“, in: Töteberg, Michael / Wenders, Wim (eds.): Die Logik der Bilder. Verl. d. Autoren, 233-238.

Chevalier, Fanny / Riche, Nathalie / Plaisant, Catherine / Chalbi, Amira / Hurter, Christophe (2016): „Animations 25 Years Later: New Roles und Opportunities“, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, 280-287.

D'Ignazio, Catherine / Klein, Lauren F. (2016): „Feminist Data Visualization“, in: VIS4DH: Workshop on Visualization for the Digital Humanities.

Deleuze, Gilles (1996): Die Falte. Leibniz und der Barock, übers. von Ulrich Johannes Schneider. Frankfurt: Suhrkamp.

Dörk, Marian / Feng, Patrick / Collins, Christopher / Carpendale, Sheelagh (2013): „Critical InfoVis: Exploring the Politics of Visualization“, in: alt.chi 2013: Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems, 2189-2198.

Drucker, Johanna (2011): „Humanities Approaches to Graphical Display“, in: Digital Humanities Quarterly 5.1, 1-21. www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Freitas, Elizabeth de (2016): „The New Empiricism of the Fractal Fold“, in: Cultural Studies Critical Methodologies 16.2, 224-234.

Hullman, Jessica / Diakopoulos, Nick (2011): „Visualization Rhetoric: Framing Effects in Narrative Visualization“, in: IEEE Transactions on Visualization and Computer Graphics 17.12, 2231-2240.

Laerke, Mogens (2008): „Four Things Deleuze Learned from Leibniz“, in: Tuinen, Sjoerd van / McDonnell, Niamh (eds.): Deleuze and the fold. Palgrave Macmillan, 25-45.

Leong, Tuck Wah / Harper, Richard / Regan, Tim (2011): „Nudging towards serendipity: a case with personal digital photos.“ Proceedings of the 25th BCS Conference on Human-Computer Interaction, 385-394.

Leibniz, Gottfried (1898): Monadology, übers. von Robert Latta. Clarendon Press.

Thudt, Alice / Hinrichs, Uta / Carpendale, Sheelagh (2012): „The bohemian bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization“, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1461.

Wagner, Thomas (1995): „Die Falten des Geistes: Barock werden mit Gilles Deleuze“, in: Frankfurter Allgemeine Zeitung, vol. 1995, Nr. 283. https://www.faz.net/-gr0-6q9tr

Die Kanonfrage 2.0

Dziudzia, Corinna

corinna@dziudzia.de
KU Eichstätt - Ingolstadt

Hall, Mark

mark.hall@work.room3b.eu
Martin-Luther-Universität Halle-Wittenberg

Die Kanonfrage 1.0

Maßgeblich leitete die amerikanische Literaturwissenschaft in den 1970er Jahren eine kritische Revision mit der Frage ein, wer eigentlich anhand welcher Kriterien entscheidet, welches literarische Werk zum Kanon gehört (vgl. Ziolkowski 2009). Diese Kritik findet in einem anhaltenden Prozess des ‚Entdeckens' und ‚Sichtbarmachens' nichtkanonisierter Autor_innen und Werke Niederschlag (vgl. u.a. Brinker-Gabler 1978; Hilger 2015) und fördert ein breites Spektrum heterogenen Literatur- und Kunstschaffens zu Tage. Die im Rahmen der Kanondebatte im Wesentlichen formulierte Kritik am zu ‚weißen' und zu ‚männlichen' Kanon als „Machtinstrument" (Winko 1996, 500) wird auf der theoretischen Ebene mit dem Nachdenken über diskursive Machtpraktiken ebenso reflektiert wie mit der Konzeption des Archivbegriffs (vgl. Foucault 1990; Derrida 2009) und für den digitalen Raum erweitert. Abigail De Kosnik

spricht entsprechend von einer potentiellen Verschmelzung von Kanon und Repertoire:

[…] digital archives potentially redefine what A. Assmann calls a ‚active memory' and ‚passive memory', in the sense that these become highly individualized: all materials contained in an online database are equally available to the user – no materials are any more ‚hidden' or ‚stored away' than any other materials, all materials that are indexed can be retrieved from the database – and so users of an Internet Archive may ‚activate' whichever of the materials they wish, constructing their own personal canons based on the materials that they use. […] Digital archives erect no physical barriers between categories of information, so conceivably any piece of information, any archived data, can enter into one person's repertoire and canon; thus, there are as many possible canons as there are archive users, and no possibility for a single canon, achieved by a consensus of cultural archive users, that would be distinct from the culture's archive. (De Kosnik 2016, 66)

So wie durch die frühe Kanondebatte die grundlegende Frage gestellt worden ist, wessen Werke eigentlich publiziert, rezensiert und damit potentiell kanonisiert werden können, muss diese Frage heute aktualisiert werden: Wessen Werke werden wie digitalisiert, um im Sinne De Kosniks überhaupt derart aktiviert werden zu können?

Die Kanonkritik verweist zudem auf grundlegende Reflexionen der Wissenschaftsgeschichte. Wesentlich ist hierfür etwa die Wiederentdeckung Ludwig Flecks und sein durch Thomas Kuhn verstärktes Betonen, wie sehr das jeweils tradierte Wissen Ausweis von Selektionsprozessen ist, das konkreten Rahmenbedingungen genauso wie Irrtümern und Denkzwängen unterliegt (Fleck 1980, 31; Kuhn 1996): Insofern die vorherige Generation das tradierenswerte Wissen in Lehrbüchern für die nachfolgende Generation auswählt, erscheint die jeweils angebotene bzw. fehlende Wissensrepräsentation aufschlussreich. Denn ungeachtet der seit mittlerweile Jahrzehnten geäußerten Kritik an der Homogenität des Kanons ist noch in jüngerer Zeit, etwa durch die Frauenforschung, festgestellt worden, wie überraschend wenig, teilweise gar nicht, schreibende Frauen in Form ihrer Namen und Werke aktuell in deutschen Leselisten, Literaturgeschichten und Lehrbüchern repräsentiert sind (vgl. Sylvester-Habenicht 2009). Die Stabilität des kleinen Kernkanons (vgl. Braam & Hagstedt 2017, 83), in dem sich immer noch vor allem männliche Autoren präsent zeigen, scheint sowohl durch die Debatte als auch die große Zahl an Entdeckungen vergessener schreibender Frauen (vgl. u.a. die Forschungsarbeiten Brinker-Gablers oder Becker-Cantarinos) wenig veränderbar, das Wissen um die Existenz der Texte ist nach wie vor marginal und auf kleine Expert_innenkreise beschränkt.

Digitale Kanonkritik

In der digitalen Welt erweist sich Google Books zwar als umfangreiches Archiv der Texte auch von Autorinnen, stellt allerdings die Werke vorrangig als Scans zur Verfügung, weswegen diese für die automatisierte DH-Analyse weniger nutzbar sind, aber zumindest für die Lektüre verwendet werden können (vorausgesetzt, sie werden gefunden). Das wird ergänzt durch (allerdings nicht textverlässliche) Primärliteratur-Archive wie das *BYU Scholars Archive Sophie* (Evanson2003) , welche die Texte allerdings auch nicht in wiederverwendbarer Form anbieten, aber zumindest stärker erkennen lassen, wie viele deutsche Autorinnen und Texte es gibt (im Augenblick sind es über 1.000 PDFs). Deutsche digitale Archive wie *Textgrid* (Schmunk & Funk 2016) oder *Das deutsche Textarchiv* (Geyken 2013) sind texteditorisch fundierter und damit hinsichtlich der Qualität für den Einsatz in der universitären Lehre besser verwendbar, in ihrer Auswahl allerdings stark kanonorientiert, d.h., weibliches Schreiben ist dort sehr wenig präsent. *Das Deutsche Textarchiv* etwa basiert seine Auswahl vorrangig auf einschlägigen älteren Literaturgeschichten, die einen männlichen Kanon repräsentieren (der Anteil an Frauen in der Standardliteraturgeschichte von Gottfried Gervinus, *Geschichte der poetischen National-Literatur des Deutschen*, liegt bei weniger als 3%). Nach eigener Aussage ist es bei *Textgrid* explizites Interesse „nahezu alle wichtigen kanonisierten Texte und zahlreiche weitere literaturhistorisch relevante Texte" bereitzustellen (https://textgrid.de/digitale-bibliothek). Von den im Augenblick 697 Autoren, deren Texte sich auf *Textgrid* finden, sind 63 Frauen (d.h. immerhin im Vergleich 9%). Davon allerdings sind einige Werke der frühen Frauenbewegung zuzurechnen – und damit nicht der Belletristik und nur bedingt von explizit literaturhistorischem Interesse. Das Korpus umfasst zudem nicht nur deutsche Autor_innen. Andere Angebote, wie die *Deutsche Digitale Bibliothek* (https://www.deutsche-digitale-bibliothek.de/), sind noch im Aufbau begriffen.

Es gibt allerdings digitale Archivprojekte, welche das erklärte Ziel haben, die Arbeiten von Frauen sichtbarer zu machen, unter anderem das „Women Writers Project" (Connell et al. 2017), „Orlando: Women's Writing in the British Isles from the Beginnings to the Present" (Booth 2017) oder „DaSind - Die Datenbank Schriftstellerinnen in Deutschland, Österreich, Schweiz 1945-2008" (Schulz 2008). Die ersten zwei sind jedoch kostenpflichtig und zielen nur auf englischsprachige Texte, und das dritte Projekt ist seit über einem Jahr nicht mehr online verfügbar (Stand: Dezember 2019).

Digitale literaturwissenschaftliche Forschung der deutschen Literatur scheint entsprechend vorrangig mit jenen Digitalisaten unternommen zu werden, die prominent zur Verfügung stehen, leicht zugänglich sind und in entsprechend nutzbaren Formaten vorliegen, daher überrascht es nicht, dass sich die Digital Humanities tendenziell eines recht kleinen und männlichen Kanons deutscher Literatur bedienen (Hall, 2019). Um diese Schieflage zunächst aufzuzeigen und dann potentiell zu korrigieren, wurde das *Unter der Oberfläche*-Projekt ins Leben gerufen.

Das Projekt verfolgt vier Ziele:

1. Die Kanonfrage vor dem Hintergrund der Digitalisierung neu zu stellen
2. Lücken in der Digitalisierung von Werken außerhalb des Kanons aufzuzeigen
3. Den Zugang zu vorhandenen Digitalisaten zu vereinfachen, um die Schwelle zur Nutzung dieser Werke zu reduzieren
4. Autor_innen und ihre Werke als bisher eher marginalisiertes Wissen auch für die nicht-wissenschaftliche Öffentlichkeit zugänglich zu machen

Unter der Oberfläche

Um diese Ziele erreichen zu können, wird im Rahmen des Projekts ein Online-Portal entwickelt, zusammen mit den notwendigen Werkzeugen, die darin enthaltenen Daten zu verwalten. Eine Grundidee in der technischen Umsetzung ist es, nicht noch ein weiteres Archiv für Digitalisate bereitzustel-

len, sondern die in den verschiedenen existierenden Archiven vorhandenen Digitalisate mit allgemeinen Informationen über tendenziell vergessene Autor_innen zusammenzuführen. Der These folgend, dass Vieles ‚unter der Oberfläche' schlummert, wenngleich nicht immer in optimalen Formaten, geht es dem Projekt primär um das Sichtbarmachen dessen, was da ist und seien es zunächst nur die Namen von Autorinnen. Es geht dem Projekt nicht um die Digitalisierung oder Archivierung von Werken, die noch nicht vorliegen, vielmehr um das Aufzeigen von potentiell systematischen Leerstellen.

Den Ausgangspunkt für das Vorhaben bildet eine erste Liste an Namen von Autor_innen, die von den Projektmitgliedern entwickelt wurde. Langfristig ist das Projekt so angelegt, dass aus der DH-Community Namen hinzugefügt werden können und das Portal so langsam wächst. Basierend auf den Namen werden Werke sowie weitere relevante Informationen in verschiedenen Archiven identifiziert. Zur Zeit werden dazu vor allem vier Quellarchive genutzt: *VIAF (*https://viaf.org/*)* als Personenquelle, *TextGrid* und *Das Deutsche Textarchiv* als Textquellen, *Wikidata (https://www.wikidata.org)* als Quelle für grundlegende Daten über die Autor_in, und *Europeana (https://www.europeana.eu)* primär als Quelle für kontextuelle Bilddaten. Zur korrekten Identifikation der Autor_innen in den jeweiligen Quellarchiven wurde eine Reihe von Heuristiken entwickelt, welche die Archive durchsuchen, potentielle Daten finden und diese gegen bereits vorhandene Daten abgleichen um sicherzustellen, dass die Daten auch der gleichen Person zugehören. Die Heuristiken machen hierbei nur Vorschläge, welche dann manuell akzeptiert oder abgelehnt werden müssen. Erste Erfahrungen aus der Praxis zeigen, dass eine Erstzuordnung über VIAF und dann die Integration von Daten aus *Wikidata* am besten funktioniert, da die dort vorhandenen Eckdaten für Suche und Validierung in anderen Archiven nützlich sind. Insbesondere listet *Wikidata* für die meisten Personen unterschiedliche Namenschreibweisen auf, was den Sucherfolg erhöht. Gerade Autorinnen wechseln über ihre Lebensspanne bisweilen mehrfach die Namen, meist durch Heirat, was ein Problem in der Eindeutigkeit ihrer Identifikation darstellt. Zusätzlich haben sich die groben Lebensdaten als nützlich zur Disambiguation der Suchergebnisse herausgestellt.

Die *Wikidata*-Daten und die Metadaten der Suchergebnisse aus anderen Archiven werden im Projekt gespeichert, aber die eigentlichen Daten (Texte, Bilder, Scans, …) werden nicht dupliziert, sondern es wird nur auf sie verwiesen. Dadurch entsteht natürlich die Möglichkeit, dass die Inhalte des Projekts und der Quellarchive divergieren und dadurch Unklarheit und Unsicherheit im Umgang mit den Daten entsteht. Um dem entgegen zu wirken, werden für alle Daten und Metadaten detaillierte Provenance-Informationen gespeichert, damit jederzeit nachvollzogen werden kann, was die jeweilige Ursprungsquelle ist.

Basierend auf den derart identifizierten Daten, wird dann das Online-Portal generiert. So vorhanden, wird dann im Rahmen des zweiten Projektziels für alle Autor_innen eine Liste der bekannten Werke geführt – unabhängig davon, ob und in welcher Form diese digitalisiert sind. Daraus generiert das Portal eine Reihe von Statistiken, welche einen Überblick darüber geben, in welchem Grad die Werke einzelner Autor_innen, bzw. der Gesamtbestand, digital bereits aufgearbeitet sind. Diese Statistiken sind auch über das Suchsystem zugänglich, es ist also möglich, zum Beispiel nach Autor_innen zu suchen, welche während eines gewissen Zeitraums an einem bestimmten Ort gewirkt haben. Potentiell können darüber Verbindungen von Autor_innen in Form von Netzwerken erkennbar werden. Dies unterstützt Geisteswissenschaftler_innen nicht zuletzt in der Identifikation potentiell interessanter Forschungsfragen. Parallel dazu werden für alle Daten und Metadaten maschinenlesbare Versionen bereitgestellt. Dies unterstützt das dritte Projektziel, da die Daten des Portals nahtlos in digitale Arbeitsabläufe der DH-Forschung integriert werden können.

Um die nicht-wissenschaftliche Öffentlichkeit anzusprechen (Johnson 2008), bietet das Portal eine Reihe von Funktionalitäten an. Es ist bekannt, dass es Nicht-Experten schwer fällt, erfolgreich zu suchen (Geser 2004; Wilson & Elsweiler 2010) und sie eine Präferenz für Browsing haben (Walsh et al. 2018). Daher entwickelt das Projekt eine Reihe von browsing-basierten Schnittstellen. Unter anderem „ein Werk/eine Autorin des Tages", welches entweder zufällig oder basierend auf Lebensdaten der Autorin ausgewählt und den Nutzern als Impuls vorgeschlagen wird. Auch sollen die Werke, basierend auf ihren Themen, automatisch gruppiert werden, damit Benutzer_innen durch die daraus entstehende Themenstruktur stöbern können. Zusätzlich wird das Portal eine Lesefunktion für jene Textdokumente anbieten, welche in den Quellarchiven in maschinenlesbarer Form vorhanden sind. Ziel ist es dabei nicht, eine Schnittstelle zur wissenschaftlichen Arbeit mit den Texten zu bieten, sondern eine Komponente, mit denen Texte wie ein gedrucktes Buch gelesen werden können.

Ausblick

Die gewählte Lösung eines (weiteren) Portals birgt natürlich die Frage, wie macht man es sichtbar und warum sollten Nutzer_innen es nutzen? Die Sichtbarkeit innerhalb der DH-Community soll über Beiträge in Konferenzen und Zeitschriften erreicht werden. In einem ersten Pilotversuch wurde das Portal in der universitären Lehre zum Gegenstand eines Seminars mit Studierenden der germanistischen Literaturwissenschaft gemacht, darüber stellt sich idealerweise perspektivisch ein Multiplikatoreneffekt ein. Der Zugriff auf Nutzer_innen aus der nicht-wissenschaftlichen Öffentlichkeit ist natürlich wesentlich schwieriger. Es ist aber so, dass soziale Medien von den unterrepräsentierten Gruppen oft stark genutzt werden und wir sehen das als die primäre Methode, um eine breitere Sichtbarkeit des Projekts zu erreichen (McLean and Maalsen 2013).

Die Problematik des Bias im Kanon kann natürlich nicht vom Projekt direkt gelöst werden. Unser Ziel ist es vielmehr, einen ersten Schritt zu unternehmen, um die Sensitivität für die Kanonfrage in den DH zu unterstützen und einen kritischen Diskurs zur Frage, welche Texte in welcher Form eigentlich digitalisiert werden, bzw. darüber hinaus, woran digitale literaturwissenschaftliche Forschung erfolgt, bzw. erfolgen kann, zu fördern. Durch eine breitere Aufstellung der für DH-Forschung genutzten Daten wäre entsprechend zu hoffen, dass sich der mutmaßlich bisher unbewußte Bias der Daten reduziert. Momentan wird digitale Forschung tendenziell an einem verengten und homogenen Kanon deutscher Literatur betrieben, während literarische Werke jenseits einer ‚männlichen' Auswahl, teils durch fehlende Digitalisate, teils durch mangelnde Sichtbarkeit, kaum berücksichtigt werden. Damit werden die Bemühungen der einzelnen Fachdisziplinen um Heterogenität und Diversität konterkariert, denn es betrifft nicht nur das Schreiben von Autorinnen, sondern ein breiteres Spektrum an unterrepräsentierten Gruppen. Langfristig ist

das Ziel des Projekts, sich selbst unnotwendig zu machen, insofern der Bias in den digitalen Quellarchive behoben ist, aber bis dahin will das Projekt die Leerstellen sichtbarer und greifbarer machen.

Bibliographie

Braam, Hans / Lutz Hagstedt (2017): „Lyrische Wunderkammern der ‚Sattelzeit': Gedichtsammlungen als Instrument bürgerlicher Kanonstiftung", in: *German Life and Letters* 70.1: 79-99.

Brinker-Gabler, Gisela (1978): *Deutsche Dichterinnen vom 16. Jahrhundert bis zur Gegenwart*. Frankfurt am Main: Fischer Taschenbuch Verlag.

Booth, Alison (2008): „Orlando: Women's Writing in the British Isles from the Beginnings to the Present", 725-734.

Connell, Sarah / Flanders, Julia / Keller, Nicole Infanta / Polcha, Elizabeth / Quinn, William Reed (2017): „Learning from the Past: The Women Writers Project and Thirty Years of Humanities Text Encoding", in: *Magnificat Cultura i Literatura Medievals*4: 1-19.

Derrida, Jacques (2009): „Dem Archiv verschrieben", in: Knut Ebeling/Stephan Günzel (eds.): *Archivologie. Theorien des Archivs in Philosophie, Medien und Künsten*. Berlin: Kadmos 29–60.

Evanson, Blaine Hill (2003): *The Sophie Digital Library of Early Women's Research*: A Blueprint for Mentored Undergraduate Online Research.

Fleck, Ludwig (1980): *Entstehung und Entwicklung einer wissenschaftlichen Tatsache:* Einführung in die Lehre von Denkstil und Denkkollektiv. Frankfurt am Main: Suhrkamp.

Foucault, Michel (1990): *Archäologie des Wissens*. Frankfurt am Main: Suhrkamp.

Geyken, Alexander (2013): „Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv", in: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. Hafemann, Ingelore: 221-234.

Geser, G. (2004): „Resource discovery – position paper: Putting the users first", pp. 7–12.

Hilger, Stephanie M. (2015): *Gender and Genre:* German women write the French Revolution. Newark: University of Delaware Press.

Johnson, A. (2008): "Users, use and context: supporting interaction between users and digital archives. What Are Archives?", in: *Cultural and Theoretical Perspectives:* A Reader. 145-164.

Kosnik, Abigail De (2016): *Rogue Archives:* Digital Cultural Memory & Media Fandom. Cambridge, Mass.: MIT Press.

Kuhn, Thomas S. (1996): *The Structure of Scientific Revolutions*. Chicago: UCP.

McLean, Jessica / Maalsen, Sophia (2013): "Destroying the joint and dying of shame? A geography of revitalised feminism in social media and beyond", in: *Geographical Research*51(3): 243-256.

Schmunk, Stefan / Funk Stefan E. (2016): „Das DARIAH-DE-und das TextGrid-Repositorium: Geistes-und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern", in: *Bibliothek Forschung und Praxis*, 40(2): 213-221.

Schulz, Marion (2008): „Die ‚virtuelle Heimat' deutscher Schriftstellerinnen: DaSinD online", in: *Zeitschrift für Germanistik*: 370-372.

Sylvester-Habenicht, Erdmute (2009): *Kanon u. Geschlecht:* eine Re-Inspektion aktueller Literaturgeschichtsschreibung aus feministisch-genderorientierter Sicht. Sulzbach/T.: Helmer.

Walsh, David / Hall, Mark Michael / Clough, Paul / Foster, Jonathan (2018): „Characterising online museum users: a study of the national museums liverpool museum website", in: *International Journal on Digital Libraries:*1-13.

Wilson, Max L. / Elsweiler, D. (2010): „Casual-leisure searching: the exploratory search scenarios that break our current models", in: *Proceedings of HCIR*. 28-31.

Winko, Simone (1996): „Literarische Wertung und Kanonbildung", in: Heinz Ludwig Arnold/Heinrich Detering (eds.): *Grundzüge der Literaturwissenschaft*. München: dtv 585-600.

Ziolkowski, Theodore (2009): „Zur Politik der Kanonbildung", in: Robert Charlier/Günther Lottes: *Kanonbildung:* Protagonisten und Prozesse der Herstellung kultureller Identität. Hannover: Wehrhahn 33-50.

Differenz und Ähnlichkeit in der computergestützten Filiation von Renaissancemusik. Zur datenbasierten Evaluation von Substitutionsmodellen mithilfe von Surrogatdaten

Plaksin, Anna
plaksin@maxweberstiftung.de
Max Weber Stiftung, Deutschland

Zur Filiation von Musik um 1500

Die Filiation als Methode zur Rekonstruktion der Überlieferungslinien eines Textes wird in der Renaissancemusikforschung nicht nur im Rahmen der Edition zur Bewertung von Lesarten verwendet, sondern auch in repertoiregeschichtlichen Studien. So diente sie als methodischer Grundpfeiler sowohl in Atlas' Studie (Atlas 1975) zum Cappella Giulia Chansonnier als auch in Urchueguías Studie (Urchueguía 2003) zu Messvertonungen des Siglo de oro. In dem sich aus den spezifischen Herausforderungen dieses Repertoires resultierenden methodischen Diskurs zu den Voraussetzungen und dem Potential der Filiation in der Renaissancemusikforschung, tritt insbesondere die starke hermeneutische Prägung zutage. Be-

zogen auf ein Material, bei dem eine bloße Recensio zumeist nur zu Variantenträgern führt, wird der Examinatio ein besonderer Stellenwert zugesprochen. (Just 1983: 130) In dieser Konsequenz ist die Filiation von Musikquellen insbesondere vom Begriff der Signifikanz geprägt – auf den Punkt gebracht von Margarent Bent:

> „It has often been said that manuscripts – and variants – should be weighed and not counted. Statistical counts of readings tell us nothing unless it is clear that the versions are stemmatically independent. However, although the strongest evidence for relating sources comes from variants that are not only shared but ‚significant'[…]" (Bent 1981: 307)

Dass eben diese starke Fokussierung auf die Signifikanz von Lesarten ein tieferes inhaltliches Verständnis des zu untersuchenden Textes erfordert, stellt damit ein wesentliches Charakteristikum dar. So beruht bereits der Variantenbegriff in der Musikphilologie auf der Unterscheidung zwischen substantiellen – d.h. Tonhöhe und -dauer betreffenden – und akzidentiellen Varianten. (Feder 1987: 60f.) Werden über methodentheoretische Beiträge hinaus noch kommentierte Stemmata der New Josquin Edition konsultiert, lassen sich wiederkehrende Argumentationsmuster beobachten:

Die inhaltliche Gewichtung von Lesarten folgt zumeist einer klaren Hierarchie. Die Klassifikation einer Lesart als Fehler, Variante – im Sinne einer Abweichung von Tonhöhe und/oder Rhythmus – oder als minor variant hat einen erheblichen Einfluss auf die ihr zugesprochene Beweiskraft. So zeigt sich, dass üblicherweise ein komplexes Geflecht aus kleineren und größeren Fehlern wie auch Varianten gebildet wird, das argumentativ gegeneinander abgewogen wird. Wird hierbei einer Lesart Leitcharakter zugesprochen, übertrifft deren Beweiskraft immer die der, stellenweise zahlreichen, anderen Befunde. Steht infolgedessen die Direktionalität der Überlieferung zur Diskussion, basieren die Argumentationsmuster zumeist auf Konzepten wie der Lectio difficilior, der musikalischen Plausibilität von Überlieferungsrichtungen, oder Vorannahmen über Eigenschaften des Archetypus, zumeist begleitet von ästhetischen Werkansprüchen – so werden beispielsweise Hyparchetypen in ein Stemma eingefügt um falsche Lesarten des Archetypus zu vermeiden. In letzter Konsequenz kann beobachtet werden, dass zwei Stemmata derselben Überlieferung erheblich abweichen können, wenn unterschiedliche Vorannahmen zugrunde gelegt werden.

Mensuralnotation als mehrdeutiges Zeichensystem

Gerade in Bezug auf die Musiküberlieferung vor 1600 erscheinen einige dieser Befunde aus der Perspektive einer Digitalen Musikwissenschaft als erstaunlich. So sind nicht nur Implikationen bezüglich der Stabilität des Werkbegriffs zu hinterfragen. In dem Bewusstsein über die prinzipielle Zeichenhaftigkeit von Musiknotation und eine dieser inhärenten semantischen Mehrdimensionalität, rückt auch das semiologische Gefüge der Überlieferungsform an sich, die Mensuralnotation, in den Fokus.

So weist Selfridge-Field bereits in ihrem Sammelband zu Musikkodierungsformaten auf die verschiedenen Kontexte hin, die eine musikalische Information beschreiben kann. (Selfridge-Field 1997: 7) Im Kodierungsformat der Music Encoding Initiative (MEI) kommt diesen Kontexten, übernommen aus der Standard Music Description Language, in Form von semantischen Domänen eine zentrale Rolle zu. Damit sind Informationen über Höhe und Dauer eines Tones vornehmlich der logischen Domäne zuzuordnen, während die visuelle Ebene den graphischen Befund, die gesturale Domäne die praktischen Ausführung bzw. klanglichen Ebene und die analytische Domäne Annotationen und Analysen umfasst. Betrachtet man nun Mensuralnotation auf dieser Basis, tritt insbesondere ihre strukturelle Ambiguität hervor. So ist gerade das Verhältnis zwischen einem Zeichen und dessen Bedeutung in der logischen Domäne nicht klar einander überführbar. Vielmehr können sich kontextbedingt Bedeutungen ändern, wenn das mensurale Gesamtgefüge die Dauer eines einzelnen Tons beeinflussen kann, oder im Falle der Alternation von Tonstufen der notierte Befund Leser*innen voraussetzt, die auch implizite Alterationen als solche zu erkennen vermögen. Darüber hinaus spiegeln sich auch Entwicklungsprozesse der Musiktheorie in der Notation wider, insbesondere im Falle von Mensur- und Proportionszeichen. Hier ergeben im äußersten Fall nicht nur verschiedene Symbole dasselbe Resultat auf der logisch-konzeptionellen Ebene. Gleichzeitig ist es möglich, dass ein und dasselbe Zeichen unterschiedliche Bedeutungen haben kann und man lediglich anhand musikalischer Gesichtspunkte abzuwägen vermag, welche der Möglichkeiten zu den geringsten strukturellen Schwierigkeiten führt. Vor diesem Hintergrund erscheint es damit geradezu als konsequent, dass Dumitrescu und van Berchum mit Blick auf die Edition des Occo Codex im CMME-Projekt die Existenz von nicht-substantiellen Varianten grundsätzlich infrage stellen:

> „The extent to which these can be considered 'non-substantive' is questionable: the positioning of line breaks, for instance, will have an effect on an editor's interpretation of the duration of manuscript accidentals, or stem direction may actually have an effect on rhythm in certain notational styles (as in some brands of 14th-century notation)." (Dumitrescu / van Berchum 2009: 143)

Filiation als Distinktion von Differenz

Im hier umrissenen Ansatz soll die Perspektive eröffnet werden, Repertoirestudien zur Renaissancemusik durch den Einsatz computergestützter Analyseverfahren auf eine breitere Basis, jenseits von Fallstudien, zu stellen. Entsprechend der seit den 1990er Jahren erprobten Übernahme von Verfahren der bioinformatischen Phylogenie, wird auch hier ein solcher Ansatz verfolgt. Doch während beim Alignment von Musik bisher insbesondere Ansätze verfolgt wurden, die explizit repräsentationsbedingte Abweichungen zu minimieren suchten, oder auf Retrieval-Szenarien ausgelegt sind, ergibt sich bei der computergestützten Filiation eine andere Konstellation. Vielmehr lässt sich diese als Methode beschreiben, mit der eine Gruppe ähnlicher Texte entsprechend ihrer Differenz in Relation gebracht werden soll. Die Anforderung an einen Prozess lässt sich somit in der spezifischen Distinktionsfähigkeit verorten. Damit stehen die Kodierung, die Datenaufbereitung sowie die Analyse sowohl in Hinblick auf die Leitthemen der traditionellen Filiation als auch in der Auseinandersetzung mit dem Material vor grundsätzlichen Fragen: Welche Rolle spielt das Verhältnis von visuellem Quellentext und dessen

logisch-konzeptioneller Ebene? Wie kann das dem Leseprozess inhärenten Interpretationsniveau methodisch so behandelt werden, dass ein schlüssiger Untersuchungsprozess möglich wird? In Bezug auf die Filiation als Verfahren, das sich per se Textzeugen widmet, stellt sich somit die Frage, wie die Integrität dieser auch bei der Analyse einer maschinenlesbaren Repräsentation gesichert werden kann.

Basierend auf einem Konzept von Kodierung als Datenerhebung, wurde an einem begrenzten Korpus von Quellen – Überlieferungen von Josquins *Missa D'ung aulte amer* und der damit verwandten Motette *Tu solus qui facis mirabilia* – ein Ansatz gewählt, der begonnen mit der Kodierung das Verhältnis von visuellem und logisch-konzeptionellem Befund in den Blick nimmt. Im Rahmen der Verfahrensentwicklung wurden diese beiden semantischen Domänen als Parametersets formalisiert, die als Grundlage für ein Alignment des gewählten Materials dienen. In diesem Zuge wurde, da sich der ausschließliche Vergleich der Ergebnisse mit traditionellen Stemmata aufgrund konkurrierender Ansichten über die rekonstruierte Überlieferungsgeschichte als nur wenig überzeugend erwies, eine datenbasierte Evaluationsmethode entwickelt.

Datenbasierte Evaluation von Substitutionsmodellen

Die wesentliche Herausforderung bestand hierbei in den noch äußerst unzureichenden Erfahrungen im Alignment von Renaissancemusik. Es wurden deshalb differenzbasierte Substitutionsmodelle und ein globales Sequenzalignment verwendet, obwohl statistische Theorien auf ähnlichkeitsbasierten Modellen und lokalen Alignments aufbauen.[1] Ohne zumindest basales Vorwissen über Substitutionsprozesse in der Renaissancemusik, sind die qualitativen Anforderungen an ähnlichkeitsbasierte Substitutionsmatrizen allerdings nicht erfüllbar. Indem gefordert ist, dass der Erwartungswert negativ und gleichzeitig mindestens ein positiver Wert im Scoringmodell möglich sein soll (Altschul 1991: 556), verlangt dies letztendlich die Festlegung einer neutralen Ähnlichkeit: Die Stelle auf der Skala, an der die Ähnlichkeit nicht groß genug ist, um einen positiven Wert zuzulassen, aber die gleichzeitig nicht different genug ist, um mit einem negativen Wert quantifiziert zu werden. Doch auch wenn eine derartige Zuweisung nicht vorgenommen wird, kann noch immer ein globales Sequenzalignment auf der Basis eines distanzbasierten Modells verwendet werden. Da in diesem Fall aber keine Modelle angewendet werden können, die auf Maximal Segment Scores[2] beruhen, wurde ein rein datenbasierter Ansatz zur Evaluation gewählt.

Der hier vorgestellte Ansatz baut auf dem Konzept des Vergleichs mit einem Zufallsmodell auf. Im Zentrum steht das Verfahren der Surrogatdatenanalyse, einem Verfahren aus der Zeitreihenanalyse (Theiler u.a. 1992: 77-78). Dabei werden basierend auf realen Reihen künstliche Daten erzeugt, bei denen gezielt ausgewählte Eigenschaften randomisiert und im Nachgang mit den Originaldaten in einem Hypothesentest verglichen werden können. Im konkreten Fall wird durch Shuffling realer Sequenzen ein Szenario erzeugt, in dem sich die vorliegenden Daten vollständig durch einen unabhängig und identisch verteilten Zufallsprozess beschreiben lassen. Hierbei wird im Wesentlichen der Grundannahme gefolgt, dass sich die Ähnlichkeit zweier Sequenzen in deren Binnenstrukturen manifestiert, also in der konkreten Reihenfolge, in der die Elemente innerhalb von Sequenzen angeordnet sind. Indem diese Binnenstrukturen in den geshuffelten Sequenzen zerstört werden, bleibt die Auftrittswahrscheinlichkeit der Elemente identisch, allerdings werden sämtliche Korrelationen mit der Reihenfolge, in der diese Elemente auftreten, zerstört:

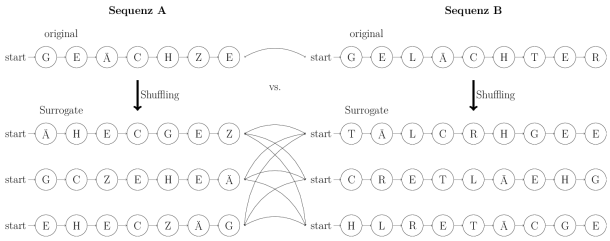


Abbildung 1: Zerstörung der Binnenstrukturen durch Shuffling.

Die Abweichung zwischen den Alignments von realen Sequenzen und Alignments von aus diesen erzeugten randomisierten Sequenzen kann hierdurch quantifiziert und als Grundlage für eine Evaluierung unterschiedlicher Substitutionsmodelle genutzt werden. In dem Fall, in dem durch einen Binnenvergleich ähnlicher Sequenzen diese entsprechend ihrer Differenz in Relation zu setzen sind, kann die Distinktionsfähigkeit von verschiedenen Ähnlichkeits- bzw. Differenzniveaus als zentrale Anforderung an ein Substitutionsmodell formuliert werden.

Diese Distinktionsfähigkeit kann im Rahmen eines Versuchsaufbaus überprüft werden, in dem eine Zahl von Vergleichen anhand der erwarteten Ähnlichkeit in Gruppen eingeteilt werden, die die Differenzniveaus widerspiegeln. So bilden beinahe identische Sequenzen eine Gruppe, etwas weniger ähnliche eine weitere etc. Dazu kommen Gruppen von Vergleichen, die als gerade so noch ähnlich bewertet werden, und wiederum Vergleiche mit erwarteter Unähnlichkeit – zusätzlich nimmt eine Gruppe von Vergleichen mit einem völlig anderen Stück die Funktion einer Kontrollgruppe ein. Für all diese Gruppen kann nun der Abstand des Vergleichs zwischen den beiden Originalsequenzen und den Surrogaten ermittelt werden.

Dieser Aufbau ermöglicht es, Mindestkriterien für die Eignung eines Substitutionsmodells zu formulieren. So ist es beispielsweise naheliegend, dass der Abstand zwischen den Originaldaten und den Surrogaten in der Kontrollgruppe möglichst nicht signifikant sein sollte. Ebenso sollte definiert werden, zwischen welchen beiden Gruppen nach Möglichkeit die Signifikanzgrenze liegen sollte, um die so gewünschte Sensitivität festzulegen. Auch kann basierend auf der vorgenommenen Gruppierung eine Varianzanalyse durchgeführt werden, um den Zusammenhang zwischen der vorgenommenen Gruppierung und dem Abstand zwischen Originaldaten und Surrogaten zu überprüfen. Basierend auf den zuvor formulierten Mindestkriterien und der Entwicklung des Abstandes von Originalen und Surrogaten, können somit Modelle verglichen und eine informierte Entscheidung über die Eignung in einem gewählten Einsatzszenario getroffen werden.

Fazit

Letztendlich kann konstatiert werden, dass die Rekonstruktion von Überlieferungsprozessen immer auf Vorannahmen beruht, egal welches Verfahren angewendet wird. Im Rahmen der traditionellen Stemmatologie führen Annahmen über die Bedingungen des Überlieferungsprozesses oder die Gestalt des Archetypus zu mitunter deutlichen Abweichungen in Stemmata, die ohne eine begleitende Erläuterung nicht zu eruieren sind. Automatisierte Verfahren basierend auf dem Alignment von Sequenzen können hierzu eine fruchtbare Ergänzung darstellen. Bei diesen stellt die Wahl des Substitutionsmodells einen wesentlichen Einflussfaktor dar. Indem allerdings gezielt Analyseparameter in Substitutionsregeln überführt werden, kann deren Einfluss überprüfbar gemacht werden. Wenn so konkurrierende Modelle an identischen Daten verglichen werden, wird deren Einfluss auf die Rekonstruktion von Beziehungen zwischen Textzeugen überprüfbar. (vgl. Plaksin 2019). Die Methode, das Verhalten eines Substitutionsmodells anhand des Vergleichs von echten und künstlich erzeugten Daten sichtbar zu machen, dient hierbei als Evaluationsverfahren und stellt damit einen wesentlichen Bestandteil in der Modellentwicklung dar.

Fußnoten

1. Während ein globales Alignment zwei vollständige Sequenzen einbezieht, während in einem lokalen Alignment Teile von Sequenzen möglichst optimal aneinander ausgerichtet werden.
2. Der Ähnlichkeitsscore des längsten Segments in einem lokalen Sequenzalignment. (Altschul 1991: 556)

Bibliographie

Altschul, Stephen F. (1991): "Amino acid substitution matrices from an information theoretic perspective" in: *Journal of Molecular Biology* 219.3: 555–565 10.1016/0022-2836(91)90193-a.

Atlas, Allan W. (1975): *The Cappella Giulia Chansonnier. (Rome, Bibliotheca Apostolica Vaticana, C.G. XIII.27)*. Brooklyn, NY: Inst. of Mediaeval Music, 1975.

Bent, Margaret (1981): "Some criteria for establishing relationships between sources of late-medieval polyphony" in: Fenlon, Iain (ed.): *Music in medieval and early modern Europe.* Cambridge: Cambridge Univ. Press: 295-317.

Berger, Karol (1987): *Musica ficta. Theories of accidental inflections in vocal polyphony from Marchetto da Padova to Gioseffo Zarlino.* Cambridge: Cambridge Univ. Press.

Boorman, Stanley (1981): "Limitations and extentions of filiation technique" in: Fenlon, Iain (ed.): *Music in medieval and early modern Europe.* Cambridge: Cambridge Univ. Press: 319–346.

Busse Berger, Anna Maria (1993): *Mensuration and proportion signs. Origins and evolution.* Oxford monographs on music. Oxford: Clarendon Press.

Dumitrescu, Theodor / van Berchum, Marnix (2009): "The CMME Occo Codex Edition. Variants and Versions in Encoding and Interface" in: Stadler, Peter / Veit, Joachim (eds.) *Digitale Edition zwischen Experiment und Standardisierung.* Beihefte zu editio. Tübingen: Niemeyer: 129–146.

Feder, Georg (1987): *Musikphilologie. Eine Einführung in die musikalische Textkritik, Hermeneutik und Editionstechnik.* Einführungen. Darmstadt: Wiss. Buchgesellschaft.

Field, Andy / Miles, Jeremy / Field, Zoë (2012): *Discovering statistics using R*. London: Sage.

ISO/IEC DIS 10743: *Standard Music Description Language (SMDL)*. https://www.lim.di.unimi.it/IEEE/SMDL/INDEX.HTM [letzter Zugriff 20. September 2019].

Josquin / Noblitt, Thomas (ed.) (1997): *Masses based on secular polyphonic songs: Critical commentary.* New Josquin Edition 7. Utrecht: Vereniging voor Nederlandse Muziekgeschiedenis.

Josquin / Blackburn, Bonnie (ed.) (2003): *Motets on nonbiblical texts 2: Critical commentary.* New Josquin Edition 22. Utrecht: Vereniging voor Nederlandse Muziekgeschiedenis.

Just, Martin (1983): "Zur Examinatio von Varianten" in: Finscher, Ludwig (ed.): *Datierung und Filiation von Musikhandschriften der Josquin-Zeit.* Wolfenbütteler Forschungen. Wiesbaden: Harrassowitz: 129–152.

Kranenburg, Peter van (2010): A Computational Approach to Content-Based Retrieval of Folk Song Melodies. PhD. Universiteit Utrecht. http://hdl.handle.net/20.500.11755/8436a210-ceeb-4c66-ba02-ffe5c7e66a42 [letzter Zugriff 20. September 2019].

Mongeau, Marcel / Sankoff, David (1990): "Comparison of Musical Sequences" in: *Computers and the Humanities* 24.3: 161–175. http://www.jstor.org/stable/30200223 [letzter Zugriff 26. September 2019].

The Music Encoding Initiative: *An Introduction to MEI.* https://music-encoding.org/resources/introduction.html [letzter Zugriff 19. September 2019].

O'Hara, Robert / Robinson, Peter (1993): "Computer-assisted Methods of Stemmatic Analysis" in: Blake, Norman Francis / Robinson, Peter (eds.): *The Canterbury Tales Project occasional papers.* Oxford: Office for Humanities Communication: 53–74.

Plaksin, Anna (2019): *Modelle zur computergestützten Analyse von Überlieferungen der Mensuralmusik. Empirische Textforschung im Kontext phylogenetischer Verfahren*. Diss., Technische Universität Darmstadt.

Selfridge-Field, Eleanor (ed.) (1997): *Beyond MIDI. The Handbook of Musical Codes.* Cambridge, Mass. [u.a.]: MIT Press.

Theiler, James u. a. (1992): "Testing for nonlinearity in time series. The method of surrogate data" in: *Physica D: Nonlinear Phenomena* 58.1-4: 77–94. 10.1016/0167-2789(92)90102-S.

Urchueguía, Cristina (2003): *Die mehrstimmige Messe im „goldenen Jahrhundert". Überlieferung und Repertoirebildung in Quellen aus Spanien und Portugal; (ca. 1490 - 1630).* Würzburger musikhistorische Beiträge, Bd. 25. Tutzing: H. Schneider.

3D-Rekonstruktion als Werkzeug der Quellenreflexion

Messemer, Heike
heike.messemer@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Clados, Christiane
clados@staff.uni-marburg.de
Philipps-Universität Marburg, Deutschland

Ausgangslage

3D-Rekonstruktionen tragen zu einem besseren Verständnis bestimmter Aspekte des kulturellen Erbes bei. Da sie bei der Präsentation des Themas behilflich sind, spielen sie eine wichtige Rolle als Instrument zur „Übersetzung" wissenschaftlicher Daten, um sie der Öffentlichkeit zugänglich zu machen.

Spätestens seit der London Charter (2009) und den Seville Principles (2011) ist die Verwendung digitaler Rekonstruktionen als wissenschaftliches Werkzeug und zur Visualisierung wissenschaftlicher Erkenntnisse in der Wissenschaftsgemeinde anerkannt. Allerdings formulieren beide Chartas nur Rahmenbedingungen für die Erstellung und Anwendung digitaler Rekonstruktionen in Wissenschaft, Forschung und Vermittlung. Auch aufgrund einer zunehmenden Vielfalt von Visualisierungstechniken scheint man heute weiter denn je von allgemeingültigen Regeln, Prinzipien und Normierungen in den Visualisierungsmethoden entfernt zu sein. Besonders zurückhaltend wird mit wenigen Ausnahmen (Grellert/Svenshon 2010; Pfarr 2010; Bruschke/Wacker 2016) bis heute die Dokumentation von 3D-Rekonstruktionen betrieben, obgleich besonders die London Charter (2009: 8-9) dieser im vierten Leitsatz besondere Aufmerksamkeit widmet.[1]

Ziel

In diese Problematik ist die Präsentation der folgenden zwei 3D-Rekonstruktionen zu verorten, die einen Beitrag zum Leitsatz 4 der London Charter leisten soll, eine digitale Rekonstruktion umfassend und nachvollziehbar zu dokumentieren. Die beiden Objekte zeigen einen hohen Grad an Unterschiedlichkeit, sodass ein breites Spektrum an Quellen, Dokumentationsmethoden und Visualisierungsmodi abgedeckt und Spielräume auf den Ebenen von Thema, Technologie und Visualisierungsstil genutzt werden können. Der Schwerpunkt liegt nicht allein auf der Datenaquirierung und Quellendokumentation, sondern auf der eigentlichen Entscheidungsfindung anhand der Quellenreflexion und dem Erkenntnisgewinn durch die fertiggestellte Rekonstruktion. Ziel ist es, die Bandbreite bisheriger Dokumentationsformen zu erweitern und sie an nicht herkömmlichen Objekttypen zu erproben.

Methodik

Die Dokumentationsmethode folgt der von Mieke Pfarr-Harfst und Marc Grellert (2016), die die Erstellung einer 3D-Rekonstruktion mit drei Wissensarten konnotieren: direkt im 3D-Modell abgebildetes Wissen, Kontextinformationen zum Modell und Informationsmehrwert, der sich aus dem 3D-Rekonstruktionsprozess ergibt (Pfarr-Harfst/Grellert 2016: 40-41). Die Idee, aus dem eigentlichen Prozess einer 3D-Rekonstruktion Erkenntnis zu erlangen, ist bereits von Forte und Siliotti (1997: 13) benannt worden. Ebenso bedeutend sind Überlegungen zu Rekonstruktionen, die ohne Befund sind (Grellert/Svenshon 2010).

Aufbauend auf diesen Ansätzen, sind mit der Erstellung der im Folgenden vorgestellten 3D-Rekonstruktionen drei Fragen verknüpft:

- Inwieweit sind die oben genannten Dokumentationsformen für Objekte mit eingeschränkter Befund-und Quellenlage anwendbar?
- In welchen Momenten der Visualisierung ergibt sich ein Erkenntnismehrwert?
- Auf welchen Ebenen findet Quellenreflexion statt?

Anwendungsbeispiele

Ruine der Sophienkirche, Dresden

Im Rahmen der in der Nachwuchsforschungsgruppe HistStadt4D erfolgten Entwicklung eines interaktiven 4D-Browsers, der ein digitales Modell der Stadt Dresden von ca. 1850 bis zur Gegenwart umfasst, wurde die heute nicht mehr existierende Sophienkirche mit der Open-Source-Software *SketchUp* 3D-rekonstruiert (Dewitz et al. 2019: 410).[2] Mit der Zeit als vierten Dimension ist es das Ziel des Stadtmodells, zu zeigen, welche baulichen Veränderungen sich in diesem Zeitraum an der im 13. Jahrhundert errichteten Sophienkirche vollzogen (Schreier/Lauffer 2014). Ende des Zweiten Weltkriegs brannte die Sophienkirche komplett aus; Gewölbe und Pfeiler stürzten später ein, stand hielten die Umfassungsmauern, Turmstümpfe sowie der Helm des südlichen Turms (Abb. 1). In diesem zerstörten Zustand, der sich im Laufe der Zeit weiter veränderte, prägte die Ruine der Sophienkirche das Dresdner Stadtbild bis zu ihrem Abriss 1964.[3]

Bei der digitalen Rekonstruktion von Ruinen kann nicht auf ein bewährtes Darstellungsrepertoire zurückgegriffen werden, denn 3D-Rekonstruktionen von nicht mehr existierenden Bauwerken, deren Zustand als Ruine modelliert wird, sind kaum Gegenstand von Forschungsprojekten.[4]

Abbildung 1: Ruine der Sophienkirche in Dresden nach der Zerstörung im Zweiten Weltkrieg, SLUB/Deutsche Fotothek, Walter Hahn, nach 1945.

Für den 4D-Browser sollten einfache Geometriemodelle, die das Bauwerk nur von außen zeigen, ohne Texturen oder modellierten Bauschmuck erstellt werden. Als Grundlage für die digitale Rekonstruktion diente eine Abbildung des Grundrisses, der den baulichen Zustand nach 1864 zeigt, da sich an dieser Grundstruktur bis zum Abriss nichts verändert hat (Schreier/Lauffer 2014: 11, Abb. 8). Zudem wurden historische Fotografien in der Bilddatenbank der Deutschen Fotothek herangezogen, die die Erscheinungsweise der Sophienkirche seit Ende des 19. Jahrhunderts gut dokumentieren.[5] Allerdings zeigt sich, dass das Bauwerk vornehmlich aus Westen und Südwesten fotografiert wurde. Der Chorbereich findet sich – sofern er überhaupt abgebildet ist – vor allem auf Fotos des kriegszerstörten Dresden, wodurch dessen Rekonstruktion erschwert wird (Abb. 2).

Abbildung 2: Sophienkirche aus Richtung Nordosten (links, SLUB/Deutsche Fotothek, Walter Möbius, 1934) und Südosten (rechts, SLUB/Deutsche Fotothek, Detail: 06.11.1951).

Im Rekonstruktionsprozess zeigten sich weitere Unsicherheiten im Wissen: Der Zustand der Ruine veränderte sich bis zum Abriss und ist fotografisch nur lückenhaft dokumentiert. Die Fotografien zeigen zudem teils nur bestimmte Abschnitte der Kirche, sodass nicht immer ersichtlich ist, welche Teile noch intakt und welche bereits zerstört waren. Auch die Datierung der Fotografien ist teils nicht gesichert. Daher ist es nicht möglich, die Ruine in einem 3D-Modell zu einem bestimmten Zeit**punkt** darzustellen. Vielmehr zeigt das resultierende digitale Modell Facetten der Ruinenwerdung über einen Zeit**raum** hinweg, summiert also einzelne Elemente und lässt einen hypothetischen Ruinenbau entstehen. Es stellt sich die Frage, welche Möglichkeiten für die Darstellung von Ruinen im 3D-Modell bestehen und welche sich hierfür am besten eigenen. Als Grundlage zur Diskussion wurden vier unterschiedliche Darstellungsweisen entwickelt (Abb. 3).[6]

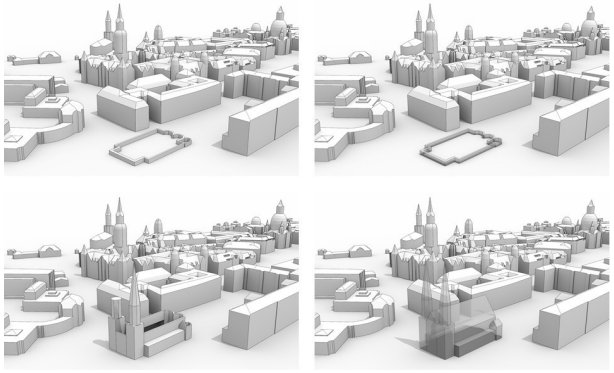


Abbildung 3: Digitale 3D-Rekonstruktion der Ruine der Sophienkirche in Dresden: verräumlichter Grundriss opak und transparent (oben), zerstörtes Bauwerk (unten links), vollständige Kirche transparent (unten rechts), Heike Messemer und Jonas Bruschke, *HistStadt4D,* 2019.

Nasenschmuck, Museo Del Oro, Bogotá

Fallbeispiel zwei ist die Rekonstruktion eines altkolumbianischen Nasenschmucks (22,1 cm x 21,1 cm) der Calima-Malagana-Kultur des 5. Jahrhunderts (Abb. 4 links). Die Herkunft des Objektes, das sich heute im Museo Del Oro, Bogotá, befindet (Reg. 016637), wird mit Cauca-Tal angegeben, obwohl es aus keiner kontrollierten Grabung stammt. Es ist aus hochwertigem Gold, Smaragden und Pyriten gearbeitet und besteht aus drei Hauptteilen, 115 Metallplättchen und acht Röhrchen, die durch Metallringe beweglich miteinander ver-

bunden sind. Das rechte Segment weist einen alten Bruch auf, der schon in der vorspanischen Zeit mit Klammern repariert wurde. Angaben zur Goldlegierung sowie die Maße konnten dem Online-Katalog des Museums entnommen werden. Ziel im Hinblick auf die Visualisierung war zum einen, einen Nasenschmuck zu rekonstruieren, wie er kurz nach der Fertigstellung ausgesehen haben mag: ohne Bruch, durch fehlende Teile ergänzt, mit unverbrauchter materialer Oberfläche. Weitere Ziele waren das Objekt aufgrund seiner Größe in Bezug zu einem menschlichen Kopf zu zeigen, um das Zusammenspiel von Gesicht und Nasenschmuck auszutesten sowie aufgrund der Vielteiligkeit des Objektes eine Simulation der sich bewegenden Plättchen zu erstellen (Abb. 4 rechts).

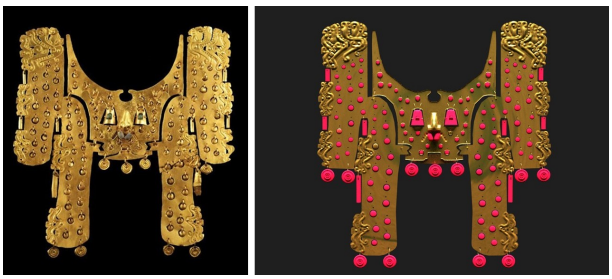


Abbildung 4: Nasenschmuck im Museo Del Oro, Bogota (links), und digitale Rekonstruktion mit Anzeige der beweglichen Teile in rot (rechts), © Christiane Clados.

Das Objekt war vor Ort nicht zugänglich, jedoch eine große Anzahl von Fotografien im Bestand des Museo Del Oro. Mit einer durch die General Public Lizence lizensierten 3D-Grafiksuite konnte das Objekt maßstabsgetreu rekonstruiert werden. Der gute Erhaltungszustand und einige wenige Vergleichsobjekte erlaubten die sachgerechte Ergänzung fehlender Plättchen. Mehr Unsicherheiten zeigt die Rekonstruktion des Objekts in seinem Kontext, d.h. im Moment, wenn es getragen wurde. Mehrere Figurinen derselben Datierung veranschaulichen den Tragemodus des Nasenornaments. In Bezug zu einem menschlichen Kopf gesetzt wurde ersichtlich, dass der Nasenschmuck einstmals große Teile des Kopfes verdeckte (Abb. 5 links). Die Vielteiligkeit des Ornaments impliziert, dass sich während des Tragens alle Teile in ständiger Bewegung befanden. In einer Animation wurde deswegen der Bewegungsverlauf derselben simuliert und mit einer Tonspur unterlegt, die der Klangkulisse sich bewegender Metallplättchen entspricht. Ferner wurde ein Licht-Umfeld erzeugt, das Tageslicht und tropische Vegetation imitiert, die vom Metall reflektiert werden (Abb. 5 rechts).[7]

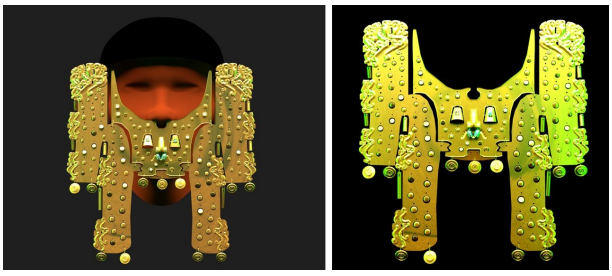


Abbildung 5: Nasenschmuck im Kontext (links), und mit Einzelteilen in Bewegung, Sonne und Vegetation reflektierend (rechts), © Christiane Clados.

Schlussfolgerungen

Die schon bestehenden Dokumentationsmethoden sind auch auf Objekte mit eingeschränkter Befund- und Quellenlage anwendbar. Jedoch ergibt sich aufgrund der Unterschiedlichkeit von Objekttyp, Visualiserungsziel und Arbeitsumfeld ein breites Feld von Ausgangslagen. Ein Erkenntnismehrwert zeigte sich in unterschiedlicher Weise in den Rekonstruktionsprozessen der beiden hier vorgestellten Fallbeispiele:

Da die Quellenlage zur Sophienkirche sehr heterogen beziehungsweise lückenhaft war, konnte das 3D-Modell zur Ruine nicht einen bestimmten Zeit**punkt** darstellen. Vielmehr entstand eine Rekonstruktion eines Zeit**raums**. In Projekten zur 3D-Rekonstruktion von historischer Architektur wird ein solcher Umstand bislang nicht thematisiert, obwohl die zur Modellierung verwendeten Quellen nicht immer einen identischen Zeitpunkt darstellen.[8] Zukünftig ist es notwendig diese Diskrepanz zu reflektieren. Die Erkenntnis führte in der Folge zur Untersuchung von Ruinendarstellungen in digitalen Rekonstruktionen. Auch in diesem Kontext wurde deutlich, dass es weiterer Forschung bedarf: Ruinen, die zum Zeitpunkt der 3D-Rekonstruktion nicht mehr existieren, werden bisher kaum 3D-modelliert. Im Fall der Sophienkirche konnte deutlich gemacht werden, dass es zur Veranschaulichung der Stadtgestalt Dresdens in den 1950er-Jahren erforderlich ist, den Zustand der Kirche als Ruine im 3D-Model darzustellen.

Im Falle des Nasenschmucks lässt sich ein deutlicher Erkenntnismehrwert in der Bewegungssimulation erkennen, insbesondere hinsichtlich der Licht- und Klangeffekte, die das Ornament im Moments des Tragens und im Zusammenspiel mit der Sonne entfaltet. Die Simulation der Bewegung des virtuellen Nasenschmucks verdeutlicht audiovisuelle Effekte, die aus konservatorischen Gründen am originalen Nasenschmuck nur bedingt aufgezeigt werden können. Die Simulation ist Interpretationshilfe, ein Umstand, der bislang, wenn überhaupt, nur in wenigen Publikationen zur Diskussion gestellt wird. Der rekonstruierte Bewegungsverlauf von mehr als 200 metallenen Elementen, ihre Reflexion im Sonnenlicht, bei gleichzeitiger Reflexion einer tropisch-grünen Umgebung lässt den Schluss zu, dass das Objekt so konzipiert wurde, dass es erst in Interaktion mit Licht und Wind seine volle Funktion erfüllte. Die Rekonstruktion bestätigt damit die Aussagen spanischer Chronisten, die von ähnlichen Licht- und Klangeffekten altkolumbianischer Goldornamente berichteten. Bei der Veranschaulichung des Tragemodus zeigt sich zudem, dass der Mund des Trägers beim Sprechen ganz bewusst verdeckt bleibt, was ihm eine nicht-menschliche Wirkung verlieh (Clados 2019).

Das Vorhandensein/Fehlen von Befunden/Quellen für die digitale Rekonstruktion eines Artefakts hat unweigerlich Auswirkungen auf den Rekonstruktionsprozess: Inwiefern sich dies auch visuell im 3D-Modell wiederspiegelt, hängt stark davon ab unter welcher Maßgabe die finale Visualisierung gestaltet sein soll. So hätten Teile der Sophienkirche, die fotografisch gut dokumentiert sind, mit architektonischen Details (wie Fenster, Portal, Bauschmuck) zwar ansatzweise rekonstruiert werden können, allerdings war für das 4D-Modell, in das die Ruine implementiert wurde, nur ein einfaches Geometriemodell notwendig. Im Falle des Nasenschmucks war es durch die Unzugänglichkeit nicht möglich, eine Klangprobe der Plättchen zu nehmen.

Die hier vorgestellten 3D-Rekonstruktionsprojekte und die darin verhandelten Fragen verdeutlichen, dass es 10 Jahre

nach Formulierung der London Charter an der Zeit ist, deren Leitsätze neu zu denken, sie an die Weiterentwicklung digitaler Technologien und neuen Fragestellungen anzupassen und mit Blick auf die Vielfältigkeit der rekonstruierten Objekte zu konkretisieren.

Forschungsförderung

Die dem Abschnitt 4.1 zugrundeliegende Forschung ist Teil der Aktivitäten der Nachwuchsforschungsgruppe *HistStadt4D*, die vom Bundesministerium für Bildung und Forschung im Rahmen der Fördervereinbarung Nr. 01UG1630 gefördert wird.

Fußnoten

1. Vgl. London Charter, Fassung 2.1, Februar 2009, online verfügbar: http://www.londoncharter.org/ (letzter Zugriff: 11.09.2019).
2. Für die 3D-Rekonstruktion mit *SketchUp* zeichnet Heike Messemer verantwortlich, für die Erstellung der finalen Visualisierung und Einbettung in den 4D-Browser Jonas Bruschke. Webseite von *HistSadt4D*: http://4dbrowser.urbanhistory4d.org/ (letzter Zugriff: 18.09.2019).
3. Wie historische Fotos aus der Zeit nach der Zerstörung der Kirche zeigen, wurde die Ruine zugemauert, der Helm des südlichen Turms entfernt, das Portal abgerissen (Lerm 2001: 85; Schreier/Lauffer 2014: 12).
4. Eine der wenigen Ausnahmen umfasst die schematisch angelegte 3D-Modellierung der Ruine des buddhistischen Klosters Sompur Mahavihara in Paharpur, Bangladesch (Rashid/Rahaman 2016).
5. Vgl. Webseite der Deutschen Fotothek: http://www.deutschefotothek.de/ (letzter Zugriff: 18.09.2019). Es existieren keine historischen Aufrisse, die die Sophienkirche nach ihrem umfassenden Umbau zeigen.
6. Um die Möglichkeiten für die Darstellungsweisen von Ruinen als 3D-Modelle weiter zu untersuchen, wurde eine Nutzerstudie zu deren Wahrnehmung im 4D-Browser durchgeführt, die noch ausgewertet wird (Dewitz et al. 2019: 410). Es wird erwartet mit der Studie grundlegende Tendenzen aufzuzeigen, wie die einzelnen Darstellungsweisen im Kontext von wissenschaftlichen 3D-Modellen bewertet werden.
7. Video zur digitalen Rekonstruktion des Nasenschmucks in Bewegung: https://www.youtube.com/watch?v=uJg3D-W_XMk0 (letzter Zugriff: 26.09.2019).
8. Beispielsweise werden für die Rekonstruktion von Synagogen auch Zeitzeugenberichte herangezogen, die in der Regel nicht die Eindrücke zu einem bestimmten Zeitpunkt wiedergeben, sondern eine Zeitspanne umfassen. Das resultierende 3D-Modell reflektiert dies allerdings nicht visuell. Auch in Berichten zur 3D-Rekonstruktion wird dieser Umstand nicht aufgegriffen. Vgl. Grellert 2007.

Bibliographie

Bruschke, Jonas / Wacker, Markus (2016): "Simplifying Documentation of Digital Reconstruction Processes. Introducing an Interactive Documentation System", in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (Hrsg.): *3D Research Challenges in Cultural Heritage II. How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage.* Lecture Notes in Computer Science, Bd. 10025. Cham: Springer 256-271.

Clados, Christiane (2019): "The Golden Ones", in: *Amerindian Socio-Cosmologies between the Andes, Amazonia and Mesoamerica. Toward an anthropological understanding of the Isthmo-Colombian Area,* Hrsg. Ernst Halbmayer. Routledge Verlag (im Druck).

Denard, Hugh (Hrsg., 2009): *London Charter. For the Computer-based Visualisation of Cultural Heritage. Draft 2.1. 7 February 2009*. London *http://www.londoncharter.org/* [letzter Zugriff 20.09.2019].

Dewitz, Leyla / Kröber, Cindy / Messemer, Heike / Maiwald, Ferdinand / Münster, Sander / Bruschke, Jonas / Niebling, Florian (2019): "Historical Photos and Visualizations: Potential for Research", in: *27th CIPA International Symposium "Documenting the past for a better future"* 405-412 (= Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Bd. XLII-2W15)

Forte, Maurizio / Siliotti Alberto (Hrsg., 1997): *Virtual Archaeology. Re-creating Ancient Worlds*. New York: Harry N. Abrams Inc.

Grellert, Marc (2007): *Immaterielle Zeugnisse. Synagogen in Deutschland. Potentiale digitaler Technologien für das Erinnern zerstörter Architektur*, Bielefeld: transcript.

Grellert, Marc / Svenshon, Helge (2010): „Rekonstruktion ohne Befund?", in: *Befund und Rekonstruktion. Mitteilungen der Deutschen Gesellschaft für Archäologie des Mittelalters und der Neuzeit* 22: 189-198.

Lerm, Matthias (2001): *Abschied vom alten Dresden. Verluste historischer Bausubstanz nach 1945*. Rostock: Hirnstorff.

o.A. (2011): *Principles of Seville. International Principles of Virtual Archaeology*. o.O. *http://smartheritage.com/wp-content/uploads/2015/03/FINAL-DRAFT.pdf* [letzter Zugriff 20.09.2019).

Pfarr, Mieke (2010): *Dokumentationssystem für Digitale Rekonstruktionen am Beispiel der Grabanlage Zhaoling, Provinz Shaanxi, China*, Dissertation, Technische Universität Darmstadt.

Pfarr-Harfst / Grellert, Marc (2016): „The Reconstruction – Argumentation Method: Proposal for a Minimum Standard of Documentation in the Context of Virtual Reconstructions", in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 6th International Conference, EuroMed 2016, (Proceedings Part 1)* 39-49.

Schreier, Dietmar / Lauffer, Manfred (2014): „Die Sophienkirche", in: Landeshauptstadt Dresden, die Oberbürgermeisterin (Hrsg.): *Verlorene Kirchen. Dresdens zerstörte Gotteshäuser. Eine Dokumentation seit 1938* (2. Auflage). Dresden: 8-13.

Ein Schritt zurück: Distinktive Eigenschaften im deutschsprachigen Drama

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

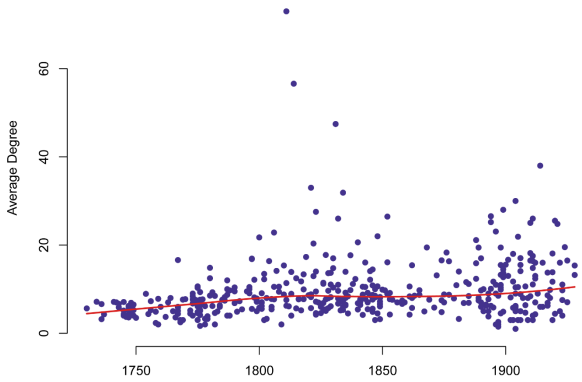


Abbildung 1: *Average Degree* im historischen Verlauf; *LOESS* Kurve.

Einführung in die Fragestellung

Topic Modeling und soziale Netzwerkanalysen zählen zu den etabliertesten quantitativen Methoden innerhalb der *Computational Literary Studies* (vgl. Du 2019; Trilcke u.a. 2016, Jannidis 2017: 148–161). Beide gelten insbesondere als geeignet, um große literarische Korpora auf Muster – semantischer oder struktureller Art – zu untersuchen, die durch lineares Lesen nicht oder nur schwer zu greifen sind (vgl. Willand 2017: 86). Auf diese Weise versprechen sie literarhistorische Entwicklungslinien aufzuzeigen, die die traditionelle, auf symptomatische Beispiele fußende Literaturgeschichtsschreibung zu übersehen neige (vgl. Moretti 2017: 6f.; Jockers 2013: 9). Epistemologisch sind die durch *Topics* oder Netzwerkmaße erschlossenen Textmodelle jedoch vom ursprünglichen literarischen Text deutlich zu unterscheiden. So konstatieren Peer Trilcke und Frank Fischer, dass es sich um andere ‚epistemische Dinge' handle (vgl. Trilcke, Fischer 2018). Denn beide Methoden reduzieren den literarischen Text auf spezifische Eigenschaften: im Fall der Netzwerkanalyse zumeist auf Figurenbeziehungen, die durch Knoten und Kanten repräsentiert werden; im Fall des *Topic Modeling* auf Wortkollokationen, die als *Topics* interpretiert werden. Positiv ließe sich diese Reduktion als Notwendigkeit der Operationalisierung fassen, die ein – in Abhängigkeit der Fragestellung gewähltes – theoretisches Konzept messbar machen soll (vgl. Moretti 2013). So könnte sich, wie Franco Moretti postuliert, die Handlung eines Textes näherungsweise als die Summe an Interaktionen der Figuren im Netzwerk operationalisieren lassen (vgl. Moretti 2011: 2–4). Die Zentralitätsmaße von Figurennetzwerken ermöglichen dann einen auf quantitativen Werten basierenden Vergleich der literarischen Texte.

Ein mögliches Anwendungsszenario zeigt *Abbildung 1*. Sie stellt das *Average Degree* (durchschnittlicher Grad) von insgesamt 443 Dramennetzwerken dar. *Degree* ist ein simples Zentralitätsmaß, das für jede Dramenfigur bemisst, mit wie vielen anderen sie im Verlauf des Stücks interagiert. Trilcke und Fischer nutzen eine ähnliche Darstellung, um literarhistorische Erkenntnisse zu bestätigen. Basierend auf der Vorstellung, dass *Degree* ein „Indikator für soziale Komplexität" (Trilcke, Fischer: 2018) sein könnte, formulieren sie die geläufige These, dass das Drama seit Mitte des 18. Jahrhunderts gesellschaftliche Modernisierungsprozesse widerspiegeln würde.

Ziel dieses Beitrags ist es jedoch, einen Schritt hinter solche makroanalytischen Befunde zurück zu treten. In einem vorgelagerten Arbeitsschritt möchte ich erörtern, welche quantitativ erfassbaren Merkmale dramatischer Texte überhaupt geeignet sind, um eine literarhistorische Einordnung und Unterscheidung der Dramen vorzunehmen. Anders formuliert sollen also die Kriterien ermittelt werden, die mit Blick auf die Entstehungszeit der Dramen unterscheidungstragend sind. Zu diesem Zweck dient im vorliegenden Fall eine einfach gehaltene Klassifikationsaufgabe. Ließen sich die dramatischen Texte erfolgreich ihrem Veröffentlichungszeitraum zuweisen, könnte daraus auf die Kriterien rückgeschlossen werden, die den entscheidenden Beitrag zu dieser Klassifikation leisten. Daran anschließend wäre eine Rückübersetzung der ermittelten Merkmale denkbar – analog zur Operationalisierung –, die eine Interpretation anleiten könnten.

Korpus

Die folgenden Untersuchungen konzentrieren sich auf 443 deutschsprachige Dramen zwischen 1730 und 1930 aus dem *German Drama Corpus* (Fischer u.a. 2019). *Abbildung* 2 gibt einen Überblick über die zeitliche Verteilung der Dramen. Es handelt sich um ein recht heterogenes Korpus, das auf unterschiedlichen poetologischen Vorstellungen fußt, die wiederum an verschiedene Produktions- und Rezeptionsbedingungen geknüpft sein können. Sowohl versifizierte als auch in Prosa gehaltene Damen sind enthalten. Sehr kurze Stücke, in denen die Figurenrede weniger als 3000 Tokens umfasst, wurden für die Analysen entfernt, so dass die Länge der Stücke zwischen noch immer kurzen 3017 und längst nicht mehr ungekürzt auf der Bühne darstellbaren 146248 Tokens liegt (median: 22548; Standardabweichung: 13304). Die Zahl der auftretenden Figuren liegt zwischen zwei und 183 (median: 17; Standardabweichung: 19,2). Das Korpus enthält Stücke von insgesamt 166 Autor*Innen, darunter sowohl hochkanonische als auch heutzutage kaum noch wahrgenommene.

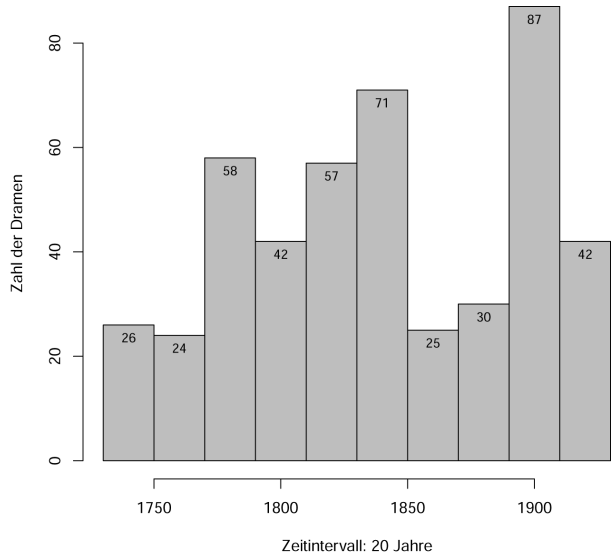


Abbildung 2: Korpusübersicht.

Methode

Die historische Verortung der dramatischen Texte fasse ich als basal gehaltene Klassifikationsaufgabe. Ziel der Klassifikation ist es, mittels maschineller Lernverfahren näherungsweise den Veröffentlichungszeitraum der Dramen zu bestimmen, um daran anschließend die Einflussfaktoren identifizieren und untersuchen zu können. Dazu greife ich auf Metadaten zurück, die Angaben zur Erstaufführung und zur Erstpublikation umfassen. Diese Metadaten werden genutzt, um jedes Drama einer von vier heuristisch gesetzten Zeitspannen zuzuordnen, die jeweils circa 50 Jahre umfassen: 1730–1785 (93 Dramen), 1786–1832 (116 Dramen), 1833–1881 (105 Dramen) und 1882–1930 (129 Dramen). Die dadurch entstehenden Zeiträume dienen als Zielpunkt der Klassifikation und orientieren sich an wichtigen literaturgeschichtlichen Zäsuren: Aufklärung, Goethezeit, Realismus und literarische Moderne (vgl. etwa Brenner 2011).[1] Als Features der Klassifikation nutze ich verschieden komplexe Netzwerk- und Zentralitätsmetriken sowie durch *Topic Modeling* trainierte *Topics*.

Basis der Netzwerk- und Zentralitätsmetriken sind Netzwerkgraphen, die auf Präsenz- bzw. Adjazenzmatrizen fußen. Knoten und Kanten repräsentieren hierbei Dramenfiguren und deren Interaktion, wobei Interaktion als das gemeinsame Sprechen innerhalb einer Szene operationalisiert ist (vgl. Trilcke 2013: 238f.). Eine Kante zwischen zwei Knoten wird also genau dann instanziiert, wenn die beiden fraglichen Figuren innerhalb derselben Dramenszene sprechen. Das bedeutet auch, dass verschiedene poetologische Vorstellungen von Akt und Aufzug sowie Szene und Auftritt, die im Verlauf der Dramengeschichte einem Wandel unterliegen, einen Eingang in die Graphen findet (vgl. etwa Vogel 2012). Für die Klassifikation nutze ich die folgenden acht Maße: *Degree, Weighted Degree, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality*, *Average Path Length*, *Clustering Coefficient* und *Density*.[2]

Topic Modeling gilt als Technik, die – in einem weiteren Sinn gefasst – semantische Strukturen in größeren Textkorpora zu identifizieren vermag (vgl. etwa Schöch 2017: 42). Die von mir verwendeten *Topics* wurden auf dem gesamten in *Abbildung 2* dargestellten Dramenkorpus trainiert, wobei die Figurenrede eines jeden Dramas nochmals in Segmente von je 1000 Tokens unterteilt ist. Vorab wurde das Wortmaterial auf Nomen, Adjektive und Vollverben beschränkt. Um die *Topics* zu trainieren, greife ich auf das von David M. Blei, Andrew Y. Ng und Michael I. Jordan (2003) vorgeschlagene probabilistische Modell *Latent Dirichlet Allocation* zurück. Für die maschinelle Klassifikation setze ich ein Modell mit 20 *Topics* (T1–T20) ein, das einen guten Kompromiss bietet zwischen der Interpretierbarkeit einzelner *Topics* und ihrer Eignung, die Texte diachron zu unterscheiden.

Das maschinelle Klassifikationsverfahren selbst nutzt den Algorithmus *Random Forest* (Ho 1995, Breiman 2001). *Random Forest* fügt mehrere unkorrelierte Entscheidungsbäume zusammen und berechnet mittels mathematischer Regression die Parameter. Beim Trainieren wurde das Korpus in zehn Segmente gegliedert (10- *fold cross validation*), wodurch verzerrte Ergebnisse durch die zufällige Verteilung von Trainings- und Testkorpus vermieden werden sollen.[3]

Ergebnisse

Tabelle 1 zeigt die Ergebnisse der Klassifikation anhand von drei Modellen. Zusätzlich zum Gesamtmodell, das Netzwerkanalysen und *Topic Modeling* zusammenführt, wurden die acht Netzwerkmetriken und die 20 *Topics* auch isoliert betrachtet. Die *Baseline* berechnet sich, angelehnt an die einleitenden Überlegungen, anhand des *Average Degree*.[4] Die Werte in *Tabelle* 1 verdeutlichen zweierlei: Einerseits erreicht bereits die *Baseline* annehmbare Ergebnisse. 309 Dramen werden anhand ihres *Average Degree* richtig klassifiziert. Andererseits scheint insbesondere das trainierte *Topic Model* zur Leistungsfähigkeit des gesamten Modells beizutragen. Letzteres zeigt sich mit einem F_1-Wert von 0.921 angemessen performant – lediglich 34 Dramen werden einem falschen Zeitraum zugewiesen. Wirklich überraschen kann dieser Umstand jedoch nicht, sind die heuristischen Klassifikationszeiträume doch recht groß gewählt. Kleiner gefasste Zeiträume führen das Modell hingegen recht schnell an seine Grenzen. Teilt man die durch das Textkorpus abgedeckte Dramengeschichte in feingliedrigere Segmente, beispielsweise in zehn Zeiträume von nurmehr 20 Jahren, sinkt der F_1-Wert auf 0.431 (Precision: 0.585, Recall: 0.451).

Tabelle 1: Modelle trainiert auf 443 Dramen; 10-*fold cross validation*, *SMOTE-sampling*.

	Precision	Recall	F_1
Baseline (*Avg. Degree*)	0.701	0.702	0.698
Netzwerkmetriken	0.829	0.821	0.814
Topic Model	0.899	0.903	0.899
Gesamtes Modell	0.921	0.925	0.921

Die in *Tabelle* 1 dargestellten Klassifikationsergebnisse erlauben nun einen Einblick in die Unterscheidungskraft der eingespeisten Features. *Abbildung* 3 zeigt die sogenannte *Feature Importance*. Sie vergleicht nach und nach die Leistungsfähigkeit des Modells, wenn jeweils eines der Features nicht beachtet wird. Die Abnahme an Performanz entspricht dann

der relativen Wichtigkeit des nicht einbezogenen Features für die Klassifikation. Die Abbildung verdeutlicht, dass das *Topic Model* – insbesondere die *Topics* 8 und 5 – einen starken Einfluss auf die Klassifikation nimmt. Doch bei weitem nicht jedes *Topic* (etwa T17, T7, T10) ist von solch großer Bedeutung. Als gewichtigste Netzwerkmetrik lässt sich die *Betweenness Centrality* identifizieren. Der durchschnittliche Grad (*Average Degree*) ist mittig platziert, wobei der Einfluss der weniger entscheidungstragenden Features insgesamt ähnlich gering ausfällt.

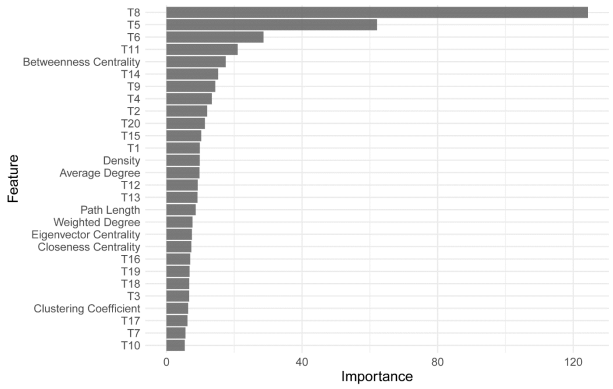


Abbildung 3: *Feature Importance* des Klassifikationsmodells.

Operationalisierung und Interpretation

Auf Basis dieser Daten lassen sich nun analog zu *Abbildung* 1 Werte berechnen und visualisieren, die die zeitliche Entwicklung der als relevant erscheinenden Merkmale darstellen, etwa von *Topic* 8 oder der *Betweenness Centrality*. Diese sollten – vertraut man der Klassifikation – einen höheren Aussagegehalt haben, als der zu Beginn diskutierte *Average Degree*.

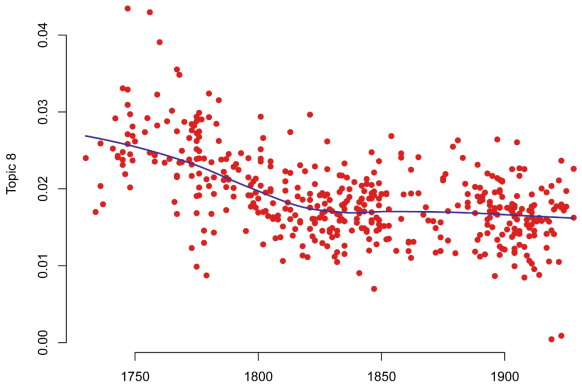


Abbildung 4: *Topic* 8 im historischen Verlauf, normalisiert nach Dramenlänge; *LOESS* Kurve.

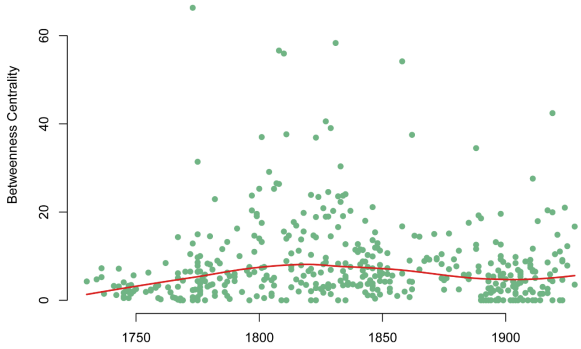


Abbildung 5: *Betweenness Centrality* im historischen Verlauf; *LOESS* Kurve.

Abbildung 4 zeigt die Frequenzen, mit denen sich die Wörter aus *Topic* 8 auf die einzelnen Dramen verteilen. Tatsächlich veranschaulicht die Visualisierung eine recht deutliche Entwicklung. Von 1730 ausgehend scheint *Topic* 8 bis etwa 1830 recht stark an Einfluss einzubüßen, ehe die Werte fortan auf einem stabilen Niveau bleiben. Betrachtet man die zehn ausschlaggebendsten Wörter des *Topics* – ‚liebe', ‚herz', ‚machen', ‚lassen', ‚sagen', ‚vater', ‚schwester', ‚sehen' und ‚weiß' –, lässt sich dieser Verlauf auch literaturgeschichtlich plausibilisieren. Zu einem großen Teil können diese Begriffe mit bürgerlichen Trauerspielen und Rührstücken in Verbindung gesetzt werden, die für die Dramengeschichte des 18. Jahrhunderts prägend sind.

Auch der in *Abbildung* 5 dargestellte diachrone Verlauf der *Betweenness Centrality* macht eine Entwicklung der Werte sichtbar. Die lokal gewichtete Regression veranschaulicht einen Höhepunkt zwischen 1810 und 1825. Die *Betweenness* Centrality bemisst, in welchem Maß ein Knoten im Netzwerk selbst zum Teil eines Pfades wird, also die indirekte Verbindung von zwei anderen Knoten ermöglicht (vgl. Newman 2010: 185–193). Auf Dramennetzwerke übertragen ließen sich dadurch Figuren identifizieren, die als Brückenfiguren agieren und voneinander getrennte Figurengruppen verbinden. Die Darstellung lässt sich somit als weiterer Indikator für die zu Beginn skizzierte These von Trilcke und Fischer lesen. Die zusehende Abkehr von der Regelpoetik in der zweiten Hälfte des 18. Jahrhunderts scheint eine komplexere Struktur der Dramen nach sich zu ziehen, die sich in den Netzwerkdaten wiederfinden lässt.

Die größte Schwierigkeit bei der Interpretation dieser Daten bleibt jedoch nach wie vor bestehen und ist den hier gezeigten Analysen vorgelagert. Es ist die Operationalisierung der Fragestellung, die zumeist mit einem großen konzeptuellen Aufwand verbunden ist (vgl. Gius 2019: 2f.; Reiter / Willand 2018). Denn unklar bleibt, wie sich Zentralitätsmetriken in einem Figurennetzwerk oder Wahrscheinlichkeitsverteilungen von Worthäufigkeiten zu literaturwissenschaftlichen Kategorien verhalten. Die bemessenen Werte müssten sich konzeptionell so rückübersetzen lassen, dass sie auch mit Blick auf spezifisch literaturwissenschaftliche Fragestellungen interpretiert werden können. Dass etwa die Handlung literarischer Texte nicht einfach durch ein Figurennetzwerk abzubilden ist, muss auch Moretti erkennen, weshalb er die Netzwerktheorie letztlich nurmehr als Vorstufe, als „beginning of the beginning" (Moretti 2011: 2) zu einer quantifizierbaren Handlung einordnet.

Fazit und Ausblick

Das vorgestellte multidimensionale Modell liefert sinnvolle Ergebnisse und kann den recht weit gefassten Veröffentlichungszeitraum deutschsprachiger Dramen mit angemessener Genauigkeit (F_1 -Wert 0.921) klassifizieren. Da die Entstehung neuer literarischer Epochen und Strömungen zumeist als fließender Prozess zu beschreiben ist, muss die hier vorgestellte Methode aber als Heuristik eingestuft werden, die vor allem zum Ziel hat, die entscheidungstragenden Merkmale der Klassifikation zu identifizieren. Erst dadurch bietet die Klassifikationsaufgabe Anschlusspotential für literarhistorische Studien. Für künftige Arbeiten erscheint es einerseits lohnend, eine metrische Vorhersage der Veröffentlichungsjahre zu erproben. Dadurch würde die Vorhersage deutlich an Präzision gewinnen. Andererseits würde es sich anbieten, neben *Topic Modeling* und Netzwerkanalysen auch stilometrische Maße in die Klassifikation zu integrieren. So könnte auch der Stil der Stücke – in einem quantitativen und damit weiten Sinn – Teil der Voraussage werden.

Fußnoten

1. Da literarische Epochen und Strömungen eine fließende Entwicklung nehmen, ergeben sich durch diese Setzung zwangsläufig Überschneidungen. So fallen beispielsweise Lessings *Emilia Galotti* (1772) und Schillers *Die Räuber* (1781), zwei schon strukturell sehr verschieden gebaute Stücke unterschiedlicher Strömungen, in die gleiche Klasse.
2. Für eine Übersicht über die verschiedenen Metriken vgl. Newman (2010): 168–204.
3. 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/. Caret bietet vielfältige Optionen für das *Preprocessing* und *Sampling* der Daten. Ich nutze die Methoden *center* und *scale* zur Kalibrierung und *SMOTE-sampling*, um die zahlenmäßige Ungleichverteilung der Dramen in den verschiedenen Zeiträumen auszugleichen (eine genauere Beschreibung eines ähnlichen Versuchsaufbaus findet sich in Krautter / Pagel / Reiter / Willand 2018: 19–29).
4. Die Leistungsfähigkeit einer *Majority Baseline* wäre aufgrund des *multiclass* Klassifikation sehr eingeschränkt (Precision: 0.291).

Bibliographie

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003). „Latent Dirichlet Allocation", in: *The Journal of machine Learning research* 3: 993–1022.

Breiman, Leo (2001): „Random Forests", in: *Machine Learning* 24 (2): 5–32.

Brenner, Peter J. (2011): *Neue deutsche Literaturgeschichte. Von ‚Ackermann' zu Günter Grass*. Berlin / New York: De Gruyter.

Du, Keli (2019): „A Survey on LDA Topic Modeling in Digital Humanities", in: *DH 2019. Conference Abstracts*. https://dev.clariah.nl/files/dh2019/boa/0326.html [letzer Zugriff 05. Januar 2019].

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): „Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor", in: *DHd 2019. Konferenzabstracts:* 194–197. DOI: 10.5281/zenodo.2596094 [letzer Zugriff 05. Januar 2019].

Gius, Evelyn (2019): „Computationelle Textanalysen als fünfdimensionales Problem: Ein Modell zur Beschreibung von Komplexität", in: *Litlab Pamphlet* 8. https://www.digitalhumanitiescooperation.de/pamphlet-8-computationelle-textanalysen/ [letzter Zugriff 05. Januar 2019].

Ho, Tin Kam (1995): „Random Decision Forests", in: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*: 278–282.

Jannidis, Fotis (2017): „Netzwerke", in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): *Digital Humanities: Eine Einführung.* Stuttgart: J.B. Metzler 148–161.

Jockers, Matthew L. (2013): *Macroanalysis. Digital Methods and Literary History.* Urbana / Chicago / Springfield: University of Illinois Press.

Krautter, Benjamin / Pagel, Janis / Reiter, Nils / Willand, Marcus (2018): „Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen", in: *Litlab Pamphlet* 7. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/06/p07_krautter_et_al-1.pdf [letzter Zugriff 05. Januar 2019].

Moretti, Franco (2011): „Network Theory, Plot Analysis", in: *Literary Lab Pamphlet* 2. https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzer Zugriff 05. Januar 2019].

Moretti, Franco (2013): „‚Operationalizing': or, the Function of Measurement in Modern Literary Theory", in: *Literary Lab Pamphlet* 6. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf [letzer Zugriff 27. September 2019].

Moretti, Franco (2017): „Patterns and Interpretation", in: *Literary Lab Pamphlet* 15. https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf [letzer Zugriff 05. Januar 2019].

Newman, Mark E. J. (2010): *Networks: An Introduction*. Oxford: UP.

Reiter, Nils / Willand, Marcus (2018): „Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse. Shakespeares natürliche Figuren im deutschen Drama des 18. Jahrhunderts, in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (eds.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin / Boston: De Gruyter 45–75.

Schöch, Christof (2017): „Gattungen des Kriminalromans: Ein quantitativer, Topic-basierter Zugang", in: Koch, Corinna / Schmitz, Sabine / Lang, Sandra (eds.): *Dialogische Krimianalysen. Fachdidaktik und Fachwissenschaft untersuchen aktuelle Repräsentationsformen des französischen Krimis*. Frankfurt a. M.: Peter Lang 37–64.

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.

Trilcke, Peer / Fischer, Frank (2018): „Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen", in: Huber, Martin / Krämer, Sybille (eds.): Sonderband 3 der ZfdG: Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegen-

stände und Methoden. http://www.zfdg.de/sb003_003 [letzter Zugriff 05. Januar 2019].

Trilcke, Peer / Fischer, Frank / Göbel, Matthias / Kampkaspar, Dario (2016): „Theatre Plays as ‚Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930", in: *DH 2016. Conference Abstracts*: 385–387.

Vogel, Juliane (2012): „Aus dem Takt: Auftrittsstrukturen in Schillers *Don Karlos*", in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 86 (4): 532–546.

Willand, Marcus (2017): „Hermeneutische Interpretation und digitale Analyse: Versuch einer Verhältnisbestimmung", in: Banki, Luisa / Scheffel, Michael (eds.): *Lektüren. Positionen zeitgenössischer Philologie*. Trier: WVT 77–98.

Erzählerische Spielräume. Medienübergreifende Erforschung von Narrativen im Mittelalter mit ONAMA

Nicka, Isabella
isabella.nicka@sbg.ac.at
Interdisziplinäres Zentrum für Mittelalter und Frühneuzeit, Universität Salzburg

Hinkelmanns, Peter
peter.hinkelmanns@sbg.ac.at
Interdisziplinäres Zentrum für Mittelalter und Frühneuzeit, Universität Salzburg

Landkammer, Miriam
miriam.landkammer@sbg.ac.at
Interdisziplinäres Zentrum für Mittelalter und Frühneuzeit, Universität Salzburg

Schwembacher, Manuel
manuel.schwembacher@sbg.ac.at
Interdisziplinäres Zentrum für Mittelalter und Frühneuzeit, Universität Salzburg

Zeppezauer-Wachauer, Katharina
Katharina.Wachauer@sbg.ac.at
Interdisziplinäres Zentrum für Mittelalter und Frühneuzeit, Universität Salzburg

Einleitung

Wer J.R.R. Tolkiens Erzählung von der Begegnung Bilbo Beutlins mit dem Feuerdrachen Smaug, der tief im Einsamen Berg einen gigantischen Schatz bewacht, las oder auf der Leinwand opulent inszeniert sah, hatte wohl mit großer Wahrscheinlichkeit das Gefühl, einer ähnlichen Geschichte irgendwann schon einmal begegnet zu sein. Und dies zu Recht, ist doch das Aufeinandertreffen eines Helden mit einem gefährlichen Drachen, das oftmals auf einen Kampf auf Leben und Tod hinausläuft, ein weitverbreitetes, lange tradiertes Narrativ, welches sich auch im Mittelalter großer Beliebtheit erfreute und dementsprechend häufig aufgegriffen wurde. So stellen sich Siegfried, Beowulf, Tristan, Georg und Lancelot – um nur einige wenige zu nennen – erfolgreich gefährlichen Drachen entgegen. Neben Beowulf ergeben sich besonders zu Siegfried explizite Verbindungen: Siegfried, der auch außerhalb des Nibelungenliedes in einer Vielzahl von Texten präsent ist, tötet in der *Völsunga Saga*[1] – hier den Namen Sigurd tragend – den Drachen Fáfnir, der in einer Höhle in der Wildnis haust, um in den Besitz des Drachenhortes zu gelangen. Richard Wagner hat diese Episode in seinem *Ring des Nibelungen*[2] aufgegriffen; J.R.R. Tolkien hat sich für *The Hobbit or There and Back Again*[3] davon inspirieren lassen.

Um nicht auf zufällige Entdeckungen von ähnlichen Narrativen angewiesen zu sein, sondern einen systematischen Vergleich der Strukturen und Bausteine von Erzähltem in der Literatur und in Bildern des Mittelalters zu ermöglichen, wurde das Projekt ONAMA – *Ontology of the Narratives of the Middle Ages* – ins Leben gerufen.

Forschungsstand

Die Erforschung von Narrativen hat in den Literaturwissenschaften eine lange Tradition, wenngleich narratologische Ansätze für Texte aus der Zeit des Mittelalters im Vergleich zu Texten der Neuzeit in geringerem Maße vorhanden sind (vgl. Störmer-Caysa 2007; Contzen/Kragl 2018). Für Erzählungen in (unbewegten) Bildern besteht hier hingegen Aufholbedarf, der unter anderem in der weit geringeren Präsenz erzähltheoretischer Forschungsansätze in der Kunstgeschichte und den Bildwissenschaften begründet ist (einen Überblick bietet Speidel 2018; konkret zu mittelalterlichen visuellen Medien vgl. Niehr 2015; Suckale 2009, Bd. 1: 427f.; Franzen 2002: 14–19). Die Ansätze der Intermedialitätsforschung bzw. Bild-Textforschung (vgl. u.a. Schellewald 2011 und Wolf 2017) wurden bis dato vor allem für Quellen genutzt, die per se schon unterschiedliche mediale Aspekte beinhalten (z.B.: illuminierte Handschriften). Übernahmen von Narrativen oder bestimmten Bausteinen eines Narrativs in unterschiedlichen Quellen bleiben dabei meist außen vor.

Die Methoden der Digital Humanities werden bis dato nur am Rande für die Erforschung von Narrativen genutzt. So entwickeln beispielsweise verschiedene Projekte zur Erforschung von Erzähltexten ontologische Repräsentationen narrativer Strukturen (z.B. Ciotti 2016, Khan et al. 2016). Narrativ-Ontologien, die auf die semantische Verknüpfung medial heterogener Quellen und Artefakte über Elemente der Erzählungen abzielen, sind als Recherche- und Explorationstools im Museums- und Medienarchivbereich angesiedelt (exemplarisch Damiano 2019 bzw. Damiano/Lieto 2013, Metilli et

al. 2019, Mulholland et al. 2004). Im Bereich der Germanistik wurden Konzepte für eine narratologische Textauszeichnung digitaler Corpora entwickelt (Dimpel 2019, Gius 2015). Modelle, die Spezifika des Erzählens in Bildern berücksichtigen, sind rar (Xu et al. 2017). Dies ist nicht zuletzt am Mangel an weiter verarbeitbaren Basisdaten für solche Analysen begründet.

Datengrundlage

ONAMA ist ein interdisziplinäres Joint Venture, welches sich auf die breite Datenbasis von zwei Langzeitprojekten aus dem Bereich der Digital Humanities stützt: einerseits die Mittelhochdeutsche Begriffsdatenbank (MHDBDB, Universität Salzburg) und andererseits die Bilddatenbank REALonline des Instituts für Realienkunde des Mittelalters und der frühen Neuzeit in Krems, welches ebenfalls Teil der Universität Salzburg ist.

In REALonline sind visuelle Medien unterschiedlicher Gattungen und Techniken, die schwerpunktmäßig vom 14.–16. Jahrhundert entstanden, in über 20.000 Datensätzen derart erfasst, dass alle semantischen Bestandteile eines Bildes und ihre Eigenschaften sowie Beziehungen zwischen diesen einzelnen Bildelementen dokumentiert werden (vgl. Matschinegg/Nicka et al. 2019, Matschinegg/Nicka 2018). Aktuell gibt es bereits über 1,2 Millionen semantische Annotationen in REALonline. In der MHDBDB bietet ein onomasiologisches Begriffssystem den Zugang zu derzeit über 650 Texten, die von Heldenepen über religiöse Kleindichtung bis hin zu Fabeln reichen (vgl. Hinkelmanns 2019, Dimpel/Zeppezauer-Wachauer/Schlager 2019). Die Annotationen der mehr als 10,5 Millionen tokens ermöglichen extensive semantische, morphologische, lexikalische und metrische Suchanfragen. Aus diesen beiden Datenpools werden im Projekt ONAMA exemplarisch mittelalterliche Narrative ausgewählt und an ihnen ein Modell für eine medienübergreifende Beschreibung von Handlungen, Aktanten, Settings und zeitlichen Strukturen entwickelt.

Methode

ONAMA zielt auf die formale Darstellung sowohl von transmedial fassbaren Bausteinen von Narrativen als auch von den jeweiligen Umsetzungen dieser Grundelemente in konkreten Bildern und Texten des Mittelalters ab. Die im Projekt erarbeitete Ontologie auf Basis der Web Ontology Language (OWL) bildet die Grundlage für den Vergleich von Narrativen. Dabei können Muster und Besonderheiten ihres Aufbaus durch Abfragen identifiziert werden, deren Ursachen und Funktionen in weiterer Folge untersucht werden können. Es wird damit weit mehr als nur der allgemeine „Plot" einer Geschichte oder eines Bilderzyklus erfasst. Die Entwicklung des ONAMA-Grundmodells und seine Verfeinerung sind dabei die ersten Schritte im Projekt. Wir definieren (wenn möglich in Anlehnung an bestehende Klassifkationssysteme wie Motif-Index [Birkhan/Lichtblau/Tuczay 2005-2010] oder ICONCLASS[4]) zunächst Narrativ-Konzepte (*concepts*) (siehe Abb.1). Diesen werden dann die jeweiligen Narrativ-Realisierungen (*realisations*) zugeordnet. Gemeint sind mit letzteren die Narrative in der Form, wie sie in dem zu annotierenden Werk (Bild oder Textstelle) tatsächlich vorkommen. Um in weiterer Folge sowohl bei *concepts* als auch bei *realisations* nach einzelnen Narrativelementen und ihren Kombinationen suchen zu können, nutzt das ONAMA-Modell das ursprünglich aus der Linguistik stammende Konzept semantischer Rollen. Damit wird ermöglicht, die Zusammensetzung und Art des Zusammenhangs zwischen Akteuren, Handlungen, Objekten und Settings für jedes einzelne Narrativ/jede Handlungseinheit zu spezifizieren. Die Handlungseinheiten werden in jeder Überlieferung zu Abfolgestrukturen verbunden.[5] Wo möglich, werden Verbindungen zwischen ONAMA und dem CIDOC Conceptual Reference Model[6] hergestellt. Am Ende des Entwicklungsprozesses steht die Publikation der Narrativ-Ontologie, die Anfang 2020 geplant ist. Im Rahmen der Annotation des ausgewählten Korpus wird daran weitergearbeitet und bei Bedarf werden weitere Adaptionen des Grundmodells veröffentlicht.

Obgleich sich die im Projekt bearbeiteten Beispiele aus einem Pool von deutschsprachigen Texten und mittelalterlichen Kunstwerken speisen, lässt sich die in ONAMA entwickelte Ontologie prinzipiell auch auf andere Sprachen und Medien hin erweitern. ONAMA erfasst die Ebene der Geschichte mit den Bausteinen *Handlung, Person, Objekt* und *Ort* und ist damit grundsätzlich als intermediales Modell angelegt.

Als Ausgangsmaterial haben wir einerseits mit dem „Trojanischen Krieg" einen spezifischen Erzählstoff herangezogen, der durch Realisierungen in beiden Datenbanken dokumentiert ist und sich somit gut für eine erzähltheoretische Auswertung unterschiedlicher Umsetzungen eines Narrativs in mehreren Versionen bzw. in Bild und Text eignet. Die Bilder stammen aus dem Cod. 2773 der Österreichischen Nationalbibliothek, einer reich illustrierten Prachthandschrift mit Guido de Columnis' *Historia destructionis Troiae*[7] in einer deutschen Übersetzung (Mitte des 15. Jh.); die literarischen Bearbeitungen des Trojastoffes sind Herborts von Fritslâr *Liet von Troye*[8] (um 1190–1200) sowie Konrads von Würzburg *Der Trojanische Krieg*[9] (letztes Viertel des 13. Jh.). Andererseits werden entlang eines bestimmten Motivs, das auch Überschneidungen zur Trojaliteratur aufweist – dem „Bekämpfen oder Zähmen wilder Tiere/Wesen" – unterschiedliche verbal oder bildlich überlieferte Narrative aus beiden Datenbanken ausgesucht und modelliert, um eine Datenbasis zur Untersuchung der konkreten sprachlichen oder bildlichen Umsetzungen dieser jeweils geschilderten Interaktionen im Kontext ihrer verschiedenen Einbettungen zu schaffen. Der Nutzen der digitalen Ontologie für die mediävistische Erforschung von Narrativen wird im Projekt anhand von Fallstudien evaluiert, die auf den generierten Daten basieren.

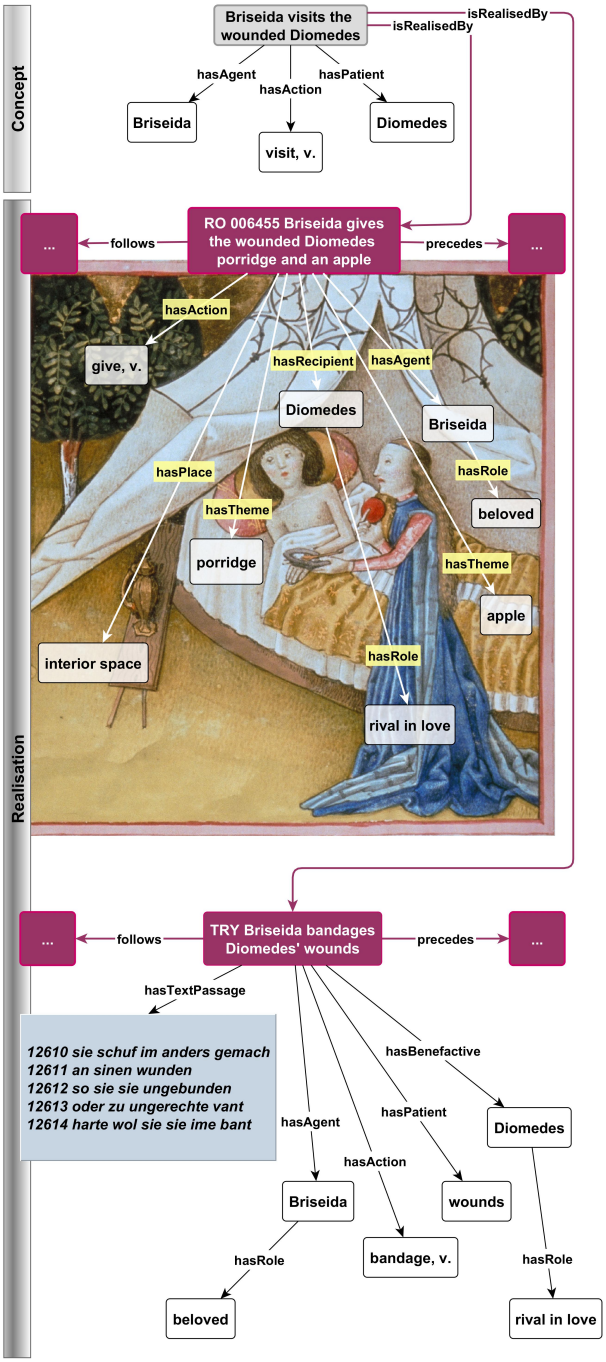


Abbildung 1: Beispiel für die Narrativ-Modellierung mit ONAMA auf Konzept- und Realisierungsebene. Briseida erhört Diomedes und besucht den durch Troilus im Kampf Verwundeten an seinem Krankenlager: eine in mittelalterlichen Adaptionen des Troja-Stoffkreises eingeführte Begebenheit, bildlich umgesetzt z.B. durch REALonline Archivnr. 006455 (Historia destructionis Troiae, Wien, Österreichische Nationalbibliothek, Cod. 2773, fol. 175v.), episch verarbeitet im Liet von Troye (MHDBDB-Text „TRY"). Durch semantische Rollen sind die jeweils beteiligten Entitäten mit dem Narrativ verbunden. Mit der Verbindung „hasRole" können hingegen über die unmittelbar dargestellte Handlung hinausreichende Rollen von Akteuren erfasst werden.

Forschungsfragen und Ergebnisse

Mit ONAMA entsteht eine sprach- und medienunabhängige Ontologie zur Erschließung mittelalterlicher Narrative, die als ein neues digitales Werkzeug althergebrachte und immer noch weit verbreitete fachliche Grenzen zwischen bildlicher und textlicher Überlieferung überwindet und so der Beantwortung interdisziplinärer sowie intermedialer Forschungsfragen dient. So interessiert sich ONAMA beispielsweise dafür, wie Narrative in Bild und Text realisiert beziehungsweise materialisiert werden; in welchem Kontext die materiellen Umsetzungen stehen und wie sich Wechsel von Medien und materiellen Informationsträgern auf das vermittelte Narrativ auswirken. Mit Hilfe der Narrativ-Ontologie können Bild- und Textquellen derart annotiert werden, dass Abfrageergebnisse sowohl Rückschlüsse auf Genese und Tradierung von Erzählkernen, Figurenkonstellationen, Handlungsmuster etc. im jeweiligen Medium als auch in der medienübergreifenden Zusammenschau ermöglichen.

Da die Nutzer*innen über das im Laufe des Projekts umgesetzte ONAMA-Frontend gleichzeitig auf umfangreiche Annotationen zu Narrativen in Bildern und Texten zugreifen können, werden Bezüge oder Unterschiede innerhalb der breit gefächerten Korpora zu mittelalterlichen Quellen in den beiden Datenbanken einfach identifizierbar. Die narrativen Muster, die Texten und Bildern inhärent sind, werden nach zeitgemäßen digitalen Standards annotiert, visualisiert und können damit besser empirisch bewertet werden. Darüber hinaus werden sämtliche im Rahmen von ONAMA generierten Daten auch für komplexe Abfragen via SPARQL zugänglich sein und der Scientific Community unter Creative Commons-Lizenz zur Verfügung gestellt, damit sie beispielsweise als Basis für Fragen zu Narrativen in anderen digitalen Korpora weiterverwendet werden können.

Im Rahmen des Vortrags werden sowohl das Projekt ONAMA (Laufzeit März 2019 – Februar 2021) vorgestellt, das aus Mitteln des Förderprogramms *go!digital* der Österreichischen Akademie der Wissenschaften finanziert wird, als auch erste Ergebnisse präsentiert.

Bibliographie:

Primärliteratur:

Herbort von Fritslâr: *Liet von Troye*. Hrsg. v. Karl Frommann (= Bibliothek der gesammten deutschen National-Literatur von der ältesten bis auf die neuere Zeit, Bd. 5). Quedlinburg / Leipzig: Basse 1837.

Konrad von Würzburg: *Der Trojanische Krieg*. Hrsg. v. Adelbert von Keller. Stuttgart: Litterar. Verein 1858.

The Saga of the Volsungs. The Icelandic Text according to MS Nks 1824 b, 4°. With an English Translation, Introduction and Notes by Kaaren Grimstad (= Biblioteca Germanica Series Nova Vol. 3) Saarbrücken: AQ-Verl. 2000.

J. R. R. Tolkien: *The Hobbit or There and Back again*. London: HarperCollins 2006.

Richard Wagner: *Der Ring des Nibelungen. Ein Bühnenfestspiel für drei Tage und einen Vorabend. Zweiter Tag: Siegfried.* Textbuch mit Varianten der Partitur. Hrsg. v. Egon Voss. Stuttgart: Reclam 2007.

Forschungsliteratur:

Birkhan, Helmut / Lichtblau, Karin / Tuczay, Christa (2005-2010): Motif-Index of the German Secular Narratives from the Beginning to 1400, 7 Bde., Berlin u. a. Online-Ausgabe im Verlag der Österreichischen Akademie der Wissenschaften (ÖAW), Wien 2009. http://hw.oeaw.ac.at/motifindex?frames=yes (06.11.2019)

Ciotti, Fabio (2016): "Toward a Formal Ontology for Narrative", in: *Matlit* 4 (1): 29–44 DOI: 10.14195/2182-8830.

Contzen, Eva von / Kragl, Florian (eds.) (2018): *Narratologie und mittelalterliches Erzählen.* Autor, Erzähler, Perspektive, Zeit und Raum (= Das Mittelalter: Beihefte 7) . Berlin / Boston: de Gruyter.

Damiano, Rossana (2019): "Investigating the Effectiveness of Narrative Relations for the Exploration of Cultural Heritage Archives", in: Papadopoulos, George Angelos / Samaras, George / Weibelzahl, Stephan / Jannach, Dietmar / Santos, Olga C. (eds.): *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. Larnaca, Cyprus, 09.-12.06.2019. New York: ACM Press 417–423.

Damiano, Rossana / Lieto, Antonio (2013): "Ontological representations of narratives. A case study on stories and actions", in: Finlayson, Mark A. / Fisseni, Bernhard / Löwe, Benedikt / Meister, Christoph (eds.): *2013 Workshop on Computational Models of Narrative*. CMN 2013. Hamburg, 04.–06.08.2013. Wadern: Schloss Dagstuhl - Leibniz-Zentrum für Informatik 76–93 http://drops.dagstuhl.de/opus/volltexte/2013/4149/pdf/p076-damiano.pdf [letzter Zugriff am 1. August 2019].

Dimpel, Friedrich Michael (2019): „Narratologische Textauszeichnung in Märe und Novelle. Mit Annotationsbeispielen und exemplarischer Auswertung von Sperber und Häslein durch MTLD und Sozialer Netzwerkanalyse", in: *Zeitschrift für digitale Geisteswissenschaften* 4. text/html Format. DOI: 10.17175/2016_012 .

Dimpel, Friedrich Michael / Zeppezauer-Wachauer, Katharina / Schlager, Daniel (2019): „Der Streit um die Birne: Autorschafts-Attributionstest mit Burrows' Delta und dessen Optimierung für Kurztexte am Beispiel der ‚Halben Birne' des Konrad von Würzburg", in: Bleier, Roman / Fischer, Franz / Hiltmann, Torsten / Viehhauser, Gabriel & Vogeler, Georg (eds.).: *Digitale Mediävistik* (= Das Mittelalter. Perspektiven mediävistischer Forschung; Band 24, Nr. 1) 71–90. DOI: 10.1515/mial-2019-0006.

Franzen, Wilfried (2002): *Die Karlsruher Passion und das „Erzählen in Bildern".* Studien zur süddeutschen Tafelmalerei des 15. Jahrhunderts. Berlin: Lukas Verlag.

Gius, Evelyn (2015): *Erzählen über Konflikte. Ein Beitrag zur digitalen Narratologie* (= Narratologia 46). Berlin / Boston: de Gruyter.

Hinkelmanns, Peter (2019): „Mittelhochdeutsche Lexikographie und Semantic Web: Die Anbindung der ‚Mittelhochdeutschen Begriffsdatenbank' an Linked Open Data", in: Bleier, Roman / Fischer, Franz / Hiltmann, Torsten / Viehhauser, Gabriel & Vogeler, Georg (eds.).: *Digitale Mediävistik* (= Das Mittelalter. Perspektiven mediävistischer Forschung; Band 24, Nr. 1) 129–141. DOI: 10.1515/mial-2019-0009 .

Hinkelmanns, Peter / Landkammer, Miriam / Nicka, Isabella / Schwembacher, Manuel / Zeppezauer-Wachauer, Katharina (in Vorbereitung): „Beyond the Plot. Der Vergleich mittelalterlicher Narrative im Semantic Web mit ONAMA", erscheint in: *Narrare-Producere-Ordinare*. New Approaches to the Middle Ages (agora. Wiener philologisch-kulturwissenschaftliche Studien/Vienna Philological and Cultural Studies). Wien: Praesens.

Khan, Anas Fahad / Bellandi, Andrea / Benotto, Giulia / Frontini, Francesca / Giovannetti, Emiliano / Reboul, Marianne (2016): "Leveraging a Narrative Ontology to Query a Literary Text", in: Miller, Ben / Lieto, Antonio / Ronfard, Rémi / Ware, Stephen G. / Finlayson, Mark A. (eds.): *7th Workshop on Computational Models of Narrative*. CMN 2016. Krakau, 11.-12.7.2016. Wadern: Schloss Dagstuhl - Leibniz-Zentrum für Informatik 10:1–10:10 http://drops.dagstuhl.de/opus/volltexte/2016/6711/ [letzter Zugriff am 1. August 2019].

Matschinegg, Ingrid / Nicka, Isabella (2018): „REALonline enhanced. Die neuen Funktionalitäten und Features der Forschungsbilddatenbank des IMAREAL", in: *MEMO 2: Digital Humanities & Materielle Kultur* 10–32 http://memo.imareal.sbg.ac.at/wsarticle/memo/2018-matschinegg-nicka-realonline-enhanced [letzter Zugriff am 5. September 2019].

Matschinegg, Ingrid / Nicka, Isabella / Hafner, Clemens / Stettner, Martin / Zedlacher, Stefan (2019): „Daten neu verknoten. Die Verwendung einer Graphdatenbank für die Bilddatenbank REALonline", in: Blümm, Mirjam / Kollatz, Thomas / Schmunk, Stefan / Schöch, Christof (eds.): *DARIAH-DE Working Papers* Nr. 31. Göttingen: DARIAH-DE 1-36 http://nbn-resolving.org/urn:nbn:de:gbv:7-dariah-2019-3-5 [letzter Zugriff am 9. September 2019].

Metilli, Daniele / Bartalesi, Valentina / Meghini, Carlo (2019): "A Wikidata-based tool for building and visualising narratives", in: *International Journal on Digital Libraries* 3 (2) DOI: 10.1007/s00799-019-00266-3.

Mulholland, Paul / Collins, Trevor / Zdrahal, Zdenek (2004): "Story fountain: intelligent support for story research and exploration", in: Vanderdonckt, Jean (ed.): *Proceedings of the 9th international conference on intelligent user interfaces*. New York: ACM 62–69.

Niehr, Klaus (2015): „Erzählebenen, Erzählformen, Erzählmotive im Bild. Die Festtagsseite des Göttinger Barfüßerretabels", in: Aman, Cornelia / Hartwieg, Babette (eds.): *Das Göttinger Barfüßerretabel von 1424*. Akten des wissenschaftlichen Kolloquiums, Landesmuseum Hannover, 28.–30. September 2006. Ergebnisband des Restaurierungs- und Forschungsprojektes (= Niederdeutsche Beiträge zur Kunstgeschichte N.F. 1). Petersberg: Imhof 161–176.

Schellewald, Barbara (2011): „Einführung, I. Bild und Text im Mittelalter", in: Krause, Karin / Schellewald, Barbara: *Bild und Text im Mittelalter*. Köln: Böhlau 11–21.

Speidel, Klaus (2018): "How single pictures tell stories. A critical introduction to narrative pictures and the problem of iconic narrative in narratology" [engl. Manuskript eines Beitrags, der in polnischer Sprache unter dem Titel "Jak pojedyncze obrazy opowiadają historie. Krytyczne wprowadzenie do problematyki narracji ikonicznej w narratologii" in: Kaczmarczyk, Katarzyna (ed.), *Narratologia transmedialna.* Wyzwania, teorie, praktyki. Krakow: Universitas 2017, 65–148, erschienen ist] https://www.researchgate.net/publication/327653026 [letzter Zugriff am 14. August 2019].

Suckale, Robert (2009): *Die Erneuerung der Malkunst vor Dürer*, 2 Bde. (= Historischer Verein Bamberg für die Pflege der Geschichte des Ehemaligen Fürstbistums e.V.: Schriftenreihe 44). Petersberg: Imhof.

Störmer-Caysa, Uta (2007), *Grundstrukturen mittelalterlicher Erzählungen*. Raum und Zeit im höfischen Roman. Berlin / New York: de Gruyter.

Wolf, Werner (2017): „Intermedialität: Konzept, literaturwissenschaftliche Relevanz, Typologie intermedialer Formen [2014]", in: Bernhart, Walter (Hg.): *Selected Essays on Intermediality by Werner Wolf (1992–2014).* Theory and Typology, Literature-Music Relations, Transmedial Narratology, Miscellaneous Transmedial Phenomena (= Studies in Intermediality Online 10). Leiden / Boston: Brill Rodopi 173–211. DOI: https://doi.org/10.1163/9789004346642_008

Xu, Lei / Meroño-Peñuela, Albert / Huang, Zhisheng / van Harmelen, Frank (2017): "An Ontology Model for Narrative Image Annotation in the Field of Cultural Heritage", in: Adamou, Alessandro / Daga, Enrico / Isaksen, Leif (eds.): *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II)*. Wien, 22.10.2017 (=CEUR Workshop Proceedings Vol. 2104). CEUR 117–122.

Fußnoten

1. Volsunga saga, vgl. Grimstad 2000, S. 125, 129, 137–143.
2. Richard Wagner, Der Ring des Nibelungen, vgl. Voss 2007, Zweiter Aufzug, V. 1565–1610, S. 73–75.
3. J. R. R. Tolkien ([1937] 2006), The Hobbit or There and Back again, S. 249–250; 257–263. Von zentraler Bedeutung für Tolkien war darüber hinaus die Drachenepisode im Beowulf.
4. http://www.iconclass.org/.
5. Eine nähere Beschreibung der Anwendung semantischer Rollen und der Darstellung von Abfolgen in ONAMA wird – ebenso wie ein Beispiel für ein Nutzungsszenario – in dem in Vorbereitung befindlichen Beitrag Hinkelmanns/Landkammer/Nicka/Schwembacher/Zeppezauer-Wachauer gegeben (erscheint voraussichtlich 2020).
6. http://www.cidoc-crm.org.
7. http://www.handschriftencensus.de/6504; REALonline - Archivnummern: 006340–006479.
8. Herbort von Fritslâr, Liet von Troye, vgl. Frommann 1837. http://mhdbdb.sbg.ac.at/mhdbdb/App?action=TextInfoEdit&text=TRY
9. Konrad von Würzburg, Der Trojanische Krieg, vgl. von Keller 1858. http://mhdbdb.sbg.ac.at/mhdbdb/App?action=TextInfoEdit&text=TRO

Friends with Benefits: Wie Deep-Learning basierte Bildanalyse und kulturhistorische Heraldik voneinander profitieren

Hiltmann, Torsten
hiltmann@uni-muenster.de
Humboldt-Universität zu Berlin, Deutschland

Thiele, Sebastian
s.thiele@uni-muenster.de
Westfälische Wilhelms Universität Münster, Deutschland

Risse, Benjamin
b.risse@uni-muenster.de
Westfälische Wilhelms Universität Münster, Deutschland

Wappen und ihre Überlieferung als kulturhistorische Herausforderung

Wappen zählen zu den am häufigsten gebrauchten visuellen Zeichen und Kommunikationsträgern des Mittelalters und der Frühen Neuzeit. Allein für das mittelalterliche Westeuropa sind über eine Million unterschiedliche Wappen bekannt (Pastoureau 2018, 42). Von fast allen sozialen Schichten gebraucht, konnten diese in den verschiedensten Techniken auf den unterschiedlichsten Materialien dargestellt werden. Dabei waren die Wappen nicht nur einfache Identifikationsmarken für ihre Besitzer, sondern Träger komplexer Kommunikationsakte, die Identität, Besitz und Herrschaft ebenso ausdrücken konnten wie Parteizugehörigkeit, (behauptete) Herkunft und Verwandtschaft, oder auch politische Konzepte, Schutz, Ehre, Schande usw. (Paravicini 1998). Sie bildeten damit ein zentrales Kommunikationsmittel, dessen Analyse umfangreiche Einblicke in die vormoderne Kultur und Gesellschaft erlaubt. Dass dies in der bisherigen Forschung jedoch kaum geschah, mag an drei Gründen liegen, die mit den traditionellen Methoden der Geschichts- und Kulturwissenschaften nur schwer zu überwinden sind: die schiere Menge der Überlieferung, die Diversität der unterschiedlichen Gebrauchskontexte, sowie die Komplexität der Wappen als Medien an sich (Hiltmann 2019).

Während für die Frage der Komplexität und der Analyse umfangreicher heraldischer Daten bereits erste digitale Lösungsansätze entwickelt wurden (Hiltmann/Riechert 2019), sind die Möglichkeiten zur Erfassung der breiten und diversen Überlieferung noch ungeklärt. Dabei ist dies die grundlegende Voraussetzung dafür, die Entwicklung der Wappen und ihres Gebrauchs über Zeiten, Räume und soziale Gruppen hinweg nachvollziehen zu können.

Wappen als Bilddomäne für Deep Learning Algorithmen

Eine mögliche Lösungsstrategie zur systematischen Erschließung visuell transportierter Wappendaten kann die maschinelle Bildanalyse liefern. Insbesondere aktuelle Durchbrüche im Bereich des Deep Learnings konnten erstaunliche Ergebnisse in der visuellen Objekterkennung erzielen. Neben der hohen Performanz und Genauigkeit dieser lernenden Algorithmen ermöglicht deren Training via Backpropagation auch die Disambiguierung komplexer Strukturen in großen Datensätzen (LeCun et al. 2015).

Dem kommt entgegen, dass sich die visuelle Struktur der Wappen im Unterschied zu den meisten anderen Bilddomä-

nen relativ einfach formalisieren läßt. Denn bei den Wappen handelt es sich nicht um Bilder im klassischen Sinne, sondern um abstrakte Codes aus Formen und Farben. Für die Darstellung und Wiedererkennung eines Wappens ist es wichtig, dass die mit den einzelnen Komponenten verbundenen Konzepte (z.B. Löwe, Kreuz, Lilie) erkennbar sind. Wie diese jedoch konkret dargestellt wurden, kann von Abbildung zu Abbildung variieren. Das heißt, dass es für die Darstellungen des gleichen Wappens einen breiten Spielraum gab.

Für die automatische Bildanalyse stellen die Wappen daher eine neue und besonders interessante Bilddomäne dar. Auf maschinellem Lernen basierende Bildanalysealgorithmen sind bislang besonders sensitiv für Texturmerkmale (Geirhos et al. 2019). Da die Textur eines Wappens jedoch häufig durch das Trägermedium und die jeweilige Darstellungstechnik determiniert ist, kann diese hier im Allgemeinen vernachlässigt werden. Stattdessen sind Wappen primär durch geometrische Primitive sowie deren Anordnung charakterisiert. Damit bietet sich hier für die maschinelle Bildanalyse die Möglichkeit, die Funktionsweise der Algorithmen für verschiedene Bildabstraktionen (z.B. Geometrie vs. Textur; Form vs. Farbe, etc.) zu untersuchen und diese entsprechend weiterzuentwickeln.

Zielstellung

Der vorliegende Beitrag beschreibt die konkrete Entwicklung neuer digitalen Ressourcen und Methoden für die kulturhistorische Forschung. Dabei macht er zugleich deutlich, wie die computergestützte Analyse von Wappendarstellungen die Entwicklung und das Verständnis von Bildanalysetechniken erweitern kann. Tatsächlich ist es erst das enge Ineinandergreifen kulturhistorischer und bildanalytischer Kompetenzen und Fragestellungen, das für beide beteiligten Domänen neue und hoch innovative Potentiale für die weitere Forschung eröffnet (Abbildung 1).

Abbildung 1: Interaktion zwischen Heraldik und Bildanalyse. Farbige Pfeile zeigen die interdisziplinären Abhängigkeiten, graue Boxen listen die Vorteile für die jeweilige Disziplinen auf. Die in dieser Arbeit adressierten Aspekte sind mit einem Haken markiert.

In einem ersten Schritt soll es dabei darum gehen, Wappendarstellungen auf unterschiedlichen Medien automatisch erfassen und verzeichnen zu lassen. Dabei konzentriert sich der vorliegende Beitrag zunächst auf die Detektion heraldischer Abbildungen in mittelalterlichen und frühneuzeitlichen Handschriften unter Rückgriff auf die Methoden der Deep Learning basierten Bildanalyse.

Konkrete Umsetzung

Erstellung des Datensatzes

Der Einsatz von Deep Learning Algorithmen auf Wappendarstellungen erfordert eine hinreichend große Bilddatenmenge, welche eine spezifische, für diese Algorithmen verständliche Struktur aufweisen muss. Da bis heute keine solche Datengrundlage verfügbar ist (Sustek 2018), bestand unser erster Schritt in der Erstellung einer entsprechenden Datenbank. Diese speist sich dabei aus 34 einschlägigen Handschriften aus der Bibliothèque nationale de France (BnF), der Bibliothèque municipale de Bourges und der Bayerischen Staatsbibliothek (BSB) München, wie sie von den betreffenden Bibliotheken über deren Internetportale als Digitalisate bereitgestellt werden (siehe zukünftig: http://digitalheraldry.org).

Die ausgewählten Handschriften stammen aus dem 14. bis 17. Jahrhundert und wurden so ausgewählt, dass sie die mögliche Bandbreite von Wappendarstellungen in Handschriften abbilden (Hofman 2019), von einfachen Einzeldarstellungen (Besitzeinträge) und ihrer Verwendung in bildlichen Darstellungen über ungeordnete Skizzensammlungen (Epitaphiensammlungen, Genealogien) bis hin zu geordneten Wappenbüchern. Dabei sind sowohl Wappendarstellungen in Schildform enthalten, mit unterschiedlichen Neigungen, Skalierungen und Ausgestaltungsformen, als auch deren freie Darstellung auf Kleidung und Fahnen. Die zugrundegelegten Textgenres reichen dabei von unterschiedlichen Traktaten und Wissenssammlungen über literarische und historiographische bis hin religiösen und liturgischen Texten.

Wie Abbildung 2A zeigt, wurden die Wappen auf den entsprechenden Digitalisaten der Handschriftenseiten mittels einer Boundingbox markiert, wozu das Labeling Tool LabelImg (Tzutalin 2015) verwendet wurde. Insgesamt beinhaltet die Datenbank damit 7.182 Wappendarstellungen, welche sich auf 1.568 Seiten verteilen.

Data Augmentation mit Style Transfer zur Texturabstraktion

Da Wappen vornehmlich durch die Geometrie und nicht durch die Textur determiniert sind, haben wir in einem zweiten Schritt die Datenbank um visuell augmentierte Varianten der Wappenbilder ergänzt (*Data Augmentation*). Insbesondere haben wir hier auf die Methoden des *Style Transfer* zurückgegriffen, mit denen der Stil eines Bildes verändert wird während der eigentliche Bildinhalt gleich bleibt. Anfang 2019 konnten Geirhos und Kollegen zeigen, dass mittels Style Transfer trainierte neuronale Netze mehr von der Textur eines Bildes abstrahieren und stärker Geometrien und Formen erlernen (Geirhos et al. 2019). In unserem Fall wurde der *Style Transfer* zur Vergrößerung der Datenbank mit Hilfe von AdaIN (Huang 2017) realisiert. Als Referenz dienten hierbei die Bilder der Datenbank “Painter by Numbers” von Kaggle.com, welche ca. 80.000 Gemälde verschiedener Künstler umfasst. Die Variationsbreite der enthaltenen Stile verhindert, dass das neuronale Netz, welches mit diesen Daten trainiert wird, den Stil eines spezifischen Künstlers lernt. Jedes Bild der Datenbank wurde in einen zufällig ausgewählten Stil transferiert und damit die Gesamtmenge der Trainingsdaten verdoppelt. Ein Beispiel für einen solchen *Style Transfer* ist in Abbildung 2 gegeben.

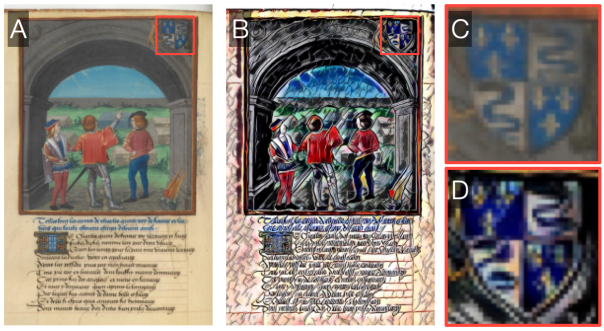


Abbildung 2: Data Augmentation mit Hilfe von Style Transfer. **(A)** Originalbild aus der Handschrift (Wappenposition via roter Box eingezeichnet). **(B)** Augmentiertes, synthetisch erstelltes Bild. **(C), (D)** Vergrößertes Wappen aus (A) und (B).

Deep Learning basierte Wappendetektion

Anschließend wurden zwei aktuelle Verfahren zur Objektdetektion auf dieser augmentierten Datenbank trainiert und analysiert. In unserer Studie haben wir YOLOv3 (Redmon 2018) und RetinaNet (Lin 2017) untersucht. In einer ersten qualitativen Analyse hat insbesondere das RetinaNet eine sehr hohe Genauigkeit erzielt, weshalb wir uns im Folgenden auf RetinaNet beschränken. Die Architektur dieses 2017 publizierten Deep Learning Objekt Detektor Modells ist in Abbildung 3 dargestellt. Dabei handelt es sich um einen einstufigen Detektor, der auf einem sogenannten *Feature Pyramid Network* aufbaut, das wiederum auf einem ResNet (He 2015) als Feature Extractor basiert. Dieser Aufbau ermöglicht es, Vorhersagen auf verschiedenen Auflösungsstufen eines Eingabebildes durchzuführen. Der Output dieser verschiedenen Stufen wird jeweils durch ein Klassifikations- und ein Regressionssubnetz zur Vorhersage der Boundingboxen verarbeitet. Außerhalb des Trainings werden diese Ausgaben zusätzlich noch durch *Non-Maximum-Suppression* gefiltert.

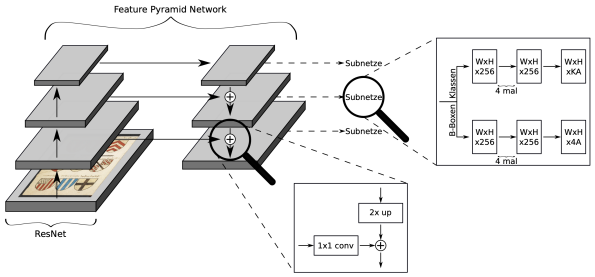


Abbildung 3: RetinaNet Architektur. Mittels Feature Pyramid Network werden relevante Merkmale auf verschiedenen Skalierungsstufen extrahiert. Anschließend werden via Subnetze Vorhersagen für Boundingboxen (B-Boxen) errechnet (angelehnt an (Lin 2017)).

Vorläufige Ergebnisse

Zum Training wurde die Datenbank in ein Trainings- und Testdatensatz aufgeteilt. Der Trainingsdatensatz stammte aus 20 Handschriften, enthielt insgesamt 1.090 Seiten mit Wappendarstellungen und wurde mittels Style Transfer augmentiert. Die übrigen 14 Handschriften im Testdatensatz enthielten 478 Seiten. Nach dem Training des RetinaNets auf dem Trainingsdatensatz erzielten wir eine Average Precision von 0,80 und einen F1 Score von 0,78 auf den Testdaten (Flach 2015).

Abbildung 4: Ergebnisse des RetinaNet Wappendetektors. Grüne Boxen zeigen manuell gesetzte Wappenpositionen, die roten Boxen die vom Wappendetektor erkannten Regionen.

In der Analyse der Ergebnisse ist festzustellen, dass Wappen sowohl über verschiedene Skalierungsstufen (vgl. Abb. 4 C und D) als auch über verschiedene Wappenstile und damit über verschiedene Texturen hinweg (vgl. Abb. 4 D und E) zuverlässig erkannt wurden. Dabei können Wappen auch in unübersichtlichen Szenen (z.B. Abb. 4A unten) detektiert werden.

Ferner zeigt sich, dass Wappen in Schildform sehr gut erkannt werden (z.B. Abb. 4C) , während Wappen auf Kleidung und Fahnen deutlich schlechter abschneiden und maßgeblich den Fehler der Average Precision und des F1 Scores beeinflussen (Abb. 4B). Dies lässt sich möglicherweise damit erklären, dass die Trainingsdaten zwar auch Wappen auf Kleidung und Fahnen enthalten, diese jedoch im Vergleich zu der deutlich weiter verbreiteten Schildform stark unterrepräsentiert sind. Um diese Fehlerquelle zu beheben, soll in einem nächsten Schritt ein Klassifikator trainiert werden, der nicht nur die Position und Größe eines möglichen Wappens erkennt, sondern auch entscheidet, ob es sich um ein heraldisches Schild, eine heraldische Fahne oder um heraldische Kleidung handelt.

Einordnung der Ergebnisse

Das Projekt macht nachvollziehbar, wie im Rahmen des maschinellen Lernens mit ganz unterschiedlichen Spielräumen umgegangen werden kann. Während durch den *Style Transfer* in den Trainingsdaten der Erkennungsraum von der Textur gelöst und auf die Geometrie der Wappen umgeleitet wurde, muss er hinsichtlich der konkreten Operationalisierung dessen, was hier als Wappen verstanden wird, für die Maschine wiederum deutlich eingeschränkt bzw. präzisiert werden. Für die maschinelle Bildanalyse ergeben sich damit neue methodische Potenziale, da die Komplexität von Wappendarstellungen zwischen populären Trainingsdatensätzen wie handgeschriebenen Ziffern (z.B. MNIST) und allgemeinen Fotos (z.B. COCO) liegen und die es an weiteren Daten (z.B. Wappendarstellungen auf anderen Materialien) zu präzisieren gilt. Neben der

Notwendigkeit der konzeptionellen Schärfung der gebrauchten Konzepte zeichnet sich mit dem Projekt für die kulturhistorische Heraldik zugleich die Möglichkeit ab, die bereits umfassend digitalisierten Handschriftenbestände von Bibliotheken wie der BnF und der BSB unter Hinzunahme der jeweils hinterlegten Metadaten erstmals umfassend auf Fragen der Verbreitung und Verwendung von Wappendarstellungen in Handschriften zu untersuchen. Ein Ansatz, der im weiteren Projektverlauf dann auch auf Wappendarstellungen in anderen Medien (Siegel, Münzen, Museumsobjekte, Wandmalereien etc.) übertragbar ist.

Bibliographie

Flach, Peter / Kull, Meelis (2015): "Precision-recall-gain curves: PR analysis done right" in: *Advances in Neural Information Processing Systems* 28: 838-846.

Geirhos, Robert / Rubisch, Patricia / Michaelis, Claudio / Bethge, Matthias / Wichmann, Felix / Brendel, Wieland (2019): "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" in: *Proceedings of the International Conference on Learning Representations* (ICLR).

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.

Hiltmann, Torsten (2019): "Zwischen Grundwissenschaft, Kulturgeschichte und digitalen Methoden. Zum aktuellen Stand der Heraldik" in: *Archiv für Diplomatik* 65: 287-319.

Hiltmann, Torsten / Riechert, Thomas (2020): "Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies" in: Christelle Balouzat-Loubet (ed.), *L'édition en ligne de documents d'archives médiévaux*, Turnhout: Brepols 102-125 [im Druck].

Hofman, Elmar (2019): *Armorials in medieval manuscripts. Collections of coats of arms as means of communication and historical sources in France and the Holy Roman Empire (13th – early 16th centuries)*, PhD University of Münster.

Huang, Xun / Belongie, Serge (2017): „Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization" in: *Proceedings of the IEEE International Conference on Computer Vision*: 1501-1510.

LeCun, Yann / Bengio, Yoshua / Hinton, Geoffrey (2015): "Deep learning" in: *Nature* 521 (7553): 436–444.

Lin, Tsung-Yi / Goyal, Priya / Girshick, Ross / He, Kaiming / Dollár, Piotr (2017): "Focal loss for dense object detection" in: *Proceedings of the IEEE international conference on computer vision*: 2980-2988.

Paravicini, Werner (1998): "Gruppe und Person. Repräsentation durch Wappen im späteren Mittelalter" in: Oexle, Otto Gerhard / Hülsen-Esch, Andrea von (eds.): *Die Repräsentation der Gruppen. Texte - Bilder - Objekte*, Göttingen: Vandenhoeck&Ruprecht 327-389.

Pastoureau, Michel (2018): *L'art héraldique au Moyen Âge*, Paris:Seuil.

Redmon, Joseph / Farhadi, Ali (2018): "Yolov3: An incremental improvement." in: *arXiv preprint* arXiv:1804.02767.

Sustek, Martin / Vidensky, Frantisek / Zboril, Frantisek / Zboril, Frantisek (2018); "Family Coat of Arms and Armorial Achievement Classification." in: *International Conference on Intelligent Systems Design and Applications*: 577-586.

Tzutalin (2015): "LabelImg" in: Git code https://github.com/tzutalin/labelImg.

Game On! Digitale Archäologie und Edition zu(m) Spielen

Roeder, Torsten
torsten.roeder@leopoldina.org
Berlin

Rettinghaus, Klaus
klaus.rettinghaus@gmail.com
Leipzig

Digitale Edition für Born-Digital-Texte?

Kulturelles Erbe wird zunehmend vielfältig in digitalen Formen hervorgebracht. Für einen stetig wachsenden Teil dieses Born-Digital-Materials werden besondere Zugangsarten benötigt: Lesegeräte für veraltete Medienspeicher müssen konserviert und bereitgehalten werden, Emulatoren für nicht mehr unterstützte Betriebssysteme entwickelt und Daten in aktuellere Formate übertragen werden – was manchmal nur noch mithilfe der Computerarchäologie gelingt.

Der Erhalt digitalen Kulturerbes wird jedoch nicht nur durch deren technische Verfügbarkeit garantiert, sondern bedarf zusätzlich einer kritischen Beleuchtung und wissenschaftlich fundierten Kommentierung. Im Bereich des "analogen" kulturellen Erbes werden beispielsweise jahrhundertealte, zerfallende Handschriften für die Wissenschaft und für die Öffentlichkeit in textkritischen (digitalen, analogen und hybriden) Editionen aufbereitet und verfügbar gemacht: Dies müsste mit denselben Konsequenzen auch für jedwede Überlieferung relevanten kulturellen Erbes in digitaler Form gelten. Benötigen wir also (digitale) textkritische Editionen von Born-Digital-Texten?

Digitale Fachmagazine der 8-Bit-Ära

Ein Beispiel, das konservatorische und textkritische Aspekte verbinden kann, ist in digitalen Fachmagazinen überliefert, die in den späten 1980er und frühen 1990er Jahren auf 5¼-Zoll-Disketten ("Floppy Disk") publiziert wurden. Ein beachtenswerter Teil dieser sogenannten "Diskmags" richtete sich an Nutzer des 8-Bit-Rechners *Commodore 64* (kurz: C64) und bediente primär eine frühe User- und Gamer-Community. Diskmags enthielten einerseits benutzbare Software (meist Public Domain) und andererseits Besprechungen neu erschienener Software, Rezensionen aktueller Spiele, Hardwaretipps und -bauanleitungen, Editorials, Leserbeiträge und

mehr. Schätzungsweise existierten über 30 Diskmags (vgl. Diskmag-Archiv o. D.), unter denen im deutschsprachigen Raum die Titel *Magic Disk 64* (1987–1993)[1] und *Game On* (1988–1995)[2] zweifelsohne zu den bekanntesten gehörten.[3]

Die Überlieferungssituation der Diskmags ist prekär. In öffentlichen Institutionen sind kaum Bestände erhalten,[4] und viele Exemplare werden nur noch in Privatsammlungen konserviert. Zudem drohen die Datenträger ihre Inhalte zu verlieren,[5] und inwieweit die gedruckten Titelseiten, mit denen die Diskmags ausgeliefert wurden, überhaupt noch greifbar sind, ist bislang nicht erforscht. Es ist einer sehr lebendigen 8-Bit-Szene zu verdanken, dass zumindest ein Teil der Binärdaten noch verfügbar ist; deren Erschließung allerdings vorrangig die Interessen der Fangemeinde bediente und bislang höchstens ansatzweise unter wissenschaftlichen Vorzeichen geschah. Bei anderen "professionellen" Bewahrern ist die Diskmagazin-Thematik offenbar noch nicht weiter ins Bewusstsein gerückt.[6]

Computerspielekritik als Forschungsobjekt

Diskmags sind unter anderem für die Kulturwissenschaft interessant, da die Spiele-Rezensionen (auch "Spieletests") unmittelbar das zeitgenössische Erleben von Computerspielen widerspiegeln. Die damaligen Kriterien hinsichtlich Konzept, Ästhetik (Grafik, Animationen, Sound, Musik) und Spielerfahrung sind aus heutiger Sicht kaum noch nachvollziehbar und können allenfalls von Zeitzeugen mit überzeugender Authentizität geschildert werden: Die C64-Hardware war mit 320x200 Bildpunkten, 16 Farben, 64KiB RAM, einer Rechenleistung von ca. 1MHz[7] und einem dreistimmigen SID-Soundchip[8] "state of the art", und bot Programmierern sowie Konsumenten völlig neue Nutzungsmöglichkeiten und Erlebnisräume. Und das sogar mit "Haushaltsgeräten": Es war gängig, anstelle eines Monitors den Fernseher[9] als Bildschirm zu verwenden, und die Mono-Tonausgabe über die eigene Soundanlage laufen zu lassen. Über den kreativen Umgang mit der technischen Grundausstattung des Commodore 64 und der Auslotung ihrer Grenzen wurde in der Fachcommunity ausführlich diskutiert. Beispielsweise konnte der Soundchip dazu gebracht werden, Sprachausgaben zu erzeugen (z. B. im Spiel "Ghostbusters", vgl. Rettinghaus 2018), und man fand heraus, dass durch eine besondere Behandlung der sogenannten "Sprites" (vgl. Morrow 2019) bildschirmfüllende Animationen kreiert werden konnten (z. B. im Spiel "Katakis", C64 Wiki 2019c, Abschnitt "Katakis-Entwicklungs-System"). Die in den Diskmags (und Printmagazinen, z.B. *64er*[10]) überlieferten Rezensionen sind für Studien im Gebiet der 8-Bit-Ära eine unersetzliche Quelle.

Ein Blick in die *Magic Disk 64* veranschaulicht deren Potenzial als Quelle kulturgeschichtlicher Untersuchungen. In frühen Ausgaben sind die Spieleberichte sehr kurz und lesen sich mehr wie Teasertexte, weshalb die Bezeichnung "Rezension" im Sinne einer kritisch-reflexiven Betrachtung noch nicht angemessen ist (z. B. in "The Last Ninja", Magic Disk 64 1987a). Die Texte vermitteln die Thematik und die Atmosphäre eines Spiels und übernehmen die Aufgabe, in der 8-Bit-Umgebung eine Vorstellungswelt zu stimulieren.[11] In späteren Ausgaben werden die Texte ausführlicher und kritischer, und ein Bewertungsraster tritt hinzu.[12] Trotz dieser Objektivierungstendenz bleibt das subjektive Spielerlebnis – so lassen die bislang gesichteten Texte vermuten – der entscheidende Faktor für die Bewertung. Anhand der Entwicklung der Spieleberichte über einen längeren Zeitraum (und den Vergleich mit weiteren Textkorpora) ließen sich sowohl die Herausbildung einer 8-bit-spezifischen Spieleästhetik als auch eine Genese der Computerspielekritik beobachten und nachvollziehen.

Herausforderungen einer digitalen Edition

Als Quellenmaterial stellen Diskmags eine Herausforderung dar. Die einerseits historische und andererseits technische Distanz sprechen für eine zeitgemäß aufbereitete (digitale) Präsentation. Im Unterschied zu analogem Material stellt beim Diskmag die Medialität eine besondere Herausforderung dar, zumal es für diese keine editionswissenschaftlichen Standards gibt. Die Medialität des Editionsobjekts zwingt den Bearbeiter förmlich dazu, sich mit der spezifischen Benutzungsweise auseinanderzusetzen und diese als "Erlebensparameter" in die Edition mit einzubeziehen – und auch Vorzüge der Emulation gegenüber der Edition abzuwägen.

Ein Versuch, die Texte zugänglich zu machen, wurde bereits durch ein anonymes Underground-Portal unternommen (http://magicdisk.untergrund.net, keine autoritative Quelle). Die dort wiedergegebenen Texte wurden aus Originaldaten generiert, welche der Beschreibung nach im Floppy-Image-Format ".d64" vorlagen.[13] Die Übertragung offenbart jedoch einige Desiderata in text- und medienkritischer Hinsicht:

1. Wie waren die Texte insgesamt zu einem Magazin angeordnet? Wie war das Layout der einzelnen Texte gestaltet? Wie wurden Grafiken eingebunden? Könnten exemplarische Screenshots helfen, um die Originaldarstellung nachvollziehen zu können? Und sind die einzelnen "Screens" statisch (wie eine konventionelle Seite) oder sind sie scrollbar (ähnlich einer Website)?
2. Das originale Diskmag zeichnete sich durch eine animierte Menüführung und eine Begleitmusik aus, die im Hintergrund abgespielt wurde.[14] Beides wurde nicht abgebildet. Wie könnten diese aber in einer Edition präsentiert werden?
3. Ferner ist das Layout von einem Zeichenraster abhängig, das aus 25 Zeilen zu 40 PETSCII-Zeichen besteht. Sollten die Texte also mit einer dicktengleichen Schriftart dargestellt werden? Wie können die originalen Zeilenumbrüche dokumentiert und berücksichtigt werden, zumal sie manchmal für die korrekte Darstellung von Rastergrafiken (z.B. Magic Disk 64 1987b) notwendig sind?
4. Hinzu kommt die Tatsache, dass der Grafikchip des C64 für eine analoge Bildschirmausgabe entwickelt wurde und eine (wie auch immer geartete) Farbtreue im Digitalen nur schwer zu erzielen ist (vgl. Pepto 2017).
5. Der teilweise sehr spezielle Fach- und Szenejargon bedarf einer historischen Erläuterung. Zudem fehlt jeglicher Kommentar zu offensichtlichen sachlichen Fehlern im Text. Beispielsweise wurde bei einer Besprechung des sehr erfolgreichen Spiels "Pirates!" der englische Ausdruck "Dutch" mit "Deutsch" verwechselt (vgl. Magic Disk

64 1987a). Spätestens an einer solchen Stelle ist ein textkritischer Kommentar notwendig.

Des Weiteren stellen sich allgemeine methodische Fragen:

1. In welches Format ließen sich die Texte generell sinnvoll übertragen, um sie langfristig zu erhalten und darauf aufbauende Studien zu ermöglichen? Finden sich beispielsweise in TEI bereits Lösungen dafür, oder bedarf die "neodigitale" Form neuer Elemente? An einem Versuch ließ sich feststellen, dass zwar keine grundsätzlichen neuen Elemente notwendig sind, sich aber teilweise die Begrifflichkeiten verschieben: So ist z.B. die "Bildschirmzeile" (die auch leer sein kann) von der "Textzeile" abzugrenzen.
2. Wie können Diskmag-spezifische interaktive oder mediale Elemente, z.B. Menüführung, Animationen und Hintergrundmusik nachnutzbar dokumentiert werden? Sollte z.B. die Musik vom SID-Format nach MEI übertragen werden, bzw. welche anderen nicht-binären Formate bieten sich an?[15]
3. Wie ist das Verhältnis zur mitgelieferten Software zu gestalten? Gibt es außerdem Printanteile, die den Magazinen beilagen (Intertextualität)?[16]
4. Ist es möglich – und wenn ja, wieweit sinnvoll – die ursprüngliche Nutzererfahrung nachzubilden? Möchte der heutige Nutzer beispielsweise die ursprünglich langen Ladezeiten nachvollziehen können?
5. Welche juristischen Modelle könnten im Hinblick auf die Urheberschaft und die Verwertungsrechte von Texten, Musik und Grafiken, die noch in persönlicher bzw. privatwirtschaftlicher Hand liegen, zur Anwendung kommen?

Zusammenfassung

Für eine kritische digitale Edition von Diskmags sprechen sowohl die kulturgeschichtliche Relevanz der Texte, gerade im Hinblick auf die wenig erforschte Entwicklung einer Ästhetik und professionellen Software- und Spielekritik, als auch die Notwendigkeit einer Vermittlung des Materials, das hinsichtlich Inhalte und (8-Bit-)Medium als historisch anzusehen ist. Der Vortrag präsentiert ausgewählte Diskmags im Original sowie eine exemplarische digitale Diskmag-Edition (TEI mit HTML-Ausgabe), anhand derer die aufgezeigten Phänomene demonstriert und die verwendeten Methoden zur Diskussion gestellt werden. Welche Grenzen weisen die gewählten Verfahren auf, und welche Aspekte – analog zur unersetzlichen Begutachtung eines echten Manuskripts – kann wiederum nur die Emulation abdecken? Welche methodischen Aspekte gelten allgemein für Born-Digital-Material, welche sind materialspezifisch für Diskmags? Bedarf es schließlich einer Erweiterung der "textkritischen Edition" zu einer "medienkritischen Edition"?

Fußnoten

1. Magic Disk 64 1987–1993. Erscheinungsjahre bei Wikipedia 2019 und C64 Wiki 2019a.
2. Game On 1988–1995. Erscheinungsjahre bei Wikipedia 2018 und C64 Wiki 2019b.
3. Eine systematische Auswertung von Auflagen, Verkaufzahlen und Verbreitung steht aus. Zudem ist eine Sekundärleserschaft zu berücksichtigen: Diskmags wurden mehrfach weitergegeben und auch durch Kopien vervielfältigt (sofern kein Kopierschutz dies verhinderte).
4. In der Zeitschriftendatenbank (ZDB) sind von den Diskmags *Magic Disk 64* und *Game On* lediglich einige Ausgaben aus den Jahren 1992 und 1993 in München nachgewiesen. Wenigstens von der *Golden Disk 64* sind Exemplare aus den Jahren 1993 bis 1996 in München greifbar.
5. Magnetische Aufzeichnungen erleiden mit der Zeit eine sogenannte "data degradation". Speziell zur Floppy Disk vgl. die Ausführungen in National Semiconductor Corporation 1989, 5-30.
6. Eine konkrete Anfrage beim Computerspielemuseum Berlin blieb bislang leider unbeantwortet.
7. Abhängig vom verwendeten Farbkodiersystem (NTSC/PAL).
8. Zu technischen Aspekten und der kulturgeschichtlichen Bedeutung des SID vgl. Rettinghaus 2018.
9. Der C64 hatte einen eingebauten HF-Modulator, sodass man das Bild über ein Antennenkabel (Koaxialkabel) in einen Fernseher einspeisen konnte.
10. 64er 1984–1996. Digitalisate des Magazins sind unter anderem im Internet Archive 2013 verfügbar.
11. Dies impliziert nicht, dass dies angesichts der grafischen und auditiven Möglichkeiten des Commodore 64 notwendig war. Selbst rein textbasierte Spiele (Textadventures) oder Spiele mit extrem einfacher Grafik können eine hohe Spielqualität aufweisen.
12. Das Schema umfasste die Kategorien Grafik, Musik [vermutlich auch Soundeffekte], Motivation [gemeint: "Wiederholungsdrang"], Preis/Leistung und Overall [Gesamtbewertung]; vgl. Magic Disk 64 1993: "Game-Test: First Samurai".
13. Eine detaillierte Beschreibung des d64-Image-Formats (entspricht dem 1541 Floppy Disk Format) ist unter http://unusedino.de/ec64/technical/formats/d64.htmlverfügbar.
14. Die Musik wurde zumeist von Thomas Detert beigesteuert und ist in der High Voltage SID Collection (HVSC) erhalten. Das Sammlungsprojekt versteht sich als "attempt to accurately archive the most popular C64 SIDs into one complete collection". Hinter SID verbergen sich die Musikdateien, die in den zwei Datenformaten PSID und RSID vorliegen, vgl. die Dokumentation bei HVSC 2019.
15. Unberücksichtigt bleibt hier der Horizont technischer Reproduzierbarkeit des Klangerlebnisses, vgl. Rettinghaus 2018, hier bes. S. 275.
16. Aufgrund des begrenzten Arbeitsspeichers waren Printbeilagen eine häufige Ergänzung, teilweise wurden z. B. (wie auch bei Printmagazinen) Poster beigelegt.

Bibliographie

64er (1984–1996): *64er – Das Magazin für Computerfans*, Haar (bei München): Markt+Technik Verlag. (ZDB: 50387-3)

C64 Wiki (2019a): "Magic Disk 64", Website: *C64 Wiki*, 17. März 2019, 21:11 Uhr. URL: https://www.c64-wiki.de/wiki/Magic_Disk_64

C64 Wiki (2019b): "Game On", Website: *C64 Wiki*, 28. März 2019, 21:28 Uhr. URL: https://www.c64-wiki.de/wiki/Game_On

C64 Wiki (2019c): "Katakis", Website: *C64 Wiki*, 29. August 2019, 18:25 Uhr. URL: https://www.c64-wiki.de/wiki/Katakis

C64 Wiki (2019d): "Ghostbusters", Website: *C64 Wiki*, 4. September 2019, 10:45 Uhr. URL: https://www.c64-wiki.de/wiki/Ghostbusters

Game On (1988–1995): *Game On. Das C64-Spielemagazin auf Diskette*, Nürnberg: CP Computer Publications, 1988–1995. (ZDB: 1276125-4)

HVSC (2019): *The High Voltage SID Collection. Commodore 64 music for the masses*, URL: https://www.hvsc.c64.org/

Internet Archive (2013): "64'er Magazine", Website: *archive.org*, 9. Januar 2013. URL: https://archive.org/details/64er_magazine

Magic Disk 64 (1987–1993): *Magic Disk 64. Das C64-Magazin auf Diskette*, Nürnberg: CP Computer Publications. (ZDB: 1275979-X)

Magic Disk 64 (1987a): "3.1 Software" [Autor: Alexander Wiederhold?], *Magic Disk 64* 12/1987. (Textversion abrufbar auf: magicdisk.untergrund.net, zuletzt geändert: 15. August 2008, 12:08 Uhr, archiviert unter: Internet Archive Wayback Machine)

Magic Disk 64 (1987b): "Intern" (= Rubrik 8.1), *Magic Disk 64* 12/1987. (Textversion abrufbar auf: magicdisk.untergrund.net, zuletzt geändert: 15. August 2008, 12:08 Uhr, archiviert unter: Internet Archive Wayback Machine)

Magic Disk 64 (1993): "Game-Test: First Samurai" [Autor: Florian Brich?], *Magic Disk 64* (02/1993). (Textversion abrufbar auf: magicdisk.untergrund.net, zuletzt geändert: 15. August 2008, 12:03 Uhr, archiviert unter: Internet Archive Wayback Machine)

Steve Morrow (2019): "C64 Sprites Defined", Website: *C64 Brain*, 4. Februar 2019. URL: https://www.c64brain.com/graphics/c64-sprites-defined/

National Semiconductor Corporation (1989): Mass Storage Handbook, Section 5: Floppy Disk Controller. Santa Clara: National Semiconductor. (archiviert unter: Internet Archive)

Philip „Pepto" Timmermann: (2017): "Calculating the color palette of the VIC II", Februar 2017. URL: https://www.pepto.de/projects/colorvic/; siehe auch die Fassung von 2001: "Commodore VIC-II Color Analysis (Preview)". URL: http://unusedino.de/ec64/technical/misc/vic656x/colors/

Klaus Rettinghaus (2018): "Sidology. Zur Geschichte und Technik des C64-Soundchips", in: *Digitale Spiele*. DOI: 10.14361/9783839440025-018

Wikipedia (2018): "Game On", Website: *Wikipedia. Die freie Enzyklopädie*, 2. Dezember 2018, 18:35 Uhr. URL: https://de.wikipedia.org/wiki/Game_On

Wikipedia (2019): "Magic Disk 64", Website: *Wikipedia. Die freie Enzyklopädie*, 9. Mai 2019, 21:06 Uhr. URL: https://de.wikipedia.org/wiki/Magic_Disk_64

www.c64.at (o. D.): "Diskmag-Archiv", Website: *www.c64.at. Das Verzeichnis von deutschsprachigen Magazinen für den C64*, ohne Datum. URL: http://www.c64.at/pages/diskmag-archiv.php

Geschichte aus erster Hand – Der Aufbau eines nationalen Zeitungsportals unter Berücksichtigung der Bedürfnisse verschiedener Nutzergruppen

Landes, Lisa
l.landes@dnb.de
Deutsche Nationalbibliothek, Deutschland

Dinger, Patrick
p.dinger@dnb.de
Deutsche Nationalbibliothek, Deutschland

Ziele und technische Grundlagen des Infrastrukturprojektes „Deutsches Zeitungsportal"

Historische Zeitungen wurden in den letzten Jahren von deutschen Kulturerbe-Einrichtungen verstärkt digitalisiert und zugänglich gemacht. Dadurch stehen der Forschung hunderte Millionen digitalisierter Zeitungsseiten zur Verfügung – ein Reichtum an Primärquellen, dem man mit den herkömmlichen geisteswissenschaftlichen Forschungsmethoden („Close Reading") kaum gerecht werden kann. Zunehmend gewinnen daher Analysemethoden der Digital Humanities, wie z.B. „Distant Reading" (Burckhardt u.A. 2018), an Bedeutung, um die bei der Massendigitalisierung entstandenen Daten auswerten zu können.

Die Aufgabe eines nationalen Zeitungsportals ist es jedoch, nicht nur für Forschende, sondern auch für eine interessierte Öffentlichkeit einen niedrigschwelligen Zugang zu entwickeln. Während mit ANNO, Delpher oder dem British Newspaper Archive im europäischen Raum mehrere Projekte – teilweise in Kooperation mit kommerziellen Partnern – entstanden sind, die große digitale Zeitungsbestände in einem nationalen Portal verbinden, existieren in Deutschland bislang nur lokale und regionale Portale einzelner Bibliotheken oder Regionen.[1] Ein übergreifendes Portal, das einen zentralen Zugriff bietet, ist bislang noch ein Forschungsdesiderat (Blome 2018: B.6-33). Dank verschiedener DFG-Förderinitiativen zur Digitalisierung historischer Zeitungen sowie der Förderung der Errichtung eines nationalen Zeitungsportals wird sich das nun ändern. Anfang 2019 wurde das Projekt „Deutsches Zeitungsportal" gestartet. Es wird im Rahmen der Deutschen Digitalen Biblio-

thek (DDB) umgesetzt. Ende 2020 soll das Portal, an dessen Aufbau vier Projektpartner[2] beteiligt sind, online gehen.

Das entstehende Portal wird eine auf die Besonderheiten von Zeitungen zugeschnittene zentrale Präsentations- und Rechercheoberfläche bieten und die folgenden Kernfunktionalitäten umsetzen:

- übergreifende Volltextsuchen in den digitalisierten Zeitungsbeständen
- unterschiedliche Einstiegspunkte, z.B. über Kalender und Zeitungstitel
- einen unmittelbar in die Portalumgebung integrierten Viewer für Volltexte und Images
- persistente Referenzierbarkeit und damit Zitationsfähigkeit der digitalen Zeitungsobjekte

Das Deutsche Zeitungsportal ist ein Infrastrukturprojekt, das auf den bestehenden Netzwerken, Techniken und Workflows der DDB aufbaut. Das heißt, die Zeitungen, die im Zeitungsportal zur Verfügung gestellt werden, stammen aus verschiedenen Einrichtungen, zumeist Bibliotheken, die Erschließungsinformationen, Bilddateien und Volltexte der Zeitungen aus ihren Beständen an das Zeitungsportal liefern. In der ersten Ausbaustufe des Zeitungsportals stellen Bibliotheken ihre Metadaten im METS/MODS-Format bereit, sodass über Verlinkungen in der Datenstruktur sowohl die Bilddateien (i.d.R. im jpg-Format) als auch die Volltexte (i.d.R. im ALTO-xml-Format) in das Portal übernommen werden. Dort werden diese Bestände zusammengeführt und können über einen Suchschlitz durchsucht werden. Basis der Suche bilden die Volltexte, welche vollständig in einen zentral vorgehaltenen SOLR-Index aufgenommen werden. Durch die standardisierten, maschinennutzbaren Inhalte des Zeitungsportals, die über eine Schnittstelle (API) heruntergeladen werden können, sollen neue Nutzungsszenarien, wie z.B. Big-Data-Analysen, ermöglicht werden (Altenhöner 2018: 146). Da es sich bei den Zeitungstexten und -bildern ausschließlich um rechtefreies Material handelt, können die Dokumente in der eigenen Forschung nachgenutzt und weiterverarbeitet werden.

Zum Start des Portals werden Bestände aus mindestens sechs deutschen Bibliotheken[3] verfügbar sein; vorsichtigen ersten Schätzungen zufolge wird es sich um 250 verschiedene Zeitungstitel im Umfang von ca. 15 Mio. Zeitungsseiten handeln. Nach der Inbetriebnahme des Portals sollen die Inhalte beständig wachsen, um möglichst viele historische Zeitungen in diesem Portal zu vereinen. Das Fernziel des Unterfangens ist es also, den Forschenden (und der allgemeinen Öffentlichkeit) einen einheitlichen und stabilen Zugang zu vielen (wenn möglich: allen) historischen Zeitungen aus deutschen Kulturerbe-Einrichtungen zu geben. Zudem wird angestrebt, in einer zweiten Ausbaustufe ab 2021 das Zeitungsportal für andere Datenformate wie TEI-XML zu öffnen, um z.B. digitale Editionen und Annotationen in den Korpus aufnehmen zu können.

Um Zeitungen aus vielen unterschiedlichen Quellen zusammenzuführen, sind bei der Aufbereitung der (Meta-)Daten vielfältige Standardisierungsprozesse notwendig. So wird im Rahmen des Projekts ein zeitungsspezifisches Anforderungsprofil für das METS/MODS-Metadatenformat entwickelt und eine Verknüpfung der Metadaten mit der Zeitschriftendatenbank (ZDB) umgesetzt, deren Identifier für die eindeutige Identifizierung der Zeitungstitel benutzt werden.

Durch den Aufbau der Datenverarbeitungsstrukturen für Zeitungen und die dafür nötigen Standardisierungsprozesse sollen auch Impulse für die Digitalisierung, Erschließung und Referenzierung der Zeitungsbestände der kooperierenden Datenpartner ausgehen, ein Effekt, der sich bereits beim Aufbau der Deutschen Digitalen Bibliothek gezeigt hat. Darüber hinaus soll der Aufbau eines Zeitungsportals Kulturerbe-Einrichtungen dazu anregen, weitere Digitalisierungsvorhaben von historischen Zeitungen anzugehen. Um ein Portal mit für die Zukunft gerüsteten offenen Schnittstellen zu schaffen, ist die perspektivische Integration der IIIF-Technologie sowie eine Anbindung an die Europeana Newspapers Collection geplant.

Ein Zeitungsportal für die Wissenschaft und die interessierte Öffentlichkeit

Das Medium Zeitung zeichnet sich dadurch aus, dass es alle Bereich des Lebens abdeckt. Digitalisierte historische Zeitungen bieten den (Geistes-)Wissenschaften die Möglichkeit, eine Vielzahl von Forschungsfragen zu adressieren (Blome 2018). Das Interesse aus der Wissenschaft fließt beim Aufbau des Zeitungsportals auf mehreren Wegen in die Konzeption ein: So waren die oben genannten vier Kernfunktionalitäten Ergebnis eines Workshops mit WissenschaftlerInnen, der im Herbst 2014 im Rahmen des DFG-Pilotprojektes „Digitalisierung historischer Zeitungen“ in Bremen stattfand. Auch der laufende Entwicklungsprozess des Zeitungsportals wird von einem internationalen wissenschaftlichen Beirat begleitet.[4] Im Gegensatz zu führenden Zeitungsprojekten der Digital Humanities wie Impresso, NewsEyes oder Oceanic Exchanges, welche ausschließlich ein wissenschaftliches Publikum im Blick haben, richtet sich das Angebot des Zeitungsportals auch an allgemeininteressierte NutzerInnen. Wie bereits bestehende, internationale Zeitungsportale zeigen, werden ihre Angebote sehr gut angenommen und von unterschiedlichsten Nutzergruppen besucht (für Nutzergruppen des österreichischen Zeitungsportals ANNO vgl. z.B. Müller 2016: 86). Vor allem die Volltextsuche, also die Möglichkeit, nicht nur in den Zeitungstiteln und anderen Metadaten, sondern in den eigentlichen Zeitungstexten zu suchen, macht das Zeitungsportal zu einem sehr niedrigschwelligen Angebot: Auch ohne große Recherchekenntnisse können NutzerInnen Artikel oder Informationen zu allen vorstellbaren Themen finden – sei es zum eigenen Sportverein, zu einer berühmten Persönlichkeit oder zur Geschichte der eigenen Familie.

Anforderungen der unterschiedlichen Nutzergruppen

Das Projekt „Deutsches Zeitungsportal“ steht somit vor der Herausforderung, Anforderungen aus der Wissenschaft und aus der allgemeinen Öffentlichkeit zu analysieren, sie soweit wie möglich zu vereinen und ihnen allen bei der Umsetzung des Zeitungsportals möglichst gerecht zu werden.

Einer der ersten Schritte bei der Entwicklung des Zeitungsportals war es darum, die Anforderungen von Seiten der Wissenschaft sowie von allgemeininteressierten NutzerInnen zu erheben. Dabei kamen zu Beginn des Projekts drei unterschiedliche Methoden zum Einsatz: eine webbasierte Nutzerumfrage, fünfzehn, jeweils einstündige Interviews mit ausge-

wählten NutzerInnen und ein zweitägiger Workshop mit dem wissenschaftlichen Beirat.

Die Umfrage hatte als Ziel, potenzielle Zielgruppen des Zeitungsportals und ihre Anforderungen und Erwartungen kennenzulernen. Sie umfasste 23 Fragen zu Themengebieten wie Rechercheanlässe, Funktionalitäten und thematische Interessen und wurde für fünfeinhalb Wochen über die Homepage der DDB und über 27 weitere digitale Kanäle verbreitet. Das große Interesse, das dem Zeitungsportal entgegengebracht wird, zeigte sich hierbei schon rein zahlenmäßig – mit 2.422 ausgefüllten Fragebogen.

Die Usability-Tests mit fünfzehn ausgesuchten potenziellen NutzerInnen wurden als halbstrukturierte Interviews durchgeführt, bei denen Methoden wie Protokolle lauten Denkens, Beobachtung und Aufzeichnung des Klickverhaltens, Aufzeichnung der Gestik und Mimik zum Einsatz kamen. Der Schwerpunkt der Interviews lag auf der Interaktion mit einem Klick-Dummy des Zeitungsportals, der aufgrund gezielter Fragen und kleiner Aufgaben auf seine Verständlichkeit und Usability einerseits und die Erwartungen der Interviewten andererseits überprüft wurde.

Die Nutzerumfrage und die Nutzerinterviews richteten sich an die allgemeine Öffentlichkeit und wurden in Zusammenarbeit mit einer Marktforschungsagentur durchgeführt, die die Ergebnisse analysiert und aufbereitet hat.

Bei dem Workshop mit dem wissenschaftlichen Beirat handelte es sich um ein zweitägiges Treffen, bei dem im Juni 2019 die vorliegenden Konzeptpapiere und der Klick-Dummy vorgestellt und mit den WissenschaftlerInnen unter Einbeziehung anderer Zeitungsportale diskutiert wurden. Die Empfehlungen, die im Lauf des Workshops entwickelt wurden, sind ebenfalls in die Konzeption des Portals eingeflossen.

Der Vortrag widmet sich der Auswertung dieser Erhebungen: Wo finden sich Gemeinsamkeiten? Wo gibt es Unterschiede? Was ist zu tun, wenn sich die Anforderungen aus Wissenschaft und allgemeiner Öffentlichkeit widersprechen? Welche Wünsche lassen sich überhaupt realisieren und wo liegen Grenzen, seien diese technisch oder urheberrechtlich bedingt?

Ein Beispiel für eine übereinstimmende Anforderung ist der Wunsch nach möglichst umfassenden Inhalten. Sowohl die allgemeine Öffentlichkeit als auch die Wissenschaftscommunity wünscht sich ein Zeitungsportal, das weitreichende Bestände anbietet, sodass sich die Recherche im besten Fall über ein einziges Portal erledigen lässt. Zwar ist genau dies der Anspruch und das angestrebte Alleinstellungsmerkmal des Deutschen Zeitungsportals, aber die Umsetzung dieses Zieles kann nicht allein vom Zeitungsportal erreicht werden. Viele Faktoren und Stakeholder – wie die schiere Menge des Materials, die Zerstreuung der Bestände in viele unterschiedliche Einrichtungstypen, die von der DFG überhaupt nicht erreicht werden, und nicht zuletzt das Urheberrecht, das die Digitalisierung und Verbreitung der besonders interessanten Bestände aus dem 20. Jahrhundert einschränkt – spielen hier eine Rolle. Ein kompletter Nachweis aller deutschen digitalisierten Zeitungsbestände kann darum eher als Vision der Community der Kulturerbe-Einrichtungen und ihrer Träger beschrieben werden, denn als ein kurzfristig erreichbares Ziel (Bürger 2018: 131f.).

Eine wichtige Erkenntnis der Diskussion war es, dass das Zeitungsportal möglichst transparent mit seinen Inhalten umgehen muss: Wenn nicht alle historischen Zeitungen zu finden sind, muss für die NutzerInnen deutlich werden, welche Inhalte verfügbar sind und nach welchen Kriterien das vorhandene Zeitungskorpus zusammengestellt wurde.

Unterschiedliche Erwartungen der Nutzergruppen wurden besonders im Bereich Nutzerinterface formuliert. Während die allgemeinen NutzerInnen sich hier eher eine einfach gestaltete Oberfläche wünschen und erwarten, alle Suchanfragen über einen Suchschlitz eingeben zu können, wurde in der Diskussion mit der wissenschaftlichen Begleitgruppe der Wunsch nach anspruchsvolleren Funktionen laut, z. B. nach der Möglichkeit, sich ein individuelles Zeitungskorpus zusammenzustellen und dieses gezielt durchsuchen zu können. Diese Anforderungen werden teilweise von der aus dem DDB-Hauptportal übernommenen Funktion „Meine DDB“ erfüllt, mit der sich Favoriten und Suchanfragen speichern lassen. Zudem könnten zukünftig erweiterte Funktionen implementiert werden, die, ohne die Startseite zu überladen, für WissenschaftlerInnen und erfahrene NutzerInnen einen komfortablen Zugang bieten, der komplexere Suchanfragen erlaubt. Der Suchschlitz für die Volltextsuche soll jedoch über alle Iterationen das bestimmende Element bleiben. Im Suchschlitz selber können auch komplexe Abfragen gemäß der von der Suchmaschine vorgegebenen Syntax eingegeben werden.

Ausblick

Zum Abschluss des Vortrags wird der aktuelle Projektstand vorgestellt, inklusive eines ersten Blicks auf den Prototypen des Zeitungsportals, das Ende 2020 für die Öffentlichkeit freigeschaltet werden soll. Der Prototyp ist nicht nur aus technischen Gesichtspunkten (Tests, Qualitätssicherung) wichtig, sondern soll als Grundlage für eine zweite Runde der Nutzerforschung dienen: Die Erkenntnisse, die zu Beginn des Projektes gewonnen wurden, sollen bis Ende 2020 anhand einer weiteren Nutzerbefragung überprüft werden. So ist geplant, den Prototyp sowohl von den allgemeininteressierten NutzerInnen als auch vom wissenschaftlichen Beirat evaluieren zu lassen und daraus Erkenntnisse für die Schwerpunkte der nächsten, für 2021–2022 geplanten Projektphase zu gewinnen.

Fußnoten

1. Hervorzuheben sind insbesondere das Portal der Staatsbibliothek zu Berlin ZEFYS http://zefys.staatsbibliothek-berlin.de [letzter Zugriff 05.09.19], digiPress der Bayerischen Staatsbibliothek https://digipress.digitale-sammlungen.de/ [letzter Zugriff 05.09.19] und das Zeitungsportal NRW https://zeitpunkt.nrw/ [letzter Zugriff 05.09.19].
2. Deutsche Nationalbibliothek (Projektleitung), Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur.
3. Dabei handelt es sich um die folgenden sechs Bibliotheken, von denen die ersten fünf auch am vorgeschalteten DFG-Pilotprojekt beteiligt waren: Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB), Bayerische Staatsbibliothek (BSB), Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB), Staats- und Universitätsbibliothek Bremen (SuUB), Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky (SUB).
4. Mitglieder des wissenschaftlichen Beirats sind: Astrid Blome (Institut für Zeitungsforschung Dortmund), Christa

Müller (Österreichische Nationalbibliothek), Claudia Resch (Österreichische Akademie der Wissenschaften), Estelle Bunout (Universität Luxemburg), Fotis Jannidis (Universität Würzburg), Günter Mühlberger (Universität Innsbruck), Jana Keck (Universität Stuttgart), Jörg Lehmann (Universität Tübingen), Marc Priewe (Universität Stuttgart), Maria Elisabeth Müller (Staats- und Universitätsbibliothek Bremen), Marian Dörk (Fachhochschule Potsdam/Urban Complexity Lab), Marten Düring (Universität Luxemburg), Pim Huijnen (Universität Utrecht), Thomas Werneke (Zentrum für Zeithistorische Forschung Potsdam).

Bibliographie

ANNO (AustriaN Newspapers Online): http://anno.onb.ac.at [letzter Zugriff 29. August 2019].

Altenhöner, Reinhard (2018): „Auf dem Weg zu einem nationalen Zeitungsportal. Eine materialspezifische Kooperation als Treiber eines neuen Dienstes für Wissenschaft und Forschung“ in: Bonte, Achim / Rehnolt, Juliane (eds.): *Kooperative Informationsinfrastrukturen als Chance und Herausforderung: Festschrift für Thomas Bürger zum 65. Geburtstag.* Berlin, Boston: De Gruyter 144-160, DOI: 10.1515/9783110587524-019.

Bürger, Thomas (2016): „Zeitungsdigitalisierung als Herausforderung und Chance für Wissenschaft und Kultur“ in: *Zeitschrift für Bibliothekswesen und Bibliographie* 63, H 3, 123-132, DOI: 10.3196/186429501663332.

Blome, Astrid (2018): „Zeitungen“ in: Busse, Laura / Enderle, Wilfried / Hohls Rüdiger u.A. (eds.): *Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften*. Berlin: Clio-online und Humboldt-Universität zu Berlin (Veröffentlichungen von Clio-online, Nr. 2), B.6-1-D.6-36, DOI: 10.18452/19244.

The British Newspaper Archive: https://www.britishnewspaperarchive.co.uk [letzter Zugriff 29. August 2019].

Burckhardt, Daniel / Geyken, Alexander / Saupe, Achim / Werneke, Thomas (2019): „Distant Reading in der Zeitgeschichte. Möglichkeiten und Grenzen einer computergestützten Historischen Semantik am Beispiel der DDR-Presse“ in: *Zeithistorische Forschungen/Studies in Contemporary History* 16: 177-196, DOI: 10.14765/zzf.dok-1345.

Delpher: https://www.delpher.nl/ [letzter Zugriff 29. August 2019].

DFG-Antrag „Errichtung eines nationalen Zeitungsportals auf der Basis der organisatorischen und technischen Infrastruktur der Deutschen Digitalen Bibliothek (DDB) – DDB-Zeitungsportal. Online verfügbar unter: https://pro.deutsche-digitale-bibliothek.de/node/985 [letzter Zugriff 29. August 2019].

DFG-Ausschreibung „Digitalisierung historischer Zeitungen des deutschen Sprachgebiets“ im Rahmen des LIS-Förderprogramms von 2018 (https://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_zeitungsdigitalisierung.pdf) [letzter Zugriff 10. September 2019].

DFG-Ausschreibung „Digitalisierung historischer Zeitungen des deutschen Sprachgebiets“ im Rahmen des LIS-Förderprogramms von 2019(https://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_zeitungsdigitalisierung_2019.pdf) [letzter Zugriff 10. September 2019].

digiPress: https://digipress.digitale-sammlungen.de/ [letzter Zugriff 05. September 2019].

Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung): http://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf [letzter Zugriff 05. September 2019].

Impresso: https://impresso-project.ch/ [letzter Zugriff 29. August 2019].

Müller, Christa (2016): „ANNO – Der Digitale Zeitungslesesaal der Österreichischen Nationalbibliothek: Aktuelle und zukünftige Entwicklungen im Überblick“ in: *BIBLIOTHEK – Forschung und Praxis* 40(1). Berlin: de Gruyter 83-89, DOI: 10.1515/bfp-2016-0012.

NewsEye: https://www.newseye.eu/ [letzter Zugriff 29. August 2019].

Oceanic Exchanges: http://oceanicexchanges.org/ [letzter Zugriff 29. August 2019].

ZDB (Zeitschriftendatenbank): https://zdb-katalog.de/ [letzter Zugriff 29. August 2019].

ZEFYS: http://zefys.staatsbibliothek-berlin.de [letzter Zugriff 05. September 2019].

Zeitpunkt NRW: https://zeitpunkt.nrw/ [letzter Zugriff 05. September 2019].

… hungere schon nach dem nächsten Band. Eine Untersuchung von Metaphern für Leseerfahrungen in Web 2.0 Literaturrezensionen

Herrmann, J. Berenike
berenike.herrmann@unibas.ch
Universität Basel, Schweiz

Messerli, Thomas
thomas.messerli@unibas.ch
Universität Basel, Schweiz

Einleitung

Kaum ein geisteswissenschaftlicher Forschungsgegenstand hat eine so intensive Diskussion erfahren wie die Metapher (Eggs, 2000). Doch existieren nur wenige Ansätze zu ihrer Formalisierung innerhalb der Digital Humanities. Unser Beitrag stellt einen einfachen Ansatz der Metaphernanalyse auf grösseren Datenmengen vor, um auch in nicht-annotierten Texten metaphorische Mappings zu finden. Mit diesem Ansatz analysieren wir konzeptuelle Strukturen des Leseerlebens von Laien-RenzensentInnen.

Mittels einer Verschränkung korpusbasierter und korpusgetriebener Methoden (Tognini-Bonelli, 2001) untersuchen wir ein Korpus von Laienrezensionen (ca. 1,3 Mio Beiträge) ex-

plorativ auf den Metapherngebrauch mit der Zieldomäne 'Leseerleben'. Metaphern werden mit der Kognitiven Theorie der Metapher (KTM) als Denk- bzw. Erfahrungsfiguren (Lakoff & Johnson, 1980, S. 4) gefasst.

Ausgehend von Befunden zu Laienbuchrezensionen im Englischen (Stockwell, 2009; Nuttall & Harrison, 2018) und zu feuilletonistischen Rezensionen (Köhler, 1999) operieren wir auf der Sprachoberfläche und inferieren von dort konzeptuelle Mappings zwischen Ziel- und Quelldomänen (Herrmann, im Druck; Shutova, 2017; Steen et al., 2010), wobei besonderes Augenmerk auf das Mapping LESEN IST NAHRUNGSAUFNAHME gelegt wird. Ausgangspunkt ist der Befund Nuttall und Harrisons (2018), dass Nahrungsmetaphern in englischsprachigen *Goodreads*-Laienrezensionen einen der häufigsten Metapherntypen darstellen (siehe auch Radway, 1986). Von dort konstruieren wir die Arbeitshypothese, dass in den sozialen Settings des als 'nicht-solitär' kommunizierten social reading eine Nahrungsmetaphorik – mit ihrer experientiell-sozialen Grundierung – besonders geeignet ist, um Anschlusskommunikation zu evozieren (vgl. Narula, 2014; Peplow et al., 2016).

Methode

Daten

Das LoBo-Korpus (extrahiert von der Social Reading-Plattform "Lovelybooks") beinhaltet ca. 1,3 Mio. deutschsprachige Laienrezensionen von 54.000 NutzerInnen, die sich auf jeweils ein Buch beziehen. Die Bücher sind kategorisiert nach 15 Genres, die der Plattform selbst entnommen sind. Das Korpus ist PoS-annotiert (Tree-Tagger), lemmatisiert und in CWB (http://cwb.sourceforge.net/) indiziert. Für die manuelle Annotation wurde das UAM CorpusTool (Version 2.8.16) (http://www.corpustool.com/) verwendet.

Metaphernidentifikation

Angesichts der Herausforderungen einer reliablen automatischen Metapherndetektion (Veale, Shutova, & Klebanov, 2016) wählen wir bewusst eine korpusstilistische Herangehensweise (Deignan & Semino, 2010). Wir verschränken als induktiven Schritt A Kookurrenzanalyse und manuelle Identifikation mit einem deduktiven Schritt B (regelbasierte Suche nach spezifischen Quelldomänen-Indikatoren). Ziel ist eine möglichst hohe Vollständigkeit und Genauigkeit der Identifikation potenzieller Metapherntypen, wobei eine formale Evaluation der Methode im gegenwärtigen Stadium mangels Goldstandard jedoch nicht möglich ist.

Um in Schritt A die Metaphern zu finden, die sich auf Leseerleben beziehen, müssen zunächst Objekte des Leseerlebens (OdL) identifiziert werden. OdL sind Referenten des Leseerlebens (*literarische Werke* wie *Buch, Geschichte, Roman,* sowie Teilaspekte wie *Ende, Seite, Handlung, Spannung, Autor, Figur*). Sie sind die Nodes, deren Kontext wir auf metaphorische Sprachverwendung untersuchen: Als Indikatoren der Zieldomäne ('Lesen') werden sie in Ausdrücken wie *hungere schon nach dem nächsten Band* mit Quelldomänen (hier 'Nahrungsaufnahme') verknüpft. In der Korpusanalyse überprüften wir mittels Kookkurrenzen die zusammen mit den Nodes signifikant häufig auftretenden Inhaltswörter (Nomen, Adjektive, Verben) auf metaphorische Verwendung.

Ergänzend zur Korpusanalyse annotieren wir eine Stichprobe auf metaphorischen Sprachgebrauch (Herrmann, Woll, & Dorst, 2019). Unsere erste Fallstudie untersuchte insgesamt 18 randomisiert ausgewählte Rezensionen zu sechs Büchern (je drei pro Buch). Dieses Subkorpus enthält zu gleichen Teilen Rezensionen von "anspruchsvollen Bestsellern" und Fantasy-Romanen. Ziel war es, die Sequenz metaphorischer Ausdrücke sowie die lexikalische und konzeptuelle Variation abzuschätzen.

In Schritt B untersuchten wir ausgehend von der Annahme eines systematischen Mappings LESEN IST NAHRUNGSAUFNAHME (ausgehend von entsprechenden Befunden durch Nuttall & Harrison, 2018, Köhler, 1999) die je hundert häufigsten Lemmata der drei "Inhaltswortklassen" Substantiv, Adjektiv und Verb auf mögliche Indikatoren. Innerhalb eines Fensters von zehn Wörtern um die in Schritt A festgelegten OdL (Ausdrücke, die Zieldomäne LESEN indizieren, z.B. *Band* in *hungere schon nach dem nächsten Band*) wurde dabei nach Lexemen mit einer Grundbedeutung in der Domäne NAHRUNGSAUFNAHME (etwa *hungere*) gesucht. Mangels eines out-of-the-box semantischen Taggers (vgl. Demmen et al., 2015) nutzten wir Dornseiffs (2004) semantisches Feld 'Essen und Trinken' zur Erstellung einer nach Häufigkeit sortierten Lemmaliste. Von den resultierenden 1.386 Lemmata finden sich 993 mindestens einmal im LoBo-Korpus. Ein Problem, das besonders die häufigen Lemmata betrifft, ist Domänengeneralität. Zum einen ist die Zuordnung zur Zieldomäne LESEN nicht immer gegeben, zum anderen ist der metaphorische Wortgebrauch im Einzelfall nicht gesichert (es kann sich z.B. um eine Inhaltsangabe handeln, in der unmetaphorisch von Nahrungsaufnahme die Rede ist). Diese Schwäche haben wir reduziert, indem besonders häufige *false positives* sowie besonders periphere Mitglieder des semantischen Feldes aus der Liste entfernt wurden (*gar, langweilig, hart, Atmosphäre, Pferd, zusagen,* etc.). Die resultierende Liste (s. Tabelle 1 für Ausschnitt) erlaubt zwar selbst keine Rückschlüsse auf die tatsächliche Metaphernverwendung, dient jedoch als Zwischenschritt um Nahrungsmetaphern im Korpus zu finden, die sich in der Folge qualitativ untersuchen lassen.

Tabelle 1: 25 häufigste Lemmata aus dem semantischen Feld 'Essen und Trinken' (Dornseiff, 2004) innerhalb von 10 Wörtern eines Objekts des Leseerlebens

Lemma (semantisches Feld: Essen & Trinken)	**Freq. innerhalb von 10 Wörtern eines OdL**
verschlingen	29827
genießen	16665
Geschmack	15992
fein	5885
zart	5316
bitter	3369
Kost	2865
kosten	2690
köstlich	2042
Essen	1682
Koch	1500
lecker	1461
essen	1100
servieren	1094
riechen	1042
schlucken	1016
Gin	1013
herzhaft	965
Duft	956
kochen	910
Hunger	881
Schokolade	846
scharf	845
fressen	718

Ergebnisse

Die Ergebnisse der Schritte 1 und 2 zeigen eine grosse Vielfalt metaphorischer Ausdrücke auf, die sich auf verschiedene Objekte des Leseerlebens beziehen. Aufschlussreich ist dabei nicht die absolute Häufigkeit der Metaphernkandidaten im Korpus – zumal keine zuverlässigen Vergleichsdaten zur Verfügung stehen –, wohl aber die quantitative Analyse der relativen Verteilung auf Rezensionen verschiedener Ratings und Genres. Das Auftreten der hier untersuchten stark wirkungsbezogenen Metaphorik gibt etwa Aufschlüsse über Rezensionsmuster, die sich je nach quantitativer Bewertung und je nach literarischer Gattung unterscheiden. Die qualitative Untersuchung ermöglicht dagegen eine erste Typologie von Mappings, die wir im Folgenden mit Beispielen für konventionelle und kreative Metaphern illustrieren.

Zwischen Konvention und Kreativität

Viele Ausdrücke sind erwartete, stark konventionalisierte Metaphern, wie etwa *verschlingen* und *Geschmack*:

- Leider kommt man erst ab der zweiten Hälfte so richtig rein und hat das <Buch innerhalb weniger Stunden verschlungen>
- Die <Geschichte kam für meinen Geschmack> zu langsam in Fahrt…

Es finden sich darüber hinaus aber auch viele Beispiele, die einen kreativen Umgang mit Metaphern illustrieren:

- Daher empfehle ich ihn gerne weiter an alle , die es ab und an mal etwas romantischer mögen und Lust auf eine <Geschichte haben , die nach Sommer schmeckt> :)
- Die Entwicklungen um Mia gefallen mir sehr gut und das Buch ist auch so beendet worden , dass der <Lesehunger>auf den zweiten <Teil gut genährt> hinterlassen wird .

Eine erste Typologie von Mappings

Unsere Resultate zeigen bislang fünf verschiedene Typen von LESEN IST NAHRUNGSAUFNAHME auf.

Lesen wird etwa (A) als eine Form von Nahrungsaufnahme konzeptualisiert, bei der Lesende als 'Essende' und literarische Werke und deren Bestandteile als 'verzehrbar' positioniert werden.

- Die ersten <Seiten habe ich gefressen> , bis ich nach 300 Seiten ins Stocken geriet…
- Ich fand weite Strecken des <Buches recht fad>, in die Länge gezogen oder nicht wirklich wichtig .
- Der <Roman ist definitiv keine leichte Kost> , insbesondere , wenn es um Annas körperlichen und seelischen Zustand geht .
- Hatte sehr lange insgesamt an dem <Buch genagt> , aber es hat sich letztendlich doch gelohnt
- Und doch habe ich es in einem Zug durchgelesen und <hungere schon nach dem nächsten Band> .
- Ich kann es zumindest gar nicht erwarten , die <Romane neun und zehn zu verzehren> oder wie seht ihr das ?
- Eine <Geschichte , die sauer> aufstößt , die einen wütend macht , die man aber doch nicht eine Minute zur Seite legen kann .

Weiter wird (B) Schreiben als 'Kochen' und 'Bewirten' dargestellt, wobei Autoren als Köche, Lesende als Gäste und Lektüre als bekocht/bewirtet erscheinen.

- Mit "Dark Wonderland : Herzkönigin" <serviert Howard dem Leser> eine düstere und noch phantastischere Version der Alice-Geschichte in bester Teegesellschaftsmanier.
- Auch die mittelschwere Portion Liebe und Romantik hat die <Autorin in meinen Augen schmackhaft> verpackt.
- Die Geschichte wirkte auf mich irgendwie zwanghaft konstruiert nach dem Motto packen wir alle Wunderland <Figuren in einen Topf> geben etwas " Grusel" und Blut dazu rühren einmal um fertig - Schade !

Darüber hinaus findet sich aber auch (C) ein anderes Mapping, das zwar auf die gleiche Quelldomäne zurückgreift, aber dem Objekt des Leseerlebens selbst als Agnes konstruiert.

- Was für den Leser immer besonders schön ist , denn genau solche Paare und <Geschichten verschlucken> einen förmlich .
- Ein interessanter und unterhaltender <Roman , genährt> durch seriöse Quellen , aber auch durch die Fantasie der Autorin selbst .

Andere Mappings (D) beziehen sich dagegen direkt auf die Objekte des Leseerlebens, werden sprachlich aber als Vergleich realisiert, der nach Steen et al. (2010) stärker intentional markiert ist.

- Das <Buch entfalltet sich wie ein guter Wein> erst am Ende.

- Anfänglich betrachtete ich ein abgebrochenes <Buch wie ein nicht aufgegessenes Essen> , ein verfehltes Ziel beim Joggen , eine Niederlage .
- Diese <Story wärmt wie eine leckere> Tasse Punsch

Schliesslich findet sich eine Reihe (E) von Mappings, die zwar auf die Quelldomäne NAHRUNGSAUFNAHME und Zieldomäne LESEN rekurrieren, sich dabei aber nicht auf Vorgänge der Nahrungsaufnahme, sondern auf das Embodiment von Emotionen zu beziehen scheinen.

- Als ich die ersten Seiten aufgeschlagen habe , musste ich erst einmal <schlucken , denn die Story> hat sich meines Erachtens ziemlich in die Länge gezogen .
- Man kann die Angst der <Protagonisten förmlich riechen> .
- Ein <Buch zum an den Fingernägel knabbern> .

Fazit

Unsere Studie leistet einerseits einen methodischen Beitrag zur Metaphernidentifikation mit einfachen korpusstilistischen Mitteln und gibt andererseits Aufschluss über die Produktivität der Quelldomäne NAHRUNGSAUFNAHME für die Konzeptualisierung von Leseerleben. Schliesslich zeigt unsere manuelle Annotation weitere Mappings auf, die den Umgang mit Objekten des Leseerlebens nicht als Nahrungsaufnahme konzeptualisieren, sondern als 'Reisen' und 'Bewegung', oder auch auch als 'Interaktion mit externen Kräften'. Bücher und Geschichten werden einerseits als 'Behälter', andererseits wie 'Personen' mit Qualitäten und Intentionen konzeptualisiert, ja als 'Freunde' der Lesenden, wobei 'gegenseitige Kompatibilität' als axiologischer Wert - situiert zwischen den inhaltlichen und (hedonistisch sowie praktisch) wirkungsbezogenen Werten nach Heydebrand und Winko (1996) - erscheint. Folgestudien sollen die Verbesserung der automatisierten Detektion leisten, unter anderem durch die Einbindung von semantischen Informationen aus GermaNet. So sollen durch systematische Untersuchung der häufigen konzeptuellen Metaphern und ihre Korrelation mit der Lovelybooks-Sterne-Wertung weitere Rückschlüsse auf zugrundeliegende Wertmassstäbe bei der Bewertung von online-Laienrezensionen ermöglicht werden.

Bibliographie

Demmen, J. / Semino, E. / Demjén, Z. / Koller, V. / Hardie, A. / Rayson, P. / Payne, S. (2015): "A computer-assisted study of the use of Violence metaphors for cancer and end of life by patients, family carers and health professionals" in: *International Journal of Corpus Linguistics*, 20/2: 205–231.

Dornseiff, F. (2004): *Der deutsche Wortschatz nach Sachgruppen*. Mit einer lexikographisch-historischen Einführung und einer ausführlichen Bibliographie zur Lexikographie und Onomasiologie (8. völlig neu bearb. u. m. einem alphabet. Zugriffsreg. vers. Aufl. Reprint 2010). https://doi.org/10.1515/9783110901009

Eggs, E. (2000): «Metapher» in: Gert Ueding (Hg.), Historisches Wörterbuch der Rhetorik, Bd. 5, Tübingen 1109-1183.

Genette, G. (1989): *Paratexte: Das Buch vom Beiwerk des Buches*. (D. Honig, Übersetzung). Frankfurt am Main: Campus.

Herrmann, J. B. (im Druck): «Operationalisierung der Metapher zur quantifizierenden Untersuchung deutschsprachiger literarischer Texte im Übergang vom Realismus zur Moderne» in: Tagungsband DFG-Symposium „Digitale Literaturwissenschaft" (Villa Vigoni, 09.-13.10.2017). Berlin: De Gruyter.

Herrmann, J. B. / Woll, K. / Dorst, A. G. (2019): "Linguistic Metaphor Identification in German" in: S. Nacey / A.G. Dorst / T. Krennmayr / W. G. Reijnierse (Hg.), MIPVU in Multiple Languages, Amsterdam: John Benjamins.

Heydebrand, R. / Winko, S. (1996): *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. Paderborn, Zürich [etc.]: Schöningh.

Köhler, M. (1999): *Wertung in der Literaturkritik: Bewertungskriterien und sprachliche Ausdrucksmöglichkeiten des Bewertens in journalistischen Rezensionen zeitgenössischer Literatur* (PhD Dissertation, Universitätsbibliothek der Universität Würzburg). https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/784

Lakoff, G. / Johnson, M. (1980): *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Narula, S. K. (2014): "Millions of people reading alone, together: The rise of Goodreads" in: *The Atlantic*. https://www.theatlantic.com/entertainment/archive/2014/02/millions-of-people-reading-alone-together-the-rise-of-goodreads/283662/

Nuttall, L. / Harrison, C. (2018): Wolfing down the Twilight series: Metaphors for reading in online reviews in: H. Ringrow / S. Pihlaja (Hg.), *Contemporary Media Stylistics*. New York: Bloomsbury Academic.

Peplow, D. / Swann, J. / Trimarco, P. / Whiteley, S. (2016): *The discourse of reading groups: Integrating cognitive and sociocultural perspectives*. London: Routledge.

Radway, J. A. (1986): "Reading is not eating: mass-produced literature and theoretical, methodological, and political consequences of a metaphor" in: *Book Research Quarterly*, 2/3: 7–29.

Steen, G. J. (2009): "Deliberate Metaphor Affords Conscious Metaphorical Cognition" in: *Cognitive Semiotics* 5/1-2: 179-197.

Steen, G. J. / Dorst, A. G. / Herrmann, J. B. / Kaal, A. A. / Krennmayr, T. / Pasma, T. (2010): A Method for Linguistic Metaphor Identification: From MIP to MIPVU. Amsterdam and Philadelphia: John Benjamins.

Stockwell, P. (2009): Texture: A cognitive aesthetics of reading. Edinburgh: Edinburgh University Press.

Tognini-Bonelli, E. (2001): Corpus Linguistics at Work. Amsterdam: John Benjamins.

Veale, T., Shutova, E. / Klebanov, B. B. (2016): "Metaphor: A Computational Perspective" in: *Synthesis Lectures on Human Language Technologies*, 9/1: 1–160.

Ikonizität als Erkenntnismittel – Vollständigkeit, Verständlichkeit und Kontextualisierung als Grundprinzipien der Visualisierung

Freyberg, Linda
linda.freyberg@gmx.de
Leuphana Universität Lüneburg; Fachhochschule Potsdam, Deutschland

Einleitung

Das hohe Aufkommen und die vermehrte Relevanz digitaler Information haben den Bereich der Erschließung, Organisation und Vermittlung von Wissen nachhaltig verändert. Vilém Flusser wies bereits 1978 auf das „Ansteigen der Wichtigkeit von zweidimensionalen Codes“ (Flusser 1998: 22) in unserer Kultur hin, womit er die Kommunikation mittels Oberflächen im Gegensatz zu den linearen Medien der „eindimensionalen Codes“, wie das Alphabet meinte. Diese Oberflächen sind graphisch konstituiert und auch das Web selbst ist ein visuelles Medium auf struktureller Ebene. In diesem Sinne konstatiert auch Horst Bredekamp: „Die hochtechnisierter Gesellschaften durchleben eine Phase der kopernikanischen Wende von der Dominanz der Sprache zur Hegemonie des Bildes.“ (Bredekamp 2000: 102). Die Beschreibung und Analyse des Phänomens Bildlichkeit erfordert eine Bildtheorie, die sowohl in der Lage ist, traditionelle Bildformen als auch digitale Bilder zu adressieren und alle gesellschaftlichen Anwendungsbereiche, von künstlerischen über alltäglichen hin zu wissenschaftlichen Ausdrucksformen einzuschließen oder wie John Michael Krois es formuliert: „Eine Bildtheorie können wir dadurch testen, dass wir nachschauen, ob sie die sonderbarsten Eigenschaften von Bildern verständlich machen kann.“ (Krois 2011: 140)

Der Schwerpunkt dieses Beitrages liegt auf der Rolle der Bildlichkeit innerhalb des Erkenntnisprozesses in digitalen Umgebungen. Generell wird eine Ikonizität aller Arten von Information angenommen, wobei hierbei sowohl eine implizite Dimension auf einer strukturellen Ebene sowie explizite Ausdrucksformen wie Visualisierungen, die topologisch Relationen darstellen, gemeint sind. Insbesondere wird in diesem Vortrag die Methode der Informationsvisualisierung als Erkenntnismittel für die geisteswissenschaftliche Forschung dargestellt. Diese Forschung agiert an der Schnittstelle zwischen den Geisteswissenschaften (speziell der Kunstgeschichte, Kultur- und Medienwissenschaften) und der Informatik und ist somit als theoretische Grundlagenforschung der Digital Humanities (DH) aufzufassen. Die Digital Humanities sind laut Markus Schoepf „ein junges Feld innerhalb der Geisteswissenschaften und Technologien, das noch nicht klar definiert ist. Die digitalen Geisteswissenschaften nutzen die Vorteile der Anwendung mathematischer Methoden zur Analyse kultureller Phänomene.“ (Schoepf 2012) Visualisierung stellt eine solche Methode dar, die auf statistischen oder generell abstrakten Daten basiert und sich durch den vermehrten Einsatz in den digitalen Geisteswissenschaften als Darstellungs- und Analyseinstrument der DH etabliert hat. Die Methoden können hierbei jedoch nicht unreflektiert übernommen werden, unter anderem da die etablierten Darstellungskonventionen nicht in der Lage sind, vielschichtige geisteswissenschaftliche Fragestellungen darzustellen. Nach Johanna Drucker stellt die Anpassung der digitalen Werkezeuge an geisteswissenschaftliche Forschung ein grundlegendes Ziel der DH dar. (siehe Drucker 2011: 2)

Visualisierungen reichen von Abbildungen über Modelle bis hin zu Simulationen. Sie können unter anderem gezeichnet, fotografiert, geometrisch konstruiert oder durch Sensorik vermittelt und digital prozessiert werden. Der Begriff Informationsvisualisierung bezieht sich auf die Darstellung abstrakter Daten und wird als "distanzierter Blick" auf große Informationsräume wahrgenommen. (siehe Drucker 2014: 7 und Glinka/ Dörk 2018: 236 ff.) Eine engere Definition von Informationsvisualisierung macht die graphische Datenverarbeitung zur Voraussetzung und sieht die Informationsvisualisierung als direkten Anschluss an traditionelle wissenschaftliche Visualisierungen. (siehe Wagner 2005: 57 ff.) Eine etwas breiter gefasste oftmals zitierte Definition von Informationsvisualisierung, die sich ebenso auf die Anwendung im Digitalen bezieht, lautet: „The use of computer-supported, interactive, visual representations of abstract data to amplify cognition.” (Card/Mackinlay/Shneiderman 1999: 7), wobei hier der Schwerpunkt auf der Abstraktion liegt, welche Informationsvisualisierungen von Visualisierung im Allgemeinen unterscheidet.

Bildlichkeit als Erkenntnismittel

Die Funktionen von Visualisierungen erstrecken sich von der Orientierung bis hin zur (hypothetischen) Voraussage und somit auch vom Überblick bis zur Evidenzsuggestion. Generell handelt es sich um vereinfachte (und vereinfachende) Darstellungen von teilweise sehr komplexen Sachverhalten, zu deren Verständnis sie beitragen sollen; daher können sie als Erkenntnismittel eingesetzt werden. Diese Funktion kommt ihnen nun nicht nur zu, weil sie – wie in den mittelalterlichen Mnemotechniken – als Gedächtnisstützen für bekannte Sachverhalte dienen, sondern resultiert vor allem aus ihrem Potenzial für die Entdeckung von *neuen* Zusammenhängen.

Für das Verständnis von Bildlichkeit als Erkenntnismittel fungiert das Werk von Charles Sanders Peirce, speziell seine Zeichentheorie sowie seine Theorie des *diagrammatic reasoning* als theoretische Grundlage.

Peirces universelle Zeichentheorie bildet einen breiten Analyserahmen für alle Bildarten und kann somit als Grundlage einer umfassenden Bildtheorie herangezogen werden. Er erweitert das semiotische Modell, bestehend aus Objekt – Zeichen – Interpretant um zahlreiche triadische Zeichenrelationen, die die drei grundlegenden Zeichenarten präzisieren (siehe Abbildung 1) und somit potentiell auf alle Arten von Phänomenen anwendbar machen.

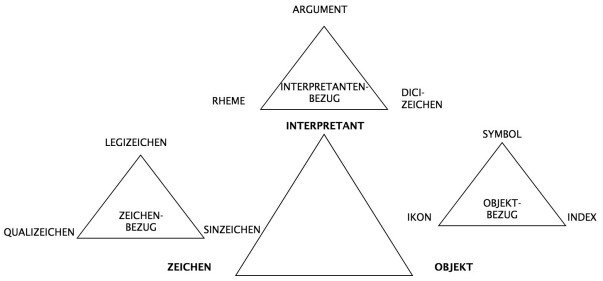


Abbildung 1: Die Peirceschen Zeichenkategorien

Dieses ausführliche Zeichenmodell soll auf historische sowie zeitgenössische Ansätze der Visualisierung angewendet werden, indem auf Zeichenebene, die Relationen der Repräsentation in den jeweiligen Visualisierungen analysiert werden. Dabei steht das ikonische Zeichen, als prinzipielle Kategorie der Bildlichkeit im Mittelpunkt.

Das Diagramm als spezifische Form des Ikon-Zeichens spielt in Peirces Werk eine zentrale Rolle vor allem aufgrund seiner Relevanz innerhalb des Erkenntnisprozesses. Diagramme sind nicht rein ikonische Zeichen, sondern besitzen aufgrund der kausalen Relation zu ihren repräsentierten Objekten auch einen indexikalischen Status und können gleichermaßen als Symbole fungieren, da Konventionen etabliert und angewendet werden. Somit handelt es sich um eine Hybridform des Zeichens. Peirces sehr weit gefasstes Konzept der Diagramme basiert auf der Grundannahme, dass "[a]ll necessary reasoning without exception is diagrammatic". (CP 5.162) Basierend auf diesen Prämissen entwickelte Peirce, das Konzept des *diagrammatic reasoning*, welches sich auf den Erkenntnisprozess generell (Abduktion – Deduktion und Induktion) bezieht und grundsätzlich feststellt: „What purpose are the diagrams fitted to subserve? They may help to analyze reasonings, and this either in a practical way by aiding a person in rendering his ideas clear, or theoretically". (CP 4.355.)

Durch die Betrachtung von Mustern und das Erkennen von Ähnlichkeiten, die in Frederik Stjernfelts Annahme "the very source of new ideas" (Stjernfelt 2007: 77) sind, wird der Erkenntnisprozess unterstützt oder nimmt eine neue Richtung. Im Kontext des *diagrammatic reasoning* identifiziert Hoffmann einen kreativen Teil, nämlich in Peirces Konzept des „theoric reasoning", welches „the power of looking at facts from a novel point of view", also das Einnehmen einer neuen Perspektive, bezeichnet. (siehe Hoffmann 2003: 138). Diese Möglichkeit bieten Visualisierungen: Im Anwendungsbereich der digitalen Kulturdaten werden multidimensionale Zugänge auf Objekte ermöglicht, die nach bestimmten Aspekten geordnet und im besten Fall dynamisch präsentiert werden. Die Objekte können dabei sowohl in einem Gesamtkontext als auch in semantischer Relation zu anderen Objekten dargestellt werden. Diese Abstraktion der Objekte in übergeordnete Dimensionen ermöglicht sowohl einen Überblick, auch über große Datenmengen, zu gewinnen als auch konkrete Forschungsfragen zu beantworten. Die digitale Repräsentation von Objekten und deren semantische Umgebung bieten also eine neuartige Art der Kontextualisierung und die Möglichkeit, sich fundiertes Wissen über ein Objekt anzueignen. Darüber hinaus kann mit dem Mittel der Visualisierung eine große Anzahl von Objekten und deren Relationen, z.B. die Zugehörigkeit zu bestimmten Epochen, dargestellt werden und zusätzliche Informationen abgeleitet werden. Die Unterscheidung von Peirce zwischen einem *type* und einem *token* kann in diesem „Zusammenhang angewendet werden, um zwischen wiederkehrenden Mustern und einzelnen Elementen zu unterscheiden. Peirce definiert ein token als „a single object or thing which is in some single place" und einen type als „definitely significant form". (CP 4.537/siehe auch Bakker/Hoffmann 2005). Für die Forschung sind beide Aspekte gleichermaßen relevant, denn das Zusammenspiel von Mustern und Singularitäten führt zum Beispiel zum Verständnis von Epochen, Kunstwerken und stilistischen Entwicklungen. Wenn type und token in einer Visualisierung miteinander in Beziehung gesetzt werden können, können Struktur und Genese der Objekte und ihr Kontext auf einen Blick erfasst werden. erkannt werden. Dieses Mittel der digitalen Repräsentation kann also helfen, die Objekte als Konfigurationen von Singularität und Replikation, d.h. von Differenz und Wiederholung zu verstehen.

Ansätze der Visualisierung

Die wichtigsten Operationen der Visualisierung als Repräsentationsform sind Simplifikation und Komplikation. Bezogen auf die Intention der Darstellung, soll die Komplexität eines Datensatzes entweder reduziert werden oder zielt auf Vollständigkeit, wobei zu berücksichtigen ist, dass eine Visualisierung eine Repräsentation und keine Replikation ist und daher immer eine gewisse Reduktion der Komplexität einhergeht. Diese Ausdrucksform changiert zudem zwischen Effektivität und Expressivität (siehe Abb. 2), wobei sich diese beiden Aspekte nicht zwangsläufig graduell zueinander verhalten.

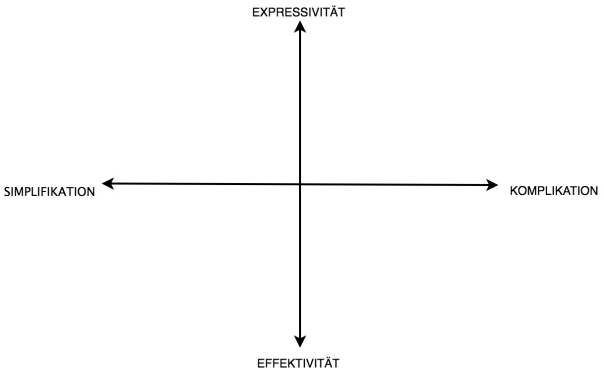


Abbildung 2: Operationen der Visualisierung

In diesem Vortrag werden historische Ansätze vorgestellt und als Modelle betrachtet, wie man Wissen ordnen, darstellen und visuell präsentieren kann. Prinzipiell ist Begriff der Visualisierung in diesem Beitrag sehr weit gefasst und bezeichnet sowohl einfache topologische Anordnungen im physischen Raum als auch elaborierte multidimensionale Visualisierungen im Digitalen. Abhängig vom Kontext und der Intention der Visualisierung werden als die drei Hauptmodi Vollständigkeit, Verständlichkeit und Kontextualisierung vorgeschlagen. Diese grundsätzlichen Ansätze der Visualisierung werden mit zeitgenössischen Beispielen in Bezug gesetzt, die auf den gleichen Prinzipien beruhen. Hierbei sollen einerseits die semantischen Repräsentationsbeziehungen auf Zeichenebene der jeweiligen Visualisierung untersucht und andererseits die historische Kontinuität der Darstellungsformen analysiert werden.

Der erste Ansatz, der der Vollständigkeit, findet beispielsweise im Bereich der Dokumentation oder im Bibliothekswesen Anwendung. Die Vision, das Wissen der Welt in einem großen Wissensorganisationssystem darzustellen, spiegelt sich bereits in der Idee einer umfassenden Enzyklopädie beispielsweise von Denis Diderot oder in dem Projekt einer universellen Klassifikation im 19. und 20. Jahrhundert wieder. Diese Ideen wirken immer noch in den Entwicklungen der digitalen Wissensorganisation nach. Speziell die Versuche von Paul Otlet und Henri La Fontaine, ein universelles Repositorium aufzubauen, und Otlets Idee einer umfassenden Visualisierung von Wissen erscheinen aus heutiger Sicht visionär und gleichermaßen wird deren Realisierung mit digitalen Methoden umfassend umsetzbar. Im Bereich der kulturellen Sammlungen verfolgen vor allem Institutionen wie Nationalbibliotheken oder große Museen mit einem breiten Sammlungsschwerpunkt, diversen Objektarten und unterschiedlichster inhaltlicher Kontexte, die Intention, ihre Bestände vollständig zu präsentieren, so dass möglichst alle Objekte sichtbar oder auffindbar sind. Als zeitgenössisches Beispiel für diesen Ansatz werden unter anderem die Visualisierungen des Urban Complexity Labs der Deutschen Nationalbibliothek (UCLAB 2017) sowie der Deutschen Digitalen Bibliothek herangezogen. Durch die Einbindung der Objekte in ein übergeordnetes Bezugssystem zum Beispiel einer Klassifikation sind bei den Beispielen hauptsächlich konventionelle generalisierte Zeichenformen wie Legizeichen, Symbole oder Argumente involviert.

Den Zugang zu Wissen zu vereinfachen und Informationen niedrigschwellig zugänglich zu machen, unabhängig von sprachlichen oder sozialen Barrieren findet sich beispielsweise bei Otto Neurath wieder. Insbesondere durch seine universelle Bildsprache (Isotype), die als Modell für eine visuelle Wissensvermittlung fungiert, die mit Vereinfachung operiert, um Wissen pragmatisch zu präsentieren. Dieser Ansatz findet sich heute vielfach in der aktuellen digitalen Kommunikation in Emoticons und Icons wieder, die als Symbole auf Sachverhalte oder Gefühlszustände verweisen. Der Bereich der Infographiken weist eine historische Kontinuität auf, da die Darstellungskonventionen sich nicht entscheidend verändert haben und die Graphiken weiterhin in nicht-digitalen Formaten breite Verwendung finden. Auf Repräsentationsebene sind die Graphiken hauptsächlich als Ikon-Zeichen einzuordnen, da sie auf eine visuelle Ähnlichkeit mit dargestellten Sachverhalten abzielen und deren Objekteigenschaften teilen.

Zuletzt werden Aby Warburgs Ideen zur Organisation multimodalen Wissens als Modell für den Ansatz der Kontextualisierung präsentiert. Einerseits fungiert die topologische und strukturelle physische Ordnung der Kulturwissenschaftlichen Bibliothek Warburg (K.B.W.) als Beispiel für semantische Kontextualisierung und die damit verbundene Erforschung guter Nachbarschaften sowie die Ermöglichung neuer Entdeckungen (*Serendipity*). Darüber hinaus stellt der Mnemosyne-Bildatlas ein Beispiel für eine anspruchsvolle Visualisierung dar, die die Grenzen von Zeit und Raum überwindet, indem sie Beständigkeit und Übergang gleichzeitig darstellt und verschiedene Arten von Relationen ausdrückt sowie den Forschungsverlauf dokumentiert. Hier fungieren als zeitgenössische Beispiele dynamische Visualisierungen des UCLABs, einmal der Vikus Viewer (UCLAB 2017/18), der explorativ kulturelle Sammlungen erschließbar machen, intuitive Zugänge auf Museumsobjekte zum Beispiel durch farbliche Anordnungen (Vane 2018) sowie das Projekt „Meta-Image“ (Meta-Image 2009-11), welches durch die Beschreibung, Annotation und Neuanordnung von Bildern und Bilddetails analog zum Warburgschen Bildatlas, den Forschungsprozess visualisiert und unterstützt.

Zusammenfassend sind Visualisierungen in der Lage, „to expand perception by adding understanding beyond that of the textual narrative or data“. (Smiraglia 2015: 42) Die Funktion des Verstehens bzw. des Vermittelns beruht auf der Eigenschaft von Visualisierungen, dass Beziehungen zwischen den Objekten oder bestimmten Entitäten visuell ausgedrückt werden und dadurch topologische und morphologische Strukturen entstehen: „A single dot in a painting does not mean anything at all. [...] The accumulation of dots builds a form, which has meaning, which can be recognized”. (Warnke/Dieckmann 2016: 113) Diese Prinzipien gelten gleichermaßen für die historischen Beispiele und zeitgenössische Datenvisualisierungen und somit weisen die gezeigten Darstellungsmodi und Hauptansätze der Visualisierung eine historische Kontinuität auf, so die zugrundeliegende These. Einige Ideen lassen sich im Digitalen jedoch weitaus übersichtlicher und dynamischer präsentieren, vor allem im Falle der Ansätze der Vollständigkeit und der Kontextualisierung, wie in dem Vortrag herausgearbeitet werden soll. Zudem soll aufgezeigt werden, dass sich Peirces umfassende Zeichentheorie hierbei als Analyseinstrument eignet, um die Funktions- und Wirkungsweise von Visualisierungen zu klären.

Bibliographie

Bakker, Arthur / Michael H. G. Hoffmann (2005): “Diagrammatic Reasoning as the Basis for Developing Concepts: A Semiotic Analysis of Students’ Learning about Statistical Distribution”, in: *Educational Studies in Mathematics* 60, no. 3 (November 2005): 333–58.

Bredekamp, Horst (2000): *Antikensehnsucht und Maschinenglauben.* Die Geschichte der Kunstkammer und die Zukunft der Kunstgeschichte. Berlin: Wagenbach.

Card, Stuart K. / Mackinlay, Jock D. / Shneiderman, Ben (1999): “Information Visualization”, in: *Readings in Information Visualization.* Using Vision to Think. San Francisco: Morgan Kaufmann: 1-34.

Drucker, Johanna (2014): *Graphesis* . Visual Forms of Knowledge Production. metaLABprojects. Boston: Harvard University Press.

Drucker, Johanna (2011): “Humanities Approaches to Graphical Display”, in: *DHQ* . Digital Humanities Quarterly, Volume 5, No.1, 1-21.

Flusser, Vilém (1998): *Kommunikologie* . Frankfurt/M.: Fischer.

Glinka, Katrin / Dörk, Marian (2018): „Zwischen Repräsentation und Rezeption. Visualisierung als Facette von Analyse und Argumentation in der Kunstgeschichte“, in *: Computing Art Reader* . Einführung in die digitale Kunstgeschichte: 234-50.

Hoffmann, Michael (2003): “Peirce's "Diagrammatic Reasoning" as a Solution of the Learning Paradox” in: Debrock, G. (eds.): *Process Pragmatism.* Essays on a Quiet Philosophical Revolution. Amsterdam, New York: Rodopi: 138.

Krois, John M. (2011): “Für Bilder braucht man keine Augen. Zur Verkörperungstheorie des Ikonischen“ in: Bredekamp, Horst/Lauschke, Marion (eds.): *John M. Krois. Bildkörper und Körperschema*. Actus et Imago, Berliner Schriften zur Bildaktforschung und Verkörperungsphilosophie, Band II. Berlin: Akademie Verlag: 132-161.

Meta-Image (2009-11): http://www2.leuphana.de/meta-image/Idee.php [letzter Zugriff 27. September 2019].

Peirce, Charles Sanders (1931-35/1958): *The Collected Papers of Charles Sanders Peirce* . Vols. I-VI ed. Charles Hartshorne and Paul Weiss. Harvard University Press: Cambridge, MA 1931-1935; Vols. VII-VIII ed. Arthur W. Burks Harvard University Press: Cambridge, MA 1958. *Zitierform (CP Nr. des Papers)*

Schoepf, Markus (2012): https://whatisdigitalhumanities.com/ [letzter Zugriff 27. September 2019].

Smiraglia, Richard P . (2015): *Domain Analysis for Knowledge Organization* . Tools for Ontology Extraction. Witney, Oxfordshire: Chandos Publishing.

Stjernfelt, Frederik (2007): *Diagrammatology.* An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics. Amsterdam: Springer Netherlands.

UCLAB (2014): Deutsche Digitale Bibliothek visualisiert https://uclab.fh-potsdam.de/ddb/ [letzter Zugriff 27. September 2019].

UCLAB (2017): DNBVIS, https://uclab.fh-potsdam.de/projects/dnbvis/ [letzter Zugriff 19. Dezember 2019].

UCLAB (2017/18): Vikus-Viewer, https://uclab.fh-potsdam.de/projects/vikus-viewer/ [letzter Zugriff 27. September 2019].

Wagner, Kirsten (2005): "Computergrafik und Informationsvisualisierung als Medien visueller Erkenntnis", in: *IMAGE: Bildwissenschaft als interdisziplinäres Unternehmen* . Eine Standortbestimmung, AUSGABE 1/2005: 46-63.

Vane, Olivia (2018): "Dive into color. Visualising the Cooper Hewitt collection by colour and time" http://olivia-vane.co.uk/research/CooperHewitt2.html [letzter Zugriff 19. Dezember 2019].

Warnke, Martin / Dieckmann, Lisa (2016): "Prometheus meets meta-image: implementations of aby warburg's methodical approach in the digital era", in: *Visual Studies* 31(2) 2016: 109-120.

Integrating user-specified Knowledge for semi-automatic Coreference Resolution

Schmidt, David
david.b.schmidt@uni-wuerzburg.de
Universität Würzburg, Deutschland

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Introduction

Coreference Resolution is a challenge which has seen the interest of researchers for half a century, but at the current point in time, even state-of-the-art algorithms (Joshi et al., 2019)cannot produce reliable results when applied in a fully automatic manner. The main problem of an automatic coreference resolution system is that decisions that occur late in the text can heavily influence prior decisions and so the resulting clustering and its composition is hard to understand. We found that, when applying the end-to-end coreference resolution algorithm (Lee et al., 2017), the algorithm tends to reset its understanding of the text every couple of paragraphs, which results in a mixture of grave errors when aggregated over the document. Coreference resolution is a necessary and very important step however, when analysing the content of a literary text. Different engines detect knowledge on a local level and coreference resolution aggregates this knowledge to the document scope of the text. This means that, without a reliable coreference resolution module, most applications that require an aggregated view of the texts (which is very important for most distant reading experiments) cannot be researched efficiently. In this work, we try to overcome the challenge of coreference resolution by presenting a semi-automatic mechanism instead of a fully automatic system. This mechanism allows the users to integrate their prior knowledge about characters of a literary text and their relations. We do this by parsing this knowledge into a machine-readable data structure and integrate this knowledge into our rule-based coreference resolution system, which was extended from (Krug et al., 2015).

Related Work

There have been various works trying to integrate world knowledge into coreference resolution. Ng (2007) added several new features to a coreference resolution system based on machine-learning, e.g. a feature for the semantic similarity of mentions and a feature based on patterns extracted from the training data. Others used knowledge that was extracted from knowledge bases like YAGO (Suchanek et al., 2007), among them Bryl et al. (2010), who model synonyms and mention types using logical constraints in a Markov Logic Network (Richardson and Domingos, 2006), and Rahman and Ng (2011), who extract relation triples between two mentions and convert them into features for their algorithm.

Aralikatte et al. (2019) use knowledge bases in a slightly different way. They apply the neural coreference resolution system of Lee et al. (2018) in a reinforcement learning setting and reward it, the more valid relations can be extracted from it. A valid relation is evaluated against knowledge bases like Wikidata and Wikipedia.

Unfortunately, these approaches are impractical for historic novels since the knowledge bases used there mostly contain knowledge about real-world entities while the novels deal with fictional characters.

Method

Our approach basically allows the user to model the knowledge which Wikipedia or Wikidata contain about real-world

entities for fictional characters. The knowledge is represented as character sheets. For each character, the user needs to determine a unique name (e.g. Richard Landsfeld) and can optionally provide a first name (Richard), last name (Landsfeld), the character's gender (male), a list of strings which are used as synonyms for the character (e.g. Baron von Landsfeld) and a list of strings which do not refer to this character. In addition to that, the user can specify the character's relations to other characters by providing the name of the other character and a list of possible labels for this character (e.g. Lydia - Ehefrau/Gattin).

The knowledge provided by the user is used in several different sieves of the algorithm. First name, last name and gender are used in a sieve which already existed prior to this work. It uses first and last names to merge clusters if they are compatible with respect to several other meta data fields (like gender). For this work, it was expanded to also merge clusters which contain mentions that have been identified to belong to the same character from user-specified knowledge. This identification is done in one of the first sieves by comparing the mentions' texts to a character's unique name, first name, synonyms and, if no other character has the same last name, last name. Mentions that have been identified to belong to a character this way are from then on prevented from being merged into a cluster together with mentions belonging to other characters as well as mentions which have one of the character's non-coreferent strings as their text.

The relations contained in the user-specified knowledge are used in one of the last sieves of the algorithm. It looks for constructions where one mention is (a) a possessive, demonstrative or relative pronoun used as an attribute, (b) a genitive or (c) preceded by a token 'von' which means that one mention is a prepositional modifier of another, like 'seine Ehefrau', 'Richards Ehefrau' or 'Ehefrau von Richard'. These constructions can be used if the first mention belongs to a character and this character has a relation that has the text of the second mention as a possible label (in our example, the character 'Richard' needs a relation that contains the label 'Ehefrau'). If this is the case, it can be determined that the second mention ('Ehefrau' in the example) belongs to the character which is the target of the relation. The cluster of the second mention is therefore merged into another cluster which has previously been identified to belong to the target character (ideally, there should only be one such cluster left by the time this sieve is applied).

Finally, we use the knowledge about relations for speculative merges of mentions that have a relation word as their text and are not in a cluster with any mention which is recognised as a name. If we find one of these mentions we go backwards in the text until we encounter a mention which belongs to a character that is the target of a relation with the relation word as a possible label (e.g. if we find a mention 'Ehefrau', encounter a mention which belongs to the character 'Lydia' and know that another character 'Richard' has a relation labelled 'Ehefrau' with 'Lydia' as the target, we merge both mentions' clusters).

Results and Discussion

We evaluated our approach on six documents that were randomly picked from the documents of DROC [1] (Krug et al., 2018) for which we have summaries: *Die Hosen des Herrn von Bredow* by Willibald Alexis, *Stilpe* by Otto-Julius Bierbaum, *Der Stechlin* by Theodor Fontane, *Amerika* by Franz Kafka, *Anna Karenina* by Lev-Nikolaevic Tolstoj and *Uli der Pächter* by Jeremias Gotthelf. For each of these documents, two annotators, who were unfamiliar with the texts, separately created a document for userspecified knowledge by first reading the corresponding summary and then skimming the actual document. The time required for the creation of the meta data was between 5 to 15 minutes per file. Table 1 shows some characteristics of this user-specified knowledge: the number of characters, the total number of synonyms, the total number of relations and the total number of relation labels.

Dokument	Characters	Synonyms	Relations	Labels
Alexis - Hosen (D)	7	14	8	12
Alexis - Hosen (M)	7	21	3	5
Bierbaum - Stilpe (D)	4	7	0	0
Bierbaum - Stilpe (M)	3	4	1	3
Fontane - Stechlin (D)	9	10	2	4
Fontane - Stechlin (M)	8	12	3	3
Kafka - Amerika (D)	5	5	0	0
Kafka - Amerika (M)	4	7	0	0
Tolstoj - Karenina (D)	9	14	7	11
Tolstoj - Karenina (M)	5	8	5	9
Gotthelf - Uli (D)	4	6	2	3
Gotthelf - Uli (M)	3	4	0	0

Tabelle 1: Characteristics of the user-specified knowledge created by the annotators: Number of characters, synonyms, relations and relation labels.

The results of the algorithm with this userspecified knowledge are depicted in table 2 alongside the baseline results (the algorithm without any additional knowledge). We report the scores of the MUC metric (Vilain et al., 1995) which evaluates based on links between mentions and we report the results of the B-Cubed metric (Bagga and Baldwin, 1998) which evaluates based on cluster overlap between system entities and gold entities.

	MUC				B^3			
Dokument	P	R	F1	Δ	P	R	F1	Δ
Alexis - Hosen	84.89	71.74	77.76	-	74.87	29.01	41.81	-
Alexis - Hosen (D)	85.45	75.00	79.88	+2.12	73.02	32.48	44.96	+3.17
Alexis - Hosen (M)	84.33	73.1	78.31	+0.55	74.00	39.96	51.90	+10.09
Bierbaum - Stilpe	94.00	85.90	89.77	-	78.78	56.13	65.55	-
Bierbaum - Stilpe (D)	94.32	86.68	90.34	+0.57	78.10	59.89	67.80	+2.25
Bierbaum - Stilpe (M)	94.29	86.16	90.04	+0.27	78.06	59.21	67.34	+1.79
Fontane - Stechlin	92.29	83.44	87.64	-	83.24	45.79	59.08	-
Fontane - Stechlin (D)	92.33	83.88	87.90	+0.26	81.12	55.84	66.15	+7.07
Fontane - Stechlin (M)	92.40	84.75	88.41	+0.77	79.14	62.31	69.72	+10.64
Kafka - Amerika	94.75	88.11	91.31	-	85.17	65.59	74.11	-
Kafka - Amerika (D)	95.15	89.63	92.31	+1.00	85.01	76.02	80.26	+6.15
Kafka - Amerika (M)	95.11	89.02	91.97	+0.66	85.12	70.76	77.27	+3.16
Tolstoj - Karenina	84.73	79.49	82.03	-	64.94	51.00	57.13	-
Tolstoj - Karenina (D)	84.80	81.46	83.09	+1.06	61.54	53.37	57.17	+0.04
Tolstoj - Karenina (M)	84.57	80.06	82.25	+0.22	62.35	51.76	56.56	-0.57
Gotthelf - Uli	86.26	79.37	82.67	-	57.80	42.22	48.80	-
Gotthelf - Uli (D)	86.36	80.03	83.07	+0.40	58.26	41.52	48.48	-0.32
Gotthelf - Uli (M)	85.69	78.70	82.05	-0.62	63.06	42.67	50.90	+2.10
Average Improvement (D)	+0.25	+1.44	+0.90	-	-1.29	+4.90	+3.06	-
Average Improvement (M)	-0.09	+0.62	+0.31	-	-0.51	+6.16	+4.54	-
Average Improvement (Avg)	+0.08	+1.03	+0.61	-	-0.90	+5.53	+3.80	-

Tabelle 2: Results of the rule-based algorithm without user-specified knowledge and with user-specified knowledge provided by two different annotators (D and M) on six randomly picked documents of DROC. The last three lines show the average improvement of the annotators.

Depending on the text snippet, the improvements range from 0% up to almost 11% B-Cubed and up to 2% MUC score with an average improvement of about 4% B-Cubed. The small improvement of the MUC metric means, that with the help of the meta data, only relatively few links are improving, but these links reveal to be among the important ones, when the results of the B-Cubed metric is consulted. The improvements are mainly due to the improvements of the Recall, our algorithm is tuned to produce a conservative output and therefore does not attempt to merge references or entities that pose a

high risk of failure. The usage of meta data adds to the confidence for these merges and subsequently increases the Recall.

With user-specified knowledge at hand, we examined whether it just serves as an addition to our rule-based algorithm, or if it is able to replace the parts of our algorithm, which handle names and non-pronominal noun phrases (the algorithm cannot be replaced completely since user-specified knowledge does not help with pronouns). To assess this theory, we created the following algorithm: In the first step, all mentions which are identified to belong to the same character from user-specified knowledge (how this is done is described in the previous section) are merged into the same cluster. After that, only the parts of our algorithm which handle pronouns are applied. Table 3 shows the results of this string-matching baseline algorithm and the difference to the rule-based algorithm using the same user-specified knowledge. While the precision of the baseline is higher in all cases, its recall is lower, often by a rather large margin. This leads to the MUC score being slightly better in three cases but being worse in all other cases. The difference is even more noticable when looking at the B-Cubed scores: With two exceptions, they are always more than 5% worse than the results of the rule-based algorithm with user-specified knowledge. To summarize, one can say that the algorithm cannot be replaced by string-matching without a significant loss of quality.

	MUC				B^3			
Dokument	P	R	F1	Δ	P	R	F1	Δ
Alexis - Hosen (D)	92.41	61.96	74.18	-5.70	89.05	18.59	30.76	-14.20
Alexis - Hosen (M)	91.67	71.47	80.32	+2.01	83.10	37.08	51.28	-0.72
Bierbaum - Stilpe (D)	97.52	84.60	90.60	+0.26	84.97	51.25	63.94	-3.86
Bierbaum - Stilpe (M)	97.12	81.72	88.76	-1.28	82.23	41.65	55.29	-12.05
Fontane - Stechlin (D)	95.30	77.56	85.52	-2.38	88.50	33.51	48.61	-17.54
Fontane - Stechlin (M)	95.15	79.74	86.77	-1.64	85.72	38.11	52.76	-16.96
Kafka - Amerika (D)	96.15	81.10	87.99	-4.32	89.50	40.93	56.17	-24.09
Kafka - Amerika (M)	96.81	86.59	91.41	-0.56	87.38	50.11	63.69	-13.58
Tolstoj - Karenina (D)	89.22	75.28	81.66	-1.43	74.88	38.43	50.79	-6.38
Tolstoj - Karenina (M)	86.42	66.85	75.39	-6.86	74.16	29.01	41.70	-14.86
Gotthelf - Uli (D)	91.40	72.55	80.89	-2.18	78.74	29.76	43.20	-5.28
Gotthelf - Uli (M)	92.02	75.87	83.17	+1.12	75.22	32.12	45.01	-5.89

Tabelle 3: Results of the String-Matching baseline using the user-specified knowledge created by the two annotators (D and M) and the difference to the results of the rule-based algorithm using the same knowledge.

During the annotation of DROC, our experiments towards inter-annotator agreement revealed, that even human annotators only had an agreement of about 76% B-Cubed (Krug et al., 2018). Achieving a B-Cubed score of about 75% is therefore a milestone where the data seems reliable and we would expect the results to be usable for downstream tasks. An interesting aspect is that there is a large variance of improvement on different texts. Whenever relatively few named-entities that are communicating in a dialog are available in the text, the improvement is high (see *Amerika* or *Stechlin*) but the inverse effect occurs, when the author either is very vague with using names and aliases for characters or if there are many characters in the text in general. The quality of the results also depends on the summaries used. Using longer summaries (the ones we used were mostly rather short) or several different summaries per novel will likely lead to better results.

Fußnoten

1. Note that in DROC, only persons are annotated. Coreference resolution usually also deals with other entities.

Bibliography

Aralikatte, R., Lent, H. / Gonzalez, A. V. / Hershcovich, D. / Qiu, C. / Sandholm, A. / Ringaard, M. / Søgaard, A. (2019). Rewarding coreference resolvers for being consistent with world knowledge. *arXiv preprint arXiv:1909.02392*.

Bagga, A. / Baldwin, B. (1998): Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

Bryl, V. / Giuliano, C. / Serafini, L. / Tymoshenko, K. (2010): Using background knowledge to support coreference resolution. In *ECAI*, volume 10, pages 759–764. Citeseer.

Joshi, M. / Chen, D. / Liu, Y. / Weld, D. S. / Zettlemoyer, L. / Levy, O. (2019): Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Krug, M. / Puppe, F. / Jannidis, F. / Macharowsky, L. / Reger, I. / Weimar, L. (2015): Rule-based coreference resolution in german historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Krug, M. / Puppe, F. / Reger, I. / Weimer, L. / Macharowsky, L. / Feldhaus, S. / Jannidis, F. (2018): Description of a corpus of character references in german novels - DROC [Deutsches ROman Corpus]. In *DARIAH-DE Working Papers*. DARIAH-DE.

Lee, K. / He, L. / Lewis, M. / Zettlemoyer, L. (2017): End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Lee, K. / He, L. / Zettlemoyer, L. (2018): Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Ng, V. (2007): Shallow semantics for coreference resolution. In *IJcAI*, volume 2007, pages 1689–1694.

Rahman, A. / Ng, V. (2011): Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language TechnologiesVolume 1*, pages 814–824. Association for Computational Linguistics.

Richardson, M. / Domingos, P. (2006): Markov logic networks. *Machine learning*, 62(1-2):107–136.

Suchanek, F. M. / Kasneci, G. / Weikum, G. (2007): Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Vilain, M. / Burger, J. / Aberdeen, J. / Connolly, D. / Hirschman, L. (1995): A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Interpretationsspielräume. Undogmatisches Annotieren literarischer Texte in CATMA 6

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

Jacke, Janina
janina.jacke@uni-hamburg.de
Universität Hamburg, Deutschland

Spezifika literaturwissenschaftlichen Annotierens

Werden *Spielräume* in der hermeneutischen Textarbeit als legitime – jedoch gewissen Regeln unterliegende – Pluralität der Zugänge zu Literatur verstanden, zeichnet sich literaturwissenschaftliches Annotieren in mindestens drei Aspekten durch Spielräume aus: 1) Literarische Texte können in vielerlei Hinsicht erforscht werden (etwa strukturell, inhaltlich oder inhaltstranszendierend, vgl. Shusterman 1978; Folde 2015). Dabei unterscheiden sich oft inhaltlicher Fokus und Methode der Texterforschung (vgl. Danneberg 1999; Bühler 2003). 2) Aufgrund ihrer Ambiguität lassen sich literarische Texte selbst innerhalb eines Zugangs unterschiedlich verstehen (vgl. Jannidis 2003; Bauer et al. 2010). 3) Der Texterforschungs-Workflow kann je nach Forscher*in unterschiedliche Methoden-Phasen zyklisch durchlaufen (vgl. Nünning & Nünning 2010, 10–21; Puhl et al. 2015, 42–46).

Da DH-Tools möglichst an disziplinspezifische geisteswissenschaftliche Theorien, Methoden und Praktiken rückgebunden werden sollen (vgl. Sahle 2015), sollten digitale Zugänge zur Literaturerforschung diese Spielräume berücksichtigen. Undogmatisches Annotieren mit CATMA 6 (https://catma.de) ist eine Möglichkeit, wie dies umgesetzt werden kann. Das webbasierte kollaborative Annotations- und Analysetool CATMA – seit 2008 an der Universität Hamburg entwickelt mit derzeit gut 8.000 aktiven Nutzer*innen weltweit – integriert Textannotation, Analyse und Visualisierung innerhalb einer webbasierten Arbeitsumgebung vor dem Hintergrund einer konzeptionellen Rückbindung an Theorien der ('undogmatischen') hermeneutischen Texterforschung. Dies ist im Bereich der DH-Tools einmalig (vgl. Meister, im Erscheinen). Mit CATMA 6 werden innerhalb des DFG-Projektes forTEXT (https://fortext.net) neue Funktionalitäten[1] und ein noch intuitiver nutzbares Interface auf Basis einer grundlegend neu gestalteten, projektzentrierten Systemarchitektur zur Verfügung gestellt.

Wie CATMA 6 geisteswissenschaftlichen Anforderungen (und damit den genannten Spielräumen) gerecht zu werden sucht, soll anhand von vier Funktionskomplexen demonstriert werden: unterschiedlichen Annotationsmodi, Mehrfachannotation, Metaannotation und kollaborativem Annotieren. Dieser Beitrag kann somit auch als exemplarische Umsetzung der Forderung verstanden werden, einen Brückenschlag zwischen DH- und traditionell-geisteswissenschaftlichen Methoden zu schaffen.

Vom ersten Zugang zur komplexen Interpretation

Zum Verhältnis zwischen Annotation und Interpretation

Die Interpretation literarischer Texte wird gemeinhin als eine Kernaufgabe literaturwissenschaftlichen Arbeitens betrachtet. Regeln der Textinterpretationen oder Gütekriterien für Interpretationshypothesen sind aufgrund der Theorie- und Methodenvielfalt nicht eindeutig festgelegt. Zwei Überzeugungen scheinen jedoch über unterschiedliche Ausrichtungen hinweg in der literaturwissenschaftlichen Forschungsgemeinschaft (relativ) allgemein anerkannt: (a) Trotz der Pluralität zulässiger Interpretationen gibt es auch Interpretationen, die einem Text *nicht* angemessen sind. Und, damit zusammenhängend: (2) Interpretationen sollten (in gewisser Hinsicht) an das sprachliche Material des Textes angebunden sein.

Angesichts dieser Sachlage lässt sich leicht erkennen, warum die Methode der Annotation im Zusammenhang mit Interpretationen fruchtbar angewandt werden kann: Der Prozess des Annotierens geht mit textnahem Arbeiten einher, und Annotationen werden grundsätzlich bestimmten Textstellen zugewiesen. Damit ist Annotation besonders geeignet für kleinschrittige textdeskriptive bzw. -analytische Vorhaben, die dann eine Grundlage für Interpretationen liefern können. Interpretationen selbst werden dann – auch in DH-Projekten – häufig in Form zusammenhängender Texte erstellt, innerhalb derer auf textanalytische Ergebnisse (Annotationen) Bezug genommen wird. Hier kann Annotation demnach als Werkzeug von Interpretation gelten.

Es kann allerdings durchaus sinnvoll sein, auch Interpretationshypothesen selbst im Prozess des Annotierens zu entwickeln und als Annotationen festzuhalten. Denn zum einen verschwimmt sogar bei gemeinhin als deskriptiv geltenden Operationen wie der narratologischen Analyse otf die Grenze zu (inhaltsspezifizierender) Interpretation.[2] Zum anderen sollte darauf geachtet werden, auch Interpretationshypothesen möglichst an Textstellen rückzukoppeln. Geschieht dies nicht, bleibt der Übergang zwischen kleinschrittiger-analytischer Textbeschreibung und holistisch-synthetischer Interpretation letztlich oft unklar.

Wie groß die Spielräume im Rahmen von Annotation sein müssen – sowohl hinsichtlich des Grads der Formalisiertheit der Annotation (vgl. 2.2) als auch der Einigkeit unter verschiedenen Annotator*innen –, hängt entsprechend davon ab, ob es sich um deskriptiv-analytische Annotationen handelt, die als Vorarbeit für Interpretationen fungieren, oder um genuin interpretative Annotationen (vgl. 3).

Drei Annotationsmodi

Viele Annotationstools (bisher auch CATMA) ermöglichen ausschließlich die Annotation mithilfe von Tagsets, also hierarchisch gegliederten Kategorien. Dafür müssen Forschende allerdings schon ein formalisiertes Kategoriensystem haben, mit dem sie den Text untersuchen wollen. Annotation sollte abern auch zur noch unstrukturierten Textexploration nutzbar sein. Zudem sollte auch Interpretation selbst mithilfe von Annotation ermöglicht werden, was aber meist nicht (allein) unter Nutzung von Kategorien umsetzbar ist. In CATMA 6 werden deshalb folgende Annotationsmodi implementiert:

(1) *Highlight*: Die Highlight-Annotation dient zunächst ausschließlich der Hervorhebung einer interessanten Textstelle. Nutzerïnnen können die annotierte Passage als relevant auszeichnen, auch wenn sie noch keine konkrete Hypothese haben. Mithilfe der Analysefunktionen in CATMA können gehighlightete Passagen gesucht und als Liste angezeigt werden. So lassen sie sich beispielsweise mit anderen Annotationsmodi weiter annotieren, wenn Texterforschung und Interpretation weiter fortgeschritten sind.

(2) *Comment*: Im Comment-Modus können Textstellen frei kommentiert werden. Dies ermöglicht es, Gedanken zu einer Textstelle festzuhalten, ohne ein strukturiertes Konzeptrepertoire zu nutzen. So können auch nicht in Kategorien überführbare Interpretationen als Annotation umgesetzt werden. Geplant ist zudem, die Erstellung von Tagsets auf der Basis von Kommentaren zu vereinfachen, beispielsweise indem die Kommentartexte (teil-automatisiert) ausgewertet werden.

(3) *Annotation*: Hierunter fallen in CATMA tagbasierte Annotationen, bei denen Textstellen mithilfe hierarchisch gegliederter Konzeptontologien mit einem Tag versehen werden (vgl. Fig. 1). Die Annotationsmodi können iterativ ineinandergreifen und bilden damit auch auf dieser Ebene den sog. ‘hermeneutischen Zirkel’ der Texterschließung ab. Die tagbasierte Form der Annotation setzt am meisten Strukturierung und Formalisierung voraus und ist nicht für alle Formen interpretativer Annotationen nutzbar. Als strukturiertes Werkzeug bzw. als Heuristik für Interpretation ist sie dafür umso fruchtbarer. In CATMA können Tagsets frei erstellt und laufend verändert werden; die Erzeugung einer solchen Konzeptontologie führt dabei zu sehr textnahem Arbeiten und erfordert in produktiver Weise die Reflexion literaturwissenschaftlicher Theorien und Methoden. Inwieweit tagbasiertes Annotieren mit der von Spielräumen geprägten literaturwissenschaftlichen Erforschung von Texten kompatibel ist, wird im Folgenden erörtert.

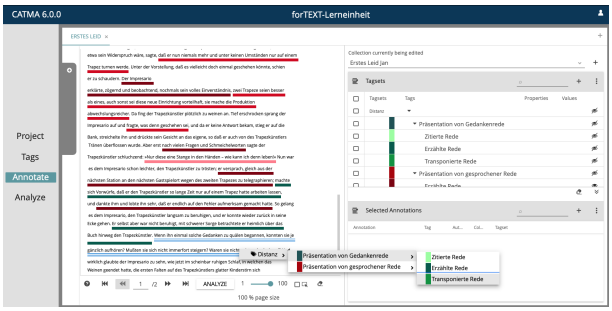


Figure 1: Annotation in CATMA 6

Voraussetzungen für gewinnbringendes taxonomiebasiertes Annotieren im Rahmen von Textauslegung

Damit ein strukturiertes Annotieren mit Tagsets nicht nur im Rahmen heuristischer Text*beschreibung* genutzt werden kann, sondern auch die Spielräume der Text*auslegung* abbildet,[3] müssen einige Bedingungen erfüllt sein.

Mehrfachannotation

Neben freiem Generieren und iterativem Überarbeiten von Tagsets ist eine Bedingung für die Nutzung tagsetbasierten Annotierens als interpretationsunterstützender Methode die Möglichkeit der (diversen oder sogar widersprüchlichen) Mehrfachannotation derselben Textstelle. Dies trägt zum einen dem Umstand Rechnung, dass ein Text aus unterschiedlichen Perspektiven untersucht werden kann: Beispielsweise kann eine Textpassage zugleich intermediale Bezüge enthalten und *Gender*-Themen adressieren. Eine mehrdimensionale Kategorisierung der Textstelle muss daher möglich sein.

Passagen literarischer Texte sind zudem häufig interpretationsoffen, weshalb unterschiedliche, teilweise auch widersprüchliche Interpretationen gleichermaßen gültig sein können. So mögen etwa (inkompatible) Thesen darüber, wer/was durch eine im Text auftretende Figur verkörpert werden soll, plausibel sein.[4] In CATMA 6 sind freie Tagsetgenerierung und Mehrfachannotationen einer Textpassage möglich, indem Textpassagen, Annotationen und Tags als Knoten in einer Graphstruktur modelliert sind, die sehr flexible Verknüpfungsmöglichkeiten erlaubt.[5]

Metaannotationen

Da bei der Interpretation von Literatur die Spielräume nicht grenzenlos sind und nach diversen Regeln gespielt werden muss (vgl. bspw. Jannidis 2003), benötigt eine Annotationsumgebung, die taxonomiegestütztes Interpretieren ermöglicht, auch Optionen zur Einordnung, Erläuterung und Aushandlung von Interpretationen. Diese Rolle erfüllen in CATMA 6 Metaannotationen, die wiederum taxonomiebasiert (als *Properties* und *Values*) bzw. als Metakommentar eingesetzt werden können:

Annotationskategorien lassen sich mit Properties versehen, denen pro Annotation feste oder ad hoc vergebbare Values zugeordnet werden können, um Annotationen genauer zu qualifizieren. Die gleiche Funktion erfüllen freitextbasierte Metaannotationen, die erstmalig in CATMA 6 nutzbar sind. Ob Metaannotationen als freie Kommentare oder auf Taxonomiebasis zur Anwendung kommen, kann vom Grad der theoretischen Ausarbeitung der genutzten Interpretationsheuristik abhängen oder eine Frage des Anwendungskontextes bzw. der persönlich präferierten Arbeitsweise sein.

In technischer Hinsicht sind Annotationen gemäß dem Web Annotation Data Model[6] modelliert und haben als *body-type* die Klasse *Dataset*. Die Struktur ist eine Liste von *key/multi-*

value-Paaren. Der Tag der Annotation gibt die möglichen *keys* vor.

Während Metaannotationen eingesetzt werden können, um einem Tagset Analysekategorien auf einer horizontalen Gliederungsebene hinzuzufügen,[7] lassen sie sich auch zur Einordnung der Interpretationsentscheidung nutzen. Forscherïnnen können beispielsweise angeben, welche Literatur- oder Interpretationstheorie (z. B. Rezeptionsästhethik oder Poststrukturalismus, vgl. Köppe & Winko 2013) sie herangezogen haben, um eine bestimmte (strittige) Interpretationsentscheidung zu treffen. Ebenso können in die Interpretation einfließende Kontextinformationen aufgeführt (z. B. Wissen über andere Texte), oder Interpretationen auf einer Sicher-unsicher-Skala verortet werden (vgl. Drucker 2011). Solche Metaannotationen helfen, hermeneutische Annotationen im Kontext theoretischer und subjektiver Einbettung zu verstehen; sie ermöglichen, zumindest in Ansätzen, die Angabe von Argumenten für Interpretationsentscheidungen und schaffen die Bedingungen einer literaturwissenschaftlichen Auseinandersetzung über die Plausibilität interpretativer Hypothesen. Besonders sind Metaannotationen notwendig, wenn Textstellen tatsächlich mehrere scheinbar widersprüchliche Annotationen aufweisen – speziell im Kontext kollaborativer Annotation und Textauslegung.

Kollaborative Annotation

Kollaboratives Annotieren ist in der Linguistik eine etablierte Methode, um Annotationsentscheidungen abzusichern (vgl. Wissler et al. 2014). In der Literaturwissenschaft ist es noch wenig etabliert (vgl. Röcke 2016); auch ist ein behutsameres Vorgehen angebracht, wenn es darum geht, Annotations- bzw. Interpretationsentscheidungen abzusichern. Als fruchtbar hat sich ein iteratives Vorgehen erwiesen, bei dem Forscherïnnen diskrepant annotierte Passagen diskutieren, um Gründe für unterschiedliche Entscheidungen herauszustellen (vgl. Gius & Jacke 2017). Durch eine gründliche Metaannotation kann dieser Workflow verschlankt werden. Je nach Grund kann abgewogen werden, ob es sich um eine legitime Uneinigkeit handelt. So können Interpretationsspielräume bei kollaborativem Annotieren zugleich gewahrt und sinnvoll eingegrenzt werden.

Kollaboration wird in CATMA ermöglicht durch die mit GitLab per API verknüpfte projektzentrierte Systemarchitektur (vgl. Fig. 2). Eine GitLab-Group[8] wird im CATMA Web UI als CATMA-Projekt gespiegelt, dem – entsprechend dem GitLab-Schema – einzelne Nutzerïnnen mit unterschiedlichen Rollen und Rechten hinzugefügt werden können. CATMA-Projekte werden mit Textdokumenten, Tagsets, Annotationsdaten und potentiell mehreren Projektmitgliedern ausgestattet. Da unterschiedliche Projektkontexte (etwa wissenschaftliche Forschungsprojekte mit mehreren Projektleiterïnnen, Mitarbeiterïnnen und Hilfskräften; Seminarprojekte in der universitären Lehre; Unterrichtsprojekte in der schulischen Lehre) die Festlegung unterschiedlicher Entscheidungsspielräume für die Mitarbeitenden erfordern, können in CATMA 6 jeweils projektbezogen die folgenden Rollen vergeben werden: Projekteigentümerïn/-leiterïn, Partnerïn, Assistentïn, Beobachterïn und Studentïn/Gast. Die Rollen sind dabei mit festen Rechte-Konfigurationen in den Feldern der Projekt- und Mitgliederverwaltung sowie der Erstellung, Bearbeitung und Löschung von Textdokumenten, Tagsets und Annotationsdaten versehen. Durch die Individualisierung von Kooperationsmodi kann festgelegt werden, wie viel Spielraum jedes Projektmitglied haben soll. Arbeitet man beispielsweise in einem Forschungsprojekt kollaborativ und möchte, dass alle Teilnehmenden die gleichen Rechte haben, vergibt man als Projekt-Owner Partner-Rollen. In Seminarkontexten könnte man Assistierenden-, Beobachtenden- oder Studierenden-Rollen vergeben, die jeweils weniger Zugriffsrechte haben.

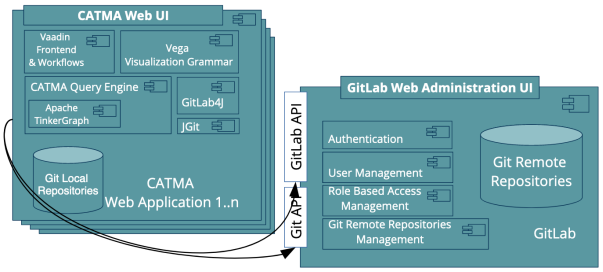


Figure 2: Systemarchitektur CATMA 6

Das neue Rollen- und Rechtesystem erlaubt somit eine differenzierte Festlegung von Spielräumen auch bei der Konzeption eines kollaborativen Annotationsprojekts. Um einem Projekt weitere Mitglieder hinzuzufügen bietet CATMA 6 zwei Möglichkeiten: (1) das manuelle Hinzufügen einzelner CATMA-Nutzerïnnen durch Eingabe des CATMA-Usernames. Diese Funktion bietet sich für den asynchronen Arbeitsmodus etwa in Forschungsprojekten an. (2) Die *live*-Einladung per generiertem Zahlencode, der nur für den Moment einer geöffneten Einladung gültig ist. Dies ist besonders für Seminar- oder Workshopkontexten eine zeitsparende Option.

Die Verknüpfung mit GitLab bietet zudem Versionierungs- und damit einhergehende Konfliktlösungsfunktionen. Denn während CATMA beispielsweise inhaltlich widersprüchliche Mehrfachannotationen erlaubt, stellt das zeilenbasiert arbeitende Versionierungssystem Git einen Konflikt fest, wenn Nutzerïnnen inkompatible Änderungen in ihren Projekten vorgenommen haben, die dieselbe Zeile des zugrundeliegenden Codes betreffen. Nehmen wir beispielsweise an, in einem kollaborativen Forschungsprojekt wurde dieselbe Metaannotation von Nutzerïnnen 1 und 2 je unterschiedlich geändert. Sobald Nutzerïn 2 die Arbeit mit dem Team synchronisiert, meldet CATMA einen Konflikt, anstatt eigenmächtig einer Version den Vorzug zu geben (siehe Fig. 3). Dabei werden eigene und fremde Version nebeneinandergestellt und Nutzerïn 2 kann sich informiert für eine Version entscheiden, ohne tiefergehende Kenntnisse über die zugrunde liegenden technischen GitLab-Prozesse haben zu müssen. Diese Funktionalität unterstützt den – im kollaborativen Modus noch stärker im Vordergrund stehenden – diskursiven Aushandlungsprozess von Annotation und Interpretation und reagiert somit auf die Forderung nach flexiblen disziplinspezifischen Arbeitsabläufen.

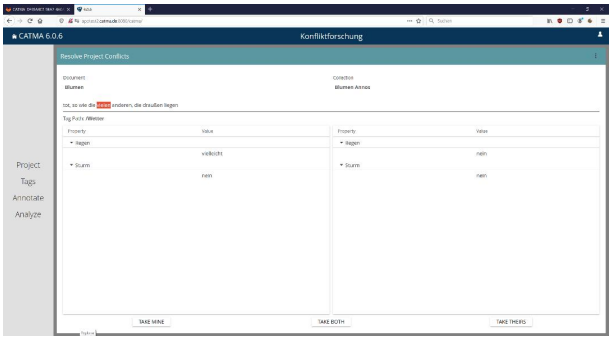


Figure 3: GUI-Unterstützung zur Konfliktlösung

Das Schaubild (vgl. Fig. 4) verdeutlicht die identifizierten literaturwissenschaftlichen Spielräume (linke Spalte) und die daraus erwachsenden generellen Anforderungen an digitale Arbeitsumgebungen (mittlere Spalte). Wie CATMA 6 diese Anforderungen konkret umsetzt, findet sich in der rechten Spalte.

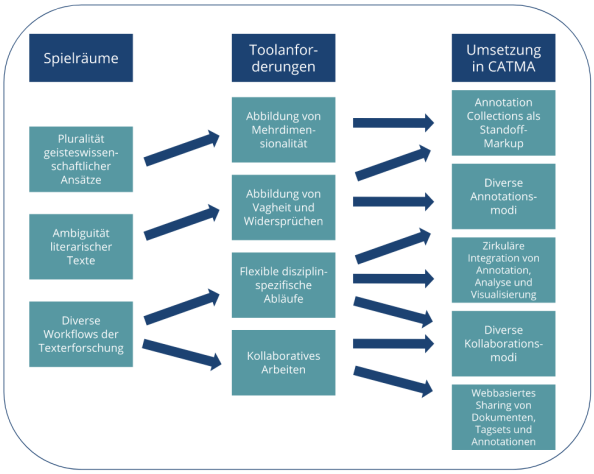


Figure 4: Technische Umsetzung der Anforderungen interpretatorischer Spielräume

Fußnoten

1. Dazu gehören u.a. umfangreiche Versionierungsfunktionalitäten und ein Rollen-/Rechtemanagement (s. Abschnitt 3.3) sowie die Integration eines geisteswissenschaftlich orientierten interaktiven Visualisierungskonzepts auf Grundlage der *Vega Visualization Grammar* gemäß der im Projekt 3DH (http://threedh.net) formulierten Kriterien einer *Dynamic Data Visualisation and Exploration for Digital Humanities Research*.
2. Dies gilt zum einen für narratologische *histoire*-Kategorien, aber auch für Kategorien, die das Verhältnis zwischen Darstellung und Handlung betreffen (vgl. Genette 2010). Je nach Gestaltung eines literarischen Textes können die inhaltlichen Komponenten, die für eine Kategorisierung identifiziert werden müssen, offen bleiben oder ambig umgesetzt sein, so dass ihre Identifikation eine Interpretation notwendig macht.
3. Zur Unterscheidung von Deskription und Interpretation vgl. Kindt & Müller 2003.
4. Vgl. hierzu bspw. Føllesdal 1979, der unterschiedliche Deutungen des fremden Passagiers bei Ibsens *Peer Gynt* vorstellt, von denen die letzten beiden (Verkörperung Lord Byrons oder des Teufels) überzeugend sind.
5. Vgl. http://tinkerpop.apache.org/.
6. Vgl. https://www.w3.org/TR/annotation-model/.
7. Beispielsweise könnte Ironie in einem literarischen Text mithilfe eines Tagsets annotiert werden (vgl. Horstmann / Kleymann 2019): Die Tags bilden diverse Formen von Ironie ab, und pro Annotation wird per Property mithilfe von Values wie "Autorïn" oder "Erzählerïn" bestimmt, welche Instanz Subjekt" bzw. Objekt der Ironie ist.
8. Vgl. https://docs.gitlab.com/ee/user/group/ (Zugriff: 19.12.2019).

Bibliographie

Bauer, Matthias / Knape, Joachim / Koch, Peter / Winkler, Susanne (2010): "Dimensionen der Ambiguität", in: *Zeitschrift für Literaturwissenschaft und Linguistik* 40 (158): 7–75.

Bühler, Axel (2003): "Die Vielfalt des Interpretierens", in: Bühler, Axel (ed.): *Hermeneutik. Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*. Heidelberg: Synchron Wissenschaftsverlag der Autoren 99–120.

Danneberg, Lutz (1999): "Zum Autorkonstrukt und zu einem methodologischen Konzept der Autorintention", in: Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (eds.): *Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs*. Tübingen: Niemeyer 77–105.

Drucker, Johanna (2011): "Humanities Approaches to Graphical Display", in: *Digital Humanities Quarterly* 5 (1), http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html (letzter Zugriff 10. September 2019).

Folde, Christian (2015): "Grounding Interpretation", in: *British Journal of Aesthetics* 55 (3): 361–374.

Føllesdal, Dagfinn (1979): "Hermeneutics and the Hypothetico-Deductive Method", in: *Dialectica* 33: 319–336.

Genette, Gérard (2010): *Die Erzählung*. 3., durchges. und korrigierte Aufl. Paderborn: Fink.

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.

Horstmann, Jan / Kleymann, Rabea (2019): "Alte Fragen, neue Methoden – Philologische und digitale Verfahren im Dialog. Ein Beitrag zum Forschungsdiskurs um Entsagung und Ironie bei Goethe", in: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2019_007 .

Jannidis, Fotis (2003): "Polyvalenz – Konvention – Autonomie", in: Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (eds.): *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin, New York: de Gruyter 305–328.

Kindt, Tom / Müller, Hans-Harald (2003): "Wieviel Interpretation enthalten Beschreibungen? Überlegungen zu einer umstrittenen Unterscheidung am Beispiel der Narratologie", in: Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (eds.). *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin, New York: de Gruyter 286–304.

Köppe, Tilmann / Winko, Simone (2013): *Neuere Literaturtheorien*. Stuttgart: Metzler.

Meister, Jan Christoph (im Erscheinen): "Annotation als Mark-Up avant la lettre", in: Jannidis, Fotis / Winko, Simone / Rapp, Andrea / Meister, Jan Christoph / Stäcker, Thomas (eds.): *Digitale Literaturwissenschaft. DFG-Symposium Villa Vigoni, 2017*. Berlin, New York: de Gruyter.

Nünning, Vera / Nünning, Ansgar (eds. 2010): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse: Ansätze – Grundlagen – Modellanalysen*. Stuttgart, Weimar: Metzler.

Puhl, Johanna / Andorfer, Peter / Höckendorff, Mareike / Schmunk, Stefan / Stiller, Juliane / Thoden, Klaus (2015): *Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften*. DARIAH-DE Working Papers Nr. 11. Göttingen: DARIAH-DE. URN: urn:nbn:de:gbv:7-dariah-2015- 4-4 .

Röcke, Werner (2016): "Geleitwort", in: Stockhorst, Stefanie / Lepper, Marcel / Hoppe, Vinzenz (eds.): *Symphilologie. Formen der Kooperation in den Geisteswissenschaften*. Göttingen: V & R unipress.

Sahle, Patrick (2015): "Digital Humanities? Gibt's doch gar nicht!", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities*. Wolfenbüttel (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 1).

Shusterman, Richard (1978): "The Logic of Interpretation", in: *The Philosophical Quarterly* 28 (113): 310–324.

Wissler, Lars / Almashraee, Mohammed / Monett, Dagmar / Paschke, Adrian (2014): "The Gold Standard in Corpus Annotation", in: *Proceedings of the 5th IEEE Germany Student Conference*. DOI: 10.13140/2.1.4316.3523 .

Multimodaler Bedeutungstransfer vom Text zum Bild. Granulare Bildklassifikation durch Verteilungssemantik.

Donig, Simon
simon.donig@uni-passau.de
Universität Passau, Deutschland

Maria, Christoforaki
christoforakimaria@gmail.com
Universität Passau, Deutschland

Bernhard, Bermeitinger
bernhard.bermeitinger@unisg.ch
Universität Sankt Gallen, CH; Universität Passau, Deutschland

Handschuh, Siegfried
siegfried.handschuh@unisg.ch
Universität Sankt Gallen, CH; Universität Passau, Deutschland

Einleitend

In den letzten Jahren hat die Verwendung von Bildklassifizierungverfahren wie Neuronalen Netzen auch im Bereich der historischen Bildwissenschaften und der Heritage Informatics weite Verbreitung gefunden (Lang, Ommer 2018). Diese Verfahren stehen dabei vor einer Reihe von Herausforderungen, darunter dem Umgang mit den vergleichsweise kleinen Datenmengen sowie zugleich hochdimensionalen Datenräumen in den digitalen Geisteswissenschaften. Meist bilden diese Methoden die Klassifizierung auf einen vergleichsweise flachen Raum ab. Dieser „flache" Zugang verliert im Bemühen um ontologische Eindeutigkeit eine Reihe von relevanten Dimensionen, darunter taxonomische, mereologische und assoziative Beziehungen zwischen den Klassen beziehungsweise dem nicht formalisierten Kontext. Eine in (Donig, Christoforaki, Bermeitinger, Handschuh 2019) vorgeschlagene Lösung, diese Beziehungen wieder in den Prozess der Klassifizierung zurückzubringen, ist, sich die größere Ausdruckskraft von textbasierten Modellen zunutze zu machen, um die Fähigkeiten visueller Klassifikatoren zu erweitern.

Dabei wird ein Convolutional Neural Network genutzt, dessen Ausgabe im Trainingsprozess anders als herkömmlich nicht auf einer Serie flacher Textlabel beruht, sondern auf einer Serie von Vektoren. Diese Vektoren resultieren aus einem Distributional Semantic Model (DSM), welches aus einem Domäne-Textkorpus generiert wird. Ein DSM ist ein multidimensionaler Vektorraum, in dem Wörter als Vektoren abgebildet werden (Lenci 2018). Wir stellen hier eine frühe Implementierung des Verfahrens vor und analysieren deren Ergebnisse.

Wir stellen hier eine frühe Implementierung des Verfahrens vor und analysieren deren Ergebnisse.

Das durchgeführte Experiment beruht auf der Kollation von zwei Korpora, einem textbasierten und einem visuellen. Mit dem Textkorpus wird zunächst ein DSM erzeugt und diesem dann eine Auswahlliste von Zielwörter zugeführt (die funktional den Annotationslabeln der Bilder entspricht). Als Ergebnis erhalten wir Vektoren, die mit diesen Wörtern korrespondieren und mit denen die Bilder annotiert werden. Mit diesen Vektorannotationen wird ein neuronales Netzwerk trainiert, das anschließend dem Klassifikator unbekanntes Bildmaterial identifizieren soll. Als Ergebnis dieses Klassifikationsprozesses erhalten wir einen Vektor, der mit Hilfe des DSMs in natürlichsprachige Wörter zurückgewandelt wird. Da wir nach reicheren Repräsentationen im Zuge dieses Vorgangs suchen, wählen wir die fünf Vektoren aus, die dem Ausgangsvektor am ähnlichsten sind (Top-5 Nearest Neighbours). Als Ähnlichkeitsmaß legen wir die Kosinus-Ähnlichkeit zwischen vorhergesagtem Vektor und jenem Vektor zugrunde, der dem ursprünglich dem Bild von uns zugewiesenen Label (Goldlabel) entspricht. Wir gehen davon aus, dass ein Bild korrekt klassifiziert wurde, wenn das Goldlabel unter den Top-5 erscheint.

Darüber hinaus vergleichen wir die Ergebnisse des vorgeschlagenen Klassifizierungsverfahrens mit einem herkömmlichen Verfahren auf der Grundlage flacher Label unter Verwendung desselben CNNs, das für das Vektor-Experiment genutzt wurde. Wir können zeigen, dass das Vektor-Verfahren (bezo-

gen auf die Metriken) ebenso effizient und in einigen Aspekten sogar besser ist.

Aufbau des Experiments

Das Experiment beruht auf je einem Bild- und Textkorpus aus dem Bereich Sachkulturforschung mit einem Fokus auf klassizistische Artefakte.

Das Textkorpus besteht aus 44 Quellen, die unter einer freien, permissiven Lizenz verfügbar sind, und umfasst englischsprachige Fachpublikationen zu Mobiliar und Raumkunst, die von der Jahrhundertwende bis zur Mitte des 20. Jahrhunderts erschienen sind. Das Textkorpus wurde in mehreren Schritten gereinigt und aufbereitet. Zum einen wurden Standard-Natural-Language-Processing-Verfahren (NLP) angewandt, darunter Tokenisierung, Satz- und Worttrennung, die Normalisierung von Zahlenwerten und die Erkennung von benannten Entitäten (NER). Da wir retrodigitalisiertes Material aus verschiedenen Quellen genutzt haben, implementierten wir manuelle Korrekturen für die häufigsten der vorkommenden Fehler (etwa Ligaturen wie II, die als U fehlinterpretiert wurden). Eine weitere Ebene der Vorverarbeitung bestand aus inhaltsbezogenen Augmentierungen. Insbesondere normalisierten wir zusammengesetzte Wörter und Synonyme gemäß einer spezifizierten Liste, die anhand einer Ontologie, der Neoclassica-Ontologie (Donig, Christoforaki, Handschuh 2016) zusammengestellt wurde. Dies resultierte in einem Korpus von 3.067.237 Wörtern aus 107.518 Wortgrundformen.

Das DSM wurde von uns mit Hilfe des Indra Frameworks (Sales, Souza, Barzegar, Davis, Freitas, Handschuh 2018) und Gensim (Řehůřek und Sojka 2010) erzeugt.[1]

Das Bildkorpus besteht aus 1231 Ansichten klassizistischer Möbel in deren Gesamtheit, die permissiv lizenziert sind[2] und die sowohl historisches Bildmaterial als auch Fotos aus der modernen Bestandsdokumentation umfassen. Es repräsentiert 28 Klassen.

Da es sich um ein *Proof-of-Concept* Experiment handelt, kam zum Zweck des Rapid Prototyping ein an die VGG-Architektur angelehntes, „simples" neuronales Netzwerk zum Einsatz.[3] Die Unabhängigkeit der Trainings- und Testbeispiele wurde durch einen Train/Test/Eval-Split von 55:20:25 garantiert.

Da durch Sammlungspraxis der Gedächtnisinstitutionen (Sammelwürdigkeit, geographischer Schwerpunkt) und Zugänglichkeit des Materials (Lizenzierung, Grad der Sammlungsdigitalisierung) die Verteilung der Artefakte nach Klassen unbalanciert ist (Abb. 1), haben wir die Klassengewichte dementsprechend angepasst (seltene Klassen werden höher gewichtet als häufig vorkommende Klassen (Johnson, Khoshgoftaar 2019: 27). Um eine Situation zu vermeiden, in der ein Machine Learning Modell derart an ein Eingabedaten-Set angepasst wird, dass es darin scheitert, auf ähnlichen Daten zu generalisieren (Overfitting), wurde die übliche Early-Stopping-Methode verwendet.

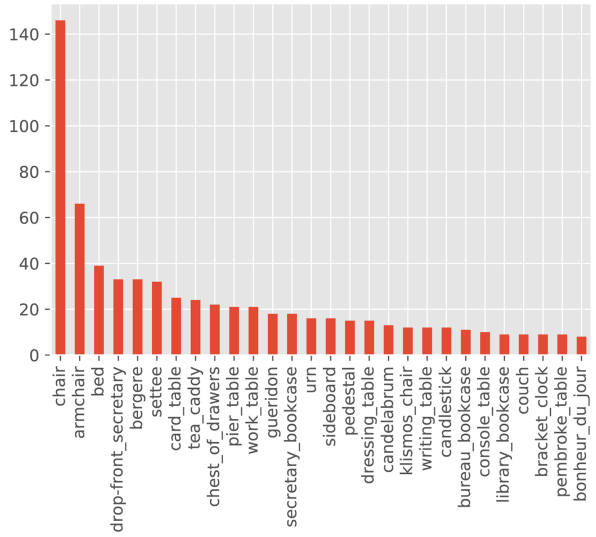


Abbildung 1: Verteilung der Klassen im Bildkorpus

Ergebnisse

Die Top-5-Richtig-Positiv Rate betrug 0.59. Das bedeutet, dass das Goldlabel in 59% der Fälle unter den fünf nächsten Nachbarn erschien.

Das mathematische Qualitätskriterium gibt für sich genommen jedoch nur einen Teil des Gesamtbilds wieder. Wir haben deshalb zugleich eine qualitative Analyse der Ergebnisse im Testset durchgeführt.

Eine Reihe von richtig-positiven Ergebnissen zeigen etwa, dass die Klassifizierung keinesfalls zufällig erfolgt, sondern dass die Top-5-Begriffe tatsächlich jeweils denselben semantischen Nachbarschaften entstammen. Sie drücken eine Reihe von Beziehungen taxonomischer und assoziativer Natur aus.

Beispielsweise wird der Roentgen-Schreibtisch aus dem Bestand des V&A in Abb. 2 mit Labeln (in der Reihenfolge) von dressing_table, writing_table und work_table assoziiert (Ankleidetisch, Schreibtisch, Nähtisch). Diese Trias ist schon deshalb sinnhaltig, weil viele dieser Artefakte multifunktional waren und mehrere dieser Funktionen erfüllten. Daneben ähneln auch jene Artefakte, die dezidiert nur einem einzigen Zweck dienten, konstruktiv den jeweils anderen Möbeltypen. Die Nähe der drei Konzepte entsteht also sowohl auf semantischer Ebene (Nähe der Wörter im DSM, die wiederum das Produkt lebensweltlicher Nähe ist), als auch auf einer visuellen Ebene im CNN (visuelle Formähnlichkeit). Ein weiteres Bild desselben Objekts (Abb. 3) zeigt einerseits, dass die Methode in sich konsistent ist (die Top4 sind identisch, obwohl eine andere Perspektive vorliegt) und andererseits, dass auch die visuellen Merkmale innerhalb des CNNs eine Auswirkung auf den Klassifizierungsprozess haben. Da Schreibschränke (*secrétaires à abbatants*) häufig frontal, hochaufrecht und mit einer geöffneten Schreibklappe oder -schublade abgebildet werden, scheint deren Vorkommen im Bild eine Klassifizierung als Sekretär getriggert zu haben. Im ersten Bild könnte dagegen die Anwesenheit von Schubladen (*drawers*) zu einer Klassifizierung als Kommode geführt haben, die naheliegenderweise auf semantischer Ebene mit Schubladen assoziiert ist.

Abbildung 2: Abweichungen bei der Klassifizierung desselben Objekts

Abbildung 3: Abweichungen bei der Klassifizierung desselben Objekts

Während die Label in den bisher betrachteten Fällen die taxonomischen Beziehungen reflektieren und alle den aus der Ontologie abgeleiteten Target Words entstammen, zeigt Abb. 4, dass das Verfahren auch aus sich selbst, rein datenzentriert Label generieren kann. Die abgebildete Kratervase wurde als Urne (urn) gold-klassifiziert. Die Top-2 Wörter reflektieren demnach auch taxonomischen Beziehungen (urn, vase). Die anderen Konzepte spiegeln dagegen assoziative Beziehungen wider. Das Label „bell“ ist ein Artefakt des Reinigungsprozesses, da im Korpus Wörter wie „bell-shaped, bell-crater“ (mit und ohne Bindestrich) existieren, um diese Art von Artefakten zu beschreiben. „Ovoid“ bezieht sich demgegenüber wohl auf die Eierstabdekoration des oberen Wulsts, die oft mit diesem Adjektiv beschrieben wird. Diese Ornamentik scheint zugleich die Assoziation zur Rosette (Patera) mitbedingt zu haben. Auf diese Weise erscheint das Target Word „patera_element“ unter den Top-5, obwohl im Bildkorpus ausschließlich ganze Artefakte, nicht aber deren Dekor annotiert wurden.

Abbildung 4: Eine Sèvres-Kopie der Medici-Vase stößt die Klassifizierung mit assoziativen Labeln an

Nicht auszuschließen ist hier zudem ein Effekt des visuellen Klassifikators, wie auch Abb. 5 zeigt. Die Fehlklassifizierung des Objekts, eines Nähtischchens, führte zu konsistenten Zuschreibungen im Bereich der Sitz- und Liegemöbel. Betrachtet man die äußere Form des Artefakts auf einer abstrakteren Ebene, kann man eine visuelle Nähe zu z.B. einem (Double-)Camel-back Sofa durchaus nachvollziehen.

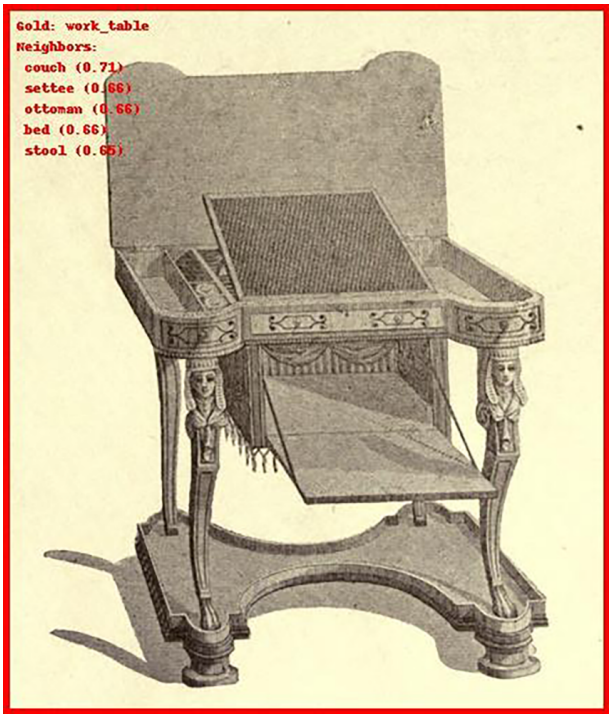


Abbildung 5: Fehlklassifizierung eines Nähtischs in ein Wortumfeld aus der Sitzmöbel-Hierarchie

Vergleichsexperiment mit einem CNN mit flachen Labeln

Um die Unterschiede zwischen beiden Zugängen besser abschätzen zu können, haben wir weiter ein Vergleichsexperiment durchgeführt, bei dem wir dasselbe CNN wie im verktorbasierten Verfahren für eine herkömmliche Klassifizierung mit flachen Labeln heranzogen.[4]

Tabelle 1: Vergleich zentraler Metriken beider Zugänge

	Vektor-Label	Flache Label
Top-1-Treffergenauigkeit	0.50	0.40
Top-1-Genauigkeit	0.32	0.29
Top-1-Trefferquote	0.25	0.26
Top-1-F1-Wert	0.26	0.25
Top5-Falsch-Positiv-Rate	0.41	0.27
Top5-Richtig-Positiv-Rate	0.59	0.73

Wie ersichtlich, ist nicht nur die Top-1-Treffergenauigkeit im Fall der Klassifizierung mit Vektoren besser, sondern auch die übrigen Metriken vergleichbar gut sind. Durch den hier vorgeschlagenen Zugang wird also nicht nur die Treffergenauigkeit verbessert, sonder er liefert zugleich eine reichhaltigere Beschreibung des Bildes.

Schlussfolgerungen und Ausblick

In dem hier vorgeschlagenen Beitrag haben wir ein neues, multimodales Verfahren für die Klassifizierung von Bildinhalten vorgestellt, das auf der Kombination von NLP-Methoden mit Bildklassifizierungsverfahren beruht. Ziel war, Objekte nicht alleine nach einem Schema flacher Label, sondern in einer kontextgerechteren Weise zu klassifizieren, wobei dieser Kontext von einschlägigen historischen Domänepublikationen gebildet wird. Dieses Klassifizierungsverfahren bietet einen Zugang zur multidimensionalen Einbettung der Artefakte in die Lebenswelt und deren sprachlicher Widerspiegelung. Dieser Umstand ist von besonderem Nutzen, um multifunktionale Objekte zu klassifizieren, ohne dabei auf mehrere Klassifikatoren und einen komplexen Annotationsprozess mit mehreren Labeln zurückgreifen zu müssen. Die Ergebnisse sind ermutigend. Auch mit einem sehr einfachen CNN erreichten wir eine Genauigkeit von 0,59. Als nächsten Schritt möchten wir mit einem komplexeren CNN und einem ausgeweiteten Bildkorpus trainieren (um bekannte Probleme wie overfitting zu reduzieren). Unser Vergleichsexperiment mit einem herkömmlichen, auf flachen Labeln beruhenden Zugang hat gezeigt, dass unter Effizienzgesichtspunkten, d. h. im direkten Vergleich der Metriken, unser Verfahren nicht nur vergleichbare Resultate liefert, sondern zugleich auch in einer reichhaltigeren Beschreibung des Bildes resultiert.

Wir werden weiter daran arbeiten, besser zu verstehen, wie ein bestimmtes Textkorpus sich in den Labeln widerspiegelt, die das DSM automatisch zuweist und die nicht Teil der Liste der Target Words sind. Ein besseres Verständnis dieser Prozesse scheint insbesondere im Hinblick auf die relativ überschaubaren Textkorpora relevant, die in den Geisteswissenschaften zu spezifischen Themenkomplexen kollationiert werden können. Nicht zuletzt werden wir aus diesem Grund die Nutzung von Thesauri und Wörterbüchern in Betracht ziehen, um Synonymlisten für Target Words zu erstellen. In ähnlicher Weise ziehen wir in Betracht, benannte Entitäten zu URIs zusammenzufassen. Das würde uns erlauben, spezifische Entitäten (z. B. Werkstätten, Ebenisten, Eigentümer) mit bestimmten Objekten zu assoziieren.

Wir denken, dass dabei der multimodale Zugriff einen besonders effizienten Zugang zu geistes- und kulturwissenschaftlichen Korpora bietet, die, verglichen mit den Korpora anderer Disziplinen in den Natur- und Sozialwissenschaften, klein und Domäne-restringiert sind.

Fußnoten

1. Das DSM wurde mit einer Vektorgröße von 50, einer Wortfenstergröße von 10 und einer minimalen Wortzahl von fünf erstellt. Als Word2Vec-Modell kam Skipgram (Mikolov, Chen, Corrado, Dean 2013) mit Negative Sampling zum Einsatz.
2. Das Korpus wurde aus den Sammlungen des Metropolitan Museum, New York, des Victoria & Albert Museum, London, der Wallace Collection, London sowie mehreren zeitgenössischen Musterbüchern zusammengestellt.
3. Das Netzwerk besteht aus drei Convolutional-Blöcken mit jeweils zwei Convolutional-Layers mit 32/64/64 Filter der Größe 3x3. Nach jedem Block folgt ein Maximum-Pooling-Layer der Größe 2x2 sowie ein Dropout-Layer mit einer Dropoutwahrscheinlichkeit von 0.25. Ein Fully-Connected-Block, bestehend aus zwei Fully-Connected-Layers mit jeweils 256 Knoten, steht im Anschluss sowie nochmals ein Dropout Layer mit 0.5 Dropoutwahrscheinlichkeit. Jeder Convolutional- und Fully-Connected-Layer bis dahin wurde zufällig initialisiert und benutzt ReLU als Aktivierungsfunktion. Der letzte Layer ist ein Fully-Connected-Layer mit 50 Ausgabeknoten und benutzt eine lineare Aktivierungsfunk-

tion. Es ist in den beiden von uns genutzten Frameworks Keras (https://keras.io) und TensorFlow (https://tensorflow.org) implementiert. Beim Training wird der durchschnittliche absolute Fehler durch die Optimierungsfunktion RMSprop minimiert.
4. Das verwendete CNN unterscheidet sich von dem im Vektor-Experiment verwendeten Netz lediglich durch den letzte Layer, der ein Fully-Connected-Layer mit 28 Ausgabeknoten ist, korrespondierend mit den 28 flachen Labeln, und die Nutzung von softmax als Aktivierungsfunktion.

Bibliographie

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 [Letzter Zugriff 25. 09. 2019].

Donig, Simon / Christoforaki, Maria / Bermeitinger, Bernhard / Handschuh, Siegfried (2019): „Vom Bild zum Text und wieder zurück", in: Sahle, Patrick (ed.): DHd 2019 - Digital Humanities: multimedial & multimodal – Konferenzabstracts. Mainz & Frankfurt a. M.: 227–232, https://zenodo.org/record/2596095/files/2019_DHd_BookOfAbstracts_web.pdf [Letzter Zugriff 25. 09. 2019].

Johnson, Justin M. / Khoshgoftaar, Taghi M. (2019): „Survey of deep learning with class imbalance", in: Journal of Big Data 6 (27): 2-54, https://doi.org/10.1186/s40537-019-0192-5 [Letzter Zugriff 25. 09. 2019].

Krizhevsky, Alex / Sutskever, Ilya / Hinton Geoffrey E. (2012): „ImageNet Classification with Deep Convolutional Neural Networks". In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, 1097–1105. NIPS'12. USA: Curran Associates Inc. http://dl.acm.org/citation.cfm?id=2999134.2999257 [Letzter Zugriff 25. 09. 2019].

Lang, Sabine / Ommer, Björn (2018): „Attesting Similarity: Supporting the Organization and Study of Art Image Collections with Computer Vision". In: Digital Scholarship in the Humanities 33 (4): 845–56. https://doi.org/10.1093/llc/fqy006 [Letzter Zugriff 25. 09. 2019].

Lenci, Alessandro (2018): "Distributional models of word meaning", in: Annual review of Linguistics, 4 (1) : 151-171.

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. 2019]

Řehůřek, Radim / Sojka, Petr (2010): „Software Framework for Topic Modelling with Large Corpora". In: LREC 2010. Valletta, Malta, 2010: 46–50.

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.

Abb. 2 / Abb. 3: Victoria & Albert Museum, London: Writing table, Neuwied, workshop of David Roentgen ca. 1774-1780. Ascension number: 1059:1 to 9-1882. http://collections.vam.ac.uk/item/O117298/writing-table-roentgen-david/ [Letzter Zugriff 25. 09. 2019].

Abb. 4: Victoria & Albert Museum, London: Vase [Sèvres Copy of the Medici Vase], Paris, 1813. Ascension Number 396-1874. http://collections.vam.ac.uk/item/O8978/vase-sevres-porcelain-factory/ [Letzter Zugriff 25. 09. 2019].

Abb. 5: **Sheraton, Thomas / Bell, J. Munro (arr.)** (1910): The furniture designs of Thomas Sheraton. London: Gibbings and Co., Ltd.

m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts

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

Flüh, Marie
marie.flueh@uni-hamburg.de
University of Hamburg, Deutschland

Während in den Digital Humanities bereits erste korpusbasierte Analysen von Figurengender in der Literatur vorgelegt wurden (Underwood 2019: 111 ff), wird in den Kulturwissenschaften zu diesem Thema selten korpusbasiert gearbeitet. Stattdessen sind Theorien zur Genderthematik häufig philosophisch-soziologisch (z. B. bei Beauvoir oder Bourdieu), diskurstheoretisch (z. B. Foucault) oder dekonstruktivistisch (z. B. Butler)[1] motiviert. Die Lücke zwischen einer an technischen Methoden ausgerichteten Modellierung und der theoretischen Betrachtung der Genderthematik schließen wir mit dem Projekt m*w, von dem wir erste Pilotstudien hier vorstellen. In einem theoriegeleiteten Mixed-Methods-Ansatz operationalisieren wir zunächst klassische Ansätze aus den Genderstudies und wenden das entstandene Modell sowohl mit Hilfe von Named Entity Recognition als auch mittels digitaler Annotation auf ein literarisches Korpus an. Das Zusammenspiel von theoretischer und datenbasierter Modellierung und die Bereitstellung von Zwischenergebnissen und "losen Enden" auf der *Projektwebseite* (https://msternchenw.de) eröffnen Spielräume, um weiter- und umzudenken und sich mit anderen Projekten zusammenzuschließen.

Konzeptioneller Rahmen und methodische Desiderata

Die leitende Fragestellung des m*w-Projektes ist: Wie werden Genderrollen in der Literatur des 19. Jahrhunderts dargestellt und bewertet? Um uns dieser Fragestellung zu nähern, erstellen wir mithilfe einer Auswahl theoretischer Ansätze ein Modell. Dieses nutzen wir, um im überwachten Machine-Learning-Verfahren der Named Entity Recognition (NER) ein Tool darauf zu trainieren, Figuren und ihre Genderzuschreibungen automatisch zu erkennen. Die Ergebnisse des NER-Verfahrens nutzen wir um unser Modell weiter zu schärfen. Im abschließenden Close Reading der Novellen werden schließlich Genderbeschreibungen und -bewertungen analysiert und mit dem Modell abgeglichen. Grundsätzlich nehmen wir sowohl Modell als auch Korpus als variable bzw. dynamische Größen wahr, die sich für den Praxistest der Operationalisierung theoriebasierter Modelle zur Erforschung literarischer Genderrollen eignen.

Korpus

Forschungsgegenstand ist der *Deutsche Novellenschatz*, der 1871–1876 von Paul Heyse und Hermann Kurz herausgegeben wurde und 87 Novellen umfasst.[2] Da in dieser Textsammlung die Thematik der Ehe von großer Bedeutung ist (vgl. Weitin/Herget 2017), gehen wir davon aus, dass stereotype Genderrollen sich potentiell häufen. Zur tatsächlichen, möglicherweise diversen Genderausrichtung der Autor*innen ist nichts bekannt. Um einerseits einen möglichst ausgewogenen Forschungsgegenstand zu bekommen, andererseits aber möglichst wenig auf das Korpus einzuwirken, wurden zunächst alle 12 aus der Feder von Schriftstellerinnen* stammenden Novellen in ein Teilkorpus übernommen. Ergänzend dazu wurden 12 per Zufallsgenerator ausgewählte Novellen von Autoren* ergänzt. Die verbleibenden Novellen wurden als Trainingsmaterial genutzt. Bei der Zusammenstellung des Korpus war uns bewusst, dass wahrscheinlich überwiegend Autorenpersönlichkeiten einbezogen wurden, die sich einem binären Geschlechtermodell zugehörig fühlten. Um Trainingsmaterial für Anschlussforschung über die Pilotstudien hinaus zur Verfügung zu stellen, planen wir ein weiteres Korpus zu erstellen, das Texte enthält, die von Personen geschrieben wurden, die sich nachweislich nicht in ein binäres System integrieren lassen (wollten).[3]

Genderstereotype

Davon ausgehend, dass sich sowohl in der Theorie als auch in Erzähltexten Spielräume als Zwischenräume auftun, die dadurch sichtbar werden, dass sie sich von den sie umgebenden normierten Räumen unterscheiden[4], untersuchen wir stereotype Darstellungen und solche, die sich davon abheben in einem relationalen Ansatz. Anschließend an Butler (vgl. Butler 1990: 190-219) differenzieren wir den Begriff Gender nach Geschlecht, Gender Identität und Gender Performanz und übersetzen diese Trias für die Anwendung auf das literarische Korpus in Geschlechtszuordnung durch Personalpronomen, Ausdruck der Gender-Identität durch Figureneigenschaften und Beschreibung der Gender-Performanz durch Aufzeigen (nicht) rollenkonformer Figurenhandlungen. Die Betrachtung der Geschlechtszuordnung durch Personalpronomen schließen wir aus, da der dieser binären Zuordnung zu Grunde liegende Biologismus fraglich ist, wie z.B. die Ausführungen Foucaults (vgl. Foucault 1998: 7-18) zu Herculine Barbin[5] sowie dessen eigene Lebenserinnerungen (vgl. ebd. 19-126) zeigen.

Um entscheiden zu können, ob Eigenschaften und Handlungen von Figuren stereotypen Rollenbildern zugeschrieben werden können oder nicht, haben wir zunächst möglichst viele Rollenbilder in unser Modell integriert. Jede dieser Rollen kann sowohl im Sein (Gender-Identität) als auch im Handeln (Gender-Performanz) von Figuren verankert sein. Die sechs in Abb. 1 abgebildeten Oberkategorien von Eigenschaften sind ebenfalls der Theorie entnommen. Allerdings wurden hier die Beschreibungen stärker kondensiert, um die zahlreichen genannten Einzeleigenschaften für die digitale Annotation besser handhabbar zu machen.

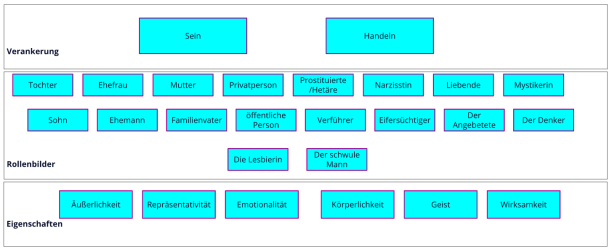


Abbildung 1: Theoriebasiertes Gender-Modell

Erste Ergebnisse der Analysen

Alle drei eingesetzten Methoden - NER, Annotation von Stereotypen und Emotionsanalyse haben sich im ersten Proof of Concept als fruchtbar für die digitale Erforschung von Figurengender erwiesen. Besonders eklatante Zwischenergebnisse fassen wir im Folgenden zusammen.

NER

Im Laufe des NER-Trainingsprozesses erwies sich die Kategorie "divers" nicht als dem Korpus angemessen, weshalb wir die NER-Kategorien auf "männlich", "weiblich", "genderneutral"[6] festlegten. Mit einer schrittweise Ausweitung des Trainingsmaterials von 40.000 auf 68.000 Tokens[7] erreichten wir eine Steigerung der anfänglich bei rund 0,47 liegenden F1-Score-Werte um 0,03 in einem ersten und 0,07 in einem zweiten Testtext (*Irrwisch-Fritze* von Reinbold und *Die drei Schwestern* von Kähler). In beiden Testtexten zeigte sich, dass genderneutrale Figuren am zuverlässigsten und weibliche Figuren am schlechtesten erkannt werden (für detailliertere Beschreibungen der ersten NER-Tests vgl. Schumacher 2020 [2]).

Um Hinweise auf eine mögliche Verzerrung durch die Zusammensetzung des Trainingskorpus zu bekommen, haben wir einen dritten Testtext hinzu genommen (*Eine fromme Lüge* von Gall). Die Ergebnisse des Tests weichen von den vorherigen ab. Der F1-Score liegt bei 0,6202 und ist damit um 0,08 höher als der höhere der beiden vorherigen Testtexte. Außer-

dem werden weibliche Figuren besser erkannt als männliche (F-Scores von 0,7093 und 0,4293). Ein unerwarteter Nebenbefund aus dem wir nach intensiver Prüfung des Ergebnisses (mehr darüber in Schumacher 2020 [1]) schließen, dass das Tool hier einen Text mit sehr stereotypen Rollenbeschreibungen erkannt hat. Dass das im Hinblick auf Autorengender zu homogene Trainingskorpus verzerrend wirkt, konnte durch diesen Test hingegen nicht endgültig bestätigt werden. Ein inhaltlicher Vergleich der NER-Ergebnisse zeigt aber, dass die Anzahl männlicher Figurenbenennungen insgesamt höher ist (Abb. 3).

Abbildung 3: Die NER-Ergebnisse der Testtexte visualisiert; links steht Hellblau, mittig Dunkelblau und rechts rosa für männliche Figuren.

Digitale Annotation von Genderstereotypen

Im Annotationstool CATMA (Meister et all. 2019) wurden die NER-Ergebnisse zuerst ergänzt, sodass alle tatsächlich vorkommenden weiblichen, männlichen und neutralen Figurenbezeichnungen im Testtext *Die drei Schwestern* annotiert wurden. Der Eindruck, dass es hier am meisten männliche Figuren gibt, wird bestätigt (Abb. 4).

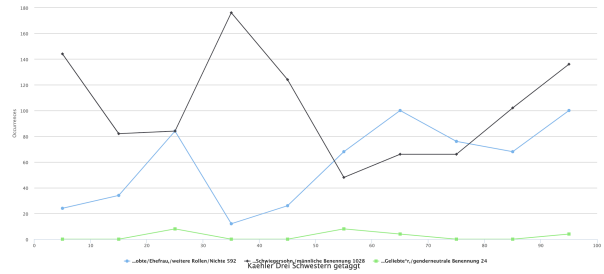


Abbildung 4: Gender-Kategorien im Verlauf des Beispiel-Textes (männlich (schwarz) , weiblich (blau) und genderneutral (grün)).

Anschließend wurden die NER-Kategorien mit Unterkategorien versehen, die den Benennungen der Genderrollen des Modells entsprechen (Abb. 5).

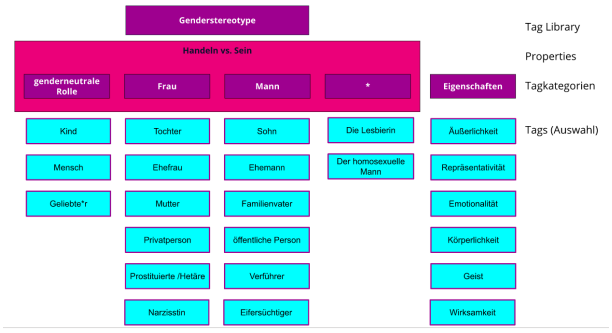


Abbildung 5: Vorläufiges Tagset (Auswahl der angewandten weiblichen und männlichen Rollen/ nur Oberkategorien der Eigenschaften).

Sofern es keine adäquate, in der Theorie erwähnte Rolle oder Eigenschaft gab, wurden zusätzliche Annotationskategorien erstellt und den Oberkategorien “unsortierte Rollen” und “unsortierte Eigenschaften” zugewiesen.

Die digitale Annotation des ersten Beispieltextes zeigt, dass stereotype Beschreibungen vor allem zu Beginn der Erzählung häufig und somit besonders für die Etablierung der Figuren von Bedeutung sind. Stereotype Eigenschaften sind hier zwar divers, für männliche und weibliche Figuren gibt es aber jeweils einige wenige, die quantitativ herausstechen. Für weibliche Figuren ist das vor allem Äußerlichkeit/Schönheit für männliche sind es Körperlichkeit/Trinkfestigkeit und Wirksamkeit/Herrschaft. Unsortierte Rollen und Eigenschaften sind zumeist in einem binären Rollensystem verankert und dennoch werden in dieser Novelle zum Teil Stereotype aufgebrochen. Dies wird hauptsächlich durch die Zuschreibung einzelner Eigenschaften zu einer Figur eines Genders erreicht, die in der theoretischen Literatur eher dem Stereotyp des anderen zugeschrieben werden (ausführlichere Auswertung des ersten Beispieltextes in Schumacher 2020 [3]).

Emotionsanalyse mit CATMA

Die Auswertung der digitalen Annotation des Beispieltextes *Gemüth und Selbstsucht* von Margarethe von Wolff macht folgende Einsichten besonders deutlich: Die Emotionen Zorn, Ekel, Trauer und Liebe – die sich wiederum aus unterschiedlichen Vertretern dieser Emotionsfamilien zusammensetzten – treten am häufigsten auf.

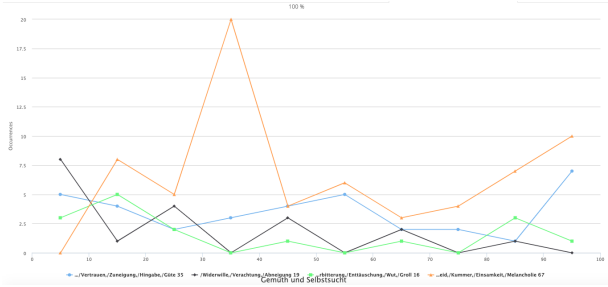


Abbildung 6: Der Verteilungsgraph zeigt den Emotionsverlauf in Gemüth und Selbstsucht: TRAUER (orange), LIEBE (blau), EKEL (schwarz), ZORN (grün)

Die Emotionen werden in den allermeisten Fällen verbal (407 Annotationen) von den Figuren ausgedrückt. Über die Veränderung des körperlichen Zustands (19 Annotationen), und nonverbal (107 Annotationen) werden Emotionen vergleichsweise selten repräsentiert. Die Auswertung der Properties ergibt genderspezifische Emotionsinformationen (s. Abb. 7 und 8). Weibliche Figuren treten ängstlicher auf als männliche. Männliche Figuren reagieren häufiger zornig als weibliche. Sie empfinden außerdem häufiger Ekel – hier meistens im Sinne von Abneigung – als weibliche Figuren. Diese leiden häufiger unter gedrückter Stimmung und empfinden deutlich häufiger negative Basisemotionen als männliche Figuren. Diese treten im Schnitt fröhlicher auf als die weiblichen Figuren. Die männlichen Figuren zeigen häufiger positive Basisemotionen als die weiblichen Figuren und auch die positive Basisemotion LIEBE überwiegt seitens der männlichen Figuren. Scham – als Unterkategorie der Problemfälle – ist seitens der weiblichen Figuren deutlich stärker ausgeprägt (für eine

ausführlichere Auswertung des ersten Beispieltextes vgl. Flüh 2020).

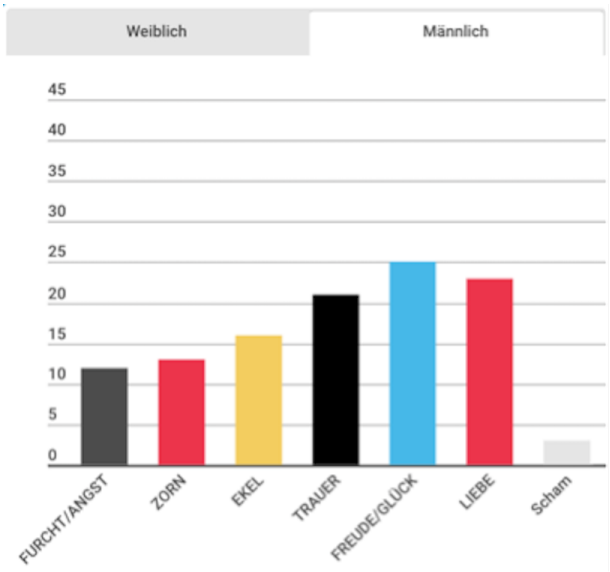


Abbildung 7: Gesamtvorkommen der mit der Value männlich ausgezeichneten Emotionen

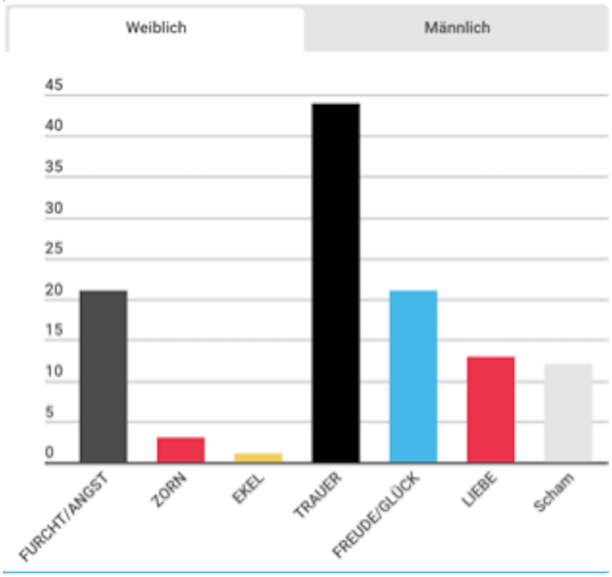


Abbildung 8: Gesamtheit der mit der Value weiblich ausgezeichneten Emotionen

Theorien hinterfragen, aufbrechen und weiterdenken

Im Hinblick auf die leitende Fragestellung können zu diesem Zeitpunkt drei Zwischenergebnisse festgehalten werden. Der Trainingsprozess des NER-Modells zeigt, dass im deutschen Novellenschatz hauptsächlich weibliche, männliche und genderneutrale Figurenbeschreibungen eine Rolle spielen. Um diverse Genderrollen in der Literatur des 19. Jahrhunderts erforschen zu können, muss weiteres Trainingsmaterial erstellt werden. Die Tests des NER-Modells machen deutlich, dass die Methode der CRF-Vorhersage geeignet ist, um innerhalb des Novellenschatzes Texte mit besonders stereotypen Genderzuschreibungen ausfindig zu machen.

Erste Auswertungen der digitalen Annotation von Genderstereotypen zeigen, dass in Beschreibungen von Figuren zwar eine Vielzahl von Eigenschaften genutzt wird, dass genderspezifisch aber jeweils einige wenige Kategorien vorherrschen. Für unseren Beispieltext konnten wir belegen, dass stereotype Genderrollen nicht durch die Einführung diverser Figuren aufgebrochen werden, sondern durch die Kennzeichnung einzelner Figuren mit einzelnen stereotypen Eigenschaften eines anderen Genders. Die digitale Annotation zeigt auch, dass genderneutrale Figurenbezeichnungen eher selten sind. Erste Auswertungen der Emotionsanalysen zeigen, dass genderspezifische Unterschiede v. a. hinsichtlich der Wertung der referierten Emotionen bestehen; männliche Figuren greifen auf ein positiveres Gefühlskonzept bzw. Bewertungskonzept zurück als weibliche. Die strukturorientiert fundierten Tagsets eignen sich, um die vielfältigen Ausdrucksweisen genderspezifischer Emotionsmanifestationen und -familien aus literarischen Texten herauszufiltern. Im weiteren Verlauf gilt es, die übrigen Texte auszuwerten und die Analyseergebnisse miteinander in Bezug zu setzen.

Fußnoten

1. Die hier ausgewerteten Ansätze der Genderstudies beziehen sich nicht primär auf literarische Korpora, die in diesen Theorien ausgewerteten Beispiele sind aber häufig literarischer Art. Die daran erkennbare Nähe zu den Literaturwissenschaften erleichtert die im m*w vorgenommene Domänenadaption der Ansätze.
2. Wir nutzen die digitale Version des deutschen Novellenschatzes, die vom Discourse Lab herausgebracht wurde und unter https://www.discourselab.de/cqpweb/ heruntergeladen werden kann. Die von uns zum Zweck des Machine Learning Trainings erstellen Daten (sowohl Trainingsdaten als auch Modell) werden auf der Projektwebseite unter http://msternchenw.de/m*w-insiders/ zur Nachnutzung bereitgestellt.
3. Dieses “Diversitäts-Korpus” wird ebenfalls auf der Projektseite unter http://msternchenw.de/m*w-insiders/ zur Verfügung gestellt werden.
4. Deutlich wird dies in der Genderforschung z.B. bei Beauvoir, die zunächst eine Reihe traditioneller Rollenmuster der Frau beschreibt, um sich dann der lesbischen Frau zuzuwenden, die hier ausbricht (vgl. Beauvoir 1949) oder bei Bourdieu, der zunächst ein binär organisiertes Herrschaftssystem beschreibt, um sich dann abschließend der Liebe als Möglichkeit, dieses zu überwinden und der Homosexuellen-Bewegung, die in seinen Augen die männliche Herrschaft eher stützt, zuzuwenden (vgl. Bourdieu 2005). Interessant ist in diesem Zusammenhang auch der gar nicht auf Geschlechterfragen ausgerichtete Ansatz Foucaults zur Heterotopologie, in dem er sich der Identifizierung von Heterotopien als Räume des anderen widmet, die Ausdifferenzierung des “normalen” Raumes aber unterlässt, wodurch eine Leerstelle entsteht (vgl. Foucault 2005). Letzteres soll hier vermieden werden.

5. Die Lebenserinnnerungen Herculine Barbins zeigen einen Fall von Hermaphroditismus im 19. Jahrhundert in Frankreich.
6. Auch diese Kategorien sind für uns nur als vorläufig gedacht und können jederzeit an unterschiedliche Korpora angepasst bzw. mit Hilfe von anderen, zusätzlichen Korpora weiterentwickelt werden.
7. Das Trainigskorpus wird auch weiterhin laufend erweitert.

Bibliographie

Argyle, Michael (1996): *Körpersprache und Kommunikation.* Paderborn: Junfermann.

Barbin, Herculine / Foucault, Michel (1998): *Über Hermaphroditismus* . Frankfurt am Main: Suhrkamp.

Beauvoir, Simone (1949): *Das andere Geschlecht.* Reinbek: Rowohlt.

Bourdieu, Pierre (2005): *Die männliche Herrschaft.* Frankfurt am Main: Suhrkamp.

Borod, Joan C. (2000): *The Neuropsychology of Emotion* . New York: Oxford University Press.

Butler, Judith (1990): *Das Unbehagen der Geschlechter.* Frankfurt am Main: Suhrkamp.

Connell, Raewyn (1999): *Der gemachte Mann. Konstruktion und Krise von Männlichkeiten.* Wiesbaden: Springer.

Darwin, Charles (1884): *Der Ausdruck der Gemüthsbewegungen bei dem Menschen und den Thieren.* Stuttgart: Schweizerbart.

Ekman, Paul (1999): *Basic Emotions.* In: Dalgleisch, Tim und Michael J. Power (Hrsg.): *Handbook of Cognition and Emotion* . Chichester: John Wiley & Sons, 45–60.

Erhart, Walter (2001): *Familienmänner. Über den literarischen Ursprung moderner Männlichkeit.* München: Wilhelm Fink.

Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher (2005): Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* , Seiten 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf [zuletzt geprüft: Dezember 21, 2019]

Flüh, Marie (2020): "Emotionsinformationen analysieren: 3. Erste Ergebnisse – Gemüth und Selbstsucht (1787) von Margarethe von Wolff." In: *m*w* , Januar 4, 2020, https://msternchenw.de/?p=249 , [zuletzt geprüft: Januar 4, 2020].

Foucault, Michel (2005): *Die Heterotopien. Der utopische Körper.* Frankfurt am Main: Suhrkamp.

Foucault, Michel (2008): *Die Ordnung der Dinge.* In: Michel Foucault. Die Hauptwerke. Frankfurt am Main: Suhrkamp.

Heydebrand, Renate von / Simone Winko (1996): *Einführung in die Wertung von Literatur* . Paderborn/München/Wien u. a.: Ferdinand Schöningh.

Irigaray, Luce (1980): *Speculum. Spiegel des anderen Geschlechts.* Frankfurt am Main: Suhrkamp.

Izard, Carroll, E. (1992): "Basic emotion, relations among emotions, and emotion-cognition relations". In: *Psychological Review* 99 (3), 561–565.

Luhmann, Niklas (2008): *Schriften zur Kultur und Literatur.* Frankfurt am Main: Suhrkamp.

Meister, Jan Christoph / Horstmann, Jan / Petris, Marco / Jacke, Janina / Bruck, Christian / Schumacher, Mareike / Flüh, Marie (2019): CATMA 6.0.0 (Version 6.0.0). Zenodo. DOI: .

Moretti, Franco (2013): *Operationalizing: or the function of measurement in modern literary theory.* https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf [zuletzt geprüft: August 8, 2019].

Oatley, Keith / P.N. Johnson-Laird (1987): "Towards a cognitive theory of emotions". In: *Cognition and Emotion* 1, 1987, 29–50.

Plutchik, Robert (1984): "Emotions. A general psycho-evolutionary theory". In: Klaus Scherer und Paul Ekman (Hrsg.): *Approaches to emotion.* Hillsdale: Erlbaum, 197–200.

Schumacher, Mareike (2020 [1)]: "Named Entity Recognition und Reverse Engineering". In: *Lebe lieber literarisch*, Januar 2, 2020, http://lebelieberliterarisch.de/named-entity-recognition-und-reverse-engineering/, [zuletzt geprüft: Januar 3, 2020].

Schumacher, Mareike (2020 [2)]: "Automatische Erkennung von Figuren-Gender – das erste Modell." In: *m*w* , Januar 3, 2020, https://msternchenw.de/automatische-erkennung-von-figuren-gender-das-erste-modell, [zuletzt geprüft: Januar 3, 2020].

Schumacher, Mareike (2020 [3)]: "Genderstereotype in der Literatur - erste Analysen". In: *Lebe lieber literarisch* , Januar 7, 2020, http://lebelieberliterarisch.de/genderstereotype-in-der-literatur-erste-analysen/, [zuletzt geprüft: Januar 7, 2020].

Schwarz-Friesel, Monika (2017): "Das Emotionspotenzial literarischer Texte." In: Betten, Anne, Ulla Fix und Berbeli Wanning (Hrsg): *Handbuch Sprache in der Literatur* . Berlin, Boston: De Gruyter, 351-370.

Simmel, Georg (1986): "Das Relative und das Absolute im Geschlechter-Problem." In: *Philosophische Kultur. Über das Abenteuer, die Geschlechter und die Krise der Moderne.* Berlin: Klaus Wagenbach, S. 64–93.

Stephan, Inge (2003): "Im toten Winkel. Die Neuentdeckung des ‚ersten Geschlechts' durch men's studies und Männlichkeitsforschung." In: dies., Claudia Benthien (Hrsg.): Männlichkeit als Maskerade. Kulturelle Inszenierung vom Mittelalter bis zur Gegenwart. In: Stephan, Inge, Sigrid Weigel: *Literatur – Kultur – Geschlecht. Studien zur Literatur- und Kulturgeschichte.* Kleine Reihe, Band 18. Köln/Weimar/Wien: Böhlau Verlag, 11–35.

Underwood, Ted (2019): *Distant Horizons.* Chicago: The University of Chicago Press.

Weitin, Thomas / Katharina Herget (2017): "Falkentopics. Über einige Probleme beim Topic Modeling literarischer Texte." In: Hartmut Bleumer, Rita Franceschini, Staphan Habscheid, Constanze Spieß und Niels Weber (Hrsg.): *Scalable Reading* . Zeitschrift für Literaturwissenschaft und Linguistik. 1/2017 (47), 1–20.

Netzwerkanalyse spielerisch vermitteln mit DraCor und forTEXT: Zur nicht-digitalen Dissemination einer digitalen Methode in Form des Kartenspiels „Dramenquartett“

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

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Deutschland

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

Fischer, Frank
frank.fischer@dariah.eu
Higher School of Economics Moskau, Russland

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Postdam, Deutschland

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

Die Komponenten: DraCor und forTEXT

DraCor

Mit ELTeC (European Literary Text Collection; https://github.com/COST-ELTeC) und DraCor gibt es mittlerweile zwei europäische Initiativen, die eine korpusbasierte Infrastruktur für die digitalen Literaturwissenschaften aufbauen, wobei sich DraCor (Drama Corpora Project; https://dracor.org) der Sammlung TEI-kodierter Dramen in verschiedenen Sprachen widmet (vgl. Fischer u.a. 2019). DraCor liefert über seine API etwa Netzwerkdaten zu Dramen aus, die auf der Kookkurrenz von Sprecherïnnen basieren und es ermöglichen, die Kommunikationsstrukturen mithilfe von Network-Analysis-Metriken zu erforschen. Darüber hinaus bietet die Plattform mit ezlinavis (Easy Literary Network Analysis Visualisation) ein didaktisches Tool an, das den Einstieg in die systematische Erhebung von Netzwerkdaten erleichtert. Außerdem wurde aus DraCor heraus mit dem *Dramenquartett* (vgl. Fig. 1) ein Kartenspiel entwickelt, mit dem das Verständnis von Netzwerkmetriken ebenso wie die typologische und historische Vielfalt von Dramennetzwerken spielerisch entdeckt und erlernt werden kann (vgl. Fischer u.a. 2018 und Fischer/Schultz 2019).

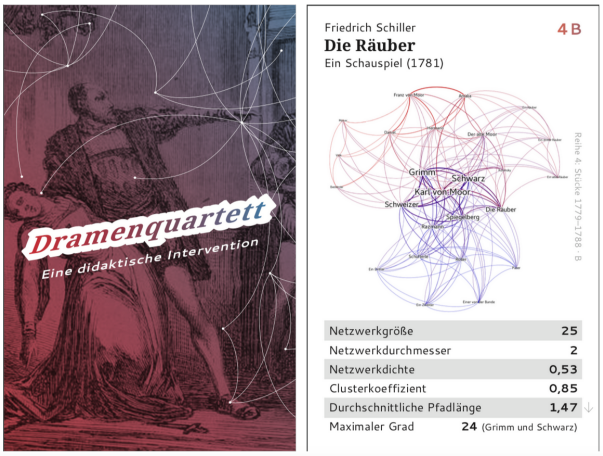


Figure 1: Beispielkarte aus dem Dramenquartett (Rück- und Vorderseite)

forTEXT

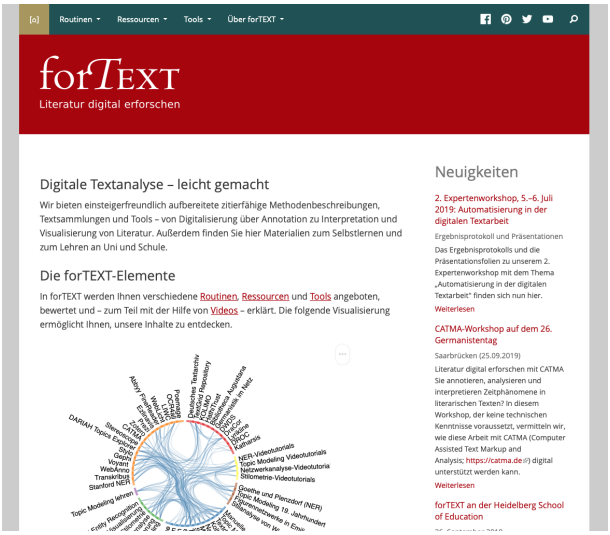


Figure 2: Die Startseite des Disseminationsportals forTEXT.net

Das DFG-Projekt *forTEXT* (https://fortext.net; vgl. Fig. 2) bietet in diesem Zusammenhang einen Methodeneintrag (vgl. Schumacher 2018b), eine (Gephi-)Lerneinheit (vgl. Schumacher 2019b), ein Fallstudien-Video[1], vier Tutorialvideos[2] sowie praktische Anwendungen in der Lehre. Hierbei sind jeweils unterschiedliche Abstraktionsgrade und verschiedene

Arten der Vermittlung abgedeckt: Der Methodeneintrag ist eine abstrakte, sprachlich-theoretische Beschreibung der Methode mit dem Schwerpunkt der Anschlussfähigkeit an die traditionelle Literaturwissenschaft. Die Lerneinheit ist eine konkrete Klick-für-Klick-Einführung für Autodidakten aus der Zielgruppe junger Geisteswissenschaftlerïnnen in Form einer Text-Bild-Kombination.

Die Videos vermitteln die Methode über eine Text-Bild-Audio-Kombination: Das Methodenvideo bietet eine Fallstudie zum Figurennetzwerk von *Emilia Galotti*. Es vermittelt die Methode an eine autodidaktische Zielgruppe geisteswissenschaftlicher Studentïnnen und wählt einen Einstieg über das ‚traditionelle' Thema, taucht in technisch-theoretische Hintergründe ein und schließt dann wieder an das literarische Thema an. Die textbasierte Thematik wird so auf eine Bildebene überführt, die DH-Tool-Grafiken mit strichmännchenartigen figurativen Darstellungen koppelt (vgl. Fig. 3).

Figure 3: Vorschaubild eines forTEXT-Fallstudien-Videos

Angesprochen werden hier theoretische, strukturelle und emotionale autodidaktische Vermittlungsmuster (zur Bedeutung von Emotionen für autodidaktisches Lernen vgl. Mega u.a. 2014). Auf der Tonebene ist ein erklärender Duktus vorherrschend. Die ‚selfmade'-Anmutung der Videos vermittelt, dass die autodidaktische Erarbeitung der Inhalte Betrachterïnnen und Erstellerïnnen des Videos miteinander verbindet (vgl. Horstmann & Schumacher 2019).

Die Tutorial-Reihe schließlich funktioniert ähnlich wie die Lerneinheit als Schritt-für-Schritt-Anleitung und bietet die Möglichkeit, die Arbeit mit Gephi als Screencast zu erlernen. Das Tutorial-Video zur Nutzung des DraCor-Tools ezlinavis verknüpft die praktische Erstellung von Netzwerken mit der Nutzung der Ressource TextGrid Repository (vgl. Horstmann 2018) und den Methoden Named Entity Recognition (vgl. Schumacher 2018a) und Annotation in CATMA (vgl. Jacke 2018 und Schumacher 2019a).

Das Dramenquartett als Erweiterung des Disseminationsmodells in forTEXT

Die Dissemination einer digitalen Methode wie der Netzwerkanalyse durch ein nicht-digitales Kartenspiel bietet Möglichkeiten, die die bisher genannten digitalen Medien nicht abdecken konnten. Die Spielerïnnen werden in einer nicht-digitalen Umgebung mit den funktional reduzierten Ergebnissen einer digitalen Analyse konfrontiert, können diese visuell und haptisch erfahren und spielerisch explorieren. Der empfohlene Spielmodus ist ‚Supertrumpf'[3], bei dem Werte der Netzwerke verglichen werden. Die Spielregeln sind online veröffentlicht (https://dramenquartett.github.io/). Neben einem neuen und vor allem kompetitiven Blick auf Dramen – der die relationale Perspektive auf figürliche Kopräsenzen hervorhebt – wird zusätzlich die Neugier auf die den Netzwerken zugrunde liegende Methode geweckt, sodass in didaktisch-produktiver Hinsicht der Prozess einer Art *Reverse Engineering* im Sinne einer Mustererkennung auf unterschiedlichen Komplexitätsstufen angestoßen werden kann. Der Weg hin zu einem Umgang mit digitalen Ressourcen und Tools wie DraCor, ezlinavis und sogar die Anwendung eines komplexeren Tools wie Gephi ist damit geebnet, die kritische Methodenreflexion kann folgen. Dieser niedrigschwellige Zugang fügt sich in das auf Zugänglichkeit und Benutzerfreundlichkeit konzentrierte Disseminationsmodell von forTEXT ein und erweitert dieses durch den zusätzlichen Abbau von Schwellenängsten oder Vorbehalten gegen digitale Methoden.

Die im Folgenden vorgestellte, reflektierte und erprobte Pipeline geht von einer ersten theoretischen Annäherung durch forTEXT-Tutorials aus, auf die eine spielerische Vertiefung der spezifischen Objektkonstitution qua Netzwerkanalyse und der entsprechenden Metriken mittels des Dramenquartetts folgt. Anschließende Arbeitsphasen könnten, wie in 3. skizziert, z. B. die formalisierte Erstellung, Gestaltung und Analyse von Dramennetzwerken mittels ezlinavis und Gephi oder die konkrete Bearbeitung von literarhistorischen Forschungsfragen mittels DraCor umfassen.

Epistemische und didaktische Implikationen

Epistemische Dimensionen des Medienwechsels

Der quantifizierende Zugriff auf Dramentexte kann als „radikale ‚Anästhetisierung' der Objekte" (Trilcke, im Erscheinen) beschrieben werden. Auf die qua Formalisierung erfolgende Anästhetisierung, bei der die ursprüngliche ästhetische Dimension des literarischen Kunstwerks zunächst ausgesetzt wird, folgt jedoch eine reästhetisierende Transformation im Zuge der Diagrammatisierung (vgl. ebd.).[4] Die Dramen werden somit zunächst zwar nicht mehr primär als textuelle Artefakte wahrgenommen, dennoch aber als ästhetische Artefakte in Form ihrer netzwerkartigen Repräsentation, wodurch andere epistemische Dimensionen angesprochen und andere epistemische Praktiken vollzogen werden können (vgl. Trilcke und Fischer 2018). Dabei ist der Weg zurück zum Dramentext vom Kartenspiel über die digitale Darstellung der entsprechenden Netzwerke dadurch geebnet, dass Medien in Form „transkriptiver Bezugnahmen" (Jäger 2010, 301) generell intermedial aufeinander Bezug nehmen und Übersetzungsprozesse somit keine einseitig vorgegebene Richtung haben.[5]

Ein entscheidender Vorteil digitaler Diagramme ist die Möglichkeit der Interaktion (vgl. Horstmann, im Erscheinen): Netzwerke lassen sich je nach Wahl des Layoutalgorithmus unterschiedlich darstellen, ein semantischer Zoom ermöglicht überdies, zusätzliche Informationen des Ausgangsmaterials zu visualisieren. Dramennetzwerke in einer festgelegten (und damit nicht mehr veränderbaren) Form als Spielkarte zu drucken, bedeutet daher in erster Linie eine funktionale Reduktion. Gerade diese funktionale Reduktion eröffnet jedoch

didaktische Spielräume: Das Wissen, dass die abgedruckten Netzwerke ebenfalls in digitaler Form vorhanden und dort sogar manipulierbar sind, wird im Laufe des Spielprozesses die Neugier auf diese Funktionsvielfalt steigern, sodass der Übergang in die ‚digitale Arbeit' fließend stattfinden kann und nicht mehr als etwas kategorial anderes empfunden wird. Die Interaktion zwischen Benutzerïnnen und Netzwerken als konzeptioneller Bestandteil digitaler Netzwerkdarstellungen wird übertragen auf die Interaktion zwischen den Spielerïnnen, wodurch nicht zuletzt die von Jenkins (2006, 2) sog. *participatory culture* im nicht-digitalen Bereich eine Entsprechung erfährt.

Ansprechen unterschiedlicher Lerntypen

Das Kartenspiel entfaltet seinen didaktischen Mehrwert auch, weil es situational gerahmt ist: Es wird in kollektiven Unterrichtsphasen eingesetzt, die darauf abzielen, sich einem abstrakten Unterrichtsgegenstand auf spielerische Weise anzunähern. Da Menschen in ihrer Rolle als *visual beings* vor allem ihren Sehsinn als einen wichtigen Wahrnehmungskanal nutzen, um Informationen zu verstehen (vgl. Ward et al. 2010), stellen Visualisierungen bei der Präsentation von wissenschaftlichen Erkenntnissen ein wichtiges, den Verstehens- und Erinnerungsprozess begünstigendes Element dar. Der Einsatz des Kartenspiels greift darauf zurück und spricht unter den vier Lerntypen (auditiv, haptisch, kommunikativ und visuell) v. a. visuelle, aber auch kommunikative Lerntypen an, indem das Spiel die Kommunikation über Fachinhalte fokussiert und Sprache als Medium des Lernens einsetzt (vgl. Anselm und Werani 2017).

Im Fokus steht der Versuch, nicht nur kumulatives bzw. assimilatives Lernen zu initiieren, wodurch v. a. begrenztes, anwendungsorientiertes Wissen oder thematisch, anwendungsorientiertes Wissen produziert werden würde (vgl. Illeris 2010). Die – von der konkreten Kenntnis des Spielprinzips ‚Supertrumpf' unabhängige – spielerische Aktivierung unterschiedlicher Sinneskanäle und die damit einhergehende Diskussion über Fachinhalte zielt auf die Einleitung akkommodativer und transformativer Lernprozesse und darauf, über Fachwissen in relevanten Kontexten frei verfügen zu können.

Anwendung in der universitären Lehre und Lehrerïnnenbildung

Das im Wintersemester 2019/2020 an der Universität Hamburg durchgeführte Seminar „Digitale Literaturwissenschaft und pädagogische Praxis" hat unterschiedliche Standardverfahren und Werkzeuge erprobt, die gegenwärtig in der digitalen Literaturwissenschaft eingesetzt werden. Dieses Feld wird zunehmend auch für Lehrerïnnen insbesondere im gymnasialen Bereich relevant: Bereits die heutige Schülerïnnengeneration zählt zu den *digital natives*, für die der Umgang mit digitalen Medien und Werkzeugen selbstverständlich ist, die aber zugleich in Schule und/oder Studium in eine vertiefte *data literacy* eingeführt werden müssen. Der Transfer von Digital-Humanities-Methoden in den schulischen Bereich kann deshalb als wichtige Herausforderung identifiziert werden. Gleichzeitig geht es darum, das vernetzte Denken zu fördern, mithin literaturwissenschaftliche und fachdidaktische Zugänge zu *einem* Gegenstand stark zu machen. Um in Seminaren kein starres Wissen zu produzieren, auf das die angehende Lehrkräfte in der nächsten Phase ihrer Ausbildung – dem schulischen Alltag – nicht zugreifen können, muss die Kooperation zwischen Fachdidaktik und Fachwissenschaft gefördert werden. Neben der Einarbeitung in die Methoden steht deshalb die Frage der Komplexitätsreduktion und des schulischen Anwendungsbezuges im Zentrum des Seminars, wofür das DraCor-Kartenspiel exemplarisch herangezogen und getestet wird. Die konzeptionelle Einbettung des Kartenspiels in eine didaktische Heranführung an digitale Methoden ergänzend, wurde damit sowohl in diesem als auch im Seminar „Gender modellieren – Genderrollen und -stereotype in der Literatur des 19. Jahrhunderts", das ebenfalls im Wintersemester 19/20 an der Universität Hamburg angeboten wurde, eine praktische Anwendung durchgeführt, deren Erfolg qualitativ evaluiert wurde. Damit soll auch ein Beitrag zur Evaluation konkreter DH-Lehrformen geleistet werden.

Erste Ergebnisse

Um den Effekt des Dramenquartetts auf den Lernerfolg der Studierenden zu untersuchen, wurde eigens ein Testverfahren entwickelt, das die Wissensstände vor und nach dem Einsatz des Quartetts mess- und v. a. vergleichbar macht. Das Verfahren setzt sich aus fünf aufeinander aufbauenden Phasen zusammen:

(1) *Vorab: Gruppeneinteilung und eigenständige Vorbereitung* (Gruppe 1: Methodenbeitrag/Lerneinheit, Gruppe 2: Video-Tutorials)

Vorbereitend befasst sich ein Teil der Lerngruppe mit schriftlichen forTEXT-Lernmaterialien zur digitalen Netzwerkanalyse, während der andere Teil die Video-Fallstudien und -Tutorials konsultiert.[6]

(2) *Praxisphase 1: Erste Umfrage*

Ausgangspunkt der Erhebung stellt folglich ein gruppenspezifisch relativ homogener Wissensstand dar, der grundlegende Kenntnisse über die Methode der digitalen Netzwerkanalyse beinhaltet. Um die Wissensstände beider Gruppen vor dem Einsatz des Quartetts zu erfassen, wurde eine Umfrage entworfen und zu Beginn des Seminars in Einzelarbeit mit dem Audience Response System ARSnova durchgeführt. Die Umfragen adressieren mit jeweils neun Fragen drei Anforderungsbereiche (I: Reproduktionsleistung, II: Reorganisation- und Transferleistung, III: Reflexion und Problemlösung). Den Anforderungsbereichen entsprechend beinhalten sie Single-Choice-, Multiple-Choice- sowie Freitextfragen.

(3) *Praxisphase 2: Einsatz des Dramenquartetts*

Nach der ersten Quizphase wurde die gesamte Testgruppe in Kleingruppen eingeteilt, die im Supertrumpf-Modus das Dramenquartett spielen.

(4) *Praxisphase 3: Zweite Umfrage*

Eine zweite Umfrage erfasst den Wissensstand beider Gruppen, nachdem sie das Dramenquartett gespielt haben.

(5) *Auswertung der Umfrage: Erste Ergebnisse und Ausblick*

Die Auswertung des ersten Testdurchlaufs, der mit 11 Teilnehmenden durchgeführt wurde, verweist auf einen lernförderlichen Effekt des Dramenquartetts. Im Rahmen der ersten Quizrunde wurden 43% der Fragen, nach der zweiten Umfrage 52% der Fragen richtig beantwortet. Darüber hinaus verweist ein erster Blick auf die Freitextantworten darauf, dass der spielerische Zugang die intrinsische Motivation, sich über den Seminarkontext hinaus mit digitaler Netzwerkanalyse auseinanderzusetzen, steigert. Das erarbeitete Verfahren

zur vergleichenden Lernstandserhebung hat sich bewährt und wird in einem weiteren Seminar eingesetzt, um den Einfluss einer spielerischen Wissensvermittlung auf Kompetenz- und Wissensstand zu untersuchen.

Ausblick: zukünftig mögliche Arbeitsfelder

Das Projekt lotet das didaktische Potenzial von Gamification-Ansätzen in den DH konzeptionell und praktisch aus, indem es das DraCor-Kartenspiel mit Tools und Tutorials in einer didaktischen ‚Pipeline' verbindet und damit in die Disseminationsstrategie von forTEXT integriert. Der damit entwickelte Prototyp eines Konzepts, das auch fachdidaktisch Weiterentwicklungspotenzial birgt, ermöglicht diverse Adaptionen und Transformationen: in Hinblick auf die Netzwerkanalyse literarischer Texte, in Hinblick auf andere Methoden der Digital Humanities sowie in Hinblick auf das didaktische Szenario einer Verzahnung von analogen und digitalen Ansätzen.

So ließen sich auf der Grundlage der Netzwerkdaten aus anderen DH-Projekten, etwa zu Romanen, andere generische Karten-Sets entwerfen, wobei auch die – durch ezlinavis in Kombination mit Gephi ermöglichte – kollaborative Erstellung eigener Sets denkbar ist. Diese selbstständige Erstellung von Karten-Sets würde nicht zuletzt auch den haptischen Lerntyp ansprechen. Eine Weiterentwicklung der didaktischen Engführung von Analogem und Digitalem ließe sich über eine Verzahnung des Kartenspiels mit der digital-interaktiven Repräsentation der einzelnen Dramen auf DraCor vornehmen (z. B. über QR-Codes). Unter didaktischen Gesichtspunkten bietet sich des Weiteren die Möglichkeit, kreativ-produktionsorientierte Elemente in die skizzierte Pipeline einzubauen, etwa indem die Lernenden Netzwerke ‚erfinden', die sie zunächst händisch zeichnen und dann – den Schritt in den digitalen Raum machend – mittels ezlinavis formal erfassen müssen.

Der im Projekt durchgeführte Testlauf soll in diesem Sinne zu einer weiteren Diskussion über didaktische Potenziale sowohl von Gamification-Ansätzen als auch der Verzahnung von analogen und digitalen Lehrmitteln anregen und damit grundsätzlich der Reflexion über didaktische Szenarien dienen, die den spielerischen, kreativen Übergang zwischen lebensweltlich vertrauten Situationen und der Abstraktion digitaler Forschungsprozesse gestalten.

Fußnoten

1. Vgl. https://fortext.net/ressourcen/videos/fallstudien/analyse-der-figurennetzwerke-in-lessings-emilia-galotti.
2. Vgl. https://fortext.net/ressourcen/videos/tutorials/netzwerkanalyse-und-literaturanalyse.
3. Vgl. https://de.wikipedia.org/wiki/Supertrumpf.
4. Zum Diagrammatikbegriff vgl. etwa Krämer 2016.
5. Zu intermedialen Übersetzungsprozessen vgl. Schmid, Veits und Vorrath 2018.
6. Beide Seminare richten sich ausdrücklich an Studierende ohne technische Vorkenntnisse. Es ist also davon auszugehen, dass die Personen beider Testgruppen über keinerlei Vorbildung bezüglich Methoden der digitalen Netzwerkanalyse verfügen.

Bibliographie

Anselm, Sabine / Werani, Anke (2017): *Kommunikation in Lehr-Lernkontexten*. Bad Heilbrunn: Klinkhardt.

Fischer, Frank / Kittel, Christopher / Milling, Carsten / Trilcke, Peer / Wolf, Jana (2018): „Dramenquartett – Eine didaktische Intervention", in: *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts*, 397–398. DOI: https://doi.org/10.6084/m9.figshare.5926363.v1.

Fischer, Frank / Schultz, Anika (2019): „Dramenquartett – Eine didaktische Intervention". Unter Mitarbeit von Christopher Kittel, Carsten Milling, Peer Trilcke und Jana Wolf. 32 Blatt in Kartonbox, Farbdruck. Bern: edition taberna kritika 2019. (Spielanleitung: https://dramenquartett.github.io/)

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): „Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor", in: Sahle, Patrick (ed.): *DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 194–197.

Horstmann, Jan (2018): „TextGrid Repository", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/ressourcen/textsammlungen/textgrid-repository [letzter Zugriff 12. September 2019].

Horstmann, Jan (im Erscheinen): „Textvisualisierung: Epistemik des Bildlichen im Digitalen", in: Huber, Martin / Krämer, Sybille / Pias, Claus (eds.): *Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen*, CompaRe: Fachinformationsdienst Allgemeine und Vergleichende Literaturwissenschaft.

Horstmann, Jan / Schumacher, Mareike (2019): „Social Media, YouTube und Co: Multimediale, multimodale und multicodierte Dissemination von Forschungsmethoden in forTEXT", in: Sahle, Patrick (ed.): *DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 207–211. DOI: 10.5281/zenodo.2596095.

Illeris, Knud (2010): *Lernen verstehen. Bedingungen erfolgreichen Lernens*. Bad Heilbrunn: Klinkhardt.

Jacke, Janina (2018): „Manuelle Annotation", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/manuelle-annotation [letzter Zugriff 12. September 2019].

Jäger, Ludwig (2010): „Intermedialität – Intramedialität – Transkriptivität: Überlegungen zu einigen Prinzipien der kulturellen Semiosis", in: Deppermann, Arnulf / Linke, Angelika (eds.): *Sprache intermedial: Stimme und Schrift, Bild und Ton*. Berlin, New York: de Gruyter 299–324. DOI: 10.1515/9783110223613.299.

Jenkins, Henry (2006): *Convergence Culture: Where Old and New Media Collide*. New York, London: New York University Press.

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis. Grundlinien einer Diagrammatologie*. Berlin: Suhrkamp.

Mega, Carolina / Ronconi, Lucia / De Beni, Rossana (2014): „What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement", in: *Journal of Educational Psychology,* Vol 106(1), 121–131.

Odebrecht, Carolin / Burnard, Lou / Navarro Colorado, Borja / Eder, Maciej / Schöch, Christof (2019): „The European Literary Text Collection (ELTeC)", in: *DH 2019. Complexities.* Utrecht University. [Poster.]

Schmid, Johannes C. P. / Veits, Andreas / Vorrath, Wiebke (eds. 2018): *Praktiken medialer Transformationen. Übersetzungen in und aus dem digitalen Raum*. Bielefeld: transcript. DOI: 10.14361/9783839441145.

Schumacher, Mareike (2018a): „Named Entity Recognition (NER)", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/named-entity-recognition-ner [letzter Zugriff 12. September 2019].

Schumacher, Mareike (2018b): „Netzwerkanalyse", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/netzwerkanalyse [letzter Zugriff 12. September 2019].

Schumacher, Mareike (2019a): „CATMA", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/tools/tools/catma [letzter Zugriff 12. September 2019].

Schumacher, Mareike (2019b): „Netzwerkanalyse mit Gephi", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/lerneinheiten/netzwerkanalyse-mit-gephi [letzter Zugriff 12. September 2019].

Trilcke, Peer / Fischer, Frank (2018): „Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen", 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_003.

Trilcke, Peer (im Erscheinen): „Small Worlds, Change Rates und die Netzwerkanalyse dramatischer Texte. Reflexionen aus dem Rabbit Hole", in: Jannidis, Fotis / Winko, Simone / Rapp, Andrea / Meister, Jan Christoph / Stäcker, Thomas (eds.): *Digitale Literaturwissenschaft. DFG-Symposium Villa Vigoni, 2017*. Berlin, New York: de Gruyter.

Ward, Matthew / Grinstein, Georges / Keim, Daniel (2010): *Interactive Data Visualization. Foundations, Techniques, and Applications.* Wellesley: Peters.

OMMR4all - ein semiautomatischer Online-Editor für mittelalterliche Musiknotationen

Wick, Christoph
christoph.wick@uni-wuerzburg.de
Universität Würzburg, Deutschland

Hartelt, Alexander
alexander.hartelt@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Insbesondere für Musikwissenschaftler im Bereich von historischen Manuskripten besteht der Wunsch nach digitalen Bibliotheken, die die gewaltigen Mengen an Material in maschinenlesbare Form (z. B. MEI) speichern. Die Kodierung der alten Werke ist jedoch oft sehr mühselig, da großer menschlicher Einsatz erforderlich ist. Das Aufkommen von künstlicher Intelligenz offenbart hier neue Ansätze, um die Arbeitsprozesse größtmöglich zu automatisieren, indem Algorithmen aus dem Bereich der optischen Musikerkennung (OMR) eingesetzt werden.

Die im Folgenden vorgestellte Software OMMR4all (Optical Medieval Music Recognition For All) realisiert diesen Ansatz für mittelalterliche Manuskripte, die in verschiedenen Neumennotationen, z. B. Quadratnotation, geschrieben sind. Der semi-automatische Workflow erwartet eine einzelne eingescannte Seite als Eingabe und erzeugt als Ausgabe die kodierte Musik z. B. als MEI oder in einer graphischen Anzeige. Hierbei werden verschiedene existierende OMR-Werkzeuge zur Notenlinien-, Notensystem- und Symbolerkennung eingesetzt, die mit einem Overlay-Editor zur Korrektur kombiniert werden. Neumen werden gemäß dem aktuellen MEI-Standard (4.0.1) repräsentiert. Eine manuell in einem modernen Stil gerenderte Beispieltranskription zeigt Abbildung 1.

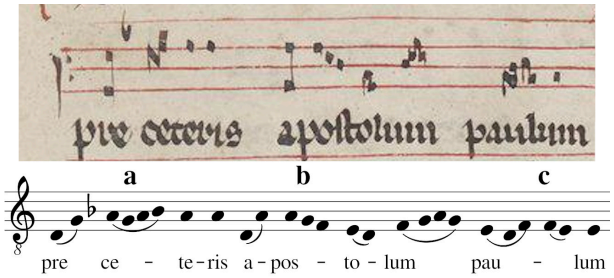


Abbildung 1: Beispieltranskription. Eine oder mehrere Neumen werden zu Silben zugeordnet. Eine Neume besteht wiederum aus einzelnen Notenkomponenten, die entweder graphisch (Bindebogen in a) oder logischen (kleiner Abstand in b) zur vorherigen Komponente verbunden sind. Mehrere Neumen, die zu einer Silbe gehören, werden mit einem größeren Abstand dargestellt (c).

Vigliensoni et al. (2019) stellten bereits einen vergleichbaren OMR-Workflow, der im SIMSSA-Projekt (Fujinaga 2014) eingebettet ist, für Musik des Mittelalters und der Renaissance vor. Hierbei arbeitet die OMR mittels eines Neuronalen Netzes (Calvo-Zaragoza 2018), das jeden Pixel des Originalmanuskripts in verschiedene Klassen wie z. B. Note, Notenzeile, Hintergrund oder Text einteilt. Die automatische Ausgabe kann durch das Webtool Pixel.js korrigiert werden (Zeyad 2017). Für die Weiterverarbeitung werden die klassifizierten Bilden in Ebenen gleichen Typs separiert, sodass ein nachfolgender Algorithmus nicht mehr das Originalbild, sondern ein Binärbild, das z. B. nur noch Notenlinienpixel oder Musiksymbolpixel umfasst. Ein weiterer Algorithmus kann

so die Musiksymbole separat erkennen, welche anschließend im Overlay-Editor Neon.js (Regimbal 2019) korrigiert werden können. Der Workflow teilt mehrere Gemeinsamkeiten mit OMMR4all, dessen Umsetzung zeigt jedoch auch Grenzen auf. So entfällt in OMMR4all die erforderliche, teils mühsame pixelgenaue Korrektur zum Erzeugen der separaten Typebenen, da die in OMMR4all verwendeten OMR-Algorithmen direkt auf dem Originalmanuskript arbeiten. Auch arbeitet der von uns vorgestellte neue Overlay-Editor näher am Original, da Notensysteme und Musiksymbole akkurat an die Positionen im Manuskript gezeichnet werden, was einen sehr schnellen Abgleich von Vorlage und Vorhersage ermöglicht.

OMMR4all

Workflow

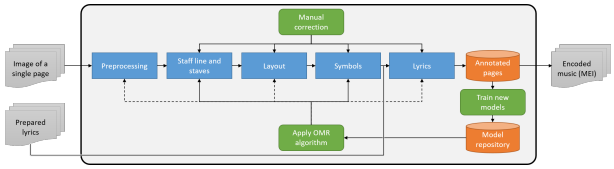


Abbildung 2: Der Workflow von OMMR4all.

Der Workflow von OMMR4all ist in Abbildung 2 gezeigt. Der hochauflösende Scan einer Seite dient neben dem vorbereiteten Liedtext (z. B. in einer Textdatei) als Eingabe. Das Bild wird zunächst durch eine automatische Vorverarbeitung geradegestellt und binarisiert. Anschließend werden Notenlinien und Notensysteme mittels eines Fully-Convolutional Neuronalen Netzes (FCN) erkannt. Darauf aufbauend werden Musik- oder Textregionen separiert, wobei im Standardfall keine exakten Regionen erforderlich sind und so die Layoutanalyse vollständig automatisch abläuft. Basierend auf den Notensystemen werden nun durch ein weiteres FCN Neumen, Notenschlüssel und Vorzeichen erkannt. Abschließend werden die Silben des vorgefertigten Liedtextes an die passenden Neumen gesetzt. Die Algorithmen zu Notenlinien, Notensystem- und Symbolerkennung wurden hierbei direkt von Wick et al. (2019) übernommen.

Iterativer Trainingsansatz

OMMR4all ermöglicht das Training von individuellen Modellen für die Notenlinien- und Symboldetektion, um die automatische Erkennung auf einen spezifischen, aber noch unbekannten Stil eines Buches anzupassen. Das Training erfordert wenige Seiten an manuell ausgezeichneten Material. Auch dieser Schritt kann durch Verwendung existenter ähnlicher Modelle semiautomatisch erfolgen.

Im Allgemeinen ist zu erwarten, dass die Notenzeilenerkennungsmodelle sehr gut auf verschiedene Notationsstile generalisieren, da Linien in allen Notationen sehr ähnlich sind. Die Symbolnotationen weisen hingegen eine größere Varianz auf, wie beispielsweise an Gotischer- oder Quadrat-Notation zu sehen ist.

Softwarearchitektur

OMMR4all[1] ist eine quelloffene Software[2], die ein auf einer REST API basierendes Client-Server-Modell implementiert und eine Benutzerverwaltung umfasst. Dies ermöglicht eine niedrige Einstiegshürde für den Einsatz der Software in der Forschung von Musikwissenschaftlern, da keine Installation nötig und die Web-Applikation plattformunabhängig ist. Zusätzlich wird die Last des Rechnersystems zum Trainieren neuer Modell auf den Server ausgelagert. Dieser kann hierbei mit high-end GPUs ausgestattet werden, um die Rechenzeiten weiter zu reduzieren. Auch werden die Daten zentralisiert gespeichert, was einen weltweiten Zugang von jedem internetfähigen Arbeitsplatz ermöglicht. Demnach ist jeder einfache Laptop oder Desktop PC als Zugriffspunkt zu OMMR4all vollkommen ausreichend und sofort einsetzbar.

Overlay-Editor

Da nicht zu erwarten ist, dass die automatischen Tools von OMMR4all perfekt arbeiten, müssen die Ergebnisse in eleganter und benutzerfreundlicher Art korrigiert werden. Dies ermöglicht der integrierte Overlay-Editor, der eine Überlagerung der Annotationen und der Originalseite anzeigt. Unterschiede von Musiksymbolen können so leicht und mit einem Blick festgestellt und anschließend korrigiert werden. Außerdem können Kommentare hinzufügt werden, um kritische oder unklare Stelle zu markieren, die dann auch in den kritischen Apparat aufgenommen werden können oder die zur Kommunikation zwischen dem Editor und einem Reviewer, der die Nachkorrektur durchführt, dienen können.

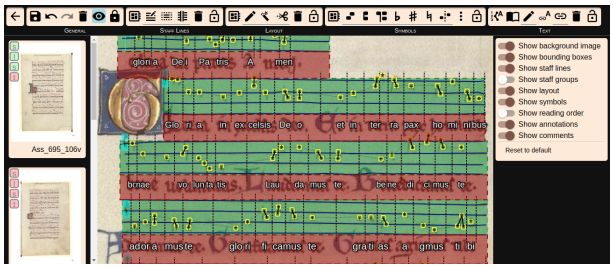


Abbildung 3: Benutzeroberfläche des Editors. Grüne Regionen definieren Notensysteme, rote Regionen Liedtext. Individuelle Notenkomponenten werden als gelbe Boxen dargestellt, grafische Verbindungen von Neumen werden als durchgezogene schwarze Linien, die zwei Noten verbinden gezeichnet, wohingegen die gestrichelten vertikalen Linien den Start einer neuen Neume angeben. Notenschlüssel sind türkis markiert. Die Silben des Liedtextes sind unter der zugehörigen Neume in der jeweiligen Textregion ausgerichtet. Die verschiedenen Knöpfe der Werkzeugleiste dienen zur Korrektur der Annotationen oder um die automatischen Algorithmen zu starten. Nicht gezeigt sind Lesereihenfolge oder Kommentare des Korrektors.

Abbildung 3 zeigt die Benutzeroberfläche des Editors. Der Editor ist konzipiert, damit er ohne steile Lernkurve leicht bedient werden kann: Die Interaktionen zur Selektion, zur Bewegung, zum Ziehen oder zum Einfügen von Elementen erfolgen mit der Maus. Erfahrene Nutzer können jedoch auf Tastaturkürzel zurückgreifen, um den Bearbeitungsprozess zu beschleunigen.

Evaluation

Zur Evaluation der Transkriptionszeit verglichen wir OMMR4all mit Monodi+ (Eipert 2019), einem Tool, das unter anderem einen hochentwickelten Editor für eine tastaturbasierte Eingabe von Cantus Planus (monophone Musik der Westlichen Kirche) anbietet. Die Bedienung der beiden Softwaresysteme erfolgte stets durch Experten des jeweiligen Programms. Als Material wählten wir jeweils fünf Seiten in gotischer und Quadratnotation (eine Beispielseite wird in Abbildung 3 bearbeitet). Wir maßen und verglichen die mittleren Zeiten, die bei Vorlage des Liedtextes nötig waren, um eine korrekte Transkription der Manuskripte zu erzielen. Somit wurden nur die Zeiten zur Korrektur oder Eingabe von Notenlinien und Musiksymbole gemessen. Die Modelle für die Quadratnotation waren auf 49 Seiten trainiert, die aus einem weiteren Buch stammen. Die gotischen Modelle basieren auf vier weiteren Seiten aus dem identischen Buch. Tabelle 1 fasst die Ergebnisse zusammen.

Tabelle 1: Evaluation der Transkriptionszeiten in Minuten. Wir listen die Anzahl der Symbole, die erforderlichen Zeiten für die Korrektur der Notenlinien, der Symbole, der Silbenzuordnung, und die Gesamtzeit. Alle Werte sind gemittelt und relativ zu einer Seite angegeben.

Notation	#Symbole	OMMR4all	Monodi+	Speed-up			
		Notenlinien	Symbole	Silben	Gesamt		
Gotisch	158	0,3	2,3	1,8	4,5	5,6	1,3
Quadrat	267	0,6	3,3	2,9	6,9	8,5	1,2

OMMR4all zeigt einen Speed-Up von 1,3 und 1,2. Hierbei ist zu beachten, dass die Bearbeitungszeit mittels Monodi+ bereits am Limit ist, da jedes Symbol manuell eingegeben werden muss, worauf das Interface perfekt zugeschnitten ist. OMMR4all hingegen kann sich stets weiterentwickeln und selbständig genauere Modelle lernen. Insbesondere wenn die verwendeten Modelle ausgehend von der aktuellen Fehlerrate von 10% genauer arbeiten, ist mit einer deutlichen Reduktion der Bearbeitungszeit für die Symboleingabe zu rechnen. Weiterhin verspricht die Entwicklung einer automatischen Silbenerkennung und -zuweisung eine weitere Beschleunigung. Natürlich können, falls die Modelle menschliche Genauigkeit erreichen, alle Seiten vollständig automatisch verarbeitet werden. Im Allgemeinen liefert OMMR4all im Vergleich zu einer manuellen Eingabe des Notentextes eine inhärente Erklärungskomponente für den Ursprung eines jeden Symbols, was insbesondere für einen kritischen Apparat relevant ist.

Geplante Erweiterungen

Trotz der vielen Funktionen von OMMR4all, existieren etliche weitere mögliche Verbesserungen oder Erweiterungen. Einige anstehenden Aufgaben werden im Folgenden vorgestellt.

Zunächst planen wir verschiedene Werkzeuge und Algorithmen, um den Liedtext automatisch zu erfassen. Das Hauptproblem hierbei ist, dass derzeit keine OCR-Engine verlässlich mit handgeschriebenen Text ohne spezielles Training umgehen kann. Deswegen soll in einem ersten Schritt zunächst die automatische Silbenzuordnung gelöst werden, wobei immer noch der vorgefertigte Liedtext vorliegen muss. Hierzu werden die fehlerhaften Ergebnisse der OCR-Engine Calamari (Wick 2018) als Vorschläge für die Position verwendet indem die bestmögliche Übereinstimmung von erkanntem und korrektem Text gefunden wird. Vorläufige Ergebnisse sind vielversprechend, selbst wenn ein großer Prozentsatz der erkannten OCR-Zeichen falsch ist. Die eigentliche automatische Erfassung des handgeschriebenen Textes stellt eine große Herausforderung dar, da eine Seite meist nur wenige Zeilen umfasst, jedoch viele manuell transkribierte Zeilen für ein Training erforderlich sind. Alle modernen Ansätze verwenden hierfür ein Sprachmodell mit n-Grammen oder zumindest einem Wörterbuch sowie tiefe Neuronale Netze, meist bidirektionale LSTMs, die mit CNNs gekoppelt werden. Chammas et al. (2018) verwendeten mehrere tausend Seiten mit reinem Text von beliebiger Handschrift und erhielten eine Wortgenauigkeit von etwa 80%. Auf sauberer geschriebenen mittelalterlichen Manuskripten erzielten Fischer et al. (2014) eine Wortgenauigkeit von etwa 93% mit etwa 11.000 Wörtern im Trainingsdatensatz und einem eingeschränkten Vokabular von etwa 5.000 Wörtern. In einem Szenario, in dem nur die Wörter des Trainingsdatensatzes bekannt waren, wurde eine Wortgenauigkeit von unter 78% erreicht, da etwa 15% der Wörter unbekannt waren. Eine Umsetzung der Techniken für Liedtexte steht noch aus, denn hier besteht ein zusätzliches Problem, dass viele Worte in Silben aufgeteilt sind, was den Einsatz von Wörterbüchern erschwert.

Andere Pläne umfassen weitere monophone Notationsstile zu unterstützen. Hierunter fallen sowohl ältere Neumennotationen, sowohl solche ohne Notenlinien, als auch spätere Mensurnotationen. Hierzu bedarf es nur kleinere kosmetischer Änderungen am Editor, jedoch ist größerer Aufwand beim Entwickeln neuer Algorithmen nötig. Polyphone Notationen, auch solche bei denen die Stimmen jeweils in separaten Notensystemen vorliegen, stellen semantische bzw. hierarchische Änderungen der Musiknotation dar, da z. B. zwei oder mehrere gleichzeitig klingende Notenzeilen zu einer Akkolade zusammengefasst werden müssen, was Änderungen am Datenformat und somit auch am Editor erfordert.

In der Praxis wird OMMR4all im Corpus Monodicum Projekt[3] der Universität Würzburg eingesetzt, um den Prozess der Transkription der Bestände von einstimmiger Musik des lateinischen Mittelalters zu beschleunigen. Der Overlay-Editor, der mit einem Blick erlaubt Fehler zu erkennen, hebt den Transkriptionsprozess bereits auf ein hohes Qualitätsniveau. Trotzdem ist zur Qualitätssicherung ein zweistufiger Prozess notwendig, indem ein musikwissenschaftlicher Reviewer das Ergebnis des Transkriptionsprozesses überprüft, das in Zusammenarbeit von OMMR4all und einem menschlichen Editor erzeugt wurde. Technisch wird dies durch die Möglichkeit unterstützt, pro Seite eine Freigabe zu dokumentieren oder, wie oben erwähnt, Kommentare zur Nachbearbeitung anzugeben.

Fußnoten

1. Eine Demo-Anwendung, die das testweise Bearbeiten zweier unterschiedlicher Bücher erlaubt ist unter https://ommr4all.informatik.uni-wuerzburg.de/ verfügbar.
2. Der Quellcode kann auf https://github.com/OMMR4all/ eingesehen werden.
3. http://www.musikwissenschaft.uni-wuerzburg.de/forschung/corpus-monodicum/

Bibliographie

Calvo-Zaragoza, Jorge / Castellanos, Francisco / Vigliensoni, Gabriel / Fujinaga, Ichiro (2018): "Deep neural networks for document processing of music score images", in: *Applied Sciences 8: 654.*

Chammas, Edgard / Mokbel, Chafic / Likforman-Sulem, Laurence (2018): "Handwriting Recognition of Historical Documents with Few Labeled Data", in: *13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna:. 43-48.*

Eipert, Tim / Herrmann, Felix / Wick, Christoph / Puppe, Frank / Haug, Andreas (2019): "Editor Support for Digital Editions of Medieval Monophonic Music", in: *Proceedings of the 2* nd *International Workshop on Reading Music Systems (submitted to).*

Fischer, Andreas / Baechler Michael / Garz, Angelika / Liwicki, Marcus / Ingold, Rolf (2014): "A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents", in: 11th IAPR International Workshop on Document Analysis Systems, Tours, 2014: 71-75.

Fujinaga, Ichiro / Hankinson, Andrew / Cumming, Julie E. (2014): "Introduction to SIMSSA (single interface for music score searching and analysis)", in: *Proceedings of the 1st International Workshop on Digital Libraries for Musicology: 1-3.*

Saleh, Zeyad / Zhang, Ké / Calvo-Zaragoza, Jorge / Vigliensoni, Gabriel / Fujinaga, Ichiro (2017): "Pixel. js: Web-based pixel classification correction platform for ground truth creation.", in: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR): 39-40.*

Regimbal, Juliette / McLennan, Zoé / Vigliensoni, Gabriel / Tran, Andrew / Fujinaga, Ichiro (2019): "Neon2: A verovio-based square-notation editor", in: *Music Encoding Conference 2019.*

Saleh, Zeyad / Zhang, Ké / Calvo-Zaragoza, Jorge / Vigliensoni, Gabriel / Fujinaga, Ichiro (2017): "Pixel.js: Web-based pixel classification correction platform for ground truth creation.", in: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR): 39-40.*

Vigliensoni, Gabriel / Daigle, Alex / Lui, Eric / Calvo-Zaragoza, Jorge / Regimbal, Juliette / Nguyen, Minh Anh / Baxter, Noah / McLennan, Zoe / Fujinaga, Ichiro (2019): "O vercoming the challenges of optical music recognition of early music with machine learning", in: *DH2019* .

Wick, Christoph / Hartelt, Alexander / Puppe, Frank (2019): "Staff, Symbol and Melody Detection of Medieval Manuscripts written in Square Notation using Deep Fully Convolutional Networks", in: *Applied Sciences 9* .

Wick, Christoph / Reul, Christian / Puppe, Frank (2018): "Comparison of OCR Accuracy on Early Printed Books using the Open Source Engines Calamari and OCRopus", in: *JLCL: Special Issue on Automatic Text and Layout Recognition: 79-96.*

Partizipatives Design in Digital Humanities Projekten: Checklist, Maßnahmenkatalog und Use-Case

Dogunke, Swantje
swantje.dogunke@gmail.com
HTWK Leipzig, Deutschland

Während die Begriffsbestimmung für Virtuelle Forschungsumgebungen weitestgehend abgeschlossen scheint (Arbeitsgruppe Virtuelle Forschungsumgebungen In Der Allianz Der Deutschen Wissenschaftsorganisationen 2011), diese bereits längst selbst Untersuchungsgegenstand geworden sind (Klein 2012), fehlt bisher eine methodische Auseinandersetzung, wie der Aufbau einer solchen digitalen Infrastruktur tatsächlich die Anforderungen und Bedürfnisse der potentiellen Nutzerschaft treffen könnte.

Denn der Erfolg für digitale Infrastruktur und Services, die unter dem Dach der Digital Humanities entstehen, wird häufig an Nutzer*innen- oder Zugriffszahlen gemessen. Hieran wird entschieden, ob Projekte weiter gefördert oder in den Betrieb überführt werden. Seit fast zehn Jahren wird für den Aufbau virtueller Forschungsumgebungen empfohlen, mit Nutzer*innen gemeinsam oder zumindest nutzer*innenzentriert diese fachspezifische digitale Infrastruktur zu entwickeln (Kommission Zukunft der Informationsinfrastruktur 2011). Die Bedarfsanalyse stellt einen zeitaufwendigen und kaum abzuschließenden Teil in jedem Projekt dar. Eine Option, um das im Call for Papers genannte Problem der Umwandlung von geistes- und kulturwissenschaftlichen Fragestellungen in Anforderungen an digitale Infrastruktur und Services anzugehen, wäre der Einsatz von partizipativem Design. Versteht man Design als eine Schnittstelle zwischen Technologie und Gesellschaft, ist eine starke und frühe Beteiligung späterer Nutzer*innen am Designprozess eine naheliegende Idee, um potentielle Fehlentwicklungen bereits zu Beginn zu vermeiden (Cross und Design Research Society 1972: 6).

Im Beitrag werden zunächst partizipative Entwicklungsansätze vorgestellt und nach ihrem Partizipationsgrad anhand eines Schemas der International Association for Public Participation eingeordnet. Das Schema sieht fünf Stufen der Partizipation „inform“, „consult“, „involve“ „collaborate“ und „empower“ vor und stellt diese mit einem implizitem Versprechen an die Beteiligten in Beziehung (International Association for Public Participation 2018). Konzepte wie User Experience Design oder User-Centred-Design sehen eine scharfe Rolleneinteilung zwischen den Usern, Forscher*innen und Designer*innen vor und bewegen sich häufig in einem Spektrum von „consult“ und „involve“. Co-Design scheint nicht nur aufgrund einer hohen Partizipationsstufe „collaborate“, sondern auch aufgrund seiner offeneren Organisationsstruktur am ehesten der fächerübergreifenden Herangehensweise in den Digital Humanities zu entsprechen.

In anderen Design-Konzepten wie z.B. Design Thinking oder Service Design werden Daten über User, ihr Verhalten und

ihre Emotionen gesammelt, ausgewertet und dienen den Designer*innen im weiteren Design-Prozess als Grundlage. Im Co-Designprozess verschwimmen diese Rollen, alle Beteiligten durchlaufen gemeinsam die verschiedenen Phasen eines Co-Designprozesses. Während also beispielsweise in den verwandten Konzepten Daten ausgewertet werden, um Personas zu erstellen, die als Repräsentationen für typische Nutzer*innen dienen (Tomitsch u. a. 2018: 100), sind im Co-Design reale Personen am Entwicklungsprozess beteiligt. Es geht also nicht um die Frage, was würde ein*e Nutzer*in tun, sondern Nutzer*innen bringen im Prozess ihre Bedürfnisse sowie Ideen ein und können dadurch – je nach Partizipationsgrad - den Entwicklungsprozess beeinflussen oder gar steuern. Hier werden auch die Grenzen der Skalierbarkeit von Co-Design deutlich. Sollen Nutzer*innen stellvertretend für andere Nutzer*innen an dem Co-Design-Prozess teilnehmen, wird die Auswahl dieser Nutzer*innen das Ergebnis stark beeinflussen. Daher wurden als Anwendungsgebiete vor allem die Entwicklung spezialisierter Services identifiziert, so z.B. virtuelle Arbeitsumgebungen für bestandsbezogene Forschungsprojekte an Bibliotheken, Archiven und Museen.

Als Ursprünge für das heutige Verständnis von Co-Design werden in der Forschungsliteratur Projekte partizipativen Designs in Skandinavien ab den 1970er Jahren genannt, in denen gemeinsam mit Beschäftigten verbesserte Arbeitsplätze entwickelt wurden (Sanders und Stappers 2008: 7). Die wissenschaftliche Auseinandersetzung zur Notwendigkeit von partizipativem Design lässt sich durch eine der ersten Konferenzen 1972 in Manchester belegen (Cross und Design Research Society 1972). In beiden Fällen wird betont, dass das Erfolgsversprechen von partizipativem Design nur eingelöst werden kann, wenn Endnutzer*innen am gesamten Design-Prozess beteiligt sind. Als Vorteile einer solchen Herangehensweise werden u.a. in der Literatur eine gesteigerte Anzahl innovativer Ideen und Vorschläge von Nutzer*innen (Mitchell u. a. 2016), ein erweitertes Wissen um ihre Bedürfnisse, der positive Einfluss auf interdisziplinäre Zusammenarbeit innerhalb der Organisation, eine höhere Qualität der Services sowie ein vermindertes Risiko des Scheiterns genannt. Zudem könnten Entscheidungen schneller und besser getroffen werden und somit die Entwicklungszeit verkürzen. Im Gesamtergebnis sei mit einer erhöhten Zufriedenheit und Bindung von Nutzer*innen zu rechnen (Steen, Manchot, und De Koning 2011). Es wird eine möglichst frühe Einbindung der Nutzer*innen empfohlen, da somit zudem ein hohes Potential zur Kostenersparnis vermutet wird (Kujala 2003).

Der Planungsphase eines Co-Design-Prozesses beginnt mit der Auswahl oder Erstellung eines Grundgerüsts, welches die Phasen des Projekts in seinem Gesamtverlauf darstellt. Die Festlegung des Partizipationsgrades und das enthaltene Versprechen an die Nutzer*innen sollte zu Beginn erfolgen. Der Grad könnte durch eine Institution vorgegeben werden oder von Nutzer*innen eingefordert werden. Es sollten sich bei Co-Design-Prozessen divergente und konvergente Elemente in den jeweiligen Phasen abwechseln. In einem divergenten Teil einer Phase besteht das Ziel darin, durch geeignete Maßnahmen eine möglichst hohe Anzahl an Ideen, Vorschlägen und Optionen zu generieren. Im konvergenten Teil einer Phase werden die Vorschläge ausgewählt, die weiterverfolgt werden. Für jede dieser Phasen werden Ziele definiert, die mit geeigneten Maßnahmen umgesetzt werden. Jede Design-Maßnahme arbeitet mit einer starken Visualisierung der erhobenen Daten, deren Ordnung und dem Herausarbeiten ihrer Zusammenhänge. Gerade in interdisziplinären Teams können so schnell unterschiedliche Kommunikationsweisen verschiedener Fachdisziplinen zusammengeführt werden (Calabretta, Gemser, und Karpen 2016: 46).

Die in der Literatur erwähnten Vorteile wurden in eine Checkliste für Digital Humanities-Projekte umgewandelt, um Potentiale für den Einsatz von Co-Design zu erkennen. Sie wird im Beitrag als Management-Tool für Institutionen vorgestellt, um bereits zu einem frühen Zeitpunkt im Projektverlauf die Rahmenbedingungen für einen Co-Design-Prozess herstellen zu können. Die Liste wurde anhand von Digital Humanities Projekten im Forschungsverbund Marbach Weimar Wolfenbüttel und der Bibliotheca Hertziana durch Experteninterviews getestet.

Beschriebener Vorteil	Fragen	Beispiel
Mehr Informationen: Häufig ist es aufgrund fehlender Informationen schwierig, das Problem der Nutzer*innen zu verstehen und den Design task zu formulieren. Das starke Einbeziehen der Nutzer*innen kann dazu beitragen, passgerechtere Lösungen zu entwickeln (Visser et al., 2005: 119).	Waren die Ergebnisse von Maßnahmen (Fragebögen, Interviews, Beobachtungen) der Anforderungsanalyse ungeeignet für eine genauere Problembeschreibung? Fällt es schwer, das Problem oder den design task Fachkolleg*innen zu erklären?	Die Anforderungsanalyse ergab, dass sich die Nutzer*innen (Geisteswissenschaftler*innen) mehr Tools zum kollaborativem Arbeiten wünschen. Diese werden vorgestellt und der Gruppe in einer VFU bereitgestellt. Sie werden jedoch kaum genutzt, Dokumente werden per Mail im Umlaufverfahren erstellt und gepflegt.
Kommunikation in heterogenen Gruppen verbessern: In heterogenen Gruppen, wie z.B. interdisziplinären Forschungsgruppen wird die Kommunikation durch unterschiedliche Forschungsperspektiven und fachspezifische Kommunikationskulturen erschwert (Muller und Druin, 2017). Co-Design arbeitet mit einer starken Visualisierung und regt die Kommunikation durch nicht textbasierte Modelle an.	Sind im Team Menschen aus verschiedenen Fachrichtungen mit unterschiedlichen Professionalisierungsgraden vertreten? Zum Beispiel Informatiker*innen und Geisteswissenschaftler*innen mit wenig Erfahrung in DH-Projekten? Oder Bibliothekar*innen, Professor*innen und studentische Hilfskräfte? Ist es schwierig eine gemeinsame Sprache zur Formulierung der Anforderungen zu finden?	In einer Forschungsgruppe wird deutlich, dass ein gemeinsames kontrolliertes Vokabular für die Verschlagwortung einzelner Dokumente nötig wird. Die Bibliothekarin im Team hat bereits einen Thesaurus entwickelt und stellt diesen zur Diskussion. Die Fachwissenschaftler*innen, die diese Verschlagwortung vornehmen werden, wollen als einfachere Lösung eine nicht-hierarchische Tagsammlung verwenden, da sie so schneller verschlagworten können. Der Informatiker wird mit seinem Vorschlag für eine Ontologie nicht gehört.
Zeitdruck: Eine Anforderungsanalyse selbst ist ein zeitintensiver Teil jedes Projekts. Durch die Einbeziehung der Nutzer*innen kann die Zeit zur Erstellung eines neuen Releases reduziert werden (Alam 2002: 254). Die am Prozess beteiligten Anwender benötigen weniger Zeit, um die Nutzung des Dienstes zu erlernen.	Wird in den nächsten Wochen ein erster Entwurf erwartet? Haben Sie bereits viel Zeit im Projekt mit der Anforderungsanalyse verbracht, ohne brauchbare Ergebnisse zu erhalten?	Ein Forschungsprojekt zur Untersuchung von NS-verfolgungsbedingt entzogenen Kulturgütern ist für drei Jahre finanziert. Es wird eine Arbeitsdatenbank benötigt. Die Beteiligten haben noch keine Erfahrung mit Datenmodellierung.
Innovationsdruck: User als "Experten ihrer eigenen Erfahrungen" generieren eine höhere Anzahl von Ideen mit einem höherem Innovationspotential. Co-Design kann hilfreich sein, diese Ideen in divergenten Phasen zu sammeln (Kristensson et al 2002: 59) und die passenden für die anschließende Weiterentwicklung auszuwählen.	Ist keine Lösung für das Problem vorhanden? Gibt es keine Vergleichsprojekte? Wurden Lösungen getestet und verworfen?	Ein Forschungsprojekt untersucht die Leihgaben von Johann Wolfgang von Goethe aus der herzoglichen Bibliothek. Zur Erfassung der Daten werden bibliographische Werkzeuge erprobt, die jedoch nicht in der Lage sind, rudimentäre Einträge ("ein Zeichnungsportfolio") und die Ausleihdaten zu erfassen.
Identifikation und Loyalität: Wenn die Nutzer in den gesamten Entwurfsprozess eingebunden sind, erhöht sich die Wahrscheinlichkeit, dass sie die Dienstleistung oder das Produkt tatsächlich nutzen, auch wenn dies eine Veränderung im Alltag bedeutet (Woods, 2017: 97).	Welche persönlichen Hürden müssen die Nutzer*innen überwinden? Was können wir tun, um die Nutzer zu halten? Wie kann die Nutzung unseres Produktes zur Routine werden?	Es wird eine Arbeitsdatenbank für ein Forschungsprojekt erstellt, da an verteilten Orten Daten erzeugt werden. Die persönlichen Datensammlungen sollen abgelöst werden. Nach einiger Zeit stellt sich heraus, dass die ursprünglichen Systeme weiter mit Daten beliefert werden, die neue Datenbank jedoch nicht genutzt wird.
Kleine Gruppe von Nutzer*innen: Ansätze wie UXP eignen sich, wenn ein Dienst für eine große Gruppe von Personen entwickelt werden soll. Dies wird durch Techniken wie die Erstellung von Persona ermöglicht. Co-Design eignet sich eher für spezialisierte Dienste für eine kleine Gruppe von Nutzer*innen.	Suchen Sie eine Lösung für eine kleine Gruppe von Nutzer*innen? Kennen Sie sie alle? Passen sie in einen Raum?	Ein Team von zehn Geisteswissenschaftler*innen hat Daten zu Autorenbibliotheken des 18. Jahrhunderts gesammelt, kommentiert und aufbereitet. Eine projektübergreifende Datenvisualisierung ist angedacht.

Des Weiteren wurde ein Katalog entwickelt, um Maßnahmen aus dem Bereich des Co-Designs hinsichtlich ihrer Eignung und dem erreichten Partizipationsgrad zu bewerten. Zur Festlegung des geeigneten Zeitpunktes einzelner Maßnahmen wurde ein angepasstes Schema aus dem Bereich des Service Design verwendet (Stickdorn u. a. 2018), welches die Phasen „Planen und Vorbereiten", „Recherche", „Ideen finden" und

„Prototyping“ umfasst. Der Start eines Co-Designprozesses wird durch einen Trigger eingeleitet, der durch die Anwendung der Checkliste erkannt werden kann. Als Endpunkt des Co-Design-Prozesses wird ein Release festgelegt. Ein weiteres Merkmal, welches den Maßnahmen zugeordnet wurde, ist die Einschätzung, ob es sich eher um eine divergente Maßnahme der Ideenfindung oder eine konvergente Maßnahme der Bewertung, Auswahl oder Konkretisierung handelt.

In einem Use-Case konnte gezeigt werden, dass mithilfe von Co-Design innerhalb eines eintägigen Workshops ein Konzept zur Erstellung einer digitalen Arbeitsumgebung für ein Forschungsprojekt erstellt werden konnte. Der Co-Design-Ansatz und die eingesetzten Maßnahmen, wie z.B. Customer Journey Mapping oder eine mithilfe von LEGO-Steinen erstellte Stakeholderanalyse unterstützten die Anforderungsanalyse und führten zu neuen Sichtweisen in der Zusammenarbeit von Geistes- und Kulturwissenschaftler*innen und Digital Humanities-Mitarbeiter*innen und könnten die Kommunikation in interdisziplinären Digital Humanities-Projekten verbessern sowie den Ressourceneinsatz verringern. Die bisweilen spielerische Herangehensweise motivierte die Teilnehmenden, sich auch intensiv mit Themen des Projektmanagements auseinanderzusetzen, die in der Vergangenheit eher als lästig angesehen wurden.

Der Vortrag stellt die Checkliste und einen erweiterten Maßnahmenkatalog als eine Art Werkzeugkasten für die iterative Entwicklung von digitaler Infrastruktur für Forschende und Institutionen vor und zeigt an einem Use-Case, wie Co-Design-Maßnahmen zu einer verbesserten Bedarfsanalyse führen können. Auch die Grenzen von Co-Design sollen beleuchtet werden. In der Fachliteratur überwiegen die positiven Berichte von Co-Design-Projekten, eine Untersuchung der Grenzen der Skalierbarkeit ist jedoch nicht zu finden. Um die Attraktivität von Co-Design für die Hochschulen, Forschungseinrichtungen oder Forschungsgruppen zu steigern, könnten umfassende Untersuchungen zur Kostenersparnis und der Erhöhung der Zufriedenheit von Nutzer*innen in Best-Practice-Projekten hilfreich sein. Eine Bedingung hierfür wäre ein vergleichbares Vorgehen.

Das Ziel des Beitrages ist es, den Austausch und die Zusammenarbeit zwischen Digital Humanities-Forschenden anzuregen, die am Aufbau digitaler Services oder Infrastruktur beteiligt sind, um gemeinsam eine Art Toolkit für Co-Design in den Digital Humanities zu erstellen. Die Vorstellung einer Co-Design-Maßnahme - der Motivation Matrix - im Rahmen des Panels „Digital Humanities from Scratch“ auf der DHd-Jahrestagung 2019 in Frankfurt am Main stieß bei den Teilnehmenden auf reges Interesse und lässt erahnen, dass der Einsatz von Co-Design in den Digital Humanities begrüßt wird und erfolgsversprechend sein könnte (Cremer, Roeder, und Söring 2019).

Bibliographie

Alam, Ian (2002): “An Exploratory Investigation of User Involvement in New Service Development.” Journal of the Academy of Marketing Science 30, no. 3 (June 1, 2002): 250. https://doi.org/10.1177/0092070302303006.

Calabretta, Giulia / Gerda Gemser / Ingo Karpen (2016): *Strategic Design: Eight Essential Practices Every Strategic Designer Must Master*. Amsterdam: BIS.

Cremer, Fabian / Torsten Roeder / Sibylle Söring (2019): „Digital Humanities ‚from Scratch‘“, Juni. https://doi.org/10.5281/zenodo.3244179.

Cross, Nigel / Design Research Society, Hrsg (1972): *Design participation: proceedings of the Design Research Society's conference, Manchester, September 1971*. London: Academy Editions.

International Association for Public Participation (2018): „IAP2 Spectrum of Public Participation“. https://cdn.ymaws.com/www.iap2.org/resource/resmgr/pillars/Spectrum_8.5x11_Print.pdf.

Klein, Julia Elisabeth (2012): „Virtuelle Forschungsumgebungen als Entwicklungsfeld für Bibliotheken am Beispiel des ‚Deutschen Textarchivs‘“. Master's Thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät I. https://doi.org/10.18452/14175.

Kristensson, Per / Peter R. Magnusson / Jonas Matthing (2002): “Users as a Hidden Resource for Creativity: Findings from an Experimental Study on User Involvement.” Creativity and Innovation Management 11, no. 1: 55–61. https://doi.org/10.1111/1467-8691.00236.

Kommission Zukunft der Informationsinfrastruktur (2011): „Gesamtkonzept für die Informationsinfrastruktur in Deutschland“, 254.

Kujala, Sari (2003): „User Involvement: A Review of the Benefits and Challenges“. *Behaviour & Information Technology* 22 (1): 1–16. https://doi.org/10.1080/01449290301782.

Mitchell, Val / Tracy Ros / Andrew May / Ruth Sims / Christopher Parker (2016): „Empirical Investigation of the Impact of Using Co-Design Methods When Generating Proposals for Sustainable Travel Solutions“. *CoDesign* 12 (4): 205–20. https://doi.org/10.1080/15710882.2015.1091894.

Muller, Michael J. / Allison Druin: “Participatory Design: The Third Space in HCI,” n.d., 70.

UX Magazine (2017): “Participatory Design in Practice | UX Magazine”. https://uxmag.com/articles/participatory-design-in-practice.

Sanders, Elizabeth B.-N. / Pieter Jan Stappers (2008): „Co-Creation and the New Landscapes of Design“. *CoDesign* 4 (1): 5–18. https://doi.org/10.1080/15710880701875068.

Sleeswijk Visser, Froukje / Pieter Jan Stappers / Remko van der Lugt / Elizabeth B.-N. Sanders (2005): “Contextmapping: Experiences from Practice.” CoDesign 1, no. 2 (April 2005): 119–49. .

Steen, Marc / Menno Manchot / Nicole De Koning (2011): „Benefits of Co-design in Service Design Projects“. *International Journal of Design* 5 (3). http://www.ijdesign.org/index.php/IJDesign/article/view/890/346.

Stickdorn, Marc / Markus Hormess / Adam Lawrence / Jakob Schneider, Hrsg. (2018): *This Is Service Design Doing: Applying Service Design Thinking in the Real World; a Practitioners` Handbook*. First edition. Sebastapol, CA: O'Reilly.

Tomitsch, Martin / Cara Wrigley / Madeleine Borthwick / Naseem Ahmadpour / Jessica Frawley / A. Baki Kocaballi / Claudia Núñez-Pacheco / Karla Straker / Lian Loke (2018): *Design. Think. Make. Break. Repeat: A Handbook of Methods*. Amsterdam: BIS Publishers B.V.

Woods, Leana / Elizabeth Cummings / Jed Duff / Kim Walker (2017): “Design Thinking for MHealth Application Co-Design to Support Heart Failure Self-Management.” In Context Sensitive Health Informatics: Redesigning Healthcare Work. IOS Press.

Passive Präsenz tragischer Hauptfiguren im Drama

Willand, Marcus
marcus.willand@gs.uni-heidelberg.de
Universität Heidelberg, Deutschland

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

Pagel, Janis
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart

Reiter, Nils
nils.reiter@uni-koeln.de
Universität Stuttgart; Universität zu Köln

Einleitung

Dramen entwerfen einen fiktiven sozialen Raum (Bourdieu 1985), dessen Bewohner sich ständig *aktiv* und *passiv* sozial verhalten, also entweder selbst dramatisch handeln oder zum *passiven* Gegenstand dramatischer (Ver-)Handlungen werden. Pointiert ließe sich sagen, sie hassen und lieben, bzw. werden geliebt und werden gehasst. Von der Forschung wurde indes nur selten betont,[1] dass die *passive Präsenz* von Figuren – also das Sprechen über Figuren, die in einer Szene nicht auftreten – ebenso interpretationsrelevant ist wie deren aktive Handlungen. So untergräbt beispielsweise die tragische Protagonistin des Stücks *Emilia Galotti* (Lessing 1772) die moralische Integrität ihrer sozialen Klasse, indem sie einen anderen Mann als ihren standesgemäß verlobten Appiani verehrt: den Prinzen. Tragödienfähige Fallhöhe (Schopenhauer [1818] 1977) erreicht Emilia aber nicht durch ihr Begehren, sondern durch ihr begehrt werden.

In unserem Beitrag möchten wir den Zusammenhang zwischen aktiver und passiver Figurenpräsenz in dramatischen Texten untersuchen, indem wir quantitative und qualitative Analysen kombinieren. In einem ersten Schritt entwickeln wir eine Operationalisierung für eine computergestützte Analyse aktiver und passiver Präsenz und werden in einem zweiten Schritt die aus den Analysen resultierenden Ergebnisse mit besonderem Fokus auf Hauptfiguren diskutieren.

Forschungsdiskussion

Seit einigen Jahren gilt die Netzwerkanalyse als eine der zentralen Forschungsgebiete innerhalb der digitalen Dramenanalyse. Typischerweise modellieren Netzwerke auf Basis von Konfigurationsmatrizen (vgl. Marcus 1973, ins. S. 308ff. und Pfister 2001, S. 235–240) aber nur die aktive (szenische) Präsenz von Figuren (Moretti 2011; Trilcke u.a. 2015; Piper u.a. 2017), obwohl Ko-Präsenz-Relationen nur einen eingeschränkten Aussagewert bezüglich der „soziale Welt" eines Dramas zulassen. Denn sie beruhen lediglich auf Informationen über die Anzahl an Szenen, in denen Figuren gemeinsam auftreten. Aktive Figuren wurden aber natürlich auch anders beforscht. Karsdorp u.a. (2015) stellen einen Ansatz zur automatischen Bestimmung von Liebesbeziehungen vor, Willand und Reiter (2017) verwenden semantische Wörterbücher, um Figurenrede und Geschlecht in einen Zusammenhang zu stellen. Nalisnick und Baird (2013) analysieren das *Sentiment* aktiver Figuren, allerdings um ihre Dialogpartner zu charakterisieren und dadurch Wendepunkte in den Figurenbeziehungen zu identifizieren. Die passive Präsenz ist bisher noch nicht eingehend untersucht worden.

Aktive und passive Figurenpräsenz

Die aktive Präsenz von Figuren lässt sich unterschiedlich operationalisieren, etwa indem der Anteil der Rede einer Figur an der Gesamtrede einer Szene oder eines Akts gemessen wird. Abb. 1 zeigt dies für die fünf Akte von *Emilia Galotti*. Jeder Balken repräsentiert einen Akt, jede Farbe den Anteil einer Figur an der Gesamtrede des Akts:

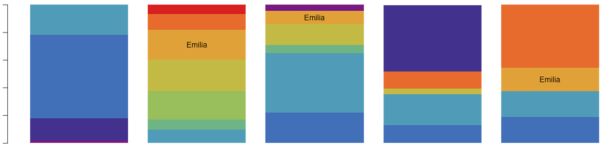


Abbildung 1: Emilias aktive Präsenz in den fünf Akten des Stücks. Die Farben indizieren unterschiedliche Figuren.

In den Akten 1 und 4 ist Emilia überhaupt nicht aktiv präsent. In den Akten 2, 3 und 5 ist sie es, aber der Anteil ihrer Rede vergleichsweise gering. Wieso aber ist sie titelgebende Protagonistin dieses Stücks, wenn sie doch kaum handelt? Deutlich wird das, betrachtet man ihre passive Präsenz:

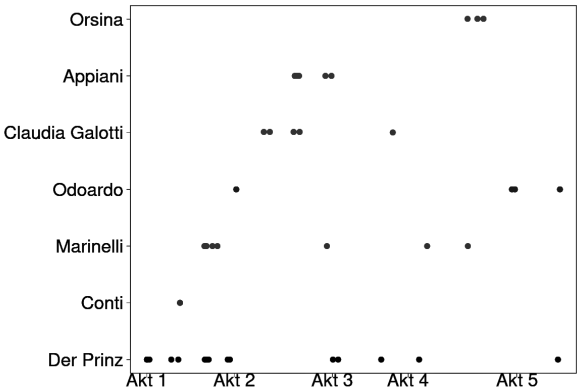


Abbildung 2: Emilias passive Präsenz im Verlauf des Stücks, gemessen anhand von Nennungen ihres Namens durch andere Figuren.

Die Punkte in Abb. 2 repräsentieren die Erwähnungen des Namens „Emilia“ in der Rede anderer Figuren (y-Achse) im Verlauf des Stücks (x-Achse). Sie zeigen, dass Emilia während des gesamten Stücks von allen Figuren wiederholt erwähnt wird. Diese *passive Präsenz* unterscheidet sie von Nebenfiguren. Betrachtet man die aktivsten Figuren des Stücks (Prinz, Marinelli), so lässt sich in Abb.3 erkennen, dass Emilia genauso oft namentlich erwähnt wird wie diese:

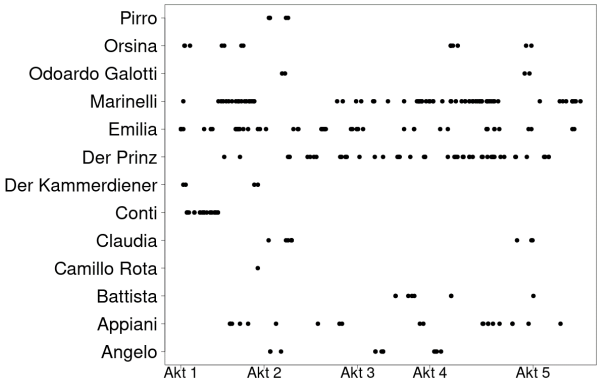


Abbildung 3: Namensnennungen aller Figuren in *Emilia Galotti*.

Zusammenfassend liefern diese Analysen Argumente für die Interpretationshypothese, dass Emilia den dramatischen Konflikt nicht selbst aktiv löst – was sie zur positiven Hauptfigur machen würde –, sondern lediglich auslöst. So wird sie zum passiv-tragischen Gegenstand der Figurenhandlung.

Präsenz messbar machen

Für jede Figur definieren wir die *aktive Präsenz* als genau die Anzahl an Szenen, in denen diese Figur spricht. So lassen sich Konfigurationsmatrizen erstellen, wie sie auch für Netzwerkanalysen verwendet werden. Die *passive Präsenz* wird anhand der Anzahl an Szenen extrahiert, in denen eine Figur namentlich erwähnt wird, aber nicht selbst präsent ist (schließlich können Figuren nicht gleichzeitig aktiv und passiv anwesend sein). Auf Figuren wird auch durch nichtnamentliche Erwähnung referiert, etwa mittels Pronomen oder Nominalphrasen – wobei Pronomen den Großteil der Erwähnungen ausmachen. Eine vollständige Annotation aller Erwähnungen erlaubt weiterführende Analysen, etwa anhand von Netzwerken, die beispielsweise darstellen können, auf welche Weise auf Figuren referiert wird (Nominalphrasen, Pronomen etc.).[2]

Gleich in mehrfacher Hinsicht handelt es sich bei der Methode um eine Heuristik: Einerseits erfasst die Analyse von Figurennamen längst nicht alle Erwähnungen einer Figur. Wir gehen aber davon aus, dass der Figurenname mindestens einmal in jeder Szene genannt wird, in der auf eine Figur referiert wird. Andererseits können Szenen sehr unterschiedlich lang ausfallen, was für die hier durchgeführten Analysen der passiven Präsenz unberücksichtigt bleibt. Anders formuliert: Jede Szene ist bei dieser Form der Analyse gleich gewichtet. Unterschiedliche poetologische Funktionalisierungen von Szenen, wie sie im Verlauf der Dramengeschichte zu beobachten sind, u.a. anhand des Abrückens von der *liaison de scène* als regelpoetischem Dogma, löst die Heuristik also nicht auf. Die verschiedene Funktionalisierung gilt es somit in der späteren Interpretation zu reflektieren.

Sowohl die aktive als auch die passive Präsenz wird der besseren Vergleichbarkeit halber über die Zahl der Szenen normalisiert. Dafür wird die Menge an aktiven Auftritten sowie passiven Erwähnungen einer Figur durch die Gesamtzahl der Szenen im Drama geteilt, sodass der Gesamtwert der Figurenaktivität immer zwischen 0 (spricht nie/wird nie erwähnt) und 1 (spricht in jeder Szene/wird immer erwähnt) liegt. Somit ergibt sich für die Berechnung:

$$Aktive\ Präsenz = \frac{Anzahl\ Szenen\ mit\ Auftritt}{Anzahl\ Szenen}.$$

Identifikation handlungskonstitutiver Figuren

Die automatische Erkennung von Hauptfiguren in dramatischen Texten ist bisher nur in Ansätzen versucht worden (Krautter & Pagel 2019; Fischer u.a. 2018), sie würde auf dem Gebiet der digitalen Dramenanalyse aber die Grundlage für erkenntnisversprechende Anschlussfragen schaffen.[3] Bloße aktive Präsenz ist für die Identifikation von Hauptfiguren aber nicht ausreichend, denn einige Figurentypen – wie der griechische Chor oder Dienerfiguren – sind häufig sehr präsent, in Bezug auf die Konfliktlösung jedoch irrelevant. Wir adressieren dieses Problem, indem wir sowohl die aktive als auch die passive Figurenpräsenz berücksichtigen.

Diese Operationalisierung erlaubt es uns, die figuren- und gattungsspezifische Verteilung der Resultate zu vergleichen und so bisher ungesehene Aspekte von Hauptfiguren zu identifizieren. In diesem Beitrag stellen wir das Genre *Bürgerliches Trauerspiel* (BT) und die Strömung *Sturm und Drang* (SD) gegenüber.

Korpus

Unser Korpus enthält deutschsprachige dramatische Texte aus der Zeit zwischen 1750 und 1800 (Fischer u.a. 2019). Es ist in zwei Teilkorpora aufgeteilt, die auf sehr unterschiedlichen Poetiken beruhen: Sechs Stücke des *Bürgerlichen Trauerspiels* (BT) und sechs Stücke des *Sturm und Drang* (SD). Diese Subkorpora erscheinen relativ klein, aber um die historische Korrektheit und Interpretierbarkeit der Ergebnisse zu gewährleisten, unterliegt unser Korpus-Design strengen Kriterien: So beschränken wir die Textauswahl auf lediglich diejenigen Tragödien, die konsensual und eindeutig einer Textgruppe zugeordnet werden können. Deshalb ist unter anderem Friedrich Maximilian Klingers Schauspiel *Sturm und Drang*, dessen Titel später zur Epochenbezeichnung wurde, nicht im Korpus enthalten. Ebenfalls nicht im Korpus vertreten sind Goethes Stücke *Clavigo* und *Stella*, Gerstenbergs *Ugolino* und Heinrich Leopold Wagners *Die Kindermörderin*. Diesen Dramen fehlt eine weitere Untergliederung der Akte in Szenen, wodurch die Präzision der vorgeschlagenen Präsenzmessung anhand von Szenengrenzen erheblich leiden würde. Durch die divergierende

Granularität der Segmentierung wären die Präsenzwerte der verschiedenen Dramen kaum mehr vergleichbar.

Ergebnisse

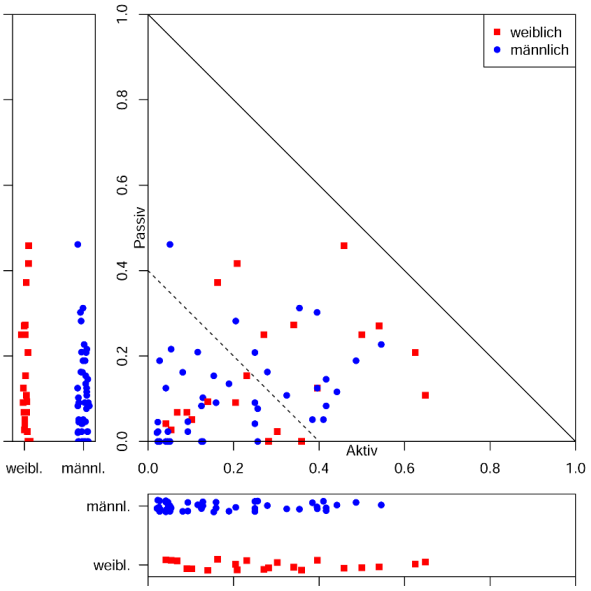


Abbildung 4: Figurenverteilung im *bürgerl. Trauerspiel.*

Abb. 4 visualisiert die präsentische Auswertung des BT-Korpus. Jeder Punkt stellt dabei eine Figur dar. In den Stücken treten fast so viele weibliche wie männliche Figuren auf, wobei die sowohl aktiv als auch passiv präsentesten Figuren überraschenderweise weiblich sind. Zudem werden keine Extremwerte erreicht: Alle aktiven Präsenzwerte liegen unter 0,7, alle passiven unter 0,5. Eine Gesamtpräsenz von 1 ist bei keiner Figur zu beobachten. Basierend auf den Präsenzwerten kann ein Schwellenwert etabliert werden, der ungefähr bei 0,4 liegt. Dieser Schwellenwert (gestrichelte Linie in Abb. 4) ergibt sich hier nicht vollständig induktiv aus den Daten, sondern wird theoriegeleitet gesetzt. An dieser Stelle greifen die formale, quantitative und die qualitative Analyse ineinander.

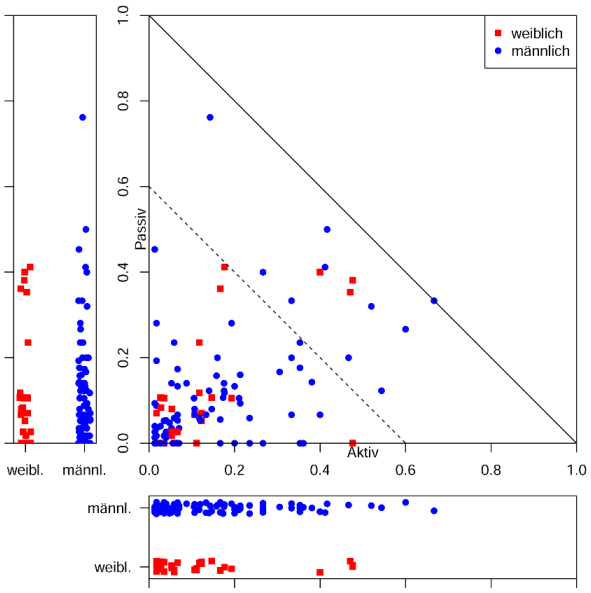


Abbildung 5: Figurenverteilung im *Sturm und Drang.*

Die Figurenverteilung in Abb. 5 (SD-Korpus) unterscheidet sich von derjenigen in Abb. 4 deutlich. Der Schwellenwert liegt hier mit 0,6 viel höher. Um als Hauptfigur zu gelten, muss eine Figur im SD also eine höhere Gesamtpräsenz aufweisen als im BT. Dies ist eine der zentralen Erkenntnisse dieses Forschungsbeitrags. Die Figuration (vgl. hierzu Elias 2002) dramatischer Hauptfiguren scheint somit textgruppenspezifisch und durch die Ermittlung des Präsenzwertes analysierbar zu sein.

Darüber hinaus ist das Geschlecht ein relevanter Faktor im Gattungsvergleich. Im SD treten insgesamt weniger weibliche Figuren auf und nur 3 von 11 Hauptfiguren sind weiblich. Zudem sind die weiblichen Figuren eher passiv präsent, während die männlichen überwiegend aktiv präsent sind. Auch die insgesamt aktivsten Figuren sind jeweils männlich. Nur eine einzige Figur erreicht den maximalen Präsenzwert von 1, nämlich Guelfo in Klingers *Die Zwillinge*:

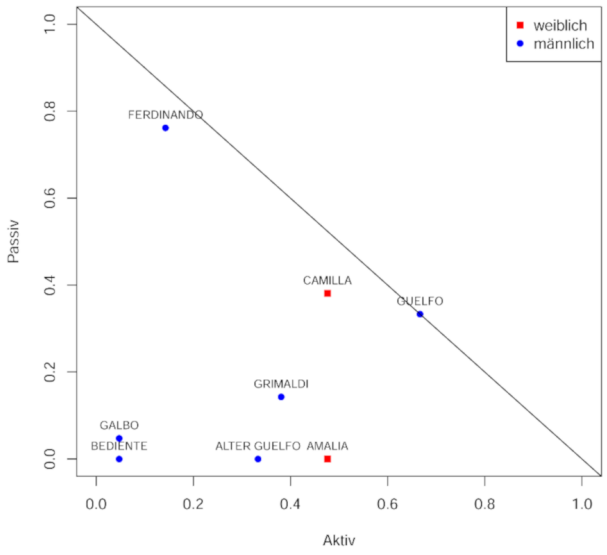


Abbildung 6: Figurenpräsenz in Klingers *Die Zwillinge* (1776).

Nicht zuletzt ist die *Korrelation* von aktiver und passiver Präsenz bei Hauptfiguren aufschlussreich: Hauptfiguren (über dem Schwellenwert 0,6) im SD zeigen eine mittelstarke positive Korrelation an ($\rho = 0{,}56$, Pearson-Korrelation). D.h., je aktiver eine Hauptfigur in Dramen des Sturm-und-Drang präsent ist, desto mehr wird auch über sie gesprochen. Im Gegensatz dazu finden wir im BT (Schwellenwert 0,4) eine schwach negative Korrelation ($\rho = -0{,}1$), d.h. hier sinkt die passive Präsenz bei zunehmender aktiver Präsenz leicht.[4]

Diese Ergebnisse sind ein gewichtiger Hinweis auf grundlegend divergierende Bauprinzipien dramatischer Texte, die sich offenbar nicht nur durch Handlungen und Themen unterscheiden, sondern auch durch die spezifische Präsenzgestaltung von Hauptfiguren. Da diese Unterschiede durch lineares Lesen jedoch kaum identifiziert werden können, möchte dieser Forschungsbeitrag als Argument für die Erweiterung der qualitativ-interpretierenden Dramenanalyse durch quantitative Methoden verstanden werden.

Weitere Möglichkeiten der Präsenzmessung

Die aktive Präsenz einer Figur lässt sich auch anhand anderer Einheiten skalieren, etwa anhand der Akte oder der Gesamtzahl der in einem Drama gesprochenen Repliken. Es wäre ebenfalls möglich, den Wert der aktiven Präsenz als die Zahl der gesprochenen Tokens (i.d.R. Wörter und Satzzeichen) und den Wert der passiven Präsenz als die Zahl namentlicher Nennungen aufzufassen. Dadurch könnten einige zuvor beschriebene Problematiken ausgeräumt werden, etwa die der differierenden Szenenlängen. Figuren, die nur kurze Passagen sprechen oder punktuell erwähnt werden, hätten dann vermutlich kleinere Aktivitätswerte, als es bei der szenisch gebundenen Präsenzberechnung der Fall ist. Die Rede- und Erwähnungsverteilung dürfte als näher an der vom Zuschauer bzw. Leser wahrgenommenen Realität des fiktiven sozialen Raums liegen. Hierbei stellen sich allerdings auch neue Herausforderungen. So kommt der Koreferenz von Figuren ein deutlich größeres Gewicht zu. Da wir zuverlässige Koreferenzen momentan nur für einzelne Stücke manuell annotiert vorliegen haben und somit auf die namentlichen Erwähnungen beschränkt sind, ergeben sich unter Umständen stark fehlerbehaftete Werte: Wenn etwa Figuren, die nur selten namentlich Erwähnung finden, überproportional stark auf andere Weise referenziert werden. Ist die namentliche Erwähnung hingegen an einzelne Szenen gebunden, hat diese Fehlerquelle geringeren Einfluss auf die Werte. Um die Auswirkungen der unterschiedlichen Operationalisierungen zumindest einer ersten Exploration zu unterziehen, nehmen wir die Präsenzanalyse von Klingers *Die Zwillinge* ein zweites Mal vor. Dazu definieren wir die aktive Präsenz als die Anzahl an gesprochenen Tokens einer Figur, die passive Präsenz als Anzahl namentlicher Erwähnungen einer Figur.

Abb. 7 zeigt die Präsenzanalyse für *Die Zwillinge* wobei für aktive und passive Präsenz gilt:

$$Aktive\ Präsenz = \frac{Anzahl\ Tokens\ einer\ Figur}{Gesamtzahl\ Tokens},$$

$$Passive\ Präsenz = \frac{Anzahl\ Erwähnungen\ einer\ Figur}{Gesamtzahl\ Erwähnungen}.$$

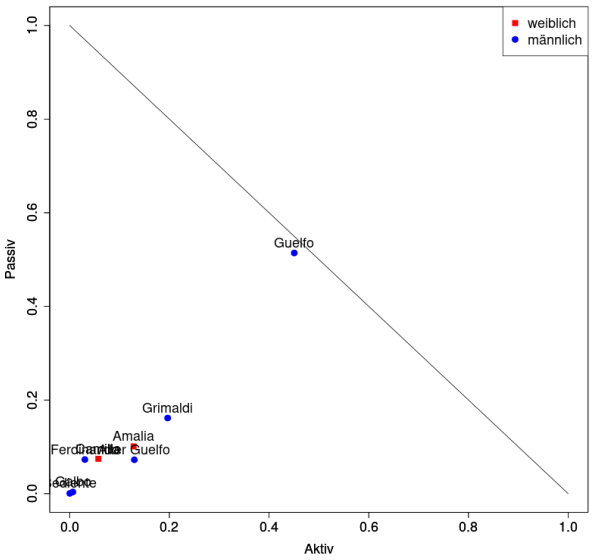


Abbildung 7: Figurenpräsenz in Klingers *Die Zwillinge* (1776) gemäß alternativer Operationalisierung.

Verglichen mit den Präsenzwerten in Abb. 6 ergeben sich erhebliche. Aufgrund der wenigen namentlichen Erwähnungen sinkt vor allem die passive Präsenz von Ferdinando von fast 0,8 auf etwa 0,1. Ferdinando wird also konsistent in vielen Szenen erwähnt, die Zahl der Erwähnung bleibt aber insgesamt vernachlässigbar, vergleicht man seinen Wert mit Guelfo. Wir gehen jedoch davon aus, dass die Fehleranfälligkeit der Koreferenzheuristik hier insgesamt ungenauere Ergebnisse liefert.

Zusammenfassung

Der Beitrag stellt eine Methode vor, die ein erweitertes Präsenzkonzept operationalisiert, das neben der aktiven Präsenz dramatischer Figuren auch die passive Präsenz umfasst. Die passive Präsenz operationalisieren wir als Zahl der Szenen, in der eine Figur namentlich erwähnt wird, ohne selbst aktiv auf der Bühne zu stehen. Die Ergebnisse unserer Korpusanalysen lassen auf unterschiedliche Bauprinzipien dramatischer Texte schließen, die an die spezifische Präsenz von Hauptfiguren gebunden sind. Für die Zukunft erscheint es fruchtbar, die hier eruierten Erkenntnisse im Lichte poetologischer Setzungen und Funktionalisierungen – etwa der Vorbildfunktion Shakespeares – zu untersuchen.

Anhang: Korpora

Sturm und Drang **(SD) Korpus**

- Klinger, Friedrich Maximilian: *Die neue Arria*
- Klinger, Friedrich Maximilian: *Die Zwillinge*
- Leisewitz, Johann Anton: *Julius von Tarent*
- Schiller, Friedrich: *Die Räuber*
- Schiller, Friedrich: *Die Verschwörung des Fiesco zu Genua*
- Goethe, Johann Wolfgang: *Götz von Berlichingen mit der eisernen Hand*

Bourgeois Tragedy **(BT) Korpus**

- Engel, Johann Jakob: *Eid und Pflicht*
- Hebbel, Friedrich: *Maria Magdalene*
- Holtei, Karl von: *Ein Trauerspiel in Berlin*
- Lessing, Gotthold Ephraim: *Emilia Galotti*
- Lessing, Gotthold Ephraim: *Miss Sara Sampson*
- Pfeil, Johann Gottlob Benjamin: *Lucie Woodvil*

Fußnoten

1. Juliane Vogel (2012) diskutiert das bereits von Johann Georg Sulzer kritisierte „Ankündigungswesen“ des höfischen Theaters (S. 536), das oft stark ritualisierte Formen der Einführung von Figuren vorschreibt; diese Einführung von Figuren durch andere sind Teil des Phänomens, das wir als *passive Präsenz* analysieren.
2. Die Veröffentlichung eines Korpus, das vollständige und teilannotierte Dramentexte enthält und auch für Experimente zur automatischen Koreferenzauflösung genutzt werden kann, ist für 2020 geplant. Auf https://github.com/quadrama/gerdracor-coref ist bereits ein Pre-Release verfügbar (DOI: 10.5281/zenodo.3559207).
3. Pfister (2001, 226f.) beschreibt das Problem der Identifikation von Hauptfiguren als eine Frage der „quantitativen Dominanzrelationen“ zwischen Figuren.
4. Die passive Präsenz hängt auch von der Aktivität der Figur selbst ab. Ihr Wert muss bei hoher aktiver Präsenz automatisch klein ausfallen, da nur Szenen beachtet werden, in denen die Figur nicht spricht.

Bibliographie

Bourdieu, Pierre (1985): *Sozialer Raum und "Klassen". Zwei Vorlesungen*. Frankfurt a. M.: Suhrkamp.

Elias, Norbert (2002): "Etablierte und Außenseiter", in: Hammer, Heike / Blomert, Reinhard (eds.): *Gesammelte Schriften*. Band 4. Frankfurt a. M.: Suhrkamp.

Fischer, Frank / Trilcke, Peer / Kittel, Christopher / Skorinkin, Daniil (2018): "To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)", in: *DH 2018 Conference Abstracts*, Mexico-City: 193–201 https://dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf [letzter Zugriff 18. Dezember 2019.

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): *"Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor"*, in: *DHd 2019 Conference Abstracts, Köln:* 194–197, DOI: 10.5281/zenodo.2596094 [letzter Zugriff 18. Dezember 2019].

Hollmer, Heide / Meier, Albert (eds.) (2001): *Dramenlexikon des 18. Jahrhunderts*. München: C. H. Beck.

Karsdorp, Folgert / Kestemont, Mike / Schöch, Christof / Bosch, Antal van den (2015): *"The Love Equation: Computational Modeling of Romantic Relationships in French Classical Drama"*, in: Mark A. Finlayson, Ben Miller, Antonio Lieto, und Remi Ronfard, (eds.), *6th Workshop on Computational Models of Narrative (CMN 2015)*: 98–107 DOI: 10.4230/OASIcs.CMN.2015.98.

Krautter, Benjamin / Pagel, Janis (2019): "Klassifikation von Titelfiguren in deutschsprachigen Dramen und Evaluation am Beispiel von Lessings *Emilia Galotti*", in: *DHd 2019 Conference Abstracts,* Köln: DOI: 10.18419/opus-10365 [letzter Zugriff 18. Dezember 2019].

Marcus, Solomon (1973 [1970]): *Mathematische Poetik*. Frankfurt a.M.: Athenäum.

Moretti, Franco (2011): *"Network Theory, Plot Analysis"*, in: *Pamphlets of the Stanford Literary Lab 2*: 2–11 https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 18. Dezember 2019].

Naumann, Karin (2007): *"Manual for the annotation of in-document referential relations"*, in: *Technical report*, Seminar für Sprachwissenschaft, Abt. Computerlinguistik Universität Tübingen.

Nalisnick, Eric T. / Baird, Henry S. (2013): *"Character-to-character sentiment analysis in Shakespeare's plays"*, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: 479–483 http://www.aclweb.org/anthology/P13-2085 [letzter Zugriff 18. Dezember 2019].

Pfister, Manfred (2001): *Das Drama. Theorie und Analyse*. München: W. Fink.

Piper, Andrew / Algee-Hewitt, Mark / Sinha, Koustuv / Ruths, Derek / Vala, Hardik (2017): *"Studying Literary Characters and Character Networks"*, in: *DH 2017 Conference Abstracts*, Montréal: https://dh2017.adho.org/abstracts/103/103.pdf [letzter Zugriff 18. Dezember 2019].

Sørensen, Bengt Algot (1984): *Herrschaft und Zärtlichkeit der Patriarchalismus und das Drama im 18. Jahrhundert*. München: C. H. Beck.

Schopenhauer, Arthur (1977): *Die Welt als Wille und Vorstellung*, in: Hübscher, Arthur (eds.), Zürich.

Trilcke, Peer / Fischer, Frank / Kampkaspar, Dario (2015): *"Digital Network Analysis of Dramatic Texts,* in: *DH 2015 Conference Abstracts*, Sydney.

Vogel, Juliane (2012): *"Aus dem Takt: Auftrittsstrukturen in Schillers Don Carlos"*, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 86/4: 532–546.

Willand, Marcus / Reiter, Nils (2017): *"Geschlecht und Gattung. Digitale Analysen von Kleists Familie Schroffenstein"*, in: Allerkamp, Ana / Blamberger, Günter / Breuer, Ingo / Gribnitz, Barbara / Lund, Hannah Lotte / Roussel, Martin (eds.): *Kleist-Jahrbuch*. Stuttgart: J. B. Metzler 177–195.

Positivismus der geistigen Gegenstände: Carnap und die Digital Humanities

Heßbrüggen-Walter, Stefan
early.modern.thought.online@gmail.com
HSE University, Russland

Kritiker der digitalen Geisteswissenschaften machen der Disziplin gerne den Vorwurf des ‚Positivismus'. Hier kann man drei Tendenzen unterscheiden: (1) Die digitalen Geisteswissenschaften verkennen die grundsätzliche Unangemessenheit quantitativer Argumente in der Auseinandersetzung mit Kunstwerken. (2) Sie ergeben sich ohne Widerstand der Unterordnung kritischer Wissenschaft unter die technokratische Verwertungslogik des Kapitalismus. (3) Sie sind geprägt von irregeleitetem Szientismus, der den ‚formalen Objekten', die unsere Kultur ausmachen, nicht gerecht werden kann (so das Referat in Eyers 2013, n. p.). Die letzten beiden Argumentationsstrategien lassen sich auch unter dem Begriff der ‚technopositivistischen Rationalität' subsumieren, der die digitalen Geisteswissenschaften verpflichtet sind (Bishop 2018, 126). Kritik am Positivsmus gibt es jedoch nicht nur von Gegnern, sondern auch von Vertretern der digitalen Geisteswissenschaften selbst. Rein quantitative, mechanische, reduktionistische und am Buchstaben klebende Verfahren müssen als positivistisch zurückgewiesen werden (Drucker 2012, 86). Statt positivistischer Ansätze bei der Sammlung und Analyse von Daten müsse ‚reflexiv' und unter Einbeziehung von ‚Theorie' vorgegangen werden (Neilson et al. 2018, 5).

Doch niemand erklärt, was genau gegen ein positivistisches Verständnis digitaler Forschungspraxen in den Geisteswissenschaften spräche. Bei genauerer Betrachtung kann gerade ein solches positivistisches Selbstbild der Theoriebildung in den digitalen Geisteswissenschaften neue Spielräume eröffnen und zugleich den längst fälligen Dialog zwischen DH und der Wissenschaftsphilosophie weiter vorantreiben. Ich werde mich im folgenden auf einen Text eines Vertreters des sogenannten ‚logischen Positivsmus' der ersten Hälfte des 20. Jahrhunderts beschränken, Rudolf Carnaps ‚Der logische Aufbau der Welt' (1928); die internen Meinungsverschiedenheiten der logischen Positivisten (Creath 2017, passim), die zum Teil auch die hier zu behandelnden Fragen berühren, müssen zunächst außer acht bleiben.

Die folgende Analyse wird sich auf zwei Aspekte der DH beschränken: ihren Gegenstand und ihre Methodik. Zu zeigen ist, dass digitale Objekte im Sinne der DH zugleich Gegenstände der Geisteswissenschaft im Sinne Carnaps sind. Die Methode der DH ist die rationale Rekonstruktion geisteswissenschaftlicher Begründungsweisen im Sinne Carnaps, die wiederum darauf beruht, dass wissenschaftliche Aussagen der digitalen Geisteswissenschaften in einem noch genauer zu bestimmenden Sinne formale Aussagen sind (nämlich auf einer bestimmten Art und Weise der Formalisierung beruhen).

Carnap und die Gegenstände der DH

Carnap erkennt die Existenz ‚geistiger Gegenstände' explizit an und weist ausdrücklich darauf hin, dass sie zum Gegenstandsgebiet der Geisteswissenschaften zu zählen sind (Carnap 1928, §23, 29). Die Selbständigkeit dieser Gegenstände sei gerade von Philosophen des 19. Jahrhunderts nicht ausreichend gewürdigt worden (Carnap 1928, §23, 30). Geistige Gegenstände sind nur auf Umwegen empirisch erfahrbar (Carnap 1928, §24, 30). Ihre Existenz setzt zwar mindestens einen Träger voraus. Jedoch können sie persistieren, auch wenn ihre Träger wechseln, und sind auch existent, wenn sie sich nicht in äußerem Verhalten ‚manifestieren' (Carnap 1928, §24, 31). Paradigmatische Beispiele solcher Gegenstände sind für Carnap Staaten oder Handlungskonventionen ("Sitten" wie etwa das Lüften des Hutes als Gruß, Carnap 1928, §23, 30). Während die Manifestation geistiger Gegenstände in erster Linie für die Sozialwissenschaften von Belang sein dürfte, ist die zweite Art und Weise, wie uns solche Gegenstände zugänglich sind, für die digitalen Geisteswissenschaften von unmittelbarem Interesse. Die Rede ist von der ‚Dokumentation': „Als Dokumentationen eines geistigen Gegenstandes bezeichnen wir dauernde physische Gebilde, in denen das geistige Leben gewissermaßen erstarrt ist, Produkte, dingliche Zeugen und Dokumente des Geistigen." (Carnap 1928, §24, 31). Sie stellen für die Geisteswissenschaften die hauptsächlichen Erkenntnisquellen dar, weil "[…] die Erforschung nicht mehr bestehender geistiger Gegenstände (und diese machen ja den größeren Teil des Gebietes aus) fast ausschließlich auf Rückschlüssen aus Dokumentationen beruht, nämlich aus schriftlichen Aufzeichnungen, Abbildungen, gebauten oder geformten Dingen oder dergl." (Carnap 1928, §24, 31f).

Carnaps Begriff des geistigen Gegenstandes ist nahezu deckungsgleich mit der in CIDOC CRM kodifizierten Klasse des ‚begrifflichen Objekts' (conceptual object, CIDOC CRM SIG 2015). Begriffliche Objekte sind wie geistige Gegenstände im Sinne Carnaps immateriell und deswegen nicht direkt erfahrbar. Sie benötigen einen Träger, entweder durch Manifestation im menschlichen Geist, oder durch Dokumentation in physischen Gegenständen. Sie sind aber von der Existenz bestimmter individueller Träger unabhängig, sofern weitere geistige oder materielle Träger desselben Begriffsgegenstandes existieren.

Geistige Gegenstände, die uns allein durch die Existenz von Dokumentationen im Carnapschen Sinne zugänglich sind, werden im Rahmen von CIDOC CRM als Informationsobjekte bezeichnet. Ein Informationsobjekt ist ein immaterieller, d. h. nicht direkt erfahrbarer Gegenstand, der eine objektiv erkennbare Struktur aufweist und als Einheit dokumentiert ist. Informationsobjekte sind somit geistige Gegenstände im Sinne Carnaps, die für uns nur durch ihre Dokumentation und nicht durch Manifestation zugänglich sind (CIDOC CRM SIG 2015a). Die digitale Provenienzontologie CRMdig geht noch einen Schritt weiter und spezifiziert digitale Objekte als diejenigen CIDOC CRM Informationsobjekte, die als Mengen von Bit-Sequenzen repräsentiert werden können (Dürr et al. 2016, 6). Sowohl digitale Objekte wie Informationsobjekte insgesamt sind begriffliche Objekte im Sinne von CIDOC CRM und somit geistige Gegenstände im Sinne Carnaps. Der Gegenstandsbegriff der DH, so wie er in grundlegenden Ontologien charakte-

risiert wird, ist also ‚positivistisch', sofern Carnap als Positivist im Sinne der eingangs dargelegten Kritik gelten soll.

Carnap nimmt weiter an, dass alle Gegenstände der Wissenschaften auf sogenannte ‚Grundgegenstände' zurückgeführt werden können und dass alle Aussagen der Wissenschaften in Aussagen über solche ‚Grundgegenstände' übersetzbar seien. In diesem Sinne müssen alle Aussagen der Wissenschaft ihrer „logischen Bedeutung nach […] von nur einem Gebiet handeln" (Carnap 1928, §41, 56). Carnap weist jedoch auch darauf hin, dass diese Umformungen in den Einzelwissenschaften nicht immer ausdrücklich vorgenommen werden. „Der logischen Form ihrer Aussagen nach hat es die Wissenschaft daher mit vielen selbständigen Gegenstandsarten zu tun." (Carnap 1928, §41, 56).

Für die ‚digital humanities' kann aber jedenfalls festgehalten werden, dass es sich bei ihren ‚primären geistigen Gegenständen' um digitale Objekte im Sinne von CRMdig handelt, also um als Objekte durch Konvention individuierte Bitströme. Hierin besteht der kennzeichnende Unterschied zu Geisteswissenschaften in Carnaps Sinne, deren primäre geistige Gegenstände durch Manifestationen konstituiert werden (Carnap 1928, §151, 201). Was nicht als digitales Objekt vorliegt, kann nicht zum Gegenstand der DH gerechnet werden. Ob diese primären geistigen Gegenstände der DH in einer projektierten Einheitswissenschaft noch weiter auf andere rückführbar wären, kann für unsere Zwecke dahingestellt bleiben.

Carnap und die Methode der DH

Der methodische Eigensinn der digitalen Geisteswissenschaften im Vergleich mit der herkömmlichen Geisteswissenschaft ist jedoch nicht auf den Primat des digitalen Objekts beschränkt. Dies erhellt aus dem Vergleich mit dem Verständnis der mathematischen Naturwissenschaft, wie es bei Carnap vorliegt. Die physikalische Welt entsteht durch Messung bzw. die Zuschreibung von ‚physikalischen Zustandsgrößen'" (Carnap 1928, §136, 180), denn: „Die Notwendigkeit der Konstitution der physikalischen Welt beruht […] auf dem Umstand, daß nur diese, nicht aber die Wahrnehmungswelt, die Möglichkeit eindeutiger, widerspruchsfreier Intersubjektivierung gibt." (Carnap 1928, §136, 180). Und ohne ‚Intersubjektivierung' gibt es keine Wissenschaft: „Die intersubjektive Welt […] bildet das eigentliche Gegenstandsgebiet der Wissenschaft." (Carnap 1928, §149, 200) Subjektive Aussagen können dann miteinbezogen werden, wenn sie ausdrücklich als solche ausgewiesen werden (Carnap 1928, §149, 200).

Digitale Objekte sind allerdings nicht bloß Quanta, sondern Träger von Informationen. Deswegen wird in den digitalen Geisteswissenschaften nicht, wie in den mathematischen Naturwissenschaften, quantifiziert, sondern formalisiert. Der Vieldeutigkeit dieses Begriffs ist die einschlägige Theoriebildung bislang aus dem Weg gegangen, was weniger verwundert, wenn man sich vergegenwärtigt, dass die philosophiehistorische Analyse nicht weniger als acht unterschiedliche Bedeutungen des Begriffs des Formalen alleine für die Logik ausweist (Dutilh Novaes 2011, 304). Eine dieser acht Bedeutungen ist jedoch hier unmittelbar einschlägig, nämlich die des Formalen als Berechenbarkeit: sowohl die wohlgeformten Ausdrücke eines Kalküls wie auch deren Transformationen lassen sich durch die mechanische Anwendung eindeutiger Regeln generieren (Dutilh Novaes 2011, 323). Formalisierung in diesem Sinne bedeutet dann, Objekte für eine Bearbeitung mithilfe solcher eindeutiger Regeln tauglich zu machen. In der Terminologie von CIDOC CRM besteht Formalisierung also darin, Informationsobjekte zu digitalen Objekten zu transformieren bzw., sehr viel einfacher gesagt, sie zu digitalisieren (mutatis mutandis dürfte Ähnliches für Bild- oder Klangobjekte gelten, auch wenn sie nicht unter den Begriff des Informationsobjekts fallen). Digitale Objekte als solche sind also schon formalisiert, weil sie als Bits vorliegen, die weiteren rechnenden Transformationen zugänglich sind. Weiter unterscheiden sich die ‚epistemischen Dinge' der DH untereinander nur durch den Grad ihrer Strukturierung (Trilcke/Fischer 2018, Abschn. 3).

Damit ist offensichtlich, dass die theoretischen Aussagen der digitalen Geisteswissenschaft von sich aus schon zum Bereich der Carnapschen intersubjektiv begründeten Wissenschaft gehören, da sie idealerweise für jeden nachvollziehbar sind. Dann aber ist digitale Geisteswissenschaft die ‚rationale Nachkonstruktion' herkömmlicher geisteswissenschaftlicher Forschung. Carnap erläutert den Sinn dieser Methodik anhand eines Beispiels aus der Biologie: „Der Botaniker muß sich bei der Nachkonstruktion der Erkennung der Pflanze fragen. Was war in der erlebten Wiederkennung das eigentlich Gesehene, und was war daran die apperzeptive Verarbeitung?; aber er kann doch diese beiden im Ergebnis vereinten Komponenten nur durch Abstraktion trennen." (Carnap 1928, §100, 139) In gleicher Weise zwingen uns die Methoden der digitalen Geisteswissenschaften, das eigentlich Gegebene auf dem Wege ‚methodischer Abstraktion' von dessen ‚apperzeptiver Verarbeitung' zu trennen, intersubjektiv zu verteidigende Aussagen von solchen, die subjektiv bleiben, explizit zu unterscheiden.

Ich habe eingangs behauptet, dass ein positivistisches Verständnis der digital humanities neue Spielräume eröffnen würde. Hierfür wäre auf dem Hintergrund des Gesagten wie folgt zu argumentieren: Die digital humanities operieren mit Gegenständen, die bereits als das Resultat einer Formalisierung, nämlich der Digitalisierung, anzusehen sind. Nur in diesem Sinne formalisierte Gegenstände können nämlich weiter durch rechnende Transformationen untersucht werden. Gegenstand der DH ist also das digitale Objekt. Transformationen von digitalen Objekten sind intersubjektiv nachvollziehbar und gehören damit unfraglich in den von Carnap umrissenen Bereich der Wissenschaft, die ‚intersubjektive Welt'. Damit ist der Bereich dessen, was in den DH nach Carnap Wissenschaft sein kann, jedoch noch nicht erschöpft. Subjektive Aussagen müssen nicht aus dem Bereich der Wissenschaft entfernt werden. Sie müssen nur explizit als solche ausgewiesen werden. Gerade positivistische digitale Geisteswissenschaftler müssen also nicht ‚am digitalen Buchstaben kleben'. Sie sind allerdings gezwungen, die nicht durch Daten selbst ausgewiesenen Aussagen explizit als solche zu kennzeichnen. Diese Form der methodischen Selbstreflexion macht es möglich, sich mit anderen über den Bereich des Subjektiven streitbar auseinanderzusetzen. Hierin, so meine ich, liegt der methodologische Gewinn der digitalen Geisteswissenschaften für die Geisteswissenschaften als ganze, wenn wir digitale Geisteswissenschaft im hier entwickelten Sinn als positivistische auffassen.

Bibliographie

Bishop, Claire (2018): „Against Digital Art History", in: International Journal for Digital Art History, no. 3 (July). https://doi.org/10.11588/dah.2018.3.49915. [Letzter Zugriff: 22. September 2019].

Carnap, Rudolf (1928): Der logische Aufbau der Welt. Berlin-Schlachtensee: Weltkreis-Verlag.

CIDOM CRM SIG (2015): E28 Conceptual Object | CIDOC CRM'. http://www.cidoc-crm.org/entity/e28-conceptual-object/version-6.2. [Letzter Zugriff: 22. September 2019]-

CIDOC CRM SIG (2015a): E73 Information Object | CIDOC CRM'. http://www.cidoc-crm.org/Entity/e73-information-object/version-6.2. [Letzter Zugriff: 22. September 2019].

Creath, Richard (2017): 'Logical Empiricism', in: The Stanford Encyclopedia of Philosophy, hg. v. Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/logical-empiricism/ [Letzter Zugriff: 22. September 2019].

Doerr, Martin, et al. (2016): Definition of the CRMdig: An Extension of CIDOC-CRM to support provenance metadata, Version 3.2.1. http://www.cidoc-crm.org/crmdig/sites/default/files/CRMdig_v3.2.1.pdf [Letzter Zugriff: 22. September 2019].

Drucker, Johanna (2012): 'Humanistic Theory and Digital Scholarship', in: Gold, Matthew K., (ed). Debates in the Digital Humanities. Minneapolis, Minn.: Univ. of Minnesota Press, 85-95.

Dutilh Novaes, Catarina (2011): 'The Different Ways in which Logic is (said to be) Formal', in: History and Philosophy of Logic, 32:4, 303-332.

Eyers, Tom (2013): 'The Perils of the "Digital Humanities": New Positivisms and the Fate of Literary Theory'. Postmodern Culture 23 (2). https://doi.org/10.1353/pmc.2013.0038. [Letzter Zugriff: 22. September 2019].

Levenberg, Lewis / Neilson, Tai (2018): Research Methods for the Digital Humanities. Cham, Switzerland: Palgrave Macmillan.

Trilcke, Peer / Fischer, Frank (2018): ‚Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen.' In: Huber, Martin/Krämer, Sybille (eds.), Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3). https://doi.org/10.17175/sb003_003[Letzter Zugriff: 22. September 2019].

Public Humanities Tools: Der Bedarf an niederschwelligen Services

Hermes, Jürgen
hermesj@uni-koeln.de
Universität zu Köln, Deutschland

Klinke, Harald
h.klinke@lmu.de
LMU, München

Demmer, Dennis
Dennis.Demmer@uni-koeln.de
Universität zu Köln, Deutschland

Spätestens mit der Herausbildung des Social Web (auch Web 2.0) seit knapp 15 Jahren, das nicht nur für die Verteilung von Information, sondern tatsächlich auch zur Mitgestaltung von Inhalten genutzt werden kann, hat das Internet die gesellschaftliche Kommunikationskultur (jedenfalls die derjenigen, die über verlässlichen Zugang verfügen und diesen nutzen) entscheidend gewandelt. Mit ResearchGate, Academia.edu, Mendeley und als neue, explizit nicht-kommerzielle Variante HCommons entstanden eine Reihe sozialer Medien spezifisch für den wissenschaftlichen Bereich, über die Forschungsergebnisse ausgetauscht und bewertet werden können und mit denen v.a. der Kontakt zu Kolleg|inn|en aufgenommen werden kann (Sugimoto et al. 2016). Jenseits dieser spezialisierten sozialen Medien nutzen Wissenschaftler|innen auch die allgemeinen Plattformen wie Facebook und Twitter, letzteres vor allem, um wissenschaftlichen Diskussionen zu folgen, Forschung zu kommentieren und auf eigene Veröffentlichungen - von Ergebnissen, jedoch auch von Daten und Software - aufmerksam zu machen (vgl. van Noorden 2014). Über die allgemein gebräuchlichen sozialen Medien ist es möglich, auch Laien zu erreichen, sei es, um die eigene Reichweite zu erhöhen oder um neue Nutzerkreise zu gewinnen, die mitunter sogar am Forschungsprozess partizipieren können. Entsprechende Programme wie *public engagement* oder *Citizen* bzw. *Crowd Science* sind institutionell erwünscht (vgl. Deutsche Akademie der Technikwissenschaften et al. 2014) und innerhalb der Wissenschaften durchaus verbreitet (vgl. Franzoni & Sauermann 2014).

Die öffentliche Publikation von Forschungsdaten

Forschung – nicht zuletzt die in den Geisteswissenschaften – generiert große Mengen an Daten, Information und Wissen, die für (Teil)Öffentlichkeiten interessant und relevant sein können. Nun ist die Publikation von Forschungsdaten – zusätzlich zu den bisher gebräuchlichen Publikationsmedien – zwar weithin erwünscht (siehe RFII 2016), zur Zeit allerdings alles andere als weitreichend umgesetzt. Dafür können sehr viele unterschiedliche Ursachen ausgemacht werden (vgl. Kaden 2018). Auf der anderen Seite bieten soziale Medien, hier vor allem Twitter, die Möglichkeit, granulare Informationshäppchen fein dosiert in den Timelines von Nutzer|inne|n erscheinen zu lassen und über diesen Weg deren Aufmerksamkeit zu gewinnen. Die Nutzung von privatwirtschaftlichen Plattformen, die vorwiegend monetäre Interessen verfolgen, für die Wissenschaftskommunikation ist nicht unproblematisch. Momentan existieren allerdings schlicht keine nichtkommerziellen Alternativen Plattformen, über die man auf relativ simple Weise ein ähnlich großes Publikum erreichen könnte.

Ein Twitterprojekt, das weitreichende Beachtung fand bis hin zu einem Artikel in der New York Times , war das Projekt @9nov38 - heute vor 75 Jahren, in dem fünf Historiker|innen die zeitliche Dimension in die Erzählung von Ereignissen der Reichspogromnacht über Twitter mit einbezogen. Nun ist die manuelle Erstellung einzelner Tweets sehr aufwendig und für

größere Datensätze eigentlich nicht ohne weiteres zu leisten. Doch im Grunde liegen die Daten, die für derartige Projekte gesammelt wurden, im Normalfall bereits in einem strukturierten Format vor, etwa in einer Datenbank oder als Spreadsheet. Auf dieser Grundlage wurde nach einem Austausch mit den am @9Nov38-Projekt beteiligten Historiker|innen auf dem Histocamp 2015 der Webservice *autoChirp* entwickelt, zunächst im Rahmen eines Projektseminars, seither weiter betreut durch das *Institut für Digital Humanities* (IDH) in Köln (Hermes et al. 2017). autoChirp ist ein Webservice, der nicht auf eine spezifische Anwendung hin entwickelt wurde, sondern eine Plattform bietet, um diversen, u.a. historischen Projekten einen niedrigschwelligen Zugang zu für sie hilfreicher Technologie zu ermöglichen. In dem bewusst einfach gehaltenen Webinterface können strukturierte Daten hochgeladen werden, um sie automatisiert auf spezifizierte Zeitpunkte zu schedulen und zu veröffentlichen. Das erste Projekt, das autoChirp nutzte, war @NRWHistory, bei dem in einem Projektseminar von Düsseldorfer Historiker|innen die Entstehung des Landes NRW um 70 Jahre zeitversetzt nacherzählt wurde (siehe http://nrw-history.de/). Kurz darauf wurde über autoChirp mit @TiwoliChirp ein weiterer Veröffentlichungskanal für bereits über eine Smartphone-App veröffentlichten Forschungsergebnisse der Literaturwissenschaft eingesetzt.

Inzwischen greifen eine ganze Reihe von Projekten, die regelmäßige Tweets publizieren, auf autoChirp zurück. Das mit mehr als 4000 Followern mit Abstand reichweitenstärkste davon ist @Die_Reklame, mit dem Akteure aus dem Projekt @9Nov38 bemerkenswerte historische Werbeanzeigen twittern. Von Interesse sind diese, weil gerade Werbung extrem gegenwartsbezogen ist, was einen Einblick in die entsprechende Zeit der ursprünglichen Publikation gibt (vgl. Hoffmann 2018). Ein weiteres Projekt mit historischem Bezug ist *Verbrannte Orte* (@pictureXnet), das die Orte von Bücherverbrennungen im Dritten Reich auf einer Karte sammelt (siehe https://verbrannte-orte.de/) und diese an den entsprechenden Jahrestagen der Verbrennungen vertwittert. Die Twitter-Plattform hilft hier dabei, Aufmerksamkeit zu generieren und auch Daten zu den Ereignissen zu sammeln. Einen sehr ähnlichen Ansatz verfolgt das Projekt @gedenkplaetze.

2019 jährte sich zum 250. Mal des Geburtstag von Alexander von Humboldt. In diesem Jahr von besonderem Interesse war daher seine Chronik (siehe https://edition-humboldt.de/chronologie/), die von der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) herausgegeben wird und inzwischen auch über autoChirp an Twitter angebunden wurde (Hermes 2017). Bemerkenswert hier ist, dass die Chronik unter den Twitter-Account @AvHChrono jahrestagsaktuell verfolgt werden kann, was von knapp 200 Leser|innen in Anspruch genommen wird. Diese tagesaktuelle Konsultation der Daten hat auch schon zur Feststellung von Fehlern geführt, die an die BBAW rückgemeldet und anschließend korrigiert wurden. Auch dieses Beispiel zeigt, dass Social Media keine Kommunikation auf der Einbahnstraße sein muss.

Die Perspektive der Kunstgeschichte

Zwei der neuesten autoChirp-nutzenden Projekte kommen aus dem Bereich der Kunstgeschichte, einer bildbasierten Wissenschaft, deren Grundlage historische visuelle Objekte sind. Daher bedeutet die Einführung digitaler Methoden in das Fach vor allem die Entwicklung von Analyseprozessen, die sich auf Bild- und Metadaten beziehen (Klinke 2018). Diese werden nicht nur in der Forschung erzeugt, sondern kommen bisher vor allem aus den Sammlungsinstitutionen (GLAM).

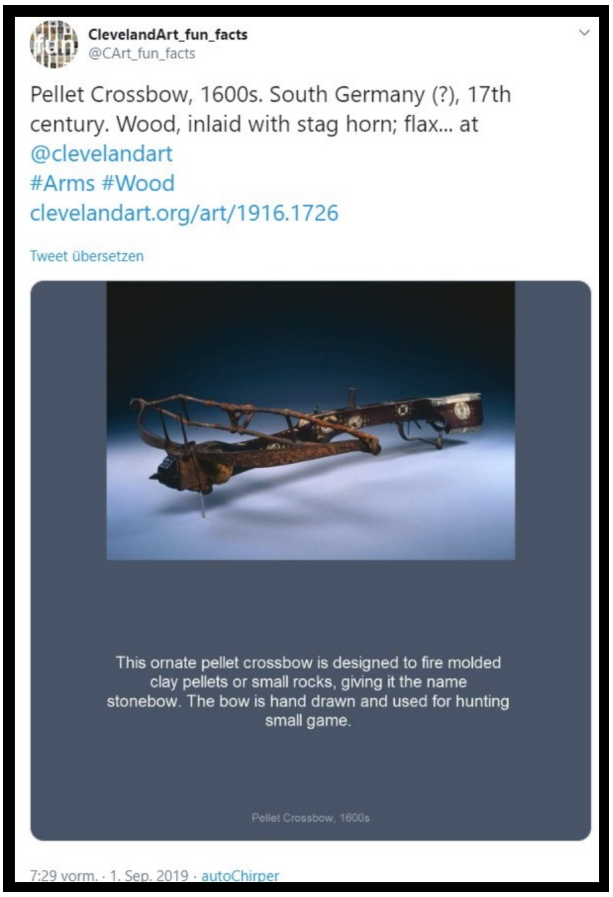


Abbildung 1: Ein Tweet aus dem Fundus des ClevelandFunFacts-Twitterbots

Museen sind einer umfangreichen Transformation unterworfen, in der sie ihre Aufgaben unter dem Vorzeichen der Digitalisierung, Social Media und Virtual Reality neu definieren müssen (Kohle 2019). So eröffnet die Publikation der Sammlungsdaten als Open Data neue Möglichkeiten, die kulturellen Artefakte in neue, zeitgenössische Zusammenhänge zu bringen, in denen sie neue Bedeutungszuschreibungen erhalten können. Durch die Verwendung von autoChirp können offene Sammlungsdaten und globale Öffentlichkeit durch das visuelle Medium Twitter zusammengebracht werden. Auch hier erlaubt Twitter nicht nur die Kommunikation in eine Richtung, sondern auch die Partizipation des Publikums durch Kommentare, Retweets und das Einbinden in neue Kontexte.

Zwei Beispiele aus dem Jahr 2019 machen dies deutlich: Der Tweetbot @cart_fun_facts baut auf der Open Data-Strategie des Cleveland Museum of Art auf. Das 1916 gegründete Museum ist eines der umfassendsten Kunstmuseen der Welt, das am 23. Januar 2019 bekannt gegeben hat, dass es sich ab sofort als eine Open-Access-Institution betrachtet, die die Bezeichnung Creative Commons Zero (CC0) für hochauflösende Bilder und Daten im Zusammenhang mit ihrer Sammlung verwendet (siehe http://www.clevelandart.org/open-access). Die Öffentlichkeit hat damit jetzt die Möglichkeit, Bilder von mehr

als 30.000 gemeinfreien Kunstwerken zu kommerziellen und nichtkommerziellen Zwecke zu teilen, neu zu mischen und wiederzuverwenden. Der von Harald Klinke (LMU München) entwickelte Tweetbot verwendet die in der Datenbank befindlichen “Fun Facts”, die täglich auf Flashcards zusammen mit den Abbildungen der Kunstwerke im Format eines visuellen Memes über den autoChirp-Service getwittert werden (siehe Abbildung 1).

Ein weiteres Beispiel ist der auf der Digital Art History Summer School 2019 in Malaga (DAHSS) durch Studierende entwickelte Tweetbot @thyssenmlgbot. Dieser twittert die Werke des dortigen Museum Carmen Thyssen unter Zuhilfenahme von NLP-Techniken und autoChirp, wodurch die Beschreibungstexte auf relevante Topics untersucht und diese in Hashtags umgewandelt werden. Dieses Projekt hat einerseits gezeigt, wie Studierende mithilfe von digitalen Kompetenzen einer GLAM-Institution helfen können, ihre Werke einer breiteren Öffentlichkeit zu vermitteln. Andererseits, wie diese Vermittlung einen Rückkanal erhalten kann, der es dem Publikum erlaubt, auf die Werke zu reagieren (beispielsweise durch in die Tweets integrierte Frage nach der vermuteten Entstehungszeit des Werks). Auf diese Weise können auch Werke, die üblicherweise nicht in der Ausstellung gezeigt werden, sondern im Depot verbleiben, sichtbar gemacht werden. Ein Online-Tool wie autoChirp ist dafür ein Hilfsmittel, das einen niederschwelligen Zugang zu neuen Formen der digitalen Museumskommunikation ermöglicht und deshalb gerade auch in der Lehre eingesetzt werden kann.

autoChirp und autoPost

Das IDH betreibt inzwischen neben autoChirp zur automatisierten Veröffentlichung auf Twitter auch *autoPost* für analoge Aufträge für Facebook-Seiten. Beide Services basieren auf Spring, einem quelloffenen Java-Framework für Web-Anwendungen. Der Quellcode ist unter Open Source-Lizenz (*Eclipse Public Licence*) auf GitHub beziehbar (siehe https://github.com/DH-Cologne/autoChirp und https://github.com/DH-Cologne/auto-post), so dass eigene Services betrieben werden können. Das IDH stellt aber auch beide Services für alle Interessierten zur Verfügung (siehe https://autochirp.spinfo.uni-koeln.de und https://autopost.spinfo.uni-koeln.de/home). Bei der Implementation wurde vor allem auf Modularität und Erweiterbarkeit geachtet, um das Programm ohne größeren Aufwand auf weitere Social Media Plattformen, wie z.B. Instagram portieren zu können, sofern diese eine entsprechende API (Application-Programming-Interface) anbieten.

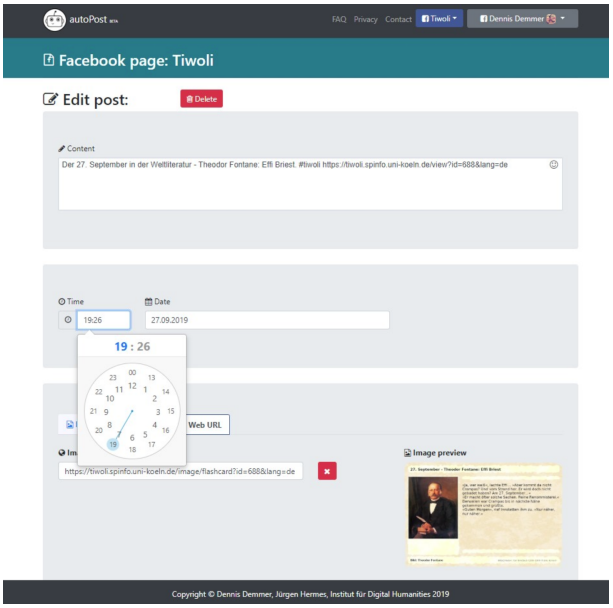


Abbildung 2: Screenshot des autoPost-Services, mit dem große Mengen von geplanten Facebook-Posts realisiert werden können (hier zum Tiwoli-Projekt).

Bei der Datenpersistenz wurde bei autoPost auf eine schwergewichtigere, aber performantere Datenbank gesetzt, da die Erfahrung mit autoChirp gezeigt hat, dass ein freier Scheduling-Service sehr gut angenommen wird und die Zahl der Datenbankeinträge dementsprechend groß werden kann. Um Nutzer|inne|n von autoChirp die Möglichkeit zu bieten, ihre Inhalte, die in autoChirp schon geplant sind, auch auf Facebook zu veröffentlichen, wurde für autoChirp eine Export-Funktion angelegt. Tweets können gruppenweise als TSV-Datei heruntergeladen und in autoPost als Facebook Posts importiert werden.

Zwischenfazit zum Nutzerzuspruch

Während autoChirp schon seit 2016 läuft und für knapp 150 Nutzer|innen-Accounts bereits über 17.500 Tweets veröffentlicht hat (weitere 10.000 Tweets sind terminiert, aktuelle Zahlen erhält man über die Statistik-Seitedes Services), startete autoPost erst im Herbst 2019. Mit *Syrian Modern History* und *Public History Weekly* konnten aber bereits zwei wissenschaftlich betreute Accounts mit kombiniert über 36.000 Facebook-Abonnent|inn|en gewonnen werden, die autoPost täglich zur Bewerbung von Archiv-Artikeln nutzen.

Die Services autoChirp und autoPost sind Beispiele, an denen sich eine der wichtigen Aufgaben für die Digital Humanities spezifizieren lässt: Die Entwicklung erfolgte, weil Wissenschaftler|innen (nicht nur) aus den Geisteswissenschaften einen Bedarf hatten, ihre Daten auf Social Media Plattformen zu teilen. Dafür benötigten sie Tools, die eine niedrige Einstiegsschwelle haben und ihnen dabei Arbeit abnehmen können, wenn sie Aspekte ihrer Forschung öffentlich sichtbar machen und Studierende sowie die interessierte Öffentlichkeit in den Forschungsprozess (hier zuvorderst: In die Datensammlung) einbinden wollen. Insofern verstehen wir die Entwicklung von autoChirp und autoPost als Hilfsmittel zur

Etablierung einer offenen, transparenten und partizipativen Wissenschaft (Open Science). Die Erfahrungen mit den hier vorgestellten Tools zeigt, dass die Methoden sowohl von den Wissenschaftler|inne|n, als auch vom Publikum angenommen werden und mithin das Potenzial haben, den Geisteswissenschaften eine größere Präsenz in der Öffentlichkeit zu ermöglichen und damit eine höhere Relevanz in der Gesellschaft zu erzielen.

Bibliographie

Deutsche Akademie der Technikwissenschaften / Union der Deutschen Akademien der Wissenschaften / Deutsche Akademie der Naturforscher Leopoldina (2014): "Zur Gestaltung der Kommunikation zwischen Wissenschaft, Öffentlichkeit und den Medien. Empfehlungen vor dem Hintergrund aktueller Entwicklungen." München: acatech – Deutsche Akademie der Technikwissenschaften e.V. / Mainz: Union der Deutschen Akademien der Wissenschaften e.V. / Halle (Saale): Deutsche Akademie der Naturforscher Leopoldina e.V. – Nationale Akademie der Wissenschaften.

Fischer, Frank / Strötgen, Jannik (2015): „When Does German Literature Take Place? – On the Analysis of Temporal Expressions in Large Corpora", in: *Proceedings of DH2015*, Sydney: Alliance of Digital Humanities Organisations.

Franzoni, Chiara / Sauermann, Henry (2014): "Crowd science: The organization of scientific research in open collaborative projects", in: *Research Policy*, Amsterdam: Elsevier Volume 43/1, 1-20.

Hermes, Jürgen (2017): „Neu: Alex von Humboldt auf Twitter!", in *TEXperimenTales*, 17/08/2017, https://texperimentales.hypotheses.org/2069. [letzter Zugriff 18.12.2019]

Hermes, Jürgen / Hoffmann, Moritz / Eide, Øyvind / Geduldig, Alena / Schildkamp, Philip (2017): „Twhistory with autoChirp" in: DHd 2017 Bern – Digitale Nachhaltigkeit. Abstractband. Bern: DHd, 277ff.

Hoffmann, Moritz (2018): „Von Funden und Schwellen: Die Reklame", https://www.moritz-hoffmann.de/2018/04/13/von-funden-und-schwellen-die-reklame/ [letzter Zugriff 18.12.2019]

Kaden, Ben (2018): „Warum Forschungsdaten nicht publiziert werden", in: *LIBREAS, Library Ideas* 33.

Klinke, Harald (2018): „Datenanalyse in der Digitalen Kunstgeschichte. Neue Methoden in Forschung und Lehre und der Einsatz des DHVLab in der Lehre", in: Harald Klinke (Hg.): *#DigiCampus. Digitale Forschung und Lehre in den Geisteswissenschaften*, München, 2018, S. 19-34.

Kohle, Hubertus (2019): *Museen digital. Eine Gedächtnisinstitution sucht den Anschluss an die Zukunft*. Heidelberg: Heidelberg University Publishing.

RfII – Rat für Informationsinfrastrukturen (2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen. URL: http://www.rfii.de/?p=1998 [letzter Zugriff 18.12.2019]

Schwarz, Ingo (2019): „Zur Alexander von Humboldt-Chronologie", in Ottmar Ette (Hg): *edition humboldt digital*, hg. v. Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. Version 5 vom 11.09.2019. URL: https://edition-humboldt.de/v5/H0000002 [letzter Zugriff 18.12.2019]

(Re-)Collecting Theatre History: Wissensdinge, Biographien, Wirkungsräume

Mertgens, Andreas
a.mertgens@googlemail.com
Universität zu Köln, Deutschland

Türkoğlu, Enes
enes.tuerkoglu@uni-koeln.de
Universität zu Köln, Deutschland

Probst, Nora
nora.probst@uni-koeln.de
Universität zu Köln, Deutschland

Fachwissenschaftliche Hintergründe

Theaterhistorische Wissensdinge sind divers und unmittelbar mit vielfältigen historisch-kulturellen Bezügen aufgeladen. Die Objekte stammen teils unmittelbar aus der Aufführungspraxis (z.B. Masken und Requisiten) und sind damit Akteure und Zeugen des Geschehens. Zum Teil stammen sie aber auch aus Produktionsprozessen (z.B. Regiebücher, Bühnenbildentwürfe und -modelle) und ermöglichen damit einen Blick in die Prozesse, Mechanismen und Materialitäten des Theaters. Andere Objekte hingegen dokumentieren das ephemere Ereignis (z.B. Fotografien oder Kritiken) und erlauben so einen Zugriff auf die performativen Spuren einer Aufführung. Und wieder andere Objekte wie Programmhefte und Theaterzettel beinhalten historisch-kulturelle Kontextinformationen. Ergänzend zu den genannten Objektarten tragen Ego-Dokumente wie Briefe und Kalender eher eine sozio-historische Bedeutung: Sie geben Hinweise auf die beteiligten Individuen sowie deren Lebensläufe und Beziehungen. Die Bedeutungsebenen und Bezüge der Objekte sind nicht statisch: Mit jeder neuen Kontextualisierung weiten sich ihre Bedeutungsräume aus. Oft befinden sich Objekte an den Schnittstellen von diversen Verwendungsmöglichkeiten – bspw. beschreiben Kritiken nicht nur, was auf der Bühne geschehen ist, sie bilden auch Publikumsreaktionen und damit einen wichtigen Teil der Rezeptionsgeschichte ab.[1]

Theateraufführungen lassen sich verstehen als das Ergebnis einer intensiven und oft monatelangen Zusammenarbeit unterschiedlichster Professionen: Akteur_innen[2] aus den Tätigkeitsbereichen Regie, Darstellung, Bühnen- und Kostümbild, Bühnentechnik und vielen weiteren bilden ein komplexes Netzwerk, das für einen begrenzten Zeitraum besteht und im organisierten Zusammenspiel eine Inszenierungsidee auf der Theaterbühne verwirklicht. Theaterhistorisch relevante Forschungsobjekte beinhalten implizite und explizite Spuren die-

ser Verwirklichung. Der Wirkungsraum der Objekte erstreckt sich dabei sowohl auf Ereignisse als auch auf die Akteur_innen, die an diesen Ereignissen beteiligt waren. Die unterschiedlichen Ereignisse, zwischen denen Beziehungen und Interdependenzen bestehen, lassen sich als das Ergebnis von komplexen Interaktionen zwischen Akteur_innen beschreiben und können somit als Modelle für personenbezogene Interaktionsnetzwerke dienen. Um diese Netzwerke erfassen zu können, müssen nicht nur die relevanten Objekte mit inhaltsreichen Daten erschlossen werden, sondern auch die Relationen zwischen diesen Daten modelliert werden.

Diesen vergänglichen Netzwerken von Ereignissen und Akteur_innen im 20. Jahrhundert spürt die Theaterwissenschaftliche Sammlung (TWS) in Zusammenarbeit mit Cologne Center for eHumanities (CCeH) nach. Das Forschungsprojekt (Re-)Collecting Theatre History zielt auf die theaterwissenschaftliche Resystematisierung personenbezogener Bestände in den theaterhistorischen Sammlungen der Universität zu Köln und der FU Berlin sowie ergänzend der Theatermuseen Düsseldorf und München ab. Die scheinbare ‚Zufälligkeit' von Lebenswegen, die sich nicht den Ordnungsbegriffen der politischen oder der Kunstgeschichte unterordnet, soll zum Ausgangspunkt genommen werden, um die Netzwerke ebenso zu erhellen wie die Frage nach personellen und ästhetischen Kontinuitäten im Theaterbetrieb.

Die im Projekt entwickelte Plattform eröffnet Querverbindungen und Vergleichsmöglichkeiten der Bestände in den wichtigsten universitären Theatersammlungen und öffentlichen Theatermuseen in Deutschland – als solche ist sie offen und auch für zukünftige Projekte erweiterbar –, und schafft dadurch ein umfangreiches Forschungsnetzwerk.

Metadaten: Modell und Erfassung

Aus methodischer Sicht lässt sich das Projekt auf den ersten Blick zwar noch als klassische Nachlassdigitalisierung und Erschließung beschreiben, doch schon auf den zweiten Blick wird deutlich, dass die einfache Erfassung von objektbezogenen Metadaten keine ausreichende Datengrundlage für die intendierte Aufdeckung und Erforschung der beschriebenen komplexen Netzwerk- und Interaktionsstrukturen bieten würde. Die Gesamtheit der Objekte und, in Konsequenz daraus, die Objektdatenbank bildet zwar das Rückgrat des Projektes, sie muss aber den gesamten sozio-historischen Wirkungsraum und Kontext der Objekte erfassen können und darf nicht auf die reine materielle Beschreibung beschränkt sein.

Für das Projekt war klar, dass ein Metadaten-Standard wie z.b. Dublin-Core hier viel zu kurz greifen würde und nicht die nötige Spezifität und semantische Tiefe wiedergeben kann. Stattdessen werden die Daten im LIDO (Lightweight Information Describing Objects) Standard erfasst. Dieser ermöglicht eine detaillierte Objektbeschreibung, bietet darüber hinaus aber auch Strukturen für die Beschreibung von "events" und "actors". Als weit verbreiteter Standard in der Museums- und Sammlungswelt bietet LIDO zum einen eine Basis für konsistente und vergleichbare Datenerhebung, zum anderen genügend Flexibilität und Spielraum innerhalb der Strukturen um die sozio-kulturellen Kontexte der Objekte beschreiben zu können. LIDO ist das Produkt einer CIDOC-Arbeitsgruppe und ist mit den Ansätzen von CIDOC Conceptual Reference Model (CRM) zu großen Teilen konform. Insbesondere bildet die Event-Actor-Struktur der CRM Ontologie einen Kernaspekt des LIDO Standards, der in diesem Projekt von besonderem Interesse und Nutzen war. Durch diese eventbasierte Modellierung können die Daten auch für Graphdatenbank- und Netzwerkansätze verwendet werden, ohne das Objekt, dessen Beschreibung natürlich der Kern eines Nachlass-Projektes bleiben muss, als Informationsträger aus dem Fokus zu verlieren. Ein LIDO Dokument kann auch jederzeit in CRM übersetzt werden, um die Daten für andere Forschungsziele zu verwenden.

Neben den unmittelbar objektbezogenen Metadaten enthält jeder Datensatz Daten zu drei vordefinierten Ereignissen: "Herstellung", "Inszenierung" und "Erwerb", welche die Biographie der Objekte rudimentär nachzeichnen. Zusätzlich können weitere nicht vordefinierte Ereignisse aufgenommen werden. Ihnen können jeweils beliebig viele Akteure zugeordnet werden. Mit diesen Strukturen können dann komplexe Aussagen erfasst werden wie z.B. "Die Maske wurde von einer bekannten Requisitenwerkstatt hergestellt" oder "Schauspieler X war der Inszenierung Y beteiligt, die auf diesem Programmzettel beschrieben wird". Es werden also teilweise bereits in den objektbezogenen Daten Beziehungen zwischen Akteuren, Ereignissen und Objekten explizit modelliert und beschrieben.

Interoperabilität und Nachhaltigkeit haben für solch ein Erschließungsprojekt maßgebliche Bedeutung. Um beides zu gewährleisten, spielen Normdaten und kontrollierte Vokabulare eine wichtige Rolle. Insbesondere personenbezogene Daten profitieren von Verknüpfungen zu externen Datensätzen. Daher werden die Personendatensätze mit GND-Nummern verknüpft (sofern vorhanden). Außerdem werden die Objekttypen unter Zuhilfenahme des Art and Architecture Thesaurus (AAT) erfasst, da sich der heterogene theaterhistorische Objektbestand mit der polyhierarchischen und multilingualen Struktur des AAT adäquat erfassen lässt. Für weitere Datenfelder und -typen, für die keine passende Standards existieren (z.B. Funktionen der Personen oder erweiterte Eventtypen), werden Theaterwissenschafts-spezifische Vokabulare entwickelt, die möglicherweise hilfreich sein können für die zukünftige Weiterentwicklung der LIDO-Terminologie.[3]

An dieser Stelle ist es wichtig anzumerken, dass LIDO als umfangreiches Datenaustauschformat entwickelt worden ist. Um LIDO für das Projekt effektiv nutzen zu können, wurde ein LIDO-Sub-Schema für die projektspezifischen Bedürfnisse entwickelt, das im Laufe des Projektes erweitert und an die Bedürfnisse des Projekts angepasst wurde. Die Eingabe der Daten findet über den LIDO-Maker statt, der auf dem Metadateneditor CMDI-Maker basiert und entsprechend weiterentwickelt wurde.[4]

Von Objektdaten zu Akteur- und Inszenierungsdaten

Verwandte Projekte wie z.B. IbsenStage oder AusStage[5] stellen die breite Erfassung von Ereignissen, Akteuren und anderen Entitäten in den Vordergrund, und erfassen Objektbezüge nur mit rudimentäre Metadatensets. Auch im institutionsübergreifenden Nachweis- und Rechercheportal *performing-arts.eu*[6] (FID Darstellende Kunst) werden nur die Kerndaten zu Objekten präsentiert. Der Ausgangspunkt dieses Projektes ist dagegen die detaillierte Erschließung von Nachlassobjekten. In der Objektdatenbank werden, wie beschrieben, bereits Informationen zu den Beziehungen von Objek-

ten zu Ereignissen und Akteur_innen explizit erfasst. Dagegen sind die Beziehungen zwischen Akteur_innen oder zwischen Ereignissen bisher nur impliziert. Um diese erforschbar und auch referenzierbar zu machen, wurden mithilfe von XSLT zwei zusätzliche Datenbanken aus der im ersten Schritt erstellten LIDO-Objektdatenbank extrapoliert. Um die komplexen XSL-Transformationen kontrollieren zu können, wurden Prinzipien der Konversionspipelines umgesetzt, so dass der Prozess der Datenkuratierung offen gelegt wird. (Barabucci 2018).

Die Personendaten orientieren sich hierbei an der Struktur des TEI:person-Modells[7], wurden aber mit einigen projektspezifischen Elementen und Attributen ergänzt. Analog dazu wurden auch die Inszenierungsdaten explizit modelliert, indem die Informationen zu einer spezifischen Inszenierung, die aus unterschiedlichen Objekten stammen, kohärent zusammengefügt wurden. In diesem Schritt wurden eindeutige Bezeichner (UUID) zu den Personen- und Inszenierungsdaten zugeordnet, so dass diese nun schon bei der Erfassung von neuen Objektdatensätzen referenziert werden können. Auch die Referenzierungen der Daten untereinander und die entsprechende Abfragen zu diesen Referenzen werden ermöglicht, um nah an der Idee der Netzwerke bleiben zu können.

Innerhalb eines Akteur_in-Datensatzes sind dann komplexe personenbezogene Informationen wie Wirkungsorte und damit auch Karriereverläufe unmittelbar abrufbar. Wenn eine Person in einem Objektdatensatz als Hersteller_in des Objektes vorkommt – bspw. als Bühnenbildner_in eines Bühnenbildmodells, so wird diese Information einerseits als Objektbezug verarbeitet, andererseits landet sie im Personendatensatz als Berufs- oder Tätigkeitsbeschreibung mit zeitlichen und örtlichen Angaben. Außerdem wird die Verknüpfung mit der entsprechenden Inszenierung gewährleistet, die mit dem jeweiligen Objekt in Beziehung steht. Somit ist die Person nicht nur als Hersteller_in eines Objektes signifikant, sondern auch als Akteur_in der Inszenierungen.

Ausblick und Fazit

Als letzter Schritt von der reinen Objektdatenerfassung hin zu einer Plattform, die Akteur_innen- und Inszenierungdatenbank vernetzt, wurden im Projekt experimentell Möglichkeiten der weiteren Datenanreicherung angedacht und bereits in Teilen umgesetzt.

So wurden zum Beispiel, um weitere Informationen über die biographische Hintergründe der Akteur_innen zu erhalten, die biographischen Artikel (ADB und NDB) von der *Deutschen Biographie* mithilfe der dort bereitgestellten API für die im Projekt erfassten Akteure abgefragt.[8] Informationen zu den familiären und beruflichen Hintergründen werden mithilfe eines Skriptes extrahiert und nach einer Überprüfung in die Datensätze der Akteur_innendatenbank eingepflegt. Weiteren biographische Informationen werden aus studentischen Dossier übernommen, die im Rahmen des Projektes von Studierenden der Universität zu Köln erarbeitet wurden.

Diese zwei Ansätze der Datenanreicherung sind als Schritt zu einer Öffnung von Diskussions- und Interaktionsräumen mit den Daten zu verstehen. Gleichzeitig besteht auch der Wunsch die im Projekt erarbeiteten Personendaten anderen Projekten zur Verfügung zu stellen, um so auch über den Projekt und Theaterkontext hinausgehende, prosopographische Studien zu ermöglichen. Hierbei erwies sich der Ansatz von einer prosopographischen Schnittstelle als sinnvoll – Personendaten profitieren von der Interoperabilität, die automatisierte Extraktion von Personenrelationen ermöglicht. (cf. Vogeler 2019).[9]

Als ein Ergebnis der Datenanalyse konnte zum Beispiel festgestellt werden, dass ein Akteur immer wieder an zentraler Stelle auftauchte: Carl Hagemann (1871-1945), der vor allem in den 1910er und 1920er Jahren als Intendant in Mannheim, Hamburg, Wiesbaden und beim frisch gegründeten Rundfunk in Berlin sowohl als Regisseur wie auch als Organisator tätig war. Darüber hinaus hat Hagemann in seiner gesamten Laufbahn intensiv als Autor gewirkt. Da sich anhand dieses Akteurs die komplexen Netzwerke der Theaterschaffenden besonders eindrücklich zeigen lassen, wurden die Bezüge zu seiner Theaterarbeit bei den zu digitalisierenden Objekten in den Nachlässen bevorzugt berücksichtigt. Hagemann repräsentiert in diesem Sinne idealtypisch auch jene im Antrag in den Blick genommenen biographischen Verläufe, die unterschiedliche historische Zäsuren überspannen.

Es lässt sich festhalten, dass es durch die Daten- und Prozessmodellierung realisierbar ist, aus den Objektdaten Datensätze zu extrahieren, die unterschiedlich modelliert und anders definiert sind - wie beispielsweise Akteure oder Ereignisse. Auf diese Weise agieren sie nicht mehr objektgebunden, stehen mit ihnen aber noch immer in engen relationalen Beziehungen. Hiermit können die Datensätze sowohl komplexe Interaktionsnetzwerke zwischen Entitäten abbilden als auch unabhängig vom intendierten Zweck alleinstehend verwendet werden. Mit dieser Aufbereitung wird eine Datenbasis generiert, die als Grundlage für eine Forschungsumgebung dient. Im Laufe des Projektes wurden bisher 1217 Objekte erfasst. Mit Hilfe dieser Objekten sind 3198 Akteur_innen und 290 Ereignisse erschlossen worden. Diese Basis erlaubt zum einen die theaterwissenschaftliche Neuperspektivierung der Bestände, zum anderen kann sie durch die generierten Netzwerke als Grundlage einer fachwissenschaftlichen Neukonfiguration der zentralen Epochen von Theater- und Kulturgeschichte der Moderne genutzt werden.

Fußnoten

1. Von dem Arbeitsbuch “100 Jahre Theaterwissenschaftliche Sammlung Köln Dokumente, Pläne, Traumreste” kann ein Überblick über die Forschung an theaterwissenschaftlich relevanten Objekten und deren Kontexte gewonnen werden (Marx, 2019)
2. Der Begriff “Akteur_in” wird hier weit gefasst und bezieht sich sowohl auf natürliche Personen als auch auf Körperschaften.
3. siehe Stand von LIDO Terminology unter: http://terminology.lido-schema.org/.
4. http://cmdi-maker.uni-koeln.de/
5. https://ibsenstage.hf.uio.no/, https://www.ausstage.e-du.au/
6. https://www.performing-arts.eu/
7. siehe https://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-person.html
8. siehe https://www.www.deutsche-biographie.de, für API siehe http://data.deutsche-biographie.de/about/
9. siehe auch https://github.com/GVogeler/prosopogrAPhI

Bibliographie

Barabucci, Gioele (2018): *Funktionale und deklarative Programmierung-basierte Methode für nachhaltige, reproduzierbare und verifizierbare Datenkuration*. DHd Konferenz 2018: Kritik der digitalen Vernunft, 214–219, doi: https://doi.org/10.18716/KUPS.8085 [letzter Zugriff 10. September 2019].

Bogatzki, Jakob (2019): »(Re-)Collecting Theatre History. Neue Perspektiven biografischer und theaterhistoriografischer Forschung«, in: *Die vierte Wand*, Heft 9, S. 26–29.

Deutsche Biographie. Hauptseite, https://www.deutsche-biographie.de [letzter Zugriff 10. September 2019]

Deutsche Biographie. Services, http://data.deutsche-biographie.de/about/ [letzter Zugriff 10. September 2019]

Illmayer, Klaus (2017): „Aufbau einer digitalen Infrastruktur für Theaterwissenschaft" (Abstract), https://www.uibk.ac.at/congress/dha2017/bilder-und-dateien/aufbau-einer-digitalen-infrastruktur-fuer-theaterwissenschaft.pdf [letzter Zugriff 7. September 2019].

Leonhardt, Nic (2014): *Digital Humanities and the Performing Arts: Building Communities, Creating Knowledge*. Keynote auf der SIBMAS/TLA Konferenz, New York (NY), 12. Juni 2014, https://mappinggth.hypotheses.org/files/2014/09/Nic-Leonhardt_DH-and-the-Performing-Arts_June-2014.pdf [letzter Zugriff 3. August 2019].

LIDO Working Group. What is LIDO, http://network.icom.museum/cidoc/working-groups/lido/what-islido/ [letzter Zugriff 10. September 2019].

LIDO Working Group. LIDO Terminology, http://network.icom.museum/cidoc/working-groups/lido/lido-technical/terminology/ [letzter Zugriff 10. September 2019].

Marx, Peter W., Hrsg. (2019): 100 Jahre Theaterwissenschaftliche Sammlung Köln. Dokumente, Pläne, Traumreste. Berlin: Alexander Verlag.

Probst, Nora / Pinto, Vito: Re-Collecting Theatre History. Theaterhistoriografische Nachlassforschung mit Verfahren der Digital Humanities. In: Wihstutz, Benjamin; Hoesch, Benjamin (Hg.): Neue Methoden der Theaterwissenschaft (Druck in Vorbereitung).

Vogeler, Georg / Vasold, Gunter / Schlögl, Matthias (2019): *Von IIIF zu IPIF? Ein Vorschlag für den Datenaustausch über Personen*. In: Sahle, Patrick (Hg.): DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts. Frankfurt / Mainz. DHd. 2019 DOI: 10.5281/zenodo.2600812. pp. 239-241.

Redewiedergabe in Heftromanen und Hochliteratur

Brunner, Annelen
brunner@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Mannheim

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

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Mannheim

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Einführung

Die Art und Weise, wie die Rede und Gedanken einer Figur im Erzähltext eingebunden werden, ist einer der traditionellen Aspekte der Narratologie (vgl. z.B. Genette 2010; Martínez/Scheffel 2016). Die vorgestellte Studie untersucht die Anteile unterschiedlicher Redewiedergabeformen im Vergleich zwischen zwei Literaturtypen von gegensätzlichen Enden des Spektrums: Hochliteratur – definiert als Werke, die auf der Auswahlliste von Literaturpreisen standen – und Heftromanen, massenproduzierten Erzählwerken, die zumeist über den Zeitschriftenhandel vertrieben werden und früher abwertend als „Romane der Unterschicht" (Nusser 1981) bezeichnet wurden. Unsere These ist, dass sich diese Literaturtypen hinsichtlich ihrer Erzählweise unterscheiden, und sich dies in den verwendeten Wiedergabeformen niederschlägt. Der Fokus der Untersuchung liegt auf der Dichotomie zwischen direkter und nicht-direkter Wiedergabe, die schon in der klassischen Rhetorik aufgemacht wurde (vgl. McHale 2014).

Die Studie geht von manuell annotierten Daten aus und evaluiert daran die Validität automatischer Annotationswerkzeuge, die im Anschluss eingesetzt werden, um die Menge des betrachteten Materials beträchtlich zu erweitern.

Zur Kontrastierung von Heftromanen und Hochliteratur mit quantitativen Methoden liegen bereits Studien vor, welche sich mit Fragen der sprachlichen und thematischen Komplexität beschäftigen (Jannidis/Konle/Leinen 2019a/2019b). Das verwendete Annotationssystem sowie die Erkenner wurden im Rahmen des Redewiedergabe-Projekts entwickelt (Brunner et al. 2019a/2019b).

Voruntersuchung und Evaluation der automatischen Methoden

Für die Voruntersuchung wurden aus 22 Hochliteratur-Texten und 22 Heftromanen zufällige Textausschnitte von ca. 1000 Tokens gezogen. Da Heftromane typischerweise in Reihen mit unterschiedlichem Fokus erscheinen, betrachten wir auch das Verhalten dieser unterschiedlichen Heftroman-Genres. Die Heftroman-Ausschnitte wurden darum je zur Hälfte aus den Genres Liebesroman und Horrorroman gewählt. Die Texte wurden von zwei Erstannotatoren unabhängig voneinander bearbeitet. Eine dritte Person erstellte dann auf dieser Grundlage eine Konsensannotation, indem sie die beiden Annotationen verglich, Unstimmigkeiten bereinigte und wenn nötig offensichtliche Fehler korrigierte.

Das Annotationssystem (Brunner et al. 2019a) erfasst sowohl die Wiedergabe von Rede als auch von Gedanken und Geschriebenem. Es umfasst vier Haupttypen von Wiedergabe:

- direkt: *Er dachte:* „Ich habe Hunger."
- frei-indirekt ('erlebte Rede'): *Er war ratlos.* Wo sollte er jetzt etwas zu essen finden?
- indirekt: *Er sagte,* dass er Hunger habe.
- erzählt: Er sprach über das Mittagessen.

Frei-indirekte Wiedergabe ist im Folgenden ausgeschlossen, da sich die automatische Erkennung für diese Form als noch zu unzuverlässig erwiesen hat. Die Formen indirekte und erzählte Wiedergabe sind zu 'nicht-direkt' zusammengefasst. Diese Form umfasst damit sowohl die klassische indirekte Wiedergabe mit Rahmenformel und abhängiger Proposition als auch strukturell abweichende und häufig stärker zusammenfassende. Sie steht im Gegensatz zur direkten Wiedergabe insofern, als die Rede, Gedanken oder schriftliche Äußerung einer Figur in den Erzähltext integriert anstatt in einem Zitat klar davon abgesetzt wird.

Die automatischen Erkenner beruhen auf DeepLearning.[1] Als Trainingsmaterial wurde hauptsächlich das Redewiedergabe-Korpus (Brunner et al. 2019b; verfügbar unter github.com/redewiedergabe/corpus) verwendet, welches historische Textdaten (19. bis frühes 20. Jahrhundert) und sowohl fiktionales als auch nicht-fiktionales Material umfasst. Da die in dieser Studie verwendeten Texte deutlich moderner sind (ca. 1950 bis Gegenwart), war es umso wichtiger, die Übertragbarkeit des Modells zu evaluieren. Es lagen speziell trainierte Erkenner für jede der Formen direkte, indirekte und erzählte Wiedergabe vor, die unabhängig voneinander auf die Testdaten angewendet wurden. Überlagerungen von Wiedergabetypen werden somit erkannt. Als ‚nicht-direkt' zählen Tokens, die entweder als Teil von indirekter oder von erzählter Wiedergabe erkannt wurden.

Um eine bessere Einschätzung der Erkennungswerte zu geben, ein paar Worte zur Vorkommenshäufigkeit der Redewiedergabetypen: Im den konsensannotierten Testdaten liegt der durchschnittliche Anteil von direkter Wiedergabe knapp unter 30% der Tokens (mit starken Schwankungen), von nicht-direkter bei ca. 15%.

Tabelle 1 zeigt die Übereinstimmungswerte zwischen den Erstannotatoren, um einen Eindruck zu vermitteln, wie verlässlich eine von Menschen durchgeführte Annotation wäre.

	F1	Precision	Recall	Fleiss' Kappa	prozentuale Anteile	
					durchschnittlicher absoluter Fehler	Standardabweichung des Fehlers
direkt						
Heftromane	0,99	0,98	0,99	0,98	0,65	1,07
Hochliteratur	0,97	0,98	0,96	0,96	1,29	2,7
Alle Samples	0,98	0,98	0,98	0,97	0,97	2,0
nicht-direkt						
Heftromane	0,77	0,78	0,77	0,74	1,96	1,46
Hochliteratur	0,78	0,78	0,78	0,74	2,60	2,64
Alle Samples	0,78	0,78	0,78	0,74	2,28	2,10

Tabelle 1: Übereinstimmung zwischen den Erstannotatoren. F1, Precision, Recall jeweils für die Kategorie direkt bzw. nicht-direkt, gerechnet auf Tokenbasis; prozentuale Anteile ebenfalls auf Tokenbasis.

Tabelle 2 zeigt nun für die Formen direkt und nicht-direkt die Übereinstimmungsquoten der automatischen Methoden im Vergleich zur Konsensannotation. Wenn man als Baseline einen Erkenner annimmt, der jedes Token mit 50% Wahrscheinlichkeit als Teil von Wiedergabe klassifiziert, käme man für die Testdaten (alle Samples) für direkt auf einen F1-Score von 0,36 (Precision: 0,28; Recall: 0,50), für nicht-direkt auf einen F1-Score von 0,23 (Precision: 0,17; Recall: 0,50), wobei die Einzelscores für Heftromane vs. Hochliteratur bei direkt gleich wären, bei nicht-direkt etwas besser für Hochliteratur (F1: 0,25).

	F1	Precision	Recall	Fleiss' Kappa	prozentuale Anteile	
					durchschnittlicher absoluter Fehler	Standardabweichung des Fehlers
direkt						
Heftromane	0,92	0,91	0,94	0,96	4,4	7,78
Hochliteratur	0,74	0,67	0,84	0,63	11,28	10,6
Alle Samples	0,83	0,78	0,89	0,76	7,84	9,91
nicht-direkt						
Heftromane	0,75	0,76	0,73	0,75	1,69	1,59
Hochliteratur	0,66	0,73	0,61	0,60	4,23	3,26
Alle Samples	0,70	0,74	0,67	0,65	2,96	2,86

Tabelle 2: Auswertung der automatischen Methoden gegen die Konsens-Annotation.

Bei direkter Wiedergabe sind die Erkennungsraten der automatischen Methoden vor allem bei den Heftromanen gut, es gibt jedoch Schwankungen zwischen den Textausschnitten. Probleme treten insbesondere bei Ich-Perspektive in Kombination mit unmarkierter Wiedergabe auf, was in Hochliteratur häufiger vorkommt. Dennoch sind die mit dem maschinellen Erkenner erzielten Ergebnisse – gerade für solche Fälle – deutlich stabiler als eine Identifikation von direkter Wiedergabe anhand von Anführungszeichen gewesen wäre. Insgesamt neigt der Erkenner dazu, den Anteil von direkter Wiedergabe eher zu über- als zu unterschätzen. Der durchschnittliche absolute Fehler bei der Abschätzung der Anteile liegt im Schnitt bei ca. 10%.

Für die nicht-direkte Wiedergabe ist zu betonen, dass die Übereinstimmungsquote auch zwischen Menschen deutlich schlechter ist (vgl. Tabelle 1). Ursache ist, dass durch die stärkere Integration in den Erzähltext sowohl die genaue Abgrenzung als auch die Entscheidung, was als Wiedergabe zu werten ist, schwieriger sind. Die automatischen Methoden erreichen bei den Heftromanen fast gleiche Verlässlichkeit, während die Hochliteratur-Abschnitte sich wiederum als etwas schwieriger erweisen. Da die Anteile von nicht-direkt geringer sind und weniger Schwankungen unterliegen als die Anteile von direkt, ist auch der durchschnittliche absolute Fehler deutlich geringer (ca. 3%), wobei der Anteil von nicht-direkt eher unterschätzt wird.

Da für die Erzählweise eines Textes auch das Zusammenspiel der beiden Wiedergabetypen von Interesse ist, visualisieren wir die Textausschnitte der unterschiedlichen Untersuchungsgruppen in einem Scatterplot (Abb. 1).

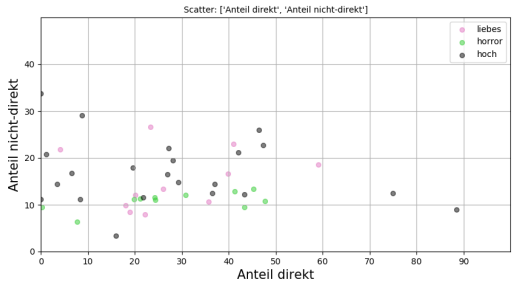


Abbildung 1: Scatterplot der 1000-Token-Samples auf Basis der manuellen Konsens-Annotation

In dieser Darstellung auf Basis der manuellen Konsens-Annotation lässt sich ein Trend der Horrorroman-Textausschnitte erkennen, mit niedrigen Werte sowohl in direkt als auch in nicht-direkt zusammenzuclustern, während Hochliteratur und Liebesroman stark gestreut erscheinen. Mit einem Permutationstest (p=0,01) (Koplenig 2019) lassen sich im Vergleich Heftroman vs. Hochliteratur allerdings auf keiner der beiden Dimensionen signifikante Unterschiede nachweisen. Bei dem Vergleich mit Genres sind lediglich die Abweichungen zwischen Hochliteratur und Horrorroman im Anteil nicht-direkter Wiedergabe signifikant. Legt man die automatisch annotierten Daten zugrunde, verschwindet auch diese Signifikanz.

Erweiterung der Studie auf Volltexte

Im nächsten Schritt erweitern wir unser Untersuchungsmaterial stark. Das Korpus wurde aus Volltexten zusammengestellt, wobei diesmal die Unterschiede zwischen Hochliteratur und den einzelnen Genres in den Fokus gerückt wurden: 50 Hochliteratur-Texte wurden mit jeweils 50 Heftromanen aus vier unterschiedlichen Reihen kontrastiert, die unterschiedliche Genres repräsentieren (vgl. Tab. 3).

Gruppe	Anzahl Texte	Anzahl Autoren	Reihe	Tokens pro Text
Hochliteratur	50	50	-	35.310 - 332.568
Liebes	50	50	Julia Extra	19.608 - 42.908
Science-Fiction	50	11 (2-5 Texte pro Autor)	Perry Rhodan	28.365 - 34.904
Horror	50	4 (2-35 Texte pro Autor)	John Sinclair	27.920 - 34.838
Krimi	50	unbekannt; mind. 2	Jerry Cotton	29.212 - 50.171

Tabelle 3: Korpuszusammensetzung

Ein Ziel war, eine größtmögliche Diversität von Autoren zu erreichen, um zu verhindern, dass das Autorensignal die Gruppenzugehörigkeiten überlagert, die uns eigentlich interessieren. Problematisch war dies bei Horrorromanen, wo ein Autor die Reihe extrem dominiert und Krimis, bei denen so gut wie keine Autoreninformationen verfügbar waren. Es ist allerdings bekannt, dass die Reihe „Jerry Cotton" von über 100 unterschiedlichen Autoren verfasst wurde (vgl. Karr 2019).[2] Da Heftromane üblicherweise unter Pseudonym veröffentlicht werden, ist die Autorenzuschreibung hier insgesamt mit Unsicherheit behaftet (vgl. Hügel 2001).

Die Texte wurden mit den automatischen Erkennern komplett annotiert. Da die Variation der Textlängen insbesondere in der Gruppe Hochliteratur stark ist, wurden die Texte in 1000-Token-Abschnitte zerlegt, für diese die Anteile von direkter und nicht-direkter Wiedergabe berechnet und die Ergebnisse anschließend für jeden Text gemittelt (analog zur standardisierten Type-Token-Ratio). Anders als bei den Testdaten zeigen sich bei der Auswertung nun klare Unterschiede in beiden Dimensionen: Der Anteil direkter Wiedergabe ist bei Hochliteratur geringer, während der Anteil nicht-direkter Wiedergabe höher ist. Die Signifikanz dieser Unterschiede, wie auch viele Unterschiede zwischen den Genres, lassen sich mit dem Permutationstest mit p=0.01 bestätigen (vgl. Abbildung 2 und 3).

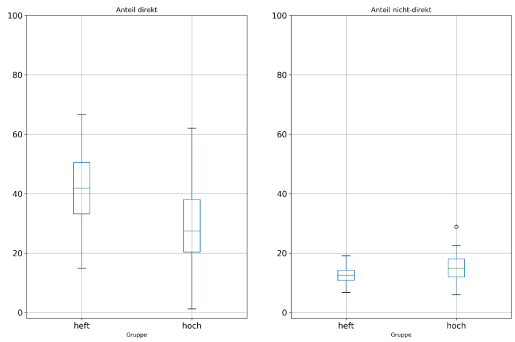


Abbildung 2: Boxplots Heftromane vs. Hochliteratur. Unterschiede sind signifikant mit p=0.01.

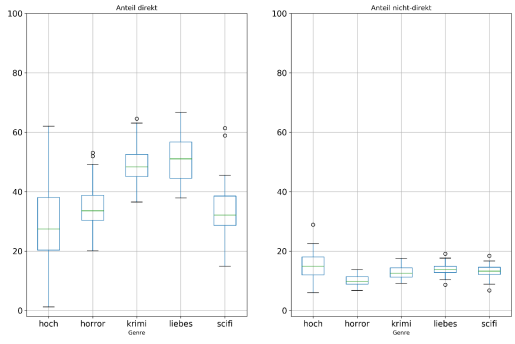


Abbildung 3: Boxplots Hochliteratur vs. Heftchen-Genres: Signifikant mit p=0,01 sind die Unterschiede hoch/krimi, hoch/horror, horror/krimi, horror/liebes (in beiden Dimensionen); hoch/liebes, scifi/liebes, scifi/krimi (nur in Anteil direkt); horror/scifi (nur in Anteil nicht-direkt).

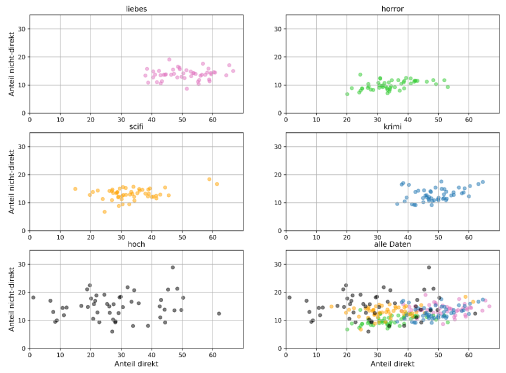


Abbildung 4: Scatterplots für die Volltexte, automatische Annotation

Bei der Betrachtung beider Dimensionen in Relation (Abb. 4) fällt sofort auf, dass die Hochliteratur-Texte eine deutliche Streuung aufweisen, während nicht nur die einzelnen Genres, sondern auch die Heftromane als Gruppe zusammenclustern. Angesichts der Tatsache, dass die Heftroman-Genres bewusst reglementierte Reihen sind, während die Hochliteratur-Gruppe nur dadurch definiert ist, dass die enthaltenen Werke als literarisch hochwertig eingeschätzt wurden, ist dieser Befund nicht erstaunlich. Es ist jedoch durchaus bemerkenswert, dass sich der Unterschied zwischen konventionalisiertem und individualistischem Erzählen auf der Dimension der Redewiedergabetypen so deutlich quantitativ nachweisen lässt.

Die Hochliteratur-Texte sind zudem die einzige Gruppe, in der ein ‚Übergewicht' an nicht-direkter im Gegensatz zu direkter Wiedergabe auftritt. Die Autoren sind also in der Art und Weise, wie sie Figurenstimmen in den Text einbinden, sowohl individualistischer als auch eher bereit, nicht das direkte Zitat zu wählen.

Innerhalb der Gruppe der Heftromane man kann für die Genres Liebesroman, Horrorroman und Krimi einen nahezu linearen Anstieg der beiden Wiedergabeformen in Relation zueinander beobachten, wobei der Anteil direkter Wiedergabe stets höher ist. Es differenziert sich recht klar das Horror-Genre mit einem insgesamt geringeren Wiedergabeanteil, während die ‚kommunikativeren' Genres Liebesroman und Krimi sich stark überlagern. Für diese beiden Genres lassen sich auch keine signifikanten Unterschiede nachweisen. Science-Fiction nimmt eine Zwischenstellung ein: Die Texte sind diverser und streuen ähnlich wie Hochliteratur, wenn auch nicht so extrem. Es ist das einzige Heftroman-Genre, für das sich auf keiner der beiden Dimensionen signifikante Unterschiede zu Hochliteratur nachweisen lassen. Dies passt zu Beobachtungen von Jannidis/Konle/Leinen (2019a), dass Science-Fiction unter den Heftroman-Genres eine Sonderstellung einnimmt und auch bei unterschiedlichen Komplexitätsmaßen wie standardisierter Type-Token-Ratio und Wortlänge höher abschneidet als die anderen Genres.

Warum zeigen sich diese interessanten Muster erst in den Volltextdaten und nicht in der Voruntersuchung? Die Erklärung ist, dass die Schwankungen in den Anteilen von Wiedergabe innerhalb eines Erzähltextes so stark sind, dass sie die beobachteten Trends überlagern. Abb. 5 zeigt einen Datenpunkt für jeden der 1000 Token-Abschnitte aus dem Untersuchungskorpus. Zwar werden in der Gesamtheit dieser Datenpunkte die gleichen Trends sichtbar wie in Abb. 4, doch wenn man – wie bei der Testauswertung – nur wenige zufällig gezogene 1000-Token-Abschnitte betrachtet, ist es unwahrscheinlich, dass sie erkennbar wären. Die Ausweitung auf mehr Material, die durch die Anwendung automatischer Methoden möglich wurde, führt hier also zu einem Erkenntnisgewinn, der sonst nur mit extremem Annotationsaufwand möglich gewesen wäre.

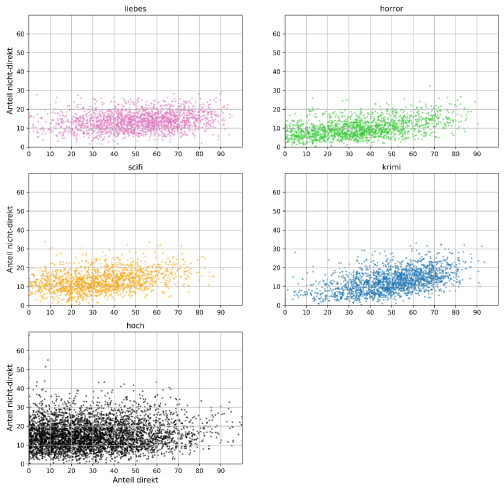


Abbildung 5: Scatterplots für die Volltexte (zerlegt in 1000-Token-Abschnitte), automatische Annotation.

Da die schlechteren Erkennungsraten des Direkte-Wiedergabe-Erkenners für Texte in der Ich-Perspektive bekannt sind, wurde für einen großen Teil der Texte die Erzählperspektive ermittelt. Die Durchmischung ist sowohl bei den Hochliteraturtexten als auch bei den Heftromanen gegeben und Texte beider Perspektiven platzieren sich an unterschiedlichen Stellen. Einzig der Bereich mit sehr niedrigem Anteil von direkter Wiedergabe (<17%) ist ausschließlich durch Texte in Er-Perspektive besetzt. Der Einfluss der Erzählperspektive ist ein Faktor, der in weiteren Untersuchungen genauer betrachtet werden sollte.

Ausblick

Mit den vorhandenen Werkzeugen ist es denkbar, die Studien auf noch mehr Textmaterial auszuweiten und dabei auch weitere Genres von Heftromanen zu untersuchen. Zudem lässt sich die Methodik leicht auf Fragestellungen zu den Anteilen von Redewiedergabe in anderen Textgruppen übertragen, z.B. zwischen fiktionalem und nicht-fiktionalem Material oder im diachronen Vergleich. Wir arbeiten zudem im Redewiedergabe-Projekt weiter daran, unsere automatischen Erkenner zu verbessern, insbesondere auch den für freie-indirekte Wiedergabe (zum aktuellen Stand vgl. Brunner et al. 2019c). Die im Redewiedergabe-Projekt entwickelten Erkenner werden nach Abschluss des Projekts im Frühjahr 2020 der Forschungsgemeinschaft zur Verfügung gestellt werden, ebenso wie große Teile des verwendeten manuell annotierten Trainingsmaterials.

Fußnoten

1. Wir verwenden eine BiLSTM-CRF-Architektur mit contextual string embeddings in Kombination mit FastText word embeddings (adaptiert von Akbik/Blythe/Vollgraf 2018). Eine detaillierte Beschreibung übersteigt den Rahmen dieses Beitrags, aber sie entspricht im Wesentlichen der Architektur in Brunner et al. 2019c.
2. Von den 50 Jerry-Cotton-Krimis konnten nur für 4 Autoreninformationen gefunden werden, von denen eine zudem unsicher ist. Gesichert kann man damit sagen, dass mindestens 2 unterschiedliche Autoren in der Gruppe Krimi vorliegen, vermutlich sind es jedoch deutlich mehr.

Bibliographie

Akbik, Alan / Blythe, Duncan / Vollgraf, Roland (2018): "Contextual String Embeddings for Sequence Labeling", in: *Proceedings of the 27th International Conference on Computational Linguistics* 1638-1649.

Brunner, Annelen / Weimer, Lukas / Engelberg, Stefan / Jannidis, Fotis / Tu, Ngoc Duyen Tanja (2019a): *Annotationsrichtlinien des Projekts „Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse".* Zenodo. http://doi.org/10.5281/zenodo.2634994 [letzter Zugriff 17. Dezember 2019].

Brunner, Annelen / Weimer, Lukas / Tu, Ngoc Duyen Tanja / Engelberg, Stefan / Jannidis, Fotis (2019b): "Das Redewiedergabe-Korpus. Eine neue Ressource", in: *Digital Humanities im deutschsprachigen Raum – Konferenzabstracts* 103-106.

Brunner, Annelen / Tu, Ngoc Duyen Tanja / Weimer, Lukas / Jannidis, Fotis (2019c): "Deep Learning for Free Indirect Representation", in: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)* 241-245.

Genette, Gérard (2010): *Die Erzählung*. 3., durchges. und korrigierte Aufl. Paderborn: Fink.

Hügel, Hans-Otto (2001): "Kommunikative und ästhetische Funktion des Romanhefts", in: Leonhardt, Joachim-Felix / Ludwig, Hans-Werner / Schwarze, Dietrich / Straßner, Erich (eds.): *Medienwissenschaft.* Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen. Berlin / New York: de Gruyter 1621-1631.

Jannidis, Fotis / Konle, Leonard / Leinen, Peter (2019a): "Makroanalytische Untersuchung von Heftromanen", in: *Digital Humanities im deutschsprachigen Raum – Konferenzabstracts* 167-173.

Jannidis, Fotis / Konle, Leonard / Leinen, Peter (2019b): "Thematic complexity", in: *Digital Humanities Conference* https://dev.clariah.nl/files/dh2019/boa/0504.html [letzter Zugriff 17. Dezember 2019].

Karr, H.P. (2019): "Cotton, Jerry", in: Ders. (ed.): *Lexikon der deutschen Krimi-Autoren – Internet Edition* http://www.krimi-lexikon.de/cotton.htm [letzter Zugriff 17. Dezember 2019].

Koplenig, A. (2019): "A non-parametric significance test to compare corpora", in: *PLoS ONE* 14(9): e0222703 https://doi.org/10.1371/journal.pone.0222703 [letzter Zugriff 17. Dezember 2019].

Martínez, Matías / Scheffel, Michael (2016): *Einführung in die Erzähltheorie*. 10. Auflage. München: C.H. Beck.

McHale, Brian (2014): "Speech Representation", in: Hühn, Peter / Pier, John / Schmid, Wolf / Schönert, Jörg (eds.): *The living handbook of narratology*. Hamburg: Hamburg University Press 434-446 http://www.lhn.uni-hamburg.de/article/speech-representation [letzter Zugriff 17. Dezember 2019].

Nusser, Peter (1981): *Romane für die Unterschicht*. Groschenhefte und ihre Leser. 5. Auflage. Stuttgart: Metzler.

Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren

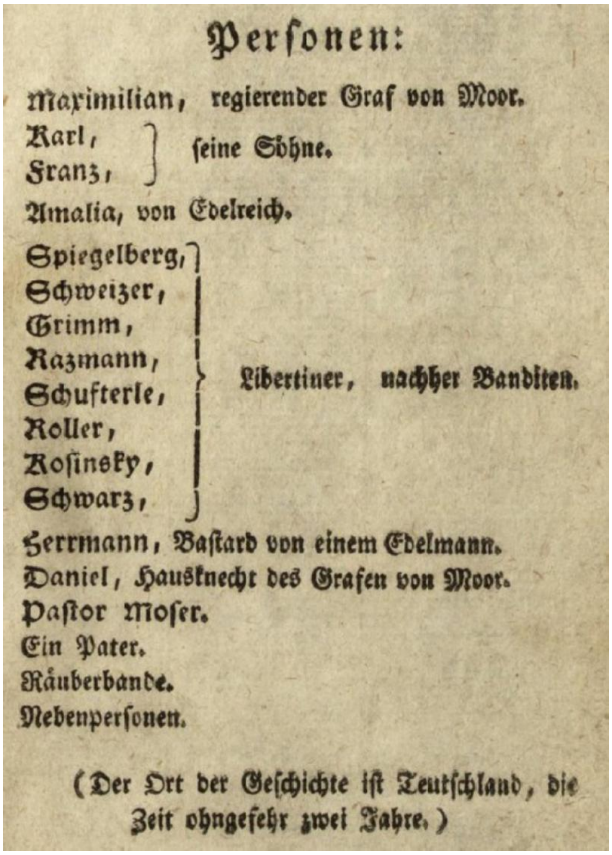
Wiedmer, Nathalie
nathalie.wiedmer@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pagel, Janis
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Reiter, Nils
nils.reiter@uni-koeln.de
Universität Stuttgart, Deutschland; Universität Köln, Deutschland

Einleitung

In diesem Beitrag stellen wir eine Methode vor, um Informationen über Figurenrelationen in dramatischen Texten, die innerhalb der *dramatis personae* (Figurenverzeichnis) sprachlich kodiert sind, zu extrahieren und maschinenlesbar im TEI/XML vorzuhalten. Das Figurenverzeichnis kann als Paratext (Genette 1993) dem Nebentext zugerechnet werden, ist jedoch literaturwissenschaftlich, von Einführungswerken abgesehen, noch so gut wie nicht erschlossen.[1] Das Figurenverzeichnis steht zwar unabhängig vom eigentlichen Text am Anfang, kann jedoch bereits Figuren- bzw. Textwissen vermitteln, indem die Figuren nach sozial-politischem Stand, Familienzugehörigkeit oder nach anderen Gruppierungen geordnet sind (vgl. Abbildung 1). Häufig lässt sich an der Positionierung eines Names im Figurenverzeichnis auch die Wichtigkeit der betreffenden Figur im Drama ablesen (Pangallo 2015, 91). Durch diese Strukturierung ist es teilweise möglich, schon vorab auf zentrale Konfliktpotentiale des Textes zu schließen (Jeßing 2015, 79–80). Darüberhinaus kann das Figurenverzeichnis laut Pfister und Asmuth auch der Ort erster auktorialer Bewertungen oder Hinweise sein und dient somit nicht nur der reinen Vorstellung der Figuren und ihrer Strukturen untereinander (Pfister 2001, 95; Asmuth 2016, 85).


Personen:

Maximilian, regierender Graf von Moor.
Karl, } seine Söhne.
Franz, }
Amalia, von Edelreich.
Spiegelberg, }
Schweizer, }
Grimm, }
Razmann, } Libertiner, nachher Banditen.
Schufterle, }
Roller, }
Rosinsky, }
Schwarz, }
Herrmann, Bastard von einem Edelmann.
Daniel, Hausknecht des Grafen von Moor.
Pastor Moser.
Ein Pater.
Räuberbande.
Nebenpersonen.

(Der Ort der Geschichte ist Teutschland, die Zeit ohngefehr zwei Jahre.)

Abbildung 1: Figurenverzeichnis in Die Räuber (Friedrich Schiller, 1781)

Das Verfahren – und dessen Implementierung in einem Python-Skript – ist auch für in Zukunft digitalisierte Dramen anwendbar, und wird von uns als quelloffene Software zur Verfügung gestellt. Es ist vergleichsweise einfach auf neue Sprachstufen oder Genres anpassbar und liefert – auch bei nicht-perfekten Ergebnissen – eine gute Vorlage. Eine Evaluation des Verfahrens erfolgt auf ungesehenen Testdaten. Außerdem veröffentlichen wir einen Datensatz mit extrahierten Figurenrelationen aus deutschsprachigen Dramen, die manuell validiert und korrigiert wurden. Diese Daten werden zur einfachen und breiten Nutzung im TEI-Format in das GerDraCor[2] eingespeist. Schlussendlich beschreiben wir beispielhaft zwei Analyseszenarien in denen die Daten neue Einblicke bieten (können).

Automatische Extraktion von Figurenrelationen

Unsere Methode unterscheidet zwischen sieben Kategorien von Figurenrelationen (Tabelle 1). Ausschlaggebend für die Zuordnung zu einer der Kategorien sind Signalwörter wie “Vater”, “Kammerdiener”, “Geschwister” etc. Diese Signalwörter werden in einer kontextfreien Grammatik der entsprechenden Kategorie zugeordnet.

Tabelle 1: Figurenrelationen

Relationen Label	gerichtet/ungerichtet	Beschreibung
parent_of	directed	Eine Figur ist Elternteil einer anderen
lover_of	directed	Liebesbeziehungen (unverheiratet)
related_with	directed	Familienbeziehungen (außer Eheleute)
associated_with	directed	Figuren, die miteinander anderweitig verbunden sind (z.B. Diener, Kindermädchen etc.)
siblings	undirected	Figuren, die mindestens ein gemeinsames Elternteil haben
spouses	undirected	verheiratete oder verlobte Figuren
friends	undirected	Freundschaftsbeziehungen

Kontextfreie Grammatiken bezeichnen in der Informatik eine Sammlung aller syntaktisch korrekten Programme einer Programmiersprache (Böckenhauer und Hromkovič 2013, 177). Die formalisierte Art, in der die Grammatik alle Regeln einer Programmiersprache enthält, erlaubt es, automatisierte Syntaxanalysen von Programmen durchzuführen (Böckenhauer und Hromkovič 2013, 177). Die Regeln werden mit Hilfe zweier Alphabete beschrieben: Das Terminalalphabet enthält alle Wörter einer Sprache, wohingegen das Nichtterminalalphabet Variablen enthält, die vorgeben, auf welche Art und Weise die Wörter kombiniert werden können (Böckenhauer und Hromkovič 2013, 178).

Wir nutzen eine solche Grammatik, um drei verschiedene Zeilenarten im Figurenverzeichnis zu unterscheiden, bei denen es sich um Nichtterminale handelt. Alle in den Sätzen vorkommenden Tokens sind Terminale, deren Kombination und Anzahl Aufschluss darüber gibt, um was für eine Art von Zeile es sich jeweils handelt. Auf diese Weise können auch zeilenübergreifende Relationen erkannt werden.

Zu Beginn des Programmablaufs werden die in GerDraCor vorhandenen Figuren-IDs zusammen mit dem Figurenverzeichnis ausgelesen und gespeichert. Da wir die Beziehungen zwischen den Figuren ausschließlich anhand der Angaben im Figurenverzeichnis konstruieren, muss der Dramentext nicht extra eingelesen werden. Daraus ergibt sich die Beschränkung, dass jegliche Beziehungen, die nicht im Figureverzeichnis explizit gemacht werden, vom Programm auch nicht erkannt werden können. Es geht demnach ausschließlich darum, das Personenverzeichnis maschinenlesbar und -interpretierbar zu machen. So ignoriert das Programm beispielsweise auch alle Zeilen, die eine Gruppe von Figuren als Kollektiv einführt, da diese als “Nummern oder als anonyme Angehörige von Untergruppen” (Schlaffer 1972, 11) meistens keine eigenen Namen haben und auch keine explizit gemachten Beziehungen.[3]

Anschließend werden alle Tokens jeder Zeile des Figurenverzeichnisses daraufhin untersucht, ob es sich dabei um Figurennennungen oder Signalwörter handelt und die Grammatik einem Parser übergeben, der die Zeilen des Figurenverzeichnisses in Baumstrukturen überführt (Abbildung 2).

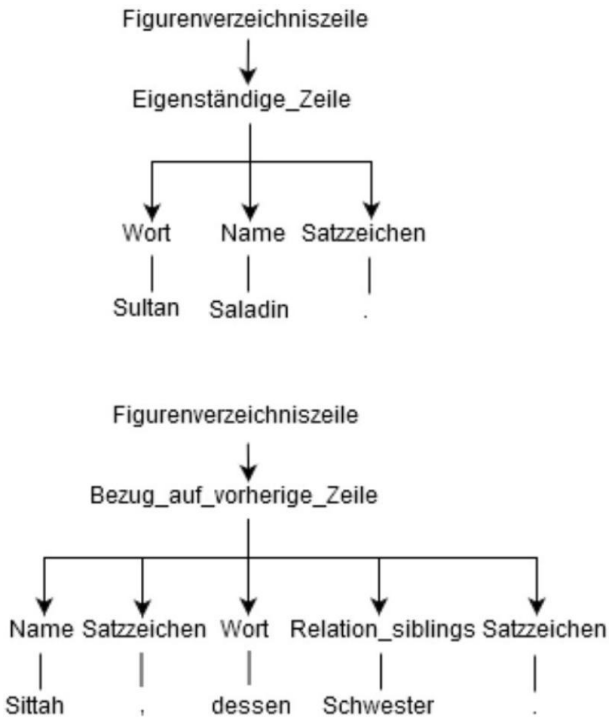


Abbildung 2: Zwei reduzierte Baumstrukturen für Figuren aus Nathan der Weise.

Aus den erstellten Baumstrukturen werden einzelne Informationen ausgelesen, die grundlegend für die Erkennung der Figurenrelationen sind. Zuerst wird überprüft, wie viele IDs sich in einer Zeile befinden. Die erste oder einzige wird zur Erstellung späterer Relationen abgespeichert. Befindet sich in einer Zeile zusätzlich zu einer ID noch ein Signalwort für eine Figurenrelation, bezieht sich die Zeile in der Regel auf die vorangegangene, wie beispielsweise in *Nathan der Weise*:

<castItem corresp="#saladin">Sultan Saladin.</castItem>
<castItem corresp="#sittah">Sittah, seine Schwester.</castItem>

Die zweite Zeile enthält neben dem Namen noch das Signalwort "Schwester", das auf die Beziehungsart siblings hinweist, eine ungerichtete Relation. Da keine zweite Figurenbezeichnung in der Zeile vorkommt, entnimmt das Programm als zweiten Part für die Geschwisterbeziehung den Namen bzw. die daraus abgeleitete ID saladin aus der vorherigen Zeile:

<relation name="siblings" mutual="#sittah #saladin" />

Wenn die beiden benötigten IDs für das Erstellen der Figurenrelation feststehen, wird die Art der Relation durch das Auslesen des Signalworts aus der Baumstruktur festgestellt. Danach werden daraus die Zeilen mit den Figurenrelationen erstellt und diese anschließend in die jeweilige TEI-Version des Textes geschrieben.

Befindet sich in einer Zeile eine zweite Figuren-ID, bezieht sich die Zeile nicht auf eine vorangegangene, sondern stellt selbst den zweite Bezugspunkt der Relation. Das ist beispielsweise bei der Figur "Camillo Rota" in *Emilia Galotti* der Fall:

<castItem corresp="#camillo_rota">Camillo Rota, einer von des Prinzen Räten.</castItem>

Die erste erkannte ID ist camillo_rota, die zweite der_prinz, abgeleitet aus "des Prinzen". Die IDs werden in gerichtete Relationen mit aktivem und passivem Part überführt:

<relation name="associated_with" active="#camillo_rota" passive="#der_prinz" />

Das Programm arbeitet dabei ausschließlich mit den IDs. Dafür ist es nicht nötig, dass Figurennamen explizit als Namen oder Adelstitel als Titel erkannt werden. Es geht ausschließlich darum aus den einzelnen Wörtern einer Zeile im Figurenverzeichnis Namen bzw. Namensteile und Titelangaben herauszufiltern, die den IDs entsprechen, um die Zeilen einer oder mehreren Figuren zuordnen zu können.

Um auch IDs zu erkennen, die sich geringfügig von den Namensnennungen im Figurenverzeichnis unterscheiden, überprüft das Programm pro Wort eine Reihe an Varianten. So trennt es beispielsweise vom oben gennannten Wort "Prinzen" das Suffix ab und überprüft, ob ein Artikel Teil der ID ist. So kann "des Prinzen" der ID "der_prinz" zugeordnet werden. In manchen Fällen funktioniert diese Abwandlung aber nicht so reibungslos. In *Der Eheteufel auf Reisen* wird eine Figur im Figurenverzeichnis mit dem Namen "Gustel" eingeführt, wohingegen die ID „gustchen" lautet. Die ID orientiert sich hier an der Namensform, die im Stück tatsächlich verwendet wird und nicht an der Bezeichnung im Figurenverzeichnis. Das führt dazu, dass das Programm die ID "gustchen" nicht dem Wort "Gustel" zuordnen kann, da sie sich zu stark unterscheiden.

Evaluation

Um die Methode zu evaluieren, wurden die automatisch erzeugten Relationen manuell nachkorrigiert und so ein Goldstandart erzeugt. Im Schnitt bearbeiteten die Korrektoren 12 Texte pro Stunde. Beim Abgleich der automatisch erzeugten Ergebnisse mit dem Goldstandart lag der Macro-Average-Recall Wert bei 0,3 (Standardabweichung: 0,3) und der Wert von Macro-Average-Precision bei 0,55 (Standardabweichung: 0,4), was einen Macro-Average-F-Score von 0,49 (Standardabweichung: 0,25) ergibt.

Korpus

GerDraCor ist ein deutsches Dramenkorpus, das nach TEI-P5 Standarts kodiert ist und im Dezember 2019 474 Dramen enthält, die im Zeitraum von 1730 bis 1940 veröffentlicht wurden (Fischer u. a. 2019). Es ist Teil des größeren DraCor (Fischer u. a. 2019), das als *Programmable Corpus* darauf ausgelegt ist, durch Community-Anstrengungen korrigiert und verbessert werden zu können (Fischer u. a. 2019, 195). Da auf einem Fork von GerDraCor gearbeitet wurde, können die automatisch erzeugten Figurenrelationen dem Korpus unproblematisch hinzugefügt werden. Zusätzlich wurden die Relationen, wie bereits beschrieben, manuell nachkorrigiert, um eine erhöhte Qualität für die Nachnutzung zu gewährleisten.

Im Rahmen der manuellen Nachkorrektur wurden außerdem interessante Fälle identifiziert. So wird etwa eine Gruppe von Figuren in dem oben abgebildeten Figurenverzeichnis von Schillers *Die Räuber* als "Libertiner, nachher Banditen" bezeichnet, wodurch Informationen aus der späteren Handlung des Stückes vorweggenommen werden. Diese Art der Vorwegnahme findet sich außerdem in Stücken von Grabbe (*Herzog Theodor von Gothland*, Panizza (*Das Liebeskonzil*) und Uhland (*Ludwig der Bayer*). In Kaisers *Stadt und Land* hingegen wird mit der Zeile "Erster Bergmann, später Michael" keine Entwicklung in der Handlung, sondern eine Veränderung der

Sprecherbezeichnung markiert. Vorwegnahmen mit Bezug auf veränderliche Beziehungen zwischen Figuren konnten nicht festgestellt werden.

Analyseszenarien

Wir stellen im folgenden zwei Analysen vor, in denen von den automatisch extrahierten Relationen Gebrauch gemacht wird, sowohl eine Einzeltext- als auch eine Korpusanalyse. Diese illustrieren Möglichkeiten, die Relationen in der Textanalyse zu berücksichtigen.

Im ersten Beispiel betrachten wir Shakespeares *Romeo and Juliet* in der derzeit auf dracor.org verfügbaren Fassung.[4] Zunächst können die Relationen visualisiert werden. Abbildung 3 zeigt das Figurennetzwerk nach Kopräsenz auf der linken und das Netzwerk, das sich aus den sozialen Beziehungen ergibt auf der rechten Seite. Zur besseren Lesbarkeit wurde ein geeigneter Layout-Algorithmus angewendet. Dabei ist zunächst interessant, dass die beiden Familien keineswegs unverbunden sind: Über Mercutio (Freund von Romeo) und Paris (Verlobter von Julia) sind beide mit dem Prinzen verbunden.

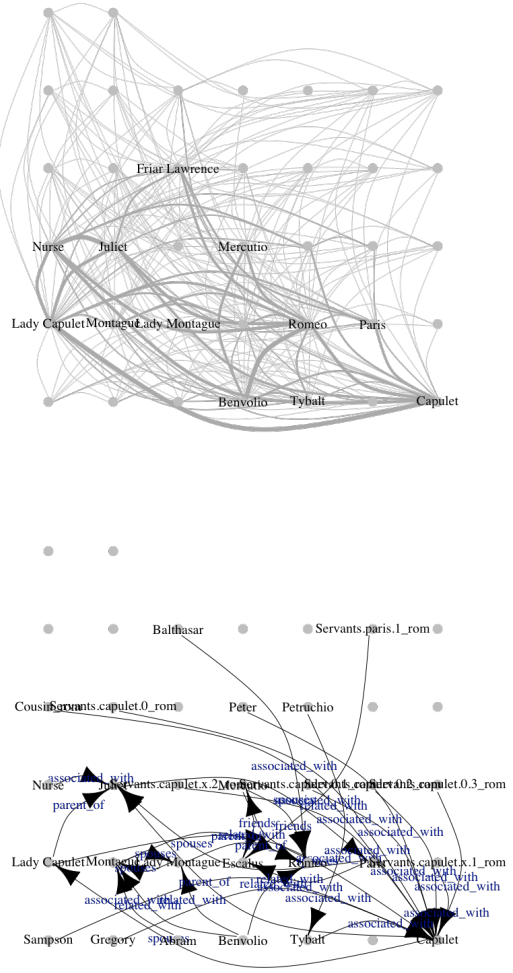


Abbildung 3: Figurennetzwerke nach Kopräsenz (oben) und Relationen (unten). Zur besseren Übersichtlichkeit wurden die Figuren auf feste Positionen gesetzt, die oben und unten gleich sind. Bezeichnungen werden nur gezeigt wenn der Grad groß genug ist (oben) oder sie an einer Beziehung beteiligt sind (unten).

Auch wenn Abbildung 3 eine gewisse Symmetrie suggeriert, ist diese keineswegs gegeben wenn wir die Redeanteile nach Familien aufschlüsseln, wie es aus den Annotationen ebenfalls direkt möglich ist. Abbildung 4 zeigt die aggregierten Redeanteile der Figuren, wobei Figuren, die durch Verwandtschaft oder Arbeitsverhältnis zu einer der Familien gehören, zusammengefasst wurden (mit Ausnahme von Mercutio und Paris, die beide mit dem Prinzen verwandt sind). Es zeigt sich, dass Angehörige der Familie Capulet etwas weniger als doppelt so viele Wörter äußern als Angehörige der Familie Montague.

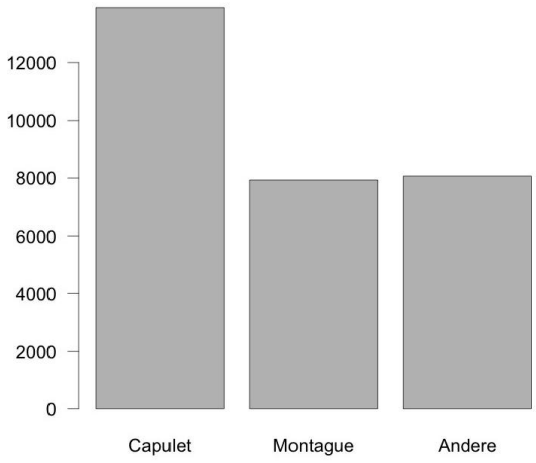


Abbildung 4: Redeanteile nach Familie

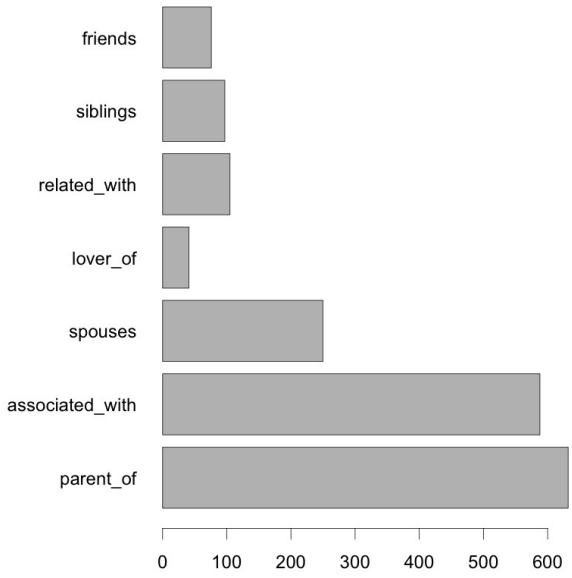


Abbildung 5: Verteilung der Relationen im Gesamtkorpus

Betrachtet man das annotierte Gesamtkorpus stellt man fest, dass die Relationen ungleich verteilt sind. Während Ehen/ Verlobungen, Elternschaft und sonstige Assoziationen relativ häufig vorkommen, spielen Geliebte, sonstige Verwandtschaften, Freundschaften und Geschwister eine vergleichsweise kleine Rolle.[5]

In Abbildung 6 sehen wir die Anzahl der Relationen bestimmter Typen ins Verhältnis gesetzt zur Großgattung (Komödie/Tragödie). Dabei wurden die Angaben auf den Titeln der Dramen übernommen und leicht vereinheitlicht (z.B. Bürgerliches Trauerspiel → Tragödie oder Zauberlustspiel → Komödie). Dabei ist zu konstatieren, dass Median und erstes Quartil bei 0 für alle Dramen bei 0 liegen: Viele Dramen weisen keine Beziehungsdefinition auf (oder sie konnten nicht automatisch identifiziert werden, siehe Fußnote). Größere oder signifikante Abweichungen zwischen den Gattungen gibt es nicht, egal welche Relation betrachtet wird. Lediglich die Relation spouses scheint im Figurenverzeichnis von Komödien häufiger genannt zu werden.

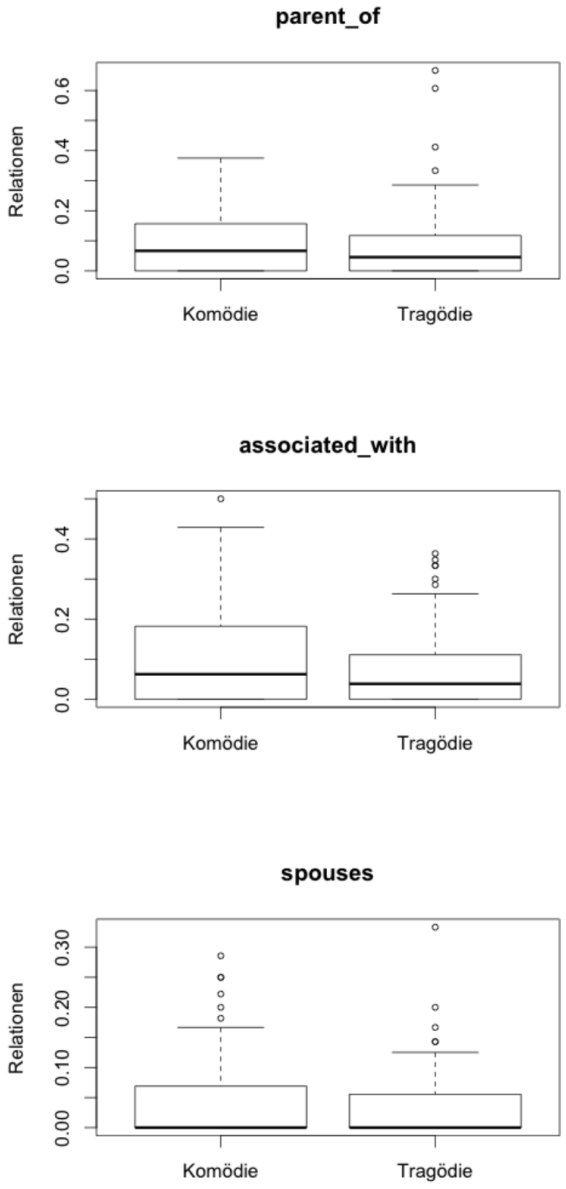


Abbildung 6: Anzahl typisierter Relationen nach Gattung

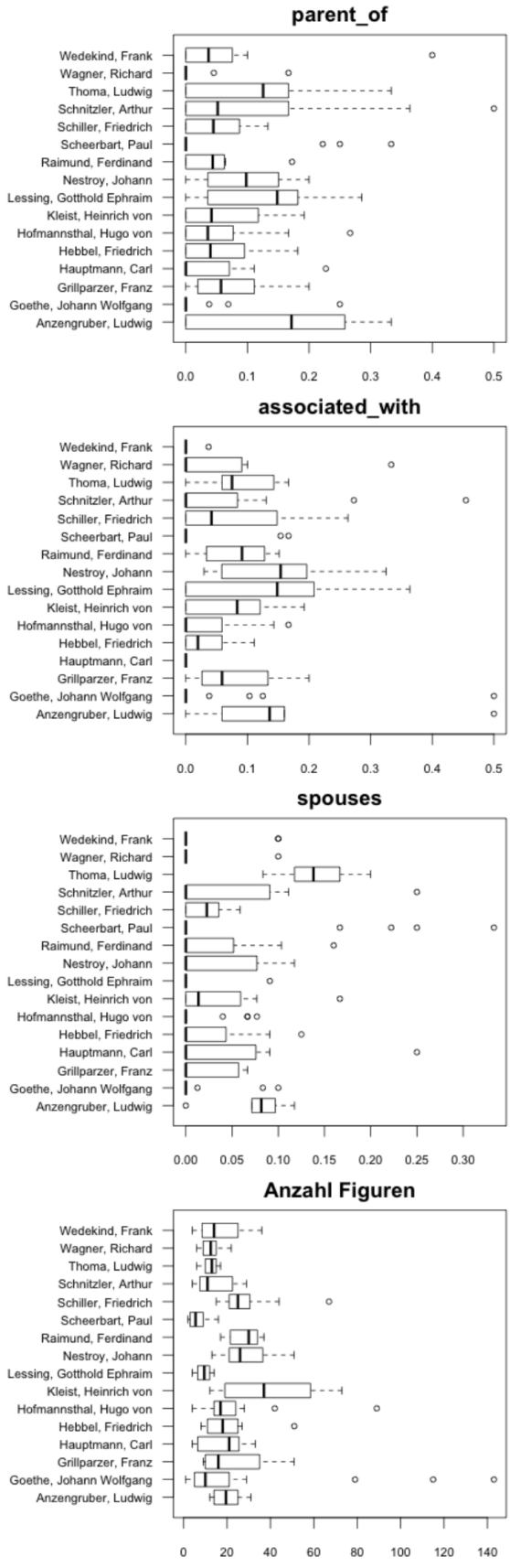


Abbildung 7: Anzahl typisierter Relationen nach Autor. Zur besseren Übersicht wurden nur Autoren berücksichtigt, die mindestens durch fünf Dramen vertreten sind

Eine Verteilung der genannten Relationen nach Autor zeichnet jedoch ein anderes Bild (Abbildung 7). Bestimmte Autoren, vor allem Ludwig Anzengruber (1839-1889) und Johann Nestroy (1801-1862), haben klare Tendenzen dazu, mehr Relationen im Figurenverzeichnis zu nennen. Beide verfassen tendenziell Possen und Komödien.

Fazit

Mit den von uns bereitgestellten maschinenlesbaren Informationen ermöglichen wir Analysen dramatischer Figuren, die die als bekannt vorausgesetzten Informationen im Figurenverzeichnis mit berücksichtigen können. Neben den oben skizzierten Analysen können die Informationen auch in inhaltliche Analysen einfließen und etwa die soziale Nähe mit der Bühnennähe korrelieren o.ä.

Kontextfreie Grammatiken haben sich hier – trotz der bekannten Schwächen im Bezug auf natürliche Sprache – als effizienter Formalismus herausgestellt, um die Figurenverzeichnisse maschinenlesbar zu machen. Wir halten dieses Verfahren für geeignet, um auch in anderen Kontexten mit semistrukturierten Textdaten zu arbeiten, wo aufgrund der begrenzten Menge ein maschinelles Lernverfahren nur bedingt zum Einsatz kommen kann.

Fußnoten

1. Beispielsweise spielt das Figurenverzeichnis im kürzlich erschienenen (Tonger-Erk, Werber, und Baum 2018), aber auch in (Genette 1993) quasi keine Rolle.
2. https://dracor.org
3. Für mehr Informationen vergleiche (Schlaffer 1972, 11)
4. https://github.com/dracor-org/shakedracor/blob/d569dc9886b3d1951f23b0454a3d7103e4cdf1bb/tei/romeo-and-juliet.xml
5. Die konkreten Ergebnisse wurden auf den *vollautomatisch erzeugten Relationen* erzielt.

Bibliographie

Asmuth, Bernhard. (2016). *Einführung in die Dramenanalyse*. Stuttgart: J.B. Metzler Verlag.

Böckenhauer, Hans-Joachim / Juraj Hromkovič (2013): *Formale Sprachen: Endliche Automaten, Grammatiken, lexikalische und syntaktische Analyse*. Zürich: Springer.

Fischer, Frank / Ingo Börner / Mathias Göbel / Angelika Hechtl / Christopher Kittel / Carsten Milling / Peer Trilcke (2019): „Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor“. In *Proceedings of DHd*. https://doi.org/10.5281/zenodo.2596095.

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

Jeßing, Benedikt (2015): *Dramenanalyse. Eine Einführung*. Berlin: Erich Schmidt Verlag.

Pangallo, Matteo (2015): „‚I will keep and character that name‘: Dramatis Personae Lists in Early Modern Manuscript Plays“. *Early Theatre* 18 (2): 87–118. https://doi.org/10.12745/et.18.2.1166.

Pfister, Manfred (2001): *Das Drama*. München: Wilhelm Fink.

Schlaffer, Hannelore (1972): *Dramenform und Klassenstruktur. Eine Analyse der dramatis persona "Volk"*. Stuttgart: J.B. Metzler Verlag.

Tonger-Erk, Lily / Nils Werber / Constanze Baum (Hrsg.) (2018): „Hauptsache Nebentext. Regiebemerkungen im Drama“. *Zeitschrift für Literaturwissenschaft und Linguistik* 48 (3).

Spiele im Spiel – Datenbankbasiertes Arbeiten zur interaktionale Sprache im Dramenwerk von Andreas Gryphius

Eggert, Lisa
lisa.eggert@uni-due.de
Universität Duisburg-Essen, Deutschland

Müller, Melissa
melissa.mueller@uni-hamburg.de
Universität Hamburg

Die Bühnensprache des Barocktheaters – insbesondere der Trauerspiele – ist normiert. Der Alexandriner als Sprechvers mit seiner festen Anzahl an Hebungen und dem Paarreim scheint trotz der dialogischen Struktur der Texte kaum Spielräume für eine Mündlichkeit im heutigen Sinne zu lassen. Die gebundene und dadurch disziplinierte Sprache der Dramen steht in Spannung zu gängigen Vorstellungen von Mündlichkeit, gilt diese doch als spontan, wenig an Normen orientiert und in der Tendenz individuell statt standardisiert.

Welche Spielräume lässt die Versifizierung dennoch zu? Welche Formen konzeptioneller Mündlichkeit lassen sich in barocken Dramentexten finden? Lassen sich Korrelationen zwischen linguistischen Phänomenen interaktionaler Sprache und den verschiedenen Möglichkeiten der Versgestaltung feststellen? Diese und weitere Fragen werden im Rahmen des DFG-Projektes *Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive* bearbeitet.

Beschäftigt sich das Projekt insgesamt inhaltlich mit Spielräumen, die im hochnormierten Trauerspiel gefunden werden können, so nimmt der Vortrag einen strukturellen Aspekt der Projektarbeit in den Blick: die Annotation von Versmaßen und ihren Abweichungen. Dabei sollen jene Spielräume in den Blick genommen werden, die an der Schnittstelle von fachlichen Anforderungen und technischen Sachzwängen entstehen und ausgehandelt werden müssen.

Projektbeschreibung und Forschungsgegenstand

Das Ziel des genannten Projektes ist es, mit Hilfe einer annotierten Datenbank (auf der Basis der Datenbankarchitektur ANNIS3; http://annis-tools.org/ – Krause/Zeldes (2016)), die das vollständige Dramenwerk von Andreas Gryphius enthält, die oben ausgeführten und weitere Fragen korpusbasiert aus literatur- und sprachwissenschaftlich übergreifender Sicht zu klären.

Hierbei wird interaktionale Sprache verstanden als Form sprachlichen Handelns, das unabhängig von seiner medialen Realisation (mündlich oder schriftlich) durch die gemeinschaftliche Erzeugung von Bedeutung von zwei oder mehr Sprecher*innen gekennzeichnet und sequenziell organisiert ist.

Das Projektkorpus besteht aus sämtlichen Dramen von Andreas Gryphius (1616-1664). Die Verschriftlichung folgt der historisch-kritischen Dramen-Ausgabe Eberhard Mannacks (1991). Die Annotation findet mithilfe des Annotationstools INCEpTION (https://inception-project.github.io/) statt.

Durch die Annotation von sowohl linguistischen als auch literaturwissenschaftlichen Phänomenen kann ein umfassenderer Blick auf interaktionale Sprache geworfen werden, da hier die sprachlichen Merkmale konzeptioneller Mündlichkeit mit literarischen Darstellungselementen konsequent enggeführt werden. Mit der Anwendung einer korpusbasierten Methodik werden aus literaturwissenschaftlicher Sicht neuartige empirische Analysemöglichkeiten erprobt, indem beispielsweise systematisch nach Regiebemerkungen, Sprecherwechseln, Versmaßen oder Reimphänomenen gesucht werden und Verteilungen erhoben werden können.

Neben den jeweiligen teilfachspezifischen Fragestellungen eröffnet das Projekt die Möglichkeit zu übergreifenden Untersuchungen, indem die zuvor jeweils disziplinspezifisch vorgenommenen Annotationen in der Abfrage miteinander kombiniert werden. Hier ergeben sich folgende mögliche Fragestellungen: (i) Welche aus der synchron orientierten Forschung zur konzeptionell mündlichen Sprachverwendung bekannten Phänomene (vgl. Schwitalla 2006. Fiehler 2016, Hennig 2006) lassen sich in den Dramentexten überhaupt lokalisieren? (ii) Welche tauchen dagegen nicht auf? Welche Gründe lassen sich für das Auftreten bzw. Ausbleiben der Phänomene finden? (iii) Wie korrelieren v.a. im hoch normierten Bereich des Trauerspiels konzeptionelle Mündlichkeit und artifizielle Versifikation?

Datenerhebung und -aufbereitung

In Vorbereitung auf die Analyse waren drei Arbeitsschritte notwendig:

1. Datenerhebung
2. Datenaufbereitung
3. Datenexport zur weiteren Analyse

Datenerhebung

Im Rahmen der Datenerhebung wurden die Texte als Abschrift digitalisiert. Dazu wurden Vorgaben erarbeitet, wie mit

Zeilen- und Versschreibweise verfahren und formale Besonderheiten verschriftlicht wurden. Nach der Abschrift wurden die Daten in ein Standarddateiformat gebracht werden, sodass die Daten annotiert werden können.

Datenaufbereitung

Gängige Tokenisierungsdienste, wie sie beispielsweise WebLicht bereitstellt, sind dabei primär auf wohlgeformte Satzstrukturen oder gesprochen-sprachliche Daten ausgelegt. Die dem Projekt zu Grunde liegenden Daten stellen allerdings eine Mischung von Vers- und Prosatexten mit Besonderheiten der Literatur des 16. Jh. dar, wodurch eine individuelle Tokenisierung notwendig wurde. Der hier entwickelte Tokenizer teilt die Sätze zeilenbasiert ein und sieht die Interpunktion somit nicht als maßgeblich für das Satzende an. Außerdem werden Satz- und Sonderzeichen separiert und der Dateiexport ist in einem validen tcf-Format möglich.

Auf Basis der Rohdaten im TCF-Dateiformat wurden mittels des DTA::CAB Web Service weitere Annotationen hinzugefügt. CAB („Cascaded Analysis Broker“) wurde vom Deutschen Textarchiv speziell für die linguistische Analyse historischer Texte entwickelt, um historische Schreibvarianten auf äquivalente „kanonische“ moderne Wortformen abzubilden. Somit werden die Daten in einem Schritt um Lemmatisierung, PoS-Tagging (STTS) und eine Normalisierung (orthography correction) ergänzt.

Die so angereicherten Daten wurden zunächst mit webAnno im späteren Verlauf mit dem auf webAnno-basierenden Tool INCEpTION korrigiert und auf verschiedenen literatur- und sprachwissenschaftlichen Ebenen annotiert.

Datenexport

Bislang gibt es zwischen dem ANNIS- und TSV3-Dateiformat, welches dem CoNNL-Format ähnlich ist, noch keine Konvertierungsmöglichkeiten. Um die annotierten Daten zur weiteren Analyse in ANNIS übertragen zu können, wird auf die Pepper-Bibliothek zurückgegriffen, welche in Zusammenarbeit mit der HU Berlin um einen Import von TSV3 erweitert wurde.

Spielräume

Das Projekt sieht sich in mehrfacher Hinsicht mit Fragen nach Spielräumen konfrontiert.

1. Wie einleitend erwähnt, ist es gerade die Suche nach Abweichungen und Spielräumen, die das Projekt inhaltlich bestimmt. Hierzu wird nach der linguistischen wie literaturwissenschaftlichen Strukturanalyse für prominente Stellen auch die jeweilige Redesituation betrachtet und so die formale und inhaltliche Untersuchung enggeführt. Ferner lassen sich aber auch über das Gesamtkorpus hinweg statistische Aussagen bspw. über die Häufigkeit von Abweichungen im Versmaß oder das Auftauchen bestimmter Formen von Gesprächspartikeln treffen. Erste Ergebnisse zum Versmaß werden im Rahmen des Vortrags vorgestellt.
2. Darüber hinaus betrifft die Frage nach Spielräumen und Interpretation auch das Design von Tagsets. Im Gegensatz zur erprobten linguistischen Methodik sind bislang kaum literaturwissenschaftliche Verfahren zur empirischen Erhebung und Interpretation interaktionaler Sprache eingeführt. Aus diesem Grund wurde in der Projektarbeit ein Set von Annotationskategorien sowie Kriterien der literaturwissenschaftlichen Auswertung solcher Phänomene neu entwickelt und kritisch reflektiert. Wesentlich sind dabei die etablierten Kategorien der Dramenanalyse, Rhetorik und Stilistik (Heudecker/Wesche 2009) sowie der Metrik. Zurückgegriffen werden kann in der methodischen Exploration auf Erkenntnisse und gängige Verfahren der Computerphilologie (Jannidis 2010a, b; Jannidis/Smith 2013), der Figuren- und Dialoganalyse im Drama (Pfister 2001: 196-264) sowie der Abweichungspoetik und Spielraumanalyse zur Barockzeit (Wesche 2004).

Dabei mussten die Phänomene so gewählt werden, dass sie einerseits interaktionale Momente der Dramentexte erfassen und andererseits für die manuelle Annotation handhabbar sind. Während sich Interaktionalität aus literaturwissenschaftlicher Sicht mit verschiedensten Kategorien auf unterschiedlichen Ebenen eines Textes beschreiben lässt, werden die Möglichkeiten der Analyse durch Annotations- und Analysetools technisch limitiert. So lassen sich beispielsweise bestimmte Phänomene der Personenkonstellation oder des Drameninhalts nur eingeschränkt einzelnen Token zuordnen und dadurch in einem Tool wie INCEpTION schwer annotieren. Auch erlaubt ANNIS zwar komplexe Suchanfragen, kann aber bestimmte Aspekte der Textgestaltung wie Einrückungen oder ähnliches nicht darstellen. Der Vortrag gibt hier einen kurzen Überblick über die Tagsets der beiden Projektteile.

1. Die Wahl der Tools unterlag hierbei allerdings insofern Sachzwängen, als diese für die linguistische Auswertung notwendig sind. So mussten innerhalb der Möglichkeiten, die diese Tools bieten, Spielräume und Lösungen gefunden werden, die für beide Projektteile zufriedenstellend sind.
2. Schließlich mussten um eine handhabbare und möglichst konsistente Annotation zu gewährleisten Guidelines entwickelt werden, die den Annotator*innen eine eindeutige Zuordnung einzelner Tags erlauben. Hier entstand häufig ein Trilemma, das auch Evelyn Gius, Nils Reiter und Marcus Willand in ihrem Shared-Task 2017 zur Erprobung von Guidelines (https://sharedtasksinthedh.github.io/) beschrieben haben: Die Kriterien „Konzeptionelle Angemessenheit“ (Wird aus der Guideline ersichtlich um welches Phänomen es geht? Orientiert sich die Guideline an gängigen Definitionen? Sind die in der Guideline beschriebenen Phänomene der Komplexität des literaturwissenschaftlichen Konzepts angemessen?), „Anwendbarkeit“ (Wie einfach kann die Guideline angewendet werden? Wie hoch ist das Inter-Annotator-Agreement?) und „Nützlichkeit“ (Wie können die nach der Guideline annotierten Daten weiterverwendet werden? Welche Fragen kann man an die Annotation stellen – und wie?) lassen sich nicht gemeinsam in gleich hohem Maß realisieren.

Am Beispiel der Annotation von Versmaßen und sowie deren Abweichungen zeigt der Vortrag, wie im Rahmen der Projektarbeit mit den letzten beiden Punkten verfahren wurde. Hierzu werden im Anschluss an eine knappe Projektvorstellung ausgewählte Phänomene, die annotiert werden, kurz er-

läutert. Danach wird der Workflow der Guideline-Entwicklung für die metrische Gestaltung der Dramentexte dargestellt. Da sich das Projekt zum Zeitpunkt des Vortrages in der letzten Phase befindet, kann hier neben Ergebnissen auch eine kritische Rückschau gezeigt werden, die mit Blick auf andere Projektentwicklungen fruchtbar sein kann.

Bibliographie

Fiehler, Reinhard (2016): Gesprochene Sprache. In: Wöllstein, A. und P. Eisenberg (Hrsg.): Duden - die Grammatik. Berlin: Dudenverlag 1181-1260.

Gryphius, Andreas (1991): Dramen. Hrsg. v. Eberhard Mannack. Frankfurt/M.: Bibliothek deutscher Klassiker.

Hennig, Mathilde (2006): Grammatik der gesprochenen Sprache in Theorie und Praxis. Kassel: Kassel Univ. Press.

Heudecker, Sylvia / Wesche, Jörg (2009): Rhetorik und Stilistik der deutschsprachigen Länder in der Zeit des Barock. In: Fix, U., A. Gardt und J. Knape (Hrsg.): Rhetorik und Stilistik. Bd. 1. Berlin: de Gruyter 97-112.

Jannidis, Fotis (2010a): Digital Editions on the Net. Perspectives for Scholarly Editing in a Digital World. In: Schäfer, J. und P. Gendolla (Hrsg.): Beyond the Screen. Bielefeld: transcript 543-560.

Jannidis, Fotis (2010b): Methoden der computergestützten Textanalyse. In: Nünning, V. (Hrsg.): Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Stuttgart: Metzler 109-132.

Jannidis, Fotis / Smith, Kathleen (2013): The Specification of User Requirements in the Design of Virtual Research Environments for the Arts and Humanities: Programming for the Future? In: Lossau, N. und H. Neuroth (Hrsg.): Evolution der Informationsinfrastruktur. Glückstadt: Verlag Werner Hülsbusch 71-84.

Krause, Thomas / Zeldes, Amir (2016): *ANNIS3: A new architecture for generic corpus query and visualization.* in: Digital Scholarship in the Humanities 2016 (31)). http://dsh.oxfordjournals.org/content/31/1/118 [zuletzt abgerufen am 14.06.2018]

Pfister, Manfred (2001): Das Drama. Theorie und Analyse. München: Fink.

Schwitalla, Johannes (2012): Gesprochenes Deutsch. Berlin: E. Schmidt.

Wesche, Jörg (2004): Literarische Diversität: Abweichungen, Lizenzen und Spielräume in der deutschen Poesie und Poetik der Barockzeit. Tübingen: Niemeyer.

Spielräume bei der retroperspektivischen Analyse der Wittgenstein-Edition und die Herausforderungen für das Semantic Clustering

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
Ludwig-Maximilians Universität München

Ullrich, Sabine
sabine.ullrich@campus.lmu.de
Ludwig-Maximilians Universität München

Still, Sebastian
sebastian.still@campus.lmu.de
Ludwig-Maximilians Universität München

Pichler, Alois
Alois.Pichler@uib.no
Wittgenstein Archiv Universität Bergen/Norwegen

Einleitung

Seit 2010 kooperieren das Wittgenstein Archiv an der Universität Bergen (WAB, Alois Pichler) und das Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians Universität München (CIS, Max Hadersbeck et. al.) in der Forschungsgruppe „Wittgenstein Advanced Search Tools" (WAST). Die WAST-Projektgruppe entwickelt die webbasierte FinderApp WiTTFind (http://wittfind.cis.lmu.de/), die einen computerlinguistisch gestützten digitalen Zugang zu WABs Wittgenstein-Edition erlaubt. Nach einer kompletten Neuscannung des Nachlasses und intensiven Verhandlungen des WAB mit den Rechteinhabern, dürfen seit 2018 WABs Edition auf der WiTTFind-Webseite durchsucht und Faksimileextrakte dargestellt werden. Nun konnten wir uns einer zentralen Frage der Wittgensteinforscher widmen: Wo finden sich in seinem Nachlass semantisch ähnliche Bemerkungen und, retroperspektivisch betrachtet, wann fanden diese Änderungen statt?

Wir entwickelten das Analysetool WiTTSim (Ullrich, 2018), das semantisch ähnliche Bemerkungen in der Edition aufspürt, zusammen mit einem vorgeschalteten semantischem Clusterverfahren (Ullrich, 2019), welches die Rechenzeit der Ähnlichkeitssuche um den Faktor 100 verkürzte. Zur retroperspektivischen Analyse der Edition entwickelten wir ein

zeitorientiertes, textgenetisches Datenmodell, das die Spielräume der Interpretation der bisher dokumentorientierten Edition auf zugelassene Lesarten reduziert.

In unserem Vortrag stellen wir die Verfahren unserer Ähnlichkeitssuche mit vorgeschaltetem semantischen Clustering und ein neues mehr textgenetisch- als dokumentorientiertes Modell einer Edition vor, das im Web-Frontend des Odysseee-Readers (www.odysseereader.wittfind.cis.lmu.de) implementiert ist und auch die Frage beantwortet: „Wann gibt es semantisch ähnliche Bemerkungen".

Die Datenbasis: Dokument- und Zeitorientierte Modelle

Die bei uns verwendete Datenbasis BNE 2015- und IDP 2016-, die am Wittgensteinarchiv an der Universität Bergen (Pichler, WAB) erstellt werden, enthalten Faksimile und Transkriptionen (auf der Basis von XML-TEI-P5) des Nachlasses von Ludwig Wittgenstein. Dieser Nachlass umfasst ca. 20.000 Seiten, welche vom WAB in Dokumente und diese wiederum in logische Textabschnitte unterteilt sind. Jeder der 54.930 Textabschnitte – eine sogenannte Bemerkung – wird mit einer eindeutigen Bezeichnung, dem sogenannten Siglum, versehen und wird in unserer Ähnlichkeitssuche als einzelnes Textobjekt definiert und semantisch analysiert.

Betrachtet man die Annotationen der BNE unter dem Aspekt der Retroperspektive, taucht folgendes Problem auf: Die BNE liefert nur auf der Ebene der Bemerkungen Informationen über ihren Erstellungszeitpunkt bzw. -zeitrahmen. Die Änderungen auf Wort und Zeichenebene sind zwar akribisch annotiert, allerdings fehlt die zeitliche Information wann diese Änderungen vorgenommen wurden. Um textgenetische Metainformationen auf Wort- bzw. Zeichenebene in das "ordered hierarchy of content objects model data" (OHCO) einer XML-Edition, wie das der BNE zu integrieren, schlägt das TEI-P5 Konsortium Fragmentierungs-, Milestone oder Standoff-Markup Annotationen vor (Jörg Hornschemeyer, 2013), die am WAB bisher nicht durchgeführt wurden. Von Geisteswissenschaftlern, deren wissenschaftliches Kerngebiet im Allgemeinen weit entfernt von der XML-Programmierung liegt, würde großer programmtechnischer Editionsaufwand verlangt. Eine Folge ist, dass von „Nachverwertern" der Edition zur Generierung der textlichen Varianten algorithmisches Ausmultiplizierten der annotierten Varianten implementiert wird, was z.B. in der Wittgenstein-Edition bei einzelnen Bemerkungen eine vierstellige Anzahl von Lesarten generiert. Betrachtet man die so automatisch generierten Lesarten, sind die meisten syntaktisch und semantisch falsch, was fatale Auswirkungen auf semantische Analysen der Textobjekte hat. Ohne zusätzliche, fein granulierte Metainformation in den annotierten Varianten sind die Spielräume der automatisierten Lesartengenerierung jedoch nicht einzugrenzen.

Im Umfeld der Wittgensteinforschung gibt es eine Edition, die bis auf Zeichenebene zeitliche Informationen zur Textgenese liefert: Die Prototractatus-Tools (PTT 2016) von Martin Pilch (Pilch 2018). Sie dokumentieren den Nutzern Ludwig Wittgensteins Schreibprozess, beginnend mit einem leeren Notizbuch im Jahre 1915 und bis zum endgültigen Diktat des Ts-204 im Sommer 1918, das zu seiner einzigen philosophischen Veröffentlichung zu Lebzeiten, der „Logisch-Philosophischen Abhandlung" führte. Leider konnten wir die Daten und Metainformationen der PTT-Edition in unserer FinderApp Infrastruktur nicht direkt analysieren, da unsere WiTTFind Infrastruktur zum einen auf das dokumentorientierte XML-TEI-P5 Datenformat aus Bergen zugeschnitten ist, und zum anderen die PTT-Edition im inkompatiblen Microsoft Word-97 Format vorliegt. Alle verfügbaren XML-TEI Importtools erfassten nur Bruchteile der Annotationen, sodass z.B. die Zeitinformationen der PTT überhaupt nicht erkannt und transformiert wurden. Um möglichst viel von der PTT-Textedition weiterzuverwenden, und damit der PTT-Hg. die Edition in seiner gewohnten Microsoft-Office Umgebung weiter optimieren kann, entwickelten wir eine mit Microsoft EXCEL leicht zu bedienende mehrdimensionale Tabellenstruktur. Die Editionsdaten und Metainformation der Word-97 Edition konnten wir größtenteils mit eigenen Programmen und Office-Macrotechniken transferieren. Zur Integration der Tabellen in die Infrastruktur unserer FinderApp verwendeten wir Libre-Office-Tools und selbst geschriebene Python Programme, die die Daten, sobald sie in das git-Repository des Projekts kopiert werden, mit Hilfe der continuous Integration automatisch transformieren und importieren. Zur Web-Präsentation werden sie an unsere neu entwickelte FinderApp, den OdysseeReader (siehe Abb. 1, odysseereader.wittfind.cis.lmu.de), übergeben. Dieses Vorgehen trennt zwar das Daten- und Repräsentationsmodell, jedoch entwickelten wir ein positionsinvariantes Siglensystem, bestehend aus dem Tupel (Zeitstempel, Dokument, Seite, Zeile, Zeichenposition), das die beiden Modelle eineindeutig verknüpft. Diese bijektive Relation zwischen den beiden Modellen definiert dem Hg., wo er in seinem Datenmodell Änderungen vornehmen muss um sie an eine bestimmte Stelle, zu einem bestimmten Zeitpunkt im Repräsentationsmodell zu platzieren.

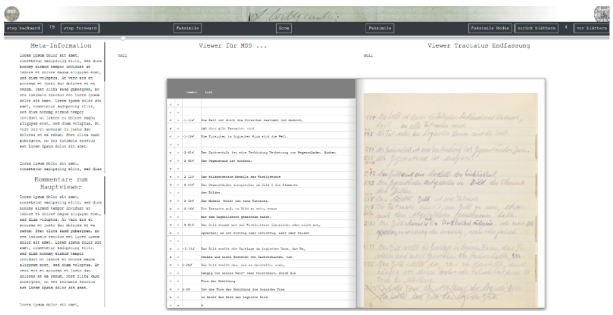


Abbildung 1: Der OdysseeReader odysseereader.wittfind.cis.lmu.de

Ähnlichkeitssuche mit vorgeschaltetem Semantic Clustering

Die Ähnlichkeitssuche WiTTSim berechnet mit Hilfe computerlinguistischer Methoden für jede Bemerkung einen „charakteristischen" Vektor, oder, intuitiv gesprochen: Man bestimmt einen "Fingerabdruck". Dieser automatisierte Prozess wird unabhängig im Voraus berechnet, was spätere Prozesse vereinfacht und beschleunigt. Dieser „Fingerabdruck" beinhaltet linguistische Informationen, wie beispielsweise Wörter, deutsche und englische Synonyme (aus Germanet und Wordnet), Wortarten (Treetagger) und Lemmata *(WiTTLex,* Röhrer 2019). Diese Informationen werden in binäre Vektoren übersetzt, welche insgesamt etwa 115.000 Features umfassen.

Zusätzlich zur Datenbasis wurden 471 Bemerkungen bereits gruppiert, also mit Ground Truth Labels versehen. Die Gruppen bestehen dabei aus 2-15 Bemerkungen und das gelabelte das Korpus umfasst 1.670 Bemerkungen, was ca. 3% des gesamten Nachlasses entspricht.

Zur Semantischen Ähnlichkeitsberechnung ist allerdings eine Reduktion des Feature Raumes zwingend nötig, da die Vektoren mit so hoher Dimensionalität semantisch „weit voneinander entfernt" sind und keine semantischen Gruppierungen auszumachen sind. Dieses Phänomen ist auch bekannt als *Curse of Dimensionality*. Daher werden die Vektoren zunächst auf eine angemessene Anzahl von Features skaliert, um sie anschließend clustern zu können. Verwendete Reduktionstechniken umfassen Singular Vector Decomposition (SVD), Principal Component Analysis (PCA), Sparse Random Projection (SRP) und Uniform Manifold Approximation and Projection (UMAP). Auf unseren Daten zeigte eine SVD Reduktion zu 1.600 Dimensionen die besten Ergebnisse, zusammen mit UMAP, welches darüber hinaus die Daten im zweidimensionalen Raum klar gruppiert. Letzteres erlaubt nur eine Zieldimension von 2 bis 100 Dimensionen, weshalb zum Erhalt der Varianz die maximale Dimensionsanzahl von 100 gewählt wurde, um einen bestmöglichen Erhalt der gespeicherten Information zu gewährleisten.

Nach erfolgter Reduktion der Dimension können die Datenpunkte, also alle Bemerkungen, geclustert werden. Verwendete Clustering Techniken umfassen den klassischen K-Means Ansatz (Mac-Queen 1967, Ball and Hall 1956, Lloyd 1982, Steinhaus 1955), aber auch Dichte-basierte Ansätze wie Mean-Shift (Duda und Hart 1973) und DBSCAN (Ester et al. 1996), das statistische Gaussian Mixture Modell (Redner und Walker 1984) und das hierarchische Ward Clustering (Ward 1963). Beste Ergebnisse konnten mit einer Kombination von SVD und K-Means mit einer Anzahl an k=150 Clustern erzielt werden. Evaluiert wurde anhand der drei unüberwachten Metriken Silhouette Score, Davies Bouldin Index, und Calinski-Harabasz Index. Zusätzlich konnte durch die verfügbaren Ground Truth Labels auch der Recall berechnet werden, welcher in den Experimenten einen maximalen Wert von 1,0 erreicht. Dies zeigt, dass alle der gelabelten Daten richtig zugeordnet werden konnten. Wird eine Suchanfrage zum Auffinden ähnlicher Bemerkungen gestartet, muss nur der charakteristische Vektor der eingegebenen Bemerkung berechnet werden und das nächstliegende Cluster bestimmt werden. Letzteres erfolgt durch eine Bestimmung des am nächsten gelegenen Cluster Mittelpunkts (Zentroids). Anschließend werden die Abstände zu allen Bemerkungen des bestimmten Clusters gemessen, welche zuletzt dem Philologen zur genaueren Prüfung „gerankt" vorgeschlagen werden.

Zusammenfassung und Ausblick

Unsere zeitgesteuerte textgenetische Edition kann von einem Wissenschaftler ohne XML Kenntnisse innerhalb einer Office Umgebung erstellt werden. Das continuous Integration System von git transferiert die Edition automatisch in unser WEB-basiertes Repräsentationssystem, den OdysseeReader. Über das von uns entwickelte eineindeutige Siglensystem verliert der Hg. niemals den klaren Zusammenhang zwischen Editions- und Präsentationsmodell.

Das von uns entwickelte Ähnlichkeitstool mit vorgeschaltetem Semantic Clustering könnte auch zur Ähnlichkeitsbestimmung zwischen zwei gegebenen Texten verwendet werden: Der Nutzer könnte einen Text eingeben, und es werden potentiell ähnliche Textpassagen in einer Sammlung von Texten gesucht, die dann „gerankt" nach Ähnlichkeiten in einer Art Hitliste ausgegeben werden. Eine derartige Sortierung nach Textähnlichkeiten könnte es dem Philologen zum Beispiel besonders erleichtern, potentielle Zitate, Einflüsse und Verweise eines Autors innerhalb seines Werkes und im Bezug auf die Literatur seiner Zeit aufzuspüren.

Bibliographie

Ball, Geoffrey H. / Hall David J. (1965): *Isodata, a novel method of data analysis and pattern classification*. Technical report, Stanford research inst Menlo Park CA.

Duda, Richard O. / Hart, Peter E. (1973): "Pattern analysis and scene classification." *J. Wiley* 1:73.

Ester, Martin / Kriegel Hans-Peter / Sander, Jörg / Xu, Xiaowei et al. (1996): „A density-based algorithm for discovering clusters in large spatial databases with noise.", in *KDD*, volume 96, pages 226–231.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Oyvind (2014): Wittgenstein's Nachlass: WiTTFind and Wittgenstein Advanced Search Tools (WAST), DATeH, Madrid.

Hadersbeck, Maximilian / Still, Sebastian (2018): *Investigating Wittgenstein's Nachlass: WiTTFind, WiTTReader, OdysseeReader and Wittgenstein Advanced Search Tools*, im Katalog zur Ausstellung „DIE TRACTATUS ODYSSEE" S.127-137, Wittgenstein Initiative, Wien.

Lloyd, Stuart P. (1982): „Least squares quantization in pcm", in: *IEEE transactions on information theory*, 28(2):129–137.

MacQueen, J. B. (1967): „Some methods for classification and analysis of multivariate observations.", in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967

Pichler, Alois / Krüger, Heinz W. / Smith, D. / Bruvik, Tone / Lindebjerg, Anne / Olstad, Vemund (Hrsg.) (2009): Wittgenstein Source Bergen Facsimile (BTE). Wittgenstein Source Bergen.

Redner, Richard A. / Walker, Homer F. (1984): Mixture densities, maximum likelihood and the em algorithm. SIAM review, 26(2):195–239.

Röhrer, Ines / Ullrich, Sabine / Hadersbeck, Maximilian (2019): *Weltkulturerbe international digital: Erweiterung der Wittgenstein Advanced Search Tools durch Semantisierung und neuronale maschinelle Übersetzung*. multimedial multimodal. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 25. - 29.03.2019 an den Universitäten zu Mainz und Frankfurt.

Steinhaus, Hans (1955): Quelques applications des principes topologiques à la géométrie des corps convexes. Fund. Math, 41:284–290.

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).

Ullrich, Sabine (2019): *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*. Master's thesis. LMU.

Pilch, Martin (2018): *Frontverläufe im Prototractatus – Zur gedanklichen Entwicklung von Krakau bis Sokal (1914/1915)*,

Wittgenstein-Studien 9 (S.101-154), Internationale Ludwig Wittgenstein Gesellschaft (ILWG).

Still, Sebastian (2018): *Ludwig Wittgenstein: 100 Jahre Traktatus. Der Odyssee-Reader, ein web-basiertes Tool zur textgenetischen Suche im Traktatus*, Masterthesis, Ludwig-Maximilians-Universität München.

Feldweg, Birgit / Feldweg, Helmut (1997): „GermaNet - a Lexical-Semantic Net for German.", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

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, pp. 2228-2235.

Hörnschemeyer, Jörg / Thaller, Manfred / Förtsch, Reinhard (2017): *Textgenetische Prozesse in Digitalen Editionen*, Köln Universitäts- und Stadtbibliothek Köln 2017, https://www.worldcat.org/title/textgenetische-prozesse-in-digitalen-editionen/oclc/1002260195

Schmidt, Alfred (2018): „Ludwig Wittgenstein's Nachlass in the UNESCO Memory of the World register.", in: *Nordic Wittgenstein Review* 7(2):209–213.

UNESCO (2017): UNESCO-Weltdokumentenerbe - Zwei Neuaufnahmen. URL: https://www.unesco.at/presse/artikel/article/unesco-weltdokumentenerbe-zwei-neuaufnahmen/ [letzter Zugriff 19. Juni 2018].

Ward, John H. (1963): „Hierarchical grouping to optimize an objective function." *Journal of the American statistical association* 58.301: 236-244.

Spielräume definieren: Cooking Recipes of the Middle Ages

Steiner, Christian

christian.steiner@uni-graz.at
Universität Graz, Österreich

Klug, Helmut W.

helmut.klug@uni-graz.at
Universität Graz, Österreich

Kochtraditionen, ob regional oder international, sind eine der herausragendsten Elemente der europäischen Kultur und ein wichtiger Bestandteil der europäischen Identität. Aber die Fragen nach ihrem Ursprung, den Einflüssen und ihrer Entwicklung sind nach wie vor unklar. In den letzten Jahrzehnten kam die Forschung zu zwei wichtigen Schlussfolgerungen, welche mittlerweile die Forschungsbestrebungen prägen: Erstens gibt es keine quantitativen Studien über den Ursprung und die Entstehung der regionalen Küche in Europa; zweitens sind erst ab dem Mittelalter Handschriften mit Tausenden von Kochrezepten überliefert, was wohl als die Geburt der modernen europäischen Küche angesehen werden kann (vgl. Flandrin & Hyman 1988, Laurioux 2005). Auf dem europäischen Kontinent stellen frühneuhochdeutsche, mittelfranzösische und mittellateinische Rezepte den größten Teil der kulinarischen Überlieferung dar, die mehr als 80 Manuskripte und etwa 8000 Rezepte umfasst. Das Projekt "Cooking Recipes of the Middle Ages" bereitet die Kochrezeptüberlieferung von Frankreich und dem deutschsprachigen Raum auf, um ihre Herkunft, ihre Beziehung und ihre Migration innerhalb Europas zu analysieren (vgl. thematisch ähnliche Studien mit unterschiedlicher Fokussetzung: Hieatt 1995 mit linguistischem Fokus, Flandrin 1984, Adamson 1995, Carlin 1989, Adamson 2002, Karg 2007 allgemein und van Winter 1989, Hyman 2005, Laurioux 2002 mit Fokus auf spezielle Gerichte). Die Partner, das Zentrum für Informationsmodellierung der Universität Graz und das Laboratoire CESR (Centre d'Etudes Supérieures de la Renaissance) der Universität Tours werden diese mehrsprachigen Texte mit Hilfe von digitalen Standards aufarbeiten und sie mit aktuellen quantitativen und qualitativen Forschungsmethoden untersuchen. Der Vergleich der französischen und deutschen Ernährungsgeschichte eignet sich besonders gut für diese Aufgabe, da Frankreich einen kulturell prägenden Einfluss auf die deutschsprachigen Völker hatte.

Kochrezepte sind kulturell aufgeladene, flüchtige Texte; die erhaltenen Niederschriften stellen daher nur ein punktuelles Zeugnis, eine individuelle Zubereitungsweise eines Gerichts (Hieatt 1985, 26) in Raum und Zeit dar. Das inhaltliche Verständnis dieser Rezepte, ihre möglichen Entstehungs- und Anwendungskontexte und ihre Überlieferung ist zudem kein einfacher Prozess, denn die Fachbegriffe, Zutaten, Utensilien, Verfahren und Bräuche der damaligen Zeit, die in den Rezepten eher öfter implizit als direkt genannt werden, sind auch für Sprach- und Geschichtswissenschaftler, die sich auf das Thema spezialisiert haben, immer wieder eine Herausforderung. Ihre Entwicklung sollte daher am besten diachron und räumlich analysiert werden, was mittlerweile mit digitalen geisteswissenschaftlichen Methoden verhältnismäßig leicht möglich ist – vorausgesetzt, die entsprechenden Daten liegen vor. Im aktuellen Projekt werden die historischen Texte auf mehreren Ebenen erschlossen: So werden die Texte nicht nur neu transkribiert und philologisch-editorisch bearbeitet (vgl. Klug, Kranich 2015), sondern auch in unterschiedlichen Wissensgebieten semantisch angereichert. Das schafft jene Spielräume, die nötig sind, um Analysen wie maschinengestützte Abgleiche von Zutaten, Kochprozessen oder Kochutensilien, die Suche nach Rezepttradition und -migration oder standardisierte philologische Vergleiche, wie z.B. Kollationierungen, durchzuführen. Die Basis unserer Daten sind customized TEI/XML-Dokumente mit einem zusätzlichen adaptierten Schema, das die semantische Annotation von Kochrezepten im Allgemeinen erleichtern soll.

Die Rezeptüberlieferung – in Form einzelner Rezeptsammlungen – wird mithilfe einer <tei:msDesc>, die sich an den Handschriftenbeschreibungen renommierter Bibliotheken orientiert und die in Kooperation mit der Abteilung für Sondersammlungen der Grazer Universitätsbibliothek entstanden ist, räumlich und zeitlich fixiert. Besonderes Augenmerk wird dabei auf Informationen zur Handschriftenentstehung (materiell wie auch inhaltlich) und auf die Schriftbeschreibung bzw. den Schreibhandbefund der Rezeptsammlungen gelegt, wobei erstere Informationen meist den Katalogen entstammen, letztere aus der Arbeit mit den Texten kommen. Die Grundlage des Projekts ist die einheitliche Erfassung der überlieferten Texte durch eine hyperdiplomatische Neutranskription der historischen Quellen: Als Arbeitsumgebung fungiert Transkribus[1], wo das Textlayout automatisch erkannt und die Texte manuell mittels proprietären Codierungen erfasst werden. Mithilfe mehrerer Transformations-

schritte wird aus den Rohdaten die Basis für die elektronische Quellenabbildung erstellt, die sich an germanistisch-editorischen Richtlinien orientiert. Die Quellentexttranskription verzeichnet dabei nicht nur das unterschiedliche Schriftzeicheninventar, sondern auch alle textstrukturierenden Elemente. Das gesamte Zeicheninventar ist in einer nach den Richtlinien der TEI erstellten Zeichenbeschreibung erfasst. Die Beschreibung stützt sich dabei auf die theoretischen Ergebnisse zur Beschreibung von Zeichen aus dem DigiPal-Projekt[2] und verwendet außerdem die Zeichenidentifikatoren der Medieval Unicode Font Initiative[3] (vgl. Böhm, Klug 2020). Die so produzierten Daten sind nicht nur der Ausgangspunkt für die wissenschaftlichen Fragestellungen im Projekt, sondern bieten eine solide Grundlage für viele weitere Forschungsfragen aus Germanistik/Linguistik, Paläographie usw. Diese Erarbeitungsstufe wird nach editorischen Richtlinien normalisiert und gibt im Rahmen des Webauftritts in Form einer Text-Bild-Synopse detaillierten Einblick in die historische Quelle.

Aus den Transkriptionsdaten wird außerdem eine auf Zeichenebene normalisierte, in Sinneinheiten untergliederte Textfassung geschaffen, in der die semantischen Informationen annotiert werden. Diese Sinneinheiten umfassen neben dem eigentlichen Rezept und Rezepttitel Eingangs- und Schlussformeln, Handlungsanweisungen, Küchen- und Serviertipps, Hinweise auf medizinische und religiöse Aspekte und selbstverständlich Zutaten, Gerichte und Küchenutensilien.

Den Kern der digitalen Forschungsstrategie bildet das Semantic Web beziehungsweise die Anbindung und Integration unserer Daten an Linked Open Data. Wir sind innerhalb der Geisteswissenschaften in der vorteilhaften Position, dass sich unser Projekt zu einem großen Teil mit Lebensmittelzutaten befasst, d.h. mit Tieren, Pflanzen und Pilzen. Das sind Forschungsgebiete, in denen sich bereits eine signifikante Menge an relevanten Ontologien etabliert haben und die gut an die Linked Open Data Cloud, einschließlich der allgemeinen Wissensdatenbanken Wikidata und DBPedia angeschlossen sind[4]. Ontologien werden, wenn auch mit unterschiedlichen Schwerpunkten und Granularität der Daten, außerdem bereits erfolgreich für die Repräsentation von Kochrezepten (Hoehndorf & Lange 2018, Sam et al. 2014, Ribeiro et al. 2006) und in deren Analyse eingesetzt (Chow & Grüniger 2019, Jovanovic et al. 2015, Vadivu & Waheeta Hopper 2010). In unserem digitalen Forschungsansatz setzen wir zwar teilweise auch auf Textähnlichkeiten, der größte Teil unserer Analyse basiert jedoch auf dem Vorkommen von Zutaten, Kochprozessen bzw. Zubereitungshinweisen und Kochutensilien. Weitere Entitäten, die wir für die Analyse der Rezepte heranziehen sind Serviervorschläge sowie medizinische, kulturelle und religiöse Aspekte in den Texten. Die Annotation dieser Entitäten gestaltet sich schon aufgrund ihrer schieren Menge in historischen Kochrezepten als sehr komplex. Neben den Möglichkeiten zuvor unbekannte Beziehungen zwischen den Quellen und deren Entitäten zu finden, war das Arbeiten außerhalb von Sprachbarrieren ein Hauptargument für die Entscheidung, Semantic Web-Technologien in den Mittelpunkt des Projekts zu stellen.

Durch die Verwendung von *Konzepten,* im Sinne einer Idee oder eines mentalen Bildes und nicht eines *Begriffes*, versuchen wir, historische und sprachliche Grenzen zu überwinden. Ein konkretes Beispiel für diese Diskrepanz zwischen Begriff und mentaler Vorstellung liefert uns die österreichische / süddeutsche Variante für Kartoffel: "Erdapfel" ("erdaphel" im Frühneuhochdeutschen) wird etwa in einem Manuskript aus der Zeit um 1488 erwähnt, lange bevor die Kartoffel von Südamerika nach Europa importiert wurde, was uns zeigt, dass das Konzept von "Erdapfel" ein anderes gewesen sein muss (wahrscheinlich jede Art von Rübe) als das heutige Konzept des Erdapfels. Wie oben bereits erwähnt, war es also nötig einen Workflow zu finden, der nicht nur die philologische, sondern auch die semantische Komplexität der Rezepte widerspiegelt. Während die phrasenartigen Informationseinheiten manuell annotiert werden, erfolgt die Annotation auf Wortebene semiautomatisch, indem die Texte mithilfe von XSLT- und Python-Skripten und individuellen Vokabularien, vorgehalten als CSV Dateien, angereichert werden, die alle darauf ausgerichtet sind, die Varianz der historischen Sprachstufen auszugleichen. Für die frühneuhochdeutschen Texte stand uns aus einem früheren Projekt eine Liste mittelalterlicher Pflanzennamen und ihrer Übersetzungen in modernes Englisch und Deutsch sowie ihrer mittelalterlichen Variantendiktionen zur Verfügung[5]. Dies gab uns die Möglichkeit, mit Hilfe der von OpenRefine bereitgestellten Reconciliation Service API[6], einen teilautomatisierten Prozess zur Annotation von Wikidata-Konzepten zu starten. Die daraus resultierenden Daten bildeten den Grundstock für die zuvor genannten Vokabularien. Ähnliche Listen wurden von den Projektpartnern in Frankreich erstellt, die mithilfe des von den französischen Kollegen entwickelten Tools “Heterotoki”[7] in einem kollaborativen Arbeitsschritt konsolidiert werden können. Sobald jeder Begriff mit einem Konzept verbunden ist, werden diese Konzepte verwendet, um die Zutaten innerhalb der eigentlichen Rezepttexte in den TEI-Dokumenten anzureichern. Ein entscheidender Faktor dieses semiautomatischen Prozesses bleibt jedoch die menschliche Interpretation der angereicherten Einheiten und die Entscheidung für ein konkretes bereits bestehendes Konzept bzw. die Erstellung eines neuen Konzepts in Wikidata.

Wir befinden uns derzeit mitten in dieser semantischen Annotationsphase. Ist diese abgeschlossen, bieten sich mannigfaltige Analysemethoden an. Sobald die Einheiten der einzelnen Rezepte mit Konzepten ausgestattet sind, kann die Analyse des Projekts übereinstimmende oder abweichende Essgewohnheiten, Textmigration sowie den Einfluss der Nachbarländer auf die jeweilige Küche aufzeigen. Die Implementierung von Ontologien aus den Naturwissenschaften wie FoodOn[8] oder SNOMED[9] ermöglicht es uns, Verbindungen von historischen Essgewohnheiten zu modernen Konzepten von Lebensmitteln herzustellen und neues Wissen für den Bereich der Ernährungsgeschichte zu generieren. Die Ontologiedaten werden zusammen mit den Entitäten in einem Triplestore gespeichert und können mit Hilfe von SPARQL Queries befragt werden. Die Ergebnisse dienen als Grundlage für eine räumliche und zeitliche Visualisierung der Daten.

Die Speicherung, Analyse und Dissemination der Projektdaten erfolgt über das vom Zentrum für Informationsmodellierung in Graz entwickelte Repository GAMS (Geisteswissenschaftliches Asset Management)[10]. Innerhalb dieser auf Langzeitarchivierung ausgerichteten Infrastruktur wird auf den Triplestore “Blazegraph”[11] über einen Webservice zur Speicherung und Abfrage von RDF-Triples zugegriffen.

Fußnoten

1. https://transkribus.eu/Transkribus/
2. Describing Handwriting I-VII; http://www.digipal.eu/blog
3. https://folk.uib.no/hnooh/mufi/

4. Für eine Übersicht an Ontologien in diesen Bereichen siehe: http://www.ontobee.org/, http://aims.fao.org/, https://ndb.nal.usda.gov/ndb/, https://agclass.nal.usda.gov/about.shtml, http://zbw.eu/stw/version/latest/thsys/70498/about.de.html.

Alle sind an die Linked Open Data Cloud angeschlossen indem eine oder mehrere Serialisierungen in OWL und/oder RDF(S) vorliegen.

5. http://medieval-plants.org
6. https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API
7. https://github.com/ponchio/heterotoki
8. http://foodon.org/
9. https://browser.ihtsdotools.org/
10. https://gams.uni-graz.at
11. https://www.blazegraph.com/

Bibliographie

Adamson, M. W. (Ed.) (1995): Food in the Middle Ages. A Book of Essays. New York, London: Garland. Adamson, M. W. (Ed.). (2002). Regional Cuisines of Medieval Europe: A Book of Essays. New York, London: Routledge.

Amoia, M. / Martínez, J.M.M. (2019): SaCoCo Diachronic Corpus [WWW Document]. URL http://fedora.clarin-d.uni-saarland.de/sacoco/ (accessed 1.7.20).

Böhm, A. / Klug, H.: Quellenorientierte Aufbereitung historischer Texte im Rahmen digitaler Editionen: Das Problem der Transkription in mediävistischen Editionsprojekten. In: [Titel steht noch nicht fest] Hrsg. von Ingrid Bennewitz und Martin Fischer (= Bamberger interdisziplinäre Mittelalterstudien.) [in Vorbereitung]

Carlin, M. / Rosenthal, J. T. (Eds.) (1998): Food and Eating in Medieval Europe. London: Hambledon Press.

Chow, A. E. / Ninger, M. G. (o. J.): *Multimodal Event Recognition with an Ontology For Cooking Recipes*. 12.

Dooley, D. M. / Griffiths, E. J. / Gosal, G. S. / Buttigieg, P. L. / Hoehndorf, R. / Lange, M. C., / ... / Hsiao, W. W. L. (2018): FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Science of Food*, *2*(1), 23. https://doi.org/10.1038/s41538-018-0032-6

Flandrin, J.-L. (1984): «Internationalisme, nationalisme et régionalisme dans la cuisine des XIVe et XVe siècles: le témoignage des livres de cuisine». In Manger et boire au Moyen âge. Actes du Colloque de Nice (15-17 octobre 1982). (pt. 2, p. 75-91). Paris.

Flandrin, J.-L. / Hyman, P. (1988): "Regional tastes and cuisines: Problems, documents, and discourses on food in Southern France in the 16th and 17th centuries". Food and Foodways 1-3, p. 221-251.

Gloning, T., (2000): Monumenta Culinaria et Diaetetica Historica. Corpus of culinary & dietetic texts of Europe from the Middle Ages to 1800. Corpus älterer deutscher Kochbücher und Ernährungslehren [WWW Document]. URL http://www.staff.uni-giessen.de/gloning/kobu.htm (accessed 1.7.20).

Hammad, R. / Hassouna, M. (2011): *Multi-Language Semantic Search Engine*. 6.

Hieatt, C. (1995): Sorting through the Titles of Medieval Dishes: What Is, or Is Not, a "Blanc manger". In M. W. Adamson (Ed.), Food in the Middle Ages. A Book of Essays. (pp. 25-43). New York, London: Garland.

Hyman, P. / M. (2005): «Les associations de saveurs dans les livres de cuisine français du XVIe siècle». In *Le Désir et le Goût. Une autre histoire (XIIIe-XVIIIe siècles). Actes du colloque international à la mémoire de Jean-Louis Flandrin* (Saint-Denis, septembre 2003). Dir. Odile Redon, Line Sallman et Sylvie Steinberg. (p. 135-150). Saint-Denis: Presses Universitaires de Vincennes.

Karg, S. (Ed.) (2007): Medieval Food Traditions in Northern Europe. Copenhagen: National Museum of Denmark.

Klug, H. W. / Kranich, K. (2015): "Das Edieren von handschriftlichen Kochrezepttexten am Weg ins digitale Zeitalter. Zur Neuedition des Tegernseer Wirtschaftsbuches." In T. Bein (Ed.), Vom Nutzen der Editionen. Zur Bedeutung moderner Editorik für die Erforschung von Literatur- und Kulturgeschichte. (pp. 121-137). Berlin, Boston: De Gruyter.

Lamé, M. / Pittet, P. (2018): *Heterotoki: Non-st / ructured and heterogeneous terminology alignment for Digital Humanities data producers*. 12.

Laurioux, B. (2005): «Les voyageurs et la gastronomie en Europe à la fin du Moyen âge ». In Le Désir et le Goût. Une autre histoire (XIIIe-XVIIIe siècles), Actes du colloque international à la mémoire de Jean-Louis Flandrin (Saint-Denis, septembre 2003). Dir. Odile Redon, Line Sallman et Sylvie Steinberg. (p. 99-117). Saint-Denis, Presses Universitaires de Vincennes.

Ribeiro, R. / Batista, F. / Pardal, J. P. / Mamede, N. J. / Pinto, H. S. (2006): Cooking an Ontology. In J. Euzenat & J. Domingue (Hrsg.), *Artificial Intelligence: Methodology, Systems, and Applications* (S. 213–221). Springer Berlin Heidelberg.

Sam, M. / Krisnadhi, A. A. / Wang, C. / Gallagher, J. / Hitzler, P. (2014): *An Ontology Design Pattern for Cooking Recipes – Classroom Created*.

Vadivu, G. / Hopper, S. W. (2010): Semantic Linking and Querying of Natural Food, Chemicals and Diseases. *International Journal of Computer Applications*, *11*(4), 35–38. https://doi.org/10.5120/1567-2093

van Winter, J. M. (1989): "Kochen und Essen im Mittelalter." In B. Herrmann (Ed.). Mensch und Umwelt im Mittelalter. (pp. 88-100). Frankfurt am Main: Fischer Taschenbuch Verl.

Spielräume modellieren. Eine digitale Edition von Giovanni Domenico Tiepolos Bildzyklus Divertimento per li Regazzi auf der Grundlage von CIDOC CRM

Tumanov, Rostislav
rostislav.tumanov@ikg.uni-stuttgart.de
Universität Stuttgart, Deutschland

Viehhauser, Gabriel
viehhauser@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Feldmann, Alina
feldmann.alina@gmail.com
Universität Stuttgart, Deutschland

Koller, Barbara
barbarakoller94@web.de
Universität Stuttgart, Deutschland

Einleitung

Kunstwerke eröffnen Spielräume für Interpretationen. Das gilt auch und gerade für den Ende des 18. Jahrhunderts vom venezianischen Maler Giovanni Domenico Tiepolo geschaffenen Bildzyklus *Divertimento per li Regazzi*, der verschiedene Motive aus dem Umkreis der aus der *Commedia dell'Arte* bekannten Figur des Pulcinella aufgreift und diese spielerisch verbindet. Die einzelnen Blätter des Zyklus scheinen eine Art Lebensgeschichte Pulcinellas zu erzählen und sich immer wieder aufeinander zu beziehen oder Werke anderer Künstler aufzugreifen, ohne dass jedoch letzte Gewissheit zu erlangen ist. Die Vieldeutigkeit und der Beziehungsreichtum haben daher ganz unterschiedliche Interpreten bis hin zum bekannten Philosophen Giorgio Agamben (Agamben 2015) auf den Plan gerufen und zu spielerischen Assoziationen verleitet.

Gleichzeitig wird am Bildzyklus jedoch deutlich, wie sehr formale Zwänge Spielräume einengen und Deutungsmöglichkeiten festlegen können. In dem von Adelheid Gealt herausgegebenen Katalog (Gealt 1986), der bis heute einen zentralen Ausgangspunkt für die wissenschaftliche Betrachtung des *Divertimento* bildet, werden die Bilder in einer festen Abfolge dargeboten, die eine bestimmte sukzessive Lektüre des Zyklus vorgibt, hinter der mögliche Querbezüge zurücktreten.

Diese Festlegung ergibt sich nicht zuletzt durch die Zwänge des Druckmediums, in dem eine lineare Abfolge von Bildseiten unvermeidlich ist. Wie in der Diskussion insbesondere im Bereich der literaturwissenschaftlichen Editorik deutlich gemacht wurde, eignen sich digitale Editionen besonders gut dazu, solche Einschränkungen zu überwinden und unterschiedliche Sichtweisen zuzulassen (Sahle 2013, 2, 166-189). Um die vielfältigen Bezugsmöglichkeiten, die das *Divertimento* bereithält, adäquat darzustellen, bietet es sich daher an, das Konzept einer offenen, digitalen Edition auch auf die Präsentation des Bildzyklus zu übertragen. An der Universität Stuttgart wird zurzeit eine solche Edition des *Divertimento* erarbeitet. Der innovative Methodentransfer von Text- auf Bilderzählungen greift damit gleichsam aus einer umgekehrten Stoßrichtung aktuelle Überlegungen zur graphbasierten Modellierung von (Text-)Editionen auf (Andrews / Mace 2013, Efer 2017, Kuczera 2016, Kuczera 2017) und vermag dadurch neues Licht auf die Modellierung von Erzähllogiken zu werfen sowie zugleich neue Wege der digitalen Darstellung von Bildkunstwerken zu eröffnen.

Bei der Einrichtung der Edition ist zu beachten, dass auch digitale Methoden formalen Zwängen unterworfen sind. Denn um die Interoperabilität und Nachhaltigkeit zu sichern, ist der Rückgriff auf ein standardisiertes Datenmodell notwendig, das den Austausch mit anderen Ressourcen erlaubt. Diese Modellierung bringt eine Einengung mit sich (Pierazzo 2019), die paradoxerweise gerade erst jene Spielräume eröffnet, in denen ein Bildzyklus wie das *Divertimento* rekontexualisiert werden kann. Um dessen komplexen Möglichkeitsräume freizulegen, greifen wir auf das im Kulturerbe-Bereich verbreitete Datenmodell des CIDOC-CRM zurück (Le Boeuf et al. 2019). Im Folgenden wollen wir zunächst die kunsthistorische Fragestellung unseres Projekts explizieren, danach dessen Workflow skizzieren und schließlich genauer auf die Datenmodellierung in CIDOC-CRM eingehen, die unter dem Aspekt der Paradoxie des Digitalen zwischen Eröffnung und Einschränkung von Spielräumen diskutiert werden kann.

Kunsthistorische Voraussetzungen

Das von Tiepolo in den letzten Jahren seines Lebens zwischen 1797 und 1804 geschaffene Korpus besteht aus 104 Zeichnungen, die allesamt die Figur des Pulcinella in verschiedenen Lebenssituationen zeigen. Nimmt man eine von Blatt zu Blatt anhaltende personelle Konstanz der äußerlich kaum unterscheidbaren Pulcinelli an, so kann man in dem Zyklus ein biographisches Narrativ erkennen, das den Weg einer oder vielleicht auch mehrerer Pulcinella-Figuren von der Wiege bis zur Bahre nachzeichnet.

Aufgrund zahlreicher variierender Wiederholungen oder sich offensichtlich widersprechender Darstellungen ist es jedoch nahezu unmöglich, alle Zeichnungen zu einer konzisen Erzählung zu verbinden (Vetrocq 1979, 19-20). So wird Pulcinella etwa in zwei Darstellungen auf unterschiedliche Weise hingerichtet, während ein weiteres Blatt seine Begnadigung zeigt. Im Gegensatz zu einem ‚gewöhnlichen' narrativen Bildzyklus gibt es also keine klar erkennbare Anordnung der Blätter, keinen eindeutig intendierten Erzählverlauf. Stattdessen wird dem Betrachter ein weiter narrativer Spielraum eröffnet, innerhalb dessen eine Vielzahl unterschiedlicher Geschichten konstruiert werden können (Gealt 1986, 16-17, Gottdang 2015, 81-86). Der Rezipient rückt somit an die Stelle des Erzählers bzw. Mitspielers, der einzelne Zeichnungen aussuchen und zu verschiedenen Sequenzen anordnen kann. Doch wird auch hier deutlich, dass es sinnvoll sein kann, Spielräume einzuschränken: Nicht jede denkbare Sequenz ist aus erzähllogischer Perspektive betrachtet gleich valide (Tumanov 2019). Im Sinne einer biographischen Anordnung erscheint es beispielsweise angebracht, Blätter, die Pulcinella als Kind darstellen, vor solche zu ordnen, die ihn als Erwachsenen zeigen.[1]

Zu dieser syntagmatischen Ordnung der Lebensgeschichte des Pulcinella kommt im Zyklus eine paradigmatische Dimension hinzu, die sich durch die Wiederkehr von bestimmten Motiven ergibt. So begegnet etwa auf mehreren Darstellungen ein Topf mit Gnocchi, der Bezüge zwischen den einzelnen Zeichnungen herstellt, die quer zum biographischen Narrativ liegen.

Diese 'internen' paradigmatischen Bezüge werden schließlich noch durch eine Reihe ‚externer' Verweise auf andere Bilder, Motive oder Praktiken ergänzt, die den Assoziationshorizont skizzieren, vor dem die Zeichnungen zu verstehen sind. So verweist beispielsweise die Szene der Erschießung des Pulcinella aufgrund einer Motivähnlichkeit auf die Erschießung der marodierenden Soldaten aus den *Grandes Misères de la gu-*

erre Jacques Callots, wodurch sich neue Bedeutungsdimensionen an den Zyklus anlagern lassen.

Aus dieser Ausgangslage ergibt sich folgendes Anforderungsprofil für die digitale Edition:

- Ziel der Edition ist es, explizit keine einheitliche Bildsequenz zu (re)konstruieren, sondern die Ordnungsstrukturen und polyvalenten Verknüpfungen aufzuzeigen, die in den Zeichnungen angelegt sind und es dem Betrachter ermöglichen, diverse Lektürewege durch den narrativen Möglichkeitsraum des *Divertimento* einzuschlagen.
- Dazu ist eine Datenmodellierung nötig, die das Material nach erzähllogischen Gesichtspunkten vorstrukturiert ohne eine genaue Abfolge festzulegen.
- Die Datenmodellierung soll neben syntagmatischen auch paradigmatische Bezugspunkte berücksichtigen.
- Neben internen sind auch externe Verweispunkte zu berücksichtigen, die auf das Semantic Web ausgreifen.

Technische Umsetzung

Zur Formalisierung dieser Zusammenhänge im digitalen Medium erscheint uns ein Graphdatenmodell besonders geeignet. Die möglichen internen und externen Verbindungen zwischen den Bildern werden demnach in RDF-Triples ausgedrückt.[2] Diese Triples werden in einem RDF-Triplestore (zurzeit kommt hierfür GraphDB[3] zum Einsatz) vorgehalten und sollen mittels des php-Frameworks ARC2[4] über SPARQL-Abfragen[5] ausgelesen und in eine HTML-Darstellung ausgegeben werden;[6] geplant ist zudem eine alternative Ausgabe in IIIF-Manifeste.[7] Die Verbindung ins Netz des *Linked Data* soll durch die Verwendung von Normdaten wie dem Iconclass-Klassifikationssystem[8] ermöglicht werden (zu Chancen und Herausforderungen beim Einsatz von Normdaten in der digitalen Kunstgeschichte vgl. Kailus / Stein 2018).

Kernstück ist dabei die Modellierung der Daten in einer vom CIDOC-CRM vorgegebenen Struktur, die die Interoperabilität und Nachhaltigkeit der Daten gewährleistet. Konkret wurde dabei auf die OWL-basierte Version Erlangen-CRM zurückgegriffen[9] und die Modellierung mit Hilfe des Ontologie-Editors Protégé ausgearbeitet.[10] CIDOC-CRM etabliert sich im DH-Bereich immer mehr als Top-Level-Ontologie (Eide / Ore 2019). Für unsere Zwecke bietet sich das CIDOC-CRM jedoch nicht nur aufgrund seines Status als Standard an, sondern auch aufgrund seiner *Event*-basierten Struktur, die der Modellierung des Bildzyklus besonders entgegen kommt.

Modellierung in CIDOC-CRM

Die Herausforderung der Modellierung bestand insbesondere darin, eine flexible Anordnung der Bilder zu erlauben, um verschiedenen Gruppierungsmöglichkeiten nach zeitlichen oder inhaltlichen Aspekten gerecht zu werden. Berücksichtigt werden musste, dass die Bilder zwar einer grundlegenden Chronologie folgen (z. B. ereignet sich die Geburt Pulcinellas vor dessen Tod), dass aber auch indifferente Einzeldarstellungen existieren, die nicht in eine eindeutige Reihenfolge gebracht werden können (z. B. verschiedene von Pulcinella ausgeübte Berufe). Zusätzlich können sie auch nach sich wiederholenden Bildmotiven gruppiert werden. Darüber hinaus sollten innerhalb des Bilderzyklus auftretende externe Bezüge auf weitere Werke Tiepolos sowie Arbeiten anderer Künstler berücksichtigt werden.

CIDOC-CRM bietet für diese Aufgabe zwei Vorteile: Zum einen ist CIDOC-CRM *Linked-Open-Data*-kompatibel, zum anderen „ereignisorientiert", d. h., Ereignisse können prozesshaft modelliert werden. Dadurch wird der Aufbau einer flexiblen chronologischen Struktur ermöglicht, die sich als Gerüst nicht an der statischen Abfolge der Bilder, sondern an einer gewissermaßen 'virtuellen' Lebensgeschichte Pulcinellas orientiert, anhand derer der Bildzyklus gruppiert werden kann. Der Lebenszyklus (E4_Period) lässt sich in Untergruppen segmentieren, die als Ereignisse (E5_Events) angelegt wurden, die sich wiederum in einzelne Handlungen (E7_Activities) aufteilen lassen. Diese Ereignisse und Handlungen können mit Hilfe von *Properties* in eine chronologische Reihenfolge gebracht werden (so steht etwa der Abschnitt 'Geburt_und_Kindheit_Pulcinellas' in der Beziehung P120_occurs_before zum Abschnitt 'Pulcinellas_Berufe'). Innerhalb einer Ereignisgruppe bleiben die Einzelbilder ohne weitere chronologische Anordnung frei sortierbar oder können durch die Verknüpfung einzelner Bildhandlungen in eine näher spezifizierte Chronologie gebracht werden. Dadurch lassen sich die grundlegende zeitliche Einteilung, zeitlich nur grob einteilbare Handlungen, aber auch spezifische chronologische Abfolgen modellieren, so dass variable Handlungswege ohne starres „Abfolgekorsett" möglich werden (Abbildung 1).

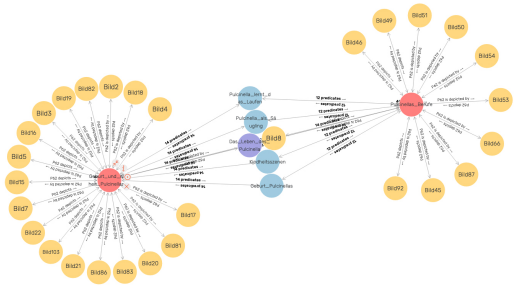


Abbildung 1: Graphische Darstellung der Modellierung (Ausschnitt), erstellt mit GraphDB. Dem roten Event-Knoten ‚Geburt_und_Kindheit_Pulcinellas' links sind die entsprechenden Bilder (gelb) zugeordnet, die Darstellungen aus diesem Lebensabschnitt zeigen. Verbindungen ergeben sich zu einzelnen Activity-Knoten (blau), der gesamten Lebensgeschichte (violett) und schließlich zum nachfolgenden Event‚Pulcinellas_Berufe' (roter Knoten) mit den jeweiligen Bildern. Bild 8, das den Titel‚Der junge Pulcinella beobachtet Landarbeiter' trägt, ist beiden Abschnitten zugeordnet.

Um die Wiederholung von einzelnen Bildmotiven zu modellieren, werden diese in den entsprechenden Bildern als vorhanden ausgezeichnet (P62_depicts) und mit der entsprechenden Iconclass-Nummern versehen, die einen ersten Anschlusspunkt an das Semantic Web ermöglichen sollen (Abbildung 2).

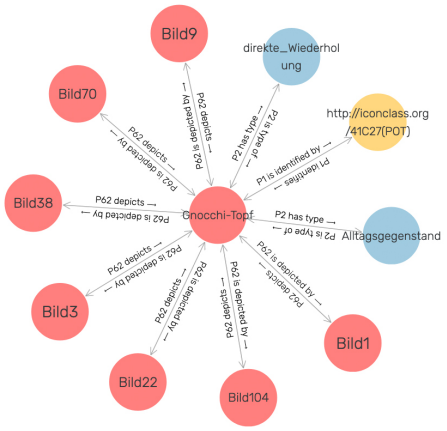


Abbildung 2: Graphische Darstellung der Modellierung (Ausschnitt), erstellt mit GraphDB. Im Zentrum steht das Motiv des‚Gnocchi-Topfs', das in unterschiedlichen Bildern dargestellt ist, die quer zur zeitlichen Abfolge stehen. Das Motiv ist zudem mit dem Iconclass-Identifier versehen, der das Ausgreifen ins Semantic Web ermöglicht.

Bezugnahmen auf externe Kunstwerke wurden durch die Anlegung eines eigenen '*Individuals*' für jedes dieser externen Werke umgesetzt, die anschließend mit den Bildern des Zyklus verknüpft werden (P130_shows_features_of).

Fazit

Als grundlegende Struktur für die Modellierung von Tiepolos variablen Bildzyklus hat sich das CIDOC-CRM aufgrund seiner *Event*-basierten Konzeption erstaunlich praktikabel erwiesen. Wie sich zeigt, sind mit dem Gerüst des CIDOC-CRM alle zuvor skizzierten Anforderungen an die Datenmodellierung des Editionsprojekts umsetzbar, ohne dass Modifikationen nötig wären. Das durch das CRM aufgespannte Beziehungsgeflecht kann in weiterer Folge mittels SPARQL-Anfragen ausgelesen und über das ARC2-Framework in eine HTML-Darstellung umgesetzt werden. Gerade die sich zunächst aus der Notwendigkeit der Standardisierung ergebende formale Einschränkung, die die Modellierung mit CIDOC mit sich bringt, eröffnet damit den Möglichkeitsraum, der das komplexe Beziehungsgeflecht des *Divertimento* erst zu Anschauung bringen kann. In dieser Hinsicht offenbart das Modell den doppelten Charakter der Spielräume-Metapher, der für die Anwendung digitaler Methoden charakteristisch sein dürfte: Digitale Modelle eröffnen Spielräume, legen aber zugleich deren räumliche Grenzen fest.

Wie weit diese Spielräume geöffnet werden können, hat sich freilich auch in der praktischen Umsetzung zu erweisen. Eine Herausforderung für unser Projekt wird es etwa sein, zu bestimmen, wie 'tief' die Querverweise auf andere Kunstwerke in die Modellierung mit aufgenommen werden können. Welche Motive sind z.B. für die paradigmatische Verknüpfung relevant? Und ist Relevanz hier binär zu denken oder gibt es unterschiedliche Grade von Relevanz? Ist demnach eine 'vollständige' Erschließung relevanter Motive überhaupt zu erreichen? Es steht zu vermuten, dass auch hier die Menge der möglichen Kontextualisierungen potentiell unbegrenzt ist, aber durch die Modellierung beschränkt werden muss. Und wie offen kann wiederum das syntagmatische Gerüst gestaltet werden, ohne dass sich die Nutzerin / der Nutzer in der Vielzahl der Möglichkeiten verliert? Solche und ähnliche Fragen stellen potentielle Limitationen unseres Ansatzes dar, die zwar formal bei der Integration in CIDOC unproblematisch sind, sich aber gewissermaßen auf inhaltlicher Ebene bei der Erstellung unseres Datenmodells abzeichnen. Mit welchen Strategien mit ihnen umgegangen werden kann, hat sich letztlich im Umgang mit dem fertigen Editionsinterface zu erweisen, entsprechende Fragen nach 'Tiefe' und 'Vollständigkeit' der Modellierung stellen aber auch Diskussionspunkte dar, die über unser Projekt auf die grundlegenden Möglichkeiten der Darstellung von Bildkunstwerken hinaus verweisen.

Fußnoten

1. Damit ist nicht gesagt, dass eine biographische bzw. erzähllogische Auffassung des Zyklus die einzige Möglichkeit seiner Rezeption darstellt. Durch die Darstellung des Pulcinella in unterschiedlichen Lebensstufen wird aber ein – von der Edition zu berücksichtigendes – Grundgerüst nahegelegt, welches auch dann die Folie bietet, wenn man (wie etwa tendenziell Agamben 2015) die Atemporalität des Dargestellten betont: Denn diese lässt sich gerade erst im Spannungsfeld zum gebrochenen biographischen Narrativ herausarbeiten.
2. https://www.w3.org/RDF/
3. http://graphdb.ontotext.com/
4. https://github.com/semsol/arc2
5. https://www.w3.org/TR/rdf-sparql-query
6. Der gegenwärtige Stand der Arbeiten ist unter https://github.com/Gabvie/Divertimento dokumentiert.
7. https://iiif.io/
8. http://www.iconclass.org/
9. http://erlangen-crm.org/
10. https://protege.stanford.edu/

Bibliographie

Agamben, Giorgio (2016): *Pulcinella ovvero Divertimento per li Regazzi.* Nuova edizione. Rom: Nottetempo.

Andrews Tara L. / Macé Caroline (2013): "Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata", in: *Literary and Linguistic Computing* 28 (4): 504–521.

Efer, Thomas (2017): *Graphdatenbanken für die textorientierten e-Humanities*. Dissertationsschrift, Universität Leipzig.

Eide, Øyvind / Ore, Christian-Emil Smith (2018): "Ontologies and data modeling", in: Flanders, Julia / Jannidis, Fotis (eds.): *The Shape of Data in Digital Humanities*. Modeling Texts and Text-based Resources. London: Routledge 178-196.

Gealt, Adelheid M. (ed.) (1986): *Domenico Tiepolo. The Punchinello Drawings*. New York: George Braziller.

Gottdang, Andrea (2015): „Das Spiel vom Leben und Sterben des Punchinello: Giandomenico Tiepolos Divertimento per li Regazzi", in: Robert Fajen (ed.): *Amüsement und Risiko.* Dimensionen des Spiels in der spanischen und italienischen Aufklärung. Halle an der Saale: Mitteldeutscher Verlag 60–101.

Kailus, Angela / Stein, Regine (2018) „Besser vernetzt: Über den Mehrwert von Standards und Normdaten zur Bilder-

schließung“, in: Kuroczyński, Piotr / Bell, Peter / Dieckmann, Lisa (eds.): *Computing Art Reader.* Einführung in die digitale Kunstgeschichte. Heidelberg: 119-139

Kuczera, Andreas (2016): “Digital Editions Beyond Xml – Graph-Based Digital Editions”, in: Preiser-Kappeller, Johannes / Düring, Marten / Jatowt, Adam (eds.): *Proceedings of the 3rd Histoinformatics Workshop on Computational History (Histoinformatics 2016)*.

Kuczera, Andreas (2017): „Graphentechnologien in den Digitalen Geisteswissenschaften“, in: *ABI Technik* 37 (3): 179–196.

Le Boeuf, Patrick / Doerr, Martin / Ore, Christian Emil / Stead, Stephen et al. (2019): “Definition of the CIDOC Conceptual Reference Model”, Version 6.2.6 May 2019, http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf [letzter Zugriff 27.9.2019].

Pierazzo, Elena (2018): “How subjective is your model?”, in: Flanders, Julia / Jannidis, Fotis (eds.): *The Shape of Data in Digital Humanities*. Modeling Texts and Text-based Resources. London: Routledge 117-132.

Sahle, Patrick (2013): *Digitale Editionsformen,* Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Norderstedt: Books on Demand.

Tumanov, Rostislav (2019): „Groteskes Korpus. Theatrale, narrative und referenzielle Aspekte des *Divertimento per li Regazzi* von Giovanni Domenico Tiepolo“, in: *Wallraf-Richartz-Jahrbuch* 80, 205-239.

Vetrocq, Marcia (1979): “Domenico Tiepolo and the Figure of Punchinello”, in: Gealt, Adelheid M. (eds.): *Domenico Tiepolo’s Punchinello drawings*. Bloomington: Indiana University Art Museum.

Sprachvarietäten-abhängige Terminologie in der neuronalen maschinellen Übersetzung: Eine Analyse in der Sprachrichtung Englisch-Deutsch mit Schwerpunkt auf der österreichischen Varietät der deutschen Sprache

Heinisch, Barbara
barbara.heinisch@univie.ac.at
Universität Wien, Österreich

Einführung

Maschinelle Übersetzung hat in den vergangenen Jahren große Fortschritte gemacht. Neuronale maschinelle Übersetzung als der aktuelle Ansatz erzielt natürlichsprachliche Ergebnisse (Hassan et al. 2018; Castilho et al. 2017; Junczys-Dowmunt et al. 2016). Allerdings können diese flüssig lesbaren Übersetzungen über inhaltliche Fehlübersetzungen hinwegtäuschen (Koehn/Knowles 2017; Forcada 2017).

Für das Training eines neuronalen maschinellen Übersetzungssystems sind große Mengen an (qualitativ hochwertigen) bilingualen Sprachressourcen erforderlich. Die Sammlung und Aufbereitung dieser Sprachressourcen, wie beispielsweise Parallelkorpora, legt damit die Grundlage für die Qualität der maschinellen Übersetzung. Allerdings gibt es auch Ansätze, die keine großen Mengen an Trainingsdaten benötigen, wie z.B. die unüberwachte maschinelle Übersetzung (Artetxe et al. 2019).

Da generische maschinelle Übersetzungssysteme mit Korpora aus unterschiedlichsten Domänen trainiert werden und daher in bestimmten Fachgebieten oft nicht die gewünschte Qualität liefern (können), werden domänenadaptierte Systeme entwickelt (Chu et al. 2017; Servan et al. 2016). Neben weiteren Besonderheiten der Fachsprache (Stolze 2009), hat insbesondere die Verwendung der korrekten Terminologie einen entscheidenden Einfluss auf die Qualitätsevaluation einer (maschinellen) Übersetzung. Im Vergleich zu regelbasierten maschinellen Übersetzungssystemen (Forcada et al. 2011; Scansani et al. 2017; Simard et al. 2007), die das Einbinden von Wörterbüchern erlauben, sind die mit dem Training des künstlichen neuronalen Netzwerkes bei der neuronalen maschinellen Übersetzung verbundenen „Lernprozesse“ intransparenter. Daher ist auch die Einbindung von terminologischen Ressourcen schwieriger.

Terminologie als Herausforderung in der maschinellen Übersetzung (Wloka et al. 2013; Reynolds 2015) hat zu Lösungsvorschlägen geführt, die unter anderem auf der Ebene der neuronalen Dekodierung (Hasler et al. 2018) ansetzen oder die Form von Terminologieangaben als Annotationen im Ausgangstext annehmen (Dinu et al. 2019). Auch eine korpus- und fallbasierte Adaption eines generischen maschinellen Übersetzungssystems (Farajian et al. 2018; Servan et al. 2016) bietet die Möglichkeit, die zielsprachliche Terminologie in der maschinellen Übersetzung an eine Domäne anzupassen. Außerdem wird an der Anpassung der maschinellen Übersetzung an das individuelle Vokabular einer Person geforscht (Michel/Neubig 2018).

Neben der Anpassung eines maschinellen Übersetzungssystems an ein bestimmtes Fachgebiet (Chu et al. 2017), spielt auch die Berücksichtigung der Sprachvarietät eine entscheidende Rolle (Costa-jussà et al. 2018; Lakew et al. 2018). Denn Benennungen, die Homonyme sind, können einen unterschiedlichen Begriffsinhalt haben. Dies kann zu Missverständnissen, vor allem in der fachsprachlichen Kommunikation, zwischen SprecherInnen unterschiedlicher Varietäten einer Sprache führen. Im Bereich der maschinellen Übersetzung sind auch Dialekte, als nicht standardsprachliche Varietäten, ein Forschungsgegenstand, z.B. Dialekte im Schweizerdeutschen, die aufgrund der Vorherrschaft der gesprochenen Sprache über keine standardisierte Schreibweise verfügen (Honnet et al. 2018).

Demnach gilt es neben der Terminologie auch die Sprachvarietät beim Training eines (neuronalen) maschinellen Über-

setzungssystems zu berücksichtigen. Obwohl sich die schriftlichen standardsprachlichen Varietäten der deutschen Sprache auch in vielen Aspekten unterscheiden (Ammon 1995), so liegt der Fokus der vorliegenden Untersuchung auf dem Bereich der Lexik, insbesondere der Terminologie.

Am Schnittpunkt zwischen Translationswissenschaft, Terminologiewissenschaft und Varietätenlinguistik kann die vorliegende Untersuchung zu den (digitalen) Geisteswissenschaften Erkenntnisse zu der Verwendung von Korpora in der maschinellen Übersetzung unter dem Aspekt der sprachvarietätenabhängigen Terminologie beitragen.

Forschungsdesign

Ausgehend von der Hypothese, dass (generische) maschinelle Übersetzungssysteme überwiegend die deutsche Standardvarietät der deutschen Sprache als Übersetzungsergebnis für Terminologie ausgeben, wurde eine Analyse in der Sprachrichtung Englisch-Deutsch mit drei (frei) verfügbaren neuronalen maschinellen Übersetzungssystemen durchgeführt. Bei diesen maschinellen Übersetzungssystemen handelt es sich um Google Translate, DeepL sowie eTranslation. Die Wahl fiel auf diese drei Systeme, da die beiden erst genannten von professionellen ÜbersetzerInnen und Studierenden der Translationswissenschaft in Österreich verwendet werden (Heinisch/Lušicky 2019; Lušicky/Heinisch (unveröffentlichtes Manuskript)). Das maschinelle Übersetzungssystem eTranslation wiederum wurde von der Europäischen Kommission entwickelt und steht MitarbeiterInnen der öffentlichen Verwaltung in allen EU-Mitgliedsstaaten kostenlos zur Verfügung, wobei es einen Schwerpunkt auf EU- und EU-relevanten Texten hat (Lösch et al. 2018).

Zum Testen der Hypothese wurden drei Domänen ausgewählt, in denen sich die Standardvarietäten des Deutschen unterscheiden: Kulinarik (Schmidlin 2011), Verwaltungssprache und Rechtssprache (Wissik 2013; Lohaus 2000), wobei hier der Universitätsterminologie besondere Berücksichtigung zukommt (Heinisch-Obermoser 2014). Der Schwerpunkt wurde außerdem auf die österreichische Standardvarietät der deutschen Sprache (im Vergleich zur deutschen) in der Übersetzung aus dem Englischen gelegt. Dafür wurden einschlägige terminologische Ressourcen, die „österreichisches Deutsch" als Ausgangssprache und – sofern vorhanden – „Englisch" (unabhängig von der Varietät) als Zielsprache hatten, verwendet. In der Domäne der Kulinarik wurde die Liste der österreichischen Ausdrücke, die im Protokoll Nr. 10 über die Verwendung spezifisch österreichischer Ausdrücke der deutschen Sprache im Rahmen der Europäischen Union im Zuge des EU-Beitritts Österreichs ausverhandelt wurde, verwendet (Markhardt 2002; EU 1995). In der Domäne der österreichischen Verwaltungssprache wurde das „Fachglossar österreichische Verwaltung. Deutsch -Englisch" dem Sprachressourcenportal Österreichs entnommen (Heinisch 2018). Die Domäne Universitätsterminologie, die sich an der Schnittstelle zwischen Rechts-, Verwaltungs- und disziplinärer Wissenschaftssprache(n) befindet, konnte mit einem Auszug aus der Terminologiedatenbank der Universität Wien (UniVieTerm) (Heinisch-Obermoser 2014, 2016) abgedeckt werden. Da es sich um originäre Ressourcen handelt, weisen diese eine unterschiedliche Anzahl an Einträgen auf: Kulinarik (23 Vorzugsbenennungen), Verwaltung (695 Vorzugsbenennungen) und Universität (779 Vorzugsbenennungen). Daher gilt es bei der Interpretation der Ergebnisse der vorliegenden Studie zu beachten, dass die Domäne der Kulinarik im Vergleich zu den beiden anderen terminologischen Ressourcen deutlich weniger Benennungen umfasst.

Die Ausgangstexte für die maschinelle Übersetzung waren daher die englischsprachigen Vorzugsbenennungen sowie sämtliche weitere zulässigen Benennungen, Abkürzungen waren ausgenommen. Die Benennungen wurden in die genannten maschinellen Übersetzungssysteme eingegeben, wobei als Ausgangssprache Englisch und als Zielsprache Deutsch gewählt wurde.

Die Analyse der maschinellen Übersetzung erfolgte mittels der in der Literatur zur maschinellen Übersetzung weit verbreiteten Multidimensional Quality Metrics (MQM) (Lommel et al. 2014), die die Bestimmung von Fehlern und damit der Qualität einer (maschinellen) Übersetzung erlaubt. In der vorliegenden Studie kommt der Fehlerkategorie „Terminologie" – und hier wiederum der Unterkategorie „Inkonsistenz mit Terminologiedatenbank" – besondere Bedeutung zu.

Die Ergebnisse der maschinellen Übersetzung wurden mit der in der terminologischen Ressource genannten deutschen Vorzugsbenennung (in der österreichischen Standardvarietät) verglichen. Hierbei wurden drei Arten der Übereinstimmung unterschieden: Vollständige Übereinstimmung bedeutet, dass nur die Vorzugsbenennung (sprich die österreichische Varietät) in der Übersetzung aufscheint. Partielle Übereinstimmung bedeutet, dass bei der Angabe von zwei (oder mehr) möglichen Benennungen im Englischen (in der terminologischen Ressource angeführte Vorzugsbenennung und weitere zulässige Benennungen im Englischen) zumindest eine der Benennungen im Deutschen die österreichische Standardvarietät (Vorzugsbenennung) ist bzw. dass es geringfügige Abweichungen von der deutschen Vorzugsbenennung gibt, z.B. Groß- und Kleinschreibung, Verwendung von Bindestrichen, Ergänzungen bzw. andere Wortstellung bei Phraseologie bzw. Komposita, die dem Begriffsinhalt entsprechen usw. Keine Übereinstimmung bedeutet, dass es weder eine vollständige noch partielle Übereinstimmung gab. Dies ist der Fall, wenn entweder ausschließlich die deutsche standardsprachliche Varietät der Terminologie (und nicht die österreichische) verwendet wurde oder gänzlich andere Benennungen in der Übersetzung aufscheinen. Dies bedeutet, dass die Benennungen, die primär der deutschen Standardvarietät zugeordnet werden können, in den beiden Kategorien „partielle Übereinstimmung" und „keine Übereinstimmung" zu finden sind.

Ergebnisse

Insgesamt wurden 1497 Benennungen aus den drei genannten Domänen (Kulinarik, Verwaltung, Universität) in Form der englischen Vorzugsbenennungen (und der zulässigen Benennungen) mit den drei erwähnten maschinellen Übersetzungssystemen (GoogleTranslate, DeepL, eTranslation) ins Deutsche übersetzt. Danach wurde die maschinelle Übersetzung der Terminologie mit den deutschen Vorzugsbenennungen in den genannten terminologischen Ressourcen, die die sprachvarietätenabhängige österreichische Terminologie abbilden, verglichen. Dieser Vergleich in Abbildung 1, in der ausschließlich vollständige und keine partiellen Übereinstimmungen mit den Vorzugsbenennungen im Deutschen berücksichtigt wurden, zeigt das Ausmaß der Berücksichtigung der österreichischen Standardvarietät der deutschen Sprache (in Form der deutschen Vorzugsbenennungen in den genannten terminologischen Ressourcen).

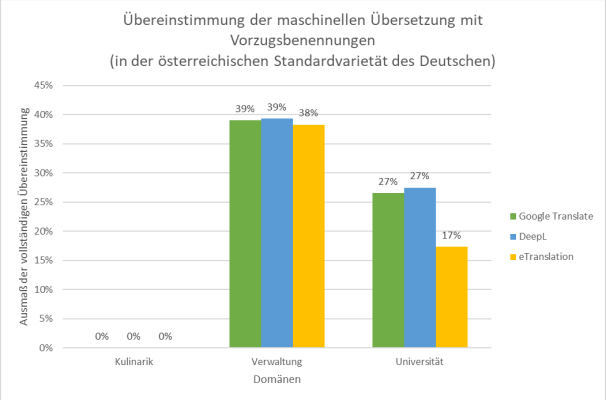


Abbildung 1: Vollständige Übereinstimmung der deutschen Übersetzung der drei verwendeten maschinellen Übersetzungssysteme mit den deutschen Vorzugsbenennungen (in der österreichischen Standardvarietät des Deutschen) in den drei terminologischen Ressourcen in den Domänen Kulinarik, Verwaltung und Universität (Prozentangaben sind gerundet)

Die Ergebnisse fallen abhängig von der Domäne unterschiedlich aus. Besonders deutlich wird dies im Bereich der Kulinarik, da hier kein einziges der untersuchten maschinellen Übersetzungssysteme die Vorzugsbenennung (im österreichischen Standarddeutsch) ausgibt. Die Analyse der Kategorien „partielle Übereinstimmung" und „keine Übereinstimmung" in der Kulinarik-Domäne zeigt, dass die deutsche Varietät in beinahe jeder Übersetzung der untersuchten maschinellen Übersetzungssysteme vorherrschend ist. Die einzigen Ausnahmen für die Benennungen Kren und Topfen fanden sich lediglich bei eTranslation. (Da diese jedoch nur durch die Verwendung von weiteren zulässigen Benennungen im Englischen (Synonyme) und in Verbindung mit der deutschen standardsprachlichen Terminologie als Synonyme aufgelistet wurden, gelten diese als partielle Übereinstimmung und scheinen nicht in Abbildung 1 auf.) Allerdings gilt es anzumerken, dass beispielsweise DeepL alternative Übersetzungen zu einer Benennung mittels Dropdown-Menü anbietet. In dieser Liste wurde in vielen Fällen die österreichische Varietät der Benennung als ein Listenelement – wenn auch teils weit unten – angezeigt. Das Ergebnis von eTranslation war überraschend, da im Protokoll 10 des EU-Beitrittsvertrags Österreichs die terminologischen Besonderheiten im Bereich der Kulinarik spezifiziert wurden und von einer umfassenderen Verwendung von österreichischen Ausdrücken hätte ausgegangen werden können.

In der Domäne der Verwaltungssprache unterschieden sich die drei untersuchten Systeme kaum (je ca. 40% vollständige Übereinstimmung). Es wurde ein Überhang der deutschen Varietät über alle drei maschinellen Übersetzungssysteme hinweg bemerkt. Nur bei wenigen Benennungen gaben alle drei Systeme zugleich die deutsche Varietät bzw. eine andere Übersetzung aus. Dies betraf in erster Linie Benennungen, die Bezirk, Gemeinde oder Magistrat enthielten. Bei anderen Benennungen waren die Ergebnisse variabel.

In der Domäne der Universitätsterminologie schließlich wurde erstmals auch ein Unterschied zwischen den maschinellen Übersetzungssystemen deutlich. GoogleTranslate und DeepL mit je 27% vollständiger Übereinstimmung gaben deutlich mehr Vorzugsbenennungen (in der österreichischen Standardvarietät) aus als eTranslation (mit 17%). Des Weiteren war grundsätzlich ein deutlicher Überhang der deutschen Standardvarietät bei den untersuchten Benennungen zu beobachten. Außerdem gab es in dieser Domäne auch die meisten Ambiguitäten in den Übersetzungen, da es unter anderem deutschsprachige Benennungen (oder Paraphrasierungen) gab, die nicht der Universitätsterminologie zugeordnet werden können und daher „keine Übereinstimmungen" darstellen. Als Beispiel hierfür kann das *Sammelzeugnis* (*transcript of records*) genannt werden, das mit *Abschrift von* bzw. *der Aufzeichnungen* oder *Transkript der Aufzeichnungen* übersetzt wurde. Interessanterweise übersetzten zwei der drei Systeme die auch im Englischen auf Deutsch belassene Funktion *Studienpräses* (entweder mit *Studienpräsenzen* oder mit *Studienbeihilfen*). Daher ist davon auszugehen, dass vor allem in dieser Domäne Missverständnisse durch die maschinelle Übersetzung entstehen können. Das schlechte Ergebnis in der Domäne der Universitätsterminologie kann allerdings auch auf die verwendete Ressource und darauf zurückzuführen sein, dass die Universität Wien teilweise ihre eigene deutschsprachige Terminologie (*Corporate Terminology*) prägt.

Zusammenfassend kann gesagt werden, dass es bei den drei (frei) verfügbaren neuronalen maschinellen Übersetzungssystemen einen signifikanten Überhang an standardvarietätenabhängiger Terminologie gibt, die der deutschen Varietät des Deutschen zugeordnet werden kann.

Diskussion

Durch die Auswahl der Daten beim Zusammenstellen der Korpora für das Training eines (neuronalen) maschinellen Übersetzungssystems kommt es bereits zu einer Verzerrung in eine bestimmte Richtung. Diese Verzerrung kann gewollt erfolgen, z.B., wenn ein maschinelles Übersetzungssystem an eine bestimmte Domäne oder bestimmte Bedürfnisse angepasst werden soll oder sie kann ungewollt aufgrund der begrenzten Verfügbarkeit von bzw. des Zugangs zu Sprachressourcen geschehen. Außerdem sind manche Varietäten in Sprachressourcen aufgrund der geringeren Anzahl der SprecherInnen unterrepräsentiert. Somit kann die Auswahl (und Verfügbarkeit) von Sprachressourcen Auswirkungen auf die sprachliche Vielfalt der maschinellen Übersetzung haben.

Für die digitalen Geisteswissenschaften sind diese Ergebnisse insofern relevant, da sie eine Reflexion über die Ergebnisse von maschineller Übersetzung erlauben, sowie ein kritisches Hinterfragen, Interpretieren und Verwenden von maschinellen Übersetzungssystemen ermöglichen.

Weiterführende Studien

Weiterführende Studien, die sich mit Sprachvarietäten in der maschinellen Übersetzung beschäftigen, sollten neben der Terminologie auch andere Unterscheidungsmerkmale in der plurizentrischen deutschen Standardsprache sowie Non-Standard (Neubarth/Trost 2017) berücksichtigen. Außerdem könnte die Übersetzung von kohärenten (originären) Texten, die relevante Terminologie enthalten, mit maschinellen Übersetzungssystemen ein anderes Ergebnis liefern. Darüber hinaus könnten sich weiterführende Studien neben der MQM-Fehlerkategorie Terminologie der Untersuchung der übrigen MQM-Kategorien, wie Genauigkeit und Stil widmen, da diesen neben der Terminologie in der Fachsprache ebenfalls Bedeutung zukommt. Des Weiteren könnte ein eigens mit der österreichischen Varietät des Deutschen trainiertes maschinelles Übersetzungssystem ebenfalls für den Vergleich genutzt

und die Ergebnisse mit einem Sprachmodell für die österreichische Varietät des Deutschen erneut analysiert werden.

Schlussfolgerung

Durch die Auswahl der Daten und die Menge der zur Verfügung stehenden Sprachressourcen für das Training von maschinellen Übersetzungssystemen kann es zu Verzerrungen in maschinellen Übersetzungen kommen, die die sprachlichen Spielräume und die sprachliche Vielfalt einschränken. (Standard-)Varietäten einer Sprache stellen aktuell eine Herausforderung im Bereich der neuronalen maschinellen Übersetzung dar. Dies konnte in der vorliegenden Studie im Bereich der Terminologie für die österreichische Standardvarietät des Deutschen in den Gebieten Kulinarik, Verwaltungssprache und Universitätsterminologie bestätigt werden, da die untersuchten (generischen) maschinellen Übersetzungssysteme eine Bevorzugung der deutschen Standardvarietät zeigten. Demnach gilt es, nicht nur die Sprachvarietät, sondern auch die sprachvarietätenabhängige Terminologie beim Training maschineller Übersetzungssysteme zu berücksichtigen.

Bibliographie

Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. Berlin / New York: De Gruyter.

Artetxe, Mikel / Labaka, Gorka / Agirre, Eneko (2019): „An Effective Approach to Unsupervised Machine Translation", in: *Proceedings of the 57* th *Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, Association for Computational Linguistics 194–203.

Castilho, Sheila / Moorkens, Joss / Gaspari, Federico / Calixto, Iacer / Tinsley, John / Way, Andy (2017): „Is Neural Machine Translation the New State of the Art?", in: *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.

Chu, Chenhui / Dabre, Raj / Kurohashi, Sadao (2017): „An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation", in: Barzilay, Regina / Kan, Min-Yen (Hrsg.): *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics 385–391.

Costa-jussà, Marta R. / Zampieri, Marcos / Pal, Santanu (2018): „A Neural Approach to Language Variety Translation", in: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)* 275–282.

Dinu, Georgiana / Mathur, Prashant / Federico, Marcello / Al-Onaizan, Yaser (2019): „Training Neural Machine Translation To Apply Terminology Constraints", in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 3063–3068.

EU (1995): *Beachtung österreichischer Ausdrücke. Protokoll Nr. 10 über die Verwendung spezifisch österreichischer Ausdrücke der deutschen Sprache im Rahmen der Europäischen Union* https://ec.europa.eu/info/sites/info/files/austrian_expressions_de.pdf [letzter Zugriff 11. November 2019].

Farajian, M. Amin / Bertoldi, Nicola / Negri, Matteo / Turchi, Marco / Federico, Marcello (2018): „Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation", in: Pérez-Ortiz, Juan Antonio / Sánchez-Martínez, Felipe / Esplá-Gomis, Miquel / Popović, Maja / Rico, Celia / Martins, André / van den Bogaert, Joachim / Forcada, Mikel L. (Hrsg.): *Proceedings of the 21* st *Annual Conference of the European Association for Machine Translation. 28-30 May 2018, Universitat d'Alacant, Alacant, Spain* 149–158.

Forcada, Mikel L. (2017): „Making sense of neural machine translation", in: *Translation Spaces* 6:2 291–309.

Forcada, Mikel L. / Ginestí-Rosell, Mireia / Nordfalk, Jacob / O'Regan, Jim / Ortiz-Rojas, Sergio / Pérez-Ortiz, Juan Antonio / Sánchez-Martínez, Felipe / Ramírez-Sánchez, Gema / Tyers, Francis M. (2011): „Apertium: a free/open-source platform for rule-based machine translation", in: *Machine Translation* 25:2 127–144.

Hasler, Eva / Gispert, Adrià De / Iglesias, Gonzalo / Byrne, Bill (2018): „Neural Machine Translation Decoding with Terminology Constraints", in: *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, June 1 - 6, 2018* 506–512.

Hassan, Hany / Aue, Anthony / Chen, Chang / Chowdhary, Vishal / Clark, Jonathan / Federmann, Christian / Huang, Xuedong / Junczys-Dowmunt, Marcin / Lewis, William / Li, Mu / Liu, Shujie / Liu, Tie-Yan / Luo, Renqian / Menezes, Arul / Qin, Tao / Seide, Frank / Tan, Xu / Tian, Fei / Wu, Lijun / Wu, Shuangzhi / Xia, Yingce / Zhang, Dongdong / Zhang, Zhirui / Zhou, Ming (2018): *Achieving Human Parity on Automatic Chinese to English News Translation* https://arxiv.org/pdf/1803.05567

Heinisch, Barbara (2018): *Dissemination of administrative terminology on Austria's language resource portal as a means of quality assurance*, Donostia / San Sebastián, Spanien: Poster presented at EAFT Terminology Summit 2018: 3M4Q: Making, Managing, Measuring Terminology. In the pursuit of Quality - Donostia / San Sebastián, Spanien, 22-23 Nov. 2018.

Heinisch, Barbara / Lušicky, Vesna (2019): „User expectations towards machine translation: A case study", in: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, 19–23 August, 2019, Dublin, Ireland* 42–48.

Heinisch-Obermoser, Barbara (2014): „University terminology: Why it is not just higher education terminology", in: Budin, Gerhard / Lušicky, Vesna (Hrsg.): *Languages for special purposes in a multilingual, transcultural world. Proceedings of the 19* th *European Symposium on Languages for Special Purposes, LSP, 2013 8-10 July, 2013 Vienna, Austria*. Vienna: Centre for Translation Studies 429–433.

Heinisch-Obermoser, Barbara (2016): „Terminology workflows at an Austrian university aimed at collaboration, terminology awareness and joint responsibility for university terminology – a study on the University of Vienna's terminological database UniVieTerm", in: European Association for Terminology (Hrsg.): *VIII EAFT Terminology Summit (2016), 14 Nov 2016 - 15 Nov 2016, Luxembourg. Revisions and Visions* 8.

Junczys-Dowmunt, Marcin / Dwojak, Tomasz / Hoang, Hieu (2016): *Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions*. https://arxiv.org/pdf/1610.01108.

Koehn, Philipp / Knowles, Rebecca (2017): „Six Challenges for Neural Machine Translation", in: *Workshop on Neural Machine Translation, Vancouver, BC. arXiv: 1706.03872*.

Lakew, Surafel M. / Erofeeva, Aliia / Federico, Marcello (2018): „Neural Machine Translation into Language Varieties", in: *Proceedings of the Third Conference on Machine Translation (WMT). Volume 1: Research Papers. Brussels, Belgium, October 31 - November 1, 2018* 156–164.

Lohaus, Marianne (2000): *Recht und Sprache in Österreich und Deutschland. Gemeinsamkeiten und Verschiedenheiten als Folge geschichtlicher Entwicklungen; Untersuchung zur juristischen Fachterminologie in Österreich und Deutschland*. Giessen: Fachverl. Köhler.

Lommel, Arle / Uszkoreit, Hans / Burchardt, Aljoscha (2014): „Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics", in: *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació* 12: 455–463.

Lösch, Andrea / Mapelli, Valérie / Piperidis, Stelios / Vasiljevs, Andrejs / Smal, Lilli / Declerck, Thierry / Schnur, Eileen / Choukri, Khalid / van Genabith, Josef (2018): „European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management", in: Calzolari / Nicoletta (Hrsg.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* 1339-1343.

Lušicky, Vesna / Heinisch, Barbara (unveröffentlichtes Manuskript): *User expectations towards machine translation: The implications of experience and disconfirmation in neural machine translation*.

Markhardt, Heidemarie (2002): *Das österreichische Deutsch im Rahmen der Europäischen Union das „Protokoll Nr. 10 über die Verwendung österreichischer Ausdrücke der deutschen Sprache" zum österreichischen EU-Beitrittsvertrag und die Folgen: eine empirische Studie zum österreichischen Deutsch in der EU*. Dissertation, Universität Wien.

Michel, Paul / Neubig, Graham (2018): „Extreme Adaptation for Personalized Neural Machine Translation", in: *Proceedings of the 56* th *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 312–318.

Neubarth, Friedrich / Trost, Harald (2017): „Statistische maschinelle Übersetzung vom Standarddeutschen in den Wiener Dialekt", in: Resch, Claudia / Dressler, Wolfgang U. (Hrsg.): *Digitale Methoden der Korpusforschung*. Wien: Verlag der österreichischen Akademie der Wissenschaften 179–203.

Reynolds, Peter (2015): „Machine translation, translation memory and terminology management", in: Kockaert, Hendrik J. / Steurs, Frieda (Hrsg.): *Handbook of Terminology*. Amsterdam: John Benjamins Publishing Company 276–287.

Scansani, Randy / Bernardini, Silvia / Ferraresi, Adriano / Gaspari, Federico / Soffritti, Marcello (2017): „Enhancing Machine Translation of Academic Course Catalogues with Terminological Resources", in: *Proceedings of the Workshop on Human-Informed Translation and Interpreting Technology: Incoma Ltd. Shoumen, Bulgaria* 1–10.

Schmidlin, Regula (2011): *Die Vielfalt des Deutschen: Standard und Variation Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin: De Gruyter.

Servan, Christophe / Crego, Josep / Senellart, Jean (2016): *Domain specialization: a post-training domain adaptation for Neural Machine Translation* https://arxiv.org/pdf/1612.06141.

Simard, Michel / Ueffing, Nicola / Isabelle, Pierre / Kuhn, Roland (2007): „Rule-based translation with statistical phrase-based post-editing", in: *Proceedings of the 45* th *Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic — June 23 - 23, 2007* 203–206.

Stolze, Radegundis (2009): *Fachübersetzen - ein Lehrbuch für Theorie und Praxis.* Berlin: Frank & Timme.

Wissik, Tanja (2013): *Terminologische Variation in der Rechts- und Verwaltungssprache eine korpusbasierte Analyse der Hochschulterminologie in den Standardvarietäten des Deutschen in Deutschland, Österreich und der Schweiz*. Dissertation, Universität Wien.

Wloka, Bartholomäus / Budin, Gerhard / Winiwarter, Werner (2013): „Machine Translation, Language Analysis, and Mobile Applications in the Terminology Domain", in: *Jusletter IT*.

SubRosa – Multi-Feature-Ähnlichkeitsvergleiche von Untertiteln

Luhmann, Jan
jan.luhmann@gmx.net
Computational Humanities Group, Universität Leipzig

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Tiepmar, Jochen
jtiepmar@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Einleitung: Filmanalyse auf Basis von Untertiteln

Mit der stetig wachsenden Verfügbarkeit von Filmen und Serien, die durch Streaming-Dienste wie Netflix und Amazon Prime in den letzten Jahren weiter befördert wurde, ergeben sich aus Perspektive der Filmanalyse ganz neue Möglichkeiten für quantitative Untersuchungen im Sinne des *distant viewing* (Arnold & Tilton, 2019). Wenngleich Film zunächst vor allem ein visuelles Medium ist, so werden in zunehmendem Maße auch Metadaten und insbesondere die Dialoge (vgl. Kozloff, 2000) in Form von online verfügbaren Untertiteln, Drehbüchern und Fan-Transkripten Gegenstand quantitativer Untersuchungen (vgl. Bednarek, 2020; Burghardt et al., 2016, 2019; Schmidt, 2014). Insbesondere die freie Datenbank *OpenSubtitles*[1] hat sich hier als ertragreiche Datenquelle bewährt. Während die Daten von *OpenSubtitles* bislang vor allem im Bereich maschineller Übersetzung (vgl. Müller & Volk, 2013; Lison & Tiedemann, 2016; Tiedemann, 2016) Verwendung fanden, schlagen wir in diesem Artikel eine Nutzung im Sinne quantitativer Filmstilanalyse basierend auf Ähnlichkeitsvergleichen vor. Wir erweitern damit bestehende Arbeiten (Blackstock & Spitz, 2008; Nessel & Cimpa, 2011; Bougiatiotis & Giannakopolous, 2017), die sich ebenfalls mit Ähnlichkeitsvergleichen von Untertiteln beschäftigen, dabei aber jeweils mit relativ überschaubaren Korpora arbeiten oder sehr spezifische Ansätze der Ähnlichkeitsberechnung umsetzen.

Wir präsentieren das experimentelle Analysetool *SubRosa*, welches Ähnlichkeitsvergleiche für mehrere tausend Untertitel über eine grafische Benutzeroberfläche erlaubt. Wir setzen dabei eine ganze Reihe von Features für die Ähnlichkeitsbe-

rechnung zwischen Untertiteln um, die zudem jeweils individuell gewichtet werden können. *SubRosa* versteht sich damit als exploratives Werkzeug, um die grundlegende Eignung unterschiedlicher Features bzw. Feature-Kombinationen für die computergestützte Ähnlichkeitsberechnung zwischen Untertiteln zu untersuchen, welche dann wiederum in einem nächsten Schritt für großangelegte Ähnlichkeitsvergleiche mithilfe statistischer Verfahren genutzt werden können.

Korpus und Datenaufbereitung

SubRosa stellt Vergleiche zwischen insgesamt 5.896 englischen Untertiteln an, die über *OpenSubtitles* bezogen wurden. *OpenSubtitles* versteht sich als offene Plattform, bei der Nutzer*innen Untertitel in unterschiedlichen Sprachen für unterschiedliche Filme hochladen können. Das Format der Untertitel entspricht dem Exportformat des *SubRip*-Tools, welches automatisiert über OCR Textzeilen aus Filmen mit bereits bestehenden Untertiteln extrahiert. Darüber hinaus werden aber auch viele von Nutzer*innen selbst transkribierte Untertitel hochgeladen. Im Ergebnis gibt es so für die meisten Filme mehrere Versionen von Untertiteln. Wir wählen jeweils die Version für unser Korpus aus, die einer automatischen Validierung in Hinblick auf Encoding- oder OCR-Fehler Stand hält. Weiterhin werden alle ausgewählten Untertitel grundlegend aufbereitet, d.h. es werden bspw. Metainformationen, Autoren-Tags, etc. entfernt, die definitiv nicht Teil des eigentlichen Filmdialogs sind. Als nächstes erfolgt eine Vorverarbeitung der Untertitel im Sinne des *natural language processing* (NLP), welche die folgenden Einzelschritte enthält: Tokenisierung, Satzsegmentierung, Lemmatisierung, POS-Tagging und *named entity recognition*. Zuletzt werden alle Untertitel mit Metadaten wie etwa „Titel", „Jahr der Veröffentlichung", „Genre", etc. verknüpft, die über die IMDb-Datenbank[2] bezogen werden.

Analyseverfahren

Mit *SubRosa* setzen wir einen parametrisierbaren Ähnlichkeitsvergleich zwischen Filmuntertiteln um, der auf ganz unterschiedlichen Features basiert. Die nachfolgenden Features sind allesamt über eine interaktiven Web-Applikation verfügbar, die eine Ähnlichkeitssuche für die eingangs erwähnten annähernd 6.000 englischsprachigen Film-Untertitel erlaubt.

- **SubRosa Code**: http://github.com/bbrause/subrosa
- **SubRosa Live-Demo**: http://ch01.informatik.uni-leipzig.de:5001/

Features auf der inhaltlichen Ebene

a) Bag of words / tf-idf („Worüber sprechen die Figuren?"): Das *bag of words*-Modell ist ein einfacher Ansatz für die Repräsentation von Textdokumenten im NLP und Information Retrieval. In unserem Anwendungskontext entspricht ein Untertitel einem „Dokument", für das die einzelnen lemmatisierten Tokens jeweils mit einer sublinearen tf-idf-Skalierung (Manning et al., 2008, S. 126-127) gewichtet werden. Durch diese Gewichtung können wir diejenigen Wörter identifizieren, die in einem bestimmten Dokument häufig vorkommen, aber insgesamt im Gesamtkorpus nur selten auftreten. Es kann davon ausgegangen werden, dass diese Begriffe für das jeweilige Dokument dann besonders aussagekräftig sind. Dementsprechend filtern wir alle Begriffe heraus, die in weniger als 2,5% und mehr als 95% aller Dokumente vorkommen. Darüber hinaus werden *named entities*, die Personen-, Orts- oder Institutionsnamen bezeichnen, entfernt, da diese die Ergebnisse stark verzerren können. Es verbleiben insgesamt 4.952 Wörter, die beim Ähnlichkeitsvergleich der Untertitel berücksichtigt werden.

b) Sentiment Analyse („Was fühlen die Figuren?"): Um Muster bzgl. der von den Figuren im Dialog zum Ausdruck gebrachten Gefühle und Emotionen automatisch zu detektieren, wurde das weitverbreitete *open source*-Tool *VADER Sentiment* (Hutto & Gilbert, 2014) verwendet. Dabei werden für beliebige Textabschnitte Sentiment-Bewertungen im Bereich -1 (maximal negativ) bis +1 (maximal positiv) berechnet. Da sich Emotionen im Laufe eines Films meist sehr divers entwickeln, ist es nicht sinnvoll, das Sentiment des gesamten Filmdialogs mit einem einzigen Wert wiederzugeben. Stattdessen berechnen wir für jede Sekunde eines Films einen spezifischen Sentiment-Wert für den dort gesprochenen Dialog, sodass sich für jeden Film eine Zeitreihe von Sentiment-Werten ergibt. Als Features dieser Zeitreihen extrahieren wir den Mittelwert und Quartilswerte, um die Verteilung der Sentiment-Werte zu erfassen. Weiterhin wird die Nulldurchgangsrate der Zeitreihenkurve sowie deren erste und zweite Ableitung ausgewertet, um Hinweise auf periodische Eigenschaften zu erlangen.

Features auf der stilistischen Ebene („Wie sprechen die Figuren?")

a) Stoppwort-Verteilung: Als weitere Features implementieren wir eine Analyse der Verteilung von Stoppwörtern, also von Wörtern, die in unserem Korpus am häufigsten auftreten und im Gegensatz zum vorherigen Ansatz nur geringe inhaltliche Aussagekraft für einen Film besitzen. Wir berücksichtigen insgesamt 87 Stoppwörter, die nach ihrer Termfrequenz gewichtet werden.

b) POS-Trigramme: Darüber hinaus setzen wir einen Ansatz von Argamon et al. (2003) und Santini (2004) um, die im Kontext stilometrischer Genreklassifikation mit POS-Trigrammen arbeiten. Wir ignorieren dabei all die POS-Trigramme, die in weniger als 90% unserer Dokumente vorkommen, was zu insgesamt 417 verbleibenden POS-Trigrammen führt. Gewichtet werden diese ebenfalls nach ihrer Termfrequenz.

c) Statistische Maße: Wir berechnen außerdem verschiedene statistische Maße, die im Bereich der Stilometrie weit verbreitet sind und die als weitere Features bei unserer Ähnlichkeitsberechnung verwendet werden können. Zu diesen Maßen zählen die Durchschnittswerte einfacher Wort- und Satzlängen sowie auch die *Entropie* (Shannon, 1948) und die *standardized type-token ratio* (Johnson, 1944; Torruella & Capsada, 2013).

d) Dialogtempo („Wie schnell bzw. wie viel wird gesprochen?"): Als letztes Feature betrachten wir das „Dialogtempo", das sich allerdings nicht auf die Sprechgeschwindigkeit einzelner Figuren bezieht, sondern vielmehr Dialoganteile pro Zeit misst. Analog zum Verfahren bei unserem Modell der Sentiment-Analyse messen wir hier pro Sekunde eines Films die Anzahl der gesprochenen Wörter, sodass sich je Film eine Zeitreihe ergibt. Als Features der Zeitreihen extrahieren wir ebenfalls Mittelwert und Quartilswerte zur Erfassung der Verteilung der Dialogtempo-Werte, sowie die Rate der Mittelwert-

durchgänge jeder Zeitreihe und Nulldurchgangsraten der ersten und zweiten Ableitung zur Abschätzung von periodischen Eigenschaften.

Ähnlichkeitsberechnung

Für alle Untertitel werden anhand der oben genannten Feature-Modelle entsprechende Ergebnisvektoren berechnet (vgl. Abb. 1). Ähnlichkeiten bzw. Distanzen werden pro Modell separat berechnet. Für das *bag of words*-Modell verwenden wir die Cosinus-Ähnlichkeit als Metrik, für alle anderen Modelle die Cosinus-Delta-Metrik, die der Cosinus-Ähnlichkeit auf standardisierten Feature-Werten (*z-Scores*) entspricht und auch häufig in der Stilometrie Verwendung findet. Ein Gesamtähnlichkeitswert zwischen zwei Filmen, wie er in *SubRosa* letztendlich ablesbar ist, wird berechnet als der gewichtete Mittelwert der Ähnlichkeitswerte aus den einzelnen Modellen. Darüber hinaus ist eine spezifische Gewichtung (jeweils 0 - 100%) der einzelnen Features über das Interface des Webtools *SubRosa* möglich.

Feature-Modell	Anzahl der Dimensionen
Bag of words / tf-idf	4952
Sentiment Analyse	6
Stoppwort-Verteilung	87
POS-Trigramme	417
Statistische Maße	10
Dialogtempo	6

Abbildung 1: Anzahl der Dimensionen je Feature-Modell.

Ergebnisse in SubRosa

Wie eingangs beschrieben versteht sich *SubRosa* als explorative Tool um die Auswirkung unterschiedlicher Features auf die Ähnlichkeitsberechnungen zwischen Untertiteln zu untersuchen. Zur besseren Illustration der Möglichkeiten des Tools zeigt Abb. 2 die grafische Benutzeroberfläche von *SubRosa* mit einer Darstellung ähnlicher Filme zum Film „Alien (1979)“. Auf der linken Seite zu sehen sind die unterschiedlichen Feature-Modelle und deren Gewichtung, die sich jeweils auf die Ergebnisdarstellung auswirken. Die Ergebnisse der Ähnlichkeitsberechnungen zwischen den Filmen werden in einem Graphen visualisiert, in dem jeder Knoten einen Film darstellt und die Länge der Kante zwischen jeweils zwei Filmen näherungsweise proportional zum Quadrat der zwischen ihnen berechneten Distanz ist.

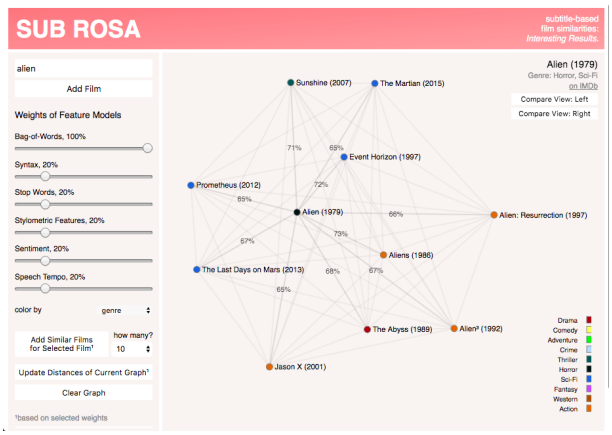


Abbildung 2: Ähnlichkeitsnetzwerk für Alien (1979) in *SubRosa*.

In Detailansichten (vgl. Abb. 3) für jedes Feature-Modell lassen sich darüber hinaus für jeden einzelnen Film seine extrahierten Feature-Daten analysieren und mit denen anderer Filme vergleichen.

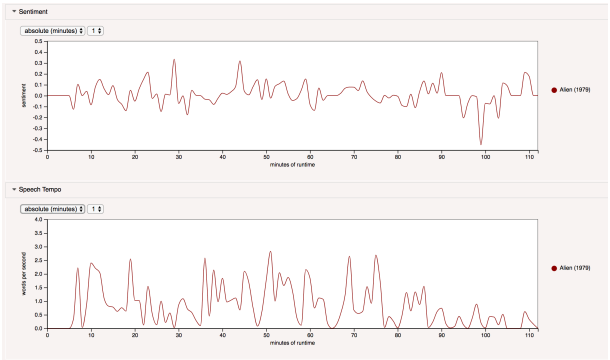


Abbildung 3: Detailsicht der einzelnen Feature-Modelle für Alien (1979), hier für die beispielhaften Features „Sentiment Analyse“ und „Dialogtempo“.

Um einen Überblick zu allgemeinen Ähnlichkeitsmustern im Sinne von Cluster-Bildung innerhalb unseres Korpus an Untertiteln zu erlangen, haben wir zudem den hochdimensionalen Vektorraum jedes Modells mithilfe einer SVD (singular value decomposition) reduziert und die Ergebnisse mittels t-SNE (t-distributed stochastic neighbor embedding) in einem zweidimensionalen Raum als Punkte visualisiert, die entsprechend der Filmgenres eingefärbt sind. Beispielhaft zeigen sich bei der Visualisierung einer gewichteten Kombination aller Modelle (50% Bag-of-Words-Modell, andere Modelle je 10%; siehe Abb. 4) interpretierbare Cluster von Filmen bestimmter Genres, am deutlichsten im Falle von Horror- und Comedy-Filmen. Bei näherer Betrachtung zeigen sich zudem Cluster von Filmen, die sich zwar im Genre stark unterscheiden, jedoch durch ein gemeinsames Setting oder Thema verbunden sind (wie z.B. Weltraum, Western, Schifffahrt, Sport, …).

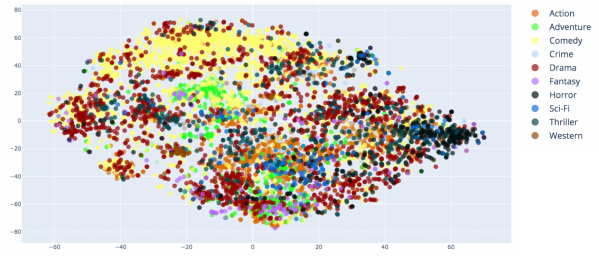


Abbildung 4: Gewichtete Kombination aller Feature-Modelle und 2D-Projektion mittels SVD und t-SNE.

Weiterhin lässt sich zeigen, dass die meisten Features nicht miteinander korrelieren, d.h. Filme die bspw. anhand des Features „Sentiment“ ähnlich sind, können sich erheblich unterscheiden was etwa das Dialogtempo angeht (vgl. Abb. 5).

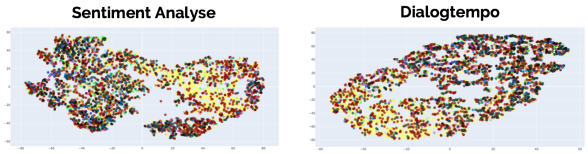


Abbildung 5: Die 2D-Projektion der Untertitel mittels SVD und t-SNE anhand der Features “Sentiment Analyse” und “Dialogtempo” zeigt sehr unterschiedliche Cluster und lässt darauf schließen, dass diese beiden Merkmale nicht korrelieren. Dies gilt im Übrigen auch für die meisten der anderen Features; die entsprechenden Diagramme finden sich online über plot.ly[3].

Die Unterschiedlichkeit der verschiedenen Features lässt sich auch gut anhand beispielhafter Analysen illustrieren. So zeigt sich etwa, dass bei der Suche nach ähnlichen Filmen zu “The Room” (2003) für jedes einzelne Feature bei den Top 5 der als ähnlich identifizierten Ergebnisse jeweils ganz unterschiedliche Filme herauskommen (vgl. Abb. 6). Einzig “Ruby Sparks” (2012) findet sich sowohl bei “Syntax” als auch bei “Sentiment” wieder. Der Film “The Disaster Artist”, der dokumentationsartig die Entstehungsgeschichte des Klassikers “The Room” schildert (und damit einen unmittelbaren inhaltlichen Bezug hat), kommt interessanterweise nur bei der *bag of words*-Methode in den Top 5 der Ergebnismenge vor. Es zeigt sich also, dass ein multifaktorieller Vergleich von Filmen anhand unterschiedlicher, dialog-basierter Features, nicht zielführend ist, sondern vielmehr unterschiedliche Merkmale unterschiedliche Ähnlichkeitsaspekte kodieren. Im nächsten Schritt planen wir eine systematische Korrelationsanalyse der unterschiedlichen Features, um gemeinsam auftretende Phänomene und Muster für spezifische Filmgenres etc. identifizieren zu können.

Methode	Unsortierte Top 5 Ergebnismenge
Bag of Words	American Reunion (2012), Bridesmaids (2011), The Disaster Artist (2017), This is Where I Leave You (2014), Funny People (2009)
Syntax (=POS-Trigramme)	Ruby Sparks (2012), Shame (2011), Addicted (2014), Pretty in Pink (1986), The Edge of Seventeen (2016)
Stop Words	Belle de Jour (1967), Jamon Jamon (1992), Frankie and Johnny (1991), Sweet November (2001), Some Kind of Wonderful (1987)
Stylometric Features (=Statistische Maße)	Stranger Than Paradise (1984), The Butterfly Effect 3: Revolutions (2009), Shallow Grave (1984), Cellular (2004), The Forgotten (2004)
Sentiment	Willy Wonka and the Chocolate Factory (1971), Night and the City (1950), Bee Movie (2007), Stuck on You (2003), Ruby Sparks (2012)
Speech Tempo	The Hangover Part III (2013), Terminal (2018), The Hateful Eight (2015), The Man Who Wasn't There (2001), Life Itself (2018)

Abbildung 6: Unterschiede in der Ergebnismenge verschiedener Feature-Konfigurationen für *„The Room“* (2003).

Fazit und Ausblick

Im hier vorgestellten Projekt dokumentieren wir aktuelle Experimente zur Identifikation von Ähnlichkeitsbeziehungen zwischen Film-Untertiteln auf Basis ganz unterschiedlicher Features, die künftig für quantitative Stil- und Genreanalyse von Filmen herangezogen werden können. *SubRosa* versteht sich zunächst als experimentelle Plattform, die es erlaubt interaktiv unterschiedliche Feature-Kombinationen für unterschiedliche Filme bzw. Fragestellungen zu erproben. Als Verbesserung auf technischer Ebene planen wir die Integration eines größeren Korpus[4] (Lison & Tiedemann, 2016), welches systematischer validiert und korrigiert wurde als es bei unserem aktuellen Testkorpus der Fall ist.

Darüber hinaus soll über eine systematische Evaluation eine Feature-Selektion und optimale Gewichtung erfolgen. Geplant ist hierzu eine Evaluation gegen eine *ground truth* auf Basis bestehender Ähnlichkeitsverbindungen, bspw. über die Empfehlungen via *collaborative filtering* bei Amazon oder über den frei verfügbaren Datensatz *MovieLens*.[5] Offen ist dabei die Frage, ob Ähnlichkeitsbewertungen auf Basis audio-visueller Features grundsätzlich mit Ähnlichkeitsbewertungen auf Dialogebene korrelieren, oder die verschriftlichte Dialogebene ggf. als isolierte Ebene betrachtet werden muss. Wir planen deshalb weitere Fallstudien mithilfe von *SubRosa*, die zusammen mit Film- und Sprachwissenschaftlern durchgeführt werden sollen.

Fußnoten

1. OpenSubtitles: https://www.opensubtitles.org/de
2. IMDb: https://www.imdb.com/
3. Feature-Visualisierungen: https://chart-studio.plot.ly/~bbrause/#/
4. OpenSubtitles 2018-Korpus: http://opus.nlpl.eu/OpenSubtitles2018.php
5. MovieLens Dataset: https://movielens.org/

Bibliographie

Aggarwal, C. C. (2001): On k-anonymity and the curse of dimensionality. In: Proc. 31st International Conference on Very Large Data Bases (VLDB), S. 901–909. ACM, 2005.

Argamon, S. / Shimoni, A. R. / Koppel, M. (2003): Automatically categorizing written texts by author gender. In: Literary and Linguistic Computing, Vol. 17, Nr. 4, S. 401– 412.

Bednarek, M. (to appear 2020): The Sydney Corpus of Television Dialogue: Designing and building a corpus of dialogue from US TV series. Corpora 15/1. Pre-Print-Version hier verfügbar: https://www.monikabednarek.com/wp-content/uploads/2019/09/Designing-and-building-a-corpus-of-US-TV-dialogue_Academia.pdf

Blackstock, A. / Spitz, M. (2008): Classifying movie scripts by genre with a MEMM using NLP-based features. M.Sc. Kurs Natural Language Processing, stud. Projektbericht, Juni 2008. Stanford University.

Bougiatiotis, K. / Giannakopoulos, T. (2017): Multimodal content representation and similarity ranking of movies. Pre-Print-Version hier verfügbar: https://arxiv.org/pdf/1702.04815.pdf

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).

Burghardt, M. / Meyer, S. / Schmidtbauer, S. / Molz, J. (2019): "The Bard meets the Doctor" – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who. In Book of Abstracts, DHd 2019.

Schmidt, B. (15.9.2014): Screen time! Published on http://sappingattention.blogspot.com/2014/09/screen-time.html (letzter Zugrifff am 24.9.2019)

Hutto, C. J. / Gilbert, E. (2014): VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: International Conference on Weblogs and Social Media.

Johnson, W. (1944): Studies in language behavior: I. A program of research. In: Psychological Monographs, Vol. 56, S. 1-15.

Kozloff, S. (2000): Overhearing Film Dialogue. University of California Press.

Santini, M. (2004): A shallow approach to syntactic feature extraction for genre classification. In Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 2004).

Shannon, C. (1948): A mathematical theory of communication. In: The Bell System Technical Journal, Vol. 27, S. 379–423, 623–656, Juli und October 1948.

Taylor, A. / Tilton, L. (2019): Distant viewing: analyzing large visual corpora. In Digital Scholarship in the Humanities, 2019. Published by Oxford University Press on behalf of EADH.

Torruella, J. / Capsada, R. (2013): Lexical statistics and topological structures: A measure of lexical richness. In: *Procedia - Social and Behavioral Sciences*, Vol. 95, S. 447-454.

Manning, C. / Raghavan, P. / Schütze, H. (2008): Introduction to Information Retrieval. Cambridge University Press.

Müller M. / Volk M. (2013): Statistical Machine Translation of Subtitles: From OpenSubtitles to TED. In: Gurevych I., Biemann C., Zesch T. (eds) Language Processing and Knowledge in the Web. Lecture Notes in Computer Science, vol 8105. Springer, Berlin, Heidelberg.

Nessel, J. / Cimpa, B. (2011): The MovieOracle-content based movie recommendations. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, S. 361-364.

Lison, P. / Tiedemann, J. (2016): OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), p. 923-929, European Language Resources Association.

Tiedemann , J. (2016): Finding Alternative Translations in a Large Corpus of Movie Subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), p. 3518–3522, European Language Resources Association.

Syntaktische Profile für Interpretationen jenseits der Textoberfläche

Andresen, Melanie
melanie.andresen@uni-hamburg.de
Universität Hamburg, Deutschland

Begerow, Anke
anke.begerow@haw-hamburg.de
Hochschule für Angewandte Wissenschaften Hamburg

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

Gaidys, Uta
Uta.Gaidys@haw-hamburg.de
Hochschule für Angewandte Wissenschaften Hamburg

Koch, Gertraud
gertraud.koch@uni-hamburg.de
Universität Hamburg, Deutschland

Zinsmeister, Heike
heike.zinsmeister@uni-hamburg.de
Universität Hamburg, Deutschland

Viele Verfahren des Text Mining und Distant Reading beschränken sich auf eine wortbasierte Auswertung von Texten. Auch wenn auf Basis der Wortformen und ihrer linearen Abfolge bereits neue Perspektiven auf Texte gewonnen werden (z. B. mittels der Voyant Tools, Sinclair / Rockwell 2016), schöpfen diese Methoden das Potential von Texten bei Weitem nicht aus. Insbesondere, wenn die Auswertungsergebnisse Gegenstand weiterführender Interpretationen werden, z. B. um soziale Phänomene zu beschreiben, sehen wir einen Mehrwert in der Auswertung zusätzlicher sprachlicher Strukturen. Konkret verwenden wir syntaktische Annotationen, die präzisere Informationen zu Wortkombinationen liefern können, etwa „X ist Subjekt von Y" anstelle von „X steht im Kontext von Y". Zudem bestehen viele syntaktische Relationen über eine längere Distanz an der Oberfläche hinweg und können deshalb nur durch eine syntaktische Perspektive erfasst werden (z. B. Andresen / Zinsmeister 2017). Dies gilt für unterschiedliche Sprachen in unterschiedlichem Ausmaß. Das Deutsche

verfügt über deutlich mehr Distanzstrukturen als das Englische, für das die meisten Analyseverfahren ursprünglich entwickelt wurden.

In diesem Beitrag vergleichen wir zwei Ansätze zur Berechnung von Kollokationen, einen oberflächenorientierten Ansatz und einen auf Dependenzannotationen basierten. An zwei Fallstudien aus den Fächern Kulturanthropologie und Pflegewissenschaft wird demonstriert, wie die beiden Ansätze eine qualitative Interpretation von Textdaten in Hinblick auf gesellschaftliche bzw. soziale Phänomene unterstützen können. Die Erstellung eines eindeutigen Goldstandards, der eine formale Evaluation erlauben würde, ist bei dieser Art Fragestellung nicht möglich. Stattdessen wird auf qualitative Weise das Potential dieser Analyse zur Bearbeitung der geistes- und sozialwissenschaftlichen Fragestellungen beschrieben.

Syntaktische Profile: Forschungsstand

Die Nutzung syntaktischer Informationen zur Charakterisierung von Wortverwendungen ist vor allem in der Lexikografie betrieben worden. Populär wurde das Konzept unter dem Namen „word sketch" besonders durch die korpuslinguistische Software SketchEngine (Kilgarriff et al. 2004, Kilgarriff et al. 2014). Für ein gegebenes Suchwort wird hier angegeben, welche anderen Wörter besonders häufig in spezifischen syntaktischen Relationen zum Suchwort stehen, z. B. *glimpse* als frequentes Objekt von *catch* (Kilgarriff et al. 2014: 9). Im Digitalen Wörterbuch der deutschen Sprache (DWDS) kann eine entsprechende Darstellung als sog. DWDS-Wortprofil abgerufen werden (Geyken 2011).

Weitere Anwendungen gibt es in der Literaturwissenschaft und Linguistik: Googasian / Heuser (2019) vergleichen die syntaktischen Kontexte von Menschen und Tieren in einem Korpus sog. „wild animal stories". Andresen (2018) nutzt syntaktische n-Gramme für einen Vergleich der Wissenschaftssprachen in den Fächern Literaturwissenschaft und Linguistik. Eine Anwendung auf sozialwissenschaftliche Fragestellungen erfolgt vor allem in den Politikwissenschaften: Anhand syntaktischer Muster wie z. B. Argumentstrukturen oder Mustern der Redewiedergabe werden hier politische Akteure und ihre Positionen identifiziert, miteinander in Relation gesetzt und so Diskurse charakterisiert (van Atteveldt et al. 2008, Kleinnijenhuis / van Atteveldt 2014, Wüest et al. 2011, Blessing et al. 2013). In den Fächern Pflegewissenschaft und Kulturanthropologie steht die Exploration des Mehrwertes syntaktischer Analysen noch aus.

Daten und Fragestellungen der Fallstudien

Das Potential syntaktisch definierter Kollokationen wird an zwei Fallstudien mit komplementärer Datenlage erprobt. Die erste nutzt ein großes Korpus geschriebener Sprache, das dadurch methodisch eine sehr sichere Grundlage bietet. Die zweite Fallstudie basiert auf einem eher kleinen Korpus mit gesprochener Sprache, was mehr methodische Herausforderungen erwarten lässt.

Die kulturanthropologische Fallstudie befasst sich mit dem Themenkomplex der Telemedizin und insbesondere der Frage nach der (Nicht-)Akzeptanz telemedizinischer Anwendungen durch unterschiedliche gesellschaftliche Akteursgruppen. Das hierfür erstellte Korpus umfasst 8.784 Texte mit insgesamt 14,8 Mio. Token und basiert auf einem Webcrawling (Adelmann / Franken 2020). Dafür wurden Webseiten von Krankenkassen, Ärzte- und Patientenverbänden als Ausgangspunkt genutzt und dann Links zu Seiten verfolgt, die mindestens eines von mehreren Wörtern aus einem Wortfeld zur Telemedizin enthalten (Koch / Franken im Druck).

Grundlage der pflegewissenschaftlichen Fallstudie ist ein Korpus aus 31 Dialogen, die mit schwerkranken und sterbenden Menschen in palliativer Versorgung geführt wurden. Es handelt sich um Transkripte gesprochener Sprache im Umfang von gut 100.000 Token. Gegenstand der Studie sind die Deutungen von Entscheidungen hinsichtlich der gesundheitlichen Versorgung der Betroffenen.

Methode

Die Texte beider Korpora werden mithilfe des Parsers MATE (Bohnet 2010), trainiert auf der Hamburger Dependency Treebank (Foth et al. 2014), mit Lemmata, Wortarten und syntaktischen Dependenzen annotiert. Unter Kollokationen verstehen wir „a combination of two words that exhibit a tendency to occur near each other in natural language" (Evert 2008: 1214). Bei der Operationalisierung von „near each other" können Kriterien an der Textoberfläche oder syntaktische Kriterien angesetzt werden: Für den einfachen Ansatz ohne Annotationen betrachten wir Wörter in einem Kontextfenster von +/- 3 Wörtern als benachbart, für den syntaktischen Ansatz Wörter mit einer direkten Dependenzrelation. In beiden Fällen wird mithilfe des Log-Likelihood-Ratios (LLR, Dunning 1993) berechnet, welche Kombinationen häufiger im Korpus vorkommen, als basierend auf den Einzelfrequenzen der Wörter zu erwarten wäre. Im Falle der syntaktischen Kollokationen werden dafür die Einzelfrequenzen in der spezifischen syntaktischen Relation genutzt. Die Ergebnisse basieren auf den Lemmata und werden nach Schlüsselwörtern gefiltert, die für die jeweiligen Fragestellungen als bedeutsam ausgewählt wurden (*T/telemed* bzw. *E/entscheid*). Das hierfür verwendete Analyseskript steht auf GitHub zur Verfügung.[1] Für die Interpretation werden die Top 10 beider Listen verglichen und nach Bedarf weitere Einträge gesichtet.

Ergebnisse der Fallstudien

Kulturanthropologie

Tabelle 1 zeigt die oberflächenbasierten Kollokationen zu Lemmata mit *T/telemed* im kulturanthropologischen Korpus mit den höchsten LLR-Werten. *Telemedizin* ist sehr stark mit dem verwandten Wort *Telematik* assoziiert, was die enge Verknüpfung der Bereiche anzeigt. Manche Wortpaare sind Bestandteil mehrteiliger Eigennamen (*Bayerische TelemedAllianz, [Zentrum für Telematik und] Telemedizin GmbH*), die für die Interpretation einen eingeschränkten Mehrwert haben, aber doch für den Diskurs potentiell relevante und ggf. bisher unbekannte Akteure sichtbar machen. Mit dem *Tag der Telemedizin* wird ein Fachkongress als wichtiger Begegnungs-

punkt dieser Akteure aufgeführt. Außerdem liegen allgemeine Konzepte wie *telemedizisch* und *Anwendung* hoch im Ranking.

Tabelle 1: Top 10 der oberflächenbasierten Kollokationen zu Lemmata mit *T/ telemed* im kulturanthropologischen Korpus

Wort 1	**Wort 2**	LLR	**abs. Frequenz**
Telematik	Telemedizin	2044,88	468
telemedizinisch	Anwendung	1753,90	465
bayerisch	TelemedAllianz	1497,94	204
Telemedizin	GmbH	1007,39	340
Tag	Telemedizin	845,74	274
telemedizinisch	Betreuung	841,88	212
bayerisch	Telemedallianz	731,35	97
der	Telemedizin	644,28	2533
Gesellschaft	Telemedizin	632,95	241
telemedizinisch	Zentrum	585,73	165

Tabelle 2: Top 10 der syntaxbasierten Kollokationen zu Lemmata mit *T/telemed* im kulturanthropologischen Korpus

Wort 1	Relation	**Wort 2**	LLR	**abs. Frequenz**
GmbH	ist Apposition von	Telemedizin	2261,71	353
telemedizinisch	ist Attribut von	Anwendung	2236,42	456
bayerisch	ist Attribut von	TelemedAllianz	2002,48	204
Telemedizin	ist Genitivattribut von	Tag	1904,26	274
telemedizinisch	ist Attribut von	Betreuung	981,87	200
bayerisch	ist Attribut von	Telemedallianz	967,93	98
DGTelemed	ist Apposition von	V.	899,83	84
telemedizinisch	ist Attribut von	Zentrum	772,92	163
Telemedizin	ist Apposition von	Fachkongress	727,37	64
Telemedizin	ist Genitivattribut von	Möglichkeit	720,75	144

Die syntaktischen Informationen in Tabelle 2 machen den Zusammenhang zwischen den Bestandteilen der Eigennamen in der Relation der Apposition explizit und bieten damit mehr Informationen zur Einordnung dieser Datenpunkte. Die Annotationen ermöglichen außerdem, die Gesamtliste nach bestimmten syntaktischen Relationen zu filtern. Die genannten Appositionen beispielsweise können anhand des Relationslabels ausgeblendet werden. Auch hier gibt es sehr allgemeine Konzepte wie *telemedizinisch* als Attribut von *Anwendung*, die zwar frequent, aber nicht sehr informativ sind. *Telemedizin* als Genitivattribut von *Möglichkeit* weist daraufhin, dass eben deren Möglichkeiten noch Gegenstand des Diskurses sind. In der Durchsicht der Kollokationen jenseits der Top 10 finden sich verwandte Themen des Potentials und der Projekthaftigkeit (*Potential der Telemedizin*, *Telemedizin-typisches Potential*, *evaluiertes Telemedizinprojekt*, *vielversprechendes Telemedizinprojekt*), die anzeigen, dass sich die Umsetzung der Telemedizin in einer frühen Phase befindet und ihre Akzeptanz als Regelversorgung noch nicht abschließend verhandelt ist.

Auch zur Kollokation *telemedizinisch* als Attribut von *Betreuung* finden sich weiter unten im Ranking ähnliche Verwendungen zum Thema Betreuung (*telemedizinisch betreuen*, *telemedizinisch betreut* …) und Unterstützung (*telemedizinisch unterstützt*, *telemedizinisch-unterstützte* (sic!) *Versorgung*, *Telematikunterstützung* …). Dies weist auf die (bisher) eher ergänzende Rolle der Telemedizin im Verhältnis zur medizinischen Regelversorgung hin.

Pflegewissenschaft

Tabelle 3 zeigt die zehn ersten oberflächenbasierten Kollokationen des Dialogkorpus zu Lemmata mit *E/entscheid*. Hier werden zunächst Probleme in den Daten deutlich: Mit *Finan/* liegt ein für gesprochene Sprache typischer Abbruch eines Wortes (vermutlich: *Finanzentscheidung*) vor. Zudem ist *Entscheidungsvariant* eine fehlerhafte Lemmaform zu *Entscheidungsvarianten*. Insgesamt sind die Frequenzen aufgrund der geringen Korpusgröße klein, lassen aber trotzdem hilfreiche Schlüsse für die Analyse zu. Im syntaxbasierten Gegenstück in Tabelle 4 sind zusätzliche Probleme erkennbar, die durch die automatische Verarbeitung gesprochener Sprache entstehen. Die Relation zwischen *hab* und *entscheiden* ist fehlerhaft als adverbial (korrekt: auxiliar) bezeichnet. Allerdings wird ein direkter Zusammenhang zwischen diesen Wörtern erst durch die syntaktischen Annotationen überhaupt erkennbar, da sie im Satz häufig nicht benachbart stehen. Zusätzlich gibt es vollständig falsche Analysen wie die Relation zwischen *entscheidend* und *Puh*.

Tabelle 3: Top 10 der oberflächenbasierten Kollokationen zu Lemmata mit *E/entscheid* im pflegewissenschaftlichen Korpus

Wort 1	**Wort 2**	LLR	**abs. Frequenz**
Entscheidung	treffen	33,41	7
richtig	Entscheidung	23,06	7
Tablettenform	entscheiden	21,28	4
der	Entscheidung	17,54	28
dieser	Entscheidung	13,02	7
selbst	entscheiden	12,88	4
Entscheidung	überlassen	10,87	2
Entscheidungsvariant	nebeneinandstellen	10,38	1
Finan/	Versorgungsentscheidung	10,38	1
entschieden	Abraten	10,38	1

Auch für die pflegewissenschaftliche Interpretation bieten die Kollokationen mit den höchsten LLR-Werten erste Anhaltspunkte, die dann durch eine Sichtung der weiteren Rangplätze ergänzt werden können. Die häufigsten Kollokatoren von Entscheidungen stehen für eine Realisierung eigener Entscheidungen der Betroffenen. Die Kollokation *richtig* macht Bewertungen der Entscheidungen sichtbar. Es zeigen sich zudem gegensätzliche Dimensionen des Phänomens, wie „selber entscheiden“ vs. „Entscheidung abgeben“, die hinter der Kollokation mit *überlassen* stehen. Insbesondere die Subjekt- und Objektrelationen (*Entscheidung treffen*, *Entschluss entstehen*, *Entscheidung überlassen*) sind durch die syntaktische Analyse adäquater und theoretisch fundierter abgebildet. Dieser Nutzen wird jedoch durch Fehler in der automatischen Annotation eingeschränkt. Zudem werden diese Relationen in der gesprochenen Sprache mit kürzeren Sätzen möglicherweise auch durch die oberflächenbasierte Analyse besser erfasst als in anderen sprachlichen Registern. Insgesamt betrachtet werden durch den quantitativen Zugang Verwendungszusammenhänge des Phänomens „Entscheidung“ transparent, die wiederum auf wichtige Handlungskontexte in der Versorgungsrealität von schwerkranken und sterbenden Menschen verweisen.

Tabelle 4: Top 10 der syntaxbasierten Kollokationen zu Lemmata mit *E/ entscheid* im pflegewissenschaftlichen Korpus

Wort 1	Relation	**Wort 2**	LLR	**abs. Frequenz**
entscheiden	ist Adverbial von	hab	33,25	9
richtig	ist Attribut von	Entscheidung	24,54	6
Entscheidung	ist Akkusativobjekt von	treffen	23,47	4
Entschluss	ist Subjekt von	entstehen	21,93	2
selbst	ist Adverbial von	entscheiden	18,74	4
entscheidend	ist Adverbial von	Puh	16,42	1
Tablettenform	ist Subjekt von	entscheiden	16,13	2
Entscheidung	ist Akkusativobjekt von	überlassen	15,59	2
Entscheidung	ist Akkusativobjekt von	treff	13,52	2
für	ist Präposition zu	entscheiden	13,46	7

Schlussfolgerungen

Es hat sich gezeigt, dass die Berechnung von Kollokationen auf der Grundlage der sprachlichen Oberfläche bzw. der Syntax für qualitative Fragestellungen informativ sein kann. Für die Auswertung einer spezifischen Fragestellung ist die Assoziationsstärke allein jedoch nicht immer das entscheidende Kriterium. Die Kollokationen geben Hinweise auf Zusammenhänge innerhalb des Korpus, die neue Fragestellungen und Perspektiven generieren können. Gleichzeitig werden durch die Relationsannotationen bereits kleine Datenmengen in erweiterter Form auswertbar.

Die beispielhaften Analysen haben gezeigt, dass die syntaktischen Annotationen eine für die Interpretation hilfreiche Differenzierung bieten, indem präziser angegeben wird, in welcher Relation zwei Wörter stehen. Das ermöglicht auch das Filtern nach interessanten Relationstypen. Zudem werden durch den Einbezug der Syntax Relationen zwischen Wörtern in Distanzstellung sichtbar, was insbesondere vom Verb abhängige Satzteile besser sichtbar macht. Andererseits erfordern die syntaktischen Annotationen eine aufwendigere Vorverarbeitung, die mehr Zeit und technische Fähigkeiten erfordert. Außerdem stellen sie eine zusätzliche Fehlerquelle dar. Dies gilt besonders für die gesprochensprachlichen Daten. Eine systematische Überprüfung und Rückbindung an konkrete Korpusbelege ist deshalb wichtig und verbessert die Interpretationsmöglichkeiten aus qualitativer Sicht.

Anschließend an diese Arbeiten ist geplant, stärker Kontexte zu aggregieren: In grammatischer Hinsicht wird das durch Koreferenzannotationen erfolgen, die für das pflegewissenschaftliche Korpus bereits vorliegen. Auf semantischer Ebene verfolgen wir den Ansatz, Wortgruppen zu Konzepten zusammenzufassen, z. B. können *Ärztin*, *Arzt*, *Hausarzt*, *Onkologin* usw. auf ein gemeinsames Konzept ÄRZT*INNEN abgebildet werden (vgl. den Ansatz von Wüest et al. 2011). Wir sehen außerdem weitere Anwendungsfälle über die Fächergrenzen hinweg, etwa zur literaturwissenschaftlichen Beschreibung von Geschlechterzuschreibungen, indem die sprachlichen Kontexte weiblicher und männlicher Vornamen verglichen werden.

Fußnoten

1. https://github.com/melandresen/DHd2020

Bibliographie

Adelmann, Benedikt / Franken, Lina (2020): Thematic web crawling and scraping as a way to form focussed web archives, in: *Engaging with Web Archives Conference Book of Abstracts*. To be published at https://ewaconference.com/.

Andresen, Melanie (2018): Sprachliche Variation in der Germanistik: eine n-Gramm-basierte Stilanalyse, in: *Book of Abstracts of DHd 2018*. Köln, Deutschland, 311–15.

Andresen, Melanie / Zinsmeister, Heike (2017): The benefit of syntactic vs. linear n-grams for linguistic description, in: *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2017).* Pisa, Italy, 4–14 http://aclweb.org/anthology/W17-6503.

van Atteveldt, Wouter / Kleinnijenhuis, Jan / Ruigrok, Nel (2008): Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles, in: *Political Analysis*, 16(4): 428–46 doi:10.1093/pan/mpn006.

Blessing, Andre / Sonntag, Jonathan / Kliche, Fritz / Heid, Ulrich / Kuhn, Jonas / Stede, Manfred (2013): Towards a Tool for Interactive Concept Building for Large Scale Analysis in the Humanities, in: *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 55–64 https://www.aclweb.org/anthology/W13-2708.

Bohnet, Bernd (2010): Very High Accuracy and Fast Dependency Parsing is not a Contradiction, in: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China, 89–97. https://www.aclweb.org/anthology/C10-1011

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence, in: *Computational Linguistics*, 19(1): 61–74. https://www.aclweb.org/anthology/J93-1003

Evert, Stefan (2008): Corpora and collocations, in: Lüdeling, Anke / Kytö, Merja (Hg.), *Corpus linguistics: an International Handbook*, Vol. 2. (Handbücher zur Sprach- und Kommunikationswissenschaft 29). Berlin, Boston: De Gruyter, 1212–1248.

Foth, Kilian A. / Köhn, Arne / Beuck, Niels / Menzel, Wolfgang (2014): Because Size Does Matter: The Hamburg Dependency Treebank, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2326–2333. Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/860_Paper.pdf.

Gaidys, Uta / Gius, Evelyn / Jarchow, Margarete / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Zinsmeister, Heike (2017): hermA: Automated modelling of hermeneutic processes, in: *Hamburger Journal für Kulturanthropologie*(7): 119–23.

Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora, in: Abel, Andrea / Zanin, Renata (Hg.), *Korpora in Lehre und Forschung*. Bozen: Bolzano Univ. Press, 129–54.

Googasian, Victoria / Heuser, Ryan J. (2019): Digital Animal Studies: Modeling Anthropomorphism in Animal Writing, 1870-1930, in: *Book of Abstracts of DH 2019* https://dev.clariah.nl/files/dh2019/boa/0458.html.

Kilgarriff, Adam / Baisa, Vít / Bušta, Jan / Jakubíček, Miloš / Kovář, Vojtěch / Michelfeit, Jan, Rychlý / Pavel. / Suchomel, Vít (2014): The Sketch Engine: ten years on, in: *Lexicography*, 1(1): 7–36 doi:10.1007/s40607-014-0009-9.

Kilgarriff, Adam / Rychlý, Pavel / Smrz, Pavel / Tugwell, David (2004): The Sketch Engine: in: *Proceedings of the*

11th EURALEX International Congress. 105–15 https://euralex.org/publications/the-sketch-engine/.

Kleinnijenhuis, Jan / van Atteveldt, Wouter (2014): Positions of Parties and Political Cleavages between Parties in Texts, in: Kaal, Bertie / Maks, Isa / van Elfrinkhof, Annemarie (Hg.), *Discourse Approaches to Politics, Society and Culture*, Vol. 55. Amsterdam: Benjamins, 1–20 doi:10.1075/dapsac.55.01kle.

Koch, Gertraud / Franken, Lina (im Druck): Automatisierungspotenziale in der qualitativen Diskursanalyse. Das Prinzip des Filterns, in: Schilling, Samuel / Klimczak, Peter (Hg.): *Die Gesellschaft im Spiegellabyrinth sozialer Medien.* Wiesbaden.

Sinclair, Stéfan / Rockwell, Geoffrey (2016): Voyant Tools. Web. http://voyant-tools.org/.

Wüest, Bruno / Clematide, Simon / Bünzli, Alexandra / Laupper, Daniel (2011): Semi-Automatic Core Sentence Analysis: Improving Content Analysis for Electoral Campaign Research, in: *International Relations Online Working Paper*(1). https://www.sowi.uni-stuttgart.de/dokumente/forschung/irowp/IROWP_Series_2011_1_Wueest_Clematide_Buenzli_Laupper_Content_Analysis.pdf.

Textanalyse mit kombinierten Methoden – ein konzeptioneller Rahmen für reflektierte Arbeitspraktiken

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

Pichler, Axel
axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Viehhauser, Gabriel
gabriel.viehhauser-mery@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Die Zahl und Intensität der Aktivitäten im interdisziplinären Spektrum der *Digital Humanities* (DH) bzw. *Computational Social Science* ist in den vergangen 5-10 Jahren enorm angewachsen. Nicht zuletzt dank eines undogmatischen Selbstverständnisses der Forschungscommunity basieren die Aktivitäten auf einem Methodenpluralismus, dem eine ebenso facettenreiche Methodenreflexion entspricht. Gleichwohl ist es für die wissenschaftliche Praxis unerlässlich, dass sich Teilcommunities, die ein vergleichbares Erkenntnisinteresse innerhalb des DH-Spektrums verfolgen, über den konzeptuellen Rahmen sowie die aus ihm folgenden Maßstäbe verständigen, die sie an ein methodisch valides Vorgehen anlegen.

Forschung mit kombinierte komputationell/geisteswissenschaftlichen Methoden

Dieser Beitrag fokussiert auf denjenigen Teilbereich der DH, der sich zum Ziel setzt, adaptierbare datenorientierte Computermodelle methodisch adäquat in kombiniert komputationell/geisteswissenschaftliche Arbeitspraktiken zu integrieren. Methodologisch zielt diese Teildisziplin also darauf ab, Forschungsfragen aus einem geisteswissenschaftlichen Kontext mit kombinierten Methoden (bzw. mit “mixed methods”) adäquat bearbeiten zu können.[1] Basierend auf eigenen Erfahrungen und dem Austausch innerhalb der DH-Community ist unsere Einschätzung, dass Mitglieder von Forschungsteams, die im Spektrum der komputationell/geisteswissenschaftlichen Methodenkombination eingehende Projekterfahrung gesammelt haben, eine ausdifferenzierte Wahrnehmung der zu kombinierenden Arbeitspraktiken entwickelt haben. Ein methodologischer Metadiskurs, der eine enge Verzahnung der kombinierten Methoden thematisiert, findet jedoch nur in engen Zirkeln – häufig projektintern – statt. Für neu etablierte Projektkooperationen ist es daher nach wie vor schwierig, Workflows aufzusetzen, die ein reflektiertes Vorgehen garantieren. Zudem wird für unterschiedliche methodische Komponenten die jeweilige Adäquatheit häufig nicht auf dem gleichen Reflexionsniveau diskutiert: so kann für die “komputationelle” Praxis, das Modellverhalten im Rahmen einer Evaluation und Fehleranalyse zu reflektieren, auf klarere Konventionen aufgebaut werden als etwa für ein Hinterfragen der Annahmen zum Interpretationskontext der untersuchten Texte, die mit der Operationalisierung zentraler Analysekategorien einhergehen.

Es erscheint uns daher an der Zeit, intensiver über einen geeigneten konzeptionellen Rahmen für Arbeiten aus einem der DH-Teilbereiche zu diskutieren, in denen kombinierte Methoden zum Einsatz kommen – einen Rahmen, der eine gleichermaßen adäquate Reflexion für alle einfließenden Vorannahmen ermöglicht und zudem einfach genug darstellbar ist, dass sich ein methodisch adäquater Workflow mit vertretbarem Aufwand und ohne Brüche konstruieren lässt.

Vier Aspekte des konzeptionellen Rahmens für die Methodenkombination

In diesem Beitrag stellen wir Kernpunkte eines generalisierten arbeitspraktischen Vorgehensmodells dar, das wir aus den Erfahrungen des Stuttgarter *Zentrums für reflektierte Textana-*

lyse (CRETA[2]) heraus entwickelt haben, in dem Beteiligte aus einer Reihe von unterschiedlichen Textwissenschaften und Computerwissenschaften kooperieren. Wir beschränken den Forschungsgegenstand auf Texte, öffnen den Raum aber für sehr unterschiedliche Arten von Fragestellungen: eine literaturwissenschaftliche Auseinandersetzung mit (poetischen) Primärtexten soll ebenso abgedeckt werden wie die Analyse von wissenschaftlichen Diskursen (etwa in der Philosophie) oder von Texten als Quellen für historische oder sozialwissenschaftliche Untersuchungen. Eine paradigmatische Auswahl erster Forschungsresultate der am Zentrum angesiedelten Projekte bietet der Band Kuhn/Pichler/Reiter (erscheint).[3]

Die Motivation für den vorgeschlagenen konzeptionellen Rahmen liegt nicht vordringlich in einer deskriptiven wissenschaftstheoretischen Betrachtung. Vielmehr soll der Rahmen Ansatzpunkte für die konkrete Praxis bieten – etwa für *Best-Practice*-Vorschläge. Wir fokussieren auf vier ineinandergreifende Aspekte, die für die DH zentral sind. Keiner dieser Aspekte ist grundsätzlich neu für die Methodendiskussion – hier geht es jedoch um eine handhabbare Gesamtkonzeption.

Als konkrete Illustration des Vorgehens mögen Arbeiten aus dem QuaDramA-Projekt dienen (Krautter/Pagel 2019, Krautter et al. 2018): ein Korpus von deutschsprachigen Dramen wird mit kombinierten Methoden erschlossen; ein exemplarischer Analyseschritt dabei liegt in der Klassifikation von Figuren nach bestimmten Typen. Einige Figurentypen sind bereits literaturwissenschaftlich etabliert (zärtlicher Vater) oder lassen sich relativ treffsicher aus der Figurentafel (Tochter) oder den Metadaten (Titelfigur) extrahieren. Andere Typen wie z.B. die Protagonistin bzw. der Protagonist entziehen sich einer aus unmittelbar verfügbaren Texteigenschaften oder Metadaten ableitbaren Zuweisung, sind jedoch von Bedeutung für literaturhistorische Betrachtungen (etwa für die Frage, inwieweit Emilia Galotti als Titelfigur aus G. E. Lessings bürgerlichem Trauerspiel (1772) den Status einer Protagonistin hat). Eine Annäherung an derartige Kategorien mit kombinierten Methoden kann ausgehend von klaren Fällen eine vorläufige Operationalisierung ansetzen, darauf aufbauend datenbasierte Computermodelle erzeugen und die Modellvorhersagen auf dem Gesamtkorpus in den Prozess einer Verfeinerung der Operationalisierung einfließen lassen.

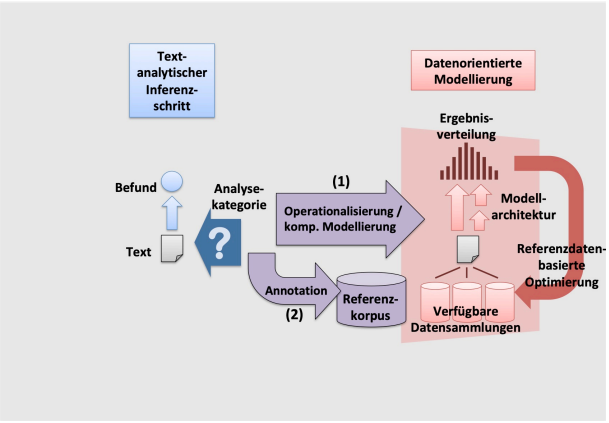


Abbildung 1: Typisches Vorgehen bei der komputationellen Modellierung nicht-trivialer Kategorien in der DH-Textanalyse

Als Ausgangspunkt skizziert Abbildung 1 ein DH-Vorgehen, das sich bei nicht-trivialen Modellierungsaufgaben etabliert hat – in Anlehnung an ausgeprägte methodische Konventionen in der Korpus- und Computerlinguistik (vgl. u.a. Hovy/Lavid 2010, Kuhn/Reiter 2015, Stefanowitsch 2018, Kuhn 2019): Die Analysekategorie, die im Rahmen einer geisteswissenschaftlichen Gesamtfragestellung angewendet werden soll, wird konzeptuell operationalisiert – gängiger Weise in Form von präzisen Annotationsrichtlinien. Der erste zentrale Aspekt für eine effektive Praxis der Methodenkombination liegt in der **(I) Anwendung der Annotationsrichtlinien auf ein geeignetes Korpus von Referenzdaten**. Die Referenzannotation kann anschließend durch die datenbasierte Entwicklungsmethodik der Computerwissenschaften für die Modelloptimierung herangezogen werden:[4] sie fixiert das Ziel für eine Optimierung der Vorhersagekraft von möglichen Modellarchitekturen und deren Parametrisierungen.[5]

Das bislang geschilderte Vorgehen fokussiert ausschließlich auf die technische Optimierung der Vorhersagemodelle für fixierte Referenzdaten. Ein effektiver konzeptioneller Rahmen für die Methodenkombination muss daneben Prozessen Raum geben, die eine sukzessive Verfeinerung der Analysekategorien vornehmen, um einem geisteswissenschaftlichen Fragenkomplex gerecht zu werden. Hier sind mehrere Aspekte zu unterscheiden: **(II)** eine **Inspektion der Vorhersageergebnisse** der (technisch optimierten) Modelle kann empirische Indikatoren zu Tage fördern, die Anlass zu einer **Revision der Operationalisierung** geben. Neben dem Entwicklungszirkel auf der rechten Seite muss es also einen Zirkel auf der linken Seite geben. Ein solches Revisionsmodell ist im Rahmen eines *Prototyping*-Ansatzes bzw. in der agilen Softwareentwicklung verbreitet – unabhängig davon, ob sich die Zielkategorisierung selbst in einem technischen Rahmen bewegt oder ob sie einem von der computerwissenschaftlichen Methodik abweichenden Rahmen entstammt.[6]

Letzteres ist allerdings bei einer komputationell/geisteswissenschaftlichen Methodenkombination der Fall. Die Gütekriterien, die zur Revision einer geisteswissenschaftlichen Analysekategorie führen, können grundsätzlich anderen methodischen Prinzipien und Vorannahmen folgen. Eine probehalber vorgenommene Operationalisierung eines vielschichtigen Konzepts in der Textanalyse (z.B. Protagonisten in Dramen) mag sich zum Beispiel als unergiebig erweisen, obgleich die komputationelle Umsetzung entsprechend den Referenzdaten eine hohe Vorhersagequalität ermöglicht. **(III)** der **textanalytische Inferenzschritt** (für den eine Computermodellierung vorgenommen wird) und die **datenbasierte Modellierung** sind jeweils in **eigene methodische bzw. arbeitspraktische Rahmen** eingebettet; zentrale Aufgabe für die „Mixed methods"-Forschung ist die Spezifikation von adäquaten Übersetzungsschritten.

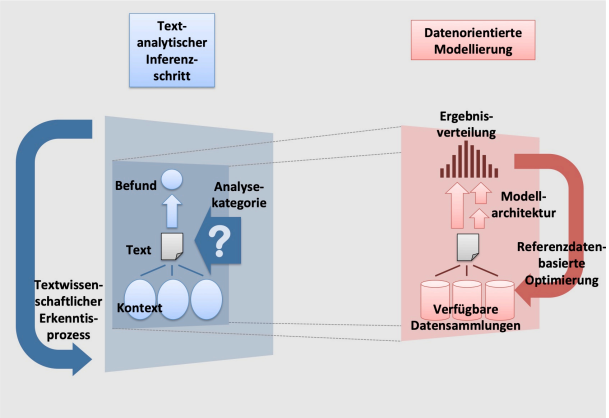


Abbildung 2: Konzeptionelle Trennung der arbeitspraktischen/methodischen Rahmen für einen textanalytischen Inferenzschritt und seine datenorientierte Modellierung; beide folgen eigenen Zyklen der Verfeinerung/Weiterentwicklung

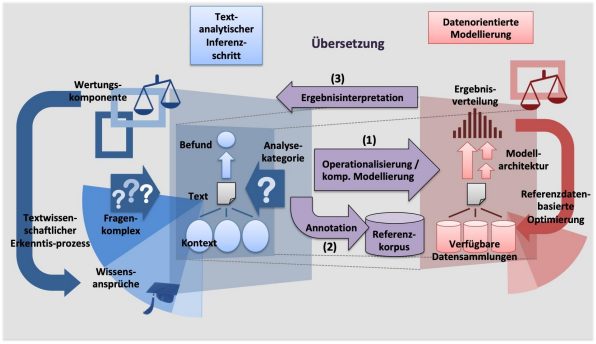


Abbildung 3: Bei Bedarf können die Wissensbereiche detailliert ausspezifiziert werden, die in die Methodenkombination gehen – aufbauend auf der wissenschaftstheoretischen Konzeption disziplinärer Praxis als Bearbeitung eines wissenschaftstheoretischen Problems nach Danneberg/Albrecht 2017 (mit den Komponenten (i) problematisierte Wissensansprüche, (ii) Wertungskomponenten, (iii) Fragenkomplex)

Abbildung 2 zeigt entsprechend eine stärker ausdifferenzierte Skizze des konzeptionellen Rahmens. Der vierte Aspekt ist hier bereits angedeutet: **(IV)** Ein bestimmter **Analysebefund zu einem Text** hinsichtlich einer Analysekategorie ist **stets in Bezug auf einen angenommen Interpretationskontext** (mit potenziell vielfältigen relevanten Dimensionen) zu sehen.[7] Ein textanalytischer Inferenzschritt (die innere blaue Box) wird also dargestellt als Ableitung eines Befundes aus einem Text, gegeben eine bestimmte Instantiierung der Kontextfaktoren, für die teilweise eigene Analysekategorien angesetzt werden müssen. (Zum Beispiel könnte die Operationalisierung der Kategorie *Protagonist* verwoben sein mit einer Analyse der aktiven und der passiven Präsenz der Figur, welche jeweils als eigene Kategorien zu operationalisieren sind.)

Wie Abbildung 2 suggeriert findet die Übersetzung aus der geisteswissenschaftlichen in die Sphäre der datenbasierten Computermodellierung sinnvollerweise für einzelne Inferenzschritte separat statt (obgleich wie in Fn. 5 angedeutet die Computerarchitektur für einen Schritt selbst technisch komplex sein kann). Es wird deutlich, dass bei der Bearbeitung von nicht-trivialen Fragestellungen rasch ein vielschichtiges Geflecht von Komponenten mit unterschiedlichem Status entsteht. Ein Hauptziel der hier vorgeschlagenen Konzeption liegt darin, das Augenmerk auf genau jene Statusunterschiede zu lenken, die für ein methodisch reflektiertes Vorgehen zu relevanten Vorannahmen relevant sind.

Strukturell sind die Elemente unserer Konzeption trotz der darstellbaren Komplexität vergleichsweise einfach – sie konzentrieren sich auf die Aufgabe der methodenübergreifenden Übersetzung mittels der referenzdatengestützten Operationalisierung und komputationellen Modellierung. Für die konkrete arbeitspraktischen Projektroutine sollte also eine verhältnismäßig übersichtliche Sicht auf die relevanten Komponenten möglich sein. Die abschließende Abbildung 3 demonstriert jedoch, dass der konzeptionelle Rahmen bei Bedarf eine Schnittstelle zu einer sehr differenziert ausgearbeiteten wissenschaftstheoretischen Konzeption wie der von Danneberg/Albrecht 2017 bietet. (Aus Platzgründen können wir in diesem Abstract nicht auf Details eingehen.) Ein reflektiertes Vorgehen kann also auch auf fundamentalere Fragen etwa zur Problematisierung von divergierenden Wissensansprüchen über Disziplingrenzen hinweg eingehen.

Fußnoten

1. Zentral für unsere methodologische Argumentation ist der Anspruch, Computermodelle für die Ableitung von einzelnen Befunden aus der Datenlage so in kombinierte Arbeitspraktiken zu integrieren, dass folgende zwei Aspekte unabhängig voneinander reflektiert werden können: (a) Wie hoch liegt die komputationelle Vorhersagequalität des Modells bei der Anwendung auf den Untersuchungsgegenstand (Texte einer bestimmten Epoche, Gattung, Inhaltsdomäne etc.) bezogen auf eine definierte Analyseaufgabe (etwa die Extraktion aller textuellen Erwähnungen der Akteure in einem Text)? (b) Wie aussagekräftig sind Befunde aus einer gegebenen, definierten Analyseaufgabe bezogen auf eine übergeordnete geisteswissenschaftliche Forschungsfrage? Eine modular aufgebaute Architektur von Analyseschritten kann entlang dieser beiden Aspekte sukzessive für die Bearbeitung eines geisteswissenschaftlichen Forschungsgegenstands angepasst werden.

Fasst man die Idee der komputationell/geisteswissenschaftlichen Methodenkombination sehr weit, fallen natürlich alle Arbeiten aus dem Spannungsfeld der DH darunter: Bevor Standard-Werkzeuge (auch ohne eine Möglichkeit der Adaptation) etwa für explorative oder heuristische Zwecke eingesetzt werden, werden sich die Forschenden ein Bild zu deren Vorhersagequalität auf dem studienspezifischen Material machen. Und bevor in einem Machine-Learning-Projekt mit großem Zeitaufwand Computermodelle für eine Analyseaufgabe und ein bestimmtes Korpus optimiert werden, vergewissern sich die Beteiligten in der Regel, dass die ableitbaren Befunde einen relevanten Beitrag zu Fachfragen leisten. Insofern schlägt sich das Nebeneinander der beiden Aspekte mittelbar auf jede DH-Methodendiskussion nieder (vgl. etwa die Diskussion von computergestützten Analyseverfahren in unterschiedlichen DH-Fachkontexten in Reiche et al. 2014 oder die Diskussion der Rolle der Informatik in den DH bei Heyer/Niekler/Wiedemann 2014 oder Deck 2018). Die praktischen und theoretischen Implikationen für das weitere Vorgehen unterscheiden sich jedoch abhängig davon, ob eine parallel komputationell/geisteswissenschaftlich reflektierte Modularisierung der Analysekonzepte im

Kern des arbeitspraktischen Vorgehensmodells für ein DH-Projekt steht oder ob beispielsweise zunächst unüberwachte Verfahren explorativ eingesetzt werden (vgl. das in Allison et al. 2011 dokumentierte Experiment). Betroffen sind also zentrale methodische Fragen, etwa zur Rolle der Visualisierung im Erkenntnisprozess (hierzu fanden und finden vielfältige Diskurse statt, vgl. z.B. Ihde 1998, Schaal/Kath/Dumm 2016, Romele et al. 2018); auch die Frage nach dem Verhältnis zwischen "building" (als Erstellen von Computerprogrammen) und "studying" in den DH (vgl. die 2011 von Stephen Ramsay ausgelöste Kontroverse, Ramsay/Rockwell 2012) erhält in einem eng verzahnten Vorgehensmodell einen differenzierten Charakter.

Gerade angesichts vielschichtiger möglicher Implikationen zum methodischen Selbstverständnis erscheint es uns für den vorliegenden Beitrag sinnvoll, die Kernideen des verzahnten Vorgehensmodells zunächst systematisch zu diskutieren. Auf eine Vertiefung spezifischer methodologischer Querbezüge müssen wir aus Platzgründen verzichten.

2. CRETA (https://www.creta.uni-stuttgart.de) wurde 2016 im Rahmen der BMBF-Förderung für Digital Humanities-Zentren etabliert.

3. Aus Platzgründen können wir die Umsetzung des konzeptionellen Rahmens in einzelnen Projekten im vorliegenden Vortragsexposé nicht diskutieren.

4. Falls hinreichend viele Daten annotiert wurden, kann überwachtes Lernen auf Basis einer Aufteilung in Trainings- und Testdaten zur Anwendung kommen. Zumeist ist bei speziellen Kategorisierungsaufgaben jedoch die Menge der Referenzdaten so klein, dass sie lediglich als Testmenge dienen kann – etwa für die Adaptation von Modellen, die für vergleichbare Kategorien trainiert werden können, oder für unüberwachte Verfahren.

5. Die interne Modellstruktur wird im Rahmen eines solchen Vorgehens getrennt vom Modellierungsziel gesehen; sie kann zwar von systematischen Annahmen über die zu modellierende Aufgabe inspiriert sein, ausschlaggebend ist jedoch die Optimierung der fixierten Referenzannotationen. Häufig erzwingt die mangelnde Verfügbarkeit von unmittelbar verwendbaren Daten bzw. Annotationen eine Modellarchitektur, die von den kausalen Zusammenhängen der modellierten Domäne abweicht.

6. Einen Prototyping-Ansatz innerhalb der DH diskutieren El Khatib et al. 2019. Ein verzahntes zyklisches Vorgehensmodell ist auch in der Visualisierung/den *Visual Analytics* verbreitet (vgl. etwa Sacha et al. 2014, El-Assady et al. 2019 und die Beiträge in Butt et al. 2020). Iterative Annotationszyklen zur Optimierung der Operationalisierung werden beschrieben bei Pustejovsky/Stubbs, 2012, Bögel et. al. 2015 Gius/Jacke 2017 und Pagel et al. 2018. Eine ausführliche Diskussion des Bezugs zu *close/distant reading* in DH findet sich in Jänicke 2016, Kap. 2.

7. Wir schließen uns hier der Konzeptualisierung an, die Danneberg/Albrecht 2017 im Rahmen einer möglichst allgemeinen wissenschaftstheoretischen Charakterisierung der Inferenzschritte vornehmen, welche einer literaturwissenschaftlichen Textinterpretation zugrunde liegen. Eine Verallgemeinerung auf andere textanalytische Problemstellungen ist ohne Weiteres möglich.

Bibliographie

Allison, Sarah / Heuser, Ryan / Jockers, Matthew / Moretti, Franco / Witmore, Michael (2011): "Quantitative Formalism: An Experiment", in: *Pamphlets of the Stanford Literary Lab* 1.. http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [Letzter Zugriff: 05.01.2020]

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: *DH-Commons Journal* 1 10.5281/zenodo.3240591

Butt, Miriam / Hautli-Janisz, Annette / Lyding, Verena (2020): *LingVis: Visual Analytics for Linguistics* . CSLI lecture notes. Stanford: CSLI Publications.

Danneberg, Lutz / Albrecht, Andrea (2017): "Beobachtungen zu den Voraussetzungen des hypothetisch-deduktiven und des hypothetisch-induktiven Argumentierens im Rahmen einer hermeneutischen Konzeption der Textinterpretation", in: *Journal of Literary Theory* 10(1): 1–37.

Deck, Klaus-Georg (2018): "Digital Humanities – Eine Herausforderung an die Informatik und an die Geisteswissenschaften", in: Huber, Martin / Krämer, Sybille (eds.): *Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden.* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3). PDF Format ohne Paginierung. 10.17175/sb003_002.

El-Assady, Mennatallah / Jentner, Wolfgang / Sperrle, Fabian / Sevastjanova, Rita / Hautli-Janisz, Annette / Butt, Miriam / Keim, Daniel A. (2019): " lingvis.io - A Linguistic Visual Analytics Framework " , in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 13-18.

El Khatib, Randa / Wrisley, David J. / Elbassuoni, Shady / Jaber, Mohamad / El Zini, Julia (2019): "Prototyping Across the Disciplines" , in: *Digital Studies/le Champ Numérique* , 8(1), 10 10.16995/dscn.282

Gius, Evelyn / Jacke, Janina (2017): „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis", in: *International Journal of Humanities and Arts Computing* 11(2): 233–254.

Ihde, Don (1998): *Expanding Hermeneutics: Visualism in Science* , Evanston, Ill.: Northwestern University Press.

Jänicke, Stefan (2016): *Close and Distant Reading Visualizations for the Comparative Analysis of Digital Humanities Data* . Dissertation, Universität Leipzig. http://www.informatik.uni-leipzig.de/~stjaenicke/dissertation.pdf [Letzter Zugriff: 05.01.2020]

Heyer, Gerhard / Niekler, Andreas / Wiedemann, Gregor (2014): "Brauchen die Digital Humanities eine eigene Methodologie? Überlegungen zur systematischen Nutzung von Text Mining Verfahren in einem politikwissenschaftlichen Projekt", in: *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)* http://asv.informatik.uni-leipzig.de/publication/file/255/Heyer-Brauchen_die_Digital_Humanities_eine_eigene_Methodologie__berlegungen-1451050.pdf [Letzter Zugriff: 05.01.2020]

Hovy, Eduard / Lavid, Julia (2010): "Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics", in: International Journal of Translation , 22(1): 13–36.

Kuhn, Jonas (2019): "Computational text analysis within the humanities: How to combine working practices from the contributing fields?", in: *Language Resources & Evaluation* 53: 565–602 https://doi.org/10.1007/s10579-019-09459-3 .

Kuhn, Jonas / Reiter, Nils (2015): "A plea for a method-driven agenda in the Digital Humanities", in: *Proceedings of Digital Humanities 2015, Sydney, Australia,* (*June 2015*).

Kuhn, Jonas / Pichler, Axel / Reiter, Nils (eds.) (erscheint): *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt.* Berlin: de Gruyter.

Krautter, Benjamin / Pagel, Janis / Reiter, Nils / Willand, Marcus (2018): "Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen", in: LitLab Pamphlets 7. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf [Letzter Zugriff: 05.01.2020].

Krautter, Benjamin / Pagel, Janis (2019): "Klassifikation von Titelfiguren in deutschsprachigen Dramen und Evaluation am Beispiel von Lessings 'Emilia Galotti'," in: *Konferenzabstracts DHd 2019 Digital Humanities: multimedial & multimodal, Frankfurt am Main, März 2019* .

Pagel, Janis / Reiter, Nils / Rösiger, Ina / Schulz, Sarah (2018) : " A Unified Annotation Workflow for Diverse Goals " , in: Kübler, Sandra / Zinsmeister, Heike (eds.) : *Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI 2018* .

Pustejovsky, James / Stubbs, Amber (2012). *Natural language annotation for machine learning* . Sebastopol: O'Reilly Media.

Ramsay, Stephen / Rockwell, Geoffrey (2012): "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities", in: Gold, Matthew K. (ed.): *Debates in the digital humanities.* Minneapolis and London: University of Minnesota Press, 2012: 75–84 10.5749/minnesota/9780816677948.003.0010.

Reiche, Ruth / Becker, Rainer / Bender, Michael / Munson, Matt / Schmunk, Stefan / Schöch, Christof (2014): *Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften* . DARIAH-DE Working Papers. Göttingen: DARIAH-DE, 2014.

Romele, Alberto / Severo, Marta / Furia, Paolo (2018): "Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines", in: *AI & Society: Knowledge, Culture and Communication* , Springer, in press. https://hal.archives-ouvertes.fr/hal-01824173 [Letzter Zugriff: 05.01.2020]

Sacha, Dominik / Stoffel, Andreas / Stoffel, Florian / Kwon, Bum Chul / Ellis, Geoffrey / Keim , Daniel A. (2014): "Knowledge Generation Model for Visual Analytics", in: *IEEE Transactions on Visualization and Computer Graphics* 20 (12) 1604–1613. 10.1109/TVCG.2014.2346481.

Schaal, Gary S. / Kath, Roxana / Dumm, Sebastian (2016): "New Visual Hermeneutics", in: *Cybernetics & Human Knowing* 23: 51–76.

Stefanowitsch, Anatol (2018): *Corpus Linguistics: A Guide to the Methodology* (Textbooks in Language Sciences 8). Open Review Version. Berlin: Language Science Press. www.langsci-press.org . [Letzter Zugriff: 05.01.2020]

Theatre-Tool: Erschließung, Verknüpfung und Web-Präsentation von Theater- und Musikbeständen mit unterschiedlichen Quellentypen

Capelle, Irmlind
irmlind.capelle@upb.de
Universität Paderborn, Deutschland

Richts, Kristina
krichts@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Schilke, Elena
eschilke@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Das sogenannte „Detmolder Hoftheater-Projekt" hat im Laufe der letzten fünf Jahre den überlieferten Musikalienbestand des Detmolder Hoftheaters aus der Zeit von 1825–1875 einschließlich der erhaltenen Aktenmaterialien erschlossen bzw. übertragen und mit einer eigenen Software im Web präsentiert (www.hoftheater-detmold.de). Die dabei eingesetzte Software „Theatre-Tool" wurde ausdrücklich so konzipiert, dass sie auf andere ähnliche Bestände übertragbar ist. Dieser Vortrag möchte die bisherigen Arbeitsergebnisse des Projekts zusammenfassen und die Erschließungsgrundsätze und den Aufbau und die Anforderungen der Software erläutern. Dabei werden auch die Möglichkeiten und Probleme der Übertragbarkeit und der Zusammenarbeit mit anderen bestehenden und neuen Erschließungsprojekten angesprochen.

Im Bereich der digitalen Edition ist es inzwischen zum Standard geworden, Kontextmaterialien im Rahmen der Edition ebenfalls als Volltext bereit zu stellen und über Markup zu verknüpfen (vgl. die für den Musiktheaterbereich vorbildliche Präsentation https://freischuetz-digital.de/). Doch im Bereich der Erschließung von Beständen wird noch sehr viel mit proprietären Datenbanken bzw. bibliothekarischen Standards gearbeitet, die nicht ohne weiteres im Web zugänglich gemacht werden können, und werden vor allem unterschiedliche Quellentypen getrennt in je eigenen Systemen erfasst. Dies zeigt sich z. B. in der zur Zeit im Bibliotheksbereich geführten Diskussion um die Erfassung von Ephemera (siehe Literaturverzeichnis), die gerade im Bereich der für die Theaterforschung wichtigen Erschließung von Theaterzetteln zu zahlreichen Insellösungen geführt hat (vgl. z. B. http://digital.ub.uni-duesseldorf.de/theaterzettel, http://www.theater-

zettel-weimar.de), was eine übergreifende Suche z. B. nach Darstellernamen unmöglich macht.

Das Hoftheater-Projekt hat ein Modell zur kontextuellen Erschließung der verschiedenen Materialien entwickelt, auf dem das in seiner Oberflächengestaltung zunächst noch rein funktionale Theatre-Tool aufbaut (https://hoftheater-detmold.de/47-2/das-modell/) .

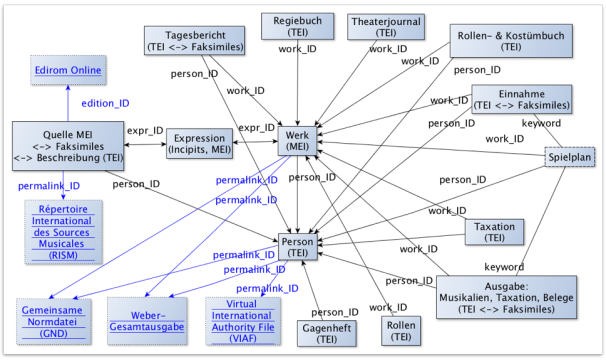


Abbildung 1: Modell des „Theatre Tool“

Dieses Modell basiert auf dem im Bibliotheksbereich allgemein angewandten FRBR-Modell, so dass die erfassten Daten sowohl bibliothekarische als auch wissenschaftliche Anforderungen erfüllen. Die Erschließung der Quellen erfolgt nach FRBR auf drei verschiedenen Ebenen: Die Werkdateien erfassen die Grunddaten ggf. mit dem Datum der Uraufführung und einer normierten Angabe zur Klassifikation. Die Quellendateien (entspricht der FRBR-Entität: manifestation) beschreiben die vorliegenden Quellen, die in unserem Fall zu einer „componentGroup“ zusammengefasst werden, da die Aufführungsmaterialien eine Einheit bilden. Bindeglied zwischen Werk und Quelle ist die expression-Datei, denn das jeweilige Aufführungsmaterial des Theaters in Detmold ist als Einheit eine expression des Werks, dasjenige eines anderen Theaters jedoch eine andere. Auch wäre beispielsweise eine Bearbeitung der Oper für Bläser-Ensemble wiederum eine weitere expression. Die Beziehungen zwischen den Dateien werden mit Relationen beschrieben, wie sie durch FRBR vorgegeben sind: „hasRealization“, „isEmbodimentOf“, „hasEmbodiment“, „isPartof“ etc.

Zusätzlich zu diesen zur Quellenerschließung notwendigen Dateien werden solche zu Personen und dramatis personae angelegt. Durch eine jeweils eindeutige ID, mit der jede Datei gekennzeichnet wird, ist bei jeglicher Wiederkehr eines Werk-, Rollen- oder Personennamens eine eindeutige Kennzeichnung möglich.

Handelt es sich bei diesen Dateien um typische Katalog-Erschließungen, die jedoch zumindest bei den Quellendateien weit über eine übliche bibliothekarische Erfassung hinausgehen, so werden die umfangreich überlieferten Kontextmaterialien z. T. als Regeste, überwiegend aber im Volltext erfasst. Beide Erschließungsformen basieren auf den XML-Standards TEI und MEI, so dass alle Daten durch ein Markup ausgezeichnet werden können.

Darüber hinaus werden für Personen, Werke und ggf. Orte Normdaten (GND, VIAF, GeoNames) verwendet, so dass externe Informationen eingebunden werden können. Da aber etliche Personen und Werke wenig bekannt oder nicht eindeutig zu bestimmen sind und damit nicht eindeutig einer Normdaten-ID zugeordnet werden können, bleibt die Verwendung von eigenen IDs notwendig.

Durch die Verknüpfung der Daten über alle Quellengrenzen hinweg ergeben sich verschiedene inhaltliche Verbindungen: So können zu den Personen des Detmolder Hoftheaters einerseits die Daten zu Gage und eventuellen Sonderzuwendungen, Beschäftigungsdauer und zusätzliche Beschäftigungen im Theaterbetrieb abgerufen werden, andererseits aber auch die Werke und sogar die Rollen, in denen sie beschäftigt waren. Zu den Aufführungsmaterialien werden aus den Akten Angaben zur Datierung und zum Schreiber verknüpft und die Einträge in den Kostüm- und Regiebüchern geben erste Hinweise auf die Darstellung einzelner Werke auf der Bühne.

Die Materialien (Digitalisate ausgewählter Quellen) und die XML-Dateien der Katalog- oder der Volltext-Erschließungen werden in einer Web-Präsenz zusammengefasst.

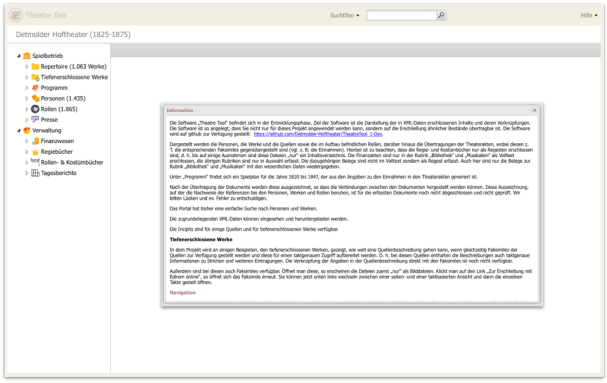


Abbildung 2: Startseite des Portals

Die hierfür eigens entwickelte Software „Theatre-Tool“ basiert auf XQuery und JavaScript.

Die Darstellung der Faksimiles erfolgt mit Hilfe eines Leaflet-Plugins (https://leafletjs.com), einer Bibliothek für die Kartendarstellung im Web.

Die Software bietet bislang eine einfache Suche nach Personen, Rollen und Werken, die mit einem Fuse.js Plugin, einer Fuzzy basierten Bibliothek, erstellt wurde.

Wie in Web-Präsenzen üblich, können die Inhalte als XML-Dateien heruntergeladen werden, um weitere Arbeiten mit den Daten zu ermöglichen (Suche, Abfrage in größerem Kontext etc.). Selbstverständlich können die Daten auch als Beispiele für eine Erschließung in anderen Projekten verwendet werden.

Die Werke, Quellen, Personen und Rollen können mit Hilfe von Permalinks von anderen Projekten direkt referenziert werden.

Da bislang im Detmolder Hoftheater-Projekt vor allem Materialien zum Musiktheater erschlossen worden sind, sind in die Software einige musik-spezifische Anwendungen integriert. So werden z. B. die Anfänge der einzelnen Musiknummern mit Noten-Incipits wiedergegeben, um sie rasch vergleichbar zu machen. Um dem Musikwissenschaftler auch Informationen zur originalen Partituranordnung, Schlüsselung, Schreibweise der Instrumente etc. zu geben, wird nicht nur – wie traditionell üblich – eine Stimme oder ein Klavierauszug wiedergegeben, sondern werden die ersten Takte und der Singstimmeneinsatz vollständig in Partitur wiedergegeben. Die Codierung der Incipits erfolgt mit MEI, die Darstellung mit einem Verovio-Plugin (http://www.verovio.org/index.xhtml).

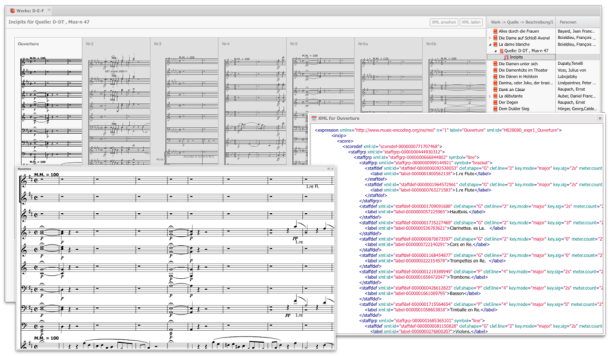


Abbildung 3: Darstellung der Incipits

Eine weitere Besonderheit des Projekts ist die exemplarische sog. Tiefenerschließung einiger ausgewählter Aufführungsmaterialien: Bei diesen werden auch die Faksimiles der Quellen zur Verfügung gestellt und zwar in einer Aufbereitung für einen taktgenauen Zugriff. Nur durch diese Form der Erschließung ist es z. B. möglich, Eingriffe in den Notentext nicht auf Grund der Materialität (also z. B.: Streichung auf Bl 4v bis 5r vorletzter Takt), sondern inhaltlich (z. B.: Streichung in Nr. 1 von T. 17–20) und damit für den Nutzer (mit Hilfe anderer Materialien, also gedruckter oder anderer handschriftlicher Quellen) nachvollziehbar anzugeben. Zur Erstellung der sog. Vertaktung wird die Software „Edirom" benutzt und zur Darstellung ist das „Theatre Tool" mit Edirom Online (https://github.com/Edirom/Edirom-Online) verknüpft. Diese Software wurde zwar für die Aufbereitung von Notenmaterial entwickelt, aber es lassen sich damit auch Textquellen z. B. nach Szenen oder sogar Zeilen kartieren.

Das Theatre-Tool ist für die Darstellung dieser komplexen Text- und Datenstrukturen entwickelt, kann aber leicht an andere Anforderungen angepasst werden: Bei dem im Projekt erfassten Material handelt es sich z. B. überwiegend um handschriftliches Material, weshalb die nach FRBR vorgesehene vierte Ebene, das Exemplar (item), nach der Regel der „manifestation singleton" nicht berücksichtigt wird. Selbstverständlich wäre aber auch diese darstellbar. Da das Hauptinteresse der Erschließung auf der Arbeitsweise und dem Personal der Detmolder Hoftheater-Gesellschaft liegt, werden die erwähnten Orte zwar ausgezeichnet, gibt es für diese aber keine eigenständigen Dateien (mit der Möglichkeit zu Referenzen) und bislang keine Suchmöglichkeit.

Mit der zunehmenden Digitalisierung der Bestände durch die Bibliotheken könnten diese über iiiF in das Theatre Tool eingebunden werden, wodurch etliche rechtliche Probleme gelöst werden könnten. Wie damit auch eine Vertaktung verbunden werden kann, wäre zu überprüfen.

Weiterer Abstimmungsbedarf, an dem aber beidseitig großes Interesse besteht, ist notwendig zwischen Wissenschaft und Bibliothek. Es ist selbstverständlich, dass die Beispiele der Tiefenerschließung des Projekts ebenso wie die Erstellung z. B. von Komponisten-Werkverzeichnissen nur durch die Wissenschaft zu leisten sind. Dennoch besteht großes Interesse, diese Detailinformationen zu einzelnen Quellen auch über die besitzende Bibliotheken zugänglich zu machen Die Verwendung von Standards und Normdaten wie sie im Hoftheater-Projekt erprobt worden sind, bildet hierzu einen erster Schritt, doch muss sicherlich auch verstärkt über Schnittstellen für den Datenaustausch nachgedacht werden.

Bibliographie

Kamzelak, Roland S. (2016): „Digitale Editionen im semantic web. Chancen und Grenzen von Normdaten, FRBR und RDF" in: Richts, Kristina / Stadler, Peter (eds.): *„ei, dem alten Herrn zoll ich Achtung gern".* Festschrift für Joachim Veit zum 60. Geburtstag. München: Allitera Verlag 423–435; online unter: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-233392

Münzmay, Andreas (2018): „Lesen und Schreiben im digitalen Dickicht, Musikwissenschaft, Digital Humanitites und die hybride Musikbibliothek" in: BIBLIOTHEK Forschung und Praxis 42; 236–246.

Münzmay, Andreas (2019): „Kulturtransferforschung und Musikwissenschaft", in: Calella, Michele / Leßmann, Benedikt (eds.): *Zwischen Transfer und Transformation: Horizonte der Rezeption von Musik* (= Wiener Veröffentlichungen zur Musikwissenschaft 51). Wien 175–190.

Richts, Kristina / Veit, Joachim (2018): „Stand und Perspektiven der Nutzung von MEI in der Musikwissenschaft und in Bibliotheken" in: BIBLIOTHEK Forschung und Praxis 42: 292–301.

Pernerstorfer, Matthias J. (2012): *Theater – Zettel – Sammlungen. Erschliessung, Digitalisierung, Forschung*. Wien (= Don Juan Archiv Wien: Bibliographica, 1)

Pernerstorfer, Matthias J. (2015): *Theater – Zettel – Sammlungen* Bd. 2: *Bestände, Erschließung, Forschung.* Wien 2015 (= Don Juan Archiv Wien: Bibliographica 2)

Veit, Joachim (2020): „Notistenspezifische Erwartungen der Wissenschaft an die Web-Präsentation digitalisierter Musikhandschriftenbestände" in *: Das Instrumentalrepertoire der Dresdner Hofkapelle in den ersten beiden Dritteln des 18. Jahrhunderts – Überlieferung und Notisten.*

Wiermann, Barbara (2018a): „Bibliothekarische Normdaten und digitale Musikwissenschaft" in: *Die Musikforschung*, 71: 338–357.

Wiermann, Barbara (2018b): „musiconn. performance: musikalische Ereignisdaten im Fachinformationsdienst Musikwissenschaft" in: Bonte, Achim / Rehnolt, Juliane (eds.): *Kooperative Informationsinfrastrukturen als Chance und Herausforderung*. Thomas Bürger zum 65. Geburtstag herausgegeben von. Berlin, Boston 398–415.

The rapid rise of Fraktur

Weichselbaumer, Nikolaus

nikolaus@weichselbaumer.info
JGU Mainz, Deutschland

Seuret, Mathias

mathias.seuret@fau.de
FAU Erlangen, Deutschland

Limbach, Saskia

limbach@uni-mainz.de
JGU Mainz, Deutschland

Hinrichsen, Lena
lhinrich@students.uni-mainz.de
JGU Mainz, Deutschland

Maier, Andreas
andreas.maier@fau.de
FAU Erlangen, Deutschland

Christlein, Vincent
vincent.christlein@fau.de
FAU Erlangen, Deutschland

Introduction

From the first experiments in 1513, Fraktur quickly became the most successful gothic font in print history. Whereas gothic fonts in most other countries went out of use in the 16th and 17th centuries, Fraktur became by far the most used font for German texts in the early modern period. The font also made it to modernity and was used frequently, almost unchanged, until the middle of the 20th century. Even today the font is often used especially when a design should appear 'historical'.

Despite its importance, fairly little is known about the famous font. The origins of Fraktur at the beginning of the 16th century and the possible creators Vincenz Rockner and Johann Neudörffer have been the subjects of several studies (Kautzsch 1922, Kapr 1993: 24, Hessel 1937). Apart from this, however, we know remarkably little about its development over the following centuries. Only the Antiqua-Fraktur dispute around 1800 gained the interest of book historians again when German intellectuals discussed which of the two fonts is more appropriate for German texts (Lühmann 1981, Killius 1999). Yet the emergence of Fraktur and its leading role in font history remains understudied.

Tracing the emergence of Fraktur is complicated by two facts: On the one hand, contemporary evidence, such as invoices, letters and type specimens, is at best fragmentary and nearly impossible to contextualise without an analysis of the books themselves. On the other hand, researchers are simply overwhelmed by the amount of material available. For the 16th century alone, the German national bibliography VD16 (www.vd16.de) lists over 100,000 titles. This makes it impractical to look at every book individually and determine its fonts or even only its main text font.

Recent research presents a solution to this problem. With the help of a newly developed pattern recognition tool, large amounts of digitised book pages can be categorised into font groups. This tool was developed in the context of a project on font-specific OCR (Weichselbaumer et al. 2019, Seuret et al. 2019) and was then used for a large dataset of digitised books from BSB Munich. This paper will present the results and provide new insights into the rapid rise of Fraktur.

Methodology

Our methodology is based on automatic document image labeling which is done by a deep convolutional neural network (CNN) trained for font classification. As artificial intelligence typically requires a great amount of data, we manually prepared a training dataset of more than 35'000 document images, each labeled with the used fonts. We recently published this dataset along with a complete description of the approach we used (Seuret et al. 2019). For these test pages, we reach an accuracy slightly higher than 98% for recognising the main font.

As CNN architecture, we employ a DenseNet-121 (Huang et al. 2017). It is composed of 121 neural layers, most of them contained in 4 densely connected blocks. To identify the main font in a document image, we split it into many overlapping 224x224 px large patches, which are subsequently passed to the CNN. The overall page result is obtained by taking the average of all classified patches. Processing pages patch-wise is significantly more memory-friendly than using fully-convolutional neural networks and does not require expensive hardware.

For this study, book processing was done in two steps. First, we extracted the production years and the language of digitised books from the available metadata. We disregarded books that were not tagged as German as well as those without a clear date of publication (using 15 processing rules for the dates, in addition to an extra-permissive roman numbers parser). Second, we identified the main font of the pages 10 to 19 of every book, thereby avoiding prefaces and title pages which can differ quite decisively from the rest of the book. In case the network did not detect the same font on at least 6 pages of the same book, we disregarded the entire 10 pages. This way we automatically labelled 10 pages of a total of 85'165 books as the basis for this study.

Library catalogues are a great resource for metadata. Yet, in many cases early printed books were collated differently, even within the same library. This is largely the result of changing bibliographical practices in the past decades in which only slowly a standard emerged. Therefore, we often find metadata that is not standardised.The date of publication is often far from straight-forward. It may just be an estimation with words like 'ca.', 'um' or 'etwa'; it may include two years, such as 1549/50; or it may be displayed in Roman numerals, which sometimes differ from modern-day practice, such as 'MDXXXX'.

In parsing this data, we attempted to keep as much usable data as possible without distorting the results. Roman numerals were transformed into arabic numbers. In case two years are given (e.g. 1549/50), the first and second year were alternated. If the year was given as a time span (e.g. 1650-1660) we computed the average of both values and rounded to the larger number when necessary. For the estimated dates we decided that omitting the records altogether would have shrunk the database considerably. So we just deleted the estimation markers like 'ca.' and kept the years. This produced larger spikes every 50 years and smaller ones every 5 and 10 years, but these can be explained easily when interpreting the results.

After we received the results of the network we double-checked unlikely results by hand. This included some 60 books printed after 1550 which were classified as fonts predominantly used in the incunabula era - Rotunda, Textura, Gotico-Antiqua and Bastarda. In most cases these were actually empty pages with text bleeding through the other side of the page. Occasionally there were also tables with arabic numerals which were classified wrongly as one of the fonts mentioned above. We decided to delete this small number of misclassified books.

Results

The resulting data[1], which shows the publication of books in the German language, seems to be fairly representative of the general print production in Germany. After a relatively slow start up to 1520, the Reformation led to a very considerable spike in print production. The Thirty Years War (1618-1648) brought the print industry almost to a standstill. After that we see a steady rise in the number of editions per year, quickly accelerating at the end of the 18th century. Interestingly, the slight drop towards the very end of the century is rather unexpected. It may just be the result of the library's preference to digitise material pre-1800.

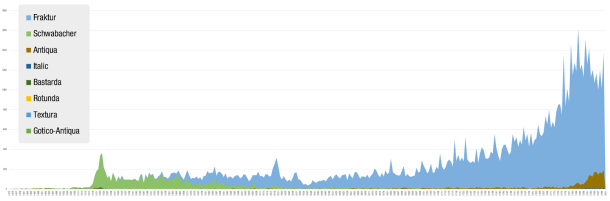


Figure 1: Main font groups in 85,223 digitised books from the Bavarian State Library, printed between 1472 and 1800[2]

With regard to font groups, the diagram showing absolute values (Fig. 1) stresses the fact that for German texts, Schwabacher and Fraktur are by far the two most important font groups. Yet, due to the much higher print production in the 18th century, Fraktur appears approximately 8.5 times as many times as a main font as Schwabacher.

In the results, you can also find a negligible number of Hebrew (4) and Greek (2) recognized as main fonts. They either actually aren't German (bsb10239978) pointing to a rare mistake in the metadata of the books, or contain pages with mainly Hebrew/Greek characters (bsb10779648 / bsb10360987). The book bsb11254779 (apparently written for Christians in Israel) is mainly Hebrew but has a German title. The *Catechismus D. Martini Lutheri minor: E lingua vernacula in Latinam & Graecam pridem translates* (bsb11229498) has been recognized as Greek although this is only true for 1/3 of the text.

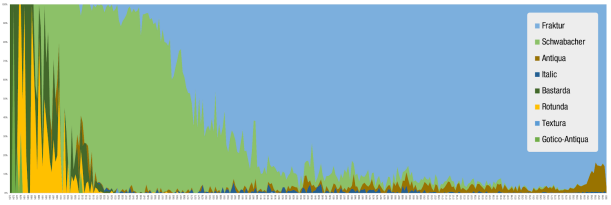


Figure 2: Main font groups in 85,223 digitised books from the Bavarian State Library, printed between 1472 and 1800, normalised in percentage

But absolute numbers only tell half the story. In order to know how important a font was at a given time, it is more fruitful to look at its share of the print production in a given year. In a normalised diagram of the same data we see much noise for the first decades of print up to the 1520ies. This is caused by the very low print production at that time and by the fact that printers often used fonts inconsistently as they did not have a complete set of font styles available. Nevertheless, the diagram shows that in the first decades of print, the two most important fonts for German texts seem to have been Bastarda and Rotunda. They were then gradually replaced by Schwabacher from the 1490ies onwards. Schwabacher reaches its largest share in the 1520ies to the 1540ies, the height of Reformation printing, before it is gradually replaced by a relatively new font - Fraktur. It is firmly established as the main font for German from about 1585 onwards. Only in the last decades of the 18th century does another font, Antiqua, become slightly more important in the production of German books. This indicates that the Antiqua-Fraktur debate had indeed some impact on contemporary book design. However, the overwhelming majority of books were still printed in Fraktur.

Conclusion

This study shows that Schwabacher dominated German language printing for the larger part of the 16th century until a fairly slow and linear change brought Fraktur to the dominant role it then kept through the 17th and 18th century. This makes the rise of Fraktur no less decisive, but significantly slower than often assumed (Kapr 1991, p 42; Killius 1999, p. 82).

These results have implications not only for the history of typography but also for OCR. When institutions use OCR engines, it is vital to choose the correct model for the specific text font. Quite commonly libraries use either Antiqua or Fraktur when a Schwabacher model or a mixed model could actually produce much better results, especially for books printed in the 16th century.

The used method promises to be a helpful and viable tool for digital book history. It paves the way for further studies on the statistical analysis of font use in early printed books and at the same time allows further research on the reasons for the change from Schwabacher to Fraktur. In addition, it offers the opportunity to shed more light on the role of type foundries in the development of book design in the early modern period.

Fußnoten

1. All data produced for this paper can be downloaded here: https://doi.org/10.5281/zenodo.3598515.
2. In this and the following figure, the font groups "other font" and "not a font" are not shown as they would mainly represent noise (images, blank pages, tables etc.).

Bibliography

Hessel, Alfred (1937): "Die Schrift der Reichskanzlei seit dem Interregnum und die Entstehung der Fraktur", in: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*. Philologisch-Historische Klasse. Fachgruppe 2: Nachrichten aus der Mittleren und Neueren Geschichte N. F. 2: 43-59.

Huang, Gao / Liu, Zhuang / van der Maaten, Laurens / Weinberger, Kilian Q. (2017): „Densely Connected Convolutional Networks", in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261-2269.

Kapr, Albert (1993): *Fraktur*. Form und Geschichte der gebrochenen Schriften. Mainz: Schmidt.

Kautzsch, Rudolf (1922): *Die Entstehung der Frakturschrift* (= Jahresbericht der Gutenberg-Gesellschaft 20, Beilage). Mainz: Gutenberg-Gesellschaft.

Kapr, Albert (1993): Fraktur. Form und Geschichte der Gebrochenen Schriften. Mainz: Hermann Schmidt.

Killius, Christina (1999): *Die Antiqua-Fraktur-Debatte um 1800 und ihre historische Herleitung* (= Mainzer Studien zur Buchwissenschaft 7). Wiesbaden: Harrassowitz.

Lühmann, Frithjof (1981): *Buchgestaltung in Deutschland 1770 - 1800*. PhD, Ludwig-Maximilians-Universität München.

Seuret, Mathias / Limbach, Saskia / Weichselbaumer, Nikolaus / Maier, Andreas / Christlein, Vincent (2019): "Dataset of Pages from Early Printed Books with Multiple Font Groups", in: *Historical Document Imaging and Processing (HIP) 2019, 5th International Workshop 21-22 September 2019, Sydney, Australia*.

Weichselbaumer, Nikolaus / Seuret, Mathias / Limbach, Saskia / Christlein, Vincent / Maier, Andreas (2019): "Automatic Font Group Recognition in Early Printed Books", in: *Digital Humanities im deutschsprachigen Raum (DHd) 2019. 6th International Conference 25-29 March 2019, Universitäten zu Mainz und Frankfurt* 84-87.

„The Vectorian" – Eine parametrisierbare Suchmaschine für intertextuelle Referenzen

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Liebl, Bernhard
Bernhard.Liebl@gmx.org
Computational Humanities Group, Universität Leipzig

Einleitung: Shakespeare, Intertextualität und computergestützte Erkennung von Zitaten

Shakespeare ist überall. Über alle zeitlichen und medialen Grenzen hinweg finden sich intertextuelle Bezüge auf die Werke von Shakespeare (vgl. Garber, 2005; Maxwell & Rumbold, 2018), der damit nicht nur der meistzitierte und meistgespielte Autor aller Zeiten, sondern auch der meistuntersuchte Autor der Welt ist (Taylor, 2016). Doch wenngleich in zahllosen Studien diverse Einzelaspekte von Shakespeares Werk aus Perspektive der Intertextualitätsforschung gründlich mittels *close reading* untersucht wurden, so gibt es bis heute keinen Überblick, kein Gesamtbild, keine systematische Karte intertextueller Shakespeare-Referenzen für größere Textkorpora. Auffällig ist zudem, dass bislang kaum Verfahren der computergestützten Erfassung intertextueller Shakespeare-Referenzen im Sinne des *distant reading* zum Einsatz kommen. Dies verwundert umso mehr, als dass sich im Bereich der Informatik und des *natural language processing* vielfältige Methoden zur Ermittlung der Ähnlichkeit zwischen Texten finden (Bär et al., 2012; Bär et al. 2015) – und nichts anderes ist Intertextualität letzten Endes. Natürlich ist hier anzumerken, dass die volle Bandbreite intertextueller Phänomene mit bloßen Mitteln der Textähnlichkeitsbestimmung nicht abgedeckt werden kann. Für unser Verständnis von Intertextualität berufen wir uns daher auf die Definition von Genette (1993) – "la présence effective d'un text dans un autre" – wobei wir unter der "effektiven Präsenz" eines Texts in einem anderen tatsächlich eine mehr oder weniger objektiv erkennbare, explizite Referenz an der Textoberfläche verstehen. Die textuelle Umschreibung einer Balkonszene mit einem Mann und einer Frau würden wir demnach nicht automatisch "Romeo and Juliet" zuordnen, was vermutlich auch nicht in allen Fällen korrekt wäre. Die folgende Variante eines bekannten Zitats aus Macbeth (Shakespeares Ursprungsvariante steht jeweils in eckigen Klammern) wäre nach unserem Verständnis hingegen objektiv aus dem Text zu erkennen und eindeutig als intertextuelle Referenz einzuordnen:

By the *stinking* [pricking] of my *nose* [thumbs], something *evil* [wicked] this way *goes* [comes]. *(Terry Pratchett: „I Shall Wear Midnight").*

Eine weitere methodische Einschränkung machen wir, indem wir Phänomene wie strukturelle Ähnlichkeit (Versmaß, Figurenkonstellation) und stilistische Ähnlichkeit[1], wie sie bspw. in der *Parodie* oder im *Pastiche* üblich sind, zunächst außer Acht lassen. In Erweiterung einer ersten Pilotstudie zur Identifizierung von Shakespearezitaten in der Fernsehserie „Dr. Who" (Burghardt et al., 2019) erproben wir in einem aktuellen Experiment das Potenzial von *word embeddings* (Mikolov et al., 2013), um so zusätzlich semantisch ähnliche oder zumindest "funktional äquivalente" (Bubenhofer, 2019) Wörter und Phrasen zu identifizieren. Durch die Auswahl unterschiedlicher *embeddings*-Modelle und weiterer, damit einhergehender Parameter (bspw. der Gewichtung anhand von Wortarten, dem Festlegen von Ähnlichkeitsschwellwerten, etc.) kann es mitunter zu sehr unterschiedlichen Ergebnissen kommen. Um hier systematisch Parameterkombinationen zu untersuchen, die möglichst optimierte Werte bzgl. *precision* und *recall* liefern, wurde im Sinne von Molnars (2019) Desiderat eines „interpretable machine learning" eine parametrisierbare Suchmaschine zur Identifizierung von Shakespeare-Referenzen als Vorstufe für einen *embeddings*-basierten Ansatz umgesetzt.

The Vectorian

Abb. 1 zeigt die Systemarchitektur der besagten Suchmaschine, die fortan als "The Vectorian"[2] bezeichnet wird. Im *Vectorian* fungieren kurze Shakespeare-Passagen (bspw. „If you prick us, do we not bleed?") als Queries; Texte, die diese Textteile (wortwörtlich oder als Variante) aufgreifen, stellen im Sinne des Information Retrieval dann die entsprechenden Ergebnisdokumente dar (für einen vergleichbaren Ansatz siehe Manjavacas et al., 2019).

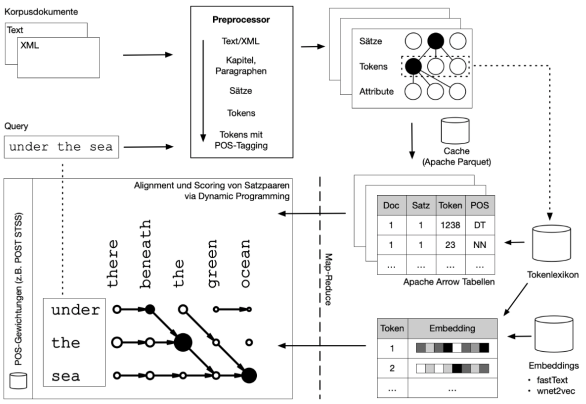


Abbildung 1: Systemarchitektur der Zitat-Suchmaschine *"The Vectorian"*.

Kern des *Vectorian* ist die Suche von optimalen semi-globalen *alignments* zwischen Satzpaaren (wobei wir einen Satz als Sequenz von Worten verstehen) über eine Variante des Needleman-Wunsch-Algorithmus (Sellers, 1974) mit sog. *free shift alignment*. Als Bewertungsfunktion nutzen wir eine über *word embeddings* errechnete Distanz zwischen Worten. Diesen Ansatz kombinieren wir mit einer Reihe experimenteller Parameter (siehe die fünf Punkte im nachfolgenden Abschnitt).

Abb. 2 zeigt das Frontend des *Vectorian*. Zu sehen ist ein Eingabefeld für beliebige Suchanfragen, d.h. die Textstellen, die man als intertextuelle Referenzen in anderen Texten finden will. Die Parameter der Suche, die nachfolgend noch näher erläutert werden, können über entsprechende Auswahlmenüs konfiguriert werden. Schließlich gibt der *Vectorian* eine Ergebnisliste zurück, deren Ranking dem jeweils höchsten Ähnlichkeitswert zwischen der Suchanfrage und einer entsprechenden Textstelle entspricht. Wortwörtliche Zitate haben demnach einen höheren Wert als stark abgeänderte Referenzen mit diversen Auslassungen und Substitutionen.

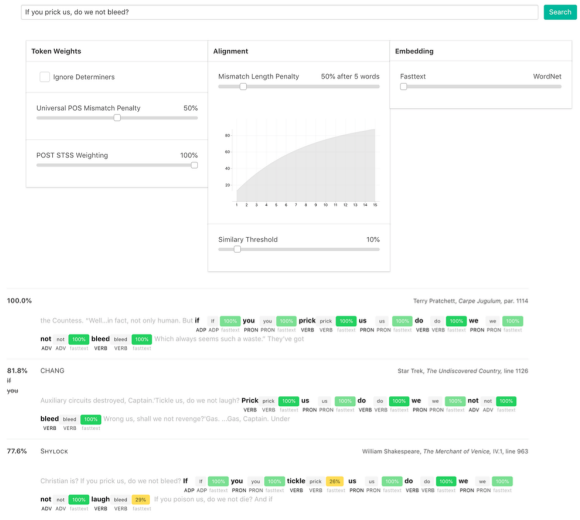


Abbildung 2: Frontend des *Vectorian* mit allen möglichen Suchparametern und einer beispielhaften Ergebnisliste für die Suchanfrage "If you prick us, do we not bleed?".

Der *Vectorian* durchsucht aktuell ein Korpus von 230 englischen Einzeltexten, darunter 50 Werke von Shakespeare (Dramen und Sonette) sowie diverse Romane aus unterschiedlichen Epochen und Transkripte von Filmen und Fernsehserien. Das Korpus enthält rund 19,5 Millionen Tokens mit POS-Annotationen (POS = *parts of speech*), die sich auf rund 2,2 Millionen Sätze verteilen. Der *Vectorian* bietet mit *fastText* (Mikolov, 2017) und *wnet2vec* (Saedi, 2018) momentan zwei *embedding*-Varianten zur Auswahl. Wir nutzen für *fastText* bestehende, vortrainierte Modelle (https://fasttext.cc/), für *wnet2vec* wurde ein eigenes *embedding* auf Basis unseres Korpus mit Hilfe einer leicht angepassten Referenzimplementierung von Saedi et al. (https://github.com/nlx-group/WordNetEmbeddings) erstellt. Im *Vectorian* kann entweder eines der beiden *embeddings* ausgewählt werden oder eine gewichtete Kombination aus beiden, bspw. 25% *fastText* und 75% *wnet2vec*. Bei der Suche wird auf dem Suchtext zunächst ein POS-Tagging durchgeführt. So können syntaktische Strukturen, die über die reine Wortreihenfolge hinausgehen, in die Suche einfließen.

Neben den beiden *embedding*-Modellen wurden zusätzlich weitere parametrisierbare Optionen umgesetzt, etwa die Berücksichtigung bzw. unterschiedliche Gewichtung von Wortarten, Einschüben sowie generell einer graduellen Anpassung des Ähnlichkeitswerts. Diese Parameter werden nachfolgend kurz erläutert.

1. "**Ignore Determiners**" entfernt alle Worte, die vom POS-Tagging als DT ("the", "this", etc.) erkannt wurden, aus der Suchanfrage.
2. "**Ensure POS Match**" ermöglicht das Ignorieren von Worten in den Korpusdokumenten, deren POS-Tags nicht dem der alignierten Worte im Suchtext entsprechen. Die Auswirkung der Einstellung kann graduell abgeschwächt werden.
3. "**POST STSS Weighting**": Nicht alle Wortarten besitzen gleiches semantisches Gewicht für die Bedeutung eines Satzes. Mittels "POST STSS Weighting" gewichten wir daher Wortähnlichkeiten bei der Suche mit einer an POST STSS („part-of-speech tag-supported short-text semantic similarity", Batanović, 2015) angelehnten Gewichtungsmatrix[3]. Die Auswirkung dieser Einstellung kann ebenfalls graduell abgeschwächt werden.
4. "**Mismatch Length Penalty**" konfiguriert, ab welcher Länge eines einzelnen *mismatch* im Ergebnis eine Abschwächung der Bewertung um 50% geschehen soll[4]. Eine Streuung von Matches ohne lokale Nähe führt in einem Ergebnis somit zur mehr oder weniger starken Abwertung. Die gesamte Abwertung für ein Ergebnis errechnet sich als Summe der Abwertungen für alle *mismatches*.
5. "**Similarity Threshold**" regelt den Schwellwert zur Ähnlichkeitsbewertung zwischen Wörtern. Ein niedriger Schwellwert erlaubt bspw. größere Abweichungen und kann dadurch auch zu einem größeren Rauschen durch mehr *false positives* führen.

Beispielabfragen

Der *Vectorian* wurde als parametrisierbare und interpretierbare Suchmaschine konzipiert, um einen explorativen Zugang zur Analyse unterschiedlicher Parameterkonfigurationen auf potenzielle Suchergebnisse, also in unserem Falle Shakespeare-Referenzen, zu ermöglichen. Nachfolgend illustrieren wir einige Auswirkungen unterschiedlicher Parame-

tereinstellungen am Beispiel der kurzen Shakespeare-Phrase "under the greenwood tree" (aus Shakespeares „As you like it").

Die am besten bewerteten Ergebnisse sind zunächst viele Varianten nach dem Schema „under the X tree", bspw. „under the *chestnut* tree". Mit dem Parameter *mismatch length penalty* kann man zusätzlich steuern, wie viele Einfügungen in den Treffern erlaubt sind. Werden Einfügungen nur in geringem Umfang erlaubt, dann erhält man vor allem Sätze bei denen die Präposition variiert wird, bspw. „**beneath** the *beech* tree". Erlaubt man hingegen mehr Einfügungen, kommt es entsprechend auch zu Ergebnissen wie "under the **dear old** *plane* tree".

Beim Parameter der *embeddings*-Wahl sieht man sehr gut, wie *FastText* und *WordNet* ganz unterschiedliche Präferenzen bei der Auswahl von alternativen „trees" liefern (*FastText*: „*chestnut*" > „*beech*" vs. *WordNet*: „*beech*" > „*oak*"). Das *mixed embedding* (also eine Aktivierung beider *embeddings* zu gleichen Teilen) scheint Vorteile beider *embeddings* optimal zu kombinieren, indem z.B. „oak tree" höher gewertet wird als „bodhi tree", wobei es sich bei Letzterem um einen spezifischen Baum aus einem religiösen Kontext handelt.

POST-STSS, ein Parameter der unterschiedliche POS unterschiedlich stark gewichtet, ist in Kombination mit dem *WordNet embedding* am aufschlussreichsten: Mit POST STSS werden im Zweifel reine Baumphrasen bevorzugt ("the fir tree", "the yew tree"). Ohne POST-STSS werden auch Substantive hoch bewertet, die mit Bäumen zwar nichts zu tun haben, dafür aber eine hohe semantische Nähe zu anderen Wörtern aufweisen, z.B. „greenwood" und „garden".

Fazit und Ausblick

Im aktuellen Stadium dient der *Vectorian* wie eingangs geschildert zunächst als Experimentierplattform, mit deren Hilfe man explorativ die Auswirkungen unterschiedlicher Einstellungsparameter erproben kann. Im nächsten Schritt soll eine systematische Evaluierung der Suchmaschine erfolgen, indem gegen eine vorab definierte *ground truth* an Shakespeare-Zitaten in einem Teilkorpus aus Fantasy-Romanen gesucht wird. Dabei werden alle möglichen Parameterkonfigurationen (insgesamt 72 Kombinationsmöglichkeiten) nacheinander durchgerechnet und die jeweiligen Bewertungen der einzelnen Sätze dokumentiert. Weiterhin soll berücksichtigt werden, wie viele *false positives* sich unter die *true positives* aus der *ground truth* mischen. Ziel ist es, diejenige Konfiguration zu identifizieren, die für möglichst viele Sätze der *ground truth* einen hohen *alignment score* aufweist und dabei die Zahl der *false positives* minimiert. Im nächsten Schritt sollen dann mit der bestbewerteten Konfiguration systematisch mehrere hundert Shakespeare-Zitate, die aus bestehenden Zitate-Datenbanken wie *WikiQuote* (https://en.wikiquote.org/) extrahiert werden, in einem großen Korpus von Fantasy-Literatur und Transkripten von Filmen und TV-Serien gesucht werden [5].

Fußnoten

1. Für eine Systematisierung von text reuse Methoden anhand der Kategorien inhaltliche, strukturelle und stilistische Ähnlichkeit vgl. Bär et al. 2012.
2. "The Vectorian" ist als Prototyp auf Anfrage verfügbar.
3. Beispiel: Eine Ähnlichkeit auf einem Adjektiv (Tag JJ) wird mit dem Faktor 0.7 gewichtet, während ein Verb (Tag VB) mit 1.2 gewichtet wird.
4. Die Abwertung über andere Längen erfolgt ausgehend vom gegebenen Basiswert exponentiell in der Länge des *mismatches*, was uns intuitiv und aufgrund der Beobachtungen in (Beeferman, 1997) sinnvoll erscheint. Der genaue Kurvenverlauf für gängige Längen wird im UI als Plot dargestellt.
5. Die Dokumentbasis des *Vectorian* kann flexibel erweitert werden solange die Texte in einem grundlegend bereinigten *plain text*-Format vorliegen.

Bibliographie

Bär, D. / Zesch, T. / Gurevych, I. (2012): Text Reuse Detection using a Composition of Text Similarity Measures. Proceedings of COLING 2012, 167-184.

Bär, D. / Zesch, T. / Gurevych, I. (2015): Composing Measures for Computing Text Similarity. Technical Report TUD-CS-2015-0017, TU Darmstadt.

Batanović, V. / Bojić, D. (2015): "Using Part-of-Speech Tags as Deep Syntax Indicators in Determining Short Text Semantic Similarity". In Computer Science and Information Systems, 12(1), S. 1–31.

Beeferman, D. / Berger, A. / Lafferty, J. (1997): A model of lexical attraction and repulsion. In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, S. 373-380.

Bubenhofer, N. (2019): Word Embeddings: Funktionale Äquivalenz statt Synonymie. Publiziert auf Sprechtakel-Blog (2.3.2019), online verfügbar unter https://www.bubenhofer.com/sprechtakel/2019/03/02/.word-embeddings-funktionale-aequivalenz-statt-synonymie/

Burghardt, M. / Meyer, S. / Schmidtbauer, S. / Molz, J. (to appear in 2019): "The Bard meets the Doctor" – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who. In Book of Abstracts, DHd 2019.

Garber, M. (2005): Shakespeare after All. New York: Anchor Books.

Genette, G. (1993): Palimpseste. Die Literatur auf zweiter Stufe. Frankfurt am Main: Suhrkamp. Translation of the revised second edition. [Genette, G. (1982). Palimpsestes. La littérature au second degré. Paris: Éditions de Seuil. Revised 2nd edition 1983.]

Kusner, M. / Sun, Y. / Kolkin, N. / Weinberger, K. (2015): "From Word Embeddings To Document Distances". In Proceedings of the 32nd International Conference on Machine Learning. Lille, Frankreich.

Manjavacas, E. / Long, B. / Kestemont, M. (2019): "On the Feasibility of Automated Detection of Allusive Text Reuse". ArXiv: 1905.02973 [Cs], 8. Mai 2019. http://arxiv.org/abs/1905.02973.

Maxwell, J. / Rumbold, K. (eds.) (2018): Shakespeare and Quotation. Cambridge: Cambridge University Press.

Mikolov, T. / Chen, K. / Corrado, G. / Dean, J. (2013): Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, Tomas, et al. "Advances in Pre-Training Distributed Word Representations." ArXiv:1712.09405 [Cs], Dec. 2017. arXiv.org, http://arxiv.org/abs/1712.09405.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. eBook online verfügbar unter https://christophm.github.io/interpretable-ml-book/

Saedi, C. / Branco, A. / Rodrigues, J. A. / Silva, J. (2018, July). Wordnet embeddings. In Proceedings of The Third Workshop on Representation Learning for NLP (pp. 122-131).

Sellers, Peter H. (1974): „On the Theory and Computation of Evolutionary Distances". *SIAM Journal on Applied Mathematics* 26, Nr. 4 (Juni 1974): 787–93. https://doi.org/10.1137/0126070.

Typisierte Varianz-Analyse von Texten

Balbach, Nico
nico.balbach@gmail.com
Zentrum für Philologie und Digitalität „Kallimachos",
Universität Würzburg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Zentrum für Philologie und Digitalität „Kallimachos",
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Zentrum für Philologie und Digitalität „Kallimachos",
Universität Würzburg, Deutschland; Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik, Universität Würzburg, Deutschland

Einleitung

Häufig gibt es unterschiedliche Quellen oder Auflagen zu einem Werk, deren Analyse Rückschlüsse auf die Entstehungsgeschichte oder auf unterschiedliche Akzentuierungen, aber z. B. auch auf Transkriptionsfehler zulässt. Dazu gehören nicht nur Differenzen in den Texten, sondern auch in der Typographie (z. B. kursive Hervorhebungen oder Fontgröße). Wir präsentieren das Open-Source Web-Tool „Variance-Viewer"[1] das, anders als die übliche Diff-Funktion in Texteditoren, nicht nur zwei Texte vergleichen und die Unterschiede markieren und hervorheben, sondern auch die Varianzen mit Regeln in Typen einteilen kann. Die verschiedenen Typen können ein- oder ausgeblendet sowie mit unterschiedlichen Farben markiert werden. Weiterhin können die zu vergleichenden Texte vor dem Vergleich normalisiert werden. Dadurch wird die Übersichtlichkeit bei vielen kleineren Unterschieden erheblich gesteigert, und es kann auf fachlich relevante Differenzen fokussiert werden. Es ist ein TEI-Export verfügbar, in dem für die Varianzen vordefinierte Tags generiert werden. Folgender Workflow soll beim Vergleich zweier Werke unterstützt werden:

1. Übersicht über Differenzen bekommen (dafür eignet sich praktisch jedes Diff-Tool).

2. Wiederhole:

a. Definition von Typen der Differenzen mittels Konfigurationsdatei.

b. Ein- und Ausblenden der Typen und Untersuchung der Restkategorie, ob weitere Typ-Definitionen sinnvoll sind.

3. Weitere editorische Arbeiten, ggf. TEI-Export der typisierten Differenzen.

Verwandte Arbeiten

Die meisten Texteditoren verfügen über eine Diff-Funktion, mit der sich zwei Texte (auch Programmcode oder DNA-Sequenzen) vergleichen und insbesondere Änderungshistorien von Dokumenten nachverfolgen lassen (vgl. z. B. die Darstellungen von Varianten in der Faust-Edition[2]). Dabei gibt es häufig zwei Darstellungen: zum einen die der Änderungen innerhalb eines Dokumentes und zum anderen die Gegenüberstellung der beiden Dokumente mit jeweiliger Hervorhebung der Änderungen. Viele Algorithmen basieren auf der Publikation von Myers (Myers 1986), der gezeigt hat, dass die Suche nach der längsten gemeinsamen Teilfolge und der kürzesten Transformation eines Strings A in einen String B als äquivalent angesehen werden können. Eine Implementierung ist die Suche nach einem kürzesten Weg in einem Edit-Graphen bzw. einer Matrix, der aus den Wörtern oder Buchstaben der beiden Dokumente als Zeilen bzw. Spalten besteht. Für literarische Texte ist im Allgemeinen eine feinere Differenzierung wünschenswert, in der Typen von Änderungen erkannt und ein- oder ausgeblendet werden können. Diese können sich sowohl auf den Text als auch die Typographie beziehen. Da die Typen von den individuellen Interessen der jeweiligen Philologen abhängen, sollten sie nicht fest vorgegeben, sondern leicht anpassbar sein. Weiterhin ist neben einer Visualisierung auch ein Export nach TEI wünschenswert. Im Folgenden präsentieren wir ein solches Tool, da wir kein vergleichbares, einfach bedienbares Werkzeug kennen (so wird z. B. in der Übersicht über Digital-Humanities-Tools und Services in (Bulatovic 2016) diese Kategorie nicht erwähnt). Ein ähnliches, aber anspruchsvolleres Tool ist CollateX[3] (Haentjens Dekker 2014), das in der Lage ist, zwei und mehr Texte zu kollationieren und das Ergebnis als Graph zu visualisieren. Dabei können auch Transpositionen, d. h. verschobene Texte gefunden werden, teilweise einschließlich Erkennung von Varianten der verschobenen Texte. Ein weiteres anspruchsvolles Tool ist Stemmaweb[4], dessen GitHub-Repository[5] jedoch darauf hindeutet, dass es nicht oder kaum noch aktiv gepflegt wird. Im Beitrag (Andrews 2014), bei dem es um eine kritische Bewertung der Entstehungsgeschichte von drei Werken geht, wird u. a. kritisiert, dass bestimmte Typen von Änderungen vorschnell als „insignifikant" bewertet werden. Entwurfsregeln zur Visualisierung von Text-Varianz-Graphen werden in (Jänicke 2014) dargestellt. Der Variance-Viewer hat einen anderen Schwerpunkt: er gibt die Typen von Änderungen nicht vor, sondern überlässt deren Definition dem Anwender durch einfache Konfiguration. Die Darstellung enthält keine Graphen, sondern eine farbige Kennzeichnung und bietet das Aus- und Einblenden von Typen von Varianten durch einfachen Klick an, so dass Editoren die Übersicht behalten, wenn Sie sich auf bestimmte Differenztypen konzentrieren wollen.

Methoden

Der Variance-Viewer verwendet für die Berechnung von Differenzen zwischen zwei Texten eine Implementierung[6] des Algorithmus von Myers und fügt dann Nachbearbeitungen zur Differenzierung verschiedener Typen von Änderungen hinzu. Die Kategorien sind frei konfigurierbar (s. Abbildung 1 links unten für einen Auszug aus der Konfigurationsdatei). Die Nachbearbeitung prüft für jede gefundene Änderung, ob die Bedingungen für einen der definierten Typen vorliegen und ordnet sie dann dem entsprechenden Typ zu. Die Änderungen werden auf Wortebene berechnet und zusätzlich die für die Änderung verantwortlichen Buchstaben identifiziert, so dass beides hervorgehoben werden kann, wobei zusätzliche Leerzeichen auch wortübergreifend gefunden werden. Alle nicht zugeordneten Typen werden einem Default-Typ (z. B. „Inhalt“ bzw. „Content“) zugeordnet, wobei noch zwischen einfachen und komplexen Änderungen unterschieden werden kann (einfache Änderungen unterscheiden sich nur in einem Buchstaben). Die Ergebnisse können in TEI ausgegeben werden, indem das „app“-Tag mit speziellen Attributen für die Änderungstypen benutzt wird. Weiterhin können sie visuell präsentiert werden, wobei den Typen verschiedene Farben zugeordnet werden und bei Bedarf jeder Typ auch ausgeblendet werden kann, um die Übersicht zu verbessern. Das Programm präsentiert beide Texte in einer synoptischen Darstellung, wobei zur Gewährleistung einer zeilenäquivalenten Darstellung in einem Dokument freier Platz auf Abschnittsebene hinzufügt wird, falls das notwendig ist.

Den Umgang mit den Typen erläutern wir an zwei philologischen Anwendungsprojekten, in denen der Variance-Viewer eingesetzt wurde: Die Analyse der Änderungen in den Schriften von Richard Wagner im Projekt RWS[7] und die Analyse der verschiedenen Auflagen von Drucken im Narragonien digital Projekt[8].

Im RWS-Projekt liegen die Texte als TEI-Dokumente vor. Bei der Analyse der Varianzen sind nicht nur textuelle Änderungen interessant, sondern auch Änderungen bzgl. der Formatierung, die in TEI im Element „rend“ hinterlegt sind. Daher wird dieses genauer analysiert. Insgesamt sind folgende Typen von Änderungen durch projektspezifische Regeln definiert (vgl. Abbildung 1):

- Satzzeichen (Punctuation): Die Änderung bezieht sich nur auf ein Satzzeichen (. , ; - ? ! usw.).
- Grapheme (Graphemics): Die Änderung bezieht sich nur auf bestimmte Schreibweisen (y i; u v; s ſ; ss ß; Groß/Kleinschreibung; th t; usw.).
- Abkürzungen (Abbreviation): Die Änderung bezieht sich nur auf Abkürzungen (z. B. Dr. Doktor; Hr. Herr Herrn; usw.).
- Typographie (Typography): Die Änderung ist keine inhaltliche, sondern bezieht sich auf das Layout oder die Typographie und wird in dem TEI-Attribut „rend“ mit entsprechenden Werten spezifiziert (kursiv; gesperrt; usw.).
- Inhalt (Content): Alle übrigen Änderungen, die keiner der obigen Kategorien zugeordnet werden können einschließlich Hinzufügen oder Löschen sowie Änderungen, bei denen mehr als eine Änderung der obigen Typen gleichzeitig vorkommt.

Im Narragonien-Projekt liegen die Drucktexte als Plain Text Dateien vor. Hier werden folgende Typen von Änderungen unterschieden (vgl. Abbildung 2):

- Grapheme (mit anderer Liste von Buchstabenersetzungen wie im RWS-Projekt).
- Abkürzungen (mit anderer Bedeutung als im RWS-Projekt; hier sind es meist einzelne Buchstaben mit Unter- oder Überstrichen, die expandiert werden).
- Leerzeichen im Wort, die ein Wort in zwei oder mehrere Wörter auftrennen. (Separation). Diese Option ist technisch aufwändiger, weil nicht einzelne Wörter sondern Wortgruppen miteinander verglichen werden müssen.
- Inhaltsänderungen mit nur einem Zeichen Unterschied (OneDifference), die nicht in der Graphem-Liste enthalten sind und anders bewertet werden als komplexere Änderungen.
- Inhalt (Content): Alle übrigen Änderungen.

Erfahrungen

Das Tool wurde in beiden Projekten erfolgreich eingesetzt, und dabei auch für die Verarbeitung sehr langer Dokumente genutzt. Im Folgenden zeigen wir zwei Screenshots aus dem RWS- und dem Narragonien-Projekt. Dabei ist besonders hervorzuheben, dass der Rest-Typ „Content“, der alle sonst nicht speziell erkannten Typen von Änderungen beinhaltet, nur noch ca. die Hälfte der Änderungen ausmacht, während die andere Hälfte spezielleren Typen zugeordnet werden konnte. Wenn das Ziel die Feinanalyse bestimmter Änderungstypen ist, können auch iterativ weitere Typen definiert und der Analysealgorithmus damit erneut ausgeführt werden.

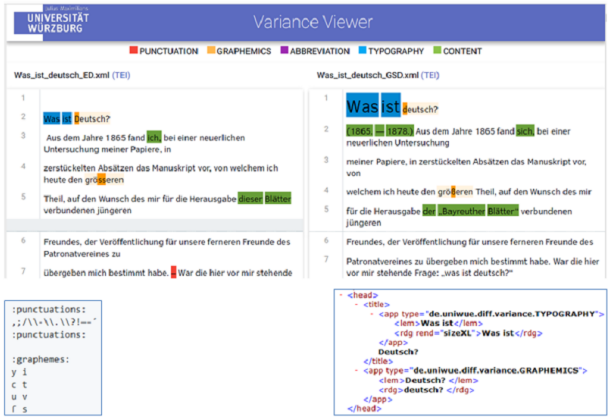


Abbildung 1: Vergleich zweier Texte aus dem Schriften-Verzeichnis von Richard Wagner mit Hervorhebung der Änderungstypen in verschiedenen Farben (Erläuterungen der Typen im Text; Auszug aus Konfigurationsdatei links unten). Die Texte liegen im Format TEI vor, wobei TEI-Attributwerte auf CSS abgebildet wurden, um die Darstellung unterschiedlicher Typen sichtbar zu machen, und die gefundenen Differenzen mit ihren Typen können auch als TEI exportiert werden (Auszug für die erste Zeile s. rechts unten).

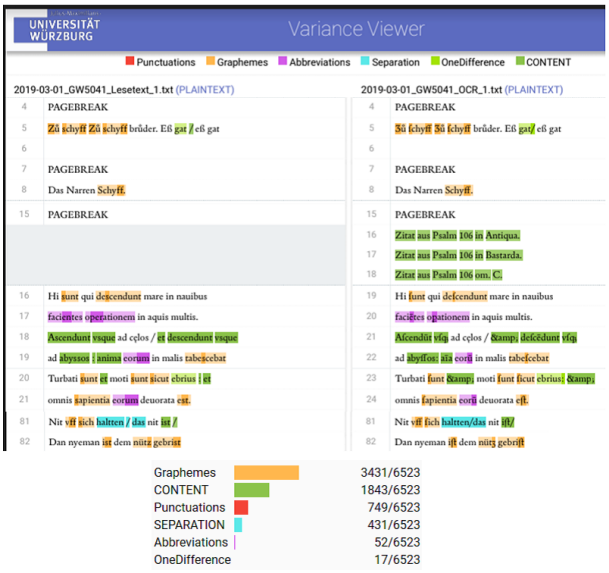


Abbildung 2: Vergleich des edierten Lesetextes der Narrenschiff-Ausgabe GW5041 (links) mit dem Ergebnis der OCR auf dem Originaltext einer anderen Druckausgabe (rechts), wobei die Änderungen sowohl OCR-Fehler als auch Normalisierungen der Schrift im Lesetext umfassen. Dies ist durch Hervorhebung der Änderungstypen in verschiedenen Farben leichter nachvollziehbar (Erläuterung der Typen im Text). Unten eine Statistik, die die 6.523 gefundenen Änderungen nach Änderungstypen aufschlüsselt. Der Gesamttext umfasste 150 Seiten mit 4.200 Zeilen, 26.000 Wörtern und 121.000 Zeichen und die zugehörige Konfigurationsdatei („Settings") ca. 100 Zeilen. Für diese Analyse brauchte der Variance-Viewer in dem serverseitig ausgeführten Demo-Modus im Web ca. 25 Sekunden (für intensive Nutzung sollte der Open-Source Code lokal installiert werden).

Die bisherigen Erfahrungen zeigen, dass noch eine Reihe von relativ einfachen Erweiterungen wünschenswert sind, wobei zu erwarten ist, dass in weiteren Projekten weitere Aspekte hinzukommen:

- Gelegentlich enthält ein Wort mehrere Änderungen (z. B. mehrere Grapheme und/oder Satzzeichen). Wenn es mehrere Änderungen desselben Typs gibt, werden diese dem Typ zugeordnet (z. B. „Schiff" in Zeile 5 in Abbildung 2 mit zwei graphemischen Differenzen). Wenn es jedoch Änderungen unterschiedlicher Klassen sind, werden diese als ein nicht näher differenzierter Unterschied („Content") betrachtet (z. B. „ist /" und „iſt/" in Zeile 81 mit zwei verschiedenen Typen von Änderungen bezüglich Leerzeichen und Graphem). Hier wäre eine Mischklasse aus den jeweiligen Ursprungsklassen wünschenswert. Das Problem lässt sich teilweise durch vorherige Normalisierung lösen, indem z. B. alle Graphem-Änderungen vorab normalisiert werden und sich dann manche komplexe Fehler zu einfachen Fehlertypen reduzieren.
- Es gibt Ausnahmen zu den Regeln, in denen die Nutzer die Änderungsklasse manuell ändern und ggf. kommentieren können sollten. Bisher zeigt der Viewer nur das automatisch generierte Ergebnis der Regelauswertung an, er sollte um eine Editierfunktion erweitert werden. Dies ist z. B. hilfreich, wenn bei automatischer OCR von verschiedenen Ausgaben eines Werkes zwischen OCR-Fehlern und Textvarianten unterschieden werden soll.

Zusammenfassung und Ausblick

Der vorgestellte Variance-Viewer ermöglicht die Feindifferenzierung und Klassifikation von Textvarianten mittels selbstdefinierter Typen. In verschiedenen Ausgaben von literarischen Texten treten oft zahlreiche „technische" Varianzen auf, die sich auf Satzzeichen, Leerzeichen, Buchstabenvarianten und ggf. auch auf das Layout oder die Typographie beziehen, die von eigentlichen inhaltlichen Änderungen zu trennen sind. Hier werden bei Verwendung eines einfachen Diff-Werkzeugs häufig so viele Änderungen angezeigt, dass der Überblick verloren geht. Ein Filtern bzw. Hervorheben bestimmter Typen von Varianzen erleichtert die philologische Arbeit beträchtlich. Wichtig ist, dass die Typen von Varianzen abhängig von den Fragestellungen und individuellen Interessen des jeweiligen Philologen leicht konfiguriert werden können. Der vorgestellte Varianz-Viewer erfüllt diese Anforderungen und hat sich in zwei größeren philologischen Projekten bewährt. Er ist Open-Source, webbasiert, leicht zu installieren und zu bedienen. Perspektiven der Weiterentwicklung umfassen eine einfachere oder sogar automatische Definition der Varianztypen sowie funktionelle Erweiterungen:

- Aus technischer Sicht sollte für die Definition von Varianztypen ein Editor bereitgestellt werden, so dass deren Definition im Vergleich zur bisherigen Konfigurationsdatei noch weiter vereinfacht wird. Dazu kann eine Regelsprache bereitgestellt werden oder ein Lernverfahren, dem einige Beispiele präsentiert werden und der das Muster dann selbständig erkennt.
- Die häufigsten Typen von Varianzen können auch durch Lernverfahren vollautomatisch erkannt werden (ohne vorgegebene Varianztypen), indem alle vom Diff-Algorithmus gefundenen Varianten auf gemeinsame Muster hin analysiert werden.
- Eine umfassende Änderung wäre die Weiterentwicklung des relativ einfachen Tools zur Erkennung komplexerer Änderungen wie Transpositionen und zur Visualisierung der Änderungen, auch von mehreren Werken, in Graphen, ggf. durch Übernahme entsprechender Funktionalitäten z. B. aus CollateX oder Stemmaweb.

Fußnoten

1. Web-Link vom Variance-Viewer: http://variance-viewer.informatik.uni-wuerzburg.de; Code Open-Source unter: https://github.com/cs6-uniwue/Variance-Viewer
2. http://faustedition.net
3. https://collatex.net
4. https://stemmaweb.net
5. https://github.com/tla/stemmatology
6. Java-diff-utils: https://code.google.com/archive/p/java-diff-utils. Die Software wird von Google gehostet und wurde von 2009-2013 von verschiedenen Autoren entwickelt (s. https://code.google.com/archive/p/java-diff-utils/source/default/commits).
7. Richard Wagner Schriften (RWS): Historisch-Kritische Gesamtausgabe: http://www.musikwissenschaft.uni-wuerzburg.de/forschung/richard-wagner-schriften.
8. „Narragonien digital": http://kallimachos.de/kallimachos/index.php/Narragonien.

Bibliographie

Andrews, Tara L (2014): Analysis of variation significance inartificial traditions using Stemmaweb, in *Digital Scholarship in the Humanities, 31(3)*.

Bulatovic, Natasa / Gnadt, Timo / Romanello, Matteo / Schmitt, Viola / Stiller, Juliane / Thoden, Klaus (2016): Usability von DHTools und Services, in *DARIAH Working Papers*: https://wiki.de.dariah.eu/download/attachments/14651583/AP1.2.3_Usability_von_DH-Tools_und-Services_final.pdf.

Haentjens Dekker, Ronald / van Hulle, Dirk / Midell, Gregor / Neyt, Vincent / van Zundert, Joris (2014): Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project, in *Digital Scholarship in the Humanities, 30(3), 452-470*.

Jänicke, Stefan / Geßner, Annette / Büchler, Marco / Scheuermann, Gerik (2014): 5 Design Rules for Visualizing Text Variant Graphs, in *DH*: https://www.informatik.uni-leipzig.de/~stjaenicke/5_Design_Rules_for_Visualizing_Text_Variant_Graphs.pdf.

Myers, Gene (1986): An O(ND) difference algorithm and its variations, in *Algorithmica, 1, 251-266*.

Unsichtbares sichtbar machen - semantische Modellierung interpretativer Vorgänge am Beispiel der historischen Bestandsaufnahme der Brandenburgisch-Preußischen Kunstkammern

Wagner, Sarah
sarah.wagner.kuhi@gmail.com
Humboldt-Universität zu Berlin, Deutschland

Einleitung

Bei der Analyse historischer Quellen sind Wissenschaftler häufig mit Unschärfen, Lücken oder auch Mehrdeutigkeiten konfrontiert. Aufgrund ihrer Historizität geht die Bearbeitung stets mit interpretativen Vorgängen einher, da die Informationen nicht objektiv gegeben, sondern subjektiv und von individuellen Erfahrungen und kulturellen Mustern geprägt sind (vgl. Büttner 2014:17). Wie können bei der Entwicklung von Datenmodellen, die auf Forschungsgegenstand und -fragestellung zugeschnitten sein sollen, historische Unschärfen und deren Interpretationsvorgänge abgebildet werden, damit die Daten semantisch erschlossen und formalisiert werden können? Wie können Auslegungen transparent und nachvollziehbar dokumentiert werden? Und wie können Spielräume geschaffen werden, um dem stetigen Neuverhandeln von Rückschlüssen sowie der Abbildung von Mehrdeutigkeiten gerecht zu werden?

Am Beispiel der historischen Bestandsaufnahme der Brandenburgisch-Preußischen Kunstkammern werden im Folgenden anhand der Beschaffenheiten des Forschungsgegenstands sowie der Zielsetzung und Grundlagen des Projekts Anforderungen an ein Datenmodell formuliert und erste Ansätze skizziert.

Ziel, Forschungsgegenstand und Grundlagen des Projekts

Im Kooperationsprojekt „Das Fenster zur Natur und Kunst – Eine historisch-kritische Aufarbeitung der Brandenburgisch-Preußischen Kunstkammer"[1] widmen sich die Staatlichen Museen zu Berlin, die Humboldt-Universität zu Berlin und das Museum für Naturkunde Berlin der Geschichte und Analyse der königlichen Kunstkammerbestände (vgl. zuletzt Dolezel 2019), die einst im Berliner Stadtschloss beherbergt waren.

Diese Bestände bilden heute den Grundstock zahlreicher Berliner Museen und umfassten einst Objekte der Kunst, der Natur und der Wissenschaft, die in ein komplexes Beziehungs- und Bedeutungsgeflecht eingebettet waren. Ihre Konstellation und räumliche Ausdehnung war von einer steten Dynamik geprägt, bedingt durch Zu- und Abgänge einzelner Objekte oder ganzer Teilbestände. Und so wie neue Sammlungskomplexe aus den Berliner Kunstkammerbeständen hervorgingen, so wurden auch sie selbst aus früheren Sammlungen geformt. Dabei lag jeder Sammlung ein individuelles Ordnungssystem zugrunde, das beispielsweise den Wissensstand oder Geschmack der Zeit reflektieren konnte, zugleich in Abhängigkeit mit der Funktion der Sammlung und der Intention ihres Eigentümers stand. Aufgrund dessen ist auch nicht von einer Brandenburgisch-Preußischen Kunstkammer auszugehen, sondern von verschiedenen Sammlungen, die sich zwischen dem 16. und der zweiten Hälfte des 19. Jahrhunderts stetig neu formierten, ihren Namen aber beibehielten. Es kann daher nicht der Anspruch sein, eine Rekonstruktion einer bestimmten Sammlungssituation zu einem bestimmten Zeitpunkt vorzunehmen. Ziel ist es vielmehr, anhand noch erhaltener, aber auch verschollen gegangener Objekte zu analysieren, inwiefern diese als materielle Träger von Bedeutung zu unterschiedlichen Zeitpunkten betrachtet und bewertet wurden, und wie sich der Zugriff auf sie gestaltete. Denn die einzige Konstante im Kreislauf von Formierung und Auflösung der Bestandskomplexe ist das Objekt in seiner physischen Gestalt. In welche taxonomische, narrative, räumliche, inszenatorische oder nutzungsbezogene Zusammenhänge wurden sie jeweils gestellt? Und inwiefern lassen sich daraus historische Sammlungspraktiken und -logiken erschließen?

Für das Vorhaben wurde eine Auswahl an Objekten zur Tiefenerschließung getroffen, um die sich aufgrund der Multi-

dimensionalität des Beziehungsgeflechts zwangsläufig Gruppierungen von weiteren Objekten bilden. Daneben stellen schriftliche historische Quellen eine weitere wesentliche Grundlage des Vorhabens dar. Diese umfassen Sammlungsinventare des 17. und 18. Jahrhunderts, Reisebeschreibungen und Stadtführer des 18. und frühen 19. Jahrhunderts, gedruckte Sammlungsführer und -geschichten des 19. Jahrhunderts sowie Museumsführer oder auch Museumsakten.

Die Rekonstruktion der Objektgeschichten erfolgt zum einen vom heutigen Standpunkt ausgehend, indem die Provenienz der Objekte zurückverfolgt wird. Zum anderen werden, ausgehend vom heterogenen historischen Quellmaterial, Geschichte, Bewertung und Zugriffspraktiken zu den einzelnen Objekten nachgezeichnet.

Anforderungen an das Datenmodell und erste Ansätze

Wie also können unsichtbare Eigenschaften von Objekten über einen Zeitraum von mehreren Jahrhunderten zurückverfolgt und vor allem sichtbar gemacht werden? Welche Eigenschaften sind das und welche Akteure und Kontexte spielen bei ihrer Zuweisung eine Rolle? Wie können die Pluralität, Gegensätzlichkeit und Ungewissheit der Informationen erfasst werden?

Im Zuge der Entwicklung eines auf die Bedürfnisse des Projekts zugeschnittenen Datenmodells, das anschließend in die virtuelle Forschungsumgebung WissKI[2] implementiert werden soll, muss zunächst der aktuell vorliegende digitale Datenbestand betrachtet werden. Die fokussierten Objekte sind heute auf u.a. die drei Projektinstitutionen aufgeteilt und damit verschiedenen Fachbereichen zugeordnet und hinsichtlich fachspezifischer Aspekte erschlossen. Aus diesem Grund muss eine Grundlage für eine disziplinübergreifende, einheitliche und strukturierte Erfassung der Objekte geschaffen werden. Während bei den im naturwissenschaftlichen Kontext befindlichen Objekten beispielsweise taxonomische Klassifizierungen relevant sind, so stehen bei jenen, die heute in Kunstsammlungen beherbergt sind, ikonographische oder stilistische Kriterien im Vordergrund. Auch diese heutigen, fachspezifischen Erschließungskriterien bilden einen weiteren Baustein in der Bewertungs- und Zugriffsgeschichte der Objekte und sind deshalb in das Vorhaben mit einzubeziehen. Neben weiteren deskriptiven und strukturellen Metadaten müssen alle vorliegenden, für die Provenienz relevanten Angaben überführt werden können. WissKI bietet dafür verschiedene Schnittstellen.[3]

Ein weiteres zentrales Konzept im Datenmodell bilden neben den Objekten die historischen Quellen, deren Entstehungs- und Funktionskontext, Informationsgehalt und Auslegung es abzubilden gilt. Dabei können bereits Art und Entstehungskontext einer Quelle in direktem Zusammenhang mit der enthaltenen Art der Information und unter Umständen auch der Art und Weise, wie diese interpretiert wird, stehen. Die jeweilige Auslegung der Information des Quelltextes muss aufgrund ihres subjektiven Charakters so modelliert werden, dass die Kriterien, die zu ihr geführt haben, möglichst transparent und nachvollziehbar dokumentiert werden können. Auch dürfen deshalb die aus dem Text extrahierten Information nicht als gegebener Fakt modelliert, sondern entsprechend des interpretativen Vorgangs hypothetisch abgebildet werden. Aus diesen Gründen sollte die Auslegung als Aktivität begriffen werden, an die eine Kombination aus Pfaden zur Verschlagwortung bzw. zur Hinterlegung von Begriffsthesauri, zur Verbindung mit Sammlungskomplexen für Standort- oder Zugehörigkeitszuweisung, zur Kommentierung für den Wissenschaftler sowie zur Transkription der betreffenden Textstelle selbst verbunden werden können.

Des Weiteren sollten die Auslegungen zeitlich verortet, mit anderen Ereignissen und auch mit Akteuren in Verbindung gebracht werden können, sofern die Aussagen nicht vom Autor der Quelle selbst, sondern von Dritten getätigt wurden. Dadurch würde gewährleistet, dass der Kontext einer Zuweisung und ihre Auslegung möglichst detailliert dokumentiert werden können, und der Wissenschaftler in den damit verbundenen Kommentarfeldern seine Auslegung begründen oder kritisch reflektieren kann. Dies gilt es jedoch zunächst zu erproben.

Durch seinen ereigniszentrierten Charakter und die Intention, einen fachübergreifenden Austausch zu ermöglichen, bietet das vom ICOM CIDOC entwickelte ISO-zertifizierte Conceptual Reference Model (CRM)[4] ideale Voraussetzungen, um als Ontologie für das zu entwickelnde Datenmodell verwendet zu werden.[5] Durch seine ISO-Zertifizierung ist die Langzeitinterpretierbarkeit der Daten garantiert, zudem wird das Modell regelmäßig von den Entwicklern aktualisiert. Das CRM stellt darüber hinaus die einzige Ontologie dar, die sich über die letzten Jahre als Erfassungsschema für den Bereich des kulturellen Erbes durchgesetzt hat. Um der Spezifizierung des zu beschreibenden Themengebiets innerhalb des kulturellen Gegenstands gerecht zu werden, kann die Ontologie angepasst bzw. erweitert werden (vgl. Hohmann / Fichtner 2015: 120). Indem das CRM ermöglicht, physische und abstrakte Konzepte über Ereignisse miteinander in Beziehung zu setzen, können insbesondere die geschilderten, notwendigen Zuweisungen von Merkmalen, Standorten usw. sowie ihre Veränderungen abgebildet und mit weiteren Akteuren, Zeit, Ort und Ereignissen in Beziehung gesetzt werden.

Für die Modellierung interpretativer Vorgänge bietet sich beispielsweise die Klasse E13 Attribute Assignment (Merkmalszuweisung) an, die es in der projektspezifischen Anwendungsontologie auszudifferenzieren gilt. Ihr Anwendungsbereich wird wie folgt beschrieben:

> „Diese Klasse umfasst die Aktionen des Feststellens von Eigenschaften eines Gegenstandes oder von Beziehungen zwischen zwei Gegenständen oder begrifflichen Konzepten. [Sie] erlaubt die Dokumentation, wie die jeweilige Feststellung zu Stande kam, und wessen Meinung es war. Alle in solch einer Aktion zugewiesenen Merkmale oder Eigenschaften können auch so verstanden werden, als ob sie direkt am jeweiligen Gegenstand oder begrifflichen Konzept fest gemacht wurden, möglicherweise auch als eine Sammlung von widersprüchlichen Werten. [...]". (Doerr / Lampe / Krause 2011: 56)

Ein Pfad für die Auslegung einer Quellinformation, konkret für die Zuweisung eines Bedeutungsbegriffs, könnte damit folgendermaßen modelliert werden[6]:

E84 Information Carrier (Quelle) → P128i carries → E73 Information Objekt (Inhalt der Quelle) → P140i was attributed by → E13 Attribute Assignment (Objektattributierung) → P17i includes → E13 Attribute Assignment (Bedeutungszuweisung) → P141 assigned → E55 Type (Begriffsthesaurus) → P149 is identified by → E75 Conceptual Object Appellation (Bezeichnung des Begriffs)

An die Klasse *E13 Attribute Assignment (Objektattributierung)* könnten über die Property *P17i includes* weitere Zuweisungen als Teil einer Objektattributierung gekettet werden, beispielsweise jene, die den Standort (Standortzuweisung) betreffen oder auf eine Sammlungszugehörigkeit (Sammlungszuweisung) hinweisen. An diese wiederum könnten Ereignisse, Akteure, Datum oder Kommentare gekoppelt werden. Auf Ebene des *E13 Attribute Assignment (Objektattributierung)* könnte das betreffende Objekt (*E84 Information Carrier*) gebunden werden, dem laut Interpretation der Quelle (*E84 Information Carrier)* Eigenschaften zugewiesen werden:

E84 Information Carrier (Quelle) → P128i carries → E73 Information Objekt (Inhalt der Quelle) → P140i was attributed by → E13 Attribute Assignment (Objektattributierung) → P140 assigned attribute to → E84 Information Carrier (Objekt) → P48 has preferred Identifier → E42 Identifier (Inventarnummer)

Die Merkmalszuweisung erlaubt aufgrund ihres Anwendungsbereichs die Darstellung interpretativer Vorgänge, an die, aufgrund ihres Aktivitätscharakters, Kontexte, Umstände und Begriffe optimal angeknüpft werden, und die durchaus plural oder mehrdeutig sein können. Zusätzlich kann die Merkmalszuweisung mit Kommentaren verbunden werden, um eine zusätzliche, ausführlichere Erläuterung des Vorgangs zu dokumentieren.

Fazit

Es muss bereits zu Beginn bewusst sein, dass allein durch die Gestaltung des Datenmodells formuliert wird, welche Informationen am Ende zu sehen sein sollen. Denn bereits die Modellierung stellt einen interpretativen Akt dar, der stets unter dem Einfluss des Wissens eines Subjekts oder einer Gruppe steht. Aus diesem Grund muss in einem stetigen Prozess immer wieder erprobt und kritisch hinterfragt werden, inwiefern besondere Aspekte zu stark akzentuiert werden oder zu wenig zur Geltung kommen. Unter der Annahme einer ständigen Überarbeitung ist es deshalb von besonderem Belang, bei der Modellierung von Daten Sackgassen zu vermeiden und stets mögliche Punkte zum Andocken bereitzustellen.

Fußnoten

1. Das Projekt wird seit 2018 von der Deutschen Forschungsgemeinschaft gefördert. Informationen zum Projekt auf der Projektseite der Staatlichen Museen zu Berlin, URL: https://www.smb.museum/forschung/forschungsprojekte/fenster-natur-kunst.html [letzter Zugriff: 02.01.2020].
2. Website WissKI, URL: http://wiss-ki.eu [letzter Zugriff: 02.01.2020].
3. Eine gängige Möglichkeit des Datenimports besteht darin, Daten aus der Quelldatenbank als CSV-Datei zu exportieren und anschließend in das stets zusätzlich zum Triplestore mit einem WissKI-System gekoppelte, relationale Datenbankmanagementsystem MariaDB (Website URL: https://mariadb.org/ [letzter Zugriff: 02.01.2020]) zu importieren. Von dort aus werden die Daten sodann über ein Import-Skript, das die Tabellendaten auf die semantischen Pfade des im System implementierten Datenmodells mappt, mit dem WissKI-ODBC-Import-Modul in die Forschungsumgebung importiert und im Zuge dessen als Tripledaten im Triplestore abgespeichert.
4. Website des CIDOC CRM, URL: http://cidoc-crm.org/ [letzter Zugriff: 02.01.2020].
5. Die einzige Implementierung in OWL stellt das sog. Erlangen CRM dar, das für die konkrete Modellierung verwendet wird. Website des Erlangen CRM, URL: http://erlangen-crm.org/ [letzter Zugriff: 02.01.2020].
6. Die in Klammern gesetzten Begriffe sollen die Subkonzepte in der projektspezifischen Anwendungsontologie darstellen.

Bibliographie

Büttner, Nils (2014): *Einführung in die frühneuzeitliche Ikonographie.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Doerr, Martin / Lampe, Karl-Heinz / Krause, Siegfried (2011): *Definition des CIDOC Conceptual Reference Model Version 5.0.1*.; autor. durch die CIDOC CRM Special Interest Group (SIG) (= Beiträge zur Museologie 1). Berlin: ICOM Deutschland.

Dolezel, Eva (2019): *Der Traum vom Museum.* Die Kunstkammer im Berliner Schloss um 1800 – eine museumsgeschichtliche Verortung. Berlin: Gebr. Mann.

Hohmann, Georg / Fichtner, Mark (2015): „Chancen und Herausforderungen in der praktischen Anwendung von Ontologien für das Kulturerbe" in: Robertson – von Trotta, Caroline Y. / Schneider, Ralf Y. (eds.): *Digitales Kulturerbe*. *Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis* (= Kulturelle Überlieferung – digital 2). Karlsruhe: KIT Scientific Publishing 115-128.

Varianz, Ambiguität, Unsicherheit. Methodische Schlaglichter zur mittelniederdeutschen Grammatikographie

Ihden, Sarah
sarah.ihden@uni-hamburg.de
Universität Hamburg, Deutschland

Eine besondere Herausforderung in der Grammatikographie historischer Sprachstufen des Deutschen stellt der Umgang mit Varianz, Ambiguitäten und Unsicherheiten dar. Hinzu kommt die Gefahr, dass durch die den Analysen für die Grammatikschreibung zugrunde gelegten Daten, insbesondere grammatische Annotationen in Korpora, die Ergebnisse dieser Analysen gewissermaßen vorgeprägt sind. Diese Besonderheiten sind auch bei der geplanten Bearbeitung der Flexionsmorphologie als Teil einer neuen wissenschaftlichen mittelniederdeutschen Grammatik zu berücksichtigen. Im Vortrag sollen die Methoden und Grundsätze dieser neuen Grammatik vorgestellt werden, wobei ein Fokus auf der Variationssensitivität und dem Korpusbezug liegt. Zudem soll beschrieben und anhand erster Analysen veranschaulicht

werden, wie in der Erforschung der mittelniederdeutschen Flexionsmorphologie dem potentiellen Risiko einer zirkulären Darstellung begegnet wird und wie auf der Basis von Daten, die möglichst oberflächenbezogen annotiert und in denen Ambiguitäten ausgezeichnet sind, flexionsmorphologische Variation ermittelt und vor dem Hintergrund potentieller außer- und innersprachlicher Parameter beschrieben werden kann.

Eine umfassende wissenschaftliche Grammatik des Mittelniederdeutschen, die modernen Ansprüchen genügt, stellt ein dringendes Forschungsdesiderat dar. Die gegenwärtig sowohl für die Forschung als auch die akademische Lehre herangezogenen Grammatiken von Colliander (1912), Lasch (1914, / ^{2}1974 / Nachdruck 2011), Lübben (1882) und Sarauw (1921–1924) sind methodisch veraltet. Ihrer Entstehungszeit entsprechend folgen sie einem junggrammatischen Paradigma und liefern Darstellungen der mittelniederdeutschen Grammatik mit einem deutlichen Fokus auf der Laut- und Formenlehre. Andere Sprachebenen wie Satz und Text bleiben weitestgehend unberücksichtigt. Auch die unflektierbaren Wortarten werden, wenn überhaupt, nur am Rand betrachtet wie bei Lübben (1882: 120–132) und Sarauw (1924: 229–234).

In einer neuen wissenschaftlichen Grammatik des Mittelniederdeutschen sollen diese Lücken geschlossen und dabei moderne Methoden der Grammatikschreibung herangezogen werden. Eine wesentliche methodische Anforderung liegt im Korpusbezug. Eine umfassende Beschreibung der tatsächlichen grammatischen Gegebenheiten im Mittelniederdeutschen benötigt ein umfangreiches empirisches Fundament. Hierfür werden die Daten des strukturierten und balancierten Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200–1650) (kurz: ReN) genutzt, das seit 2013 mit Unterstützung der DFG an den Universitäten Hamburg und Münster entstanden ist und dessen finale Korpusversion im September 2019 online veröffentlicht wird. Das ReN ist nach verschiedenen Zeit- und Sprachräumen sowie Feldern der Schriftlichkeit strukturiert (vgl. Barteld et al. 2017: 227f., Peters / Nagel 2014: 167–169). Zudem können die Texte des Korpus differenziert nach umfangreichen Metadaten, unter anderem zur Kommunikationssituation, zur äußeren Form oder zum Genre (z.B. Prosa vs. Vers), betrachtet werden.

Das nach verschiedenen Parametern strukturierte Korpus als Basis der Analysen ermöglicht die Umsetzung zweier weiterer methodischer Prinzipien der neuen Grammatik: Die Variationssensitivität und die diasystematische Differenziertheit, die in neueren grammatischen Darstellungen zunehmend als wesentliche Parameter erkannt worden sind. So spielt sprachliche Variation unter anderem in neueren grammatischen Studien zur Gegenwartssprache eine Rolle, bspw. in der Variantengrammatik des Standarddeutschen (Dürscheidt / Elspaß / Ziegler 2018) und in der Korpus-Grammatik des IDS (http://www1.ids-mannheim.de/gra/projekte/korpusgrammatik.html?L=0). Wie die Abbildung von zeitlich und räumlich sowie durch die Überlieferungsform bedingter Variation auf der Basis eines umfangreichen Korpus für eine historische Grammatik erfolgen kann, lässt sich anhand der Neuerarbeitung der Mittelhochdeutschen Grammatik beobachten (vgl. Herbers 2014), von der die Bände zur Wortbildung (Klein u. a. 2009) sowie zur Flexionsmorphologie (Klein u. a. 2018) publiziert sind. Auch im Konzept der geplanten Neuerarbeitung der Mittelniederdeutschen Grammatik stehen Korpusbezug und Variation im Mittelpunkt.

Die geplante Gesamtgrammatik des Mittelniederdeutschen soll mit Bezug auf die zu berücksichtigenden Sprachebenen (Graphemik, Phonologie, Lexemklassifizierung, Lexembildung und Flexion, Syntax und Text inklusive Pragmatik) in mehreren Schritten bearbeitet werden. In einem ersten Zugriff steht die Flexionsmorphologie im Fokus. Gerade für flexionsmorphologische Analysen bildet das ReN mit seinen Annotationen zu Wortart (PoS), Flexionsmorphologie und Lemma die ideale Basis. Von den insgesamt ca. 2,3 Mio. Token der finalen Korpusversion ReN 1.0 (http://hdl.handle.net/11022/0000-0007-D829-8) sind knapp 1,4 Mio. Token grammatisch annotiert. Damit ist eine unabdingbare Voraussetzung für eine korpusbezogene variationssensitive Grammatik erfüllt. Die Annotation der Wortarten und der Flexionsmorphologie ist mit dem im ReN entwickelten HiNTS (Historisches-Niederdeutsch-Tagset) erfolgt, das auf dem HiTS (Historisches Tagset; vgl. Dipper et al. 2013) und dem STTS (Stuttgart-Tübingen-Tagset; vgl. Schiller et al. 1999) basiert. Zusätzlich wurde auf der Basis einer digitalen Lemmaliste für das Mittelniederdeutsche eine Lemmatisierung vorgenommen (vgl. Kleymann et al. 2015).

Eine zentrale Anforderung an die neue mittelniederdeutsche Grammatik, die bereits in den grammatischen Annotationen des ReN berücksichtigt wurde, stellt die Vermeidung von Vorgriffen und bestimmten Interpretationen, welche die Ergebnisse entscheidend beeinflussen, dar. Gerade in einer korpusbasierten Grammatikschreibung ist die Auseinandersetzung mit der Frage, inwieweit die genutzten Daten und deren Annotation das Ergebnis der Analyse vorprägen, von besonderer Bedeutung. So basiert bspw. die Festsetzung der PoS-Tags im Tagset auf Vorannahmen zur Differenzierung zwischen bestimmten Wortarten in der jeweiligen Sprache. Auch im Bereich der mittelniederdeutschen Flexionsmorphologie können durch Interpretationen der Annotatorinnen und Annotatoren, z.B. aufgrund vorhandener grammatischer Darstellungen oder durch Übertragungen aus dem Gegenwartsdeutschen, die Ergebnisse vorweggenommen werden. Gleichzeitig jedoch kann keine grammatische Annotation absolut frei von bestimmten Annahmen über die Grammatik der jeweiligen Sprache erfolgen. Wie gezeigt werden soll, wird im ReN daher eine Balance zwischen notwendigen und vermeidbaren Vorannahmen sowie zwischen Oberflächenbezogenheit und Einbeziehung des sprachlichen Kontextes angestrebt. So gilt im Bereich der Valenz von Verben beispielweise die Regel, dass für das Subjekt im Satz der Nominativ annotiert werden kann (auch wenn der Beleg rein formal z.B. nach Nominativ oder Akkusativ flektiert ist), sofern keine eindeutig davon abweichende Form vorliegt. Eine Unterscheidung zwischen Genitiv-, Dativ- und Akkusativobjekt hingegen kann lediglich auf der Basis einer eindeutigen Flexionsform erfolgen. Da bislang kein Valenzwörterbuch des Mittelniederdeutschen existiert, ist bei ambigen Formen eine Auflösung nicht möglich, ohne dadurch wiederum die Ergebnisse zur Valenz mittelniederdeutscher Verben vorwegzunehmen.

Im Vortrag soll erläutert werden, wie der beschriebenen Gefahr einer zirkulären Darstellung begegnet werden kann, um eine oberflächenbezogene grammatische Analyse zu ermöglichen. Im ReN wurde mit dem HiNTS die Auszeichnung von Ambiguitäten auf flexionsmorphologischer Ebene mithilfe von Portmanteau-Tags (Leech et al. 1994) vorgenommen (vgl. Barteld et al. 2014).[1] So erhält bspw. eine Form wie *mî* (Personalpronomen), die sowohl Dativ als auch Akkusativ repräsentieren kann und bei der aufgrund des vorliegenden Kontextes keine Disambiguierung möglich ist, das Tag „Dat-Akk“. Der Vorteil eines solchen Tags im Vergleich z.B. zum Asterisk, der für ambige Formen im STTS genutzt wird, besteht darin, dass

die konkrete Überschneidung – hier zwischen Dativ und Akkusativ – abgebildet wird und weitere Alternativen ausgeschlossen werden. Statt die bestehenden Ambiguitäten auf der Basis bestimmter Vorannahmen aufzulösen und auf diese Weise mit den Annotationen dieses Vorwissen zu bestätigen, werden mehrdeutige Formen als solche gekennzeichnet.

Bei einem solchen Vorgehen ist kritisch zu hinterfragen, ob durch die Vergabe unterspezifizierter Tags die Annotationsentscheidung unnötigerweise auf einen späteren Zeitpunkt verlagert wird. Sollten nämlich bestimmte zunächst ausgewiesene Ambiguitäten nachträglich aufgelöst werden, erfordert dies eine erneute Sichtung der Belege, was insgesamt einen Mehraufwand bedeutet. Diese spätere Analyse der Sprachdaten mit potentieller Neubewertung der grammatischen Annotation ist jedoch unabdingbar, um die oben erwähnte Zirkularität zu vermeiden. Kann bspw. auf Basis der ReN-Daten die Valenz eines Verbs im Mittelniederdeutschen (z.B. *rôpen*) mit Hilfe der Fälle nicht-ambiger Kasusannotation bei Objekten dieses Verbs (z.B. *he rep dat kint* Neut.Akk.S) eindeutig bestimmt werden, wäre bei Objekten mit ambiger Form (z.B. *he rep den sone* Masc.Dat-Akk.Sg) eine Auflösung der Ambiguität zugunsten des ermittelten Objektkasus möglich (z.B. *he rep den sone* [Masc.Akk.Sg]). Während der Annotation im Rahmen der Korpuserstellung sind diese Informationen noch nicht vorhanden, sodass eine Disambiguierung ohne Vorannahmen nicht möglich ist. Neben diese Fälle, in denen die Auszeichnung der Ambiguität im ersten Schritt notwendig ist, um die Ergebnisse nicht vorzuprägen, eventuell aber in einem zweiten Schritt anhand neuer korpusbasierter Erkenntnisse eine nachträgliche Disambiguierung erfolgt, treten solche Fälle, in denen auch zu einem späteren Zeitpunkt keinerlei Auflösung der Ambiguität möglich ist. Dies betrifft unter anderem die Ebene des Genus, wo formal bestehende Ambiguitäten (z.B. *dit is en spegel* Masc-Neut.Nom.Sg) ausschließlich durch den sprachlichen Kontext (z.B. durch einen eindeutig nach einem Genus markierten Determinierer) aufgelöst werden können, was bereits in der Annotation im ReN berücksichtigt wird. Diese Beispiele belegen, dass durch die Vergabe von Portmanteau-Tags auf flexionsmorphologischer Ebene Entscheidungen zugunsten eines Wertes vermieden werden, die entweder gar nicht oder zum Zeitpunkt der Korpuserstellung noch nicht getroffen werden können, und dass die Auszeichnung von Ambiguitäten für eine möglichst vorurteilsfreie Annotation der mittelniederdeutschen Sprachdaten zwingend notwendig ist.

Um die Anwendbarkeit dieses Verfahrens zu evaluieren und potentielle Vor- und Nachteile der Auszeichnung von Ambiguitäten durch Portmanteau-Tags im ReN gegenüber der Annotation mithilfe des Asteriks wie im STTS auszumachen, wurden Inter-Annotator-Agreement-Experimente durchgeführt (vgl. Barteld et al. 2018: 3943). Diese ergaben unter anderem, dass der Wert der Übereinstimmung zwischen den Annotierenden auf flexionsmorphologischer Ebene unter Verwendung des HiNTS ähnlich hoch ausfällt, wie wenn statt der Portmanteau-Tags solche mit Asterisk gesetzt würden. Die vergleichsweise höhere Zahl an potentiellen Tags im HiNTS führt somit nicht zu einer geringeren Qualität der Annotation. Dabei bieten jedoch die Portmanteau-Tags den entscheidenden Vorteil, eine bestehende Ambiguität in konkreter Form abzubilden.

Für eine variationssensitive Analyse der mittelniederdeutschen Flexionsmorphologie können solche Auszeichnungen von Ambiguitäten auf unterschiedlichen Ebenen genutzt werden, bspw. um den Gebrauch von Substantiven in verschiedenen Genera zu untersuchen. Hierfür können in einem ersten Schritt all diejenigen Lemmata ermittelt werden, bei denen eine Genusambiguität annotiert ist. Anschließend kann für spezifische Lemmata geprüft werden, wo sie in einer eindeutigen Genusform vorkommen. Disambiguierung wird hier durch den sprachlichen Kontext erreicht, z.B. durch einen eindeutig nach einem Genus flektierten Determinierer. In Analysen exemplarisch betrachteter Substantive wie „lîf“ und „dê^{i}l“ wird eine auf der Ebene des Genus zum Teil sehr unterschiedlich stark ausgeprägte Variation sichtbar. Während bei „dê^{i}l“ der Anteil der als Maskulinum und der als Neutrum annotierten Belege annähernd ähnlich hoch ausfällt, dominieren bei „lîf“ deutlich die als Neutrum annotierten Belege. Zudem zeigen sich vereinzelt je nach Text und Sprachraum unterschiedliche Verteilungen.

Ein Beispiel für die erwähnte Ausbalancierung von Vorwissen einerseits und reiner Oberflächenbezogenheit andererseits findet sich bei der Rektion von Präpositionen. Ausgehend von den Angaben im Mittelniederdeutschen Handwörterbuch von Lasch et al. (1956ff.) wurde bei der Annotation eine Eingrenzung auf bestimmte Kasus vorgenommen, z.B. bei *in* auf den Dativ und den Akkusativ. Gleichzeitig aber wurde eine Einbeziehung des semantischen Kontextes, z.B. bei *in* die Annotation von Dativ bei lokaler und Akkusativ bei direktionaler Semantik, vermieden und stattdessen die Ambiguität mithilfe von Portmanteau-Tags abgebildet. Bei Belegen, die eine von den Angaben im Wörterbuch abweichende eindeutige Form aufweisen, wurde eine Annotation zugunsten der tatsächlich vorliegenden Form vorgenommen. Auf diese Weise kann z.B. an bestimmten Stellen der sich im Niederdeutschen vollziehende Kasussynkretismus der Substantive, bei denen Dativ und Akkusativ zu einem obliquen Kasus auf der Basis der Akkusativform zusammenfallen, beobachtet und untersucht werden. Erste Analysen, die im Vortrag vorgestellt werden sollen, liefern im ReN mehrere Belege, in denen auf die Präpositionen *nâ* und *tô,* die im Mittelniederdeutschen überwiegend den Dativ regieren, Nominalphrasen im Akkusativ als Teil der Präpositionalphrase folgen (*nâ*: 22 Belege, *tô*: 14 Belege). Hierbei zeigt sich eine deutliche Konzentration auf Texte des 16. und 17. Jahrhunderts. Dies stützt die Hypothese, dass im späteren Mittelniederdeutschen der Kasuszusammenfall einsetzt. Zudem zeigt sich ein starkes Gewicht der Belege auf einer Quelle, der Seekarte von 1577, das durch den Inhalt des Textes, der zahlreiche Richtungsangaben enthält, zu erklären ist. Bei *nâ* fällt außerdem auf, dass es mit Akkusativ vereinzelt auch in früheren Texten aus dem niederrheinischen Sprachraum vorkommt, was auf eine gewisse diatopische Variation hindeutet.

Wie anhand der Ergebnisse erster Analysen gezeigt wird, kann dank der Oberflächenbasiertheit und der Auszeichnung von Ambiguitäten im ReN Variation erfasst und vor dem Hintergrund potentieller außer- und innersprachlicher Parameter beschrieben werden. Auf diese Weise leistet die geplante korpuslinguistisch basierte variationssensitive Grammatik einen entscheidenden Beitrag für die moderne mittelniederdeutsche Grammatikographie.

Fußnoten

1. Die Herausforderung ambiger bzw. unterspezifizierter Einheiten für linguistische Annotationen wurde zunächst vor allem für die semantische und syntaktische Ebene diskutiert (s. z.B. Bunt 2007, Kountz et al. 2007). Auch für die Auszeichnung von grammatischer Ambiguität in historischen Sprach-

daten wurden bereits Vorschläge gemacht (s. z.B. Pauly et al. 2012, Dipper et al. 2013), die sich jedoch auf die Annotation von Wortarten und syntaktischen Strukturen konzentrieren und flexionsmorphologische Mehrdeutigkeiten nicht berühren.

Bibliographie

Barteld Fabian / Ihden, Sarah / Schröder, Ingrid / Zinsmeister, Heike (2014): "Annotating descriptively incomplete language phenomena", in: *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop* 99–104 http://www.aclweb.org/anthology/W14-4915 [letzter Zugriff 09. Dezember 2019].

Barteld, Fabian / Dreessen, Katharina / Ihden, Sarah / Schröder, Ingrid (2017): „Das Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200–1650) – Korpusdesign, Korpuserstellung und Korpusnutzung" in: Becker, Anja / Hausmann, Albrecht (eds.): *Mittelniederdeutsche Literatur. Mitteilungen des deutschen Germanistenverbandes* 64/3: 226–241.

Barteld, Fabian / Ihden, Sarah / Dreessen, Katharina / Schröder, Ingrid (2018): "HiNTS: A Tagset for Middle Low German", in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* 3940–3945 http://www.lrec-conf.org/proceedings/lrec2018/summaries/870.html [letzter Zugriff 09. Dezember 2019].

Bunt, Harry (2007): "Semantic underspecification. Which technique for what purpose?" in: Bunt, Harry / Muskens, Reinhard (eds.): *Computing Meaning* Bd. 3 (= Studies in Linguistics and Philosophy 83). Dordrecht: Springer 55–85.

Colliander, Elof (1912): *Mittelniederdeutsches Elementarbuch* (Photokopie der Druckfahnen des nicht zur Veröffentlichung gelangten Werkes, das mit § 365 abbricht). Heidelberg: Winter.

Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): „HiTS: ein Tagset für historische Sprachstufen des Deutschen", in: *Journal for Language Technology and Computational Linguistics*, Special Issue, 28(1): 85–137 https://jlcl.org/content/2-allissues/9-Heft1-2013/5Dipper.pdf [letzter Zugriff 09. Dezember 2019].

Dürscheid, Christa / Elspaß, Stephan /Ziegler, Arne (2018): *Variantengrammatik des Standarddeutschen*. Ein Online-Nachschlagewerk http://mediawiki.ids-mannheim.de/VarGra/index.php/Start [letzter Zugriff 09. Dezember 2019]

Herbers, Birgit (2014): „Prozessualität und Variabilität in der Grammatikographie des Mittelhochdeutschen" in: Ágel, Vilmos / Gardt, Andreas (eds.): *Paradigmen der aktuellen Sprachgeschichtsforschung* (= Jahrbuch für germanistische Sprachgeschichte 5). Berlin, Boston: de Gruyter 135–149.

IDS (o.J.): Korpusgrammatik – grammatische Variation im standardsprachlichen und standardnahen Deutsch: http://www1.ids-mannheim.de/gra/projekte/korpusgrammatik.html?L=0 [letzter Zugriff 09. Dezember 2019].

Klein, Thomas / Solms, Hans-Joachim / Wegera, Klaus-Peter (2009): *Mittelhochdeutsche Grammatik*. Teil III: Wortbildung. Tübingen: Niemeyer.

Klein, Thomas / Solms, Hans-Joachim / Wegera, Klaus-Peter (2018): *Mittelhochdeutsche Grammatik*. Teil II: Flexionsmorphologie. Berlin, Boston: de Gruyter.

Kleymann, Verena / Nagel, Norbert / Peters, Robert (2015): „Die digitale Lemmaliste für das Mittelniederdeutsche im DFG-Projekt ‚Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200–1650)'", in: *Korrespondenzblatt des Vereins für niederdeutsche Sprachforschung* 122/2: 95–100.

Kountz, Manuel / Heid, Ulrich / Eckart, Kerstin (2007): "A LAF/GrAF based Encoding Scheme for underspecified Representations of syntactic Annotations", in: *Proceedings of the International Conference on Language Resources and Evaluation 2008* 2262–2269 http://www.lrec-conf.org/proceedings/lrec2008/pdf/569_paper.pdf [letzter Zugriff 09. Dezember 2019].

Lasch, Agathe (1914 / ^{2}1974 / Nachdruck 2011): *Mittelniederdeutsche Grammatik* (= Sammlung kurzer Grammatiken germanischer Dialekte. A. Hauptreihe 9). Tübingen: Niemeyer.

Lasch, Agathe / Borchling, Conrad (1956ff.): *Mittelniederdeutsches Handwörterbuch*. Fortgef. von Gerhard Cordes und Dieter Möhn. Hg. von Ingrid Schröder. Neumünster / Kiel / Hamburg: Wachholtz.

Leech, Geoffrey / Garside, Roger / Bryant, Michael (1994): „CLAWS4: The tagging of the British National Corpus", in: Proceedings of the 15th International Conference on Computational Linguistics (COLING) 1: 622–628.

Lübben, August (1882): *Mittelniederdeutsche Grammatik*. Nebst Chrestomathie und Glossar. Leipzig: Weigel.

Pauly, Dennis / Senyuk, Ulyana / Demske, Ulrike (2012): "Strukturelle Mehrdeutigkeit in frühneuhochdeutschen Texten, in: JLCL 27(2): 65–82 https://jlcl.org/content/2-allissues/10-Heft2-2012/5Pauly.pdf [letzter Zugriff 09. Dezember 2019].

Peters, Robert / Nagel, Norbert (2014): „Das digitale ‚Referenzkorpus Mittelniederdeutsch / Niederrheinisch (ReN)'" in: Ágel, Vilmos / Gardt, Andreas (eds.): *Paradigmen der aktuellen Sprachgeschichtsforschung* (= Jahrbuch für germanistische Sprachgeschichte 5). Berlin, Boston: de Gruyter 165–175.

ReN-Team (2019): Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). Archiviert im Hamburger Zentrum für Sprachkorpora. Version 1.0. Publikationsdatum 14.08.2019. http://hdl.handle.net/11022/0000-0007-D829-8 [letzter Zugriff 09. Dezember 2019].

Sarauw, Christian (1921): *Niederdeutsche Forschungen*. Bd. 1: Vergleichende Lautlehre der niederdeutschen Mundarten im Stammlande (= Historisk-filologiske Meddelelser 5.1). Kopenhagen: Høst.

Sarauw, Christian (1924): *Niederdeutsche Forschungen*. Bd. 2: Die Flexionen der mittelniederdeutschen Sprache (= Historisk-filologiske Meddelelser 10.1). Kopenhagen: Høst.

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 [letzter Zugriff 09. Dezember 2019].

Volltexttransformation frühneuzeitlicher Drucke – Ergebnisse und Perspektiven des OCR-D-Projekts

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Engl, Elisabeth
engl@hab.de
Herzog-August-Bibliothek Wolfenbüttel, Deutschland

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Deutschland

Hartmann, Volker
volker.hartmann@kit.edu
Karlsruher Institut für Technologie

Neudecker, Clemens
clemens.neudecker@europeana-newspapers.eu
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Deutschland

Einleitung

Das schriftliche Kulturgut des deutschsprachigen Raums aus dem 16.–18. Jahrhundert wird schon seit Jahrzehnten in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke (VD) zusammengetragen. Ein signifikanter Anteil der verzeichneten Titel wurde der Forschung bereits durch die Bereitstellung von Volldigitalisaten oder einzelnen Schlüsselseiten leichter zugänglich gemacht. Die Verfügbarmachung von Volltexten ist dagegen noch ein Desiderat der Forschung. Das DFG-Projekt OCR-D nimmt sich seit Oktober 2015 im Rahmen der Koordinierten Förderinitiative zur Weiterentwicklung von Verfahren für die Optical Character Recognition (OCR) dieser Aufgabe an, indem es eine modular aufgebaute Open Source-Software entwickelt, deren Werkzeuge alle für die Texterkennung nötigen Schritte abdecken sollen. Der modulare Ansatz ermöglicht es, die technischen Abläufe und Parameter der Texterkennung stets nachzuvollziehen und maßgeschneiderte Workflows zu definieren, die jeweils optimale Ergebnisse für spezifische Titel aus dem Zeitraum des 16. bis 19. Jahrhunderts liefern. Zudem werden Antworten auf die damit verbundenen konzeptionellen, informationswissenschaftlichen und organisatorischen Fragen gefunden.

Künftig sollen mithilfe der OCR-D-Software Volltexte generiert werden, die zum einen von Forschenden zur Recherche verwendet werden können. Zum anderen könnten diese zum Ausgangspunkt für Studien im Bereich der Digital Humanities (DH) werden, wobei auch auf diese Texte die textkritische Methode anzuwenden ist. Gerade bei einer automatisierten Weiterverarbeitung der erzeugten Volltexte ist es für Forschende unerlässlich, die Genese der von ihnen verwendeten Daten kritisch zu hinterfragen. Nur so können Eigenheiten der Daten, die Resultat von zuvor genutzten “Spielräumen” sind, von DH-Forschenden erkannt und in ihrem Umgang mit der Datengrundlage berücksichtigt werden. Nicht nur diese interpretatorischen Spielräume sind zu betrachten, sondern auch, welche konkreten Implementierungen den DH die gewünschten “Spielräume“ für die Erkenntnisgenerierung geben. Im Folgenden wird in vier Thesen eine notwendige Begrenzung der Spielräume vorgenommen. Diese Begrenzung ergibt sich aus dem Vergleich mit anderen Projekten und der heute gängigen Praxis. Ziel ist es, den Forderungen der DH nach qualitativ hochwertigen Volltexten gerecht zu werden.

Im Rückblick

Das Projekt hat sich in den vergangenen vier Jahren mit verschiedenen Themen auf der DHd zur Diskussion gestellt (Boenig et al 2016; Boenig et al 2018; Baierer et al 2019). Zu Beginn standen methodische Fragen, wie die Textqualität erhöht werden kann. Dabei wurden statistische Methoden vorgestellt, die auf Basis eines Vergleichs von mindestens zwei erstellten Textfassungen entwickelt wurden. Im Rahmen des Themas “Kritik der digitalen Vernunft” wurden die DH befragt, wie in den Geisteswissenschaften Ergebnisse ohne Ground Truth und Referenzdaten gewonnen bzw. verifiziert werden. Diesem Desiderat begegnete das Projekt OCR-D mit dem Vorschlag von Transkriptionsrichtlinien für die Erfassung von Ground Truth-Daten[1] und in der Folge mit der Definition von spezifischen Metadaten. Bei dem 2019 veranstalteten Workshop konnten Wissenschaftler und Wissenschaftlerinnen sowie Interessierte Einblicke in den OCR-D-Workflow erhalten. An Beispielen konnten die Möglichkeiten der Software demonstriert und getestet werden. Die Diskussion, Hinweise und Fragen wurden soweit wie möglich in OCR-D umgesetzt.

Thesen

Das Ziel der prototypischen Implementierung des OCR-D-Workflows und damit der Generierung von Forschungsdaten, die sich durch eine erkennbare XML-Strukturierung sowie eine hohe Zeichen- und Textqualität auszeichnen, wird im ersten Quartal 2020 erreicht werden. Dies stellt jedoch nicht das Ende des Weges dar, sondern eher den Beginn der nun folgenden Volltexttransformation. Letztlich besteht die Aufgabe darin, ca. 1 Mio. frühneuzeitliche Titel mit ca. 250 Mio. Seiten, die zum Teil bereits als Bilddigitalisate vorliegen, zu Volltextdigitalisaten zu transformieren.

1. Die Volltexttransformation der Bestände stellt eine Herausforderung für Bibliotheken und Archive dar. Die vorhandenen institutionellen und interinstitutionellen Vorgehens-

weisen und Konventionen sind möglichst zentral aufeinander abzustimmen, damit die Aufgabe in absehbarer Zeit gelöst wird.[2]

Es gibt bereits einige Projekte, in denen (Teil-)Bestände und Sammlungen volltextdigitalisiert wurden.[3] Deren Nutzen für die DH wird jedoch v.a. durch zwei Faktoren begrenzt: Zum einen weisen die erstellten Volltexte aufgrund fehlender Standards bzw. Konventionen im Bereich von Text- und Strukturerkennung eine große Bandbreite in der Transkription der Texte und der Benennung von Textstrukturen auf, die deren automatisierte Auswertung und Bearbeitung durch die DH erschweren. Zum anderen gibt es bislang keine zentrale Anlaufstelle, die die Bereitstellung und auch die Erstellung von Volltexten steuert. Dadurch sind die existierenden Volltexte sowohl für die Forschung, als auch für die volltextdigitalisierenden Einrichtungen weniger sichtbar, was die Gefahr aufwändiger und teurer Doppelarbeiten erhöht.

2. Die Volltexttransformation auf Basis von Erkennungssoftware, die neuronale Netze nutzt, setzt Trainingsdaten voraus. Diese fundamental wichtigen Daten sind systematisch aus vorhandenen Ressourcen zu gewinnen und aktiv zu erweitern.

Mit ihren Förderinitiativen von 2010 und 2013 hat die DFG die Bedeutung der Forschungsdaten und des zugehörigen Managements erkannt.[4] Heute sollten Projekte von Beginn an mit entsprechenden Forschungsdatenmangementplänen aufgesetzt und die entstehenden Daten in den zuvor bereitgestellten Repositorien verwahrt werden.[5] Gerade bei der automatisierten Texterfassung im Rahmen von Editionsprojekten werden in der Regel aber nur die abschließend bearbeiteten und korrigierten Daten veröffentlicht. Eine Nachnutzung dieser Daten ist in vielfacher Hinsicht nur begrenzt möglich. Dabei spielt nicht nur das Format der Daten, sondern auch die Methodik der Datenerfassung eine entscheidende Rolle. Für die Nachnutzung ist eine Transformation dieser Daten nötig, die entweder von den Nutzenden zu leisten ist, oder von den bestandsverwaltenden Einrichtungen angeboten werden könnte. Um eine solche Transformation zu gewährleisten, sind sowohl Richtlinien als auch entsprechende Metadaten zu etablieren, damit vergleichbare und konsistente Daten bereitgestellt werden können.[6]

3. Die Volltexttransformation wird für einen Teil der Dokumente ein Prozess sein, der sich über einen größeren Zeitraum wiederholt.

Digitale Daten müssen beständig gepflegt und aktualisiert werden. Dies haben auch Bibliotheken als Herausforderung der digitalen Transformation ihrer Bestände erkannt (vgl. Kempf 2015: 277–278). Werden lernende Systeme für die Text- und Strukturerkennung genutzt, können diese in absehbaren Intervallen verbessert werden.[7] Denn die Verbesserung bestehender Algorithmen sowie die Nutzung zusätzlicher oder verbesserter Trainingsdaten führt auch zu besseren Ergebnissen in der Text- und Strukturerkennung, wie sich beispielsweise im GoogleBooks-Projekt[8] zeigt. Diese wiederkehrende Prozessierung muss konzeptionell berücksichtigt werden.

4. Die Volltexttransformation muss in ihrer Qualität von den Nutzenden beurteilbar sein.

Bibliotheken geben den Nutzenden mit ihrem Bestand und dessen Erschließung ein Qualitätsversprechen. Die Nutzenden können sich auf die vorhandenen Daten verlassen und sie z.B. in Bibliographien verwenden. Das Volltextangebot aus der automatischen Texterkennung kann dagegen häufig nur unpräzise als "schmutzige OCR"[9] bezeichnet werden. Diese pauschale Angabe ermöglicht den DH keine verlässliche Qualitätseinschätzung und führt dazu, dass Volltextbestände oft a priori als minderwertig eingeschätzt werden. Daher besteht die Gefahr, dass projektintern eine erneute Volltextdigitalisierung durchgeführt wird, die nicht immer sinnvoll ist, da die Erkennung teilweise nur durch eine Korrektur verbessert werden könnte. Oder es könnten im umgekehrten Fall auf Grund einer ungenauen bzw. zu groben Einschätzung aufwendige Korrekturen vorgenommen werden. In beiden Fällen werden finanzielle und personelle Ressourcen verschwendet.

Lösungen und Desiderate des OCR-D-Projekts

Zu 1: Die bisherigen umfassenden Bilddigitalisierungsarbeiten im VD17 wurden über einen Masterplan gesteuert, um die große Anzahl an Titeln effizient, in nachnutzbarer Form verarbeiten zu können und Doppelarbeiten zu vermeiden. Ein ähnliches Vorgehen, bei dem die zu prozessierenden Titel an interessierte Einrichtungen verteilt werden, dürfte auch für die Volltexttransformation der VD zielführend sein. Die Voraussetzungen und Rahmenbedingungen für diese Arbeiten wurden von dem OCR-D-Koordinierungsprojekt um die Jahreswende 2019/2020 durch eine Umfrage mit den VD-Bibliotheken zusammengetragen. OCR-D wird die mehrjährige Projekterfahrung im Austausch mit den verschiedenen Stakeholdern nutzen, um die Nachnutzbarkeit von Daten und Abläufen zu verbessern, sowohl mit technischer Dokumentation und Best Practices, als auch als Katalysator für einen ergebnisorientierten, inklusiven Diskurs zur Etablierung von Standards.

Zu 2: Für die Transkription von Texten gibt es unzählige Richtlinien, die von verschiedenen Fächern, Arbeitskreisen und Forschungsprojekten entsprechend ihrer jeweiligen Anforderungen aufgestellt und wiederum an die spezifischen Erfordernisse bestimmter Transkriptionsprojekte angepasst wurden. Bei diesen Gruppen ist zum einen ein Bewusstsein dafür zu schaffen, ihre Transkriptionen auch mit Blick auf deren Nachnutzbarkeit durch andere Projekte anzufertigen. Zum anderen sind interdisziplinär erarbeitete und gültige Transkriptionsrichtlinien ein großes Desiderat der Forschung. Erste Impulse hierfür könnten große Fördergeber wie bspw. die DFG geben, indem Praxisrichtlinien geschaffen werden, die von Antragstellern zu beachten sind. Das OCR-D-Projekt ist zudem darum bemüht, seine auf Grundlage des DTA-Basisformats erstellten Transkriptionsrichtlinien interdisziplinär zur Nutzung durch weitere Projekte zu kommunizieren.

Zu 3: Modelltraining mit tesstrain und okralact

Das Projekt ocropy, die Python-Implementierung von Tom Breuels OCRopus-Projekt, brachte neben Werkzeugen für die Text- und Strukturerkennung auch Werkzeuge für das Erstellen von Ground Truth und das Trainieren neuer Modelle mit sich. Mit diesen Werkzeugen und einigen Anpassungen lassen sich auch die auf ocropy basierenden Weiterentwicklungen Calamari und Kraken trainieren. Insbesondere für tesseract, die mit Abstand am meisten genutzte Open Source OCR, gab es bis 2018 kaum Dokumentation oder Tooling für das Training. Daher wurde im Rahmen von OCR-D *ocrd-train* entwickelt, eine Makefile-basierte Lösung zum Trainieren von Tesseracts LSTM-Engine, das inzwischen unter dem Namen *tesstrain* vom Tesseract-Entwicklerteam gepflegt und weiterentwickelt wird.[10] Die Aufrufe zum Training von Texterkennungsmodel-

len und insbesondere das Inventar an freien Parametern sind allerdings in hohem Maße engine-spezifisch, keineswegs trivial und erfordern zur optimalen Feinadjustierung manuelle Intervention. Daher entwickelt OCR-D seit 2019 das Werkzeug okralact,[11] das über ein komfortables Webinterface und ein skalierbares Backend ein Training aller relevanter Open Source OCR-Engines mit einem einheitlichen Interface ermöglichen wird.

Zu 4: Nachkorrektur und Qualitätsanalyse

Innerhalb des OCR-D-Projektes beschäftigen sich zwei Projekte mit der automatischen, bzw. semi-automatischen Nachkorrektur von OCR-Texten. Das Hauptproblem dabei ist es, historische Schreibweisen und Druckfehler von OCR-Fehlern zu unterscheiden. Für moderne Texte würde eine reine Rechtschreiberkennung genügen, wie sie in jedem Textverarbeitungsprogramm verfügbar ist. Die Projekte kooperieren und haben verschiedene Verfahren entwickelt, basierend auf einem Fehler-Profiler, neuronalen Netzen oder endlichen Automaten. Als trainierbare Algorithmen werden sie, analog zur Struktur- und Texterkennung, mit mehr und besseren Trainingsdaten bessere Ergebnisse liefern. Was "besser" bedeutet ist noch Gegenstand der Forschung. OCR-D bringt sich in die Entwicklung ein und unterstützt tatkräftig Projekte wie dinglehopper[12] (ein Werkzeug zur Fehlervisualisierung). Gerade im Bereich der Ground-Truth-freien Evaluation von Text und der Qualitätsanalyse von Strukturdaten gibt es noch große Lücken im Software-Portfolio, die zu schließen sich OCR-D auch weiterhin befleißigen wird.

Ausblick

Ab der ersten Jahreshälfte 2020 werden die entwickelten Software-Komponenten im OCR-D-Workflow verankert sein. Damit tritt diese Software immer mehr aus dem Projektstadium heraus und wird in den produktiven Einsatz überführt. Um kontinuierlich gute Erkennungsergebnisse mit dem aus fast vier Jahrhunderten stammenden Material zu erhalten, sind Optimierungen notwendig. Dabei wird stets darauf abgezielt, Forschungsdaten aus den digitalen Beständen der Bibliotheken zu erzeugen und nicht unstrukturierte Textdaten. So wird die Volltexttransformation in einem umfassenden Maße Grundlagen für datenzentrierte Digital Humanities schaffen.

Fußnoten

1. Im Kontext von OCR bezeichnet *Ground Truth* manuell korrigierte, fehlerfreie Transkriptionen. Diese werden zum einen für das Training von OCR-Engines, zum anderen für die Evaluation der OCR-Ergebnisse benötigt.
2. Der Gedanke folgt der neunten Empfehlung ("Establish an 'OCR Service Bureau'") aus dem Report von Smith und Cordell (2018).
3. Vgl. bspw. die folgenden Projekte, die sich auf unterschiedlich große (Teil-)Bestände beziehen: Helmstedter Drucke Online: http://www.hab.de/de/home/wissenschaft/forschungsprofil-und-projekte/helmstedter-drucke-online.html; Über 14.000 preußische Drucke des 17. Jahrhunderts online verfügbar: https://blog.sbb.berlin/ueber-14-000-preussische-drucke-des-17-jahrhunderts-online-verfuegbar/; Projekt Digi20 https://digi20.digitale-sammlungen.de/de/fs1/about/static.html
4. Nachdem im Jahr 2010 der Aufbau von Infrastrukturen für Forschungsdaten von der DFG ausgeschrieben worden war, wurde drei Jahre später das Förderprogramm „Informationsinfrastrukturen für Forschungsdaten" eingerichtet. Vgl. DFG 2019: 7.
5. Zur aktuellen Situation des Datenmanagements und der Rolle, die Bibliotheken in diesem Bereich einnehmen (könnten), vgl. Neuroth et al 2019.
6. Die Notwendigkeit einheitlicher Richtlinien wird besonders an Projekten wie "Venice Time Machine" deutlich, dessen bereits vorhandenen 8 TB an Daten aufgrund fehlender einheitlicher Richtlinien und Vorgehensweisen bei der Erfassung der Metadaten für die Forschung vermutlich wertlos sind. Vgl. Castelvecchi 2019: 607.
7. Kempf geht davon aus, dass mit OCR-Software nie völlig fehlerfreie Volltexte generiert werden können. Vgl. Kempf 2015: 274.
8. Während die OCR-Ergebnisse im Rahmen des Google-Books-Projekts zunächst insgesamt unbefriedigend, für gebrochene Schriften vollkommen unbrauchbar waren, konnten ab dem Jahr 2008 einzelne Frakturtexte in ausreichender Qualität prozessiert werden. In den letzten Jahren konnte die Erkennungsrate noch deutlich gesteigert werden. Vgl. Wikisource: Google Book Search.
9. Bspw. gibt Google die Fehlerquote im Google Books pauschal mit 1,37 % an (vgl. Kempf 2015: 272). Diese für die wissenschaftliche Nutzung hohe Fehlerrate unterscheidet sich, bedingt durch die Vielfalt an Typen und Layouts sowie den großen Publikationszeitraum der digitalisierten Bücher, von Text zu Text deutlich.
10. https://github.com/tesseract-ocr/tesstrain
11. https://github.com/OCR-D/okralact
12. https://github.com/qurator-spk/dinglehopper

Bibliographie

Baierer, Konstantin / Boenig, Matthias / Hartmann, Volker / Hermann, Elisa / Neudecker, Clemens (2019): *„Vom gedruckten Werk zu elektronischem Volltext als Forschungsgrundlage"* (Workshop) (https://zenodo.org/record/2596095/files/2019_DHd_BookOfAbstracts_web.pdf, S. 58).

Boenig, Matthias / Würzner, Kay-Michael / Binder, Arne / Springmann, Uwe (2016): *„Über den Mehrwert der Vernetzung von OCR-Verfahren zur Erfassung von Texten des 17. Jahrhunderts."* Vortrag auf der DHd 2016, 7.12.03.2016 in Leipzig (http://dhd2016.de/boa.pdf#page=103).

Boenig, Matthias / Federbusch, Maria / Herrmann, Elisa / Neudecker, Clemens / Würzner, Kay-Michael (2018): *„Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?"*. Vortrag auf der DHd2018, 28.02.2018 in Köln (http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf#page=221).

Castelvecchi, Davide (2019): *"Venice 'Time Machine' Project Suspended amid Data Row. Disagreements between International Partners Leave Plans to Digitize the Italian City's History in Limbo"* in: *Nature* 574: 607.

DFG (2019): *„Weiterentwicklung des Förderprogramms ‚Informationsinfrastrukturen für Forschungsdaten'"* https://zenodo.org/record/2650866 [6.3.2019 / 26.9.2019].

Kempf, Klaus (2015): *„Data Curation oder (Retro-)Digitalisierung ist mehr als die Produktion von Daten"* in: *o-bib* 4: 268–278.

Neuroth, Heike / Rothfritz, Laura / Petras, Vivien / Kindling, Maxi (2019): "Digitales Datenmanagement als neue Aufgabe für wissenschaftliche Bibliotheken" in: Bibliothek. Forschung und Praxis 43: 421–431.

Smith, David A. / Cordell, Ryan (2018): *"A Research Agenda for Historical and Multilingual Optical Character Recognition"* http://hdl.handle.net/2047/D20297452 [9.12.2019].

Wikipedia, Die freie Enzyklopädie (2019): *„Google Books"*. https://de.wikipedia.org/w/index.php?title=Google_Books&oldid=189583765 [16.6.2019 / 25.9.2019].

Wikisource (2019): *"Google Book Search"* https://de.wikisource.org/wiki/Wikisource:Google_Book_Search [22.8.2019 / 26.9.2019].

Wege bereiten, vermitteln und Denkräume schaffen! Digital Humanities als community-induziertes Phänomen

Wuttke, Ulrike
ulrike.wuttke@gmx.net
Fachhochschule Potsdam, Deutschland

Einleitung

Inzwischen haben computergestützte Informationstechnologien als Teil der digitalen Transformation einen festen Platz in allen Bereichen der geisteswissenschaftlichen Forschung und Lehre gefunden, oftmals unter der Bezeichnung digitale Geisteswissenschaften bzw. Digital Humanities (vgl. Borgman 2015: 37, 161-164; Thaller 2017: 3-5). Die gezielte Unterstützung der digitalen Transformation stellt alle Akteure des Wissenschaftssystems, nicht nur in den Digital Humanities, vor große Herausforderungen und findet seinen Niederschlag auf wissenschafts- und förderpolitischen Agenden (vgl. RfII 2016: 9). Neben neuen Berufsbildern, neuen Studiengängen, einer neuen Datenkultur und neuen Anreizsystemen, wird in diesem Zusammenhang auch die Forderung nach der Schaffung geeigneter infrastruktureller Rahmenbedingungen laut (vgl. RfII 2016: 49-58).

Im Kontext dieser Diskussion spielen die geisteswissenschaftliche Forschung begleitende und unterstützende lokale, nationale und internationale (digitale) Infrastrukturen eine zentrale Rolle, wobei zunächst meist deren technologische Aspekte im Mittelpunkt stehen (vgl. Thaller 2017: 11). Inzwischen wird jedoch auch verstärkt der soziale Aspekt von Infrastrukturen thematisiert und ihre Rolle als soziale Netzwerke (sogenannte Peer-to-Peer-Netzwerke) anerkannt, denen eine grundlegende Bedeutung bezüglich der Veränderung der Forschungskulturzukommen kann. Mit anderen Worten, der Begriff digitale Forschungsinfrastruktur sollte nicht darüber hinwegtäuschen, dass auch im Mittelpunkt digitaler Forschungsinfrastrukturen immer zunächst Menschen und ihre Interaktionen, Bedürfnisse und Forschungsinteressen stehen (sollten), d. h. digitale Forschungsinfrastrukturen in den Geisteswissenschaften haben wichtige soziale Komponenten (vgl. Wissenschaftsrat 2011: 70; Anne et al. 2017: 26–29).

Während sich auf nationaler und internationaler Ebene bereits verschiedene – mehr oder weniger breit angelegte – Initiativen, Projekteu. ä. auf das Ermitteln von Bedarfen und die Entwicklung und Bereitstellung entsprechender Ressourcen und Dienste für die Digital Humanities spezialisiert haben (wie z. B. der DHd-Verband, CLARIAH oder einzelne Konsortien der zukünftigen NFDI), besteht bezüglich der durch die zunehmende Digitalisierung der Geisteswissenschaften bedingten neuen Organisationsformen und infrastrukturellen Bedarfe an einzelnen Standorten noch konkreter Forschungs- und Handlungsbedarf (vgl. HRK 2014, Roeder et al. 2019).

Forschungsfrage

Unabhängig davon, ob die Digital Humanities als eigenständige Disziplin betrachtet werden oder als ein Phänomen in anderen geisteswissenschaftlichen Fächern als Ausdruck verschiedener Grade der digitalen Transformation (vgl. Sahle 2015), wird einerseits postuliert, dass sie ihnen als Motor der Transformation und Reflexion eine tragenden Rolle in diesen soeben angerissenen Prozessen zukommt, andererseits wird Kritik geäußert, dass es sich um einen Hype handelt, der seinen tieferen Sinn noch nicht bewiesen hat (vgl. DFG o. J.; Posner 2016; Underwood 2019). Wenn wir das Prinzip *In dubio pro reo* gelten lassen wollen, d. h. "Im Zweifel für den Angeklagten", dann müssen wir uns zunächst ernsthaft und vorbehaltlos fragen, welche Bedingungen den Digital Humanities zuträglich sind, um die ihnen zugesprochenen Potenziale voll zu entfalten.

Fakt ist, dass die Digital Humanities nicht in einem luftleeren Raum existieren. Ihr „Spielraum" befindet sich zwischen „traditionellen" geisteswissenschaftlichen Fächern und je nach Forschungskontext anderen relevanten Fächern, wie z. B. der Informatik oder den Medienwissenschaften. Es stellt sich damit die besondere Herausforderung, Rahmenbedingungen zu schaffen, die einerseits „genuinen" Digital Humanities-Forschungsaktivitäten (im engeren Sinn als Brückenfach zwischen Geisteswissenschaften und Informatik, vgl. Sahle 2015) und anderseits der breiteren Digitalisierung geisteswissenschaftlicher Forschungsprozessedienlich sind, d. h. Wissenschaftler*innen und Studierende bei der nachhaltigen Implementierung neuer Forschungsparadigmen zu unterstützen (vgl. Harrower 2015: 12).

Trotz der Zentralität der Fragestellung, werden diese institutionellen und infrastrukturellen Dimensionen der Digital Humanities in Deutschland noch relativ selten übergreifend reflektiert. Zwar existieren bereits verschiedene Modelle und Ansätze zur institutionellen Förderung der Digital Humanities, insbesondere aus anglo-amerikanischer Sicht (vgl. für den anglo-amerikanischen Raum u. a. Posner 2016; Anne et al. 2017), aber es liegen nur wenige auf das deutsche universitäre System bezogene Erkenntnisse vor (vgl. für Deutschland u. a. Burghardt und Wolff 2015). Angesichts des zunehmenden Bewusstseins, dass die Digital Humanities neue Anforderungen an die Organisationsformen der geisteswissenschaftlichen Forschung und Lehre in ihrer Gesamtheit stellen, greift der Beitrag die Konkretisierung dieser Anforderungen als For-

schungsdesiderat auf. Im Mittelpunkt des Beitrags stehen die Ergebnisse einer Untersuchungmit der zentralen Frage, wie aus demsogenannten *Computational Turn* der Geisteswissenschaften entstehende Bedarfe konkret in institutionellen Digital-Strategien adressiert werden können, speziell für den *Use Case* der Digital Humanities-Forschungan deutschen Universitäten (vgl. Wuttke 2019).

Methode

Ziel der Untersuchung war es, basierend auf externen Erfahrungswerten potentielle Erfolgsfaktoren für einen universitären Digital Humanities-Schwerpunkt herauszuarbeiten, die als Anhaltspunkt für institutionelle Strategieprozesse dienen und jeweils entsprechend der individuellen Rahmenbedingungen verfeinert werden können. Es sollten unmittelbare Einsichten in *Good Practices* aus der Sicht von in Digital Humanities-"Labore" involvierten Wissenschaftler*innen gewonnen werden und in Relation zu Erfahrungswerten aus dem In- und Ausland gesetzt werden. Hierfür wurden die Rahmenbedingungen und Entwicklungspfade vier erfolgreicher deutscher Digital Humanities-Standorte aus der Sicht beteiligter Wissenschaftler*innen analysiert (Expert*inneninterviews) und vor dem Hintergrund nationaler und internationaler Entwicklungen Empfehlungen abgeleitet.

Die Standorte für die Expert*inneninterviews wurden durch eine quantitative Erhebung ermittelt, die auf einem eigens für die Untersuchung entwickelten Vorschlag für die Quantifizierung der Forschungsstärke von Digital Humanities-Standorten beruht, der im Beitrag näher erläutert wird. Zentrales Element der Auswahl der Standorte für die Expert*inneninterviews war eine umfängliche Auswertung der zur Verfügung stehenden Books of Abstracts vergangener DHd-Konferenzen (2015-2018) anhand des zuvor definierten Kriteriums "Erfolg in der Digital Humanities-Forschung". Letztendlich wurden mit jeweils einem Vertreter oder einer Vertreterin der Digital Humanities-Forschung der am besten plazierten universitären Standorte, nämlich der Julius-Maximilians-Universität Würzburg, der Universität zu Köln, der Humboldt-Universität zu Berlin und der Universität Stuttgart leitfadenbasierte Expert*inneninterviews geführt. Der hierfür verwendete Interviewleitfaden wurde aufgrund des Stands der Forschung unter Einbeziehung eigener Erfahrungen formuliert und mit den Expert*innen der ausgewählten Standorte erörtert.

Ergebnisse

Im Mittelpunkt des Beitrags steht die Diskussion eines aus den Ergebnissen der Expert*inneninterviews abgeleiteten fünfstufigen Modells infrastruktureller Erfolgsfaktoren für die universitäre Digital Humanities-Forschung, insbesondere die aus Sicht der Expert*innen diesbezüglich essentielle Rolle sozialer Faktoren.

Das im Folgenden vorgestellte, abstrakte, fünfstufige Modell ist erweiterbar und modifizierbar. Seine Ebenen reichen von der Schaffung von Grundvoraussetzungen (*state of mind*) bis zur Etablierung komplexer bzw. langfristiger Strukturen. Im Beitrag werden die fünf Ebenen bzw. infrastrukturellen Erfolgsfaktoren näher erläutert:

- Interdisziplinäre Gesprächsbereitschaft und interdisziplinärer Dialog,
- Förderung eines günstigen Klimas für Kooperationen,
- Personelle und räumliche Bündelung von Digital Humanities-Aktivitäten,
- Enge Verzahnung Digital Humanities und Geisteswissenschaften und
- Stärkung der Nachhaltigkeit.

Das vorgestellte fünfstufige Modell unterstreicht, dass die Digital Humanities als *community*-induziertes Phänomenzu betrachten sind. Digital Humanities-Schwerpunkte benötigen zu ihrer infrastrukturellen Stimulierung und Konsolidierung der 1) Induktion in Form von Wegbereiter*innen bzw. Vermittler*innen und der 2) Inkubation in Form von institutionellen Denkräumen zwischen geisteswissenschaftlichen Forschungsfragen und informationstechnischen Lösungswegen (vgl. Edmond 2016: 57; Rehbein und Sahle 2013: 227). In diesem Zusammenhang wird auch der ungebrochene aber in der Forschung nicht unkritisch betrachtete Trend zur Bündelung bzw. Institutionalisierung der Digital Humanities als sogenannte Digital Humanities Center (DHC) diskutiert (vgl. Maron und Pickle 2014; Prescott 2016; Mortiz et al. 2017: 102).

Anschließend wird ein sich aus den Expert*inneninterviews extrahiertes Grundmuster als mögliche Handlungsempfehlungen für die Entwicklung von institutionellen Digital Humanities-Strategien mit dem Ziel der breiteren digitalen Durchdringung der Geisteswissenschaften zur Diskussion gestellt. Diese Empfehlungen richten sich insbesondere an Personen, die an entsprechenden universitären Strategieprozessen beteiligt sind, sie können aber auch für andere Einrichtungstypen modifiziert werden. Konkret handelt es sich um folgende Empfehlungen:

- Schaffung zentraler Vermittler*innen und institutioneller Denkräume (institutionelle Bündelung und Unterstützung),
- Gezielte, enge Verzahnung von Digital Humanities-Aktivitäten mit geisteswissenschaftlichen Forschungscommunities (Akzeptanz, Kooperationsanbahnungen, Transformationsimpulse),
- Einrichtung dedizierter Digital Humanities-Professuren und -Studiengänge (Schwerpunktsetzung in der Lehre, Ansprech- und Kooperationspartner*innen),
- Technische und personelle Nachhaltigkeit der Digital Humanities-Infrastrukturen (Ausgleich zwischen Dienstleistung und Forschung sowie Standardisierung und Innovation).

Abschließend wird thematisiert, welche Bedeutung dieses Ergebnis für den Auf- und Ausbau universitärer Digital Humanities-Schwerpunkte und darüber hinaus hat, d. h. seine Implikationen für die breitere Diskussion über infrastrukturelle Rahmenbedingungen und Organisationsmodelle der Digital Humanities, ihre zukünftigen Entwicklungspfade und ihr Selbstverständnis als Community. Hierbei ist besonders die grundlegende Wichtigkeit der ganzheitlichen Betrachtung infrastruktureller Erfolgsfaktoren hervorzuheben, d. h. die Einbeziehung sozialer und technologischer Aspekte, weil soziale und wissenschaftspolitische Prozesse eine essentielle Rolle für das Digital Humanities-Ökosystem spielen und einen Einfluss auf die Leistungsfähigkeit und die Ausstrahlung der Digital Humanities auf die breiteren Geisteswissenschaften haben (vgl. Hügi und Schneider 2013: i). Mit anderen Worten, die Digital Humanities sind ein stark *community*-induziertes Phänomen und für ihren Erfolg sind soziale Faktoren und Kompetenzen wie Kooperationsgeist, Interdisziplinarität und die

Etablierung einer Kultur des Lernens und riskanten Denkens (und Scheitern-Dürfens!) genauso wichtig wie technologische Aspekte (vgl. Lewis et al. 2015: 2). Diese Erkenntnisse haben neben der Auslotung hierfür förderlicher Organisationsmodelle auch Bedeutung für die Besetzung universitärer Curricula.

Die gezielte und enge Verzahnung von Digital Humanities-Aktivitäten mit der geisteswissenschaftlichen Forschung – im Fall von Universitäten mit den geisteswissenschaftlichen Fakultäten – ist laut dieser Studie ein wichtiger Schlüssel für ihre breitere Akzeptanz, für Kooperationsanbahnungen und Transformationsimpulse. Die vorgestellten Ergebnisse sollen darüber hinaus als Diskussionsimpuls für die Übertragung auf andere, vor ähnlichen Herausforderungen stehende, Institutionen, wie z. B. außeruniversitäre Forschungseinrichtungen, dienen.

Bibliographie

Anne, Kirk M. et al. (2017): *Building Capacity for Digital Humanities: A Framework for Institutional Planning*. ECAR Working Group Paper, https://library.educause.edu/-/media/files/library/2017/5/ewg1702.pdf [letzter Zugriff 15.09.2019].

Borgman, Christine L. (2015): *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge (Mass.), London: MIT Press.

Burghardt, Manuel / Wolff, Christian (2015): "Zentren für Digital Humanities in Deutschland", in: *Information – Wissenschaft & Praxis* 66(5–6): 312–326, DOI: 10.1515/iwp-2015-0056 https://www.degruyter.com/downloadpdf/j/iwp.2015.66.issue-5-6/iwp-2015-0056/iwp-2015-0056.xml [letzter Zugriff 15.09.2019].

DFG (o.J.): GEPRIS Detailseite Projekt Symposienreihe 'Digitalität in den Geisteswissenschafteften'. Titel der Webseite: DFG – Deutsche Forschungsgemeinschaft https://gepris.dfg.de/gepris/projekt/287972711 [letzter Zugriff 15.09.2019].

Edmond, Jennifer (2016): "Collaboration and Infrastructure" in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.), *A New Companion to Digital Humanities*, Chichester (West Sussex): Wiley Blackwell 54–65.

Harrower, Natalie (ed.) (2015): *Going Digital: Creating Change in the Humanities*. ALLEA E-Humanities Working Group Report http://dri.ie/sites/default/files/files/Going-Digital-digital-version.pdf [letzter Zugriff 15.09.2019].

HRK (2014): *Management von Forschungsdaten: Eine zentrale strategische Herausforderung für Hochschulleitungen* https://www.hrk.de/fileadmin/_migrated/content_uploads/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf [letzter Zugriff 15.09.2019].

Hügi, Jasmin / Schneider, René (2013): *Digitale Forschungsinfrastrukturen für die Geistes- und Geschichtswissenschaften*, Genf https://doc.rero.ch/record/31535/files/Schneider_Digitale_Forschungsinfrastrukturen.pdf [letzter Zugriff 15.09.2019].

Lewis, Vivian / Spiro, Lisa / Wang, Xuemao / Cawthorne, Jon E. (2015): *Building Expertise to Support Digital Scholarship: A Global Perspective*. CLIR Publication, Bd. 168, Washington D.C. http://www.clir.org/pubs/reports/pub168/pub168 [letzter Zugriff 15.09.2019].

Maron, Nancy L. / Pickle, Sarah (2014): *Sustaining the Digital Humanities: Host Institution Support beyond the Start-up Phase* https://doi.org/10.18665/sr.22548 [letzter Zugriff 15.09.2019].

Mortiz, Carolyn / Smart, Rachel / Retteen, Aaron / Hunter, Matthew / Stanley, Sarah / Soper, Devin / Vandegriff, Micah (2017): "De-Centering and Recentering Digital Scholarship: A Manifesto" in: *Journal of New Librarianship* 2(2): 101–109 https://doi.org/10.21173/newlibs/3/2 [letzter Zugriff 15.09.2019].

Posner, Miriam (2016): "Here and There: Creating DH Community" in: Gold, Matthew K. / Klein, Lauren F. (eds.): *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press https://dhdebates.gc.cuny.edu/read/untitled/section/c6b8f952-acfd-48b6-82bb-71580c54cad2#ch22 [letzter Zugriff 15.09.2019].

Prescott, Andrew (2016): "Beyond the Digital Humanities Center: The Administrative Landscapes of the Digital Humanities" in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A New Companion to Digital Humanities*, Chichester (West Sussex): Wiley Blackwell 461–475.

Rehbein, Malte / Sahle, Patrick (2013): "Digital Humanities lehren und lernen: Modelle, Strategien, Erwartungen" in: Neuroth, Heike / Lossau, Norbert/ Rapp, Andrea (eds.): *Evolution der Informationsinfrastruktur: Kooperation zwischen Bibliothek und Wissenschaft*, Glückstadt: Verlag Werner Hülsbusch 209–228.

RfII (2016): *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*, Göttingen www.rfii.de/?wpdmdl=1998 [letzter Zugriff 15.09.2019].

Roeder, Torsten / Söring, Sibylle / Dogunke, Swantje / Elwert, Frederik / Wübbena, Thorsten / Lordick, Harald / Cremer, Fabian / Klammt, Anne (2019): "Digital Humanities "from Scratch": Ein Panel-Bericht zur DHd 2019 #DHfromScratch #dhd2019", Blogbeitrag DHd-Blog, 03.07.2019 https://dhd-blog.org/?p=11804 [letzter Zugriff 15.09.2019].

Sahle, Patrick (2015): "Digital Humanities? Gibt's doch gar nicht!" in: *Zeitschrift für digitale Geisteswissenschaften* Sonderband 1 http://dx.doi.org/10.17175/sb001_004 [letzter Zugriff 15.09.2019].

Thaller, Manfred (2017): „Geschichte der Digital Humanities" in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte(eds.): *Digital Humanities: Eine Einführung*, Stuttgart: J. B. Metzler 3–12.

Underwood, Ted (2019): "Dear Humanists: Fear Not the Digital Revolution" in: *The Chronicle of Higher Education*, 27. März 2019 https://www.chronicle.com/article/Dear-Humanists-Fear-Not-the/245987/ [letzter Zugriff 15.09.2019].

Wissenschaftsrat (2011): *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften* (Drs. 10465-11). 8. Januar 2011 h [letzter Zugriff 15.09.2019].

Wuttke, Ulrike (2019), *Infrastrukturelle Erfolgsfaktoren für einen Digital Humanities-Schwerpunkt an deutschen Universitäten*, Masterarbeit TH Köln URN: urn:nbn:de:hbz:79pbc-opus-13968.

Welche Beziehungen steuern das Briefkorrespondenznetzwerk der Reformatoren? Eine Netzwerkanalyse

Roller, Ramona
rroller@ethz.ch
ETH Zürich, Schweiz

Schweitzer, Frank
fschweitzer@ethz.ch
ETH Zürich, Schweiz

Einleitung

Die europäische Reformation war eine gesellschaftliche Erneuerungsbewegung im 16. Jahrhundert, die die Spaltung der katholischen Kirche zur Folge hatte (Kaufmann, 2016; Strohm, 2017). Jene Periode zeichnete sich durch einen regen Briefverkehr zwischen Gelehrten aus, die so miteinander im Austausch blieben. Diese Briefkorrespondenzen lassen sich als Netzwerk darstellen, eine beliebte Modellierungsmethode, um Verbindungsmuster in sozialen Systemen zu repräsentieren (Newman, 2018).

Wir konstruieren ein Korrespondenznetzwerk der Reformatoren aus Briefdaten mit dem Ziel, Aspekte des sozialen Systems der Reformation zu analysieren. Dieses Netzwerk beruht auf 20.000 Briefen, welche zwischen 1510 und 1575 von 2.000 Personen in ganz Europa verschickt und empfangen wurden. Reformatoren werden zu Punkten reduziert (Knoten), welche durch Linien (Kanten) miteinander verbunden sind, wenn sie sich Briefe geschrieben haben. Die Richtung der Kanten identifiziert Sender und Empfänger.

Das Problem dieser Netzwerkkonstruktion ist, dass jene Briefdaten nicht alle Briefe, inklusive Sender und Empfänger, enthalten, die während der Reformation verfasst wurden. Dies kann verschiedene Gründe haben: Lediglich Briefe von „wichtigen“ Akteuren wurden aufbewahrt, Briefe wurden absichtlich vernichtet oder wurden noch nicht digitalisiert. Bei unseren Briefdaten handelt es sich also um eine unrepräsentative Stichprobe von *unvollständigen Daten*. Diese unvollständigen Daten führen zu verzerrten Netzwerktopologien, welche die Interpretation des Netzwerks beeinflussen und somit zu falschen Rückschlüssen auf das reale System führen können (Eliassi-Rad et al., 2019).

Netzwerkrekonstruktion, also die Wiederherstellung des globalen Netzwerks mithilfe des observierten verzerrten Netzwerks, ist ein schwieriges und noch stets ungelöstes Problem. Ein grosser Nachteil bisheriger Lösungsansätze ist, dass sie ausschließlich Eigenschaften des observierten Netzwerks verwenden, um genau jenes zu rekonstruieren (Liben-Nowell & Kleinberg, 2007; Eagle & Lazer, 2009; Wu et al., 2009). Diese eingeschränkte Sichtweise lässt außer Acht, dass das observierte Netzwerk nur eine Form sozialer Interaktion widerspiegelt, nämlich Briefkorrespondenz. In Wirklichkeit bestand das soziale System der Reformation aus vielen verschiedenen Interaktionen und Beziehungen, die sich gegenseitig beeinflussen. Neben Briefkorrespondenzen können Reformatoren zum Beispiel auch durch persönliche Gespräche, Familienbande oder Kollegenschaft miteinander verbunden sein. Wenn wir verstehen, wie verschiedene soziale Beziehungen miteinander zusammenhängen, können wir diese Zusammenhänge eventuell nutzen, um fehlende Briefkorrespondenzen zu rekonstruieren.

Hierzu sind mehrere Zwischenschritte erforderlich: Wir müssen zunächst verstehen, (i) welche Beziehungen wichtig sind, (ii) wie diese mit dem observierten, jedoch unvollständigen Netzwerk zusammenhängen (iii) und wie wir von diesem Zusammenhang auf einen vergleichbaren Zusammenhang im globalen, jedoch unbekannten, Netzwerk schließen können. Diese Zwischenschritte lassen sich mithilfe von Regressionsanalyse und Inferenzstatistik untersuchen. *Regressionsanalyse* ist eine statistische Methode, um den Zusammenhang zwischen Variablen in vorhandenen Daten zu quantifizieren. *Inferenzstatistik* benutzt Wahrscheinlichkeitstheorie, um von jenem ermittelten Zusammenhang in der Stichprobe Rückschlüsse auf die gesamte Population zu ziehen.

In diesem Beitrag veranschaulichen wir den Nutzen von Regressionsanalysen und Inferenzstatistik am Beispiel des Briefkorrespondenznetzwerks der Reformatoren. Im Speziellen sind wir an folgender Frage interessiert: **Welcher Zusammenhang besteht zwischen der geographischen Distanz zwischen Reformatoren-Paaren und der Anzahl Briefe, die sie sich schreiben?**

Wir erklären, warum etablierte Regressionsmodelle nicht auf das Korrespondenznetzwerk der Reformatoren angewandt werden können und stellen eine neue Methode vor, die Netzwerkregression, welche jene Nachteile entschärft. Zuletzt zeigen wir, wie wir die Netzwerkregression für statistische Inferenz nutzen können. Unsere Arbeit zeigt Möglichkeiten auf, Unsicherheiten in unvollständigen Daten zu quantifizieren und jene in die Interpretation der Ergebnisse mit einzubauen.

Datengrundlage und Korrespondenznetzwerk

Unsere Daten beruhen auf den Briefeditionen sieben ausgewählter Reformatoren. Martin Luther (ProQuest LLC, 2019), Philipp Melanchthon (HAW, 2019), Ulrich Zwingli (Moser, 2019), Heinrich Bullinger (Bodenmann, 2019), Martin Bucer (Simon & Friedrich, 2018), Andreas Karlstadt (Kaufmann, 2012) und Oswald Myconius (Wallraff, 2016). Unser ausschlaggebendes Auswahlkriterium bestand darin, dass die Briefdaten öffentlich digital zugänglich sind, sodass wir sie mit einem Web-Crawler aus den entsprechenden Datenbanken herausfiltern konnten. Wir verwenden den Begriff „Reformator“ als Sammelbegriff für alle Sender und Empfänger in unserem Datensatz, obwohl viele jener Personen keine protestantischen Theologen waren. „Reformator“ umfasst somit auch Adlige, Humanisten, Katholiken und andere gesellschaftliche Gruppen.

Neben den Sendern und Empfängern der Briefen beinhalten unsere Daten das Sendedatum, sowie Sende- und Empfangsort. In der Datenvorverarbeitung haben wir Synonyme in Personen- und Ortsnamen eindeutigen Entitäten zugeordnet, sowie das jeweilige taggenaue Sendedatum aus den Ausgangsdaten abgeleitet.

Abbildung 1 zeigt das Briefkorrespondenznetzwerk, welches wir aus obigen Daten erstellen. Wir stellen fest, dass die sieben ausgewählten Reformatoren als Sternzentren mit vielen Briefverbindungen erscheinen, wohingegen Knoten in der Netzwerkperipherie kaum Briefverbindungen aufweisen. Hier zeigt sich bereits, wie unsere Datenauswahl die Netzwerktopologie beeinflusst. Demnach sind die dazugehörigen deskriptiven Statistiken der Netzwerktopologie nicht aussagekräftig für die Reformationszeit. Abbildung 2 listet jene Statistiken auf, um das unvollständige Korrespondenznetzwerk besser zu beschreiben.

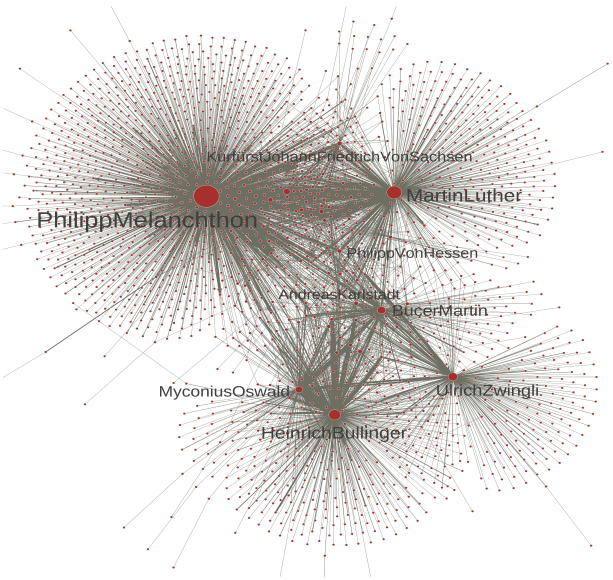


Abbildung 1: Unvollständiges Briefkorrespondenznetzwerk der Reformatoren. Die Knotengröße veranschaulicht die Anzahl gesendeter und empfangener Briefe. Die Kantendicke veranschaulicht die Anzahl gesendeter Briefe.

	Topologisches Netzwerkmass	Definition	Wert
Netzwerk-spezifisch	Größe	Anzahl Knoten (N) und Kanten (E)	$N = 2.096$, $E = 19.472$
	Dichte	Anteil möglicher Kanten, die im observierten Netzwerk existieren	$2,385 * 10^{-3}$
	Globaler Clusterkoeffizient	Ausmaß in dem sich Knoten gruppieren	$2,446 * 10^{-3}$
	Durchschnittl. Pfadlänge	durchschnittl. Anzahl von Kanten auf den kürzesten Verbindungen zwischen allen Knotenpaaren.	$2,482$, $\sigma = 5,967 * 10^{-1}$
	Durchmesser	Verbindung mit den meisten Kanten unter den kürzesten Verbindungen zwischen allen Knotenpaaren	6
Knoten-spezifisch	Degree	Anzahl ein- und ausgehender Kanten pro Knoten (k)	$\langle k \rangle = 9,992$, $\sigma_k = 159,613$
	In-degree	Anzahl empfangener Briefe pro Knoten (k_{in})	$\langle k_{in} \rangle = 4,996$, $\sigma_{k_{in}} = 49,232$
	Out-degree	Anzahl versendeter Briefe pro Knoten (k_{out})	$\langle k_{out} \rangle = 4,996$, $\sigma_{k_{out}} = 121,420$
	Betweenness	Wie stark wirkt Knoten als Vermittler zwischen anderen Knoten? (b_{norm})	$\langle b_{norm} \rangle = 1,920 * 10^{-4}$, $\sigma_{b_{norm}} = 6,080 * 10^{-3}$
	Closeness	Wie nah ist Knoten anderen Knoten in Bezug auf Anzahl zu passierender Kanten? (c_{norm})	$\langle c_{norm} \rangle = 3,232 * 10^{-2}$, $\sigma_{c_{norm}} = 3.736 * 10^{-3}$

Abbildung 2: Deskriptive topologische Netzwerkmaße im unvollständigen Briefkorrespondenznetzwerk der Reformatoren. Knoten-spezifische Maße werde als Durschnittswert (< >) zusammen mit der Standardabweichung (σ) angegeben. Das „norm“ Kürzel, gibt normalisierte Werte an (Minimum=0, Maximum=1). Wir weisen eindrücklich darauf hin, dass diese Netzwerkmaße lediglich eine **Beschreibung** des Briefkorrespondenznetzwerk darstellen. Aufgrund unserer unvollständigen Daten ist die Netzwerktopologie verzerrt und eine Interpretation der resultierenden Netzwerkmaße in Bezug auf die Reformation daher nicht möglich.

Soziale Beziehungen auswählen

Gemäß Schritt (i) (siehe Einleitung), müssen zunächst einige Beziehungstypen für die Analyse ausgewählt werden. Die Auswahl erfolgt auf Grundlage von vorhandenen Theorien und Datenverfügbarkeit.

Wir konzentrieren uns hier auf die geographische Distanz zwischen Reformatoren, eine Beziehung, die bereits zur Rekonstruktion von antiken Handelsnetzwerken genutzt wurde (Amati et al., 2019). Da unsere Korrespondenzdaten den Sende- und den Empfangsort der jeweiligen Briefe enthalten, benutzen wir die GoogleMaps API, um die Gehstrecke zwischen Sender und Empfänger per Brief zu berechnen. Für unsere Analyse ist es unbedeutend, dass die ermittelte moderne Gehstrecke nicht mit historischen Geh- oder Poststrecken übereinstimmt. Wichtig ist, dass die Größenordnungen vergleichbar sind. Unsere Methode gewährleistet zum Beispiel, dass die Strecke über die Alpen heute wie damals um den gleichen Faktor länger ist als die Strecke zwischen Wittenberg und Heidelberg.

Wir analysieren zwei mögliche Hauptszenarien wie geographische Distanz die Anzahl der Briefe zwischen Reformatoren beeinflusst: Auf der einen Seite schreiben sich Reformatoren

weniger Briefe, je weiter sie auseinander wohnen, da die Kosten steigen (z.B. teure Boten), um jene zu verschicken (Kostenszenario). Auf der anderen Seite schreiben sich Reformatoren mehr Briefe, je weiter sie auseinander wohnen, da Briefe das einzige Kommunikationsmittel waren, um lange Abstände zu überbrücken (Zweckmäßigkeitsszenario). Über kurze Abstände ist es demnach praktischer persönliche Gespräche zu führen, anstatt Briefe zu schreiben.

Neben der geographischen Distanz testen wir zwei Kontrollfaktoren. Dies sind weitere soziale Beziehungen, an deren Effekt auf die Anzahl Briefe wir nicht per se interessiert sind, die jene Anzahl aber dennoch beeinflussen und deshalb mitberücksichtigt werden müssen. Der erste Kontrollfaktor ist Reziprozität, ein Grundprinzip des menschlichen Handels, wobei eine Person die Handlung ihres Gegenübers erwidert.

Der zweite Kontrollfaktor ist religiöse Homophilie. Homophilie beschreibt ein soziologisches Prinzip, wonach Menschen verstärkt interagieren, wenn sie sich ähneln (z.B. gleicher Bildungsstatus). Religiöse Homophilie beschreibt die Annahme, dass Reformatoren, die dieselbe protestantische Strömung vertreten, sich mehr Briefe schreiben, als wenn sie unterschiedliche Strömungen unterstützen.

Etablierte Modelle und ihre Schwächen

Regression ist ein statistisches Modell, welches den Zusammenhang und dessen Stärke zwischen verschiedenen Variablen mithilfe einer mathematischen Funktion beschreibt. Auf das Briefkorrespondenznetzwerk bezogen, können wir zum Beispiel ermitteln, wie die geographische Distanz, Reziprozität und religiöse Homophilie die Anzahl Briefe zwischen Sender- und Empfänger-Paaren beeinflussen. Abbildung 3 veranschaulicht eine einfache lineare Regression.

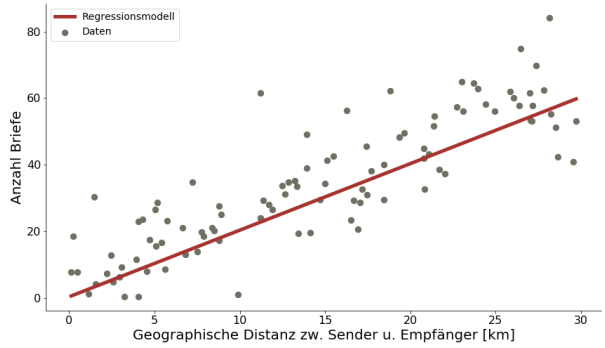


Abbildung 3: Schematisches Beispiel einer einfachen linearen Regression. Die Abhängigkeit zwischen geographischer Distanz (x-Achse) und der Anzahl Briefe per Sender-Empfänger-Paar (y-Achse) wird durch die lineare Funktion $y = \beta_0 + \beta_1 * x_1 + \varepsilon$ beschrieben. β_0 ist der y-Achsenabschnitt. β_1 ist die Steigung der Geraden und quantifiziert die Stärke des ermittelten Zusammenhangs zwischen Distanz und Briefanzahl. ε ist der Fehlerterm, der unbeobachtete Zufallsvariablen repräsentiert. Für jede weitere getestete soziale Beziehung (z.B. Homophilie), fügen wir eine Dimension in der Abbildung hinzu und ergänzen die obere Formel mit einem weiteren $\beta * x$. Das heißt, wir berechnen eine Gerade im mehrdimensionalen Raum.

Der ermittelte Zusammenhang zwischen den Variablen ist nicht absolut, sondern abhängig von Fehlern in der Datenmessung und Bestimmung der Parameterwerte (β‘s). Diese Unsicherheiten lassen sich in der Regression modellieren und können so in die Interpretation der Ergebnisse mit einbezogen werden.

Allerdings setzt die Regression voraus, dass Observationen unabhängig sind, eine Annahme, die in Netzwerken ungültig ist, da die Postion von Kanten durch die restlichen Kanten eingeschränkt ist. Klassische generative Netzwerkmodelle, wie das *Exponential Random Graph Model,* machen jene Annahme zwar nicht, sind aber sehr rechenintensiv und daher für Netzwerke mit mehreren hundert Knoten nicht geeignet (An, 2016). Für das Briefkorrespondenznetzwerk sind wir daher auf ein alternatives Modell angewiesen, die Netzwerkregression.

Netzwerkregression

Konzept und Definition

Ähnlich wie bei der klassischen Regression ermittelt die Netzwerkregression den Zusammenhang und dessen Stärke zwischen einer abhängigen und einer oder mehrerer unabhängiger Variablen (Casiraghi, 2017). Allerdings sind nun die abhängige und die unabhängigen Variablen keine einfachen Zahlen mehr (z.B. Briefanzahl, geografische Distanz in km), sondern Netzwerktopologien. Ausgangspunkt für die Netzwerkregression ist ein mehrlagiges Netzwerk, dessen unterste Lage ein Interaktionsnetzwerk darstellt (abhängige Variable) und jede weitere Lage jeweils ein soziales Beziehungsnetzwerk darstellt (unabhängige Variablen) (siehe Abbildung 4).

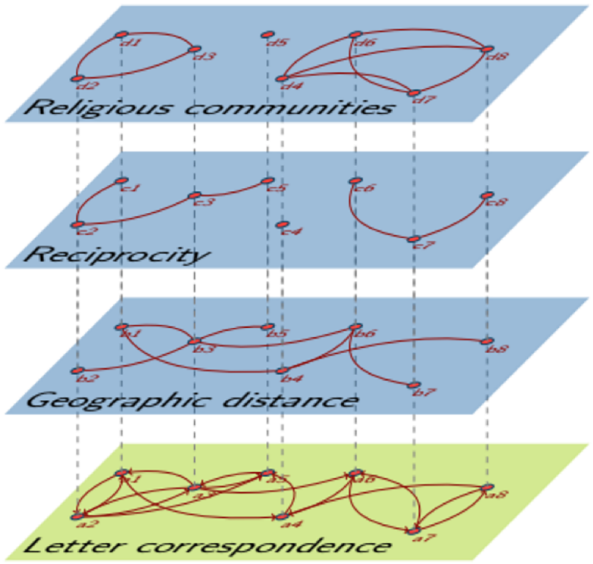


Abbildung 4: Mehrlagiges Netzwerk als Ausgangspunkt für die Netzwerkregression. Die unterste Lage (grün) repräsentiert das Briefkorrespondenznetzwerk der Reformatoren. Jede weitere blaue Lage repräsentiert jeweils eine soziale Beziehung zwischen den Reformatoren: geographische Distanz, Reziprozität und religiöse Homophilie. Alle Lagen beinhalten dieselben Knoten (Reformatoren), welche aber unterschiedlich miteinander verbunden sind, weil die Kanten verschiedene soziale Relationen darstellen. Ziel der Netzwerkregression ist es einen Zusammenhang zwischen den blauen und der grünen Lage herzustellen, um die Topologie des Briefkorrespondenznetzwerk durch soziale Beziehungen zu erklären.

Jede Netzwerklage, die eine unabhängige Variable repräsentiert, wird durch eine Matrix (**R**) dargestellt, welche die soziale Beziehung für jedes Reformatoren-Paar quantifiziert. Ähnlich wie in der klassischen Regression quantifiziert der Parameter β die Stärke des Zusammenhangs zwischen einer **R** Matrix und dem Briefkorrespondenznetzwerk.

Mithilfe der **R** Matrizen wird nun die Wahrscheinlichkeit per Reformatoren-Paar berechnet, dass dieses durch eine Kante verbunden ist. Diese Berechnung basiert auf einem statistischen Modell dem *generaliserten hypergeometrischen Ensemble* (Casiraghi & Nanumyan, 2018). Mit diesen Verbindungswahrscheinlichkeiten werden neue synthetische Netzwerke generiert, deren Topologien wissentlich durch die getesteten sozialen Beziehungen zustande kommen. Falls sich das observierte Netzwerk sehr stark von den synthetischen unterscheidet, sind die getesteten sozialen Beziehungen keine gute Erklärungen für die Topologie des observierten Netzwerks. Umgekehrt schon.

Anwendung im Briefkorrespondenznetzwerk

Wir konstruieren zunächst die jeweiligen **R** Matrizen für die sozialen Beziehungen geographische Distanz, Reziprozität und religiöse Homophilie. Im Bezug auf geographische Distanz möchten wir herausfinden, ob die Anzahl der Briefe einem Kostenszenario (größere Distanz führt zu weniger Briefen) oder einem Zweckmäßigkeitsszenario (größere Distanz führt zu mehr Briefen) folgt. Für dieses Ziel teilen wir den Effekt der geographischen Distanz auf zwei **R** Matrizen auf, die jeweils eine Exponentialfunktion der geographischen Distanz darstellen. $\mathbf{R}^{(1)}$ hat die lineare Distanz per Sender-Empfänger-Paar im Exponenten und $\mathbf{R}^{(2)}$ die quadratische.

$$\mathbf{R}_{ij}^{(1)} = e^{dist_{ij}} \quad \text{und} \quad \mathbf{R}_{ij}^{(2)} = e^{dist_{ij}^2}$$

wobei *dist* die geografische Distanz zwischen zwei Reformatoren *i* und *j* darstellt.

Abbildung 5 zeigt ein schematisches Beispiel der $\mathbf{R}^{(1)}$ Matrix.

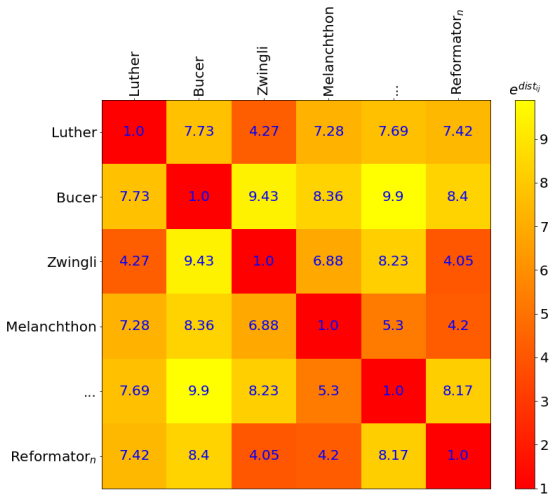


Abbildung 5: Schematisches Beispiel der linearen geografischen Distanzmatrix $\mathbf{R}^{(1)}$. Reihen repräsentieren die Briefsender und Spalten die Empfänger, also jeweils die Knoten (Reformatoren) im Netzwerk. Für jedes Sender-Empfänger-Paar berechnen wir $e^{\text{Distanz [km]}}$. Wenn Luther und Melanchthon zum Beispiel 2 km von einander entfernt wären, würden wir e^2=7,28 in die entsprechende Matrixzelle eintragen. Die Werte auf der Diagonalen sind 1, da Sender und Empfänger hier die gleiche Person sind, die sich räumlich nicht aufteilen kann (Distanz = 0 km, e^0=1). Reformator$_n$ ist der letzte Reformator im Datensatz.

Die **R** Matrizen für Reziprozität ($\mathbf{R}^{(3)}$) und religöse Homophilie ($\mathbf{R}^{(4)}$) werden jeweils mit numerischen Ersatzwerten und der change statistic konstruiert. Letztere ist ein Standardmaß für Netzwerkmuster in generativen Netzwerkmodellen (Snijders et al., 2006). Wir aggregieren die vier **R** Matrizen, um für jedes Sender-Empfänger Paar die Chance zu berechnen, dass es sich Briefe schreibt. In dieser Berechnung ermitteln wir den Wert für jeweils ein β pro **R** Matrix. Die β's beschreiben, wie sehr die entsprechende soziale Beziehung die Anzahl geschriebener Briefe beeinflusst. Abhängig vom Vorzeichen der ermittelnden β's, werden vier Verläufe unterschieden, wie sich die geographische Distanz auf jene Chance auswirkt (siehe Abbildung 6).

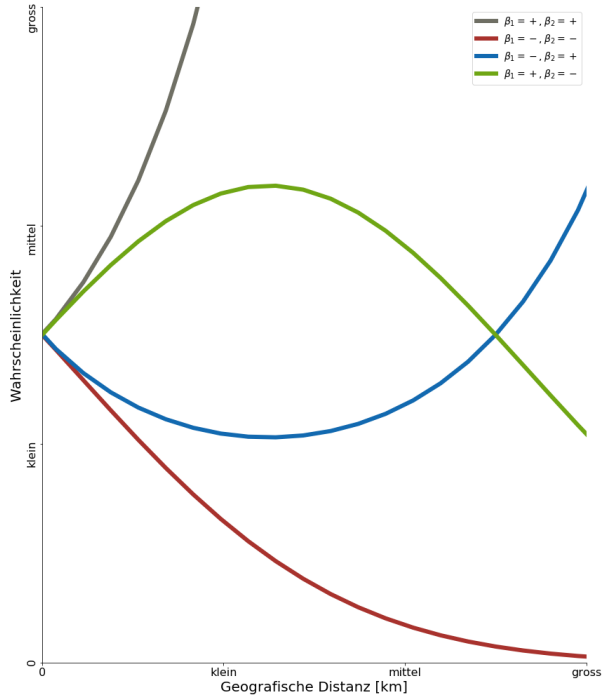


Abbildung 6: Schematische Repräsentation der Auswirkung von geographischer Distanz (x-Achse) auf die Wahrscheinlichkeit zum Briefeschreiben (y-Achse) in Abhängigkeit des Vorzeichens der ermittelten β Koeffizienten für $\mathbf{R}^{(1)}$ und $\mathbf{R}^{(2)}$. Sind beide β's positiv, steigt die Chance, dass Reformatoren sich Briefe schreiben, mit wachsender geographischer Distanz. Dies ist das absolute Zweckmäßigkeitsszenario). Sind beide β's negativ, sinkt die Chance, dass Reformatoren sich Briefe schreiben, mit wachsender geographischer Distanz. Dies ist das absolute Kostenszenario. Ist das lineare β negativ und das quadratische β positiv, ist die Chance, dass Reformatoren sich Briefe schreiben, hoch für sehr kurze und sehr lange Distanzen. Dies ist ein Kosten- oder Zweckmäßigkeitsszenario). Ist das lineare β positiv und das quadratische β negativ, ist die Chance, dass Reformatoren sich Briefe schreiben, hoch für mittlere geographische Distanzen. Das Kosten- und das Zweckmäßigkeitsszenario sind in Balance.

Abbildung 7 zeigt die Ergebnisse der Netzwerkregression. Die Vorzeichen der β Koeffizienten für geographische Distanz geben an, dass das Zweckmässigkeitsszenario bei großen geographischen Distanzen dominiert und das Kostenszenario bei kurzen Distanzen. Es ist also wahrscheinlicher, dass Reformatoren sich Briefe über sehr kurze und sehr lange Distanzen schreiben als über mittlere.

	full model
Distance	
Linear distance	$-3.354\ (0.176)^{***}$
Quadratic distance	$5.032\ (0.388)^{***}$
Controls	
Reciprocity	$0.461\ (0.004)^{***}$
Religious homophily	$0.276\ (0.016)^{***}$
AIC	33989.210
McFadden $pseudo - R^2$	0.224

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Abbildung 7: Ergebnis der Netzwerkregression. Der negative und positive Koeffizient für geographische Distanz zeigen, dass Reformatoren sich vor allem über sehr kurze und sehr lange Distanzen Briefe schreiben. Die positiven Koeffizienten der Kontrollfaktoren weisen darauf hin, dass Reziprozität und religiöse Homophilie die Anzahl geschriebener Briefe positiv beeinflussen.

Die Koeffizienten für Reziprozität und religiöse Homophilie zeigen jeweils, dass beide Faktoren einen positiven Effekt auf die Anzahl versendeter Briefe haben. Demnach steigt die Chance, dass ein Reformator A einem anderen Reformator B Briefe schreibt, falls B bereits Briefe an A geschrieben hat und beide Reformatoren dieselbe religiöse Strömung vertreten. Am absoluten Wert der β Koeffizienten sehen wir, dass Reziprozität einen größeren Effekt hat als religiöse Homophilie.

Diese Ergebnisse sind an sich nicht überraschend, könnten jedoch sehr nützlich für die Netzwerkrekonstruktion sein. Aus den Werten für Distanz, Reziprozität und Homophilie können wir die oben beschriebenen Wahrscheinlichkeit berechnen, dass Reformatoren sich Briefe geschrieben haben. Falls unsere Daten keine Briefverbindung zwischen zwei Reformatoren aufweisen, jene berechnete Wahrscheinlichkeit aber groß ist, können wir annehmen, dass jene Briefe in unseren Daten schlicht fehlen. Unter der Annahme, dass Briefe zur Verbreitung reformatorischen Gedankenguts genutzt wurden, können wir anhand der β-Werte bestimmen, welche sozialen Beziehungen die Etablierung der Reformation vorangetrieben haben. Dies erlaubt uns eine neue Perspektive auf die Reformation, welche nicht bestimmte Hauptfiguren (z.B. Luther als Wegbereiter der Reformation) oder einzelne Großereignisse (z.B. Erfindung des Buchdrucks, Augsburger Religionsfrieden) als Antriebsfaktoren für die Reformation in den Vordergrund rückt, sondern sich auf die lokalen Interaktions- und Beziehungsstrukturen der Leute von damals konzentriert.

Unabhängig von der Reformation lassen sich mithilfe der Netzwerkregression Interaktionsnetzwerke jeglicher Art beschreiben. Beispiele für weitere Interaktionsnetzwerke umfassen den Handel mit Gütern, finanzielle Transaktionen und persönliche Gespräche. Wenn wir zusätzlich zu jenem Interaktionsnetzwerk auch Informationen zu den sozialen Beziehungen zwischen den Knoten im Netzwerk haben (z.B. Verwandtschaft, Heirat, Arbeitsverhältnis und Homophilie), können wir mithilfe der Netzwerkregression ermitteln wie sehr jene Beziehungen das Interaktionsnetzwerk beeinflussen. Die Auswahl des Interaktionsnetzwerks und der sozialen Beziehungen hängt von theoretischen Überlegungen und der Datengrundlage ab. Welche Interaktionen und Beziehungen sind für den Forschungsstand von Belang? Sind die gewünschten Daten zugänglich?

Fazit

Wir beschreiben den Zusammenhang zwischen sozialen Beziehungen und einem Interaktionsnetzwerk mithilfe einer Netzwerkregression. Am Beispiel des Briefkorrespondenznetzwerks der Reformatoren haben wir aufgezeigt, dass sehr kurze und sehr lange Distanzen, Reziprozität, sowie Zugehörigkeit zur selben religiösen Strömung die Chance vergrößern, dass Reformatoren sich Briefe schreiben. Diese Ergebnisse geben mögliche Hinweise, wie externe Faktoren, soziale Normen und Gruppendynamiken Briefkorrespondenzen beeinflussen. Sie sind erste wichtige Bausteine, um das globale unbekannte Briefkorrespondenznetzwerk zu rekonstruieren.

Bibliographie

An, Weihua (2016): „Fitting ERGMs on big networks.“, in: Social Science Research 59: 107-119.

Amati, Viviana / Mol, Angus / Shafie, Termeh / Hofman, Corinne / Brandes, Ulrik (2019): „A Framework for Reconstructing Archaeological Networks Using Exponential Random Graph Models.“, in: *Journal of Archaeological Method and Theory* 29: 1-28.

Bodenmann, Reinhard (2019): „Bullinger-Briefwechsel“, unter http://www.arpa-docs.ch/SedServer/SedWEB.cgi?fld_41a=&fld_30b=&fld_41c=&fld_30c=&fld_41e=&search=&range=&Alias=Briefe&Lng=0&first=0&session=0&awidth=1440&aheight=769&PrjName=Bullinger+-+Briefwechsel (abgerufen am 12.12.2019).

Casiraghi, Giona / Nanumyan, Vahan (2018): „Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem“, *arXiv:1810.06495*.

Casiraghi, Giona (2017): „Multiplex Network Regression: How do relations drive interactions?“ *arXiv:1702.02048*.

Eagle, Nathan / Lazer, David (2009): „Inferring friendship network structure by using mobile phonedata“, in *PNAS* 106 *(36)*, 15274-15278.

Eliassi-Rad, Tina / Caceres, Rajmonda. / LaRock, Timothy (2019): „Incompleteness in Networks: Biases, Skewed Results, and Some Solutions“, in *KDD'19 Proceedings*.

HAW-Forschungsstelle Melanchthon-Briefwechsel (2019): „Melanchthon Briefwechsel – Regesten online“, unter https://www.haw.uni-heidelberg.de/forschung/forschungsstellen/melanchthon/mbw-online.de.html (abgerufen am 12.12.2019).

Liben-Nowell, David / Kleinberg, Jon (2007): „The Link Prediction Problem for Social Net- works“, in *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM)*.

Kaufmann, Thomas. (2017): „Erlöste und Verdammte - Eine Geschichte der Reformation (2. Ed.)“, C.H.Beck.

Kaufmann, Thomas (2012): *„Kritische Gesamtausgabe der Schriften und Briefe Andreas Bodensteins von Karlstadt, Teil I (1507-1518)*. Wolfenbüttel 2012. *(Editiones Electronicae Guelferbytanae),* unter http://diglib.hab.de/edoc/ed000216/start.htm (abgerufen am 12.12.2019).

Moser, Christian (2019): „Huldrych Zwingli Briefe: Digitale Texte“, unter http://www.irg.uzh.ch/static/zwingli-briefe/?n=Main.HomePage (abgerufen am 12.12.2019).

Newman, Mark. (2018): „Networks (2. Ed.)“, Oxford University Press.

ProQuest LLC. (2019): „Luthers Werke im WWW: Weimarer Ausgabe“, unter (abgerufen am 12.12.2019).

Simon, Wolfgang / Friedrich, Reinhold (2018): „Briefkorrespondenz von Martin Bucer, unter https://augustana.de/forschung-lehre/kirchengeschichte/bucer-forschungsstelle.html (abgerufen am 12.12.2019).

Snijders, Tom / Pattison, Philippa / Robins, Garry / Handcock, Mark (2006): „New specifications for exponential random graph models“, in *Sociological methodology* 36 *(1)*, 99-153.

Strohm, Christoph (2017): „Theologenbriefwechsel im Südwesten des Reichs in der Frühen Neuzeit (1550-1620)“ , Universitätsverlag WINTER Heidelberg.

Wallraff, Martin (2016): „Erschliessung des Briefwechsels von Oswald Myconius“, unter https://myconius.unibas.ch/impressum.html (abgerufen am 12.12.2019).

Wu, Yun-Jhong / Levina, Elizaveta / Zhu, Ji (2009): „Link prediction for egocentrically sampled networks“, *arXiv:1803.04084.*

Wie wir lesen könnten. StreamreaderPS 0.1

Sahle, Patrick
sahle@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Ausgangspunkte

Man sagt, Sprache sei ein lineares Ereignis in der Zeit. Schrift sei die Fixierung gesprochener Sprache. Text sei essentiell ein Strom sprachlicher Einheiten. Davon halte ich: nichts. Insbesondere ist die Idee, Text sei (letztlich nur) eine Abfolge von Zeichen, Wörtern oder Sätzen, ganz unsinnig. Auch wenn die analytisch ausgerichtete Praxis der Computerlinguistik und großer Teile der Computerphilologie sich aus der höchst produktiven Reduktion des (recodierten) Textes auf einen "stream of tokens" speist, bleibt die Vorstellung eines linearen Textes aus einer allgemeinen, medienhistorisch bewussten Sicht auf Text, Textgenres, Textualität und Textmedialität heraus ganz arm, um nicht zu sagen: schlicht falsch. Denn offensichtlich beruht der große Erfolg der etabliertesten Textmedien (z.B.: das Buch) darauf, dass die Linearität der (gesprochenen) Sprache durch eine Zweidimensionalität der Schriftsprache auf der Schreibfläche *ersetzt* ist. Text (wie wir ihn kennen) ist nicht so sehr Fixation von gesprochener Sprache, sondern bildet ein eigenes, autonomes Ausdruckssystem, das auf komplexe Medialisierung und auf eine primär humanoide Rezeption, die wir Lesen nennen, ausgerichtet ist. Unser Lesen aber ist Sehen. Ist das visuelle Erfassen von flächigen Bildern, nämlich Wörtern, auf strukturierten Flächen: dem Layout der Seite. Diese Konfiguration des Lesens als Erkennung komplexer Muster und Strukturen ist aber historisch bedingt, ist das Ergebnis einer bestimmten Lesesozialisation und die Konsequenz aus der Anwendung bestimmter Schrift- und Lese*medien*, die von der beschriebenen Fläche ausgehen. Sie ist das Ergebnis der Nutzung bestimmter Technologien.

Was jedoch würde passieren, wenn wir die (naive) Ursprungsidee des Textes als Zeichenstrom wieder aufnähmen und sie mit aktuellen technischen Möglichkeiten verbänden?

Der StreamReader

Es ist ganz einfach: Digitale Technologien medialisieren Daten ad hoc und erlauben eine dynamische, interaktiv zu kontrollierende Darstellung von Inhalten. Natürlich könnten Texte auch als "laufende" Schrift dargestellt werden, die sich als einzelne Zeile auf einer Anzeigefläche von rechts nach links bewegt. Als Strom von Zeichen, die wie gehabt als Wörter und Sätze gelesen werden können. Als Leser möchte man dann einige erste Steuerungsfunktionen haben: Start / Pause, Schneller, Langsamer, Zurück zum Absatzanfang.

Abbildung 1: Der Streamreader, Papiermodell für eine klassische Bildschirmanwendung. Das Fließen der Textzeile muss man sich hinzudenken, weil auch dieser Text (das vorliegende Abstract) im Rahmen des Drucks gefangen ist.

Darstellung und Steuerung hängen im Detail von der gewählten technischen Lösung und medialen Umgebung ab. Einige denkbare Szenarien seien hier angedeutet: (1.) In einem Webbrowser steuert man Bedienungselemente mit der Maus oder Tastatur um den Textlauf zu regeln. (2.) Bei einem Smartphone kippt man das Gerät nach rechts oder links, um die Geschwindigkeit des Stroms zu kontrollieren. Man kippt nach vorne, um zu stoppen. (3.) Bei einer VR-Brille dreht man den Kopf nach rechts um die fließende Textzeile, die gewissermaßen einen halbkreisförmigen Horizont bildet, zu beschleunigen - und nach links, um ihr hinterherzusehen und sie zu verlangsamen. Mit einem längeren Augenzwinkern würde man pausieren. (4.) Ein kleiner Projektor wirft laufende Schrift an die Zimmerdecke, wenn man im Bett liegt. Ein Steuerungsgerät hält man in der Hand. Oder die Zimmerdecke schaut zurück: Beschleunigung, wenn Kopf oder Augen sich nach rechts wenden. Verlangsamung beim Blick nach Links. Die Anwendung wird geschlossen, wenn die Augen lange geschlossen bleiben.

Das sind naheliegende Anwendungsszenarien mit grundlegenden Funktionen. Schnell fragt man nach weiteren Möglichkeiten, die sich aus den jeweiligen technischen Bedingungen ergeben oder eine Übersetzung aus den gewohnten Funktionalitäten etablierter Textmedien sind. Inhaltsverzeichnisse und Text-Makro-Strukturen wie Abschnitte, Absätze oder Seiteneinheiten sind in kompakte Visualisierungen umzusetzen, die für Überblick, Orientierung und gezieltes Einspringen in den Text sorgen. Eine Lesezeichenfunktion erlaubt das Weiterlesen nach längerer Pause an entsprechender Stelle. Illustrationen, Fußnoten und andere "Textfeatures", die sich die Zweidimensionalität der Schriftseite zunutze machen, erfordern angepasste Verhaltensweisen in der Stromdarstellung. Hinzu kommen weitere Text-Ausdrucksmittel. Mikrotypografische Phänomene wie Fettdruck, Kursivierung, Höher- bzw. Tieferstellung, Farbe, Font, Schriftgröße können erhalten bleiben, müssen aber auf ihre Funktionalität und veränderte Effekte hin überprüft werden. Andere Gestaltungselemente, wie Zeilenumbrüche oder vertikale Abstände verlangen nach ganz neuen Ausdrucksformen. Das Spacing innerhalb und zwischen Wörtern und Sätzen müsste neu bedacht werden. Dies aber sind nur einige erste Hinweise. Der schriftsprachliche, typografische Ausdrucksraum enthält noch weitere Phänomene, die zu remodellieren wären.

Prototypenentwicklung

Dies ist kein Bericht zu einem laufenden oder geplanten Projekt. Es ist der Versuch, einen vielleicht neuen, recht allgemeinen Ansatz auf der fachlich einschlägigen Konferenz vorzustellen und zu diskutieren. In den vergangenen Jahren habe ich mit verschiedenen Kolleginnen und Mitarbeitern begonnen, die technischen Möglichkeiten eines StreamReaders im Rahmen spielerischer Ansätze, ohne jede projektförmige Organisation oder Finanzierung zu bedenken. Dabei ergab sich zunächst der überraschende Befund, dass die gängigen, besonders niederschwelligen Standardtechnologien, wie Web Browser, HTML (mit dem uralten marquee-Element oder den neueren HTML5-Canvas-Möglichkeiten), CSS, Javascript, SVG, VR-Programmbibliotheken die intendierte Anwendung gar nicht gut unterstützen. Bereits hier lässt sich mit einigem technologiekritischen Gewinn herausarbeiten, wie weit unsere medialen Grundvorstellungen von Text als einem an Statik, an hierarchischer Grundstruktur und an der begrenzten Fläche als Präsentationsraum orientierten Informationsobjekt, sich in die Grundkonzepte der Technologien eingeschrieben haben und ein fundamental abweichendes Denken und entsprechende Softwarelösungen behindern. Zu den zuletzt im Rahmen einer Qualifikationsarbeit (Drach 2019) weiter verfolgten Ansätzen gehörte schließlich ein Rückgriff auf primitivste HTML-Browser-Features, die immerhin das Testen einiger Grundfunktionen des StreamReaders ermöglichen und mit ersten Anwendungsreflexionen Anstöße für die weitere Konzeptentwicklung geben können. Sviatoslav Drach (2020) wird diesen Prototyp auf einem Poster zur Konferenz vorstellen. Gemeinsam machen wir uns hier dialogisch ein niederschwelliges "critical prototyping" zunutze, bei dem die Entwicklung von Software zugleich Anstöße für die weitere konzeptionelle Entwicklung und die theoretische Auseinandersetzung mit dem Gegenstand ist.

Worum es hier (nicht) geht

Disclaimer wegen der absehbaren Reflexe: Es geht hier *nicht* darum, eine neue, bessere Lesetechnologie zu entwickeln, die das gedruckte Buch oder den digital flächig dargestellten Text ablösen könnte. Es ist klar und offensichtlich, dass diese Form der Textdarstellung und Textrezeption erhebliche Schwierigkeiten und Nachteile mit sich bringen, die vielleicht nicht nur unserem gelernten Leseverhalten geschuldet sind, sondern auch auf anthropologische Konstanten (z.B. unserer visuellen Wahrnehmung) zurückgehen. Die vielen Vorteile flächiger Textdarstellung brauchen innerhalb dieses Ansatzes auch nicht erneut diskutiert zu werden. Sie werden durch die Demonstration abweichender Formen ohnehin augenfällig und sind damit ein Abfallprodukt dieser Form praktischer Auseinandersetzung. Es sollen ausdrücklich nicht Textdarstellung oder Lesen *verbessert* oder optimiert werden, sondern mög-

liche Darstellungsoptionen und bisherige Lesepraktiken *kontrastiert* werden - um neue Einblicke in den Phänomenbereich Textualität und alternative Leseformen zu gewinnen. Es geht auch nicht um die Effizienz des Lesens (z.B. durch Schnelleseverfahren) oder gar der Informationsaufnahme aus Texten. Im Gegenteil: Der Diskurs um das Lesen digitaler Texte (das es im Übrigen nicht gibt; es gibt nur das Lesen verschiedener digital angetriebener Textmedien) dreht sich häufig um den Verlust an Konzentration und tiefergehender Beschäftigung mit Texten. Dagegen ist es einer der vielen Aspekte des StreamReaders, dem (erzwungenen?) langsamen und damit kontemplativen Lesen vielleicht eine neue Möglichkeit zu geben. Insofern schließt sich der StreamReader auch nicht historisch, technisch oder konzeptionell an bestehende Formen von Textströmen an, wie man sie von anderen Übermittlungssystemen (Morse, Fernschreiber), Präsentationsformen (Werbebannern) oder Bildmedien (Newstickern) kennt und die auf ganz andere Textsorten und Nutzungssituationen abziel(t)en. Bisher ist auf Studien zur Nutzung typischer Anwendungen des StreamReaders im Sinne einer Leseforschung ganz verzichtet worden, denn Effizienz oder Ergonomie stehen, wie gesagt, zunächst nicht im Mittelpunkt des Interesses. Stattdessen geht es eher um eine medien- und technikkritische Untersuchung zu Textualität als Medialität: Was sind in den verschiedenen Medien und Texttechnologien die Ausdrucksmöglichkeiten oder Ausdrucksräume von Schrift und wie verhalten sie sich in Übersetzungssituationen zwischen traditionellen und neuen Medien? Hinzu kommt die Frage nach den Genres: wie verhalten sich verschiedene Textsorten in den Textmedien? Welche Medien unterstützen oder behindern welche Textsorten? Welche lassen sich besser oder schlechter von einem anderen Medium adaptieren? Jenseits dieser auf Experiment und Prototyping aufbauenden deskriptiv-analytischen Betrachtungsweise geht der Blick nach vorne: Wenn Text auf eine laufende Zeile reduziert wird, dann schafft das vorgestellte neue Textmedium ganz neue Ausdrucks*räume* im engeren Sinne. Reduktion *und* Expansion! Es entstehen plötzlich Spielräume, in denen experimentell neue Formen der Textpräsentation entstehen können. Unter Erweiterung der einfachen Einzeiligkeit können verstärkt Kontexte und hypertextuelle Verbindungen sichtbar gemacht, Texte stellennah illustriert oder annotiert und mehrfache oder synoptische, aufeinander bezogene Schriftströme realisiert werden.

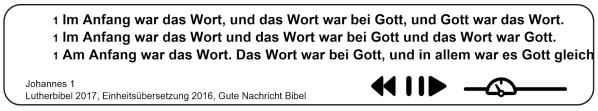


Abbildung 2: Mehrfacher Schriftstrom, hier: mehrfache Übersetzungen.

Abbildung 3: Mehrfacher Schriftstrom, hier: variante Überlieferung mit Normalisierungsstufen.

Erste Experimente zur Darstellung von varianten Fassungen, Normalisierungsschritten in der Transkription oder Visualisierung der semantischen oder Rhythmusstruktur in den Nebensatzkonstruktionen (z.B. in literarischen Werken wie bei Thomas Bernhard) deuten bereits an, dass sich hier jenseits der übersetzenden Reproduktion auch erhebliche produktive und innovative Kräfte entfalten könnten. Denn, Aphorismussektion: Alle neuen Medien müssen nicht nur ihre Formen erst noch finden, sondern auch ihre spezifischen Funktionen.

Digital Humanities?

Dies ist ein sehr offener, explorativer Ansatz. Er speist sich aus keiner Fachdisziplin. Er nimmt seinen Ausgang nicht von bestimmten sprach- oder literaturwissenschaftlichen, anthropologischen, medienwissenschaftlichen oder historischen Fragestellungen. Es ist nur Digital Humanities: Ausgehend von dem allgemeinen Interesse an "Text" und der Reflexion unserer historischen und gegenwärtigen medialen, technologischen und Informationsumwelt soll über experimentelle Anwendungen nachgedacht werden, die uns helfen, durch den Kontrast mit bestehenden Lösungen ein besseres Verständnis dieser Umwelt zu entwickeln und zugleich neue Möglichkeiten auszuloten. Wir kennen die historischen Textmedien, wir haben elektronische Texte entwickelt, die Diskussion um Hypertext und Multimedia ist geführt. Gegenwärtig sehen wir die gegensätzlichen paradigmatischen Strömungen von informationsorientiertem Text Mining und KI auf Basis linearisierter Texte und einer an Komplexität und skripto- bzw. typografischen Details interessierten "material philology" und medienbewussten Textologie. Im vorliegenden Fall dient ein (simulierter, kontrafaktischer) Medienübergang einmal mehr der Reflexion und Modellbildung zu Textualität, textlichen Informationsstrukturen und Ausdrucksformen als Kerngegenstände der Geisteswissenschaften. Überprüft werden dabei auch die Übertragbarkeit und Übersetzbarkeit textueller Information und ihrer Codierungen und damit Effekte, Möglichkeiten und Grenzen der eben genannten Paradigmen von Text-Informativität und Text-Medialität. Die Angebote neuer Technologien sind von den Digital Humanities immer wieder auf ihre Einsatzoptionen hin zu prüfen. Ob sich daraus neue Lösungen und nachhaltig neue Anwendungen ergeben ist zunächst nebensächlich. Grundsätzlich aber gilt, dass Innovation häufig nicht aus dem Versuch entsteht, ein bestehendes Problem zu lösen, sondern aus dem spielerischen Ausloten der Affordanz neu gegebener technischer Rahmenbedingungen. Let's play!

Bibliographie

Die Überlegungen zum StreamReader befinden sich hinsichtlich der Literaturbasis zugleich in einem Ozean und einer Wüste. Die Forschungen zu Schrift, Text, Textualität, Textmedien, Texttechnologien und Lesen sind uferlos, leisten aber nur sehr mittelbar Beiträge zum hier vorgestellten Ansatz. Gering an Zahl sind Texte zu "laufender Schrift" oder gar ihrer Reflexion. Mit dem Konzept und Anwendungsbereich des StreamReaders haben diese außerdem nichts zu tun, da sie z.B. niemals auf klassische literarische Genres, "Langtexte" oder Multitexte zielen. Eine sehr allgemeine Literaturbasis ergäbe

sich allenfalls in der Diskussion um „kritisches Prototyping“ in den Digital Humanities einerseits und literaturwissenschaftlichen Debatten z.B. um den Sprachfluss bei Autoren wie Thomas Bernhard. Beides wird hier aber bislang eher spielerisch aufgenommen und nicht gezielt als Ausgangspunkte genommen. Der StreamReader hat insofern zwar mit sehr vielem zu tun, baut aber auf nichts auf - außer dem allgemeinen historischen Hintergrund, dem aktuellen technischen System und der Neugier auf abweichende Ansätze. Er ist keine Fortführung anderer Arbeiten. Tatsächlich entspringt er letztlich einem Gespräch mit Tobias Kraft, das im April 2016 an der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin stattgefunden hat und sich um Textualität, Medialität und Technizität drehte. Der unmittelbare, bibliografisch manifestierbare Bezug, die (allerdings seitwärts, nicht rückwärts) genutzte Literatur reduziert sich damit auf:

Drach, Sviatoslav *(2019): Neue Leseformen in digitalen Umgebungen - StreamReader*$_{SD}$ *0.1 als Webanwendung für Text als Zeichenstrom, Masterarbeit Universität zu Köln.*

Drach, Sviatoslav (2020): *StreamReader*$_{SD}$ *0.2 - Eine prototypische Webanwendung für das Lesen von Texten als Zeichenstrom*. Book of Abstracts zur DHd2020, Paderborn.

Würgegriff oder Rettungsanker? – Interpretationsspielräume handschriftlicher (Musik-)Quellen im digitalen Kontext

Veit, Joachim
jveit@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Der Schubert-Forscher Walther Dürr veröffentlichte 2002 einen Beitrag zu Problemen der Artikulation und Dynamik bei Franz Schubert, in dem er betonte, wie stark das „Lesen“ einer Partitur von der Kenntnis der Schreibgewohnheiten eines Komponisten abhängt. Selbst erfahrenen Handschriftenkennern bereiten Phänomene wie jenes von „Schuberts so viel diskutiertem Akzentzeichen“ Schwierigkeiten: Akzente und *decrescendo*-Winkel „sind bei Schubert oft nicht leicht zu unterscheiden“ und gelegentlich handele es sich um „etwas dazwischen, das sich im Druck unserer Ausgabe nicht wiedergeben läßt“. Dies gelte auch für manche Bogensetzungen, die „offenbar nicht anzeigen, *was*, sondern nur, *daß* überhaupt gebunden werden sollte“. Er rät dem Editor daher, zwar Beliebigkeiten der Schubertschen Schreibweise zu kennzeichnen, aber wo „Präzision gemeint“ sei, „diese auch dort anzuzeigen, wo *das Manuskript* sie *nicht* hergibt“. Von einer (analogen) Edition erwarte die Aufführungspraxis „genaue Anweisungen“, und „musikalische Plausibilität“ sei dabei zweifellos ein wichtiger Orientierungspunkt (Dürr 2002: 322-326).

In der gedruckten Edition sorgen die Entscheidungen des Editors und die Normierungen des modernen Notensatzes (wie jede Übertragung eines Schriftträgers in einen anderen) zwangsläufig für eine Verengung des in der Vorlage gegebenen (oder laut Dürr bloß vom Lesenden empfundenen) Interpretationsspielraums, der nur durch verbale Erläuterungen im Kritischen Apparat wieder geöffnet werden kann.

Bei den ersten digitalen Editionen von Musik der klassisch-romantischen Epoche mit der „Edirom“-Software (Edirom 2005/2010, Reger-Werkausgabe, OPERA) ging es genau um diese Frage der Transparenz editorischer Entscheidungen, die nun durch die fallspezifische Einbindung von Digitalisaten der zur Erstellung des Edierten Textes herangezogenen historischen Quellen erreicht werden sollte. Die Kombination „eindeutiger“ Edierter Texte mit deren „Vorlagen“ sollte Interpretationsspielräume wieder öffnen, was durch eine zusätzliche Kombination mit Annotationen an Ort und Stelle (also nicht im separierten Apparat) erleichtert wurde. Das Konzept ging teilweise auf, auch wenn die Verführung durch Bilder alles andere als unproblematisch ist – die gebotenen Ausschnitte verkürzen Wirklichkeit und können bei entsprechender Auswahl (und ohne Kenntnis einschlägiger Notationsgepflogenheiten) ebenso manipulativ sein wie traditionelle verbale Erläuterungen des Editors (vgl. dazu Sahle 2013, Kap. 3.2).

Abhilfe versprach die auch für die praktische Nutzung beobachteter Alternativen notwendige Überführung des bildlich Vorgefunden in maschinenles- und verarbeitbare Repräsentationen. Für diese wurde im Bereich wissenschaftlich-kritischer Editionen in den letzten zwanzig Jahren das Format der *Music Encoding Initiative* (MEI) entwickelt, das im Gegensatz zu anderen, auf spezifische Erfordernisse zugeschnittenen Codierungsformen oder proprietären Notensatzprogrammen von Anfang an (in Anlehnung an TEI) auf die dokumentarischen Bedürfnisse der wissenschaftlichen Community zielte (vgl. Richts/Veit 2018). Mit dieser Codierung können nun Interpretationsspielräume wie die erwähnten bzw. unterschiedliche Deutungen dieser Symbolschrift erfasst und explizit festgehalten werden.

Was bedeutet dies konkret? – MEI ist keine „Auszeichnungssprache“ im engeren Sinne (wie TEI), sondern ein „beschreibendes Markup“, das die auf Konventionen beruhende konkrete graphische Gestalt durch Begriffe bezeichnet und bei deren inhaltlicher Deutung auch den Zeichenkontext berücksichtigt – so kann eine Note aufgrund der äußeren Form als „Viertel“ und durch ihre Position im zweiten Zwischenraum in Verbindung mit einem vorausgehenden Schlüssel als Tonhöhe „c2“ bzw. mit einem vorausgehenden Akzidens als „Viertelnote cis2“ bezeichnet werden. Dabei sagen historische Notensatzregeln, dass ein anschließend wiederholter Ton im gleichen Zwischenraum kein Akzidens benötigt, also graphisch wie ein „c“ aussieht, klingend aber als „cis“ realisiert wird. In die Codierung fließt also – wie im Computernotensatz – Wissen um Notationsregeln mit ein. Wenn der Rechner aber mit dem Ton arbeiten soll, muss ihm explizit mitgeteilt werden, dass die graphische Form hier durch zusätzliche Kontextinformationen in ein anderes klingendes Ergebnis verwandelt wird.

Ebenso könnte in dem genannten Schubertschen Beispiel die Ausdehnung des Akzent- bzw. *decrescendo*-Zeichens im Verhältnis zu den Notenpositionen konkret festgehalten und damit expliziter als in einer bloßen verbalen Beschreibung dokumentiert werden. Zusätzlich wären die Deutungsmöglichkeiten als Alternativen in der Codierung – eventuell in einem Apparateintrag als Lesarten – festzuhalten. Dabei erlaubt MEI Angaben zum Urheber der jeweiligen Interpretation sowie prozentuale Festlegungen der Wahrscheinlichkeit der jeweili-

gen Lösung. Die Codierung erzwingt also eine möglichst präzise Beschreibung der Alternativen – aber wie hoch ist der Erkenntnisgewinn? Und wo liegen – von dem heilsamen Zwang zur präziseren Erfassung der Phänomene abgesehen – die Vorteile eines solchen Verfahrens gegenüber der traditionellen analogen Arbeit?

Um die Frage zuzuspitzen: Menschenlesbar wird das, was hier codiert wird, erst bei einer Rücküberführung in die gewohnte Notendarstellung – dort aber sorgt die Normierung des Drucksatzes dafür, dass der Eindruck, den die Handschrift vermittelte, ein völlig anderer ist. Und die präzise Codierung etwa von Bogenlängen, die nicht mit, sondern irgendwo nach Noten beginnen oder enden (und damit ggf. Bedeutungsunterschiede suggerieren), erweist sich letztlich als verlorene Liebesmüh', denn sie wirkt in der normierten Umgebung völlig anders und bleibt als bloße Positionsbestimmung blind für eine maschinelle Auswertung, die die Werte in Beziehung zu unterschiedlichen Kontextfaktoren setzen müsste. Das Projekt „Beethovens Werkstatt", das sich mit der komplexen Genese Beethovenscher Kompositionshandschriften beschäftigt, hat daraus die Konsequenz gezogen, solche kodikologischen Aspekte nicht im Neusatz nachzuahmen (und damit zu verfälschen), sondern die Codierung sozusagen fest mit den Einträgen im Manuskript zu verdrahten – das Markup zur Beschreibung des problematischen Bogens wäre dementsprechend direkt mit einer SVG-Erfassung dieses Objekts im Handschriftendigitalisat verknüpft. Das erleichtert das Erkennen von im Apparat erfassten Mehrdeutigkeiten, dennoch bleiben diese ohne Annotation schwer nachvollziehbarund der Urteilsfähigkeit eines Betrachters anvertraut, der zudem schreibereigene Notationsgepflogenheiten zu berücksichtigen hätte. Zwar könnte man vorgefundene Phänomene unter Kategorien subsumieren und damit rascher auffind- und auswertbar machen – aber dennoch bewegen wir uns hier noch in einem Denkraum, für den das Digitale zwar Erleichterungen bringt, der aber traditionellen Herangehensweisen verpflichtet bleibt, weil die Repräsentation des Objekts, die MEI in der bisher beschriebenen Form bietet – wie bei allen derartigen Repräsentationen – nur auf einer sehr spezifischen, von bestimmten Interessen geleiteten Wahrnehmung des Gegenstands beruht (Sahle 2013, Kap. 2).

Greifen also bisherige digitale Editionsmethoden oder Codierungen zu kurz? Wie aber könnten digitale Methoden helfen, mit den genannten Interpretationsspielräumen sinnvoller umzugehen? Ist hier nicht ein radikal anderes Denken erforderlich?

Zunächst muss man sich bewusst machen, welche Fragen überhaupt mit Rechnerunterstützung sinnvoll beantwortbar sind. Bei den erwähnten Bögen mit unklarem Anfang und Ende oder der Akzent/ *decrescendo*-Unterscheidung bleibt die Geltungsdauer der Zeichen stark von individuellen Schreibgewohnheiten abhängig und dürfte kaum schreiberunabhängig beurteilbar sein. Aber Dürrs (durch Doppelautographe Carl Maria von Webers bestätigte) Hypothese, andere Bogenformen suggerierten nur, „daß" und nicht „was" genau gebunden werden solle – bezeichneten also im Sinne eines *sempre legato* nur grundsätzlich das Binden (nicht aber z.B. den Strichwechsel) –, wäre auf einem großen Korpus an Digitalisaten (ggf. epochenspezifisch) untersuchbar. Was wäre hier nötig?: Es muss zunächst händisch ein ausreichend großer Bestand (zu Festlegung seiner Größe fehlen noch jegliche Erfahrungswerte!) erfasst werden, um die Frage – auch im Hinblick auf zumindest zeichenspezifisches OMR – präzisieren zu können. Darauf aufbauend könnte eine maschinelle Auswertung umfassender Bibliotheksbestände (z.B. über den Zugriff auf im IIIF-Format zur Verfügung gestellte Digitalisate) erfolgen. Das angewandte Verfahren müsste auch in der Lage sein, aufgefundene Stellen so zu „markieren" (bzw. ihre Koordinaten zu erfassen), dass sie für spätere Einzelfallstudien rascher auffindbar wären. Je nach Aufbereitung der Trainingsdaten könnte dabei bereits eine automatisierte Sortierung der Fundstellen nach vorgegebenen Kategorien erfolgen, um darauf aufsetzende Arbeiten zu erleichtern. (Bei der Interpretation des Akzentzeichens wäre z. B. ein Einbeziehen verbaler Bezeichnungen wie *sf* und *fz* sowie weiterer orthographischer Varianten denkbar, um die Frage zu klären, inwieweit die Verwendung zeitlich, lokal oder im Hinblick auf die Faktur der Musik variiert, je nach Kontext unterschiedliche Deutungen suggeriert oder lediglich auf Synonymität hindeutet.) Bei dieser Form des Markup hilft die für Common Western Notation im Moment entwickelte automatische Taktmarkierung (Waloschek), da sie ein Festhalten von Phänomenen nicht bloß in abstrakten Koordinaten, sondern auch in Bezug auf ein inhaltliches Modell (hier die in MEI abgebildete Werkstruktur) erlaubt.

Ein zweites Beispiel, das im Falle Mozarts gar zu einem Preisausschreiben geführt hat (Albrecht): Die in der Aufführungspraxis heiß diskutierte Frage nach dem Unterschied zwischen „Punkt" und „Strich" in Artikulationsbezeichnungen bzw. die Grundsatzfrage, ob überhaupt ein Bedeutungsunterschied zu konstatieren sei (Brown: 200ff.). Die subjektive Beobachtung, dass z. B. Striche in *forte*-Abschnitten eindeutig überwiegen, wäre statistisch und im Hinblick auf bestimmte Zeitabschnitte belegbar. Ebenso die Hypothese, dass besonders deutliche Striche Akzentfunktion haben. Aber für die zahllosen, gerade bei Handschriften kaum unterscheidbaren Zwischenformen wäre zunächst ein auch begrifflich schwer zu differenzierendes Vergleichskorpus anzulegen und zusätzlich auf „Normalwerte" eines Schreibers zu beziehen (welche zudem von Schreibmittel und beschriebener Oberfläche abhängen), um die Auswertung nicht zu verfälschen. Neben die Auswertung des graphischen Befunds muss außerdem stets eine Auswertung aufführungspraktischer Hinweise in Unterrichts- und Lehrwerken oder erläuternder Texte und Selbstzeugnisse im zeitlichen Kontext treten – trotz des komplizierten Zusammenspiels, das m.E. methodisch die Grenzen unseres Faches überschreitet und auch für die Informatik interessante Modellierungsprobleme bietet, sind hier hilfreiche Erkenntnisse zu erwarten.

Letztlich wird man sich derartigen Interpretationsproblemen von zwei Seiten nähern können: Wenn etwas für bedeutungstragend gehalten wird, sollte es in der Codierung festgehalten (also „be-zeichnet") werden, um eine Sammlung von Befunden anzulegen, die maschinell leicht akkumulierbar und strukturierbar ist – auf der anderen Seite werden Materialitäts- und Schriftlichkeitsuntersuchungen im großen Stil durch neuronale Netze oder Künstliche Intelligenz erst jetzt (bzw. künftig) sinnvoll durchführbar. Dann können solche Verfahren wirklich zum Rettungsanker zumindest bei ausgewählten Interpretationsproblemen werden. Bisherige isolierte Untersuchungen bergen stets die Gefahr sehr eingeschränkter Gültigkeit, die heute möglich werdende Korpus-Analyse birgt andererseits die Gefahr, dass die Bedingungen des Einzelfalls nicht genügend berücksichtigt sind – zwischen beiden Extremen kann sich künftig eine sinnvolle Nutzung digitaler Techniken bewegen, die mit neuen Mitteln die Spielräume von Interpretation auszuloten versucht.

Links

https://beethovens-werkstatt.de/
https://music-endocing.org/

Bibliographie

Albrecht, Hans (1957): *Die Bedeutung der Zeichen Keil, Strich und Punkt bei Mozart. Fünf Lösungen einer Preisfrage*, Kassel: Bärenreiter

Brown, Clive (1999): *Classical & Romantic Performance Practice 1750–1900*, Oxford: OUP

Dürr, Walther (2002): „Notation und Aufführungspraxis: Artikulation und Dynamik bei Schubert", in: *Musikedition. Mittler zwischen Wissenschaft und musikalischer Praxis*, hg. von Helga Lühning (Beihefte zu editio 17), Tübingen: Niemeyer, S. 313-327

Edirom (2005): *Carl Maria von Weber. Sämtliche Werke*, Serie VI, Bd. 3: *Kammermusik mit Klarinette*, hg. von Gerhard Allroggen, Knut Holtsträter und Joachim Veit. Mit einer digitalen Edition des Quintetts op. 34 von Johannes Kepper u. Ralf Schnieders, Mainz: Schott Musik International

Edirom (2010): **Carl Maria von Weber. Sämtliche Werke**, Serie V, Bd. 6: *Konzertante Werke für Klarinette*, hg. von Frank Heidlberger. Mit einer digitalen Edition der Werke, erarbeitet von Benjamin Wolff Bohl, Daniel Röwenstrunk u. Joachim Veit unter Mitwirkung von Philemon Jacobsen, Mainz: Schott Musik International

Ertel, Wolfgang (2016): *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung* (Computational Intelligence), 4. Auflage, Wiesbaden: Springer Vieweg

OPERA. Spektrum des europäischen Musiktheaters in Einzeleditionen (2013ff.), hg. von Thomas Betzwieser, Kassel: Bärenreiter (bislang 3 Bde.)

Reger, Max (2008–2015): *Orgelwerke. Reger-Werkausgabe*, Serie I, Bd. 1–7, hg. von Alexander Becker et al., Kritischer Bericht auf DVD, Stuttgart: Carus

Richts, Kristina / Veit, Joachim (2018): „Stand und Perspektiven der Nutzung von MEI in der Musikwissenschaft und in Bibliotheken", in: *Bibliothek – Forschung und Praxis* 42 (2) ,S. 291-301

Sahle, Patrick (2013): *Digitale Editionsformen. Zum Umgang der Überlieferung unter den Bedingungen des Medienwandels* (Schriften des Instituts für Dokumentologie und Editorik 9), Teil 1: Das typographische Erbe, Teil 2: Befunde Theorie und Methodik, Teil 3: Texbegriffe und Recodierung, Norderstedt: BoD

Schmid, Manfred Hermann (2012): *Notationskunde. Schrift und Komposition 900–1900* (Bärenreiter Studienbücher 18), Kassel

Waloschek, Simon / Hadjakos, Aristotelis / Pacha, Alexander (2019): „Identification and Cross-Document Alignement of Measures in Music Score Images", in: *Proceedings of the 20th International Society for Music Information Retrieval Conference* (ISMIR), Delft 2019

Zu den Anforderungen einer musikalischen Stilometrie

Kepper, Johannes
kepper@edirom.de
Universität Paderborn, Deutschland

In der Literaturwissenschaft ist die Methode der Stilometrie ein etabliertes Verfahren, um Fragestellungen verschiedener Art zu bearbeiten. Die Zuordnung anonym bzw. pseudonym überlieferter Werke zu bekannten Autoren aufgrund persönlicher stilistischer Gewohnheiten steht dabei oft im Vordergrund, aber auch die Untersuchung stilistischer Merkmale selbst ist Gegenstand von Stilometrie. Mit den Möglichkeiten der Digital Humanities lassen sich nicht nur größere Texte, sondern sogar ganze Korpora automatisiert auswerten und im Hinblick auf unterschiedlichste Textmerkmale untereinander vergleichen. Eine nach wie vor bestehende Hürde stellt dabei die Verfügbarkeit zuverlässigen Ausgangsmaterials dar, also die digitale Volltexterschließung des zu untersuchenden Materials. Zwar gibt es inzwischen mehrere Repositorien, die nennenswerte Textkorpora bereitstellen,[1] eine vollständige Abdeckung auch nur einzelner Epochen der Literaturgeschichte ist jedoch trotz der stetig verbesserten Erkennungsraten im Bereich der *Optical Character Recognition* (OCR) noch immer in weiter Ferne.

Stilkritik ist auch in der Musikwissenschaft eine übliche Methode. Hier wie dort geht es zunächst darum, autor- bzw. komponistenspezifische Eigen- und Gewohnheiten durch den Abgleich mit Vergleichswerken und -komponisten herauszuarbeiten. Während aber Stilometrie im literaturwissenschaftlich geprägten Bereich der Digital Humanities dankbar und schnell aufgegriffen und als eigener digitaler Forschungsbereich etabliert wurde, bleibt dieses Feld im musikwissenschaftlichen Teil der Digital Humanities bislang weitgehend unbespielt – eine digitale musikalische Stilometrie findet zumindest in der eingangs erwähnten Breite der Fragestellungen nicht statt. Verschiedene musikbezogene Disziplinen – in erster Linie Musikinformatik und Music Information Retrieval (MIR) – forschen zwar durchaus an Fragestellungen, die der Stilometrie inhaltlich nahestehen, setzen diese jedoch nicht im Sinne einer musikwissenschaftlicher Digital Humanities um.[2]

Der vorliegende Beitrag soll versuchen, die methodischen und gegenstandsbezogenen Gemeinsamkeiten und Unterschiede zwischen (textbezogener) Stilometrie und Music Information Retrieval herauszuarbeiten. Davon ausgehend sollen Perspektiven entwickelt werden, welchen Anforderungen eine erfolgversprechende musikalische Stilometrie gerecht werden müsste. Dabei soll es nicht um statistische bzw. algorithmische Optimierungen etwa im Hinblick auf einzusetzende Distanzmaße (Evert 2017) gehen, sondern vielmehr um grundsätzliche Überlegungen zu den Anforderungen aus musikwissenschaftlicher Sicht.

Musik folgt grundsätzlich Regeln, die sich zwar von Epoche zu Epoche unterscheiden, und deren Durchbrechen oft als Innovation wahrgenommen und damit erwünscht ist, die sich aber in einer Musiktheorie beschreiben lassen, und deren Umsetzung Raum zur Entwicklung eines eigenen Stils bie-

tet. Die grundsätzliche Formalisierbarkeit charakteristischer Merkmale bildet die Grundlage für Forschungen im Bereich der Komponisten-Attribution. Eine wichtige Arbeit in diesem Feld, die sich schon in ihrem Titel *On musical stylometry: A pattern recognition approach* ausdrücklich auf Stilometrie bezieht, wurde 2005 von Eric Backer und Peter van Kranenburg vorgelegt (Backer 2005). Darin extrahieren Backer und van Kranenburg zwanzig verschiedene Features aus dreissig Fugen von Komponisten, um dann eine weitere Fuge, die unklaren Ursprungs ist und in der Literatur allen drei Komponisten zugeordnet wird, auf ihre stilistische Nähe zu den anderen Werken zu untersuchen. Tatsächlich stellen sie eine große stilistische Nähe zu einem der Komponisten fest und können so glaubhaft für dessen Autorschaft argumentieren. Allerdings lässt sich ihre Arbeit nicht ohne weiteres übertragen. Die von ihnen gewählten Features sind sehr deutlich auf die musikalische Faktur des untersuchten Repertoires – barocke Fugen – abgestellt. Die Aussagekraft der bei Backer und van Kranenburg deutlich im Vordergrund stehenden Harmonie- und Stimmführungsbezogenen Features (Backer 2005: 304) ist jedoch zunächst nur für das von ihnen untersuchte Material als gegeben anzusehen. Tatsächlich beobachten sie in einem vorgeschalteten Experiment, dass aus einer heterogeneren Mischung von Komponisten – J.S. Bach, G.P. Telemann, G.F. Händel, W.A. Mozart und J. Haydn – deutlich schlechtere Ergebnisse resultieren, als wenn sie sich auf die Barock-Komponisten Bach, Telemann und Händel beschränken. Die gewählten Features können damit zur Beschreibung eines musikalischen Stils nicht als allgemeingültig betrachtet werden.

In der Musikwissenschaft wird musikalischer Stil auf verschiedenen Ebenen untersucht: Stil in Bezug auf Epochen ebenso wie auf Gattungen, harmonischer, melodischer und rhythmischer Stil, idiomatische, d.h. instrumentenspezifische Notationsweisen und Spielfiguren usw. (Pascall 2001). Eine Möglichkeit, dem zu begegnen, wäre die Berücksichtigung zahlreicher weiterer Features, wie es etwa in der Software *jMIR* bzw. dessen Komponente *jSymbolic* geschieht. Deren Zielsetzung "is to extract statistical information from musical data", um "research in the fields of music information retrieval (MIR), musicology and music theory"[3] zu ermöglichen. Zu diesem Zweck werden insgesamt 246 verschiedene Features extrahiert, die in ihrer Breite deutlich generischer angelegt sind als bei Backer und van Kranenburg, daher auch über Gattungs- und Epochengrenzen hinweg verschiedene Stile besser abbilden und so insgesamt "robustere" Ergebnisse produzieren sollten. Sie decken unterschiedliche Aspekte eines Notensatzes ab, darunter Tonhöhen (41 Features), Melodielinien (25), Akkorde (35), Rhythmus (95), Instrumentation (20).[4] Damit wird auf den ersten Blick bereits eine hohe Bandbreite musikalischer Eigenschaften berücksichtigt, die allerdings im Vorfeld (automatisiert) extrahiert werden müssen, und die wiederum im Einzelnen nicht immer von gleichbleibender Aussagekraft sind, sondern je nach Stilbegriff und untersuchtem Material gezielt ausgewählt werden müssen, um aussagekräftige Ergebnisse zu erzielen, indem "irrelevant features that would introduce unproductive noise into classification systems" (McKay 2010: 197) vermieden werden.

Spätestens an dieser Stelle wird deutlich, dass der Datenaufbereitung im Fall von Musik eine besondere Bedeutung zukommt. In der (literaturwissenschaftlichen) Stilometrie werden demgegenüber i.d.R. vollständige Texte ausgewertet, die auf den ersten Blick nicht speziell aufbereitet werden. Allerdings finden auch hier durchaus vergleichbare Vorbereitungen statt, die allerdings im Fall des verbreiteten Programms *Stylo*[5] in den üblichen Arbeitsablauf direkt integriert sind. Üblicherweise basieren stilometrische Untersuchungen mit *Stylo* auf der Untersuchung der *most frequent words* (MFW), also der Verteilung der am häufigsten genutzten Wörter, da diese "kaum bewusst manipuliert" werden[6] und so ein gutes Kriterium zur Autorattribution sind. Hierzu werden also die Texte in ihren Schreibweisen normalisiert und dann anhand ihrer Frequenzen die Verteilung der *n* häufigsten Worte bestimmt. Die eigentliche stilometrische Untersuchung besteht nun darin, die resultierenden Histogramme statistisch abzugleichen. Dieses Verfahren ist durchaus vergleichbar mit der Untersuchung extrahierter Features im Bereich MIR. Es erscheint daher naheliegend, die Eignung von *Stylo* zur stilometrischen Untersuchung musikalischer Daten zu überprüfen. Ein geeignetes Datenset findet sich im *Haydn/Mozart String Quartet Quiz*.[7] Bei diesem Quiz werden dem Benutzer zufällig Streichquartette von Mozart oder Haydn vorgespielt, die dieser den Komponisten zuordnen soll. Die durchschnittliche Trefferquote für die 287 Sätze, die im Korpus enthalten sind, liegt bei annähernd 50.000 gehörten Sätzen bei ungefähr 58%, gemittelt über Hörer aller Grade an musikalischem Vorwissen.[8] Diese geringe Quote, die nur unwesentlich besser als eine statistische Normalverteilung ist, verdeutlicht die stilistische Nähe der beiden Komponisten: Es ist auch für menschliche Hörer alles andere als einfach, die Werke beider Komponisten zu unterscheiden. Damit stellt dieser Korpus einen interessanten Testfall dar, da die Hürde für eine nach menschlichen Maßstäben "erfolgreichere" Klassifizierung nicht allzu hoch liegt, während die Schwierigkeit gleichzeitig recht hoch ist und damit genug Raum zur Optimierung bietet.

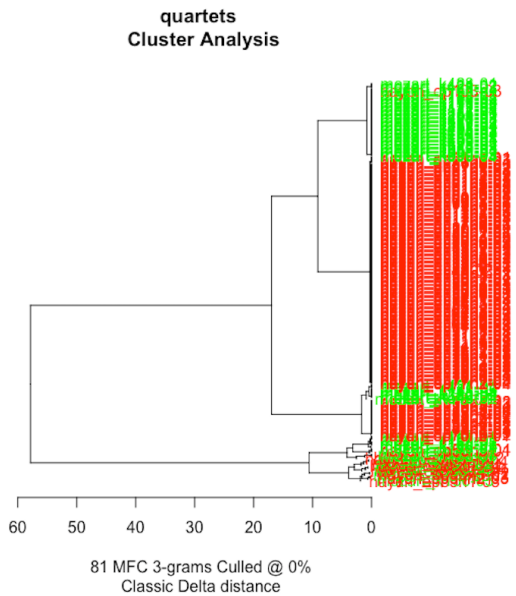


Abbildung 1: Auswertung der Daten des Haydn/Mozart-Quiz, basierend auf MusicXML. Grüne Einträge verweisen auf Mozart, rot steht für Haydn.

Das Datenmaterial findet sich gleich mehrfach im Netz. Es basiert auf Codierungen, die in den 1990er Jahren am *Center for Computer Assisted Research in the Humanities* (CCARH) der Stanford University entstanden sind, und die in verschiedenen Datenformaten von http://kern.ccarh.org/herunterge-

laden werden können. Um ein auch strukturell möglichst einfaches Format nutzen zu können, wurden die im MusicXML-Format geladenen Daten per Skript ins abc-Format konvertiert.[9] Dieses plain-text-Format nutzt eine sehr einfache, nach Stimmen geordnete Syntax, um Melodielinien zu codieren. Mit entsprechenden Parametern[10] wird hier eine Erkennungsrate erzielt, die mit ~62% etwa im Bereich der menschlichen Benutzer des Mozart-Haydn-Quiz liegt. Eine etwas bessere Unterscheidung (~69%) gelingt mit *Stylos classify()*-Funktion, wenn dort als Algorithmus *SVM (Support Vector Machines*[11]) ausgewählt werden. Die genutzten Wörter sind dabei in jedem Fall deutlich aussagekräftiger, wie das Beispiel "D E F G F E D C B A" zeigt. Trotz der besseren Ergebnisse zeigt sich, dass die Betrachtung nur eines einzelnen Aspekts der Notation – hier der Abfolge von Tonhöhen – noch keine ausreichende Aussagekraft bietet, um darauf stilometrische Analysen aufzubauen. Musik ist "mehrdimensional", d.h. mehrere Stimmen bzw. Instrumente verlaufen zeitgleich. Während es textbasierte Datenformate gibt, die sich im Fall von einstimmiger Musik recht plausibel als Text behandeln lassen, lässt sich mehrstimmige Musik nur mit einem Fokus entweder auf Harmonik (d.h. als Fortschreibung von Klängen) oder auf Melodien (d.h. als Abfolge einzelner Stimmen, deren Gleichzeitigkeit in den Daten nicht ersichtlich wird) in einem Datenformat darstellen. Das explizite Aufbereiten der Daten – die Extraktion von Features – ist also unumgänglich, um sämtliche Facetten von Musik in den Blick nehmen und im Rahmen einer stilometrischen Analyse berücksichtigen zu können.

Vor diesem Hintergrund erscheint es naheliegend, zu *jSymbolic* zurückzukehren, da dessen oben erwähnte Features ein weitaus breiteres Spektrum der musikalischen Mehrdimensionalität abbilden. *jSymbolic* arbeitet i.d.R. mit MIDI-Daten – es gibt zwar durchaus die Möglichkeit, MEI-Daten zu verarbeiten, allerdings werden lediglich zwei Features (Häufigkeit von Bindebögen bzw. Vorschlagsnoten) aus diesen Daten extrahiert, die bei einem MIDI-Input nicht zur Verfügung stehen. *jSymbolic* übernimmt dabei nur die Extraktion und Aufbereitung der Features, die eigentliche statistische Auswertung erfolgt durch *Weka*, eine Java-basierte, unter GPL lizensierte Software zum maschinellen Lernen.[12] Diese bietet eine zu *Stylo* vergleichbare Benutzeroberfläche und erlaubt es u.a., die extrahierten Features ebenfalls mit dem *SVM*-Algorithmus auszuwerten. Bei Übernahme der im *jSymbolic*-Tutorial beschriebenen Standard-Einstellungen[13] wird bereits eine korrekte Klassifizierung von knapp 82% erreicht, d.h. 233 von 286 Dateien werden korrekt zugeordnet. Dieses Ergebnis ist überaus beachtlich, insbesondere da keinerlei Optimierungen vorgenommen wurden, und auch auf genau dieses Datenset hin optimierte Lösungen aktuell nur auf Raten von gut 85% kommen (Kempfert 2019: 13).

Damit scheinen die von *jSymbolic* extrahierten Features eine gute Ausgangslage zu bieten, um stilometrische Fragestellungen im Bereich Musik zu bearbeiten – sei es mithilfe von *Stylo*, *Weka*, oder anderen entsprechenden Werkzeugen. Allerdings stellt sich die Frage, ob der geschilderte Workflow bereits das volle Potential des zur Verfügung stehenden Datenmaterials ausschöpft: MIDI taugt als Datengrundlage nur sehr bedingt, wenn es um Musiknotation geht. Einer der offensichtlichsten Gründe ist dabei die Art, Tonhöhen quasi als durchnummerierte Klaviertasten zu codieren, da in diesem System kein Unterschied zwischen z.B. den Tönen Dis und Es besteht. Will man nun das Intervall etwa von einem C zu diesem Ton bestimmen, so ergibt sich in einem Fall eine Sekunde, im anderen ein Terzabstand. Auch hier gilt wieder, dass das Ignorieren dieser sog. *Enharmonik* nicht zwangsläufig und in jedem Fall zu schlechteren Ergebnissen führen *muss*, allerdings sind derartige Informationen für eine "händische" Klassifizierung überaus hilfreich, nicht zuletzt zur Epochenzuordnung. An dieser Stelle merkt man *jSymbolic* an, dass es im Rahmen von *jMIR* nur eine Komponente darstellt, und es (neben weiteren) mit *jAudio* eine Entsprechung zur Feature-Extraktion aus Audiodaten gibt – in denen dann naturgemäß keine Informationen zur Enharmonik enthalten sein können. *jSymbolic* schöpft hier also aus Gründen der Merkmalsparität der extrahierten Features nicht das volle Potential der Datengrundlage codierter Partituren aus.[14] Da in den letzten Jahren immer mehr Editionsprojekte hochwertige Datenkorpora erarbeiten und bereitstellen, bietet sich hier eine Perspektive, wie stilometrische Fragestellungen auch im Bereich Musik sinnvoll projektiert werden können: Noch vor der Optimierung der statistischen Methoden für das zu untersuchende Material muss dessen Aufbereitung für eine solche Auswertung optimiert werden. Dazu gibt es umfangreiche Vorarbeiten, insbesondere im Bereich des Music Information Retrieval, die aber stellenweise aus musikwissenschaftlicher Sicht noch defizitär sind und das volle Potential der zur Verfügung stehenden Daten bislang nicht ausschöpfen.

Fußnoten

1. Vgl. etwa die über TextGrid frei verfügbaren Bestände von http://zeno.org.
2. Als Beispiel sei hier Cilibrasi, Rudi / Vitányi, Paul / de Wolf, Ronald (2004): "Algorithmic clustering of music based on string compression", in: *Computer Music Journal* , 28 (4), 49–67 genannt. Dieser Beitrag diskutiert auf rein statistischer Ebene mögliche Verwandtschaftsbeziehungen zwischen verschiedenen Musikcodierungen, ohne dabei jedoch deren musikwissenschaftliche Bedingungen bzw. Implikationen angemessen zu thematisieren.
3. Vgl. http://jmir.sourceforge.net/ (letzter Zugriff 06.01.2020).
4. Vgl. http://jmir.sourceforge.net/manuals/jSymbolic_manual/home.html (letzter Zugriff 06.01.2020).
5. Vgl. https://github.com/computationalstylistics/stylo (letzter Zugriff 06.01.2020).
6. Jan Horstmann (2018, § 11): „Stilometrie“. In: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/stilometrie (letzter Zugriff 06.01.2020). Das Zitat wird dort Fotis Jannidis und Gerhard Lauer: „Burrows's Delta and Its Use in German Literary History”. In: Matt Erlin und Lynne Tatlock (Hrsg.): *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Rochester, New York: Camden House, 29–54, S. 180 zugeschrieben, findet sich dort aber zumindest nicht in dieser Form.
7. Vgl. http://qq.themefinder.org/ (letzter Zugriff 06.01.2020; das Quiz selbst war zu diesem Zeitpunkt unter verschiedenen aktuellen Browsern nicht funktionsfähig).
8. Die Benutzer des Quiz müssen sich durch einige Fragen selbst auf einer Skala von 0 ("novice") bis 10 ("expert") einordnen. Die Erkennungsrate schwankt laut einer Auswertung im Frühjahr 2006 zwischen ~51% (Level 1) und ~88% (Level 8). Vgl. http://qq.themefinder.org/cgi-bin/mhstyle?submit=statistics (letzter Zugriff 06.01.2020).
9. Zur Konvertierung vgl. https://wim.vree.org/svgParse/xml2abc.html, zum Datenformat vgl. http://abcno-

tation.com (letzter Zugriff jeweils 06.01.2020). Tatsächlich lassen sich auch die MusicXML-Daten direkt in *Stylo* auswerten. Die dabei zu erreichende Erkennungsrate von ~75% ist aber irreführend: Die von *Stylo* angelegte Wortliste offenbart, dass vor allem Metadaten zur Besetzung, die in den beiden Datensätzen unterschiedlich angelegt sind, ausschlaggebend sind, und musikalische Inhalte im Grunde nicht berücksichtigt werden.
10. Auswertung von n-grams mit n=10 auf Basis vollständiger Wörter, 100 MFW, kein Culling, Classic Delta.
11. Vgl. https://de.wikipedia.org/wiki/Support_Vector_Machine (letzter Zugriff 06.01.2020).
12. Vgl. http://www.cs.waikato.ac.nz/ml/weka/ (letzter Zugriff 06.01.2020).
13. Vgl. http://jmir.sourceforge.net/manuals/jSymbolic_tutorial/usingweka.html (letzter Zugriff 06.01.2020).
14. Gleiches gilt etwa für die Dissertation von Christof Weiß ("Computational methods for tonality-based style analysis of classical music audio recordings", Ilmenau 2017,), die zwar einen überaus wichtigen Beitrag zur automatischen Klassifizierung von Musik anhand stilistischer Merkmale liefert, dabei aber ausschließlich auf Audiodaten operiert.

Bibliographie

Backer, Eric / van Kranenburg, Peter (2005): "On musical stylometry – a pattern recognition approach", in: *Pattern Recognition Letters* 26, 299–309.

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Reger, Isabella / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2017): "Understanding and explaining Delta measures for authorship attribution", in: *Digital Scholarship in the Humanities* 32, ii4–ii16. https://doi.org/10.1093/llc/fqx023

Horstmann, Jan (2018): "Stilometrie", in: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/stilometrie.

McKay, Cory (2010): "Automatic music classification with jMIR" (PhD). McGill University, Montréal.

Kempfert, Katherine C. / Wong, Samuel W. K. (2019): "Where Does Haydn End and Mozart Begin? Composer Classification of String Quartets", arXiv:1809.05075 [stat].

Pascall, Robert (2001): "Style", in: *The New Grove Dictionary of Music and Musicians*. Oxford. 24, 638–642.

Doctoral Consortium

Annotation of Non-Standard Varieties

Seltmann, Melanie E.-H.
melanie.seltmann@univie.ac.at
Universität Wien, Österreich

Das Dissertationsprojekt beschäftigt sich mit nachhaltiger Annotation sprachlicher Kategorien in der Variationslinguistik. Untersuchungsgegenstand ist gesprochene Sprache. Hierbei fokussiert das Projekt verschiedene thematische Bereiche: Hauptthemen sind dabei zum einen Standardisierungsversuche von linguistischen Annotationen, Annotation insbesondere von Nonstandardvarietäten und dabei auftretende Probleme, Annotation als Kategorisierung und die damit verbundenen Möglichkeiten des Erkenntnisgewinns sowie der Explikation theoretischer Grundannahmen, die Beeinflussung der Kategorien vorab anhand der Erhebungsmethode, technische Aspekte der Modellierung und Darstellung, Nachhaltigkeit der Annotation sowie eine kritische Reflexion verwendeter Annotationsmethoden.

Die Dissertation entsteht im Rahmen des Spezialforschungsbereichs „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (FWF F60) (im Folgenden SFB) im Teilprojekt PP11 „Kollaborative Online-Forschungsplattform ‚Deutsch in Österreich‘“ und nimmt auch deswegen die Annotationen des SFB als Ausgangs- und Referenzgröße (vgl. Budin et al. 2018). Der SFB bietet die Möglichkeit ein geschlossenes Annotationssystem, das alle linguistischen „Systemebenen“ aus verschiedenen sprachwissenschaftlichen Perspektiven abdeckt, zu entwickeln bzw. iterativ dessen Entwicklung zu beobachten und zu untersuchen. Insbesondere im Bereich Standardisierung von Annotationen werden auch andere linguistische Projekte zum Vergleich herangezogen und nationale sowie internationale Forschungsverbände und ihre Best Practices sowie De-facto-Standards berücksichtigt.

Ziel der Arbeit ist es herauszuarbeiten, welche Vorteile und Möglichkeiten in einer standardisierten Annotation liegen, welche Möglichkeiten es für eine weitestgehende Standardisierung innerhalb der (Variations-)Linguistik gibt, aber auch welche Probleme dieser inne liegen. Zeitgleich wird reflektiert, inwiefern ein Versuch eines solchen standardisierten Annotationssystems (nämlich das des SFB) gelingen kann und wo seine Grenzen liegen. Dabei wird ein neuartiges Annotationssystem vorgeschlagen, das insbesondere auf die Standardisierung der Modellierung von Tag Sets abzielt, aber auch auf die Standardisierung des verwendeten Annotationsvokabulars (vgl. Breuer/Seltmann 2018: 147).

Der Dissertation liegt die übergeordnete Forschungsfrage zu Grunde:

Was sind Anforderungen an ein Annotationssystem für die Variationslinguistik, insbesondere zur Annotation von Nonstandardvarietäten? Inwiefern können Annotationssysteme zu einer guten Forschungspraxis beitragen und den Erkenntnisgewinn bzw. die -vermittlung erhöhen oder zumindest erleichtern? Welche Erfahrungen können im SFB „Deutsch in Österreich. Variation – Kontakt – Perzeption.“ mit dem dafür entwickelten Annotationssystem gemacht werden? Welche Stärken und Schwächen hat dieses?

Um diese Fragestellung beantworten zu können, orientieren sich die verschiedenen Papers an folgenden Unterfragen:

1. Kategorisierungsprozess
   - Inwiefern handelt es sich bei der Annotation um einen Erkenntnisprozess?
   - Welche Vor- und Nachteile bringt die Annotation als eine Methode der Digital Humanities dem Erkenntnisprozess?
   - Unter welchen Voraussetzungen können auch Citizens Analysen mittels Annotation betreiben?
2. Standardisierung
   - Welche Anforderungen werden an ein Annotationssystem für Nonstandardvarietäten gestellt?
   - Inwiefern ist ein solches Framework standardisierbar bzw. welche Aspekte davon?
   - Welche Vor- und Nachteile birgt ein standardisiertes Annotationssystem?
3. Variationslinguistische Variable in der Annotationsumsetzung
   - Wie ist im Hinblick auf linguistische Annotationen systemübergreifend eine Variable zu definieren?
   - Wie ist der Variablenbegriff auf verschiedene linguistische Systemebenen ansetzbar?
   - Wie ist ein solcher Variablenbegriff technisch für ein Annotationssystem implementierbar?
   - Inwiefern kann die technische Implementierung helfen, den Variablenbegriff für die Variationslinguistik greifbarer zu machen?
   - Welche technisch-formale Modellierung eignet sich für diesen Variablenbegriff: eher eine hierarchische oder eher eine relationale?
   - Inwiefern unterstützt die technische Abbildung der Variable den Forschungsprozess? Inwiefern beeinflusst sie die Ergebnisse?
   - Inwiefern werden die Ergebnisse des Forschungsprozesses vorab durch die Erhebungsmethode beeinflusst?
4. Nachhaltigkeit
   - Was bedeutet „Nachhaltigkeit“ in Forschung und Infrastruktur bei (variations-) linguistischen Projekten im Hinblick auf Annotationen?
   - Wie kann eine solche Nachhaltigkeit in linguistischen Projekten erreicht werden?
   - Wie versucht konkret der SFB Nachhaltigkeit sowohl der Forschungsergebnisse als auch der technischen Erzeugnisse zu gewährleisten?
   - Wie kann Nachhaltigkeit auch für die (interessierte) Öffentlichkeit gewährleistet werden?
5. Usability
   - Welche Erfahrungen und Probleme sind mit der Umsetzung der Annotationsrichtlinien im SFB eingetreten?
   - Inwiefern waren Abläufe und Teilaspekte des Annotationsprozesses zielführend und hilfreich, an welchen Punkten bedarf es Verbesserung?

Der Dissertation liegen die theoretisch-methodologischen Grundlagen verschiedener Paradigmen zugrunde, einerseits variationslinguistische Grundlagen, auf denen der SFB aufbaut (u.a. Auer 2005, Lenz 2003, Kehrein 2012), zweitens theoretische Konzepte der Korpus- und Computerlinguistik sowie der Digital Humanities (u.a. Chiarcos 2009, Ide/ Pustejovsky 2017), die für die Annotationen von Bedeutung sind, und drittens unterschiedliche linguistische Theoriekonzepte, die für die spezifischen Modellierungen der einzelnen un-

tersuchten Phänomene auf den verschiedenen Systemebenen wichtig sind.

Die Dissertation wird als kumulative Dissertation geschrieben und strebt 5 Forschungspapers an (s.o.). Die Papers werden zuvor auf wissenschaftlichen Fachtagungen diskutiert. Dabei reichen die verschiedenen Papers in unterschiedliche Schwerpunkte hinein.

Die Datenbasis bildet das Korpus des SFB, insbesondere auditive Aufnahmen aus verschiedenen Erhebungssettings wie beispielsweise Sprachproduktionsexperimente, Übersetzungsaufgaben, Interviews oder Freundesgesprächen (vgl. Lenz 2018) von verschiedenen Orten (urban und rural) in Österreich.

Bibliographie

Auer, Peter (2005): Europe's sociolinguistic unity, or. A typology of European dialect/standard constellations. In: *Perspectives on Variation, Sociolinguistic, Historical, Comparative*. Berlin, 7–42.

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*. Wien, 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.

Chiarcos, Christian / Dipper, Stefanie / Götze, Michael / Leser, Ulf / Lüdeling, Anke / Ritz, Julia / Stede, Manfred (2009): A flexible framework for integrating annotations from different tools and tagsets. In: *Traitement Automatique des Langues*, 49 (2), 271–291.

Ide, Nancy / Pustejovsky, James (Hg.) (2017): *Handbook of Linguistic Annotation*. Dordrecht.

Kehrein, Roland (2012): *Regionalsprachliche Spektren im Raum* – Zur linguistischen Struktur der Vertikale (= Zeitschrift für Dialektologie und Linguistik. Beihefte 152). Stuttgart.

Lenz, Alexandra N. (2018): The Special Research Programme „German in Austria. Variation – Contact – Perception“. In: Ammon, Ulrich / Costa, Marcella (eds.): *Sprachwahl im Tourismus – mit Schwerpunkt Europa. Language Choice in Tourism – Focus on Europe. Choix de langues dans le tourisme – focus sur l'Europe*. Berlin / Boston: De Gruyter (Yearbook Sociolinguistica 32) 269–277.

Lenz, Alexandra N. (2003): *Struktur und Dynamik des Substandards.* Eine Studie zum Westmitteldeutschen (Wittlich/Eifel) (=Zeitschrift für Dialektologie und Linguistik. Beihefte 125). Stuttgart.

„Ein lebendiges psychologisches Parlament“. Lazarus‘ und Steinthals Zeitschrift für Völkerpsychologie und Sprachwissenschaft.

Reiners, Stefan
stefan.reiners@rub.de
Ruhr-Universität Bochum, Deutschland

Begründung des Vorhabens

Die Völkerpsychologie Lazarus‘ und Steinthals stellt in der zweiten Hälfte des 19. Jahrhunderts die erste Psychologie – wenn nicht gar die erste empirische Wissenschaft überhaupt – dar, die den Menschen als soziales Wesen fasst. Entgegen der psycho-physisch reduktionistischen Theorien, die zur selben Zeit entstehen,[1] betont die Völkerpsychologie den sozialen Aspekt des menschlichen Denkens und Handelns, indem sie geteilte kulturelle Inhalte und Strukturen erforscht. Sie kann so als Grundstein der Soziologie, Sozialpsychologie sowie Ethnologie (vgl. Jahoda 2007, Köhnke 2003) und als zukunftsweisend für sämtliche Kulturwissenschaften gelten (vgl. Kalmar 1987).

Gegenwärtig meist als historische Anekdote abgetan, (vgl.: etwa Eckardt 1997, Lück / Guski-Leinwand 2014) bedienten sich die Völkerpsychologen des Mediums Zeitschrift allerdings auf innovative Weise. Die Zeitschrift für Völkerpsychologie wird als „lebendiges psychologisches Parlament“ begründet: Vielerlei Perspektiven und Methoden fließen ein, sodass sich die Wissenschaft im Forschungsprozess selbst konstituiert.

Die grundlegendste Frage: „Was ist Völkerpsychologie?“ kann also nicht mit Verweis auf die einleitenden programmatischen Texte dieser Zeitschrift beantwortet werden. Vielmehr muss der gesamte Forschungsprozess, d.h. die gesamte Zeitschrift, in den Blick genommen werden. Dies ist mit traditionellen Verfahren des Close Readings (CR) bei einem Korpus von ca. 2,9 Millionen Wörtern in ca. 400 Aufsätzen, erschienen über 30 Jahre, nicht zufriedenstellend zu bewältigen. Der Gegenstand macht damit Methoden des Distant Readings (DR) erforderlich.

Methode

Als methodisches Vorbild kann die kürzlich veröffentlichte Studie „What Is This Thing Called Philosophy of Science?” (Malaterre et al. 2019) dienen: Sie stellt eine Pionierleistung im Feld der Philosophie dar. Hier wurden über 80 Jahrgänge einer Zeitschrift mithilfe von Topic-Modeling-Methoden analysiert und die Ergebnisse der Analyse mit dem bis-

herigen Kenntnisstand über die historische Entwicklung der Disziplin verglichen.

Da es aber keinerlei Forschung zur historischen Entwicklung oder zum Prozess der Konstituierung der Völkerpsychologie gibt, dient das DR hier der Exploration. Dafür erscheint Topic Modeling (z.B. mit MALLET, McCallum 2002), gegebenenfalls im Zusammenspiel mit Kollokationsanalysen (z.B. mit #LancsBox, Brezina et al. 2019), als adäquate Lösung: Ziel ist es, die Entwicklungsmuster der Wissenschaft Völkerpsychologie abzubilden und dabei auch Bruchstellen oder andere Auffälligkeiten für das CR zu identifizieren, indem die Verteilung ausgewählter Topics auf das gesamte Korpus der Zeitschrift analysiert wird. Die genannten Methoden fügen sich dabei gut in das philosophische Paradigma ein, da sie eine differenzierte strukturelle Begriffsanalyse ermöglichen. Grundsätzlich muss natürlich diskutiert werden, inwieweit die dabei generierten Topics einer philosophischen Begriffsanalyse gerecht werden, bzw. inwiefern die Topics semantischen Themen und Begriffsnetzen vergleichbar sind: Underwood merkt an, dass dies bei *non-ficiton* eher der Fall ist als bei *fiction* (vgl.: Underwood 2012), während Jockers Preprocessing-Schritte empfiehlt, um dem näherzukommen. (vgl.: Jockers 2013)

Die Rohdaten der ersten zehn Bände der Zeitschrift liegen dabei als OCR-Scans vor, welche die Bayrische Staatsbibliothek zur Verfügung gestellt hat.[2] Diese wurden aufbereitet[3] und nach UTF-8 kodiert. Die Bände elf bis zwanzig liegen als Printmedien in guter Qualität vor und werden im Verlaufe des Vorhabens mit OCR verarbeitet werden.

Aktuell stellen sich Fragen nach der Weiterverarbeitung und Visualisierung der Topic Models sowie nach der Segmentierung der Rohdaten. Grundsätzlich soll sich jedoch eines Mixed-Methods-Ansatz bedient werden: Dem DR geht ein CR der programmatischen Aufsätze voraus, dessen Ergebnis wiederum zur Diskussion des DR genutzt wird. Schließlich soll das DR helfen, Entwicklungen und Umbrüche zu identifizieren, die abschließend als Zusammenschau einem datenbasierten CR unterzogen werden. Da es sich um eine philosophische Arbeit handelt, wird der gesamte Forschungsprozess ständig wissenschaftstheoretisch-kritisch reflektiert sowie abschließend diskutiert werden.

Gliederung und Stand der Arbeit

Das Dissertationsprojekt gliedert sich wie folgt:

(1) *Historische Analyse* der Völkerpsychologie als Psychologie des 19. Jahrhunderts sowie als erster Versuch einer Psychologie als Leitwissenschaft auf Basis eines weit gefassten, empirischen Geist-Begriffs.

(2) *Wissenschaftstheoretische Analyse* der programmatischen Aufsätze: Position eines gemäßigten methodischen Materialismus, verbunden mit einem Methoden-Relativismus, was eine Art Historical Turn in der Psychologie darstellt.

(3) *Medienphilologische Analyse* der Zeitschrift für Völkerpsychologie und Sprachwissenschaft als mediale Realisierung einer öffentlichen (sozial)psychologischen Diskussionsplattform.

(4) *Distant Reading* der Zeitschrift für Völkerpsychologie; Deutung vor dem Hintergrund der vorausgehenden Analysen.

(5) *Datenbasiertes Close Reading* als Zusammenschau.

Die Bearbeitung der Punkte (1) und (2) ist in wichtigen Grundzügen bereits geleistet und als Aufsatz im Erscheinen (Reiners 2020). Ein Aufsatz zu Punkt (3) wurde im Januar eingereicht.

Das vorliegende Projekt stellt einen interessanten Diskussionsgegenstand für das Kolloquium dar, weil es Fragestellungen der Medienwissenschaften, philosophischen Wissenschaftstheorie und Geschichtswissenschaften mit Methoden der Digital Humanities verbindet.

Fußnoten

1. Etwa die Psychophysik nach Fechner, aber auch die Ansätze Hermann von Helmholtz' und später Ebbinghaus' und Wundts.
2. S. https://opacplus.bsb-muenchen.de/title/BV002529202
3. Vom Rauschen befreit, Schaft-S durch gerundetes S ersetzt, Diphtong- bzw. Ligaturschreibung der Umlaute durch Umlautbuchstaben (‚ae'-‚ä') ersetzt, Silbentrennung aufgehoben usw. (Band eins bis sieben sind in Frakturschrift gedruckt).

Raise your voice! - Über den Zusammenhang zwischen Lautstärkemerkmalen in literarischen Prosatexten und der Emanzipation der Frau von 1848 bis 1920

Guhr, Svenja
Guhr@linglit.tu-darmstadt.de
TU Darmstadt, Deutschland

Mein im Rahmen des *Doctoral Consortiums* vorzustellendes Dissertationsvorhaben befindet sich noch in seinen Anfängen. Verortet als interdisziplinäres Projekt berührt es die Forschungsbereiche der Digital Philology spezialisiert auf neuere deutsche Literaturwissenschaft, Bereiche der historischen Gender Studies mit Fokus auf weibliche Emanzipationsprozesse im 19. und beginnenden 20. Jahrhundert sowie Bereiche der deutschen Linguistik. Mein Projekt setzt sich mit der Analyse von Lautstärkemerkmalen in deutschsprachigen literarischen Prosatexten von 1848 bis 1920 im Kontext der Emanzipation der Frau auseinander.

Haupthypothese

Die Untersuchungen bauen auf der Hypothese auf, dass in der Mitte des 19. Jahrhunderts Frauenfiguren in literarischen Prosatexten prozentual weniger, kürzere und leisere Redebeiträge zugeschrieben werden als männlichen Figuren, was sich jedoch mit der ansteigenden Emanzipation der Frau verändert. Das Projekt zielt darauf herauszufinden, ob sich Frauen-

und Männerfiguren im Verlauf der betrachteten Zeitperiode in ihrer Anzahl an Redebeiträgen und ihrer Lautstärke annähern. Die steigende Lautstärke zeichnet sich dabei durch eine „lautere" Beschreibung von Redebeiträgen aus, die u.a. durch als „lauter" wahrnehmbare redeeinleitende Verben gekennzeichnet sind.

Lautstärke als (audio-)narratologisches Element

Lautstärke wird dabei als ein narratologisches Element betrachtet, das in der Literaturwissenschaft bisher nur wenig Aufmerksamkeit erhalten hat. Der Umgang mit Erzählformen und Diskursen als ein wichtiges Kriterium der Narratologie fand seine Erweiterung durch das neue Forschungsfeld der Audionarratologie. Diese widmet sich u.a. der Relation zwischen Narrativen und den in der Lesevorstellung bei der stillen Lektüre erlebten Geräuschen, Tönen und Dynamik (Mildorf / Kinzel 2016; Kuzmičová 2013). Literarische Texte beinhalten neben Beschreibungen von natürlichen und industriellen Geräuschen (z.B. Natur- und Maschinengeräusche) auch Wiedergaben von Figurenrede. Insbesondere in Prosa ordnen Autoren und Autorinnen (i. F. generisches Femininum) ihren Figuren durch beschreibende Einleitungen von Redebeiträgen Stimmen zu, die in den Gedanken von Rezipientinnen wahrzunehmen sind. Ein neuer Ansatz der Figurenanalyse findet in der Betrachtung von Tönen, Lautstärke und Stimmvolumen in literarischen Prosatexten Anwendung. Vor allem die redeeinleitenden Verben, die direkte wie indirekte Redebeiträge einleiten und somit die Art und Weise der Figurenrede beschreiben, ermöglichen den Rezipientinnen die Wahrnehmung von Figurenstimmen und -lautstärke, z.B. ob eine Figur schreit oder flüstert. In Anlehnung an Hunt (2017), die in ihrer Studie ein Korpus aus anglophoner Kinder- und Jugendliteratur auf stereotypische Rollendarstellungen untersuchte, die sie anhand der „gendered nature" von redeeinleitenden Verben herausstellte, wird auch in meinem Forschungsprojekt eine genderdifferenzierte Betrachtung dieser Verbgruppe vorgenommen. In Hunts Ausführungen stützt sie sich auf Caldas-Coulthards (1992) Unterscheidung von redeeinleitenden Verben in neutrale (z.B. *sagen*) und illokutive (z.B. *befehlen*) Verben, wobei Hunt herausfand, dass in ihrem Untersuchungskorpus männlichen Figuren durchschnittlich eher redeeinleitende Verben zugeordnet wurden, die höhere Lautstärke und Macht (z.B. *brülllen*) widerspiegeln, während Frauenfiguren „triviale Emotionsäußerungen, Schwäche und hohe Tonlagen" (Hunt 2017: 2; Übersetzung modifiziert) (z.B. *jammern*, *kreischen*) zugeschrieben würden (Hunt 2017).

Ziel der Untersuchung und Subhypothesen

Ziel der Lautstärkeuntersuchung ist es herauszufinden, ob es einen Zusammenhang zwischen der beschriebenen Sprechweise weiblicher Prosafiguren und der ansteigenden Emanzipation der Frau in der deutschsprachigen Gesellschaft ab der zweiten Hälfte des 19. Jahrhunderts bis zum Erhalt des deutschen Frauenwahlrechts 1919 gibt (erweitert um das Jahr 1920, damit die Auswirkungen der Einführung des Frauenwahlrechtes mit aufgenommen werden).

Weitere Unterhypothesen beschäftigen sich mit dem Zusammenhang zwischen der beschriebenen Lautstärke einer Frauenfigur mit ihrem Bildungsstand sowie ihrer gesellschaftlichen Stellung. Weiterhin wird untersucht, ob Frauenfiguren in der Öffentlichkeit leiser beschrieben werden als im privaten Raum (vgl. Howe 2000). Darüber hinaus soll herausgestellt werden, ob Frauenfiguren in Anwesenheit von Männerfiguren leiser dargestellt werden als in der alleinigen Gesellschaft von Frauenfiguren, wobei untersucht wird, ob die plötzliche Anwesenheit auch nur einer Männerfigur in einem weiblichen Beisammensein die Art und Weise, in der Frauenfiguren miteinander sprechen, beeinflusst und ob diese Verhaltensänderung von der Beziehung der anwesenden männlichen Figur zu den Frauenfiguren (z.B. Vater, Bruder, Cousin, Fremder, Arbeitgeber, etc.) abhängt.

Korpusaufbau

Die Studie basiert auf der Analyse eines deutschsprachigen Korpus bestehend aus literarischen Prosatexten (Ziel: ca. 500). Das betrachtete thematische Korpus (vgl. Baker 2007: 26, Gür-Şeker 2014: 585) befindet sich aktuell (Stand: Januar 2020) noch in der Erstellungsphase, wobei auch auf existierende und teilweise bereits annotierte Korpora wie z.B. auf das Redewiedergabekorpus der Kooperation zwischen dem Leibniz-Institut für Deutsche Sprache, Mannheim und der Universität Würzburg (Brunner et al.) zurückgegriffen werden wird. Die strömungsübergreifend deutschsprachigen Prosatexte werden nach den folgenden Kriterien ausgewählt: deutschsprachig, Zeit der Publikation zwischen 1848 und 1920, *histoire* (vgl. Genette 1972) spielt im zentraleuropäisch-deutschsprachigen Raum nach 1848, min. 20% der ausgewählten Texte von Autorinnen (Ziel), Vorkommen von Hoch- und Trivialliteratur.

Erste Probeanalysen und Onlineumfrage

Für einen ersten Analyseansatz wurde ein Probekorpus erstellt, das 80 deutschsprachige Prosatexte umfasst und in zwei vergleichbare Subkorpora unterteilt wurde (2x 40 Prosatexte). Bei der Erstellung wurden die Texte nach den zuvor genannten Kriterien ausgewählt, wobei der Fokus auf die zwei Zeiträume 1865-75 und 1885-95 gelegt wurde, in denen jeweils eine überregional bedeutende Frauenrechtsaktion stattfand (Richards 2004).

Das Probekorpus diente als Grundlage zur Entwicklung einer regelbasierten Methode, mit deren Hilfe redeeinleitende Verben sowie die sie umgebenden Adjektive, Adverbien (z.B. *sagte mit lauter Stimme, sagte leise*) und Namen der sprechenden Figuren extrahiert werden. Der angewandte Algorithmus gründet sich auf den sprachspezifisch morphologischen und syntaktischen Eigenschaften der deutschen redeeinleitenden Verben hinsichtlich ihrer Konstruktion und bevorzugten Anwendung in Vergangenheits- und Gegenwartsform sowie ihrer zum Teil durch Interpunktion (z.B. Anführungszeichen) aufgezeigten Nähe zur in-/direkten Rede. Darauf aufbauend wurden Lautstärkeprofile der sprechenden Figuren erstellt, die anschließend als Grundlage für einen genderorientierten Vergleich von männlichen und weiblichen Figurenbeiträgen

und Sprechweisen in literarischen Prosatexten dienen. Die Lautstärkewerte zur Erstellung der Lautstärkefigurenprofile wurden den Ergebnissen einer zuvor durchgeführten Onlineumfrage (im Sommer 2018) zu 20 redeeinleitenden Verben entnommen. In der Umfrage wurden 45 deutsche Muttersprachlerinnen gebeten, in einem Vergleich desselben Satzes, der von zwei verschiedenen redeeinleitenden Verben eingeführt wurde, das jeweils lautere redeeinleitende Verb auszuwählen (Umfragebeispiel: *Marie schreit: Die Sonne scheint.* vs. *Marie flüstert: Die Sonne scheint.*). Durch die Umfrage ergaben sich Informationen zur Unterschiedlichkeit der durch redeeinleitende Verben erzeugten dynamischen Lesewahrnehmung von Redebeiträgen durch die Rezipientinnen. In Abgleich der untersuchten Verbauswahl konnte ein erstes Lautstärkediagramm mit Werten erstellt werden, die sich aus den erhaltenen Dynamikabstufungen der Surveyergebnisse ergaben und schließlich zur Erstellung erster Lautstärkefigurenprofile verwendet wurden.

Ausblick

Im Laufe des Dissertationsvorhabens sollen die Untersuchungen zur Lautstärkewertzuweisung zu redeeinleitenden Verben in einem umfangreicheren und wissenschaftlich fundierten Verfahren wiederholt werden. Zudem werden methodische Herausforderungen wie die Auflösung von Koreferenzen und Anaphern sowie die Erkennung von (unregelmäßigen) Redebeiträgen, Szenengrenzen und Figurenkonstellationen einen großen Bestandteil meiner Forschung einnehmen.

Bibliographie

Baker, Paul (2010): *Sociolinguistics and corpus linguistics* . Edinburgh: Edinburgh University Press.

Brunner, Annelen / Engelberg, Stefan / Tu, Ngoc Duyen Tanja / Jannidis, Fotis / Weimer, Lukas (2019): Redewiedergabe-Projekt. Mannheim / Würzburg: Leibniz-Institut für Deutsche Sprache, Universität Würzburg https://github.com/redewiedergabe/corpus [letzter Zugriff 05. Januar 2020].

Caldas-Coulthard, Carmen Rosa (1992): „Reporting Speech in Narrative Discourse: Stylistic and ideological implications", in: *Revista Ilha do Desterro* 27: 67–82 https://doaj.org/article/346291b81422492392f7c48c0635f7a2 [letzter Zugriff 05. Januar 2020].

Catling, Jo (ed.) (2000): *A History of Women's Writing in Germany, Austria and Switzerland.* Cambridge: Cambridge University Press.

Genette, Gérard (1969): *Figures. Essais.* Paris: Editions du Seuil.

Gür-Şeker, Derya (2014): „Zur Verwendung von Korpora in der Diskurslinguistik", in: *Diskursforschung. Ein interdisziplinäres Handbuch. DiskursNetz* 1: 583–603. Bielefeld: Transcript Verlag.

Howe, Patricia (2000): „Women's writing 1830-1890", in: Catling, Jo (ed.): *A History of Women's Writing in Germany, Austria and Switzerland.* Cambridge: Cambridge University Press, 87–103.

Hunt, Sally (2017): „Boast and bellow, giggle or chatter: gender and verbs of speech in children's fiction", in: *9th International Corpus Linguistics Conference,* Birmingham: University of Birmingham https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper306.pdf [letzter Zugriff 05. Januar 2020].

Kuzmičová, Anežka (2013): „Outer vs. Inner Reverberations: Verbal Auditory Imagery and Meaning-Making in Literary Narrative", *Journal of Literary Theory* 7: 111–134 https://philpapers.org/archive/KUZOVI.pdf [letzter Zugriff 05. Januar 2020].

Mildorf, Jarmila / Kinzel, Till (2016): „Audionarratology: Prolegomena to a Research Paradigm Exploring Sound and Narrative", in: *Audionarratology. Interfaces of Sound and Narrative.* Berlin / Boston: De Gruyter.

Richards, Anna (2004): *The Wasting Heroine in German Fiction by Women 1770-1914,* Oxford: Oxford University Press.

Schmid, Wolf (2014): *Elemente der Narratologie.* Berlin: De Gruyter.

Posterpräsentationen

Abstract Enhancement. Potentiale der DHd-Konferenzabstracts als Daten/Publikation

Steyer, Timo
steyer@hab.de
Forschungsverbund Marbach Weimar Wolfenbüttel, Deutschland

Andorfer, Peter
Peter.Andorfer@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Cremer, Fabian
Cremer@MaxWeberStiftung.de
Max Weber Stiftung; Leibniz-Institut für Europäische Geschichte

Das Abstract im Kontext

Die Autoren haben auf der DHd2019 einen Hackathon zum Book of Abstracts durchgeführt, in dessen Rahmen sich die Teilnehmenden nicht nur mit der digitalen Publikation, sondern auch mit der Normierung der biobibliographischen Angaben und den Potentialen einer inhaltlichen Analyse der Abstracts auseinander gesetzt haben. (Andorfer et al. 2019) Dieser Beitrag greift die Erkenntnisse der Veranstaltung auf und soll sowohl die konzeptuelle Auseinandersetzung als auch die konkrete Implementierung weiterführen sowie die im Nachgang des Workshops erfolgten Arbeiten präsentieren. Dabei werden im Sinne des Konferenzthemas insbesondere experimentelle Ansätze hervorgehoben. Zwei Aspekte der Konferenzabstracts stehen im Fokus des Beitrags: 1.) das Abstract als eigenständige und reputierliche Publikation und 2.) die Abstracts als Datenquelle selbstreflexiver Untersuchungsansätze in den DH. Die wissenschaftliche Relevanz der Book of Abstracts der DHd-Jahrestagung bekräftigte zuletzt noch einmal Sahle in seiner Einführung des letzten Konferenzbandes (Sahle 2019, S.): „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."

Das Abstract als Publikation

Der Anspruch an die Veröffentlichung der Konferenzbeiträge wird in dem bereits 2018 konstatierten „Status einer wissenschaftlich nutzbaren Publikation" (Vogeler 2018) deutlich und mit dem letzten Band der DHd 2019 noch einmal unterstrichen (Sahle 2019): „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." Auf dem Weg zu einer digitalen Publikation, die sowohl informationswissenschaftliche Standards erfüllt als auch informationstechnologische Potentiale ausreizt, ergeben sich zum jetzigen Stand noch viel Raum für Entwicklung der Konferenzbeiträge (Cremer 2018). Die Book of Abstracts werden als Gesamtband in einer Druckfassung sowie als PDF publiziert. Daneben werden einzelne Beiträge von den Vortragenden auf verschiedenen Repositorien oder Webseiten unsystematisch veröffentlicht, darunter einzelne Abstracts, Poster, Präsentationen oder zugrundeliegende Daten.[1] Der Beitrag evaluiert die Möglichkeiten und Voraussetzungen für eine eigenständige Publikation der einzelnen Abstracts mit persistenter Speicherung, zitationsfähiger Adressierung, bibliografischer Erfassung und multipler Repräsentationsform (PDF, HTML, TEI). Dabei werden auch dezentrale (z.B. Zenodo-Community) und zentrale Ansätze (z.B. zentrale Redaktion, eigene Infrastruktur) verglichen. Gegenüber traditionellen Formaten und Infrastrukturkomponenten im Sinne einer reputierlichen und zitierfähigen Publikationsform sollen auch experimentelle Repräsentationsformen betrachtet werden, um die vorhandenen Spielräume digitaler Publikationen auszuloten. Als Ausgangspunkt ist hier die im Rahmen des Hackathons entwickelte Präsentationsschicht zu nennen.[2]

Das Abstract als Daten

Die Konferenzabstracts als TEI-basierte Veröffentlichungen demonstrieren ihr Potential als Untersuchungsgegenstand innerhalb des eigenen Faches (Sahle/Henny-Krahmer 2018; Hannesschläger/Andorfer 2018; Hoenen 2019; Kiefer 2019). Ein Desiderat der Untersuchungen bis dato ist die Betrachtung und Auswertung der in den Abstracts zitierten Literatur. Die Bibliographie wissenschaftlicher Artikel dient in der geisteswissenschaftlichen Forschung neben dem Nachweis der zitierten Literatur auch als Ressource für Recherche und Kontextualisierung (Andorfer, DWP 14, S. 24-25) sowie als Datenquelle für die Analyse von Publikations- und Zitationspraktiken (Nyhan/Duke-Williams 2014). Gerade in den Digital Humanities eröffnen sich durch die Verbindung mit Methoden der Netzwerkanalyse neue Untersuchungsansätze (Gao et al. 2018). Im Rahmen des Hackathons wurden die eingereichten Abstracts über Skripte automatisiert mit zusätzlichen Informationen angereichert sowie über manuelle Arbeiten in ihren Metadaten vereinheitlicht.[3] Die bibliographischen Angaben in den Abstracts lagen jedoch in zu heterogenen Formen vor, so dass Auswertungen und Visualisierungen nicht möglich waren. Für das Poster werden diese Daten mit Unterstützung der DHd-AG Digitales Publizieren vereinheitlicht und in der Folge durch die Autoren in ersten Analyseergebnissen und Visualisierungen ausgewertet. Die aufgezeigten Potentiale ließen sich zudem multiplizieren, wenn auch die Konferenzabstracts früherer und folgender DHd-Tagungen aufbereitet werden können, um so auch Entwicklungen und Tendenzen eruieren zu können.

Das Abstract in der Diskussion

Viele Jahre nach Christines Borgmans “Call to Action for the Humanities” (Borgman 2010), der auch das digitale Publizieren jenseits der simplen Konversion der Papiermedien in das PDF-Format inkludierte, werden auch in den Digital Humanities die Möglichkeiten nicht ausgeschöpft und traditionelle Praktiken gepflegt (Kaden/Kleineberg 2017) – von dem Wechsel einer layoutbasierten zu einer strukturbasierten Publikationstechnik ganz zu schweigen (Stäcker 2013). Die DHd-Konferenzabstracts bergen dabei das Potential dieses Paradigma zu durchbrechen: die XML-basierte Einreichung, die datengestützten Analysemethoden und selbstreflexiven Ansätze des Faches, die technische Expertise der Einreichenden, die Kürze der Beiträge und die enge Vernetzung mit Infrastruktureinrichtungen. Das Poster soll auf die bisher erfolgten Arbeiten und die erzielten Ergebnisse aufmerksam machen sowie vor Ort die Diskussion um Möglichkeiten und Ressourcen sowie Relevanz und Reputation einer „erweiterten Publikation“ der DHd-Abstracts weiterführen. Die Autoren werden im Vorfeld der DHd2020 mit dem Organisationskomitee zur Anreicherung der diesjährigen Abstracts sowie Nutzung der HTML-Präsentationsschicht in Kontakt treten.

Fußnoten

1. DHd-Community bei Zenodo: https://zenodo.org/communities/dhd
2. DHd 2019 Book of Abstracts Hackathon: https://dhd-boas-app.acdh-dev.oeaw.ac.at
3. Die aufbereiteten Daten sind zu finden unter: https://github.com/csae8092/dhd-boas-data.

Bibliographie

Andorfer, Peter (2015): *Forschungsdaten in den (digitalen) Geisteswissenschaften. Versuch einer Konkretisierung,* DARIAH-DE Working Papers, 14 [letzter Zugriff 30.08.2019].

Andorfer, Peter / Cremer, Fabian / Steyer, Timo (2019): DHd 2019 Book of Abstracts Hackathon, in: *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts* https://doi.org/10.20375/0000-000B-D512-0[letzter Zugriff 30.08.2019].

Borgman, Christine L. (2010): The Digital Future is Now: A Call to Action for the Humanities, in: *Digital Humanities Quarterly* 003, Nr. 4.

Cremer, Fabian (2018): Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?, in: *Blog. Digitale Redaktion (blog)* https://editorial.hypotheses.org/113[letzter Zugriff 30.08.2019].

Gao, J. / Duke-Williams, O. / Mahony, S. / Ramdarshan Bold, M. / Nyhan, J. (2017):The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis, in: *Digital Humanities 2017* http://discovery.ucl.ac.uk/id/eprint/10052270 [letzter Zugriff 30.08.2019].

Hall, Mark (2019): DH is the Study of dead Dudes, in: *Hd 2019: multimedial & multimodal Konferenzabstracts*: 111-113 https://doi.org/10.5281/zenodo.2600812[letzter Zugriff 30.08.2019].

Hannesschläger, Vanessa / Andorfer, Peter (2018): *Menschen gendern? Datenmodellierung zur Erhebung von Geschlechterverteilung am Beispiel der TEI2016 Abstracts App* https://doi.org/10.5281/zenodo.1182576[letzter Zugriff 30.08.2019].

Henny-Krahmer, Ulrike / Sahle, Patrick (2019): Einreichungen zur DHd 2018, in: *DHd-Blog* (blog), 29. März 2019 https://dhd-blog.org/?p=9001[letzter Zugriff 30.08.2019].

Hoenen, Armin (2019): Einreichungen zur DHd 2019 II, in: *DHd-Blog* (blog), 29. März 2019 https://dhd-blog.org/?p=11418[letzter Zugriff 30.08.2019].

Kaden, Ben / Kleineberg, Michael (2019): Zur Situation des digitalen geisteswissenschaftlichen Publizierens – Erfahrungen aus dem DFG-Projekt ‚Future Publications in den Humanities‘, in: *Bibliothek Forschung und Praxis* 41, Nr. 1 (2017): 7–14 https://doi.org/10.1515/bfp-2017-0009[letzter Zugriff 30.08.2019].

Kiefer, Katharina (2019): Einreichungen zur DHd 2019, in: *DHd-Blog* (blog), 29. März 2019 https://dhd-blog.org/?p=11358[letzter Zugriff 30.08.2019].

Nyhan, Julianne / Duke-Williams, Oliver (2014): Joint and Multi-Authored Publication Patterns in the Digital Humanities, in: *Literary and Linguistic Computing* 29, Nr. 3 (1. September 2014) https://doi.org/10.1093/llc/fqu018[letzter Zugriff 30.08.2019].

Sahle, Patrick (ed.) (2019): *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main: Zenodo, 2019 https://doi.org/10.5281/zenodo.2596095[letzter Zugriff 30.08.2019].

Stäcker, Thomas (2013): Wie schreibt man Digital Humanities richtig?, in: *Bibliotheksdienst* 47, Nr. 1 https://doi.org/10.1515/bd-2013-0005[letzter Zugriff 30.08.2019].

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

A Linked Open Data Platform for Historical Geographic Data

Görz, Günther
guenther.goerz@fau.de
FAU Erlangen-Nürnberg, Informatik, Deutschland;
Bibliotheca Hertziana, Rom, Italien

Seidl, Chiara
chiara.seidl0510@gmail.com
FAU Erlangen-Nürnberg, Informatik, Deutschland;
Bibliotheca Hertziana, Rom, Italien

Thiering, Martin
pd_dr_Martin_Thiering@posteo.de
TU Berlin, Deutschland; Bibliotheca Hertziana, Rom, Italien

The goal of Bibliotheca Hertziana's project "Historical spaces in texts and maps" is to investigate the relations between historic geographical texts and maps to reconstruct a historical understanding of space and the knowledge associa-

ted with it. Starting with a cognitive-semantic analysis of Flavio Biondo's "Italia Illustrata" (1474), first of all, toponyms, place descriptions and spatial relations are annotated in the text and Renaissance maps. Our contribution to Spatial Humanities is based on the conviction that all maps are cognitive maps, depicting culture-specific spatial knowledge and practices (Goerz et al., 2018, 2019).

In general, our research combines cognitive-semantic parameters such as toponyms, landmarks, spatial frames of reference, geometric relations, gestalt principles and different perspectives with computational linguistic analysis (Thiering, 2015). We designed a workflow comprising the steps of transcription, annotation, geographic verification, export and ontology-based semantic enrichment of these data, finally stored and published as Linked Open Data. We use Recogito (https://recogito.pelagios.org, 15.09.2019) as our main tool for static annotations of places and persons/peoples in text and maps. Toponyms are georeferenced with gazetteers, in our case primarily with Pleiades (https://pleiades.stoa.org, 15.09.2019), and the annotations can be exported in various formats, in particular, CSV, GeoJSON, and KML. Spatial relations in texts are annotated in terms of the cognitive-semantic parameters by means of the brat tool. These annotation data are mapped into triples encoding cognitive parameters, primarily in "figure-spatial_relation-ground" constructions. Furthermore, dependency parsing (http://ufal.mff.cuni.cz/udpipe, 15.09.2019) has been applied to the text for comparison. To achieve a generic semantic level for linguistic and map-related annotations, we perform a transition to an ontology-based representation. For this purpose, we defined a domain ontology *hmap* for historical maps and geographical texts based on the event-centered CIDOC Conceptual Reference Model (CRM, ISO standard 21127) and its spatio-temporal extension CRMgeo in OWL-DL (http://erlangen-crm.org, 15.09.2019). Using the CRM opens up a wide spectrum of interoperability and linking to many web resources. The domain ontology *hmap* for the description of historical maps and their content offers a framework for the general metadata of maps and geographical texts as well as for descriptions of their content.

As Linked Data platform we chose the Virtual Research Environment WissKI (Scholz et al., 2016; http://wiss-ki.eu, 15.09.2019), a semantic database extension of the CMS Drupal, in which we defined our data model in terms of so-called ontology paths. These are sequences of triples built from entities and properties of the ontology. As an example, in a map production event (*hmap:M9_Map_Production*) there is an actor, the Creator, defined by

```
hmap:M28_Map --> hmap:A3i_was_produced_by
--> hmap:M9_Map_Production --> hmap:A4_car-
ried_out_by_map_author --> hmap:M1_Map_Aut-
hor --> ecrm:P131_is_identified_by
--> ecrm:E82_Actor_Appellation.
```

For each map we may have several images, in which depicted objects are annotated; so there is an analogous data model for images (*hmap:M34_Image*). What the image depicts, in our case annotated places, is specified by

```
hmap:M34_Image --> hmap:A43_depicts --> hma-
p:M3_Annotated_Place --> ecrm:P1_is_identi-
fied_by --> ecrm:E42_Identifier.
```

For each annotated place (*hmap:M3_Annotated_Place* is a subclass of *crmgeo:SP6_Declarative_Place*) where the (geographical) contents of the annotations are encoded in the columns of the CSV tables, each column is transformed into a component for which similar ontology paths are defined. The annotated place is linked to the image by

```
hmap:M3_Annotated_Place --> hmap:A43i_is_de-
picted by --> hmap:M34_Image
--> ecrm:P48_has_preferred_identifier -->
ecrm:E42_Identifier.
```

So, e.g., for *QUOTE_TRANSCRIPTION*, the path is

```
hmap:M3_Annotated_Place --> ecrm:P87_is_i-
dentified_by --> hmap:M42_Transcribe-
d_Place_Appellation
```

Each annotation, represented as a row in the table, has a unique ID (UUID) and refers, if geographically verified with a gazetteer (Pleiades), via a URL to a graph containing various information such as e.g. sources, archeological data, images, etc. In some maps, annotated places are additionally represented by a visual item (E36) such as a church, a tower, or a wall. There are also further data models for image series and works like map collections or atlases. From these paths, WissKI generates automatically input forms for map and text metadata and provides an interface for importing all table-formatted annotations and converting them into triples. Ontological enrichment of our data with CRM allows for a semantic interpretation of annotations such that, e.g., for each PlaceName, an instantiated CRM description in RDF/OWL triple format is generated and stored in a triple store. Using the semantically enriched geo-information from text (and map) annotations as CRM instances, spatial entities ("figure", "ground") and relations obtained by spatial role labeling as "figure-spatial_relation-ground" triples can now be upgraded to this rich semantic level by linking data. Due to the fundamental underlying triple structure for all kinds of annotations, the data are immediately ready for publication as standardized Linked (Open) Data; WissKI provides a SPARQL query interface. These triple data constitute a huge knowledge graph; they are the "raw material" for further research steps, i.e. the exploration of the historical understanding of spaces and the associated knowledge. Interpretation of the data has just begun: Actually, a study of Biondo's spatial language – comprising the whole text for the first time – by Berthele and Thiering is being finalized (2020) and a comparative overview of the toponyms in the Latium book with different maps (traditional and "modern" Ptolemaic maps, genuine maps of Italy and portolans) as well as a representation of the hodological dimension of the text is being prepared.

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'', e-perimetron 13,4: 182-199.

Görz Günther / Seidl Chiara / Thiering, Martin (2019): Linked Biondo: Modelling Geographical Features in Renaissance Texts and Maps **.** In: Boutoura, Ch., Tsorlini, A., Livieratos, E. (Ed.): Proceedings 14th ICA Conference *Digital Approaches to Cartographic Heritage,* Thessaloniki, 8-10 May 2019.International Cartographic Association, Commission on Cartographic Heritage into the Digital. AUTH CartoGeoLab, ISSN 2459-3893, 124-140.

Michalsky, Tanja / Thiering, Martin (2020): Walking through history. An interdisciplinary approach to Flavio Biondo's spaces in the "Italia illustrata". Rome: Bibliotheca Hertziana.

Scholz, Martin / Merz, Dorian / Goerz, Günther (2016): Working with WissKI - A Virtual Research Environment for Object Documentation and Object-Based Research. Digital Humanities 2016, Conference Abstracts, Krakow, 11-16 July 2016. ISBN 978-83-942760-3-4, 944-945.

Thiering, Martin (2015): Spatial Semiotics and Spatial Mental Models: Figure-Ground Asymmetries in Language. Berlin: De Gruyter Mouton.

Anforderungen an das Forschungsdatenmanagement an einer mittelgroßen Universität und Konzeption einer prototypischen Lösung

Jegan, Robin
robin.jegan@uni-bamberg.de
Otto-Friedrich-Universität Bamberg, Deutschland

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Otto-Friedrich-Universität Bamberg, Deutschland

Henrich, Andreas
andreas.henrich@uni-bamberg.de
Otto-Friedrich-Universität Bamberg, Deutschland

Am Lehrstuhl für Medieninformatik der Otto-Friedrich-Universität Bamberg wird derzeit ein Prototyp zum Management von Forschungsdaten entwickelt (FDMS), der heterogene Daten, die momentan nicht digital katalogisiert werden, speichern, beschreiben, und veröffentlichen soll. Von besonderer Bedeutung sind hierbei die FAIR-Prinzipien, wonach wissenschaftliche Daten auffindbar (Findable), zugänglich (Accessible), interoperabel (Interoperable), und wiederverwendbar (Re-usable) sein sollten (Wilkinson 2016). Es existiert bereits ein Projekt zur Einführung eines Forschungsinformationssystems (FIS) an der Universitätsbibliothek Bamberg (Franke 2019), jedoch liegt dort der Fokus auf Publikation, Forschenden, und Projekten. Forschungsdaten werden hier bisher nur am Rande als Anhang bei der Publikation von Dissertationen und anderen Arbeiten adressiert. Weitere Forschungsdaten, die mit keiner Publikation zusammenhängen, werden hingegen in diesem FIS nicht miteinbezogen. Aus diesem Grund soll ein FDMS entwickelt werden, welches genau diese Art von Forschungsdaten ohne zugehörige Publikation, sowie auch alle anderen Forschungsdaten abdeckt.

Eine wichtige Anforderung und Herausforderung für das FDMS stellt die Heterogenität der Daten dar. Diese Heterogenität manifestiert sich durch die unterschiedlichen Fachrichtungen der Lehrstühle an der Universität Bamberg und dementsprechend in einer Vielzahl an Metadatenschemata und Datenformaten. Die Erschließung und weitere Verwendung dieser Metadaten ist von großer Bedeutung, da die effiziente Suche sowie weitere Dienste, wie Filterung oder Visualisierung, von umfangreichen Metadaten profitieren (Neuroth 2017) und zudem Fördergeber wie die Deutsche Forschungsgemeinschaft Anforderungen an ein Forschungsdatenmanagement stellen, das den FAIR-Prinzipien folgt.

Für die erste Umsetzung wurde DSpace verwendet (Smith 2003, Donohue 2018), eine Software zur Verwaltung von digitalen Forschungs- und Lehrmaterialien. DSpace wird in Bamberg auch zur Realisierung des FIS eingesetzt, dort jedoch in der Erweiterung als DSpace-CRIS (Donohue 2019). Während der Installation und Anpassung von DSpace an die Testdaten wurde eine Einschränkung und zugleich ein Nachteil von DSpace deutlich, welche die Nutzung dieser Software für den Zweck eines FDMS ungeeignet erscheinen lassen. Der Import von unterschiedlichen Metadatenschemata für Forschungsdaten wird in DSpace zwar unterstützt, jedoch ist dieser Import in der DSpace Standard-Installation nur mit den originalen Dublin-Core Elementen eingerichtet. Außerdem können keine hierarchischen Datenelemente erzeugt werden, weswegen der Import von Daten sowie deren Metadaten, welche in hierarchischen XML-Dateien vorliegen, verhindert wird.

Aufgrund dieser Einschränkung von DSpace, welche sich bei einer Vielzahl von Forschungsdaten als restriktiv darstellen würde und somit einer der Anforderungen an ein FAIR-basiertes FDMS widerspricht, wurde eine alternative Software benötigt und in Dataverse gefunden (Dataverse 2019a). Dataverse, als Projekt vom Harvard Institute for Quantitative Social Science, der Harvard Library, und anderen Partnern initiiert (Dataverse 2019b), ermöglicht die Modellierung der Metadaten in mehreren Varianten. Zum einen können über die Administrationsoberfläche der Dataverse Installation händisch Elemente der bestehenden Metadatenschemata angepasst werden. Zum anderen, und für das Umsetzungsszenario der Universität Bamberg passender, kann ein sogenannter Metadatenblock angelegt werden, mit dem ein komplett neues Metadatenschema für Datenobjekte in der Dataverse Installation zur Verfügung gestellt wird (Dataverse 2019c). Dieser selbst erzeugte Metadatenblock, in Form einer TSV (Tab-Separated Values) Datei, kann mithilfe eines Kommandozeilen Befehls importiert werden und steht daraufhin neben den in Dataverse bereits vorgefertigten Standard-Metadatenschemata zur Verfügung.

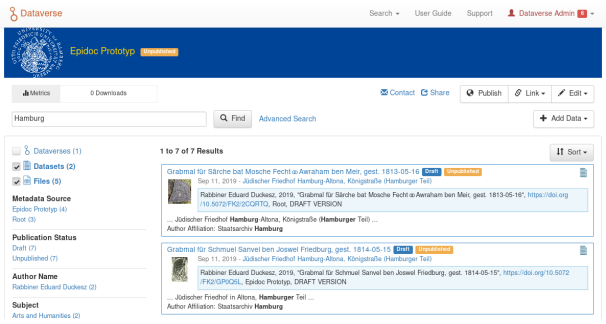


Abbildung 1: Oberfläche des FDMS-Prototyps mit einer Suchanfrage und der Ergebnisseite.

Als prototypische Anwendung wurden Bilder von Grabsteinen mit weiteren zugehörigen Forschungsprimärdaten verwendet, die mehrere Eigenschaften eines typischen Use Cases für das künftige FDMS abdecken. Diese Forschungsdaten liegen zum einen als Bilder vor, die ohne zugehörige Publikation bisher nicht in einem FIS abgespeichert werden können, zum anderen sind komplexe Metadaten vorhanden, welche in DSpace, wie oben beschrieben, nur eingeschränkt darzustellen, in Dataverse jedoch besser einzufügen sind. Die Testdaten für diesen Prototyp basieren auf dem epidat Projekt des Steinheim-Instituts (Steinheim-Institut 2019), die dem Datenbestand der Professur für Judaistik der Universität Bamberg ähneln, welcher ein konkretes künftiges Anwendungsszenario darstellt.

Auch für die Analyse der Metadatenelemente, die im FDMS verwendet werden, wurde das epidat Projekt als Testfall genutzt. Hierzu wurden die XML-Dateien, welche die Gräber einer Vielzahl von Friedhöfen in Deutschland mit Metadaten im Epidoc-TEI-Format (Elliott 2006-2017) beschreiben, analysiert und relevante Metadatenfelder identifiziert. Diese relevanten Metadatenelemente wurden als Basis für ein Dataverse im Kontext von Fotografien von Grabsteinen verwendet. Der Begriff Dataverse umfasst dabei nicht nur die Software Dataverse, sondern dient auch als Oberbegriff für eine Sammlung von Datensätzen (King 2007). Die Vorgehensweise zur Erstellung eines eigenen Metadatenschemas wurde verwendet, da das im epidat-Projekt verwendete Epidoc-Format sehr umfangreich und dementsprechend für ein initiales Anwendungsszenario zu mächtig scheint. Die Anzahl der Elemente für den Prototyp wurde daher eingeschränkt.

Die identifizierten Metadatenelemente, welche für die forschungsorientierte Beschreibung der Grabsteine notwendig sind, wurden mithilfe eines der oben erwähnten Metadatenblöcke modelliert. Das hiermit erstellte Metdatenschema „epitaph“ – betitelt in Anlehnung an den Fachbegriff für Grabinschriften – enthält 25 Elemente, die unter anderem den zugehörigen Friedhof, den Zustand des Grabmals, die Inschrift inklusive Übersetzung, und weitere Datenfelder umfassen. Weiterhin wurden hierarchische Beziehungen zwischen den Elementen aufrechterhalten, beispielsweise für die verstorbene Person, deren Todestag, und die erwähnten Verwandten dieser Person über die Datenfelder „person“, „personDeathdate“, und „personRelationship“.

In Kooperation mit dem Rechenzentrum und der Universitätsbibliothek wird aufbauend auf internen Diskussionen die Weiterentwicklung des in diesem Poster vorgestellten Prototyps geplant, welche in einer größeren Testphase 2020 vorgenommen werden soll. Die Anbindung an das bereits im Betrieb befindliche FIS soll ebenso wie die Integration der lokalen Shibboleth-Authentifizierungsverfahren dort umgesetzt werden.

Das Poster verdeutlicht die Anforderungen und Anstrengungen, die für die Konzeption eines FDMS an einer mittelgroßen Universität notwendig sind, sowie erste positive Ergebnisse. Weiterhin werden durch die Betrachtung zweier Softwarelösungen die Probleme in der Praxis, also in der Umsetzung eines derartigen Forschungsdatenmanagements, näher beleuchtet. Die Erfahrungen verdeutlichen wie wichtig eine systematische Anforderungsanalyse bei der Auswahl eines Systems zum Forschungsdatenmanagement ist. Hierbei sollten insbesondere Aspekte wie die Unterstützung unterschiedlicher Metadatenformate und deren hierarchische Ausprägungen berücksichtigt werden. Diese und weitere Erfahrungen werden am Poster geteilt und ausgetauscht.

Bibliographie

Dataverse (2019): The Dataverse Project: Open source research data repository software. Online: https://dataverse.org/ [letzter Zugriff: 19.12.2019]

Dataverse (2019): The Dataverse Project: About the project. Online: https://dataverse.org/about [letzter Zugriff: 20.12.2019]

Dataverse (2019): The Dataverse Project Admin Guide: Metadata Customization. Online: http://guides.dataverse.org/en/latest/admin/metadatacustomization.html [letzter Zugriff: 19.12.2019]

Donohue, Tim (2018): DSpace User FAQ. Online: https://wiki.lyrasis.org/display/DSPACE/User+FAQ [letzter Zugriff: 20.12.2019]

Donohue, Tim (2019): DSpace-CRIS Home. Online: https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home [letzter Zugriff: 19.12.2019]

Elliott, Tom / Bodard, Gabriel / Cayless, Hugh / et al. (2006-2017): EpiDoc: Epigraphic Documents in TEI XML. Online: https://sourceforge.net/p/epidoc/wiki/Home/ [letzter Zugriff: 19.12.2019]

Franke, Fabian (2019): Laufende Projekte unter Beteiligung der Universitätsbibliothek Bamberg. In: Projekte – Otto-Friedrich-Universtiät Bamberg. Online: https://www.uni-bamberg.de/ub/ueber-uns/projekte/ [letzter Zugriff: 19.12.2019]

King, Gary (2007): An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. In: *Sociological Methods & Research* 36(2): 173–199.

Neuroth, Heike (2017): Bibliothek, Archiv, Museum. In: *Digital Humanities:* 213–222.

Smith, MacKenzie / et al. (2003): "An Open Source Dynamic Digital Repository". In: *D-Lib Magazine* 9.1.

Steinheim-Institut (2019): Datenbank: Jüdische Grabsteinepigraphik. Online: http://www.steinheim-institut.de/cgi-bin/epidat [letzter Zugriff: 20.12.2019]

Wilkinson, Mark D., et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific data* 3.

Aufbau und Erfahrungen aus dem Digital Humanities Lab der Universität Erlangen-Nürnberg

Scholz, Martin
martin.scholz@fau.de
Universitätsbibliothek Erlangen-Nürnberg

Klusik-Eckert, Jacqueline
jacqueline.klusik@fau.de
IZdigital, Friedrich-Alexander-Universität Erlangen-Nürnberg

Die Aktivitäten in den Digital Humanities sind an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) historisch bedingt auf zahlreiche Lehrstühle über die fünf Fakultäten verteilt. Mit dem Interdisziplinären Zentrum für digitale Geistes- und Sozialwissenschaften (IZdigital) wurden diese 2014 erstmals lose organisiert. Darauf aufbauend wurde an der philosophischen Fakultät der fächerübergreifende Studiengang Digitale Geistes- und Sozialwissenschaften etabliert. Zwar wurde damit Studierenden der Geistes- und Sozialwissenschaften die Möglichkeit eröffnet, sich vertieft und gezielt mit digitalen Aspekten von Forschung, Arbeit und Gesellschaft auseinanderzusetzen. Doch bleibt abseits des Studiengangs für die breite Masse an Studierenden und Promovierenden die Aneignung von digitaler Kompetenz und der Austausch mit Gleichgesinnten weitgehend auf das Selbststudium und informellen Austausch beschränkt. Mit dem Digital Humanities Lab (DHLab) wollen die Universitätsbibliothek Erlangen-Nürnberg, das IZdigital und die Philosophische Fakultät gemeinsam diese Lücke füllen oder zumindest schmälern. Das Poster stellt den Aufbau und bisherige Erfahrungen aus dem Betrieb des DHLab vor.

Das Digital Humanities Lab versteht sich als Kombination aus Ort, Personen und Inhalten. Es möchte Studierende, Lehrende und Forschende gleichermaßen in ihren praktischen Belangen, Fragen und Problemen rund um das Thema Digital Humanities unterstützen und begreift sich als Dienstleister für die Geistes- und Sozialwissenschaften. Es verfolgt daher ein offenes Format, das Aspekte eines Helpdesks mit Schulungen, Vorträgen sowie Diskussions- und Netzwerkmöglichkeiten vereint. Das DHLab findet in Räumen der Universitätsbibliothek für je zwei Stunden pro Woche statt, in denen Bibliothekspersonal als Ansprechpartner beziehungsweise Dozent zur Verfügung steht. Die räumliche und personelle Verankerung an der Universitätsbibliothek bietet eine neutrale und institutionelle Plattform, auf der sich die unterschiedlichen DH-Akteure der FAU gleichberechtigt begegnen können. Der offene Charakter ermöglicht Interessierten das Einbringen eigener Expertise und Inhalte. Getreu dem Ziel, die Digital Literacy in die Breite zu streuen, ist das Angebot nicht als Teil bestimmter Curricula konzipiert, sondern als Ergänzung zu den fachlichen Lehrveranstaltungen. Ferner werden in den Schulungen vorwiegend niedrigschwellige Inhalte in kleinen Zeiteinheiten angeboten. Zwar werden nach Möglichkeit Querverweise hergestellt, Lerneinheiten sind aber in sich abgeschlossen, um den Einstieg zu jedem Zeitpunkt zu ermöglichen.

Das DHLab nimmt damit sowohl Anleihen bei den aufkommenden Makerspaces (Owen 2017) als auch der traditionellen Bibliotheksexpertise in der Vermittlung von Informationskompetenz (Rauchmann 2010) und erweitert diese auf digitale Forschungswerkzeuge (Brandtner 2019). Die Implementierung wird bewusst bottom-up betrieben: In einer Sondierungsphase wurden die Ideen bestehender oder im Aufbau befindlicher Angebote anderer Institutionen[1] verglichen. Dabei wurde festgestellt, dass die bereits unterschiedlichen Konzepte stark variieren und nicht einfach auf die FAU übertragen werden konnten. Im Fokus steht daher, in Zusammenarbeit mit Digital Humanists der FAU, konkrete und drängende Anliegen vor Ort anzugehen und umzusetzen: Fragen zu Räumlichkeiten, Wissensvermittlung und Austausch. Dabei konnte auf Erfahrungen vorausgegangener Initiativen des akademischen Mittelbaus aufgebaut werden. So sollen umfangreiche Planungen "am Bedarf vorbei" vermieden werden.

Nach einem Semester Betrieb zeichnen sich bereits erste Vorteile und Herausforderungen dieses Vorgehens sowie gewisse Tendenzen ab: Trotz gemeldetem Bedarf ist das Angebot kein Selbstläufer, sondern muss gezielt und wiederkehrend über verschiedene digitale und analoge Kanäle beworben werden. Als problematisch hat sich hier die Benennung als "Digital Humanities Lab" erwiesen. Der Begriff „Digital Humanities“ spricht die geistes- und sozialwissenschaftliche Zielgruppe ungenügend an, wirkt teils ausgrenzend oder zu abstrakt. Was hier nun für die geglückte Wahrnehmung der DH als Fach an der FAU zu interpretieren ist, wirkt gleichermaßen ausschließend für alle nicht Dhler. Dies mag mit dem techniklastigen DH-Studiengang zusammenhängen, der durch den Erlanger Informatikkern (Sahle 2013: 32–37) profunde Programmierkenntnisse verlangt. Geisteswissenschaftler*innen fühlen sich daher durch die Benennung nicht angesprochen und befürchten, dass die Einstiegshürde für sie zu hoch ist. Darüberhinaus finden sich sozialwissenschaftlich Forschende in dem Begriff nicht wieder, auch wenn Digital-Humanities-Projekte häufig in ihren interdisziplinären Anlagen diese implizieren.

Als Lösung dieses Problems soll eine zielgruppengerechtere Bewerbung der konkreten Inhalte erfolgen. Gut angenommen wurden die Vorträge und die Schulungen zu Software-Werkzeugen, die ohne Programmierkenntnisse einen schnellen Einstieg bieten. Durch die enge Verzahnung mit den Wissenschaftlern entwickeln diese teilweise ein hohes Engagement und bringen zahlreiche Themen und Inhalte ein. Dies wirkt sich überaus fördernd auf Austausch und Vernetzung über die Fach- und Fakultätsgrenzen hinweg aus. Gerade relativ kleine Runden erlauben das Eingehen auf persönliche Wünsche und spezielle Kompetenzen. Gleichzeitig entwickelt sich ein aktiver "harter Kern" von Personen. Ob dies den Zielen wie auch der Außenwirkung eher hinderlich oder förderlich sein wird, ist momentan noch nicht abzusehen. Das Format des Helpdesks wird momentan am wenigsten angenommen. Dies mag der oben genannten Außenwahrnehmung geschuldet sein, den Öffnungszeiten, den weiteren, etablierten Beratungsmöglichkeiten der Universitätsbibliothek, oder auch der höheren Hemmschwelle einer persönlichen Anfrage.

Kurzfristig wird das DHLab seine Aktivitäten daher auf Schulungen und Vorträge konzentrieren. Die einzelnen Inhalte

werden genauer abgestimmt, wobei Tandem-Termine aus Erfahrungsschilderungen und einführende Tutorien eine wichtige Rolle spielen und Anreiz geben sollen, sich mit den Technologien zu beschäftigen. Momentan wird auch mit Blended Learning als Erweiterung des Präsenzangebots experimentiert. Als weitere Säule seines Service-Angebots plant das DH-Lab für 2020 die Bereitstellung von Spezialgeräten wie 3D-Scanner und VR-Brille. Dies soll den Einsatz moderner Technologien auch dort in Lehre und Forschung ermöglichen, wo sich die Anschaffung im Alleingang nicht lohnen würde.

Fußnoten

1. Vgl. beispielsweise das Digital Humanities Learning Lab Marburg (https://www.uni-marburg.de/de/ub/lernen/kurse-beratung/wissen-organisieren/dll [letzter Zugriff 24.09.2019]), das Beratungsangebot der ULB Bonn (https://www.ulb.uni-bonn.de/de/service/digital-humanities [letzter Zugriff 24.09.2019]) und das Konzept des Scholarly Makerspace der Humboldt-Universität (Kaden / Kleineberg 2019).

Bibliographie

Brandtner, Andreas / Lauer, Gerhard / Reuter, Peter (2019): "Die Bibliotheken haben ihre Zukunft vor sich, aber es sind Bibliotheken des 21. Jahrhunderts." Bibliotheken als Infrastrukturen der Geisteswissenschaften und als Orte der Selbstkultivierung, in: *ABI Technik*, 39(2), S. 171-178. http://dx.doi.org/10.1515/abitech-2019-2011 [letzter Zugriff: 24.09.2019]

Kaden, Ben / Kleineberg, Michael (2019): "Scholarly Makerspaces – Ein Zwischenbericht zum DFG-Projekt FuReSH", in: *LIBREAS. Library Ideas*, 35. https://libreas.eu/ausgabe35/kaden/ [letzter Zugriff: 24.09.2019]

Owen, Ivan (2017): "3D Printing and Makerspaces in Libraries", in: IFLA Trend Report 2017 Update, https://trends.ifla.org/files/trends/assets/documents/ifla_trend_report_2017.pdf [letzter Zugriff: 24.09.2019]

Rauchmann, Sabine (2010): Bibliothekare in Hochschulbibliotheken als Vermittler von Informationskompetenz, Humboldt-Universität zu Berlin, Philosophische Fakultät I, Kap. 1.2.1, 5.1 und 5.2 http://dx.doi.org/10.18452/16133 [letzter Zugriff:24.09.2019]

Sahle, Patrick (2013): DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities, DARIAH-DE Working Papers Nr. 1, Göttingen, http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2013-1-5 [letzter Zugriff 24.09.2019]

Becoming Urban: Ein Spiel mit Räumen

Bürgermeister, Martina
martina.buergermeister@uni-graz.at
Universität Graz, Österreich

Holzer, Matthias
matthias.holzer@stadt.graz.at
Stadtarchiv Graz

Nussmüller, Antonia
antonia.nussmueller@stadt.graz.at
GrazMuseum

Scheuermann, Leif
leif.scheuermann@uni-graz.at
Universität Graz, Österreich

Sonnberger, Jakob
jakob.sonnberger@uni-graz.at
Universität Graz, Österreich

Historischer Raum konstituiert sich stets als Summe der dokumentierten Wahrnehmungen einer Zeit. Dabei spielen die beschriebenen Räume nicht nur eine dokumentarische Rolle, vielmehr waren sie in ihrer Zeit normative Dokumente, so dass sich mentale Räume nicht nur im Begehen entwickeln konnten, sondern auch im Erlesen oder Erblicken. Historischer Raum kann also als Bricolage aus Erfahrungen historischer AkteurInnen verstanden werden, welche diese selbst körperlich oder aber textuell bzw. bildlich vermittelt gemacht haben.

Das durch die Österreichische Akademie der Wissenschaft geförderte Projekt „Becoming Urban – Reconstructing the city of Graz in the long 19th century (BeUrb)“ widmet sich als Gemeinschaftsprojekt der Karl-Franzens-Universität Graz, des Stadtmuseums Graz sowie des Stadtarchivs Graz der (Re-)konstruktion dieses historischen Raums am Fallbeispiel der Stadt Graz im langen 19. Jahrhundert (1789-1914). Thematisiert werden dabei gleichfalls die Veränderung der Stadt im Kontext der Urbanisierung wie auch deren Wahrnehmung in schriftlichen, bildlichen und kartographischen Quellen.

Durch die Zusammenarbeit von Stadtarchiv, GrazMuseum und Universität konnte die Quellenbasis mit historischen Reiseberichten und -führern, Karten, Plänen und weiteren geographischen Darstellungen sowie Stadtansichten auf Fotografien, Druckgrafiken und Gemälden auf unterschiedlichsten Gattungen aufgebaut werden, was einen umfangreichen Blick auf das Verständnis von der Stadt Graz im 19. Jh. erlaubt. Diese transdisziplinäre Herangehensweise, welche Historische Geographie, Kunstgeschichte, Urbanistik und Digitale Geisteswissenschaften vereint, wird durch eine gemeinsame technische Grundlage in Form eines Geoinformationssystems ermöglicht.

Als erste Herausforderung steht dabei die Erarbeitung eines gemeinsamen Datenmodells, welches die quellenspezifischen Charakteristika abbildet und die aufgenommenen Quellen gleichzeitig auf einer topographischen Ebene miteinander in Beziehung setzt. Modellierung wird dabei als „a process of signification and reasoning in action“ (Ciula / Eide, 2017 i34) angesehen. Die Theorie Modelle als *Icons* – also als bildhafte Vertreter anzusehen, hilft dabei Blickpunkte zu verschieben und Interpretationsspielräumen zu eröffnen, denn: “The context of the interpretation changes the sign but the sign also changes the context of interpretation” (i38). Fotografien und Pläne, aber auch Texte sind zudem an sich bildhafte Modelle, die in einem spezifischen Ähnlichkeitsverhältnis zum Original stehen. Ziel des Projektes ist es also, aus diesen singulären Modellen relationale, diagrammatische Modelle in Form eigener

historischer, dynamisch abfragbarer Karten zu erzeugen, die eine, wie Kralemann und Lattmann definieren, "interdependence between the structure of the sign and the structure of the object" (Kralemann / Lattmann, 2013: 3408) aufweisen.

Umsetzung

Die Umsetzung kann in eine Datenaufbereitung und daran anschließend eine vertiefte Geoanalyse unterteilt werden. Abschließend werden beide Ergebnisse des Projektes in einem Webauftritt präsentiert und nachhaltig in einem Repositorium archiviert.

a) Ein Metadatenmodell wird für die bildlichen, textuellen und kartographischen Quellen in LIDO implementiert.

b) Historische Karten werden georeferenziert und transkribiert, d.h. die in der Entzerrung entstandenen Artefakte werden auf einen Grundplan übertragen, der auf der Basis des Franziszeischen Katasters (1824) erstellt wird. Dabei werden bauliche Veränderungen, die sich aus den späteren Kartendarstellungen ergeben, in die Grundkarte integriert. Ein in diesem Kontext aufgebauter eindeutiger Identifikator fungiert als Grundlage für die räumliche Verortung der Bild- und Textquellen.

c) Textuelle Stadtbeschreibungen (z.B. Reiseberichte) werden transkribiert und die räumlichen Objekte (Gebiete, Wege, Plätze, Gebäude) darin in TEI-P5 mit den Referenzen aus der Basiskarte ausgezeichnet. In den Texten implizit vorhandene ‚narrative' Wege bzw. Abfolgen der Nennungen der Geoobjekte finden dabei ebenfalls besondere Berücksichtigung.

d) Bildliche Darstellungen (z.B. Gemälde, Druckgraphiken, Fotos, Postkarten) werden in LIDO beschrieben, in ihrer zeitlichen und räumlichen Gestaltung verortet. Dabei werden nicht nur abgebildete Objekte, sondern auch die mögliche Standorte und Blickrichtungen einbezogen

Auf Basis der gewonnenen Daten kann nun ein ‚Spiel mit den Räumen' beginnen[1], welches die hermeneutische Aneignung durch die Projektteilnehmer ebenso beinhaltet wie die diagrammatische Analyse (vgl. Bauer / Ernst 2010, Rau 2013). In der Analyse wird nun die sich stets im architektonischen Wandel begriffene ‚fluide' Stadt stehen, wie auch das Diskursfeld Graz, welche sich in den Quellen widerspiegelt. Dabei wird nicht nur die ‚dargestellte' Stadt im Zentrum stehen, sondern in besonderer Weise auch diejenigen Bezirke, die gerade nicht in den Quellen genannt werden und so als ‚blinde Flecken' im Diskursfeld der Stadt besonderer Aufmerksamkeit bedürfen.

Dabei stellen sich fragen wie: In wieweit sind jene Gebäude Ereignisorte der öffentlichen, aber auch einer ganz privaten Wahrnehmung? Wie sind die Eindrücke einer Stadt für Bewohnerinnen und Bewohner? Wie im Gegensatz dazu für Reisende? Wie möchte sich die Stadt nach außen repräsentieren? Was bewirken diesbezüglich Bau- und Infrastrukturmaßnahmen? Welche Bedeutung kommt Plänen einer zukünftigen Stadt zu, die niemals umgesetzt wurden? Wo finden sich widersprüchliche Wahrnehmungen – oder Raumverzerrungen?

Fußnoten

1. Als Spielwiese dient das open-source geographisches Informationssystem ‚QGIS' und die ‚GAMS-Infrastruktur', die die Forschungsdaten (GML, TEI, LIDO) langzeitarchiviert und die Möglichkeit bietet Daten neu zu kombinieren (content models, GEOSPARQL) und zu vernetzen (RDF) und abzufragen (SPARQL, SOLR).

Bibliographie

Bauer, Matthias / Ernst, Christoph (2010): Diagrammatik. Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld. Bielefeld: transcript.

Ciula, Arianna / Eide, Øyvind (2017): „Modelling in digital humanities: Sings in context" in: Digital Scholarship in the Humanities 32 (suppl_1) i33-i46 https://doi.org/10.1093/llc/fqw045

GAMS: Geisteswissenschaftliches Asset Management System, http://gams.uni-graz.at/

GEOSPARQL – A Geographic Query Language for RDF DATA (2012). Version 1.0 (eds.) OGC https://www.opengeospatial.org/standards/geosparql

GML: Geographic Markup Language (2012) – Extended schemas and encoding rules. Version 3.3.0 (eds.) OGC https://www.opengeospatial.org/standards/gml

LIDO – Lightweight Information Describing Objects (2010). Version 1.0 (eds.) Coburn, Erin et al. http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

Kralemann, B. / Lattmann, C. (2013): „Models as icons: modeling models in the semiotic framework of Peirce's theory of signs" in: *Synthese* 190, 3397-3420.

QGIS: https://www.qgis.org/de/site/

Rau, Susanne (2013): Räume, Wahrnehmungen, Nutzungen. Frankfurt/New York: Campus.

RDF – Resource Description Framework (2014) (eds.) W3C https://www.w3.org/RDF/

Scheuermann, Leif (2014): "Thoughts on a Web Based Coproductive Spatio-Temporal Information System" in: Rau, Susanne / Schönherr, Ekkehard (eds.): Mapping Spatial Relations, Their Perceptions and Danymics. Cham / Heidelberg / New York: Springer 17-23.

SOLR: Apache Solr. (eds.) Apache Software Foundation https://lucene.apache.org/solr/

SPARQL 1.1 Query Language (2013) (eds.) W3C https://www.w3.org/TR/sparql11-query/

TEI P5: Guidelines for Electronic Text Encoding and Interchange (2019). Version 3.6.0 (eds.) the TEI Consortium https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Besuch im »Marstheater« – Eine Netzwerkmodellierung von Karl Kraus' Riesendrama »Die letzten Tage der Menschheit«

Fischer, Frank
frank.fischer@dariah.eu
Higher School of Economics, Moskau, Russland

Busch, Anna
annabusch@uni-potsdam.de
Universität Potsdam

Hechtl, Angelika
angelika.hechtl@wu.ac.at
WU Wien

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam

Vogel, Andreas
andrea.uccello@gmail.com
Hamburg

Karl Kraus' Endzeitdrama »Die letzten Tage der Menschheit«, 1919 zum ersten Mal vollständig erschienen (Buchausgabe 1922), ist in vielerlei Hinsicht inkommensurabel. Der schiere Umfang sprengt alle Gattungsnormen (638 Seiten in der »Volk und Welt«-Ausgabe von 1978). Die fünf Akte plus Vorspiel und Epilog sind in 220 Szenen unterteilt, es gibt je nach Zählweise um die 1.000 sprechende Figuren bzw. Instanzen (zum Vergleich: als nächstgrößtes deutschsprachiges Drama gilt Grabbes »Napoleon oder Die hundert Tage« von 1831 mit 259 Figuren).

Die Zählweise ist nicht nur deshalb kontingent, weil es zahlreiche Rufe aus der Menge gibt, die sich nicht quantifizieren lassen (wozu vor allem auch das undurchsichtige Stimmengewirr im Epilog gehört), sondern auch, weil es konkrete Gruppierungen wie die ›Fünfzig Drückeberger‹ (III/26) oder ›Die zwölfhundert Pferde‹ (V/55) gibt, die man theoretisch quantifizieren könnte, auch wenn dies nicht unmittelbar sinnvoll erscheint. Insgesamt spricht man tatsächlich besser von Sprecherinstanzen, die von historischen Personen über namenlose Zwischenrufer und allegorische Figuren (etwa den »Hyänen, die Menschengesichter tragen«) bis hin zur »Stimme Gottes« reichen.

Es ist nicht nur auf das Thema des Stücks bezogen – die Apokalypse des Ersten Weltkriegs –, sondern auch auf die Form, wenn Kraus im Vorwort schreibt: »Die Aufführung des Dramas, dessen Umfang nach irdischem Zeitmaß etwa zehn Abende umfassen würde, ist einem Marstheater zugedacht. Theatergänger dieser Welt vermöchten ihm nicht standzuhalten.« (Kraus 1978, S. 5) Die Handlung der Tragödie ist »unmöglich, zerklüftet, heldenlos« (ebd.) und erschwert jede Absicht, das Stück darzustellen, zumal vollständig. Dies betrifft sowohl Inszenierungen auf der Bühne oder als Hörspiel (obwohl es schon Kompletteinspielungen gibt) als auch digitale Modellierungen der Figurenbeziehungen.

Es ist Konsens innerhalb des Forschungszweigs der Netzwerkanalyse dramatischer Texte, dass sich eine Einzelanalyse der verhältnismäßig übersichtlichen Figurennetzwerke selten lohnt. Das Augenmerk liegt daher normalerweise auf der Untersuchung struktureller Entwicklungen hunderter oder tausender Stücke über verschiedene historische Zeiträume (Algee-Hewitt 2017, Trilcke/Fischer 2018).

»Die letzten Tage der Menschheit« gehören hier zu den Ausnahmen. Ziel dieses Projekts ist es, das Stück als soziales Netzwerk zu visualisieren, basierend auf Kookkurrenzen von Sprecherinstanzen in den einzelnen Szenen. Voraussetzung dafür ist eine brauchbare Formalisierung des Gesamttextes. Dieser ist einerseits bereits digitalisiert, in annehmbarer Qualität innerhalb des Projekts Gutenberg-DE (obwohl es in dieser Version kaum eine Seite ohne zumindest kleinere OCR-Fehler gibt). Andererseits gibt es noch keine digitale Fassung in einem Format, das die wissenschaftliche Auswertung ermöglicht.

Am Beginn dieses Projekts stand daher die Herstellung einer TEI-Version des Dramas, die vor Konferenzbeginn veröffentlicht wurde und damit der wissenschaftlichen Community zum ersten Mal eine Version des Textes zur Verfügung stellt, die auf die FAIR-Prinzipien setzt (findable, accessible, interoperable, reusable). Neben einem Qualitätssprung hinsichtlich der Textbasis im Vergleich zur Gutenberg-DE-Version stand dabei die Auszeichnung der Sprecher-IDs im Mittelpunkt. Da, wie bereits angedeutet, diese Auszeichnung kontingent ist, also je nach Formalisierungsentscheidung anders aussehen kann, wird dieser Prozess offengelegt. So werden etwa die Vielzahl an Stimmen aus Menschenmengen oder die Unzahl ausrufender Zeitungsverkäufer nachvollziehbar individualisiert, speziell die Massenszenen in Wien, etwa die Geschehnisse an der Sirk-Ecke, die das Vorspiel und jeden der fünf Akte eröffnen.

Ergebnis ist ein visualisiertes Netzwerk, das auf einem Poster im A0-Format einen Blick ins Kraus'sche »Marstheater« erlaubt, auf die schiere Masse der Auftritte und Stimmen, aus der doch eine Struktur hervorscheint, wie sie bisher im Kontext der Kraus-Forschung noch nicht visualisiert worden ist. So werden viele »innere Symmetrien« sichtbar (Matala de Mazza 2018), die das Stück strukturieren, wiederkehrende Konstellationen wie etwa die vier Offiziere am Beginn jedes Aktes oder die Szenen in der Schulklasse (I/9 und V/23).

Deutlich wird im Netzwerkgraph auch die Diskrepanz zwischen Front und Heimat, zwei Welten für sich, wobei Kraus den Fokus auf die entlarvende Sprache von nicht direkt am Krieg beteiligten Personen legt: »Wenn nicht Krieg wär, möcht man rein glauben, es is Friede.« (Kraus 1978, S. 95)

Da der Text nunmehr als Volltext-TEI-Dokument vorliegt, lässt sich auch der Word Space in das Netzwerk hineinmodellieren, d. h., die Anzahl der Wörter pro Sprecherinstanz. Auf diese Weise scheinen deutlich die (quantitativ gesehen) Hauptfiguren dieses »heldenlosen« Dramas auf (etwa der »Nörgler« und der »Optimist« sowie der »Patriot« und der »Abonnent«), die oft über dutzende Seiten als Zweierkonstellationen auftreten, die aber darüber hinaus, wie der Graph verdeutlicht, auch anderweitig vernetzt sind.

Um auch komparatistische Aspekte abzudecken, werden auf dem Poster vergleichend einige Netzwerkmetriken präsentiert, um die Gigantomanie des Dramas mit Zahlen zu verdeutlichen.

Zur Gewährleistung der Nachnutzbarkeit und Nachhaltigkeit der Modellierung wurde das Stück auch dem German Drama Corpus hinzugefügt (https://dracor.org/ger), der den Zugang zu bestimmten Formalisierungen der Textsubstanz erheblich erleichtert (Fischer et al. 2019).

Bibliographie

Algee-Hewitt, Mark (2017): Distributed Character: Quantitative Models of the English Stage, 1550–1900. In: New Literary History 48(4), 751–782. Johns Hopkins University Press. DOI: https://doi.org/10.1353/nlh.2017.0038

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor. DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts, S. 194–197. DOI: https://doi.org/10.5281/zenodo.2596094

Kraus, Karl (1978): Die letzten Tage der Menschheit. Tragödie in fünf Akten mit Vorspiel und Epilog. Ausgewählte Werke. Band 5,1. Berlin: Volk und Welt 1978.

Matala de Mazza, Ethel (2018): Der populäre Pakt. Verhandlungen der Moderne zwischen Operette und Feuilleton. Frankfurt am Main: S. Fischer 2018.

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: https://doi.org/10.17175/sb003_003

BeyondTheNotes: Ein Tool zur quantitativen Analyse in den digitalen Musikwissenschaften

Ortloff, Anna-Marie
anna-marie.ortloff@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Windl, Maximiliane
maximiliane.windl@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Güntner, Lydia
lydia-maria.guentner@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Schmidt, Thomas
thomas.schmidt@ur.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Einleitung

Mit der Einführung des Konzepts des „Distant Reading“ von Moretti (2002) wurde in den digitalen Literaturwissenschaften in den letzten Jahren ein Trend angestoßen, den Einsatz von computergestützten quantitativen Methoden zur Analyse und Visualisierung von sehr großen Mengen von Texten zu explorieren. Während dieses Konzept und der Einsatz digitaler Methoden in den Literaturwissenschaften umstritten ist, sind computergestützte und quantitative Verfahren in den Musikwissenschaften schon länger etabliert und werden meist als statistische Musikwissenschaften bezeichnet (Nettheim, 1997). In Anlehnung an den Distant Reading-Begriff aus den Literaturwissenschaften wurde in den letzten Jahren versucht ähnliche Begriffe für die Musikwissenschaft einzuführen, um die computergestützte quantitative Analyse und Visualisierung von größeren Mengen an Musikstücken zu beschreiben. In der jüngsten Forschung findet man diesbezüglich die Begriffe: *Distant Audition* (Abdallah et al., 2017), *Distant Listening* (Cook, 2013) aber auch *Distant Hearing* (Burghardt, 2018). Diese Begriffe werden in Abgrenzung des jeweiligen Close-Konzepts, also *Close Audition/Listening/Hearing* betrachtet, womit die etablierte individuelle Analyse einzelner oder sehr weniger Stücke mittels hermeneutischer und qualitativer Methoden bezeichnet wird. Die genannten Begriffe sind nicht in gleicher Weise etabliert wie der Distant Reading-Begriff. Auch wird mit Distant Reading mittlerweile eine Vielzahl komplexer Methoden wie Sentiment Analysis und Topic Modeling beschrieben. Im Folgenden werden wir jedoch den Begriff Distant Hearing verwenden und bezeichnen damit die computergestützte quantitative Analyse und Visualisierung von mehreren Musikstücken. Wir berichten im vorliegenden Beitrag über den momentanen Stand der Entwicklung des neuen Tools *BeyondTheNote*, welches Konzepte des Distant Hearing integriert.

Tools und Programme in der digitalen Musikwissenschaft

Unabhängig von der Begriffsverwendung wurden einige Tools und Programme entwickelt, um die computergestützte Analyse im Sinne von Distant Hearing zu unterstützen. Nichtsdestotrotz liegen noch einige Mängel vor, die wir im Folgenden herausarbeiten, um die Entwicklung des neuen Tools *BeyondTheNotes* zu motivieren. Bereits in den 1990er Jahren wurde das *Humdrum*-Toolkit entwickelt (Huron, 1994; Huron, 2002). Es handelt sich dabei um eine programmiersprachen-unabhängige Sammlung von Kommandozeilen-Tools. Eines der bekanntesten und meistgenutzten Programm-Pakete ist *music21* (Cuthbert & Ariza, 2010). Dies ist eine Python-Bibliothek, die Analyse-Möglichkeiten für Musikstücke bietet, die in digitalen Formaten symbolhaft repräsentierter Musik (z.B. MusicXML) vorliegen. In beiden Fällen sind jedoch fortgeschrittene Programmier- und IT-Kenntnisse notwendig, um die Tools zu verwenden. Speziell für HumDrum findet man aber auch Tools, die versuchen eine grafische Schnittstelle anzubieten, um leichter auf die Funktionen von HumDrum zuzugreifen (Taylor, 1996; Kornstädt, 1996). Die genannten Umsetzungen benötigen jedoch teils aufwendige Installationen und erhebliche Einarbeitungszeit. Wie jedoch Burghardt und Wolff (2014) in ihrem Aufsatz über Humanist-Computer Interaction schreiben, ist eine möglichst einfache Zugänglichkeit und eine geringe Schwelle bezüglich des technischen Vorwissens ein essenzielles Kriterium damit Tools in den Geisteswissenschaften breite Verwendung finden. Ferner wird argumentiert, dass auch Aspekte der Usability und User Experience besonders wichtig sind, um aufwendige Einarbeitungszeiten zu vermeiden. Ein leichter zugängliches Web-Tool ist das *Digital Music Lab VIS* (DML-VIS, Abdallah et al., 2017). Das Tool integriert

auch Ideen des Konzepts von Distant Hearing und ermöglicht Analysen und Visualisierungen auf vorgefertigten Korpora. Dennoch fehlen einige Analysen wichtiger musikalischer Metriken und es ist auch nicht möglich eigenes Material zu analysieren.

Tools und Programme, die speziell die Bedürfnisse von Geisteswissenschaftlern beachten sind bislang selten. Im Kontext der Digital Humanities findet man aktuell Arbeiten im Kontext von Jazz (Frieler et al., 2018), klassischer Musik (Condit-Schultz et al., 2018) und Volksmusik (Burghardt et al., 2015; Burghardt & Lamm, 2017). Vereinzelt bieten diese auch statistische Analysen an (Burghardt et al., 2015), sind aber insgesamt verstärkt fokussiert auf Retrieval-Aspekte.

BeyondTheNotes

BeyondTheNotes wurde mit dem Ziel entwickelt, die weiter oben genannten Probleme und Mängel bisheriger Software-Pakete aufzugreifen und sich an Bedürfnissen von Musikwissenschaftlern zu orientieren. BeyondTheNotes grenzt sich von bisherigen Tools ab indem die technischen Hürden bezüglich Programmierkenntnissen und aufwendigen Installationsverfahren umgangen werden, da BeyondTheNotes als leicht zugängliches Web-Tool geplant ist, das eine grafische Benutzeroberfläche bietet und in jedem gängigen Browser verwendet werden kann. Um Aspekten der Usability und User Experience gerecht zu werden, integrieren wir Methoden des User Centered Design-Prozesses. Als weitere Abgrenzung zu bisherigen Software-Paketen liegt der Fokus auf Distant Hearing und nicht auf der Einzelanalyse. Im DML-VIS fehlende Funktionen wie der Upload von eigenen Dateien oder die Analyse wichtiger musikalischer Metriken wurden integriert. Zielgruppe des Tools sind Musikwissenschaftler und Studierende mit Interesse an quantitativer computergestützter Musikanalyse.

Abbildung 1: Logo von BeyondTheNotes

Entwicklung

Für die Entwicklung des Tools wurden Ideen des User Centered Design-Prozesses (UCD) (Vredenburg et al., 2002) integriert. Dabei wird versucht in iterativen Entwicklungszyklen potentielle Nutzer mit Methoden des Usability Engineerings so früh wie möglich in den Entwicklungsprozess einzubeziehen.

Um den Anforderungen unserer Zielgruppen gerecht zu werden, fand gemäß UCD vor Entwicklungsbeginn eine Anforderungsanalyse statt. Diesbezüglich wurden eine Fokusgruppe mit Studierenden der Musikwissenschaft sowie zwei semi-strukturierte Interviews mit ausgebildeten Musikwissenschaftlern durchgeführt. Dadurch sollte Einblick in die Arbeitsweisen von Musikwissenschaftlern gewonnen und Bedürfnisse an ein computergestütztes Tool identifiziert werden. Die Ergebnisse werden im Folgenden zusammengefasst:

Die Teilnehmer unserer Anforderungsanalysen erläuterten, dass es kein festes methodisches Vorgehen bei der Analyse von Musikstücken gibt, jedoch steht im Mittelpunkt stets das Verfassen eines Textes. Für diesen Prozess werden statistische Visualisierungen als nützlich erachtet. Meist wird nur ein Stück oder eine überschaubar große Zahl analysiert. Größere Analysen finden für Genres und Komponisten. Als wichtige Features wurden die Analyse von eigenem Material sowie der Download der Ergebnisse genannt. Interessante Metriken für die Analyse sind aus Sicht unserer Teilnehmer Leittöne, Tonarten, Akkorde, Intervalle, der Tonumfang, Tonhöhen und jegliche Form von Motiven. Die Teilnehmer äußerten selten den konkreten potentiellen Einsatz und Nutzen eines Tools in ihrem Arbeitsworkflow und sehen den meisten Nutzen eines potentiellen Tools eher in der vielseitigen Exploration einer großen Menge an Ergebnisse. Die Ergebnisse der Anforderungsanalyse wurden in greifbare Features übertragen und die Mehrzahl dieser in das Tool eingearbeitet. In der Weiterentwicklung des Tools werden wir den UCD weiter aufgreifen indem z.B. größere Usability-Tests und Redesign-Phasen stattfinden und wir den Einsatz des Tools im konkreten Forschungsworkflow untersuchen.

Das Tool wurde in Python mit dem Framework Django implementiert. Für viele musikalische Analysen wurde im Back-End music21 (Cuthbert & Ariza, 2010) eingesetzt. Für die Visualisierung von Notenblätter und Statistiken wurde OpenSheetMusicDisplay[1], Zingchart[2] und chartist.js[3] genutzt. Andere wichtige Technologien für die Entwicklung schließen PostgreSQL, JavaScript und Jquery ein.

Funktionen

Die Funktionen des Tools gliedern sich in zwei Bereiche. Die Analyse von einem einzelnen Stück inklusive seiner Partitur („Individual Analysis“) und die statistische Analyse von einem oder mehreren Werken bezüglich der Verteilungen unterschiedlicher Metriken („Distant Hearing"; Abbildung 2).

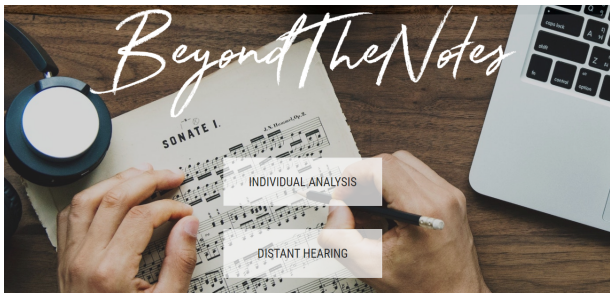


Abbildung 2: Start-Screen von BeyondTheNotes

Nach Auswahl eines Bereichs kann der Nutzer eine oder mehrerer Dateien für die Analyse hochladen. Es werden alle gängigen Dateiformate symbolhaft repräsentierter Musik akzeptiert z.B. MusicXML, MEI, Midi, ABC usw. Alternativ wird

zum Testen der music21-Korpus zur Verfügung gestellt. Es handelt sich dabei um ein freies, überschaubar großes Korpus, das unter anderem Werke von Mozart, Bach und Schubert enthält.

Über eine Suchfunktion können die hochgeladenen Dateien und das bestehende Korpus gefiltert werden (Abbildung 3).

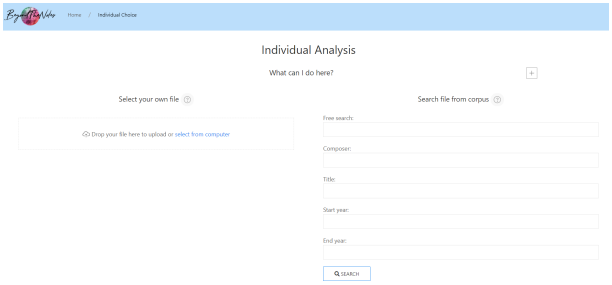


Abbildung 3: Upload und Suche

Wird die Individual Analysis gewählt, wird die Partitur des Stücks angezeigt (Abbildung 4).

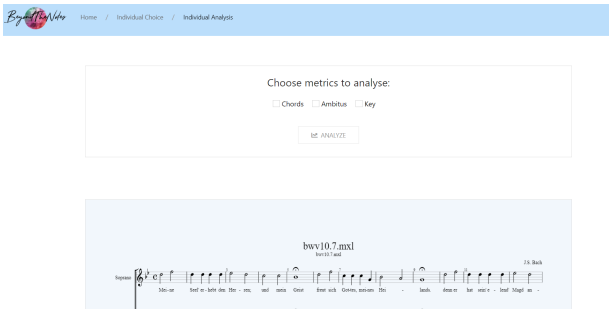


Abbildung 4: Anzeige für die Auswahl der „Individual Analysis“

Folgende Analysemöglichkeiten sind hier möglich:

- Die Akkordanalyse („Chords“): Hierbei wird die Partitur mit den Akkorden ersetzt und diese in römischen Ziffern oder ihren herkömmlichen Namen angezeigt (Abbildung 5)

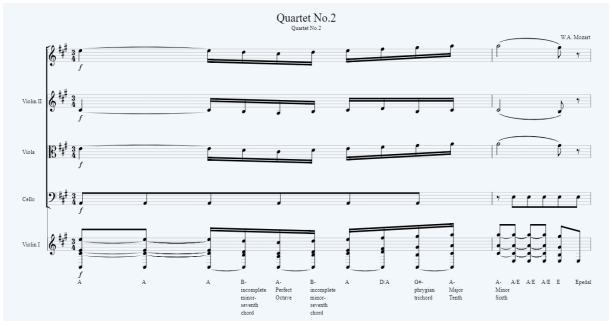


Abbildung 5: Transformiertes Notenblatt nachdem die Akkordanalyse durchgeführt wurde

- Die Analyse des Tonumfangs („Ambitus“) (Abbildung 6)

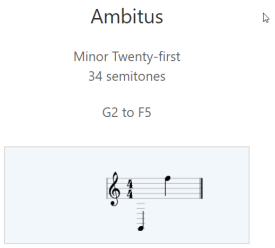


Abbildung 6: Tonumfang-Analyse (Ambitus)

- Die Analyse der Tonart: Hierbei werden die vier wahrscheinlichsten Tonarten mit ihren Wahrscheinlichkeitswerten angezeigt (Abbildung 7). Die Kalkulationen basieren auf music21.

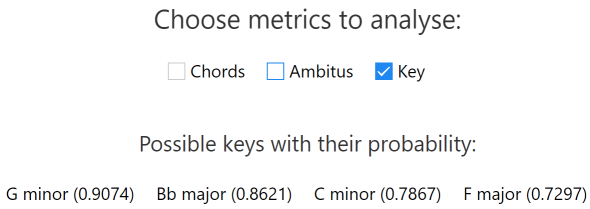


Abbildung 7: Tonart-Analyse

Die Ergebnisse der Akkord- und Tonartanalyse können auch verknüpft werden. Der Nutzer kann eine der ermittelten Tonarten auswählen und je nachdem werden die Akkorde angepasst, wenn römische Ziffern zur Anzeige verwendet werden.

Für die Distant Hearing-Funktionen muss der Nutzer zunächst die zu analysierenden Gruppen benennen. Es können dann beliebig viele Stücke der Suchleiste einer Gruppe hinzugefügt werden. Nach der Kalkulation der Daten werden fünf Visualisierungsbereiche angezeigt:

- Akkordanalyse: Über gepaarte Histogramme werden die Verteilungen der Akkorde in den einzelnen Gruppen dargestellt (Abbildung 8). Neben den Akkordverteilungen werden auch Akkord-Grundton- und Tongeschlechts-Verteilungen der Akkorde angezeigt.

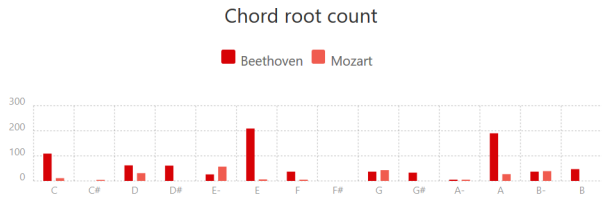


Abbildung 8: Akkordanalyse – Verteilungen von Akkorden für zwei Gruppen

- Tonhöhenanalyse: Über gepaarte Histogramme werden die Verteilungen der einzelnen Töne sortiert nach Tonname und Oktave angezeigt.

- Tondaueranalyse: Über gepaarte Histogramme werden die Verteilungen der Tondauern unterteilt in Noten und Pausen angezeigt (Abbildung 9).

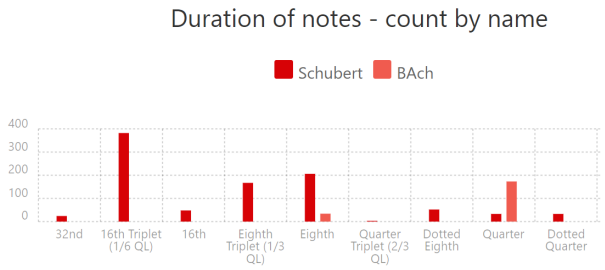


Abbildung 9: Tondaueranalyse – Verteilungen von Tondauern für zwei Gruppen

- Tonartanalysen: Hier werden über gepaarte Histogramme die Verteilung der Tonarten angezeigt. Auch wird ein Liniengraph angezeigt, der pro Gruppe die Wahrscheinlichkeiten für die einzelnen Tonarten angibt (Abbildung 10).

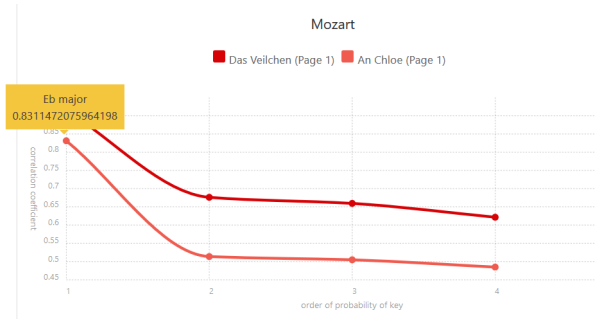


Abbildung 10: Tonartanalyse – Liniendiagramm für die Wahrscheinlichkeiten verschiedener Tonarten mehrerer Stücke

- Tonumfanganalyse: Es wird ein Reichweitendiagramm pro Gruppe angezeigt, welches den Tonumfang pro Stück in Form von horizontalen Balkendiagrammen anzeigt (Abbildung 11). Für die Gesamtgruppe wird die Menge und die Verteilung der genutzten Halbtonschritte auch noch in Form eines Boxplots angezeigt.

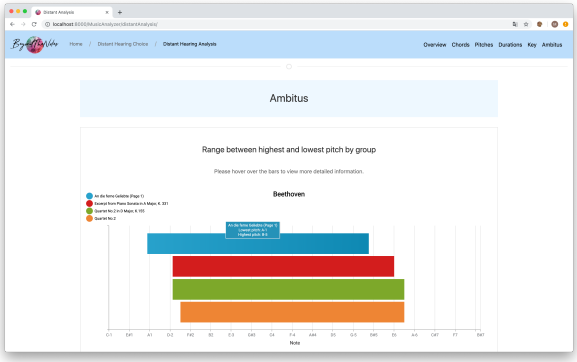


Abbildung 11: Tonumfanganalyse – Reichweitendiagramm für 4 Stücke, die der Gruppe Beethoven hinzugefügt wurden

Alle Graphen sind dabei interaktiv und bieten weiterführende Informationen an, wenn der Mauszeiger über Elemente bewegt wird. Die Diagramme können auch zusammen mit ihren Legenden heruntergeladen werden. Ebenso können die gesammelten Daten zur Weiterverwendung in einem JSON-Format heruntergeladen werden. An zahlreichen Stellen wurden Tutorials und Erklärungen eingebaut, um die Nutzung zu erleichtern.

Ausblick

Die momentane erste Version des Tools ist frei verfügbar und kann über *GitHub* heruntergeladen und genutzt werden[4]. Des Weiteren ist ein erster vorläufiger Prototyp auch online verfügbar[5].

Wir befinden uns am Ende des ersten Entwicklungszyklus und planen momentan die Evaluation des Tools gemäß dem UCD-Prozess. Des Weiteren explorieren wir weiter zusammen mit Musikwissenschaftlern, ob die gelieferten Funktionen den Analyseprozess unterstützen können und wie das Tool konkret in den Forschungsworkflow integriert werden kann. Im gleichen Schritt wollen wir auch erste forschungsrelevante Einsatzbeispiele diskutieren. Als ein Bereich für mögliche Analysen wurde von den Teilnehmern unserer Anforderungsanalyse vor allem der Vergleich von Genres, Komponisten und eigens erstellten Sammlungen bezüglich gängiger musikalischer Metriken genannt (Akkorde, Tonumfang etc.). Als eine komplexere Forschungsidee wurde die Untersuchung von Variationen diskutiert. *La Folia*, ein spanisches Motiv aus dem 16. Jahrhundert wurde von zahlreichen Komponisten als Grundlage für Variationen genutzt (Hudson, 1973). Durch die Nutzung eines geeigneten Korpus kann mit BeyondTheNotes untersucht werden, ob die Variationen dieses Motivs sich mehr nach Komponist, Zeitraum oder Ursprungsland unterscheiden. Ebenso wollen wir in den kommenden Iterationen durch die enge Zusammenarbeit mit Musikwissenschaftlern das Konzept und den tatsächlichen Nutzen des Distant Hearing kritisch reflektieren.

Fußnoten

1. https://opensheetmusicdisplay.org/
2. https://www.zingchart.com/
3. https://gionkunz.github.io/chartist-js/
4. Online verfügbar unter: https://github.com/Maxikilliane/DH_MusicAnalysis (Eine Installationsanleitung findet man im Repository)
5. Online verfügbar unter: https://beyondthenotes.herokuapp.com/

Bibliographie

Abdallah, Samer / Benetos, Emmanouil / Gold, Nicolas / Hargreaves, Steven / Weyde, Tillman / Wolff, Daniel (2017): "The Digital Music Lab: A Big Data Infrastructure for Digital Musicology", in: *Journal on Computing and Cultural Heritage (JOCCH)* 10(1).

Burghardt, Manuel (2018): "Digital Humanities in Der Musikwissenschaft – Computergestützte Erschließungsstra-

tegien Und Analyseansätze Für Handschriftliche Liedblätter" in: *Bibliothek Forschung Und Praxis* 42(2): 324–32.

Burghardt, Manuel / Lamm, Lukas (2017): "Entwicklung Eines Music Information Retrieval-Tools Zur Melodic Similarity-Analyse Deutschsprachiger Volkslieder" in: Eibl, Maximilian / Gaedke Martin (eds.): *INFORMATIK 2017*. Bonn: Gesellschaft für Informatik 87–99.

Burghardt, Manuel / Wolff, Christian (2014): "Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik" in: *DHd Workshop: Informatik und die Digital Humanities.*

Burghardt, Manuel / Lamm, Lukas / Lechler, David / Schneider, Matthias / Semmelmann, Tobias (2015): "MusicXML Analyzer. Ein Analysewerkzeug für die computergestützte Identifikation von Melodie-Patterns" in: *Hildesheimer Evaluierungs- und Retrievalworkshop* 2015: 29-42.

Condit-Schultz, Nathaniel / Ju, Yaolong / Fujinaga, Ichiro (2018): "A Flexible Approach to Automated Harmonic Analysis: Multiple Annotations of Chorales by Bach and Prætorius" in: *19th International Society for Music Information Retrieval Conference* 66–73.

Cook, Nicholas (2013): *Beyond the score: Music as performance.* Oxford University Press.

Cuthbert, Michael Scott / Christopher, Ariza (2010): "Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data" in: *11th International Society for Music Information Retrieval Conference (ISMIR 2010)* 637–642.

Frieler, Klaus / Hoger, Frank / Pfleiderer, Martin / Dixon, Simon (2018): "Two Web Applications for Exploring Melodic Patterns in Jazz Solos" in: *19th International Society for Music Information Retrieval Conference* 777–83.

Hudson, Richard (1973): "The Folia Melodies" in: Acta Musicologica 45 (1): 98–119.

Huron, David (1994): *UNIX Tools for Music Research: The Humdrum Toolkit*. Reference manual http://www.humdrum.org/Humdrum/manual07.html [letzter Zugriff 21. September 2019].

Huron, David. (2002): "Music Information Processing Using the Humdrum Toolkit: Concepts, Examples, and Lessons" in: *Computer Music Journal* 26 (2): 11–26.

Kornstädt, Andreas (1996): "SCORE-to-Humdrum: A Graphical Environment for Musicological Analysis" in: *Computing in Musicology* 10: 105–22.

Moretti, Franco (2002): "Conjectures on World Literature" in: *New Left Review* Jan / Feb: 54–68.

Nettheim, Nigel (1997): "A Bibliography of Statistical Applications in Musicology" in: *Musicology Australia* 20(1): 94–106.

Taylor, Michael (1996): *Humdrum Graphical User Interface*. Belfast, Queen's University.

Vredenburg, Karel / Mao, Ji-Ye / Smith, Paul W. / Carey, Tom (2002): "A Survey of User-Centered Design Practice" in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 471–478.

CASOTEX – Ein Projekt, das sozialwissenschaftliche, interpretative Methoden mit maschinellen Lernverfahren verschränkt

Albrecht, Jens
jens.albrecht@th-nuernberg.de
Technische Hochschule Georg Simon Ohm, Kesslerplatz 12, 90489 Nürnberg, Deutschland

Lehmann, Robert
robert.lehmann@th-nuernberg.de
Technische Hochschule Georg Simon Ohm, Kesslerplatz 12, 90489 Nürnberg, Deutschland

Einleitung

Viele Menschen leiden unter psychologischen und sozialen Problemen. Als niederschwellige Alternative zu klassischen Beratungsangeboten haben sich seit den 2000er Jahren Online-Beratungsangebote etabliert. Ratsuchende können hier ohne die Hemmschwelle eines persönlichen Kontakts ihre Probleme schildern und Lösungsvorschläge erhalten.

Im Rahmen des CASOTEX-Projektes werden Konversationen aus öffentlichen Online-Beratungsforen mithilfe statistischer Methoden und maschineller Lernverfahren analysiert. Dabei soll die Frage geklärt werden, ob eine Unterstützung der Berater durch die Identifikation erfolgreicher Beratungsmuster im Kontext einer individuellen Beratungssituation möglich ist. Erfolgreiche Beratungsmuster könnten dann sowohl im Rahmen der Ausbildung von Onlineberatern genutzt werden, als auch die Grundlage für Echtzeit-Assistenzsysteme in der Onlineberatung bilden. Im Detail wird untersucht, wie computerlinguistische Methoden und maschinelle Lernverfahren qualitative Analysen unterstützen bzw. ergänzen können, wo die Grenzen der Verfahren liegen und wie bei deren Einsatz vorzugehen ist.

Datengrundlage

Datengrundlage für CASOTEX sind anonymisierte Daten aus dem öffentlichen Beratungsforum der bke[1] (Bundeskonferenz für Erziehungsberatung). Hier gibt es seit über 10 Jahren Beratungsverläufe zu Fragen rund um Familie und Erziehung mit über 70.000 Beiträgen, in denen sowohl professionelle BeraterInnen als auch Laien in einer Beraterfunk-

tion auftreten. Bisher wurden diese großen Datensätze fast ausschließlich mit den Methoden der qualitativen Sozialforschung untersucht. Dadurch gelang es zwar, subjektive Sinnzusammenhänge zu extrahieren, allerdings waren in den entsprechenden Studien nur relativ kleine Stichproben möglich.

Darüber hinaus ist die qualitative Inhaltsanalyse, bei der menschliche Subjekte die Subjektivität anderer Menschen zu erschließen versuchen, nach wie vor dem Vorwurf ausgesetzt, keine objektiven Ergebnisse zu erzeugen (vgl. Döring/Bortz 2016). Vor diesem Hintergrund erscheint es hilfreich, umfangreiche Textquellen nicht nur von Menschen, sondern auch mit Methoden der künstlichen Intelligenz analysieren zu lassen. So könnte einerseits die Subjektivität der Texte berücksichtigt werden und andererseits die entstehenden Erkenntnisse aus Analysen sehr großer Stichproben gespeist werden.

Computergestützte Inhaltsanalyse

Die computergestützte Inhaltsanalyse bei CASOTEX umfasst verschiedene statistische und maschinelle Lern-Verfahren (s. Abb. 1).

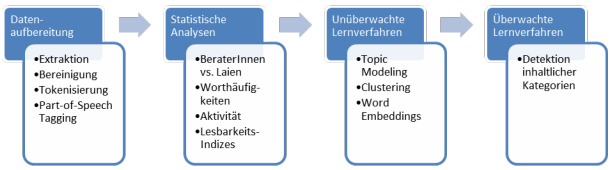


Abbildung 1: Schritte der computergestützten Analyse

Statistische Verfahren liefern erste Erkenntnisse zu Worthäufigkeiten, zum Sprachgebrauch und zur Aktivität in den Foren. Unüberwachte maschinelle Lernverfahren wie das Topic Modeling (Blei 2012) werden eingesetzt, um latente Themenkreise zu identifizieren. Eine zeitliche Betrachtung von Topic Models kann wesentliche Aufschlüsse über Veränderungen in der Art und Weise der Kommunikation liefern. Darüber, ob und inwiefern sich diese Verfahren auch für sozialwissenschaftliche Texte eignen, gibt es bisher nur wenig Erkenntnisse und die auch nur im englischen Sprachraum. Ergänzend können mithilfe neuronaler Netze kontextabhängige semantische Repräsentationen von Worten und Äußerungen erstellt werden, sogenannte Word Embeddings (Mikolov u.a. 2013). Dabei wird ein Sprachmodell trainiert, das es ermöglicht, Äußerungen semantisch zu differenzieren und Assoziationen zu bilden. Auch hierbei kann es aufschlussreich sein, zeitliche Verschiebungen zu identifizieren.

Für tiefere Analysen der Konversationen zwischen Beratenden und Hilfesuchenden ist jedoch eine präzise automatische Identifikation bestimmter sprachlicher Aspekte erforderlich. Diese Aspekte umfassen unter anderem die thematisierten Probleme, die Art der Äußerungen (Frage, Antwort, Feedback, Bestätigung, Zurückweisung, Ratschlag usw.), Emotionen (Trauer, Schmerz, Freude, Wut), psychosoziale Verhaltensweisen und Merkmale (Empathie, Depression) sowie Phasen im Beratungsprozess (Wälte/Borg-Laufs 2018, Althoff u.a. 2016). Überwachte Lernverfahren, die einen manuell annotierten Trainingskorpus benötigen, können genutzt werden, um diese Aspekte zu identifizieren. Als Grundlage für überwachte maschinelle Lernverfahren wurde daher ein Trainingskorpus generiert, der von menschlichen Textinterpreten mit einem Kategoriensystem nach Mayring strukturiert wurde (Mayring 2015). Der Korpus umfasst einige tausend annotierte Äußerungen zu 10 Kategoriebereichen. Die manuelle Annotation der Texte erfolgt über ein kooperatives webgestütztes Verfahren mithilfe des Werkzeugs Webanno[2]. Eine große Herausforderung stellt die Identifikation erfolgreicher Beratungsverläufe dar. Während in anderen Studien bewusst während der Beratung eine Erfolgseinschätzung bei den Ratsuchenden abgefragt wurde (Althoff u.a. 2016), liegen in den Forenverläufen entsprechende Informationen nicht systematisch vor. Neben der Berücksichtigung von entsprechenden Äußerungen im Kategoriensystem wurde im Rahmen der manuellen Annotation eine Experteneinschätzung des Erfolgs für die einzelnen Beratungsverläufe erstellt.

Erste Ergebnisse

Erste Ergebnisse mit den Trainingsdaten zum Kategoriensystem sind vielversprechend. So konnten bspw. die Kategorien "Problemdarstellung", "Empathie für Klient*in" und "Handlungsempfehlung" ohne spezifische Optimierungen mit einer Genauigkeit (F1-Score) von ca. 70% vorhergesagt werden. Als Lernverfahren wurde dabei eine Support Vector Machine basierend auf einem TF-IDF-Modell verwendet. Bei vielen Kategorien sind allerdings sehr wenig Trainingsdatensätze vorhanden, so dass sie nur unscharf bestimmt werden können. Durch den Einsatz aktueller Lernverfahren, insbesondere neuronale Netze mit vortrainierten Embeddings (Devlin u.a. 2018), werden deutliche Verbesserungen erwartet. In Abhängigkeit von den Ergebnissen müssen ggf. die Trainingsdaten erweitert und das Kategoriensystem geschärft werden.

Sehr gut ist eine automatische Klassifikation der Nutzer in Ratsuchende, professionelle BeraterInnen und Laien-BeraterInnen, d.h. angemeldete BenutzerInnen, die deutlich mehr Antworten geben als Fragen stellen, möglich. Darüber hinaus zeigte eine Untersuchung anhand von Lesbarkeitsindizes wie z.B. dem Flesch-Reading-Ease und der Wiener Sachtextformel (vgl. Groeben 1982), dass die Laien-BeraterInnen durchweg syntaktisch einfachere Äußerungen formulierten. Dieses Ergebnis ist eine wichtige Grundlage für die Ausbildung von Onlineberatern und entsprechende Assistenzsysteme. So ist eine praktische Umsetzung denkbar, die Beratern direkt bei der Erstellung eines Posts Rückmeldung zum syntaktischen Niveau ihrer Antwort gibt und ggf. Anpassungen vorschlägt.

Weiteres Vorgehen

Nachdem die manuelle Annotation der Trainingsdaten abgeschlossen ist, konzentriert sich die Arbeit jetzt auf die Untersuchung der überwachten Lernverfahren. Gesucht wird ein Modell, das es ermöglicht, bestimmte Äußerungen oder Teile von Äußerungen automatisch zu kategorisieren. Hierbei wird es für die praktische Anwendung in der Onlineberatung von großer Bedeutung sein, mit welcher Qualität das Erfolgskriterium modelliert werden kann. Da im annotierten Datensatz relativ selten eindeutige Erfolgsaussagen annotiert werden konnten, muss in der Auswertung mit dem Expertenurteil gearbeitet werden, das keinen konkreten Textbezug aufweist.

Unabhängig davon sollen auf Basis des Trainingskorpus bestimmte Aspekte der Beratungskonversationen, z.B. Empa-

thie-Äußerungen oder methodische Elemente wie ein von den Beratern initiierter Perspektivenwechsel rekonstruiert werden. Je nach Güte der Erkennungsleistung können diese Ergebnisse zumindest ein fundiertes deskriptives Bild über die Verbreitung der jeweiligen Aspekte im Gesamtdatensatz liefern. Darüber hinaus werden Zusammenhänge zwischen den identifizierbaren Aspekten im Datensatz untersucht. Selbst wenn eine Verbindung mit dem Erfolgskriterium nicht möglich sein sollte, kann so erstmals mit der Tiefe qualitativer Verfahren ein großer Datensatz der Onlineberatung quantitativ analysiert werden und mit bisher vorhandenen Wirkannahmen verglichen werden. So sollten zumindest in Teilbereichen Hypothesen generiert werden können, die in zukünftigen Studien genauer untersucht werden können.

Fußnoten

1. https://eltern.bke-beratung.de
2. https://webanno.sfs.uni-tuebingen.de/

Bibliographie

Althoff, Tim / Clark, Kevin / Leskovec, Jure (2016) "Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health". In: *Transactions of the Association for Computational Linguistics*, Vol. 4, S. 463–476.

Blei, David M. (2012) "Probabilistic topic models". In: *Communications of the ACM*. ACM, 55(4), S. 77ff.

Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Computing Research Repository (CoRR), arXiv:1810.04805

Döring, Nicola / Bortz, Jürgen (2016) *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg: Springer.

Groeben, Norbert (1982) *Leserpsychologie: Textverständnis - Textverständlichkeit.* Münster: Aschendorff.

Mayring, Philipp (2015) *Qualitative Inhaltsanalyse*. Weinheim: Beltz

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013) "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. USA: Curran Associates Inc. (NIPS'13), S. 3111–3119.

Wälte, Dieter / Borg-Laufs, Michael (Hrsg.) *Psychosoziale Beratung – Grundlagen, Diagnostik, Intervention. Stuttgart: Kohlhammer*.

Cooking Recipes of the Middle Ages: Nachnutzbare Ressourcen eines internationalen Forschungsprojekts

Steiner, Christian
christian.steiner@uni-graz.at
Universität Graz, Österreich

Klug, Helmut W.
helmut.klug@uni-graz.at
Universität Graz, Österreich

Böhm, Astrid
astrid.boehm@uni-graz.at
Universität Graz, Österreich

Raunig, Elisabeth
elisabeth.raunig@uni-graz.at
Universität Graz, Österreich

Laurioux, Bruno
bruno.laurioux@univ-tours.fr
Université de Tours

Ardesi, Denise
denise.ardesi@gmail.com
Université de Tours

Poirier, Corentin
corentin.poirier@univ-tours.fr
Université de Tours

Die kulinarische Tradition ist eine der prägendsten Elemente der europäischen Kultur und sie stellt einen großen Teil der nationalen Identitäten dar. In den letzten Jahrzehnten kam die Forschung zu zwei wichtigen Schlussfolgerungen in Bezug auf dieses Thema: Erstens, es gibt keine quantitativen Studien über die Herkunft und die Bildung von regionalen Küchen in Europa. Zweitens, im Mittelalter entstehen wesentliche Quellen: Manuskripte mit tausenden von Kochrezepten. Damit kann das Mittelalter als die Wiege der modernen europäischen Küche angesehen werden. Auf dem europäischen Kontinent bilden lateinische, mittelfranzösische und frühneuhochdeutsche Rezepte den Großteil der kulinarischen Überlieferung.

Das vorliegende internationale Projekt (ANR-17-CE27-0019-01, fwf I 3614) zielt darauf ab, die interkultu-

relle Forschung der mittelalterlichen Kochrezepte und deren Wechselbeziehung mithilfe eines interdisziplinären Ansatzes zu verwirklichen. Das Projekt nimmt die Kochrezeptüberlieferung von Frankreich und den deutschsprachigen Ländern als Korpus – dieses umfasst mehr als 80 Manuskripte und an die 8000 Rezepte – und untersucht sie in Hinblick auf ihre Entstehung, ihre Beziehung untereinander und ihre Migration durch Europa. Der Vergleich der französischen und deutschen Kulinargeschichte eignet sich besonders für diese Aufgabe, da Frankreich seit jeher einen kulturell prägenden Einfluss auf deutschsprachigen Völker hatte!

Die Partner, das Zentrum für Informationsmodellierung der Universität Graz und das Laboratoire CESR (Centre d'Etudes Supérieures de la Renaissance) der Universität Tours werden diese mehrsprachigen Texte nach modernen Standards aufarbeiten und sie mit aktuellen quantitativen und qualitativen Forschungsmethoden untersuchen. Für eine computergestützte Analyse werden die Rezeptsammlungen und die darin enthaltenen Texte und deren Metadaten in TEI/XML (Digitale Transkription und Edition) modelliert und mit Semantic Web Technologien analysiert (Digitale Annotation und Datenvisualisierung). Die Daten werden einer Langzeitarchivierungsinfrastruktur (GAMS, Zentrum für Informationsmodellierung Graz) zugeführt, in der sie weiter erforscht werden können. Alle Rezepte werden mithilfe von Vokabularien für Zutaten, Kochprozesse und Kochutensilien sowie kulturhistorisch relevanten Metadaten (z. B. in Bezug auf religiöse, kulturelle oder medizinische Aspekte) angereichert. Aufgrund dieser Informationen wird das Projekt über die Sprachgrenzen hinweg konkurrierende oder abweichende Essgewohnheiten, Textmigration sowie den gegenseitigen Einfluss der Nachbarländer auf ihre jeweilige Küche zu Tage fördern. Für die Analyse der deutschsprachigen Texte werden außerdem NLP-Methoden für historische Sprachstufen herangezogen, um Textverwandtschaften innerhalb dieser Überlieferung untersuchen zu können. Die Forschungsdaten und die Auswertungsergebnisse werden die Grundlage für eine räumliche und zeitliche Visualisierung und statistische Auswertung bilden, die neue Ansätze zur Interpretation des historischen und kulturellen Vermögens fördern wird.

Die im Projekt erarbeiteten Workflows und Daten werden ganz im Sinne des Open Science Gedanken und den FAIR-Prinzipien für die Nachnutzung zur Verfügung gestellt:

Der Transkriptionsworkflow und die Transkriptionsprinzipien (Theorie und Praxis, in Kooperation mit KONDE[1]) können zur Gänze nachgenutzt werden. Da die Manuskripte mit Transkribus [2] transkribiert wurden, steht ein trainiertes HTR-Modell zur Verfügung mit dem eine automatische Handschriftenerkennung von ähnlichen Texten denkbar ist. Das Annotationsvokabular (Zutaten, Speisen, Küchengeräte, Zubereitungsweisen) wird samt der zugewiesenen semantischen Wikidata-Konzepte zur Verfügung gestellt und stellt somit eine essentielle Basis für die Forschung im Bereich Kulinarhistorik dar. Die Konzepte in Wikidata[3] werden falls vorhanden kontrolliert und gegebenenfalls mit weiteren Daten (wie etwa Links zu relevanten Ontologien wie FoodO[4] oder SNOMED[5]) von unserer Seite angereichert. Noch nicht vorhandene Konzepte werden von uns neu erstellt und mit allen nötigen Daten (Statements) versehen. Die Nutzung von Wikidata verfolgt neben praktischen Überlegungen hauptsächlich das Ziel, die im Projekt gewonnenen Daten auf einfache Art und Weise für die Community bereitzustellen und eine weitere Bearbeitung dieser Daten zu ermöglichen. Überdies hinaus werden von uns auch die Annotationsskripte (Python und XSLT) für die Übertragung der Annotationsvokabularien nach TEI/XML zum Download angeboten.

Die überlieferten Texte werden durch eine hyperdiplomatische Neutranskription der historischen Quellen einheitlich erfasst und stehen als TEI/XML ebenfalls zur weiteren Nutzung zur Verfügung. Die Quellentexttranskription verzeichnet dabei nicht nur das unterschiedliche Schriftzeicheninventar, sondern auch alle textstrukturierenden Elemente. Das gesamte Zeicheninventar ist in einer nach den Richtlinien der TEI erstellten Zeichenbeschreibung erfasst. Die Beschreibung stützt sich dabei auf die theoretischen Ergebnisse zur Beschreibung von Zeichen aus dem DigiPal-Projekt[6] und verwendet außerdem die Zeichenidentifikatoren der Medieval Unicode Font Initiative[7] (vgl. Böhm, Klug 2020). Die so produzierten Daten sind nicht nur der Ausgangspunkt für die wissenschaftlichen Fragestellungen im Projekt, sondern bieten eine solide Grundlage für viele weitere Forschungsfragen aus Germanistik/Linguistik, Paläographie usw. Die Textdaten werden für eine Nutzung durch NLP Tools auch als Plaintext angeboten und die Handschriftenabbildungen sind je nach Nutzungsbedingungen der Bibliotheken frei verfügbar.

Darüber hinaus wird aus dem CoReMA-Projekt heraus ein Modell für die Integration weiterer Texte in die Forschungsumgebung bereitgestellt. Das Projekt soll fachliche Impulse für alle betroffenen Disziplinen der mittelalterlichen und frühneuzeitlichen Geschichte, Kulinargeschichte, Digitale Edition und Digital Humanities liefern.

Fußnoten

1. http://www.digitale-edition.at/
2. https://transkribus.eu/Transkribus/
3. https://www.wikidata.org
4. http://foodon.org/
5. https://browser.ihtsdotools.org/
6. Describing Handwriting I-VII; http://www.digipal.eu/blog
7. https://folk.uib.no/hnooh/mufi/

Bibliographie

Adamson, M. W. (Ed.) (1995): Food in the Middle Ages. A Book of Essays. New York, London: Garland. Adamson, M. W. (Ed.). (2002). Regional Cuisines of Medieval Europe: A Book of Essays. New York, London: Routledge.

Amoia, M., Martínez, J.M.M. (2019): SaCoCo Diachronic Corpus [WWW Document]. URL http://fedora.clarin-d.uni-saarland.de/sacoco/ (accessed 1.7.20).

Böhm, A. & Klug, H.: Quellenorientierte Aufbereitung historischer Texte im Rahmen digitaler Editionen: Das Problem der Transkription in mediävistischen Editionsprojekten. In: [Titel steht noch nicht fest] Hrsg. von Ingrid Bennewitz und Martin Fischer (= Bamberger interdisziplinäre Mittelalterstudien.) [in Vorbereitung]

Carlin, M., & Rosenthal, J. T. (Eds.). (1998): Food and Eating in Medieval Europe. London: Hambledon Press.

Flandrin, J.-L. (1984): «Internationalisme, nationalisme et régionalisme dans la cuisine des XIVe et XVe siècles: le témoignage des livres de cuisine». In Manger et boire au Moyen âge. Actes du Colloque de Nice (15-17 octobre 1982). (pt. 2, p. 75-91). Paris.

Flandrin, J.-L. & Hyman, P. (1988): "Regional tastes and cuisines: Problems, documents, and discourses on food in Southern France in the 16th and 17th centuries". Food and Foodways 1-3, p. 221-251.

Gloning, T., (2000): Monumenta Culinaria et Diaetetica Historica. Corpus of culinary & dietetic texts of Europe from the Middle Ages to 1800. Corpus älterer deutscher Kochbücher und Ernährungslehren [WWW Document]. URL http://www.staff.uni-giessen.de/gloning/kobu.htm (accessed 1.7.20).

Hieatt, C. (1995): Sorting through the Titles of Medieval Dishes: What Is, or Is Not, a "Blanc manger". In M. W. Adamson (Ed.), Food in the Middle Ages. A Book of Essays. (pp. 25-43). New York, London: Garland.

Hyman, P. & M. (2005). «Les associations de saveurs dans les livres de cuisine français du XVIe siècle». In *Le Désir et le Goût. Une autre histoire (XIIIe-XVIIIe siècles). Actes du colloque international à la mémoire de Jean-Louis Flandrin* (Saint-Denis, septembre 2003). Dir. Odile Redon, Line Sallman et Sylvie Steinberg. (p. 135-150). Saint-Denis: Presses Universitaires de Vincennes.

Karg, S. (Ed.). (2007): Medieval Food Traditions in Northern Europe. Copenhagen: National Museum of Denmark.

Klug, H. W., & Kranich, K. (2015): "Das Edieren von handschriftlichen Kochrezepttexten am Weg ins digitale Zeitalter. Zur Neuedition des Tegernseer Wirtschaftsbuches." In T. Bein (Ed.), Vom Nutzen der Editionen. Zur Bedeutung moderner Editorik für die Erforschung von Literatur- und Kulturgeschichte. (pp. 121-137). Berlin, Boston: De Gruyter.

Laurioux, B. (2005): «Les voyageurs et la gastronomie en Europe à la fin du Moyen âge ». In Le Désir et le Goût. Une autre histoire (XIIIe-XVIIIe siècles), Actes du colloque international à la mémoire de Jean-Louis Flandrin (Saint-Denis, septembre 2003). Dir. Odile Redon, Line Sallman et Sylvie Steinberg. (p. 99-117). Saint-Denis, Presses Universitaires de Vincennes.

van Winter, J. M. (1989). "Kochen und Essen im Mittelalter." In B. Herrmann (Ed.). Mensch und Umwelt im Mittelalter. (pp. 88-100). Frankfurt am Main: Fischer Taschenbuch Verl.

Das Erkenntnispotenzial Digitaler Musikedition

Iffland, Joachim

joachim.iffland@uni-paderborn.de
Universität Paderborn, Musikwissenschaftliches Seminar
Detmold/Paderborn, Deutschland

Die Entwicklungen im Bereich der digitalen Musikedition haben seit ihrer Entstehung eine Vielzahl von Projekten initiiert. Durch die Möglichkeit der Codierung und der mehrdimensionalen Darstellung musikbezogener Inhalte (vgl. Wiering 2009) konnte wesentlich zur Überwindung der Vorstellung von musikalischen Werken *einer* festen Gestalt verholfen,und das intersektionale Arbeiten der Digital Humanities in den Musikwissenschaften verankert werden.

Das Potenzial digitaler Musikedition – so zeigt es das hier vorgestellte, im *Zentrum Musik – Edition – Medien* angesiedelte und Ende 2019 abzuschließende Dissertationsprojekt – erschöpft sich jedoch nicht an der Bearbeitung des Werk-Faktors und des notenschriftbasierten Quellenmaterials. Digitale Edition eröffnet durch ihr Potenzial der tiefen und mehrdimensionalen Erschließung eines Gegenstandes auch das Potenzial, den zu edierenden musikbezogenen Gegenstand auszuweiten. Sie suggeriert somit die Möglichkeit, Musik nicht alleine mit Bezug auf ihren logischen Inhalt zu erschließen und dessen editorische Darstellung durch die optische und akustische Domäne musikbezogener Quellen zu flankieren, sondern Musik im Sinne *gelebter Wirklichkeiten* zu repräsentieren, in Musik also auch im editorischen Sinne mehr zu sehen als Notentext. *Digitale* Musikedition eröffnet im Sinne der Digital Humanities somit ein Erkenntnispotenzial, das es ermöglicht, aus editorischer Sicht die grundlegende Frage zu stellen, was Musik ist.

Die Arbeit zeigt dabei, dass der Versuch, Musikedition mit digitalen Mitteln über den Notentext hinaus auszuweiten, Erkenntnis über den Gegenstand „Musik" offenbart und geht von der kulturwissenschaftlich inspirierten Prämisse aus, dass Musik ein vom Handeln geprägtes Ereignis ist. Am Beispiel einer dichten Beschreibung eines Ausschnitts einer Konzert-Aufzeichnung des Sängers Marius Müller-Westernhagen, wird die Vielfalt des Komplexes „Musik" verdeutlicht und der Frage nachgegangen, auf welcher entitätenbezogenen Basis dieses musikbezogene Handeln in editorische Kontexte integrierbar ist, um nicht nur digitale Notenedition, sondern digitale Musikedition im umfassendsten Sinne zu betreiben – als dichte Beschreibung mit digitalen Mitteln. Neben der Beleuchtung bisheriger musikwissenschaftlicher Editionspraxis und damit verbundener Prinzipien, gilt es, das Wesen digitaler Notenedition vorzustellen, um zunächst zu verdeutlichen, dass diese unter der Nutzung der xml-basierten MEI- und TEI-Standards weitgehend die Prinzipien traditioneller Notenedition in das Digitale transferiert hat und qua der Struktur des Codes an der Edition von Meisterwerken festhält. Kulturwissenschaftliche Erkenntnisse (wie die Bedeutung musikbezogener Handlungen) sind hier kaum in editorischen Kontexten wiederzufinden oder in diese integrierbar. Diese Arbeit verdeutlicht durch experimentelle Anreicherung einer MEI-Codierung die Notwendigkeit der grundsätzlichen ontologischen Erschließung des (handlungsbezogenen) Gegenstands „Musik" sowie die Notwendigkeit des grundsätzlichen Lösens vom bisherigen werkbezogenen Blickwinkel.

Bestehende Projekte entwickeln bereits vielfältige, durch digitale Techniken ermöglichte Insellösungen, die damit beginnen, die Betrachtung des Komplexes „Musik" auszuweiten. Doch der Faktor des Werkes scheint hier ein schwer zu überwindendes Hindernis. Um in diesem Kontext die (auch editorische) Betrachtung von Musik in einen größeren Zusammenhang zu stellen, frage ich, was Musik ist und stelle im Zusammenhang mit Christopher Smalls Konzept des *Musicking* einen handlungsbezogenen Musikbegriff vor. (Small 1998) Als Verifizierung seiner These und zur Überbrückung von in seiner Arbeit vorzufindenden Defiziten, wird der Begriff des Musicking zunächst ontologisch differenziert. Das Musicking kann somit auf der Basis von fünf grundlegenden Musicking-Entitäten – Akteur, Ding, Ereignis, Text, Raum – präzisiert werden. Diese werden als ontologische Basis einer Musikedition vorgeschlagen, die den Status von Musik als Handeln anerkennt und widerspiegelt. Der Begriff der Musikedition wird dabei präzisiert und vom Komplex der Noten- oder Werkedition unterschieden. Das Projekt verdeutlicht so die Notwendigkeit, diesen Ansatz als Ontologie des Musicking weiter auszubauen, um Musik mit digitalen Mitteln einer „wirklichen" Musikedition zuzuführen und – im Sinne der Digital Humanities als „intersection" (vgl. Nyhan/Flinn 2016:1.) – Edition als digitale kulturwissenschaftliche Edition zu betreiben. Bestehende, in Insellösungen manifestierte Bestrebun-

gen zu Edition, Codierung und Erforschung musikbezogener Kontexte können durch das vorgeschlagene Prinzip aufgegriffen werden, welches mittels einer „Partitur des Musicking" u.a. mit Techniken der Graphenvisualisierung einen editorischen Rahmen für alle bisher durchgeführten Konzepte vorschlägt.

Bibliographie

Babbitt, Milton (1965): „The Use of Computers in Musicological Research", in: *Perspectives of New Music* 3/2 74–83.

Grotjahn, Rebecca / Iffland, Joachim (2018): „Digitale Musikedition und die Wissenschaft der Populären Musik", in: *Die Musikforschung* 71/IV 379–393.

Iffland, Joachim (2018): „Materialität und Schriftlichkeit", in: *Zentrum Musik – Edition – Medien* https://zenmem.de/confluence/pages/viewpage.action?pageId=33718295 [letzter Abruf 19. September 2019).

Kaden, Christian (2004): *Das Unerhörte und das Unhörbare. Was Musik ist, was Musik sein kann,* Kassel/Stuttgart.

Kepper, Johannes (2011): *Musikedition im Zeichen neuer Medien. Historische Entwicklung und gegenwärtige Perspektiven musikalischer Gesamtausgaben,* Diss., Paderborn 2009, Norderstedt.

Kepper, Johannes / Pugin, Laurent (2017): „Was ist eine Digitale Edition? Versuch einer Positionsbestimmung zum Stand der Musikphilologie im Jahr 2017", in: *Musiktheorie* 32/4 347–363.

McCarty, Willard (2003): „Humanities Computing", in: *Encyclopedia of Library and Information Science*, DOI: 10.1081/E-ELIS 120008491, New York 1224– 1235.

McCarty, Willard (2016): „Becoming Interdisciplinary", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A New Companion to Digita l Humanities*, Chichester 69–83.

Nyhan, Julianne / Flinn, Andrew (2016): *Computation and the Humanities. Towards an Oral History of Digital Humanities,* Cham.

Small, Christopher (1998): *Musicking. The Meanings of Performing and Listening* (Music/Culture), Middletown.

Veit, Joachim (2015): „Musikedition 2.0. Das ‚Aus' für den edierten Notentext?", in: *editio 29*/1 187–197.

Veit, Joachim / Richts, Kristina (2018): „Stand und Perspektiven der Nutzung von MEI in der Musikwissenschaft und in Bibliotheken", in: *Bibliothek – Forschung und Praxis* 42/2 292–301.

Walser, Robert (2016): *The Christopher Small Reader* (Music/Culture), Middletown.

Wiering, Frans (2009): „Digital Critical Editions of Music: A Multidimensional Model", in: **Crawford, Tim / Gibson, Lorna**: *Modern Methods for Musicology. Prospects, Proposals, and Realities*, Farnham 23–45.

Das gute Glas – Glasgestaltung im Zeitalter der guten Form

Kraft, Anneli
anneli.kraft@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg

Entwicklung einer digitalen Infrastruktur als Werkzeug zur Analyse und designhistorischen Betrachtung von Trinkgläsern.

Abbildung 1: Szene aus Hitchcocks "Suspicion".

Ein Glas Milch wird im Psychothriller „Suspicion"[1] (Jacobi 2014) von Alfred Hitchcock zum spannungsgeladenen Deutungsträger und Schlüssel der weiteren Handlung. Tatsächlich sind wir umgeben von den verschiedensten Arten von Trinkgläsern, die immer auch Träger von Informationen über Nutzer, Konsumverhalten und die gesellschaftlichen Normen sind. Allerdings sind sie so alltäglich, dass wir uns darüber kaum Gedanken machen, genauso wenig wie über die Form oder die Herstellungsweise. Ob Preis, Funktion, Design – es spielen unterschiedliche Kriterien eine Rolle, warum Konsumenten sich für ein ganz bestimmtes Trinkglas entscheiden. Gibt es aber auch Kriterien, die das gute Glas von einem minderwertigen unterscheiden? Um diese zu beurteilen, müssen sowohl die Hintergründe der formalen Gestaltung, das Wissen über Trends und Geschmack, die traditionelle Verwendung sowie insbesondere die Möglichkeiten der Glasherstellung untersucht werden. Unter der Prämisse der ‚guten Form' wurden in der Nachkriegszeit Objekte ausgewählt, die als besonders hochwertig angesehen wurden. Diese wurden in Ausstellungen gezeigt oder mit Preisen ausgezeichnet, um sowohl Hersteller als auch Verbraucher für die gute Gestaltung zu sensibilisieren. Die damals aufgestellten Richtlinien werden in dieser Arbeit anhand der prämierten Trinkgläser überprüft. Auf dieser Basis werden schließlich eigene Kriterien festgelegt, die eine Aussage über die Qualität von Trinkgläsern geben.

Als Werkzeug zur Sammlung und Analyse der Daten dient ein digitales Instrumentarium, das auf Basis der bereits bestehenden Informationsplattform *WissKI* (Görz 2011)[2] entwickelt und dem Forschungsgegenstand angepasst wurde. Bereits in der Planung des Instrumentariums stand der zu erforschende Gegenstand im Zentrum. Die Untersuchung des Trinkglases als Typus mit seinen Einzelteilen und Merkmalen diente sowohl der Konzeption der Datenbank als auch der Erstellung der *CIDOC-CRM*[3] basierten Domänenontologie[4] als Grundlage. Mit diesen Vorarbeiten wurde eine dem Forschungsthema entsprechende Datenbank modelliert, mit passenden Masken und einem hinterlegten Fachbegriffssystem.

Aus der inhaltlichen Bearbeitung und der Entwicklung der digitalen Infrastruktur ist der Grundstock zu einem disziplinä-

ren Repositorium[5] zum Thema Trinkglas gelegt. Ziel ist, dass diese Datenbank nicht nur für ein Forschungsprojekt genutzt wird, sondern anschließend auch in Museen speziell für den Bereich Gebrauchsglas eingesetzt werden kann. Das Repositorium dient als zentrale Datenbank in der sowohl die Daten zu den verschiedenen Trinkgläsern gebündelt als auch sämtliche zugehörigen Informationen und Dokumente, die zur Erforschung der Gläser relevant sind, an einem Ort zusammengeführt werden. Zusätzlich zu den Abbildungen von Gläsern werden Kataloge, Werbeprospekte und technische Zeichnungen bereitgestellt. Bereichert wird die Datenbank durch ein erweiterbares Glossar mit Abbildungen und der Definition von Fachbegriffen sowie einem Warenzeichenlexikon, welches die Einordnung und Zuschreibung von Gläsern erleichtert.

Für Datensammlungen im Allgemeinen und die Datenbank ‚Das gute Glas' im Speziellen stammen die Daten in der Regel aus unterschiedlichen Quellen und Kontexten, was eine Heterogenität der Daten zwangsläufig mit sich bringt. Das bedeutet, die sinnvolle Nutzung der Daten ist nicht immer gegeben und muss zunächst durch eine systematische Vereinheitlichung passieren. Zusätzlich zum Metadatenschema, die als Ontologie dem System hinterlegt ist, dient dafür die Orientierung an Dokumentationsstandards[6] sowie die Verlinkung zu Normdaten[7].

Im Gegensatz zu konventionellen Museumsdatenbanken bleibt durch die Nutzung solch eines Trinkglas-Repositoriums bereits geleistete Forschungsarbeit nicht auf eine Institution oder Person begrenzt, sondern kann auch von anderen Wissenschaftlern und Wissenschaftlerinnen genutzt werden. Insbesondere bei einem Forschungsgegenstand wie dem Trinkglas, bei dem das Material weit über Europa verstreut ist, ist eine Zusammenführung von Daten zum Erkenntnisgewinn wünschenswert. Das Dissertationsprojekt ‚Das gute Glas' stellt in der Kunstgeschichte und im Museumsalltag ein Pilotprojekt dar, welches die Methoden der Digital Humanities[8] nutzt und seine Vorteile gegenüber herkömmlicher Herangehensweisen für die wissenschaftliche Bearbeitung etabliert.

Fußnoten

1. Hitchcocks „Suspicion" (Verdacht) ist eines der filmischen Meisterwerke zum selben Grundthema. Einer der Höhepunkte ist die Szene, in der Johnnie, gespielt von Cary Grant, seiner Frau Lina (Joan Fontaine) ein Glas mit vergifteter Milch bringt, das szenisch inszeniert und von innen beleuchtet ist. (Jacobi 2014)
2. Wissenschaftliche Kommunikations-Infrastruktur
3. CIDOC CRM (CIDOC conceptual reference model) bietet eine erweiterbare Ontologie für Begriffe und Informationen im Bereich des kulturellen Erbes, als Norm für einen kontrollierten Austausch von Daten.
4. Ontologien (Informatik) sind meist sprachlich gefasste und formal geordnete Darstellungen einer Menge von Begrifflichkeiten und der zwischen ihnen bestehenden Beziehungen in einem bestimmten Gegenstandsbereich, eine Art Schlagwortsystem.
5. Repositorien (lat. Lager) sind im allgemeinen Verständnis gut sortierte und verwaltete Speicherorte für digitale Forschungsdaten, welche entweder öffentlich oder in den meisten Fällen einem beschränkten Nutzerkreis zur Verfügung stehen. In der Regel werden damit Forschungsdaten zugänglich gemacht und im besten Fall auch die Infrastruktur geboten, die eine Langzeitarchivierung der Daten sicherstellt.
6. SPECTRUM ist der britische Dokumentationsstandard für Museumsobjekte (The UK Museum Documentation Standard). Online unter: http://museumswesen.smwk.sachsen.de/download/spectrum-de-3-1_21-1-2013.pdf, [Stand: 05.07.2019]
7. Eine Verlinkung kann beispielsweise zu Wikidata, Getty oder der GND erfolgen, sie ist aber nicht vorher festgelegt.
8. Die digitalen Geisteswissenschaften versuchen über die Interessen an einem Fachgebiet hinaus Prozesse der Gewinnung und Vermittlung neuen Wissens unter den Bedingungen einer digitalen Arbeits- und Medienwelt weiter zu entwickeln. Dazu forschen und lehren sie z.B. im Bereich der Digitalisierung des Wissens und des kulturellen Erbes, der Anwendung und Weiterentwicklung von Werkzeugen, der Operationalisierung und Beantwortung von Forschungsfragen und der Reflexion über die methodischen und theoretischen Grundlagen der Geisteswissenschaften in einer digitalen Welt. (Sahle 2011: 4)

Bibliographie

Görz, Günther (2011): „WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung. In: Kunstgeschichte.", in: *Open Peer Reviewed Journal*, 2011 http://www.kunstgeschichte-ejournal.net/167/1/Goerz.pdf [letzter Zugriff am: 26.02.2015].

Jacobi, Hanna (2014): „Alfred Hitchcock: ‚Suspicion' - das wichtigste Milchglas des Kinos", Radiobeitrag am 03.05.2014 http://detektor.fm/kultur/alfred-hitchcock-suspicion [letzter Zugriff am: 13.09.2019].

Fichtner, Mark / Hohmann, Georg (2005): „Chancen und Herausforderungen in der praktischen Anwendung von Ontologien für das Kulturerbe" in: Digitales Kulturerbe : Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis: 116.

„Repositorien | Bewahren und Nachnutzen | Themen" https://www.forschungsdaten.info/themen/bewahren-und-nachnutzen/repositorien/ [letzter Zugriff: 18.10.2016].

Sahle, Patrick (2011): „Was sind die digitalen Geisteswissenschaften?": 4. in: Sahle, Patrick: Digitale Geisteswissenschaften. http://dig-hum.de/sites/dig-hum.de/files/cceh_broschuereweb.pdf" [letzter Zugriff: 06.03.2015)].

SPECTRUM The UK Museum Documentation Standard. http://museumswesen.smwk.sachsen.de/download/spectrum-de-3-1_21-1-2013.pdf [letzter Zugriff: 05.07.2019].

„Was sind Repositorien" https://open-access.net/informationen-zu-open-access/repositorien/ [letzter Zugriff: 18.10.2016].

Weller, Kathrin (2013): „Ontologien", in: Kuhlen, Rainer / Semar, Wolfgang / Strauch, Dietmar: Grundlagen der praktischen Information und Dokumentation. Handbuch zur Einführung in die Informationswissenschaft und-praxis: 207, S. 209.

„What is the CIDOC CRM" http://www.cidoc-crm.org [letzter Zugriff: 13.09.2019].

„What is WissKI?" http://wiss-ki.eu/what_is_wisski. [letzter Zugriff: 07.07.2017].

„WissKI – Wissenschaftliche Kommunikations-Infrastruktur" http://www.gnm.de/forschung/archiv-forschungsprojekte/wisski/ [letzter Zugriff: 07.07.2017].

Abbildungsnachweis: Filmstill aus dem Film „Suspicion" von A. Hitchcock. https://3.bp.blogspot.com/-y-WDR3et970/VU6WDMemTbI/AAAAAAAAMJU/fHcl4mAGpZY/s1600/cary%2Bgrant%2Bglass%2Bmilk%2Bsuspicion%2Bhitchcock.jpg [heruntergeladen am: 13.09.2018].

Das Theater mit dem Theater: Thementransfer in den Spectators

Fuchs, Alexandra
alexandra.fuchs@uni-graz.at
Universität Graz, Österreich

Geiger, Bernhard
geiger@ieee.org
Know-Center Graz, Österreich

Hobisch, Elisabeth
elisabeth.hobisch@uni-graz.at
Universität Graz, Österreich

Koncar, Philipp
philipp.koncar@tugraz.at
Technische Universität Graz, Österreich

More, Jacqueline
jacqueline.more@uni-graz.at
Universität Graz, Österreich

Saric, Sanja
sanja.saric@uni-graz.at
Universität Graz, Österreich

Scholger, Martina
martina.scholger@uni-graz.at
Universität Graz, Österreich

Die journalistische Gattung der „Spectators" des 18. Jahrhunderts stellt ein wichtiges Kulturerbe aus der Zeit der Aufklärung dar. Die Zeitschriften entsprachen dem demokratischen Ideal, kulturelle und moralische Fragen in nicht-akademischen Kreisen zu verbreiten und Werte der Aufklärung wie Weltoffenheit, Toleranz, intellektuelle Kritik und soziale Verantwortung zu popularisieren. Ausgehend von den englischen Modellzeitschriften *The Tatler* (1709-1711), *The Spectator* (1711-1712 bzw. 1714) und *The Guardian* (1713) hat sich dieses journalistische Genre über ganz Europa mittels Übersetzungen, Adaptionen und Imitationen verbreitet. Auf Basis des mehrsprachigen Korpus (derzeit Italienisch, Spanisch, Französisch, Englisch, Deutsch, Portugiesisch) der Digitalen Edition *The Spectators in the International Context* (Ertler et al.) zeigt das Poster die Ausbreitung zentraler Themen des Aufklärungsdiskurses, wie etwa Theater, Sitten und Bräuche, Frauen- und Männerbild.

Anhand der Untersuchung von populären Themen mit maschinellen Methoden präsentiert der Beitrag zentrale Linien des Transfers von Diskursen innerhalb des Genres der Spectators und reflektiert damit den Zeitgeist des 18. Jahrhunderts. Mit Hilfe einer Kombination aus *close reading*, *distant reading* und Visualisierungsmethoden wird aufgezeigt, welche Themen lokale Relevanz hatten und welche länderübergreifend in den Diskurs aufgenommen wurden. Innerhalb letzterer soll zwischen Themen kontrastiert werden, die in unterschiedlichen Sprachgemeinschaften unabhängig voneinander Relevanz erlangt hatten, und jenen, die durch Imitation und Übersetzung verbreitet wurden.

Als Fallbeispiel kann das Thema Theater angeführt werden, das in zahlreichen Zeitschriften unterschiedlicher Länder unabhängig voneinander diskutiert wurde. Während in Frankreich das neoklassizistische Theater im 18. Jahrhundert bereits vollends etabliert und somit als Thema in den Spectators weniger relevant war, erlebten Italien und Spanien eine bewegte nationale Theaterdiskussion. Seit dem 16. Jahrhundert hatte sich das italienische Theater stetig weiterentwickelt. Im 18. Jahrhundert wurde es jedoch als Repräsentationsmedium für ein modernes Italien auserkoren und erlebte eine massive Veränderung. Unabhängig davon wurde auch in Spanien der politische Streit zwischen Progressisten und Traditionalisten über das Thema des Theaters ausgetragen: Spanische Intellektuelle versuchten zunehmend durch kulturelle Reformen den intellektuellen Anschluss an Europa zu schaffen und lieferten sich mit Traditionalisten einen regelrechten Streit über eine radikale Reform des Theaters nach neoklassizistisch französischem Vorbild. (Vgl. z.B. Guinard 1973, 133-138; Ertler 2003, 120-124)

Die Basis der Untersuchungen bildet das Korpus von etwa 4000 Texten, das im *close reading*-Verfahren hinsichtlich der narrativen, thematischen und sachorientierten Inhalte annotiert und mittels des XML/TEI Standards (TEI Consortium 2019) ausgezeichnet wurde. Der Vorteil der TEI Kodierung für die Analyse ist neben der quantitativen Auswertung der Metadaten, Schlagworte und Entitäten die Möglichkeit von Untersuchungen auf Basis der narrativen Textstruktur. So können nicht nur der gesamte Text, sondern in den Zeitschriftentexten ausgezeichnete narrative Erzählformen (wie etwa Traumsequenzen oder Leserbriefe) separiert voneinander analysiert und kontrastiert werden. Neben der quantitativen Auswertung der manuell zugewiesenen Themen wird Topic Modelling (z.B. Blei et al. 2003, Jelodar et al. 2019, Kuang et al. 2015) eingesetzt, um diese Annotationen zu bestätigen und zu ergänzen und neue Perspektiven auf die Rezeption des Materials zu eröffnen. Abb. 1 und 2 zeigen einen Vergleich der manuellen und der maschinellen Auswertung von Themen: Es zeigt sich eine Übereinstimmung beider Herangehensweisen dahingehend, dass die maschinelle Auswertung das verstärkte Auftreten des Themas Theater in den Ausgaben 18-27, 37-46 und 68-78 bestätigt.

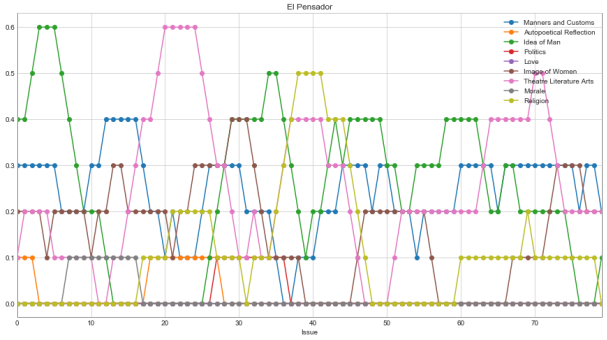


Abbildung 1: Statistische Auswertung der manuell zugewiesenen Themen (in rosa z.B. "Theatre Literature Arts") auf die Ausgaben des *El Pensador*.

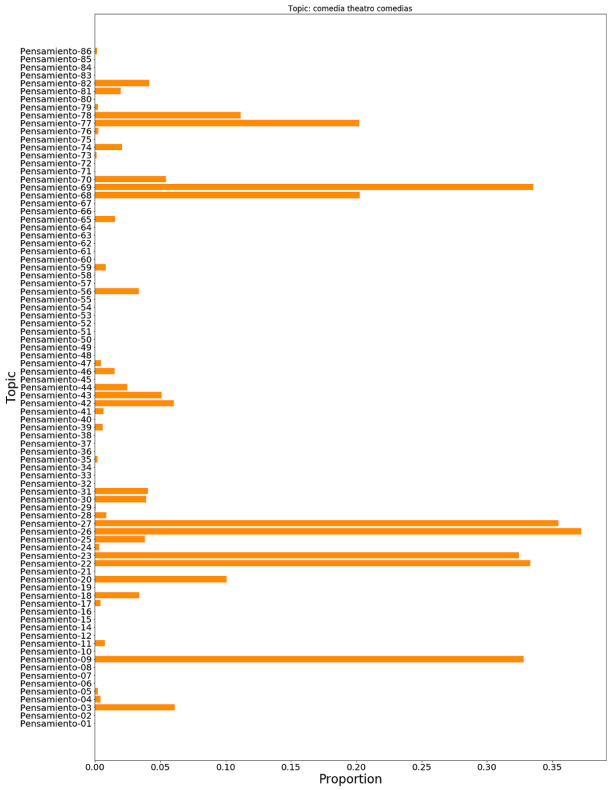


Abbildung 2: Topic Modelling: Verteilung des Topics "Comedia & Theatro" (mit den 10 häufigsten Wörtern comedia, theatro, poetas, pueblo, pieza, amor, accion, nacion, pasion, tragedia) auf die Ausgaben des El Pensador.

Darüber hinaus werden die (manuellen und maschinellen) Annotationen verwendet, um das zeitliche Auftreten ausgewählter Themen innerhalb einer Sprachgemeinschaft und sprachenübergreifend zu analysieren. Als Referenz für diese Untersuchungen dient das Zeitschriftennetzwerk der Spectators: Dieses stellt die Abhängigkeit von Zeitschriften unterschiedlicher Sprachgemeinschaften hinsichtlich Übersetzung, Adaption und Imitation dar.

Der wissenschaftliche Beitrag des Posters und der dem Beitrag zugrunde liegenden Arbeit am Projekt *Distant Spectators: Distant Reading for Periodicals of the Enlightenment* – einer Kooperation zwischen dem Zentrum für Informationsmodellierung und dem Institut für Romanistik der Universität Graz, sowie dem Institute of Interactive Systems and Data Science der Technischen Universität Graz und dem Know-Center Graz – kann in zwei Aspekte unterteilt werden. Einerseits kann durch die Anwendung von *distant reading*-Methoden die Liste der manuell selektierten Themen erweitert und feiner granuliert werden. Dieser Zugang relativiert die vorab generierte Leseerwartung und ermöglicht somit einen unvoreingenommenen Blick auf den Text sowie die Analyse größerer, ggf. noch nicht annotierter Korpora. Andererseits wird durch die vorgestellte Arbeit die öffentlich zugängliche Digitale Edition der Spectators (Ertler et al.) zusätzlich (semi-automatisiert) angereichert und durch auf den TEI-Annotationen basierende Visualisierungen augmentiert. Dies ist ein erster explorativer Schritt in Richtung intelligentes maschinelles Lernen im Kontext der Spectators und eröffnet neue Wege der Erschließung, Analyse und Präsentation dieser multilingualen literaturwissenschaftlichen Ressource.

Bibliographie

Ertler, K.; Fuchs, A.; Fischer, M.; Hobisch, E.; Scholger, M.; Völkl, Y. (Hg.) (2011-2019): *The Spectators in the International Context*, https://gams.uni-graz.at/spectators.

Ertler, K. (2003): *Moralische Wochenschriften in Spanien. José Clavijo y Fajardo: ‚El Pensador'.* Tübingen: Gunter Narr.

Guinard, P. (1973): *La Presse espagnole de 1737 à 1791. Formation et signification d'un genre*. Paris: Centre de recherches hispaniques.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003): Latent dirichlet allocation. *Journal of machine Learning research*, 3 (Jan), 993-1022.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019): Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78 (11), 15169-15211.

Kuang, D., Choo, J., & Park, H. (2015): Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms* (pp. 215-243). Springer, Cham.

TEI Consortium (2019): eds. *Guidelines for Electronic Text Encoding and Interchange.* July 2019. http://www.tei-c.org/P5/.

Der Datenpool eines frühneuzeitlichen Self-Trackers, oder: Johann Christian Senckenbergs „Observationes". Ein Distant Reading-Zugang

Faßhauer, Vera
fasshauer@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Einleitung

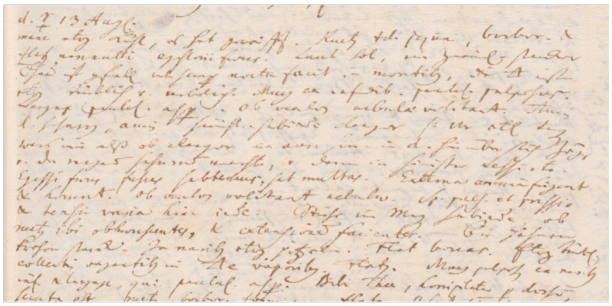
Aktuelle Diskussionen über ständig wachsende Möglichkeiten der Erfassung, Speicherung und Analyse großer Datenmengen lassen uns vergessen, dass sowohl Wissenschaft als auch Behörden bereits seit Jahrhunderten Praktiken zur Datenerhebung und -verarbeitung entwickelt haben (Borck 2017, Oertzen 2017). So wurden bereits im 17. Jahrhundert astronomische und meteorologische Beobachtungsdaten in Formularen und Tabellen erfasst, in Zahlen und Symbolen kodiert und in Karten und Diagrammen visualisiert (Daston 2011, Mendelsohn 2011, Hess 2011). Auch die empirische Erfassung von personenbezogenen Körperdaten nahm ihren Anfang in den Praxisjournalen frühneuzeitlicher Ärzte, die alle Symptome, Zustandsmerkmale und Reaktionen sowie den Krankheitsverlauf, die Medikamentierung und den Heilungsprozess ihrer Patienten genau dokumentierten (Geyer-Kordesch 1990, Stolberg 2007). Nicht einmal das von den Anhängern der *Quantified Self* -Bewegung praktizierte *Self-Tracking* (Lupton 2016) gehört gänzlich dem 21. Jahrhundert an, wie die Tagebücher zahlreicher religiöser Selbstoptimierer aus dem 17. und 18. Jahrhundert zeigen.

Eines der ambitioniertesten historischen *Self-Tracking* -Projekte waren zweifellos die *Observationes in me ipso factae* des pietistischen Arztes Johann Christian Senckenberg (1707–1772). In ihnen protokollierte er tagtäglich sein gesamtes Körper- und Seelenleben, um seinen Lebenswandel zu vervollkommnen. Neben Ernährung, Stoffwechsel, Körperaktivität sowie Schlaf- und Ruhephasen notierte er auch alle Reizempfindungen, Körperreflexe und Gemütszustände sowie alle spürbaren Umwelt- und Witterungseinflüsse (Faßhauer 2017). Zeitweise brachte er auf diese Weise täglich bis zu 5.000 Wörter zu Papier, so dass auf den Gesamtzeitraum von dreizehn Jahren gerechnet ca. 14.000 Seiten mit 12.600.000 Wörtern zusammenkamen. Der Beitrag beleuchtet zunächst den epistemischen Zweck dieses riesigen frühneuzeitlichen Datenpools, aus dem derzeit in Frankfurt ausgewählte Bände digital ediert werden (Faßhauer 2018), und setzt sie mit zeitgenössischen Positionen zum Verhältnis von Daten und Theorie in Beziehung. Anschließend wird die Möglichkeit einer Analyse dieser Daten mit modernen *Distant-reading* -Methoden und deren Vereinbarkeit mit dem epistemischen Ziel des religiösen Autors diskutiert.

Selbstbeobachtung und Anti-Rationalismus

Wenn ein Selbstoptimierer des digitalen Zeitalters beschließt, seine Lebenszeit effizient zu nutzen, seinen Körper gesund zu erhalten oder seine Finanzen zu organisieren, bezweckt er damit meist größtmöglichen persönlichen Erfolg und Selbstzufriedenheit im Diesseits. Ein religiöser Self-Tracker des 18. Jahrhunderts hatte hingegen zu allererst sein Seelenheil und seine Erlösung nach dem Tod im Sinn. Diese konnte jedoch nur erlangen, wer die ihm anvertrauten Gottesgeschenke auf Erden treulich verwaltete, pflegte und mehrte. Genau wie für materielle Güter galt dies auch für Gesundheit und Wissenskapital. Nach Auffassung religiöser Gelehrter wie Senckenberg war der Mensch seit dem Sündenfall jedoch geistig so zerrüttet, dass er durch seine Verstandeskräfte zu keinen verlässlichen Erkenntnissen gelangen konnte. Insbesondere theoretische Modelle, die durch „künstliches syllogisiren der Vernunfft“ dem Verstand der Gelehrten („ex mente Doctorum“) entstiegen sind, repräsentierten nur fragmentarisches oder abstraktes Wissen. Zudem müssten ohnehin „alle Regulae universaliores erstlich ab experientia in particularibus“ abgeleitet werden, deren Vielfalt jedoch so viele Ausnahmen aufzeige, „daß die Regulae selbst wieder darüber zernichtet werden“ (Senckenberg 1735: 1r/v, Faßhauer 2017). Sichere medizinische Erkenntnisse waren deshalb nur „ohne Praeoccupation von einer vorher gefassten Hypothesi“ durch mehrfach wiederholte unmittelbare Selbsterfahrung zu erlangen, deren Resultate möglichst vollständig aufgezeichnet und induktiv ausgewertet werden mussten. Auf diese Weise ließen sich Vergleiche mit Aufzeichnungen aus ähnlichen Situationen herstellen, wobei einzelne Faktoren miteinander korreliert, auf ihre Relevanz und Rolle im Gesamtkontext befragt und als mögliche Ursachen oder Auswirkungen anderer Faktoren in Betracht gezogen werden konnten.

Ganz ähnliche Überzeugungen wie Senckenberg äußerte der amerikanische Journalist Chris Anderson, als er im Jahre 2008 das Ende der Theorie und den Beginn des Datenzeitalters verkündete. Auch er ging dabei von der Prämisse aus, dass Theorien die Realität nur verzerrt wiedergäben und letztlich allein in den Hirnen der Wissenschaftler existierten: „The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists”. Stattdessen verwies er wie Senckenberg auf die Möglichkeit, durch Erhebung und Speicherung einer möglichst großen Datenfülle ungleich genauere und verlässlichere Aussagen über die Welt in ihrer ganzen Komplexität zu treffen. War der Verzicht auf die Suche nach letztgültigen Kausalitäten bei Senckenberg noch religiös motiviert, entspringt er bei Anderson aus der pragmatischen Erkenntnis, dass die Verfügbarkeit nie gekannter Datenmengen die Notwendigkeit zur Hypothesenbildung schlichtweg erübrige, da sie unter den verschiedensten Gesichtspunkten miteinander korreliert werden könnten. Durch ihre maschinelle Auswertbarkeit gerieten zudem auch Einzelheiten und Muster ins Blickfeld, die der theoriegeleiteten Forschung entgingen: „Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all“ (Anderson 2008). Seither wurde gegen Andersons These wiederholt eingewandt, dass bereits die Erstellung von Datensätzen auf theoretischen Prämissen und Selektionskriterien beruhe (boyd/Crawford 2012; Boellstorff 2014). Zudem stelle jede Datenanalyse eine subjektive Interpretation innerhalb bestimmter kultureller, dogmatischer oder ideologischer Kontexte dar, so dass eine hypothesenfreie Datenauswertung unmöglich sei. Die gleiche Problematik lässt sich auch für die Aufzeichnungen Senckenberg aufzeigen: Außer Thermometer und Barometer stand ihm nur sein eigenes Bewusstsein als Messinstanz zur Verfügung, das alle Empfindungen und Wahrnehmungen zwangsläufig subjektiv registrierte, filterte und interpretierte und dabei den protokollierten physischen und seelischen Befindlichkeiten selbst unterworfen war. Auch erfolgte die Aufzeichnung der Daten allein durch seine eigene schreibende Hand, die auf körperliche Irregularitäten ebenso empfindlich reagierte wie auf seelische Erschütterungen und Stimmungsschwankungen. Waren Körper und Bewusstsein anderweitig okkupiert, konnte die Datenerfassung entweder gar nicht oder nur rückwirkend und durch das Gedächtnis vermittelt erfolgen.




```
<div type="day"> <head> <date when-iso="1732-08-13"> <supplied>13.8.1732</supplied></date> </head>
  <pb facs="b0359.jpeg" n="359"/>
  <p> d<ex>ie</ex> <choice><orig>&#x263F;</orig><reg><g>Mercurii</g></reg> </choice> 13. Aug<ex>usti</ex>
    <lb/> mane etwas kühl, es hat gereifft. Ructus <choice><orig>&#x1F70A;</orig> <reg><g>aci</g></reg>
    </choice>di jejuni, borbor<ex>ygmi</ex> &amp;amp; <lb/> flatus nonnulli egestorii faeces. Lucet sol, ein
    zieml<ex>ich</ex> starcker <lb/> Thau ist gefallen ut semper noctu facit in montibus, die <choice>
    <orig>&#x1F701;</orig> <reg><g>Luft</g></reg> </choice> ist <lb/> etwas trüblich v. nebelich. Mucus
    ex infundib<ex>ulo</ex> paululum pulposus. <lb/> Larynx paululum asper. ob oculos nebulae volitant.
    Auris <lb/> d<ex>extrae</ex> susurrus, auris sinist<ex>rae</ex> subinde clangor!: vor etl<ex>ichen</ex>
    tagen <lb/> wars mir alß ob clangor ex aure sin<ex>istra</ex> in d<ex>extram</ex> hin über sich zöge
    <lb/> v. da magnum susurrum machte, v. dann in sinistra cessirte. <lb/> Egessi faeces fuscas subtenues,
    sat multas. Extrema omnia frigent <lb/> &amp;amp; horrent. ob oculos volitant nebulae. Sp<ex>asmus</ex>
    puls<ex>atorius</ex> et pressio <lb/> &amp;amp; tensio varia hinc inde. Stiche im Magen subinde, ob <lb/>
    ructus ibi obstrecentes, &amp;amp; extensionem facientes. Die bahnen <lb/> krehen starck. In naribus etwas
    pitzeln. Flat boreas. Etwas trübl<ex>ich</ex> <lb/> collectis vagantibus in <choice><orig>&#x1F701;</orig>
    <reg><g>aer</g></reg></choice>e vaporibus. Flatus. Mucus pulposus ex naribus <lb/> inf<ex>undibulo</ex>
    &amp;amp; larynge, qui paululum asp<ex>er</ex>. Bibi Thee, horripilatio p<ex>er</ex> dorsum <lb/> secuta est,
    ructus, borbor<ex>ygmi</ex> tormina, flatus. Oscitatio &amp;amp; pandi-<lb type="inWord"/> culatio, &amp;amp;
    extremorum frigus, sonderl<ex>ich</ex> pedum. Eunti im hembde <lb/> allein bey kühler <choice>
    <orig>&#x1F701;</orig><reg><g>Luft</g></reg></choice>, und horripilatione, ructibus <choice>
    <orig>&#x1F70A;</orig> <reg><g>aci</g></reg> </choice>di &amp;amp; borbor<ex>ygmi</ex> <lb/> it<ex>em</ex>
    flatibus nonnullis &amp;amp; sp<ex>asmus</ex> puls<ex>atorius</ex> hinc inde, &amp;amp; frigore manuum <lb/>
    &amp;amp; pedum, bina ex nare sinistra sternutatio, bey dem susurro <lb/> auris d<ex>extrae</ex> multo.
```


Abbildungen 1a und 1b: Selbstbeobachtung in Senckenbergs Tagebuch, Dezember 1732 (Manuskript und TEI/XML-Transkription)

Datengetriebenes *Distant Reading*?

Senckenbergs riesiger Datenpool konfrontiert seine modernen Leser mit einem so mikroskopisch detaillierten Bewusstseinsstrom, dass ein Verständnis seiner Erkenntnisse im *close reading* -Modus nahezu unmöglich ist. Die digitale Erschließung einzelner Bände im Rahmen der *Frankfurter Auswahledition* ermöglicht nun eine Annäherung an das umfangreiche Textmaterial aus der Makroperspektive (Abb. 1a–b). Der Literaturwissenschaftler Franco Moretti hat dieses Vorgehen bekanntlich als „distant reading“ bezeichnet, da Distanz hier statt eines Hindernisses eine Bedingung der Erkenntnis darstelle: „it allows you to focus on units that are much smaller or much larger than the text“. Die Reduktion von Senckenbergs umfangreichen Beobachtungsdaten auf abstrakte Schemata scheint jedoch zunächst im Widerspruch zu seiner Absicht zu stehen, die ganze Vielfalt der natürlichen Erscheinungsformen unverkürzt zu erfassen. Auch Moretti hat auf dieses Problem hingewiesen: „If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor” (Moretti 2013: 48–49). Die irreversible Reduktion von Texten auf abstrakte Schemen, die in der Forschung sogar als gewaltsame Zerstörung des eigentlichen Untersuchungsgegenstandes beschrieben worden ist (Bradley 2012), kann nur in solchen Forschungsumgebungen vermieden werden, die – wie etwa die Voyant Tools (Sinclair und Rockwell 2003) – einen flexiblen Wechsel zwischen der Text- und der Grafikebene und damit zwischen dem *close* - und *distant reading* -Modus ermöglichen (Jänicke 2016: 20–23). Ein möglicher Ausgangspunkt für eine Fernlektüre ist die Identifizierung von Schlüsselwörtern, die hier anhand von Häufigkeitskriterien erfolgt. Werden in der Liste einzelne Keywords ausgewählt, kann deren Verteilung im Korpus angezeigt werden. Eine Visualisierung der markantesten körperlichen Empfindungen Senckenbergs zwischen August und Dezember 1732 zeigt zum Beispiel, dass im November und Dezember Spannungs- und Druckgefühle vorherrschten, während er im Oktober hauptsächlich Stiche und Zuckungen verspürte, im August und September aber frei von derlei Empfindungen war (Abb. 2). Ein Blick auf die Kollokationen dieser Begriffe zeigt, in welchen Körperteilen sie am häufigsten bemerkbar waren (Abb. 3). Allerdings stellen derlei Festlegungen auf bestimmte Untersuchungszeiträume oder Körperempfindungen bereits Arbeitshypothesen dar, die mit einer bestimmten Erwartungshaltung einhergehen und das Ergebnis dadurch nicht unmaßgeblich präformieren. Die oben aufgezeigte Unmöglichkeit des von Senckenberg projektierten theoriefreien Wissenserwerbs spiegelt sich deshalb unmittelbar in der digitalen Korpusanalyse wider, die gleichfalls nicht rein datengetrieben bzw. ohne hypothetische Vorüberlegungen erfolgen kann.

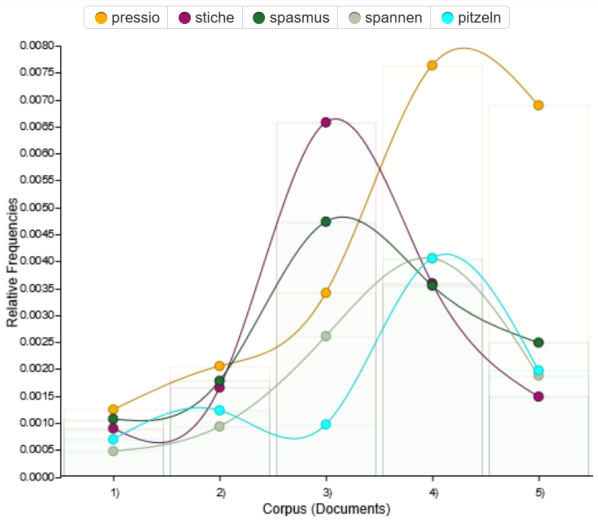


Abbildung 2: Häufigkeit von Körperempfindungen zwischen August und Dezember 1732

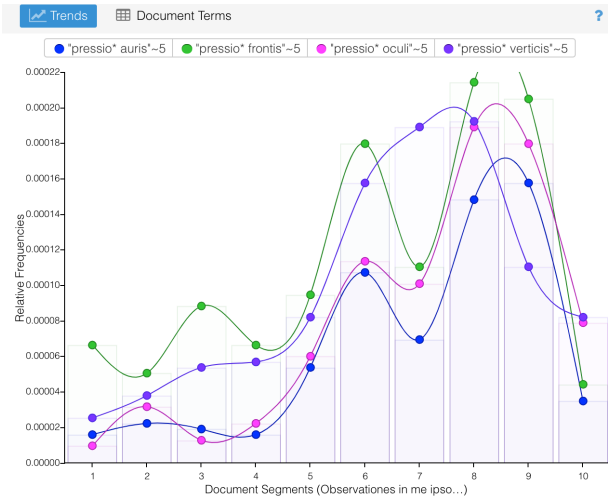


Abbildung 3: Verteilung von Druckempfindungen auf verschiedene Körperteile

Auch Senckenbergs Essgewohnheiten lassen sich nur mit begriffsbasierten Suchabfragen analysieren. Die Suche nach

dem Schlüsselwort „bibi" (ich trank) im Kontext der fünf angrenzenden Wörter zeigt beispielsweise, dass Senckenberg zwar überwiegend Wasser und Tee trank, aber bereits an dritter Stelle der Wein folgte (Abb. 4). Berücksichtigt man jedoch auch andere Getränke wie Kaffee, Bier, Alantwein und Milch sowie die entsprechenden lateinischen Begriffe, so ergibt sich aus dem Verhältnis zwischen alkoholfreien und alkoholischen Getränken das eher moderate Verhältnis von 253 zu 105. Übermäßiger Alkoholkonsum konnte jedoch Gottes Unwillen hervorrufen und durch körperliche und mentale Beschwerden bestraft werden, die sich gleichfalls in den Aufzeichnungen niederschlagen. Die Akribie der Senckenbergischen Selbstbeobachtung ermöglicht es, Vergleiche zwischen den in mehreren solcher Situationen auftretenden Symptomen anzustellen, die als Muster visualisiert und auf Ähnlichkeiten und Abweichungen durch Begleitfaktoren untersucht werden können. Abb. 5 zeigt etwa, dass sich die körperlichen Auswirkungen des Weinkonsums im warmen Monat September (a), den der Diarist für zahlreiche Freiluftaktivitäten nutzte, deutlich von denen im kälteren November (b) unterscheiden, welchen der Autor größtenteils daheim verbrachte. Noch aussagekräftiger sind die Ergebnisse einer Symptomanalyse auf Wochen- oder Tagesbasis. Dabei ist weniger bedeutsam, ob und wie die Symptome kausal zusammenhängen: Wichtiger ist es zu zeigen, dass und auf welche Weise sie gemeinsam auftreten, und wie sich verschiedene Situationen voneinander unterscheiden.

	Term	Collocate	Count (context)
☐	bibi*	gläser	140
☑	bibi*	wasser	133
☑	bibi*	thee	73
☑	bibi*	wein	55
☐	bibi*	glas	44
☐	bibi*	edi	33
☐	bibi*	butter	20
☑	bibi*	vinum	19
☑	bibi*	coffee	19
☐	bibi*	butterbrodt	18
☑	bibi*	alantwein	18
☑	bibi*	aquam	16
☐	bibi*	ructus	15
☐	bibi*	kuchen	13
☑	bibi*	bier	13
☐	bibi*	äpfel	13
☐	bibi*	asse	13
☐	bibi*	tasses	12
☑	bibi*	milch	12

Abbildung 4: Meistkonsumierte Getränke zwischen August und Dezember 1732

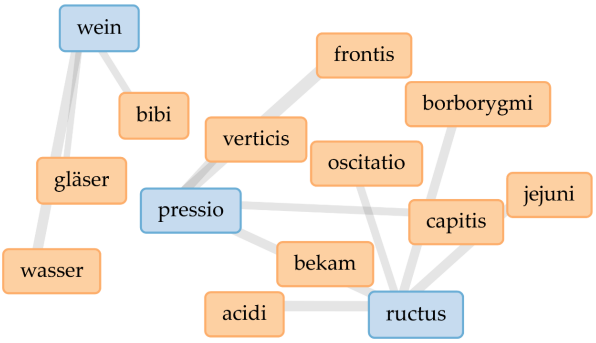


Abbildung 5a: Korrelationen zwischen Weinkonsum und Körperempfindungen in zwei verschiedenen Situationen

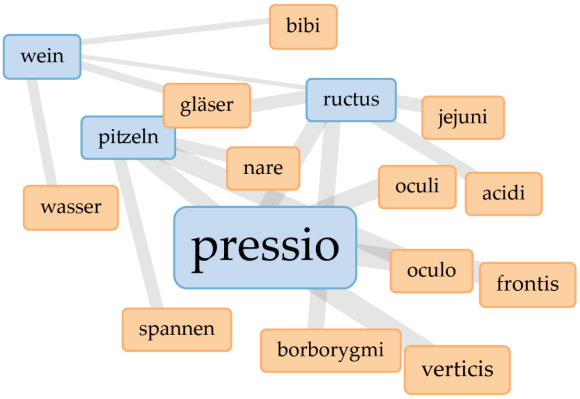


Abbildung 5b: Korrelationen zwischen Weinkonsum und Körperempfindungen in zwei verschiedenen Situationen

Bibliographie

Anderson, Chris (2008): "The End of Theory. Will the Data Deluge Make the Scientific Method Obsolete?" in: Edge, http://www.edge.org/3rd_culture/anderson08/anderson08_index.html [letzter Zugriff 20. Dezember 2019].

Borck, Cornelius (2017): „Big Data. Praktiken und Theorien der Datenverarbeitung im historischen Querschnitt", in: *Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin (NTM)* 25.4, 399–405.

Boyd, Danah und Crawford, Kate (2012): "Critical Questions for Big Data". In: *Information, Communication & Society* 15:5, 662–679, DOI: 10.1080/1369118X.2012.678878 [letzter Zugriff 20. Dezember 2019].

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, 105–131.

Bradley, Adam James (2012): Violence and the Digital Humanities Text as Pharmakon, in: *Proceedings of the Digital Humanities 2012.* http://www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf [letzter Zugriff 20. Dezember 2019].

Daston, Lorraine (2011): "The Empire of Observation, 1600–1800", in: Daston, Lorraine / Lunbeck, Elizabeth (eds.):

Histories of Scientific Observation, Chicago/London: University of Chicago Press, 81–113.

Faßhauer, Vera (2017): *Sacra à Deo in corde discenda, natura ex natura*. Die Observationes Johann Christian Senckenbergs als medico-theologische Aufzeichnungspraktik", in: *Berichte zur Wissenschaftsgeschichte* 40, 225–246.

Faßhauer, Vera (2018): "Accessing, Editing and Indexing Large Manuscript Collections: The Selected Edition of J. Chr. Senckenberg's Journals." In: *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)*, 30th November – 1st December 2017, Freie Universität Berlin, ed. Christian Wartena, Michael Franke-Maier and Ernesto de Luca, Berlin 2018, 31–36. https://refubium.fu-berlin.de/bitstream/handle/fub188/20535/ProcWissOrg2017.pdf [letzter Zugriff 20. Dezember 2019].

Geyer-Kordesch, Johanna (1990): „Medizinische Fallbeschreibungen und ihre Bedeutung in der Wissensreform des 17. und 18. Jahrhunderts", in: *Medizin, Gesellschaft und Geschichte* 9, 7–19.

Hess, Volker (2011): „Das Material einer guten Geschichte. Register, Reglements und Formulare", in: Dickson, Sheila / Goldmann, Stefan / Wingertszahn, Christof (eds.): *Fakta, und kein moralisches Geschwätz. Zu den Fallgeschichten im Magazin zur Erfahrungsseelenkunde (1783–1793)*, Göttingen: Wallstein, 115–139.

Jänicke, Stefan (2016): *Close and Distant Reading Visualizations for the Comparative Analysis of Digital Humanities Data*, Diss. Leipzig. http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-207418 [letzter Zugriff 20. Dezember 2019].

Lupton, Deborah (2016): *The Quantified Self. A Sociology of Self-Tracking*, Cambridge: Polity Press.

Mendelsohn, J. Andrew (2011): "The World on a Page: Making a General Observation in the Eighteenth Century", in: Daston, Lorraine / Lunbeck, Elizabeth (eds.): *Histories of Scientific Observation*, Chicago/London: University of Chicago Press, 396–420.

Moretti, Franco (2013): *Distant Reading*, London/New York: Verso.

Oertzen, Christine von (2017): „Die Historizität der Verdatung: Konzepte, Werkzeuge und Praktiken im 19. Jahrhundert", in: *Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin (NTM)* 25.4, 407–434.

Senckenberg, Johann Christian (1732): *Tagebücher*, Bd. 2: *Observationes in me ipso factae*, August–Dezember 1732, Senckenbergisches Archiv, Na 31, 2, UB Frankfurt am Main, Digitalisat unter http://sammlungen.ub.uni-frankfurt.de/senckenberg/content/pageview/5381525 [letzter Zugriff 20. Dezember 2019].

Senckenberg, Johann Christian (1735): *Briefentwurf an einen unbekannten Empfänger*, 24. Januar 1735, Senckenbergisches Archiv, Mp. 57, UB Frankfurt am Main.

Sinclair, Stéfan / Rockwell, Geoffrey (2003): Voyant Tools. http://voyant-tools.org.

Stolberg, Michael (2007): „Formen und Funktionen medizinischer Fallberichte in der Frühen Neuzeit (1500– 1800)" in: Süßmann, Johannes / Scholz, Susanne / Engel, Gisela (eds.): *Fallstudien: Theorie – Geschichte – Methoden*. Berlin: Trafo, 81–89.

Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte

Schmidt, Thomas
thomas.schmidt@ur.de
Universität Regensburg, Deutschland

Bauer, Marlene
marlene.bauer@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Habler, Florian
florian.habler@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Heuberger, Hannes
hannes.heuberger@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Pilsl, Florian
florian.pilsl@stud.uni-regensburg.de
Universität Regensburg, Deutschland

Wolff, Christian
christian.wolff@ur.de
Universität Regensburg, Deutschland

Einleitung

Die Idee des Distant Reading (Moretti, 2002) ist davon geprägt, durch den Einsatz von Methoden der computergestützten Textanalyse und Textvisualisierung große Mengen an Literatur zu explorieren, um Einsichten zu gewinnen, die mit herkömmlichen Methoden nicht möglich sind. Der Einsatz von Distant Reading wird dabei mittlerweile auch außerhalb der Literaturwissenschaften untersucht wie z.B. in den Religionswissenschaften (Pfahler et al., 2018). Im folgenden Beitrag wird ein Projekt vorgestellt, in dem der Einsatz und Nutzen von Distant Reading in ersten Analysen auf einer größeren Menge deutschsprachiger Songtexte exploriert wird. Ziel des Projekts ist es, mittels Distant Reading Unterschiede in gängigen Genres populärer Musik herauszukristallisieren.

Verwandte Arbeiten

Im Bereich des Text Mining wird die Analyse von Songtexten vor allem im Kontext von Retrieval- und Recommender-Aufgaben betrieben. Ziel ist meist die automatische Klassifikation und Vorhersage verschiedener Kategorien, z.B. dem Genre (Fell & Sporleder, 2014; De Sousa et al., 2016). Außerhalb dieses Arbeitsgebiets findet man in Bereichen der Kultur- und Literaturwissenschaften sowie der Psychologie Studien mit Songtexten als Untersuchungsgegenstand (Cole, 1971; Kuhn, 1999). Forschungsinteressen umfassen dabei Analysen spezifischen Musikern (*Beatles*, West & Martindale, 1996; Whissel, 1996, *Bob Dylan*, Whissel, 2008; Körner, 2012), Epochen (Pettijohn & Sacco, 2009), Emotionen (Napier & Shamir, 2018) oder Erfolg (Riedemann, 2012). Im Bereich der computergestützten Korpus-Analyse findet man vereinzelt Projekte für den englischsprachigen Bereich. Dabei werden beispielsweise quantitative und qualitative Methoden verknüpft, um Stil und historische Eigenheiten zu analysieren (Werner, 2012), Annotations- und Akquisemöglichkeiten von Korpora exploriert (Kreyer & Mukherjee, 2009) oder N-Gramme untersucht (Nishina, 2017). Die Analyse von deutschsprachigen Texten ist jedoch bislang selten und findet vor allem im Bereich von regionalem Rap statt (Hess-Lüttich, 2009) sowie eher qualitativ und hermeneutisch (Stiegler, 2009).

Korpus-Erstellung

Als Plattform für die Akquise der Songtexte wurde *LyricWiki*[1] gewählt. Ausgehend von aktuellen Umfragen zu den populärsten Genres in Deutschland[2] werden die folgenden vier Genres betrachtet: *Pop, Rock, Schlager* und *Rap/Hip Hop*. Für die Auswahl der Songs wurden manuell durch Analyse der deutschen Charts seit den 60er Jahren eine angemessene Anzahl der wichtigsten deutschsprachigen Genre-Vertreter aufgestellt. Dieser Schritt ist (auch) subjektiv geprägt, der Fokus auf berühmte und „typische" Vertreter der einzelnen Genres erlaubt jedoch trotzdem erste Analysen. Kritisch sei jedoch anzumerken, dass die Grenzen der Genres für einzelne Interpreten und Songs nicht immer eindeutig sind, insbesondere was Rock, Pop und Schlager betrifft. Wir haben versucht, für das vorliegende Korpus eine Auswahl mit möglichst eindeutigen Zuordnungen zu treffen.[3]

Für jeden gewählten Interpreten wurden über ein Skript alle Songtexte mit Metadaten von LyricWiki akquiriert. Die Akquise des Korpus wurde mittels eines frei verfügbaren angepassten ruby-Skripts durchgeführt[4].

Abbildung 1 illustriert Eckdaten zum Gesamtkorpus und den Künstlern. In der Spalte „Bekannte Vertreter" werden einige Künstler beispielhaft angegeben.

Genre	Künstler	Albums	Songs	Tokens	Bekannte Vertreter
Pop	22	96	1132	302614	Nena, Rosenstolz, Herbert Grönemeyer
Rap	33	129	1558	864925	Die fantastischen Vier, Samy Deluxe, Sido
Rock	20	126	1312	320751	Udo Lindenberg, Die Ärzte, Rammstein
Schlager	16	83	634	147833	Peter Maffay, Wolfgang Petry, Helene Fischer
Gesamt	91	434	4636	1636123	

Abbildung 1: Korpus-Zusammensetzung

Abbildung 2 zeigt die Songverteilung im zeitlichen Verlauf und Genre-Kontext auf.

	Pop	Rap	Rock	Schlager	All
60er	-	-	-	10	10
70er	10	-	51	41	102
80er	101	-	132	85	318
90er	98	108	194	141	541
00er	459	694	632	264	2049
10er	463	756	303	94	1616
All	1132	1558	1312	634	4636

Abbildung 2: Genre und zeitlicher Verlauf des Korpus

Im Bereich des Preprocessing wurden Stoppwörter entfernt und alle Wörter zu Normalisierungszwecken in Kleinschreibung gebracht.

Methoden und Ergebnisse

Für die allgemeine Textanalyse und das Topic Modeling wurden alle Analysen mittels *R* und unterschiedlichen Bibliotheken wie dem *NLP*[5]- und *topicmodels*-package[6] durchgeführt. Die Sentiment Analysis wurde mit *Python* und *SentiWS* (Remus et al., 2010) implementiert.

Allgemeine Textanalyse

Die Repetition von besonders bedeutenden Wörtern ist ein gängiges Stilmittel bei der Gestaltung von Songtexten. Aus diesem Grund betrachten wir die Analyse der häufigsten Wörter von Songtexten als besonders aufschlussreich. Die folgenden Bilder (Abbildung 3-6) illustrieren die 10 häufigsten Wörter (Most Frequently Used Words; MFWs) der einzelnen Genres.

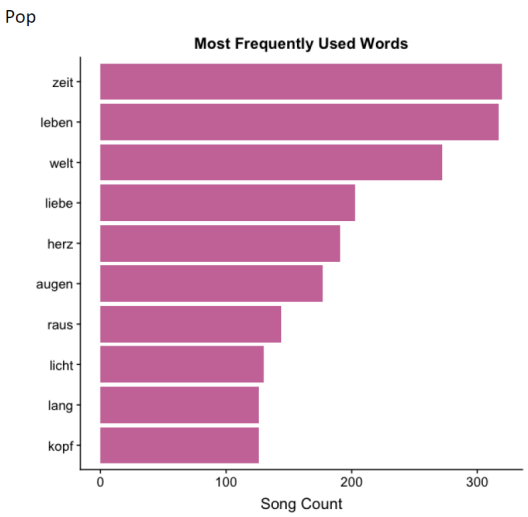


Abbildung 3: MFWs für Pop

Rap

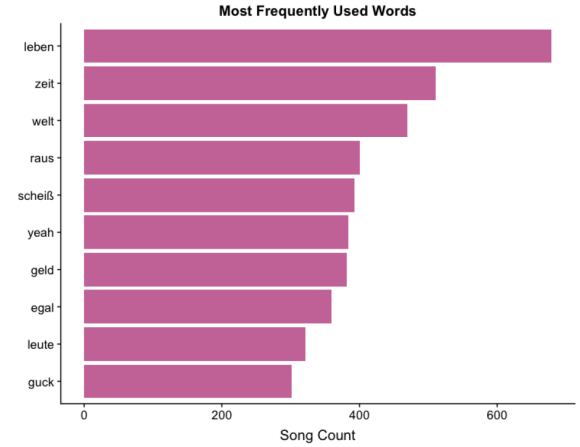


Abbildung 4: MFWs für Rap

Rock

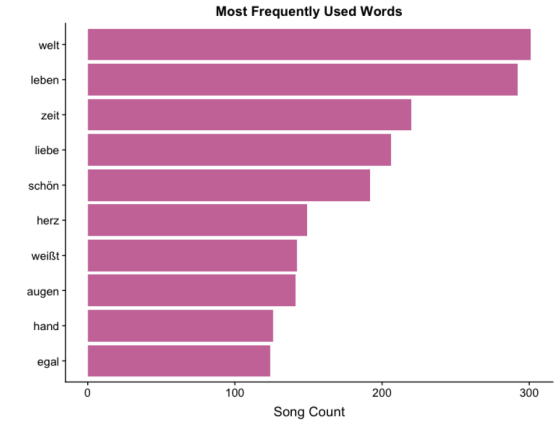


Abbildung 5: MFWs für Rock

Schlager

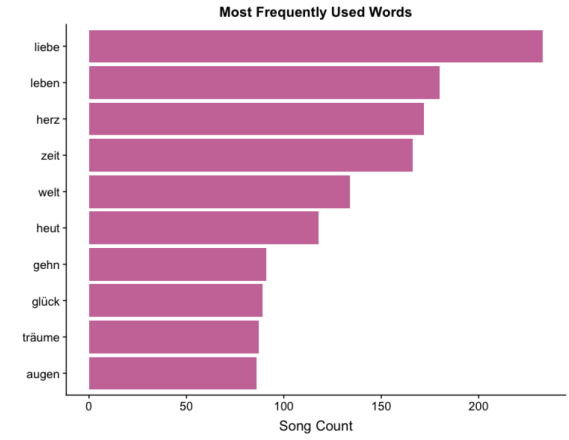


Abbildung 6: MFWs für Schlager

Man erkennt, dass es drei Wörter gibt, die in allen vier Genres gleichmäßig stark vertreten sind: „Welt“, „Leben“ und „Zeit“. Diese Konzepte sind demnach konsistenter Inhalt deutschsprachiger Liedtexte unabhängig vom Genre. Die größte Differenzierung zeigen die Genres Rap, in dem Terme der Umgangs- und Jugendsprache enthalten sind, aber auch thematische Schwerpunkte deutlich werden („Geld“) sowie das Genre Schlager, das vor allem von emotionalen Termen wie „Liebe“, „Herz“ oder „Glück“ dominiert wird.

Sentiment Analysis

Sentiment Analysis ist die Methodik zur computergestützten Analyse von Sentiments in Texten, also ob und in welchem Ausmaß Wörter eines Textes eher positiv oder negativ konnotiert ist (Liu, 2016). In den Digital Humanities werden häufig lexikonbasierte Methoden zur Bestimmung von Sentiment-Werten eingesetzt (Mohammad, 2011; Nalisnick & Baird, 2013). Dabei wird durch Summenbildung von Sentiment-Werten von Wörtern die Gesamtpolarität einer Texteinheit ermittelt. Wir verwenden dabei das etablierte Sentiment-Lexikon *SentiWS* (Remus et al., 2010). Abbildung 7 illustriert einige Ergebnisse:

Genre	Häufigste positive Wörter (Häufigkeit)	Häufigste negative Wörter (Häufigkeit)	Gesamt-Polarität	Gesamt-Polarität normalisiert an der Anzahl der Tokens
Pop	liebe (628) schön (255) lieben (116) rein (198)	schwer (186) kurz (108) feuer (104) kleine (87) fallen (85)	-1373.681	-0.004539
Rap	liebe (627) rein (399) lieber (377) schön (257) reich (207)	scheiß (718) scheiße (427) hart (317) alter (271) schwer (232)	-363.398	-0.0004201
Rock	liebe (579) schön (349) lieber (210) lieb (150) rein (146)	schwer (144) scheiß (115) feuer (112) kurz (108) kalt (108)	-424.503	-0.0013234
Schlager	liebe (547) nah (115) lieb (108) lieben (104) schön (102)	feuer (114) arm (78) schwer (78) wein (55) fehlt (50)	-344.082	-0.0023275

Abbildung 7: Ergebnisse – Sentiment Analysis

Man erkennt, dass für alle 4 Genres insbesondere Varianten von Liebe einen erheblichen Beitrag zur positivien Polarität leisten. Rap grenzt sich deutlich mit für das Genre typischen Themen ab, ausgedrückt durch Wörter wie „reich“ und mit Slang („hart“, „alter“). Alle Genres weisen insgesamt auf eine negative Polarität hin. Entgegen der naiven Intuition sind die Genres „Rap“ und „Rock“ dabei noch am positivsten (gemessen an den normalisierten Werten) bewertet. Erste Analysen machen jedoch auch Probleme der lexikonbasierten Sentiment-Analyse deutlich. Die Wörter „wein“ (weinen) und „feuer“ (das Feuer) sind in SentiWS als negativ markiert, haben aber in unseren Texten oft eher positive Konnotationen. Bei dem Wort „wein“ dann, wenn dieses durch die Normalisierung von „der Wein“ hergeleitet wird. In zukünftigen Arbeiten wollen wir mit einem domänenspezifischen Lexikon arbeiten, das für die jeweilige Anwendungsdomäne optimiert ist.

Topic Modeling

Topic Modeling ist eine Methode, um den Anteil verschiedener Themen in Dokumenten zu analysieren. Ein Thema ist dabei ein selbst definiertes Label für eine Liste von Wörtern, die besonders häufig zusammen auftreten. Als Algorithmus

wurde Latent Dirichlet Allocation (LDA) gewählt (Blei et al., 2003). Das Topic Modeling wurde separat für die einzelnen Genres durchgeführt, um Unterschiede und Gemeinsamkeiten zu untersuchen. Wir sind momentan noch am Anfang der Analyse der einzelnen Topics, aber neben Differenzen werden auch Topics gefunden, die ähnliche Konzepte widerspiegeln. Folgende Visualisierungen geben die Wortlisten wider, die wir jeweils als das Topic „Liebe“ in den einzelnen Genres benannt haben. Die Wortgröße gibt die Häufigkeit des Wortes im jeweiligen Sub-Korpus wider (Abbildung 8).

Abbildung 8: Wortlisten für das Topic „Liebe“

Auffällig ist, dass insbesondere bei Rap familiäre Begriffe wie „Mama“, „Vater“ oder auch „Bruder“ Bestandteil des Topics sind, was traditionellerweise ein häufiger Schwerpunkt im Rap-Genre ist.

Ausblick

In unseren zukünftigen Arbeiten wollen wir insbesondere das Korpus systematisch vergrößern und verbessern. Momentane Probleme sind z.B. die Ungleichverteilung in der Menge bezüglich der Genres aber auch ein Fokus auf eher aktuelle Künstler. Wenngleich wir schon erste Eigenheiten der Genres feststellen konnten, wollen wir Methoden wie Sentiment Analysis und Topic Modeling noch weiter explorieren, indem wir beispielsweise die Varianz der Sentiments untersuchen. Des Weiteren wollen wir unsere Arbeit aber auch auf andere Textanalyse-Möglichkeiten wie Kollokationsprofile von Keywords, Named Entity Recognition und Stilometrie ausweiten. Durch die Zusammenarbeit mit Musik- und Literaturwissenschaftlern wollen wir in Zukunft auch explorieren, welche weiteren Forschungsfragen mit Hilfe größerer Korpora und Distant Reading-Methoden beantwortet werden können.

Fußnoten

1. https://lyrics.fandom.com/wiki/LyricWiki
2. https://de.statista.com/statistik/daten/studie/171224/umfrage/beliebteste-musikrichtungen/
3. Das Korpus kann auf Anfrage per Mail erhalten werden.
4. https://gist.github.com/siavashs/3556469
5. https://cran.r-project.org/web/packages/NLP/index.html
6. https://cran.r-project.org/web/packages/topicmodels/index.html

Bibliographie

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

Cole, Richard R. (1971): "Top songs in the sixties: A content analysis of popular lyrics", in: *American Behavioral Scientist* 14 (3): 389-400.

De Sousa, Jefferson Martins / Eanes Torres, Pereira / Luciana Ribeiro, Veloso (2016): "A robust music genre classification approach for global and regional music datasets evaluation", in: *IEEE International Conference on Digital Signal Processing (DSP).*

Fell, Michael / Caroline Sporleder (2014): "Lyrics-based analysis and classification of music", in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics.*

Hess-Lüttich, Ernest WB. (2009): "Rap-Rhetorik. Eine semiolinguistische Analyse schweizerischer rap-lyrics", in: *Ars Semeiotica 32.*

Körner, Stefan (2012): "Bob, Pop, Bibel-die Spuren der Bibel in den Songtexten Bob Dylans." *Pastoraltheologie* 101 (12): 503-521.

Kreyer, Rolf / Joybrato Mukherjee (2007): "The style of pop song lyrics: A corpus-linguistic pilot study." in: *Anglia-Zeitschrift für englische Philologie* 125 (1): 31-58.

Kuhn, Elisabeth D. (1999): "'I just want to make love to you'–Seductive strategies in blues lyrics", in: *Journal of pragmatics* 31 (4): 525-534.

Liu, Bing (2016): *Sentiment analysis: Mining opinions, sentiments, and emotions.* New York: Cambridge University Press.

Mohammad, Saif (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 105-114.

Moretti, Franco (2002): “Conjectures on World Literature” in: *New Left Review* Jan / Feb: 54–68.

Napier, Kathleen / Lior, Shamir (2018): "Quantitative Sentiment Analysis of Lyrics in Popular Music", in: *Journal of Popular Music Studies* 30 (4): 161-176.

Nalisnick, Eric T. / Baird, Henry S. (2013): "Character-to-character sentiment analysis in shakespeare's plays.“, in *: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 479–483.

Nishina, Yasunori (2017): "A study of pop songs based on the billboard corpus." in: *Int. J. Lang. Linguist* 4 (2): 125-134.

Pettijohn, Terry F. / Donald F. Sacco Jr. (2009): "The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat", in: *Journal of Language and Social Psychology* 28(3): 297-311.

Pfahler, Lukas / Elwert, Frederik / Tabti, Samira / Morik, Katharina / Krech, Volker (2018): "Versuche zum distant reading religiöser Online-Foren“: in *Book of Abstracts, DHd 2018.*

Remus, Robert / Quasthoff, Uwe / Gerhard, Heyer (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.", in: *LREC:* 1168-1171.

Riedemann, Frank (2012): "Computergestützte Analyse und Hit-Songwriting.", in: *Black box pop. Analysen populärer Musik*: 43-56.

Stiegler, Christian (2009): *Nur ein Wort.* Dissertation, Universität Wien.

Werner, Valentin (2012): "Love is all around: A corpus-based study of pop lyrics." in: *Corpora* 7 (1): 19–50.

West, Alan / Colin Martindale (1996): "Creative trends in the content of Beatles lyrics" *Popular Music & Society* 20 (4): 103-125.

Whissell, Cynthia (1996): "Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon", in: *Computers and the Humanities* 30 (3): 257-265.

Whissell, Cynthia (2008): "Emotional fluctuations in Bob Dylan's lyrics measured by the Dictionary of Affect accompany events and phases in his life", in *: Psychological reports* 102 (2): 469-483.

Der Event Crawl als Ansatz für den Aufbau von Webarchiven am Beispiel von politischen Wahlkämpfen

Eckl, Markus

markus.eckl@uni-passau.de
Uni Passau, Deutschland

Gassner, Sebastian

Sebastian.Gassner@uni-passau.de
Uni Passau, Deutschland

Die vielfältigen internationalen Aktivitäten im Handlungsfeld Webarchivierung zeigen, dass der Aufgabe, die Inhalte im Web als eine neue Form von Quellenmaterial für die Wissenschaft dauerhaft zu sichern und zugänglich zu halten, in den Gedächtnis- und Forschungsinstitutionen inzwischen eine große Bedeutung beigemessen wird (Aubry 2010). So nimmt der Aufbau und die Erforschung von Webarchiven auch in den Sozial- und Geisteswissenschaften eine immer wichtigere Rolle ein (Brügger 2012; Milligan et al. 2019). Zu konstatieren ist aber, dass sowohl die Methoden ihrer Erzeugung, sowie die Analyse fachwissenschaftlicher Fragestellungen noch nicht hinreichend erforscht sind.

Politische Wahlkämpfe sind beispielsweise nicht nur für die Politikwissenschaft, sondern auch für andere sozial- und geisteswissenschaftliche Disziplinen ein wichtiges Forschungsfeld. Angesichts der zunehmenden Verlagerung von Wahlkämpfen auf virtuelle Arenen ergeben sich methodische Herausforderungen, wie diese politischen Diskurse beobachtet und analysiert werden können. So steigt der Umfang von zu untersuchenden Inhalten und Diskursen enorm an und herkömmliche, qualitative Analysemethoden stoßen an ihre Grenzen. Zum anderen verschwinden nach einiger Zeit immer mehr relevante Inhalte im Web und können nicht mehr abgerufen werden. In diesem Zusammenhang wird der Aufbau und die Nutzung von Webarchiven immer bedeutender, wobei angesichts des diachronen Verlaufs von Wahlkämpfen auch spezifische Strategien des Web-Crawlings notwendig sind (Eckl & Rehbein 2018). Der Event Crawl, als ein möglicher Ansatz der Webarchivierung, kann diese besonderen Anforderungen nicht nur berücksichtigen, sondern er ermöglicht auch die Archivierung digitaler Diskurslandschaften von Ereignissen (Brügger 2012, Rogers 2019).

Das Poster möchte auf Grundlage von zwei durchgeführten Event Crawls, dem bayerischen Landtagswahlkampf 2018 und dem Europawahlkampf 2019, die methodischen Herausforderungen und deren Lösungsansätze darlegen und die Möglichkeiten hinsichtlich der Analyse dieses Webarchives mit Methoden der Digital Humanities diskutieren. Bei beiden Wahlkämpfen wurden Webseiten von Medienhäusern, Parteien und Politikern (+social media accounts) mehrfach gecrawlt und daraus ein Archiv mit einem Umfang von mehr als 4 TB aufgebaut.

In Abgrenzung zum Web Scraping einerseits, bei dem Inhalte von möglichst vielen Webseiten automatisiert gecrawlt werden, und dem Selektiven Crawling andererseits, bei dem eine sehr begrenzte Anzahl von Webseiten gecrawlt werden, besteht hinsichtlich des Event Crawls die Herausforderung, die Grenzen des Events und somit die relevanten Webinhalte zu bestimmen. Durch eine sachliche, zeitliche und akteurszentrierte Eingrenzung des Gegenstandes ist es möglich, relevante Webseiten und -inhalte zu bestimmen und dieses Vorgehen kann einen wichtigen Beitrag hinsichtlich der Diskussion über die Vollständigkeit von Webarchiven liefern (Weber & Napoli 2018). Wir teilen hier die Auffassung von Brügger (2018), dass Webarchive in der Regel unvollständig sind und eine hohe Selektivität aufweisen. Auch wenn durch unsere Eingrenzungen sich kein vollständiges Webarchiv aufbauen lässt, ist es dadurch dennoch möglich, approximativ diesem Ziel näher zu kommen. Vielmehr soll durch die gewählte Abgrenzung des Untersuchungsgegenstandes, die zentralsten Diskurse der beiden Wahlkämpfe erfasst werden. Im Gegensatz zu manch anderen Webarchiven, die eine hohe Selektivität aufweisen, nicht zuletzt aufgrund anderer Crawlmethoden, sind hier vielversprechende Ergebnisse zu erwarten.

Die zweite Herausforderung ergibt sich hinsichtlich der zeitlichen Taktung der durchgeführten Crawls. Denn durch die diachrone zeitliche Entwicklung von Wahlkämpfen, wie zum Beispiel dem diachronen Posten von Inhalten auf Blogs oder Medienwebsteinen, sowie dem Verschwinden und dem Löschen von Webinhalten noch während des Beobachtungszeitraums, muss ein und dieselbe Webseite mehrfach gecrawlt werden. Zu diskutieren ist, welche Taktungen für welche Webseiten notwendig sind und inwieweit sich durch unterschiedliche Vorgehensweisen die erstellten Korpora unterscheiden. Es muss auch die Frage geklärt werden, ob gesamte Webseiten oder nur relevante Inhalte der Webseiten häufiger zu crawlen sind. Neben ökonomisch und technisch beschränkten Mitteln ist hier ein Vorgehen zu wählen, welches sich auch an mögliche fachdisziplinäre Forschungsfragestellungen orientiert.

Ebenfalls soll die Schnittstellen von WARC Dateien diskutiert werden, die genutzt werden können, um analysefähige Korpora zu erstellen. Eine WARC ist ein genormtes Archivformat, das die Inhalte der gecrawlten Webseiten, sowie Metadaten zu den spezifischen Crawls enthält. Dieser Vorgang ist von großer Bedeutung, da dabei nicht nur die WARC Datei entpackt wird, sondern es findet auch eine Filterung, Gruppierung und Extraktion der Daten statt. Auch wenn dafür bereits

einige Programme entwickelt wurden, wie zum Beispiel "ArchivSpark" (Holzmann, Goel & Anand 2017) oder das "Archive Unleashed Toolkit" (Lin et al. 2017), braucht es zum Beispiel für den automatisierten Aufbau eines hochwertigen Textkorpus mit Metadaten aus den Webseiten, eine an den spezifischen Webseiten orientierte Extraktion der Textinhalte. Diese Anforderung ergibt sich, weil Webseiten von verschiedenen Quellen in ihrem Aufbau unterschiedlich sind. Zusätzlich können Webseiten im Laufe der Zeit ihr Aussehen verändern, wodurch in einem Webarchiv für eine Quelle mehrere Layouts enthalten sein können. Wurde nun bei der Verarbeitung einer relevanten Webseite ein Layout identifiziert, wird die Position der gewünschten Daten mit Hilfe von sogenannten *CSS Paths* spezifiziert und die Extraktion kann erfolgen. Nach der Extraktion werden die Daten in einer Mongodb Datenbank zur weiteren Verarbeitung abgelegt.

Nach den methodologischen Überlegungen hinsichtlich des Aufbaus eines Webarchievs, der Beschreibung des Event Crawls und der Erstellung einer MongoDB Datenbank mit Metadaten aus den WARC Dateien (wie z.B. Datum, Überschrift, Verfasser), ist es auch unter Zuhilfenahme von Methoden der Digital Humanities nun möglich, Archive auf Basis fachwissenschaftlicher Fragestellungen zu untersuchen. Auf Grundlage des beschriebenen Archivs zum Europawahlkampf 2018 können unterschiedliche politikwissenschaftliche Forschungsfragen gestellt werden. Exemplarisch kann untersucht werden, welche Themen auf den Parteienwebsetien im Rahmen des Europawahlkampf 2018 diskutiert wurden. Weiter kann danach gefragt werden, wie die Themenkonjunktur im Laufe des Wahlkampfes war? Eine solche fachwissenschaftliche Fragestellung kann untersuchen, inwieweit die Europawahl 2018 als eine "second order election" zu verstehen ist (Hix, S. & Marsh 2011, Weber 2009). Darunter versteht man den Sachverhalt, dass in europäische Wahnen häufig nationale und nicht europäische Themen im Wahlkampf diskutiert werden. Um diese Fragestellung zu bearbeiten wurde aus der MongoDB Datenbank ein spezifischer Textkorpus mit zusätzlichen Metadaten aus den jeweiligen Webseiten erstellt. Für die Untersuchung und Ermittlung von Wahlkampfthemen fanden Methoden des Topic Modelings Anwendung, wobei hierfür das R Package "Structural Topic Modeling" von Roberts et al. (2019) genutzt wurde. Für weitere Analysen wurde zudem unter anderem auf Methoden der Netzwerkanalyse zurückgegriffen. Erste Ergebnisse können auf GitHub eingeshen werden (https://github.com/MarkusEckl/web_archive_and_stm).

Bibliographie

Aubry, Sara (2010): *Introducing Web Archives as a New Library Service: the Experience of the National Library of France*, in: *LIBER Quarterly*, 20(2), S. 179–199. DOI: http://doi.org/10.18352/lq.7987

Brügger, Niels (2012): *Historical Network Analysis of the Web*in: Social Science Computer Review 31 (3), 306–321.DOI: https://doi.org/10.1177/0894439312454267

Eckl, Markus / Rehbein, Malte (2018): Methoden der Digital Humanities in Anwendung für den Aufbau und die Nutzung von Webarchiven. Konferenz zur Bewahrung digitalen kulturellen Erbes. Deutsche Nationalbibliothek. Frankfurt am Main. https://www.dnb.de/SharedDocs/Downloads/DE/Kulturell/konferenzeDigKultErbe2018_MarkusEckl.html?nn=56454 (letzter Zugriff 12. September 2019).

Holzmann, Helge / Goel, Vinay / Anand, Avishek (2016): *ArchiveSpark*in: Nabil R. Adam / Boots Cassel / Yelena Yesha / Richard Furuta / Michele C. Weigle (Hg.): JCDL'16. Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries : June 19-23, 2016, Newark, NJ, USA. DOI: https://doi.org/10.1145/2910896.2910902

Lin, Jimmy / Milligan, Ian / Wiebe, Jeremy / Zhou, Alice (2017): *Warcbase* in: Journal of Cultural Heritage 10 (4), 1–30.

Roberts, Molly E., / Stewart, Brandon M. / Tingley, Dustin (2019): stm: An R Package for Structural Topic Models. *Journal of Statistical Software*,**91** (2), 1–40. DOI: 10.18637/jss.v091.i02.

Rogers, Richard (2019): „Periodizing Web Archiving: Biographical, Event-Based, National and Autobiographical Traditions" in: Brügger, Niels / Milligan Ian (Hg.): *The SAGA Handbook of Web History*. London: SAGE.

Weber, Matthew S. / Napoli, Philip M. (2018): *Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism* in: Digital Journalism, 6:9, 1186-1205. DOI: https://doi.org/10.1080/21670811.2018.1510293

Der Haken von Frazier, der Ali 1971 zu Boden schickte, hat jetzt eine URI: Zur Modellierung, Transkription und Visualisierung performativer kultureller Objekte

Geißler, Nils
nils.geissler@uni-koeln.de
Universität zu Köln, Deutschland

Sport ist ein wesentlicher Bestandteil unserer Kultur (Hitzler 1991: 485f), der Boxsport war bereits bei den olympischen Spielen der Antike in Form des antiken Faustkampfes vertreten (Rudolph 1965: 8ff). Im letzten Jahrhundert wurden Boxkämpfe teilweise stark politisiert und erlangten dadurch Bedeutung auch außerhalb der sportlichen Domäne selbst (Hughes 2018 und Bloom & Willard 2002). Als komplexe Ereignisse werden die Kämpfe zunächst durch die Performativität der einzelnen Handlungen ihrer Akteure konstituiert: (Meid-) Bewegungen im Ring und Faustschläge (Fischer-Lichte 2004). Die Regeln, die das Boxen ausmachen, bieten den Akteuren im konkreten Ereignis eines Boxkampfes einen Spielraum, um unterschiedliche Techniken auszuführen, die ihnen im kompetitiven Vergleich zum Sieg verhelfen sollen (Fiedler 1983: 63). Die Bandbreite der nach den Regeln des modernen Boxens möglichen Techniken ist bereits 1963 von Horst Fiedler (Fiedler 1963) in einem Schema erfasst und dabei gewissermaßen nicht-formal modelliert worden. Auch

wenn Boxtechniken in konkreter Anwendung gegenüber den üblichen kulturellen Objekten martialisch anmuten, erscheint es aus oben genannten Gründen sinnvoll, sie dennoch als solche anzusehen und das Ziel ihrer Analyse auf der Grundlage ihrer Verdatung auszugeben. Dazu wurde ein Workflow erarbeitet, der es ermöglicht, die Boxtechniken als konkrete Umsetzungen der von Horst Fiedler ermittelten Schlag- und Bewegungsvarianten zu interpretieren und formal zu erfassen, um sie einer weiteren Auswertung zugänglich zu machen. Auf der Basis des Fiedler-Schemas wurde ein Modell erstellt, das bei der Transkription von Videoaufzeichnungen als Grundlage für mehrere Annotationsebenen und das kontrollierte Vokabular dient. Als Datengrundlage werden dabei Videos von Boxkämpfen auf der Plattform YouTube genutzt. Im Ergebnis entsteht eine Annotation der Aufnahmen von Boxtechniken, die durch das verwendete Modell im Kontext des entsprechenden Boxkampfes und der einzelnen Runden verknüpft sind. Dadurch erhält jede erfasste Aktion einen eigenen Unique Resource Identifier (URI), über den nicht nur ein sekundengenauer Zugriff auf die jeweilige Stelle im Video möglich wird, sondern auch eine Verknüpfung mit anderen Instanzen der jeweiligen Boxtechnik. Die Modellierung von Boxkämpfen bringt somit den Vergleich von Aufnahmen von Umsetzungen bestimmter Techniken zu Wege, beispielsweise von linken Haken in den 1970er Jahren mit denen in heutigen Kämpfen. Dies wird durch die Verwendung von CIDOC CRM (http://www.cidoc-crm.org/) als für das Modell zugrundeliegende Ontologie realisiert. CIDOC CRM ist eine etablierte Technologie zur Erfassung kultureller Objekte (Doerr 2003).
Der Workflow besteht aus den folgenden Schritten: Die Transkription beziehungsweise Annotation der Videos erfolgt in ELAN (https://tla.mpi.nl/tools/tla-tools/elan/), die erzeugten Daten liegen in einem XML-Format vor. Die Transkriptionsdaten werden in oXygen mittels XSLT vorverarbeitet und als Triples formuliert. Die Codierung selbst geschieht in RDF/XML (unter Verwendung von CIDOC CRM Version 6.2). Die erzeugten RDF-Triples werden am Ende des Workflows in ein RDF4J-Repository (https://rdf4j.eclipse.org/) überführt und können von dort aus mit SPARQL abgefragt werden. Die Befüllung des Triple Store geschieht über die zugehörige RDF4J API, die durch eine Java-Anwendung angesprochen wird.

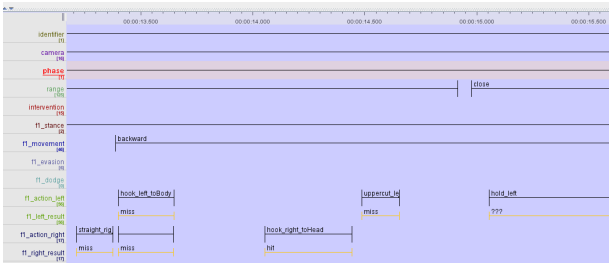


Abbildung 1: Screenshot von Annotation in ELAN mit den verschiedenen Annotationsebenen (Tiers).

Anschließend können auf Basis der verknüpften Daten Analysen durchgeführt und Visualisierungen erstellt werden, anhand derer an die der ausgeführten Technik entsprechenden Stelle in der Videoaufnahme gesprungen werden kann. Zur Veranschaulichung der Daten wurden zwei verschiedene Visualisierungsformen gewählt: In der ringförmigen Visualisierung werden die einzelnen Aktivitäten pro Runde auf einem Kreis dargestellt. Dabei werden die unterschiedlichen Aktivitätstypen durch entsprechende Symbole verdeutlicht. Die zweite Visualisierung ordnet die durch die Akteure durchgeführten Ereignisse auf einer Art Zeitstrahl an. Zudem wird die zugrundeliegende Videoaufnahme von YouTube direkt auf der Seite mit eingebunden.

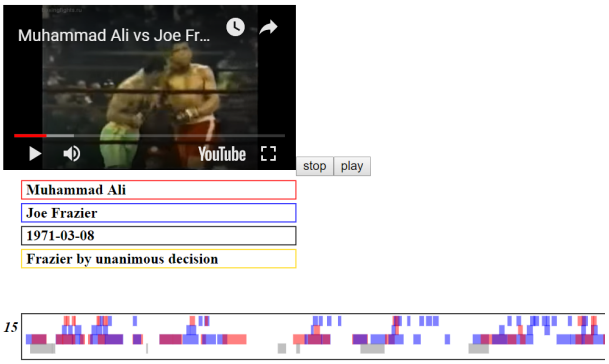


Abbildung 2: Screenshot von Visualisierung als horizontale Timeline mit eingebettetem Video.

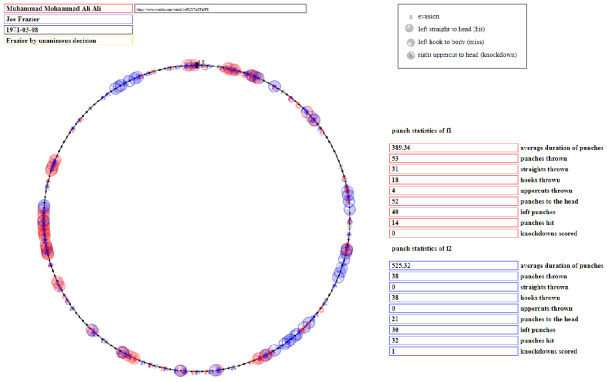


Abbildung 3: Screenshot von Visualisierung als Kreis mit Auswertungen der annotierten Entitäten.

Am Beispiel des Boxsports wird gezeigt, wie durch Videoaufnahmen festgehaltene Ereignisse und deren Teilereignisse in einem Datenmodell erfasst werden können. Diese werden anschließend über Abfragen zugänglich gemacht, zudem wird über Visualisierungen die Exploration der Daten ermöglicht. Durch die Verwendung von CIDOC CRM werden die Daten semantisch modelliert. Dieser Ansatz sollte auf andere Bereiche performativer Kulturobjekte übertragbar sein, für die ähnliche Daten vorliegen. Dabei müsste lediglich das Modell beziehungsweise die Modelle an die entsprechende Domäne angepasst werden.
Das Poster soll den Workflow diagrammatisch darstellen und Einblicke in die verschiedenen Stationen, vor allem in das verwendete Modell und die am Ende entstehenden Visualisierungen, liefern.

Bibliographie

Bloom, John / Willard, Michael Nevin (2002): "Introduction. Out of Bounds and between the Lines: Race in Twentieth-Century American Sport" in: Bloom, John / Willard,

Michael Nevin (eds.): *Sports Matters: Race, Recreation, and Culture*. New York & London: New York University Press 1–10.

Doerr, Martin (2003): "The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata", in: *AI Magazine* 24.

Fiedler, Horst (1983): *Boxsport*, 2nd ed. Berlin: Sportverlag Berlin.

Fiedler, Horst (1971): *Boxen.* Eine Anleitung für die Ausbildung der Anfänger, 3rd ed. Berlin: Sportverlag Berlin.

Fischer-Lichte, Erika (2004): *Ästhetik des Performativen*. Frankfurt am Main: Suhrkamp.

Hitzler, Ronald (1991): "Ist Sport Kultur?", in: *Zeitschrift für Soziologie* 20: 479–487.

Hughes, Jon (2018): *Max Schmeling and the Making of a National Hero in Twentieth-Century Germany*. Cham: Springer International Publishing.

Die „Hans Kelsen Werke" (HKW) – eine rechtswissenschaftliche Hybridedition

Reinthal, Angela
angela.reinthal@jura.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg

Tscheu, Amelie
amelie.tscheu@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Trautmann, Marjam
marjam.trautmann@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Seit März 2006 entsteht in der Hans Kelsen-Forschungsstelle (bis 2011 in Erlangen, seitdem in Freiburg i. Br.) die historisch-kritische Ausgabe der Werke des österreichischen Rechtstheoretikers Hans Kelsen (1881–1973). Mit der Aufnahme in das Programm der Akademie der Wissenschaften und der Literatur | Mainz im Jahr 2018 wurde eine weitere Arbeitsstelle für die digitale Komponente der Edition in Frankfurt a. M. etabliert.
Die im Mohr Siebeck Verlag erschienenen Druckbände werden nach dem *Moving-Wall*-Prinzip ebenfalls als digitale Edition aufbereitet und unter einer CC BY 2.0 Lizenz zur Verfügung gestellt. Fünf der geplanten 35 Bände (Jestaedt 2007–2013) sind bereits vor 2018 als Print erschienen und werden nachträglich für das Digitale aufbereitet, während zukünftig die digitale und die analoge Form der Edition *single source* in einer XML-basierten digitalen Editionsumgebung erarbeitet wird. Somit werden für die (Rechts-)Wissenschaften maschinenlesbare Forschungsdaten zum Werk einer zentralen rechtshistorischen Figur des 20. Jahrhunderts nachnutzbar.

Wie bei vielen Editionsprojekten, die bisher in rein analoger Form zur Verfügung standen und nun eine „digitale Wende" vollziehen, stellen sich für die HKW vielschichtige Herausforderungen. Diese betreffen insbesondere die Beschaffenheit der Datengrundlage und deren Aufbereitung, die Umstellung bisher etablierter Redaktionsprozesse sowie Lern- und Lehrabläufe der digitalen und der klassischen Geisteswissenschaften. In unserem Poster stellen wir dar, inwiefern etablierte Workflows und Standards der Digital Humanities für die „Hans Kelsen Werke" eingesetzt werden und geben einen Ausblick auf den Wert für die (digitale) Rechtswissenschaft insgesamt.

Datengrundlage und -modellierung

Vor Abschluss der Redaktionsumstellung in eine XML-basierte digitale Editionsumgebung bilden die Drucksatzdaten der nach bisherigem Verfahren erstellten Bände (innerhalb des Textverarbeitungsprogramms Microsoft-Word) die Datengrundlage für die digitale Edition der entsprechenden Werke Hans Kelsens. Die Datenmodellierung der Texte und der Register anhand der Guidelines der *Text Encoding Initiative* (TEI) in XML orientiert sich am Basisformat des Deutschen Textarchivs[1] und wurde um projektspezifische Anforderungen erweitert. Eine Besonderheit der Texte im Vergleich zur herkömmlichen Quellenedition besteht beispielsweise im doppelten Fußnotenapparat – demjenigen Hans Kelsens und den Anmerkungen der HKW-Editor*innen – sowie der Heterogenität der Texte an sich (Buchbesprechungen, Gesetzestexte, Aufsätze, Monografien). Auch der vielfältige Einsatz des editorischen Fußnotenapparats der HKW birgt für die Übertragung in eine semantische Kodierung der Texte kreative Möglichkeiten.

Die Bände der HKW liefern umfangreiche und für die Kelsen-Forschung unverzichtbare Personen- und Sachregister. Die einzelnen Personenregister wurden zusammengeführt und mit entsprechenden Normdaten der *Gemeinsamen Normdatei* (GND) versehen. Sie bilden ein digitales Register, welches bereits vor der endgültigen Umstellung des Redaktionsprozesses in eine digitale Editionsumgebung in die reguläre Editionsarbeit integriert wird. Ebenso wurden die Schriftenverzeichnisse der bisher publizierten Bände homogenisiert in das Literaturverwaltungssystem *Zotero*[2] übertragen, in der zukünftig die Literatur nicht nur verwaltet, sondern auch in die neue Arbeitsumgebung integriert wird. Eine besondere Herausforderung stellt das heterogene und komplexe – und dafür umso bedeutendere – Sachregister dar, welches perspektivisch eine Grundlage für die Erarbeitung einer Ontologie in den digitalen Rechtswissenschaften und damit einen Einstieg der Fachrichtung in das Feld der *Linked Open Data* darstellen kann. Die datenbasierte Modellierung der Hans Kelsen Werke bietet somit vielfältige innovative Spielräume für die Digital Humanities hinsichtlich ihrer Wirkung auf die Rechtswissenschaften.

Digitale Infrastruktur

In der digitalen Infrastruktur und Editionsumgebung der HKW werden etablierte Standards und Angebote aus den DH zur Anwendung gebracht und weiterentwickelt. Die zu edie-

renden Quellen und Forschungsdaten werden in einer Instanz der XML-Datenbank eXist-db[3] verwaltet und über eine Integration in den oXygen XML-Editor[4] im Author-Modus editorisch bearbeitet. Zum Einsatz kommt hierbei ein projektspezifisches Erweiterungsframework auf Basis von *ediarum* sowie die eXistdb-App *ediarum.db*[5].

Die Präsentation der digitalen Edition findet sich perspektivisch auf *kelsen.online*, zunächst werden hier nähere Projektinformationen, die PDF der bisher analog publizierten HKW-Bände und ein kumuliertes Gesamtregister der entsprechenden Bände zur Verfügung gestellt. Die Präsentationsschicht basiert auf dem Content Management System TYPO3[6], in das die Forschungsdaten aus der eXist-db importiert werden und redaktionelle Arbeiten an der Website stattfinden. Neben einer ansprechenden und benutzerfreundlichen Präsentation der Forschungsdaten und korpusinterner sowie -externer Interaktion werden diese über Schnittstellen beziehbar und für weitere maschinengestützte Forschungen nutzbar sein.

Von der digitalen Redaktionsumstellung profitiert auch die gedruckte Buchausgabe der Edition, welche weiterhin ein gleichwertiger Bestand des Projektes bleibt. So beispielsweise durch die Reduzierung bisheriger Arbeitsschritte und einheitliche Ansetzungen in den Verzeichnissen.

Ausblick

Die digitale Edition der Hans Kelsen Werke wird in der 25-jährigen Laufzeit des Projektes die Entwicklungen und Standards der Digital Humanities verfolgen, gegebenenfalls adaptieren und sich dem Forschungsgegenstand “Hans Kelsen” mit dem Einsatz digitaler Methoden nähern. Auch für die rechtswissenschaftliche Forschung insgesamt hoffen wir durch die Erarbeitung von Standards für die digitale Aufbereitung fachspezifischer Daten einen nachhaltigen Beitrag zu leisten.

Fußnoten

1. www.deutschestextarchiv.de/doku/basisformat/.
2. https://www.zotero.org/.
3. http://exist-db.org/exist/apps/homepage/index.html.
4. https://www.oxygenxml.com/.
5. https://github.com/ediarum.
6. https://typo3.org/.

Bibliographie

Jestaedt, Matthias (eds.) (2007): Hans Kelsen Werke. Band 1. Veröffentlichte Schriften 1905–1910 und Selbstzeugnisse. Tübingen: Mohr Siebeck.

Jestaedt, Matthias (eds.) (2008): Hans Kelsen Werke. Band 2: Veröffentlichte Schriften 1911. Tübingen: Mohr Siebeck.

Jestaedt, Matthias (eds.) (2010): Hans Kelsen Werke. Band 3: Veröffentlichte Schriften 1911–1917. Tübingen: Mohr Siebeck.

Jestaedt, Matthias (eds.) (2013): Hans Kelsen Werke. Band 4: Veröffentlichte Schriften 1918–1920. Tübingen: Mohr Siebeck.

Jestaedt, Matthias (eds.) (2011): Hans Kelsen Werke. Band 5: Veröffentlichte Schriften 1919–1920. Tübingen: Mohr Siebeck.

Reinthal, Angela (2014): „InterNationalität und InterDisziplinarität der Hans Kelsen Werke (HKW)“ in: Stolz, Michael / Chen, Yen-Chun (eds.): Internationalität und Interdisziplinarität der Editonswissenschaft (= Beihefte zu editio 38). Berlin: De Gruyter 303-314.

Nečaský, Martin / Knap, Tomáš / Klímek, Jakub / Holubová, Irena / Vidová-Hladká, Barbora (2013) Linked Open Data for Legislative Domain – Ontology and Experimental Data, in: Abramowicz W. (eds) Business Information Systems Workshops (= Lecture Notes in Business Information Processing 160). Berlin / Heidelberg: Springer Verlag 172–183.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe. (= Schriften des Instituts für Dokumentologie und Editorik 7). Norderstedt: BoD https://kups.ub.uni-koeln.de/5351/.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. (= Schriften des Instituts für Dokumentologie und Editorik 8). Norderstedt: BoD https://kups.ub.uni-koeln.de/5352/.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung. (= Schriften des Instituts für Dokumentologie und Editorik 9). Norderstedt: BoD https://kups.ub.uni-koeln.de/5353/.

Staab, Steffen / Studer, Rudi (2009) (eds.): Handbook on Ontologies. Berlin / Heidelberg: Springer Verlag.

TEI Consortium (eds.) (2019): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0 vom 17.07.2019. TEI Consortium. https://tei-c.org/guidelines/P5/ [letzter Zugriff 25.09.2019].

Digitale Editionen im Spannungsfeld zwischen Formalisierung und Interpretation: Rezensionen der Online-Zeitschrift RIDE als Gradmesser für die Zukunft

Resch, Claudia
claudia.resch@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Rastinger, Nina
ninaclaudia.rastinger@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Einleitung

Vorliegender Posterbeitrag geht davon aus, dass digitale Editionen Produkte teils formalisierender, teils interpretierender Prozesse sind und damit in ein herausforderndes „Spiel- und Spannungsfeld“ geraten. Sie sind einerseits Standards verpflichtet, vermitteln jedoch andererseits – bedingt durch Editionsentscheidungen wie Auswahl der Quellen, Modellierung und Präsentation – eine bestimmte Sicht auf das edierte Material, welche durch Wissen und Erkenntnisinteressen der jeweiligen Herausgeber*innen geprägt ist.

Somit stellt sich aber die Frage, welche Spielräume den Herausgeber*innen und welche den Benutzer*innen bei der digitalen Erschließung von Quellen zugestanden werden und wie letztere damit umgehen, dass die ihnen zur Verfügung gestellten Ressourcen bereits durch andere vorgeformt sind. Um hierauf Antworten zu finden, haben die Autorinnen das etablierte Rezensionsorgan „RIDE – A review journal for digital editions and resources“ (https://ride.i-d-e.de/) herangezogen und die Äußerungen der Rezensent*innen als stellvertretend für die Perspektiven von Nutzer*innen untersucht.

RIDE als Untersuchungskorpus

RIDE wird vom Institut für Dokumentologie und Editorik verantwortet und möchte gemäß der Eigendefinition, „ExpertInnen ein Forum zur kritischen Auseinandersetzung“ mit Editionen bieten und damit dazu beitragen, „die gängige Praxis zu verbessern und die zukünftige Entwicklung voranzutreiben“ (RIDE 2019). Insofern gehen aus den bereits erschienenen Rezensionen auch Überlegungen hervor, wie Editionen in Zukunft konzipiert werden könnten. Zudem ist der Kriterienkatalog von Patrick Sahle (in Zusammenarbeit mit Georg Vogeler und anderen Mitgliedern des IDE erstellt, vgl. http://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/), an dem sich Gutachter*innen orientieren, einerseits „Bewertungsgrundlage“ und andererseits „Checkliste für Wissenschaftler [d.h. für die Ersteller*innen digitaler Editionen]“ (vgl. Henny 2017). Nicht zuletzt sind dadurch auch Minimalanforderungen für das zeitgemäße Edieren formuliert (vgl. Schnöpf 2013: 75), welche dazu beitragen, die Qualität digitaler Editionen zu sichern.

Zum Zeitpunkt der Untersuchung waren auf der Website „ride.i-d-e.de“ im Zeitraum von 2014 bis 2019 bereits sieben Bände zu „wissenschaftlichen Editionen“ mit insgesamt 35 Rezensionen in deutscher und englischer Sprache veröffentlicht worden. Die Daten dieser Rezensionen stehen auf GitHub (https://github.com/i-d-e/ride) zur Verfügung und wurden für die Untersuchung heruntergeladen und als Textkorpus mit insgesamt 161.553 Token aufbereitet. Die Auswertung erfolgte korpusbasiert mithilfe der Suche nach ausgewählten Keywords (etwa: „leider“, „Vorteil“ oder „wünschenswert“), um relevante Textstellen schnell auffinden zu können, sowie über ein „close reading“-Verfahren, um das Verständnis der einzelnen Kontexte zu sichern. Dabei wurden die Belegstellen in Anlehnung an Sahles Kriterienkatalog inhaltlich sortiert und ausgehend von mehreren Fragen ausgewertet:

- In welchen Bereichen digitaler Editionen sehen sich Rezensent*innen durch Vorannahmen und Interpretationen eingeschränkt?
- In welchen Bereichen digitaler Editionen wäre aus Sicht der Rezensent*innen mehr Formalisierung/Standardisierung wünschenswert?
- Welche Maßnahmen schlagen Rezensent*innen vor, um neue Spielräume zu eröffnen und verschiedene Interpretationsmöglichkeiten offen zu halten?

Ergebnisse und Ausblick

In der Auswertung der 35 Rezensionen zeigt sich unter anderem, dass Gutachter*innen es zunehmend schätzen, wenn den Nutzer*innen digitaler Editionen möglichst viele unterschiedliche Perspektiven auf die jeweiligen Daten ermöglicht werden. So wird etwa die „Möglichkeit verschiedener Präsentationsmodi“ als positiv hervorgehoben, wohingegen das Fehlen von Faksimiles als Defizit gewertet wird. Der als ideal angenommene Zugang zu den Daten beinhaltet zudem in fast allen Rezensionen die Downloadbarkeit der XML/TEI-Dateien.

Diese Beobachtungen sprechen dafür, dass großes Interesse daran besteht, Daten nachzunutzen und damit zu eigenen Einschätzungen und Interpretationen zu gelangen – ein potentieller Mehrwert, welcher teils auch deutlich formuliert wird: „Die Zurverfügungstellung der Transkription in XML oder einem anderen für die Nachnutzung der Daten geeigneten Format wäre wünschenswert und wertvoll.“ Gegeben ist diese Option im überwiegenden Teil bislang rezensierter Editionsprojekte jedoch nicht, wie die von RIDE selbst generierten Auswertungen offenbaren (vgl. Chart Nr. 20 und Nr. 23 unter https://ride.i-d-e.de/data/charts/). Somit ergibt sich hier ein erster möglicher Ansatzpunkt für die Optimierung zukünftiger digitale Editionen.

Genauso zeigt sich aber auch im Bereich der Dokumentation digitaler Editionsprojekte noch Verbesserungsbedarf. Schließlich machen Rezensent*innen mehrfach auf das Fehlen editorischer Richtlinien aufmerksam und thematisieren – wie in folgendem Fall – die fehlende Transparenz und „Selbstverortung“ (Schnöpf 2013: 72) der ihnen gegebenen Ressourcen: „I do not doubt that the transcribers and editors had a clear idea of what they were doing, but they have not documented it in the edition and so the user (and the reviewer) can only retrospectively deduce what that idea might have been.”

Zu diesen (und weiteren gefundenen) Kritikpunkten kommt freilich erschwerend hinzu, dass Nutzer*innengruppen mit ihren Forschungspraktiken, Forschungsprozessen und Motivationen sehr heterogen sein können (vgl. Kramer 2016, Hewing/Mandl/Womser-Hacker 2016) und „[k]omplexe digitale Ressourcen [...] auch an die Benutzer höhere Anforderungen [stellen]“ (Sahle 2013: 262). Es gilt hier also, neue Interaktionsmuster zu entwickeln, um die Kluft zwischen Editor*innen und Nutzer*innen zu überbrücken. In diesem Sinne soll der Posterbeitrag durch abschließende Empfehlungen abgerundet werden, wie Nutzer*innen in Zukunft (noch) erfolgreicher an digitale Editionen herangeführt werden könnten.

Bibliographie

Henny, Ulrike (2018): „Reviewing von digitalen Editionen im Kontext der Evaluation digitaler Forschungsergebnisse“, in: Kamzelak, Roland S. / Steyer, Timo (eds.): *Digitale Metamorphose: Digital Humanities und Editionswissenschaft* (= Son-

derband der Zeitschrift für digitale Geisteswissenschaften 2) 10.17175/sb002_006.

Hewing, Ben / Mandl, Thomas / Womser-Hacker, Christa (2016): "Methods for User-Centered Design and Evaluation of Text Analysis Tools in a Digital History Project", in: *ASIST Annual Meeting Proceedings. Joining Research and Practice* 53: 1–10 https://onlinelibrary.wiley.com/doi/full/10.1002/pra2.2016.14505301078 [letzter Zugriff 30. Dezember 2019].

Kramer, Michael J. (2016): *The Digital Humanities Reader*. http://www.michaeljkramer.net/the-digital-humanities-reader/ [letzter Zugriff 30. Dezember 2019].

Sahle, Patrick (2013): *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1–3: Befunde, Theorie und Methodik* (= Schriften des Instituts für Dokumentologie und Editorik 7–9). Norderstedt: BoD.

Schnöpf, Markus (2013): „Evaluationskriterien für digitale Editionen und die reale Welt", in: *HiN - Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien* 27: 69–76.

Digitales Publizieren im Spiegel der Zeitschrift für digitale Geisteswissenschaften: Eine Standortbestimmung

Fricke-Steyer, Henrike

henrike.fricke@hab.de
Forschungsverbund MWW, Herzog August Bibliothek Wolfenbüttel, Deutschland

Klaffki, Lisa

klaffki@hab.de
Forschungsverbund MWW, Herzog August Bibliothek Wolfenbüttel, Deutschland

Die Zeitschrift für digitale Geisteswissenschaften 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 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 versteht sich als Organ, das all diese Entwicklungen disziplinenübergreifend begleitet und auch die philosophischen, politischen, sozialen und kulturellen Implikationen und Konsequenzen beleuchtet, die der digitale Wandel mit sich bringt. 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. Die mittlerweile fünfjährige Publikationstätigkeit der Zeitschrift für digitale Geisteswissenschaften bietet Anlass, auf die bisherige Arbeit zurückzublicken und auszuloten, inwieweit die (technisch möglichen) Spielräume genutzt werden. Das Poster möchte daher im Spiegel der bisher veröffentlichten Artikel auf die Landschaft des Digitalen Publizierens blicken und dabei folgende Aspekte thematisieren:

1. Autorschaften (kollaborativ oder einzeln)
2. Einbindung multimedialer Inhalte
3. Publikation / Verbindung mit Forschungsdaten
4. Fachliche Zuordnung(en) der Artikel
5. Gewähltes Peer Review-Verfahren (Post- oder Prepublication) (Amsen 2014)
6. Metriken (Zugriffszahlen und Downloads)
7. 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, Überarbeitung in Vorbereitung).

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, um einen Status Quo des Digitalen Publizierens zu bestimmen.

Mit Blick auf sich aktuell abzeichnende Entwicklungen, beispielsweise im Bereich Open Peer Review (Ross-Hellauer 2017), möchte das Poster auch Kommunikationsanlass sein, um künftige Spielräume mit der Fachcommunity zu durchmessen und zu öffnen.

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: 26.09.2019].

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].

Ross-Hellauer, Tony (2017): "What is open peer review? A systematic review", in: Blogpost auf F1000 Research Blog https://f1000research.com/articles/6-588/v2 . [letzter Zugriff 26.09.2019].

Digitalisierung und Erschließung arkaner Quellen im Virtuellen Archiv „Sachsen und das östliche Europa“

Kunze, Kristina
kristina.kunze@leibniz-gwzo.de
Leibniz-Institut für Geschichte und Kultur des östlichen Europa (GWZO), Deutschland

Das Projekt Virtuelles Archiv „Sachsen und das östliche Europa“ – Erschließung arkaner Quellen für die Osteuropaforschung beschäftigt sich mit der Digitalisierung und Erschließung zweier spezieller Quellengattungen, Dia-Kleinbild und Film. Dabei ist es Teil des sächsischen Verbundprojektes „Virtuelle Archive für die geisteswissenschaftliche Forschung“ welches sich dem Thema digitaler Archive von unterschiedlichen Standpunkten aus annähert. Ziel des Projektes am Leibniz-Institut für Geschichte und Kultur des östlichen Europa (GWZO) ist es, die Sammlungen für die Forschung zugänglich zu machen und die erarbeiteten Workflows in einem Best-Practice Leitfaden zu dokumentieren. Anhand der zwei sehr unterschiedlichen Quellengattungen sollen Fragen zur Nutzung geeigneter Metadatenstandards, Möglichkeiten der Präsentation der Quellen, Eignung digitale Methoden zur Beantwortung von Forschungsfragen, sowie zur Schaffung nachhaltiger Strukturen für ähnliche Projekte beantwortet werden. Die Ergebnisse und erarbeiteten Workflows sollen auf dem Poster präsentiert und zur Diskussion gestellt werden.

Im Projekt wird die Dia-Sammlung aus dem Nachlass des Prähistorikers und Archäologen Joachim Herrmann digitalisiert und erschlossen. Auf Basis einer bereits im Projekt erstellten Übersicht über den Bestand von ca. 5400 Dias werden die Digitalisate mit Metadaten versehen. Die Auswahl geeigneter Standards stellte sich allerdings als nicht trivial heraus. Die Vielzahl an Standards, die Abhängigkeit von der genutzten Software und den darin unterstützten Standards sowie begrenzte Nutzungsrechte für die Sammlung erschwerten diese. Als Anhaltspunkt für die Digitalisierung wurden die DFG-Praxisregeln „Digitalisierung“ (Deutsche Forschungsgemeinschaft 2016) herangezogen sowie weitere Checklisten (u.a. Wendel und ETH-Bibliothek). Hinsichtlich der Nutzung von Normdaten konnte sich im Verbundprojekt auf die GND Daten zu Personen und Orten geeinigt werden, mit dem Ziel die Teilprojekte über diese Normdaten mithilfe des Beacon Service[1] miteinander zu verknüpfen. Bei den Dias wurden vor allem Normdaten zu Geografika und speziellen Kulturerbestätten verwendet, da es sich vorwiegend um Fotografien von Ausgrabungsstätten handelt. Um die Wahl einer geeigneten Software zur Präsentation der Digitalisate flexibel zu bleiben, wurden die Metadaten bisher nur in .csv Dateien und als IPTC[2], EXIF[3] und XMP[4] Daten direkt im Bild abgespeichert. Die im Bild enthaltenen Metadaten können so später direkt ausgelesen, weiterverarbeitet und bei Bedarf in andere Formate übertragen werden. Die Digitalisierung der Dias ermöglicht die Zugänglichmachung der Fotografien für Forscher, die letztendlich die Relevanz dieser Sammlung erst bewerten können. Dafür ist es weiterhin notwendig die Digitalisate mit weiteren Metadaten anzureichern. Zwar gibt es auf einem Großteil der Diarahmen Beschriftungen zur jeweiligen Abbildung, diese müssten jedoch noch mit aussagekräftigen Schlagworten ergänzt werden. Eine Idee dafür ist die Einbindung der Diasammlung in die Lehre und die Erarbeitung eines Thesaurus zur Verschlagwortung in einem Seminar. Außerdem sollen über die Orte/Geografika weitere Daten durch Abfrage von Wikidata[5] zu den Dias ergänzt werden um diese beispielsweise auf einer Karte anzeigen zu lassen.

Abbildung 1: Ein Diakasten aus dem Nachlass von Joachim Herrmann

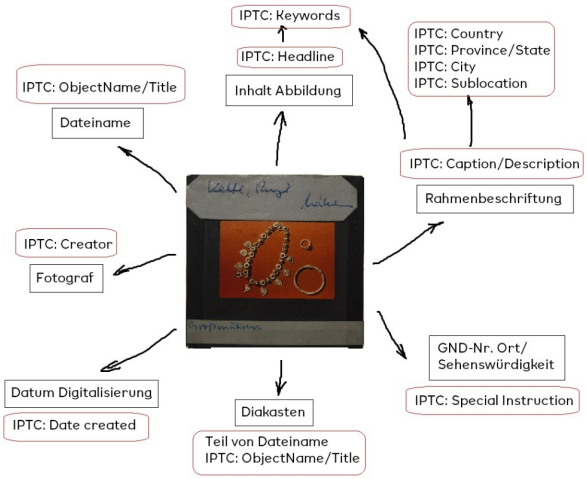


Abbildung 2: Metadaten zu den Dias und zugehöriges Mapping auf IPTC Datenfelder

Eine zweite Sammlung, etwa 500 gesammelte DVDs mit Dokumentar-, Animations-, und Spielfilmen aus dem östlichen Europa, darunter bisher unveröffentlichte Einreichungen zu Filmfestivals, hat der Filmwissenschaftler Hans-Joachim Schlegel hinterlassen. Die Filme werden nach RDA[6] erfasst und sollen über den eigenen Bibliothekskatalog zugänglich gemacht werden. RDA basiert auf dem Datenmodell FRBR[7] für bibliographische Metadaten und nutzt verschiedene Entitäten, beispielsweise Personen, Werk, Manifestation oder Exemplar, bei der Titelaufnahme. Wie schon bei der Dia-Sammlung spielen auch bei dieser Sammlung die GND Daten zu den an einem Film beteiligten Personen eine große Rolle. Diese können vor allem später zur weiteren Analyse der Filme hilfreich sein um bspw. Netzwerke osteuropäischer Filmschaffender zu untersuchen. Mit Hilfe von OpenRefine[8] konnten bereits für ca. ¼ der Filmschaffenden GND Daten gesucht und vorhandene Daten zugeordnet werden.

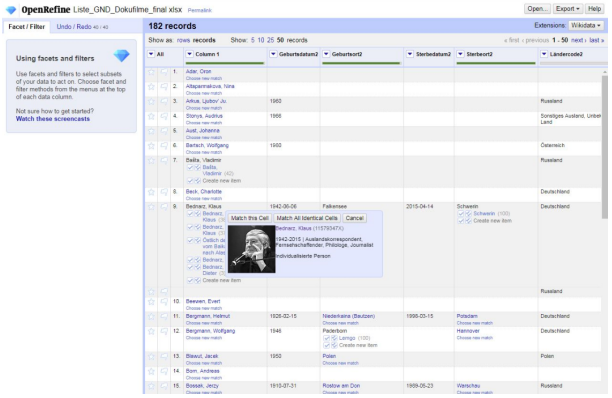


Abbildung 3: Nutzung von OpenRefine für die Zuordnung von GND Daten zu Filmschaffenden

Anhand dieses Pilotprojektes sollen am Institut Erfahrungen zur Bearbeitung unterschiedlicher wissenschaftlichen Sammlungen gesammelt werden um daraus eine Strategie für das Institut und seine bisher unbearbeiteten Nachlässe zu erarbeiten. Die Ergebnisse dieses Projektes, dabei aufgetretene Schwierigkeiten, sowie die erarbeiteten Workflows sollen auf dem Poster dargestellt werden.

Fußnoten

1. https://de.wikipedia.org/wiki/Wikipedia:BEACON
2. IPTC-IIM-Standard, Information Interchange Model(IIM) das vom International Press Telecommunications Council (IPTC) zusammen mit der Newspaper Association of America (NAA) entwickelt wurde
3. Exchangeable Image File Format – Metadatenstandard für das Speichern von hauptsächlich technischen Metadaten in digitalen Bildern
4. Extensible Metadata Platform – Standard um Metadaten in digitale Medien einzubetten
5. Durch den Einsatz der Wissensdatenbank Wikidata können weitere frei verfügbare Informationen abgefragt und für die Anreicherung der eigenen Daten genutzt werden
6. Ressource Description and Access – ein bibliothekarisches Regelwerk zur Katalogisierung von Veröffentlichungen
7. Functional requirements for bibliographic records Datenmodell welches auf dem Entity-Relationship-Modell basiert
8. https://openrefine.org/

Bibliographie

Deutsche Forschungsgemeinschaft (2016): DFG-Praxisregeln „Digitalisierung“ [12/16]: https://www.dfg.de/formulare/12_151/12_151_de.pdf [letzter Zugriff 16.12.2019]

ETH-Bibliothek DigiCenter (2016): Best Practices Digitalisierung, Version 1.1: https://www.library.ethz.ch/de/ms/DigiCenter/Best-Practices-Digitalisierung [letzter Zugriff 16.12.2019]

Wendel, Klaus (2013): „Checkliste" zur Bewertung von Angeboten zur Digitalisierung von Kulturgut, Version 1.1: https://www.digis-berlin.de/wp-content/uploads/2016/07/Checkliste_Digitalisierung_v1.1.pdf [letzter Zugriff 16.12.2019]

Discovery-Service TRIPLE

Schulte, Judith

schulte@maxweberstiftung.de
Max Weber Stiftung, Deutschland

Welche Spielräume eröffnen sich den einzelnen geisteswissenschaftlichen Disziplinen, wenn sie ihre Daten für die Forschung digital verfügbar machen möchten? Sollten sie auf möglichst weitgehenden Standards basieren, oder sollten sie vor allem die spezifischen Methoden, Zugänge und Sprachen in den jeweiligen Fächern widerspiegeln?

Der Reichtum der Sozial- und Geisteswissenschaften (SSH) besteht darin, dass sie eine Vielzahl von Disziplinen und Sprachen umfassen. Die daraus resultierende Spezialisierung ermöglicht es, eine immense Bandbreite von SSH-Themen mit unterschiedlichen Methoden und aus unterschiedlichen Perspektiven zu untersuchen. Allerdings führt dies zu einer Fragmentierung in einzelne Disziplinen und Bereiche, die einen inter- und transdisziplinären Austausch erschwert und somit einer vollen Ausschöpfung des Potenzials der SSH-Forschung im Weg steht. Wie geht man also mit der Varianz der Ansätze in den Geisteswissenschaften um, wenn man Daten sichtbar machen will? Konzeptionell gibt es eine Fülle von transdisziplinären Kooperationen, doch bedarf es hier einer Möglichkeit, die anfallenden Forschungsdaten für diese übergreifenden Ansätze aufzubereiten und nutzbar zu machen.

TRIPLE (*Targeting Researchers through Innovative Practices and multiLingual Exploration*), die europäische Discovery-Plattform, setzt an dieser Leerstelle an. Bei der Zusammenführung der Daten setzt sie auf fachspezifische und multilinguale Vielfalt einerseits und ermöglicht als Meta-Suche andererseits Forschenden, über Fach- und Sprachgrenzen hinweg sowohl Daten, aber auch andere WissenschaftlerInnen und Projekte im europäische Forschungsraum zu identifizieren und die in diesen Projektkontexten angefallenen Daten nachzunutzen und weiterzuverwenden. Der Service basiert auf der von Huma-Num entwickelten Suchmaschine Isidore, die bereits Daten von Forscherteams, Dokumentationszentren und Bibliotheken enthält. Er wird im Zuge des Projektes TRIPLE fortentwickelt und erweitert um Daten aus zahlreichen Bibliothekskatalogen, Repositorien, Archiven etc. Dabei werden unter anderem die Daten aus Forschungsprojekten der Max Weber Stiftung in das Discovery-System eingespeist wie beispielsweise Schatullrechnungen Friedrichs des Großen und Korrespondenzen zwischen Henri Fatin-Latour und Otto Scholderer (Arnoux, Gaehtgens, Tempelaere-Panzani 2014) oder der Constance de Salm, die bereits auf der Plattform perspectivia.net Open Access bereitgestellt werden. Um dies möglich zu machen, werden die Daten angereichert und standardisiert. Das Vokabular wird auf Basis von mulitlingualen Thesauri erweitert werden. Semantische Auszeichnung soll auf Basis der RAMEAU (für Französisch), LCSH (für Englisch), Spanish Biblioteca Nationale Espana und der Deutschen National Bibliothek (für Deutsch) basieren. Außerdem wird mit *named-entity-recognition-tools* (NERD) gearbeitet, um ein *discovery-tool* für Citizen Sciences (wie Wikidata) zu ermöglichen. Um eine Verbindung zur europäischen Open Science Cloud herzustellen, setzt TRIPLE auf FAIRe Daten (*Findable*, *Accessible*, *Interopable*, *Reusable*). Der Service soll SSH-Resourcen mehrsprachig zur Verfügung stellen. Derzeit sind neun Sprachen geplant. Nutzerinnen und Nutzer werden sich auf der Plattform ein Profil anlegen, nach unterschiedlichen Kriterien suchen, Suchen speichern und ausgeben lassen können.

Im weiteren Diensteportfolio sind Tools für die Visualisierung und Annotation von Daten sowie für Crowdfunding von Projekten und einem soziales Netzwerk mit Empfehlungssystem geplant, dass es NutzerInnen ermöglicht, die Daten und Literatur zu kommentieren und zu bewerten und sich zu vernetzen. Ziel ist es, WissenschaftlerInnen, BürgerInnen und Wirtschaftsorganisationen den Zugang zu wissenschaftlichen Publikationen, Daten, Datenverarbeitungsplattformen und Datenverarbeitungsdiensten im Sinne von Open Science deutlich zu erleichtern.

Das Projekt wird im Rahmen der Horizon 2020 Förderlinie von der Europäischen Kommission gefördert seit Oktober 2019 gefördert und läuft über 42 Monate. 18 europäische Partner aus zwölf Ländern (Universitäten, außeruniversitäre Forschungseinrichtungen und Infrastruktureinrichtungen, europäische Infrastrukturen, kommerzielle Verlage) sind an der Entwicklung der Plattform und Einspeisung der Daten beteiligt (u. a. Huma-Num, University of Aberdeen, Institute of Literary Research of the Polish Academy of Sciences, Max Weber Stiftung – Deutsche Geisteswissenschaftliche Institute im Ausland, Dariah-EU, CESSDA (ERIC), CLARIN (ERIC), Net7, Open Knowledge Maps). Derzeit werden über Umfragen die Bedürfnisse von WissenschaftlerInnen eruiert, um den Dienst möglichst nutzer- und bedarfsgerecht zu gestalten.

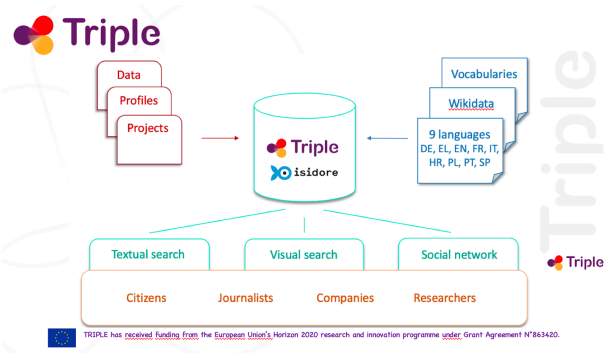


Abbildung 1: TRIPLE

TRIPLE wird als Dienst der Forschungsinfrastruktur OPERAS (*open scholarly communication in the european research area for social sciences and humanities*) entwickelt. OPERAS baut Dienste auf, die einen transnationalen Zugang zu Publikationsservices ermöglichen, die auf der Annahme gemeinsamer Normen, der Interoperabilität zwischen Verlagsdiensten und der Anbindung an den europäischer Clouddienst EOSC basieren. So soll WissenschaftlerInnen im Bereich SSH der Zugang zu wissenschaftlichen Publikationen, Daten, Datenverarbei-

tungsplattformen und Dienstleistungen für die Datenverarbeitung wesentlich erleichtert werden.

Bibliographie

Arnoux, Mathilde / Gaehtgens, Thomas W. / Tempelaere-Panzani, Anne (eds.) (2014): Correspondance entre Henri Fantin-Latour et Otto Scholderer (1858-1902), Centre allemand d'histoire de l'art, http://quellen-perspectivia.net/fantin-scholderer [letzter Zugriff 27.09.2019].

Diskursive Strukturen als öffentliche Spielräume in Graphenstrukturen. Konzeption, Modellierung und Auswertung ideengeschichtlicher Netzwerke am Beispiel der Spätaufklärung im Fürstentum Lippe

Schneider, Philipp
schneider.philipp@uni-muenster.de
Universität Münster, Deutschland

Einführung und Forschungsdesign

Das Poster diskutiert die methodischen Herausforderungen und Erkenntnismöglichkeiten bei der Modellierung und Auswertung heterogener ideengeschichtlicher Netzwerkstrukturen mit den Methoden der Sozialen Netzwerkanalyse. Dies soll im Rahmen der Vorstellung eines Projekts erfolgen, bei dem mittels einer landesgeschichtlichen Perspektive die Spätaufklärung im Fürstentum Lippe untersucht wurde. Grundlage hierfür Listen der innerhalb dieses Territoriums gelesenen und gedruckten Literatur.

Unter dem Begriff der Aufkärung wird eine Vielzahl von Diskursen zu philosophischen, politischen, religiösen und anthropologischen Ideen gefasst, die alle Lebensbereiche in den Gesellschaften des 18. Jahrhunderts betrafen. An diesen Diskursen waren unterschiedliche Akteure und soziale Gruppen beteiligt, die nicht nur verschiedene inhaltliche Schwerpunkte setzten, sondern auch variierende, zum Teil gegenläufige, Ziele verfolgten (Stollberg-Rilinger 2011: 10). Ausgehend von dieser Bandbreite und Diffusität bewegen sich nahezu alle Studien zu diesem Themenkomplex bei der Wahl ihrer Untersuchungsperspektive in der Nähe von zwei Polen:

1. Aufklärung wird aus einer *mikroskopischen* Perspektive betrachtet, wobei meist ein einzelner aufgeklärter Diskurs intensiv untersucht wird.
2. Aufklärung wird aus einer *makroskopischen* Perspektive betrachtet. Hierbei wird versucht eine möglichst allgemeine, idealtypische Begriffsbestimmung von Aufklärung zu schaffen, bei der zwar der großen Vielfalt des Themas Rechnung getragen wird, Details jedoch (zwangsläufig) ausgeblendet werden.

Die Bedeutung dieser Perspektiven für die Aufklärungsforschung ist nicht zu bestreiten. Dennoch bleiben mit der Auslassung einer mesoskopischen Sichtweise einige Fragen unbeantwortet. Zwar wurde an vielen Stellen die interne Struktur einzelner Diskurse offen gelegt und zugleich deutlich gemacht, welche Bedeutung diese für die Epoche der Aufklärung insgesamt darstellten. Unklar bleibt jedoch häufig, wie sich die einzelnen Teildiskurse zueinander verhielten; wie sie sich gegenseitig beeinflussten und wo inhaltliche sowie personelle Überschneidungen existierten. Auf Grund des umfassenden, gesamtgesellschaftlichen Durchdringungspotentials der Aufklärung waren diese diskursiven Bezüge jedoch vorhanden (Stollberg-Rilinger 2011: 10). Größere Entwicklungsverläufe in der aufgeklärten Diskurslandschaft lassen sich durch die skizzierten Perspektiven nicht vollständig, sondern nur in Ausschnitten darstellen; ebenso ist die Identifizierung von Teildiskursen oder auch aufgeklärten Denkschulen meist nur unter Bezugnahme auf Meistererzählungen sowie zeitgenössische Deutungen möglich. Eine Adressierung dieser Problemfelder würde jedoch nicht nur unser generelles Verständnis dieser Epoche schärfen, sondern könnte auch allgemeine Erkenntnisse zur Funktionsweise großer sozialer Bewegungen liefern, die von einer ähnlich heterogenen Diskurslandschaft getragen werden.

In der dem Poster zu Grunde liegenden Arbeit[1] wurde die aufgeklärte Diskurslandschaft in einem Nebenland der Aufklärung – dem Fürstentum Lippe[2] – im Zeitraum zwischen 1796 und 1820 untersucht. Auch die Aufklärungsforschung zu diesem Territorium des Alten Reiches bewegte sich vor allem in der Nähe der oben skizzierten makroskopischen (z.B. Arndt 1992) und mikroskopischen (z.B. Behrisch 2016 oder Wehrmann 1972) Dimensionen. Um die beschriebenen Anforderungen einer mesoskopischen Perspektive auf Aufklärung zu berücksichtigen wurden daher die einzelnen historischen Quellen, Personen und Ideen in ihrer singulären Bedeutung zurückgestellt und sie stattdessen in ihrer Relationalität zueinander untersucht um Rückschlüsse auf die Struktur der aufgeklärten Diskurslandschaft als Ganzes innerhalb des Untersuchungsraumes zu gewinnnen.

Modellierung ideengeschichtlicher Netzwerkdaten

Der Modellierung ebendieser Diskurslandschaft in einer Graphenstruktur lag insbesondere eine kulturwissenschaftliche Vorannahme über die betrachteten historischen Konzepte zu Grunde: Die Operationalisierung der Diskursstruktur erfolgte über ein praxeologisches Verständnis von *Aufklärung*. Hierbei wurden einerseits das Verfassen von Büchern und

Journalartikeln als Teilhabe an aufgeklärten Diskursen sowie andererseits deren Rezeption durch Lesen dieser Beiträge innerhalb von Sozietäten als zentrale Praktiken der Aufklärung identifiziert (Vgl. für eine allgemeine Operationalisierung aufgeklärter Kommunikation auch Bödeker 1987). Diese Praktiken bieten den Rahmen für die Spielräume in denen die verschiedenen, heterogenen Diskurse der Aufklärung öffentlich ausgetragen wurden.

Die wichtigsten Quellengrundlagen waren hierbei zum Einen für die innerhalb des untersuchten Fürstentums Lippe *rezipierten* Publikationen die Auktionslisten der lippischen Lesegesellschaft.[3] Zum Anderen bot für die im Untersuchungsraum *publizierten* Druckwerke das Verzeichnis des im Untersuchungszeitraum einzigen Verlags Lippes den Beleg (Weißbrodt 1914). Diese Literaturlisten wurden als strukturierte Informationen automatisiert in Netzwerkdaten überführt.

Die in den einzelnen Publikationen behandelten Themen wurden dann als Teildiskurse bzw. *Diskursfelder* aufgefasst, die in ihrer Gesamtheit den aufgeklärten Diskurs im Untersuchungsraum abbildeten. Ein Diskursfeld konnte hierbei einzelne Themenbereiche darstellen – etwa die Verbesserung der Situation der Bauern, aufgeklärte Pädagogik oder die Rolle des Adels in der Gesellschaft. Diese Themenfelder stehen somit auch für sich genommen als eigene, diskursanalytisch relevante ideengeschichtliche Entitäten. Ihre Modellierung als Netzwerkknoten erfolgte auf Grundlage der bisherigen Forschung zur Aufklärung. Diese nicht-automatisierte Erstellung der Diskursfelder bedingt dementsprechend besonders stark die Erkenntnismöglichkeiten des Netzwerks. Zugleich wird so bereits bei der Modellierung sichergestellt, dass die Netzwerkdaten im Rahmen der bisherigen Aufklärungsforschung kontextualisiert sind.

Diskursfelder und Publikationen wurden dann als Knoten in einem bimodalen Netzwerk modelliert (Abbildung 1). Die Verknüpfungen zwischen den Publikations- und Diskursfeld-Knoten erfolgte in einem hermeneutischen-interpretativen Prozess auf Grundlage der Titel der Bücher und Journalartikel.[4] Das Netzwerk wurde dann zu einem unimodalen Netzwerk aus Diskursfeldern transformiert um eine Auswertung vornehmen zu können (Abbildung 2). Durch die Verwendung der Publikationen als heuristische Grundlage konnten so nach der Transformation ideengeschichtliche Ähnlichkeiten und Abhängigkeiten zwischen den einzelnen Diskursfeldern abgebildet werden. Um dabei auch Veränderungen innerhalb des Untersuchungszeitraums erfassen zu können wurden die Kanten des Netzwerks mit zeitlichen Informationen versehen. Grundlage waren die Jahre, in denen die einzelnen Bücher und Journalartikel veröffentlicht wurden – in diesem Zeitraum lässt sich durch den Druck einer Publikation ihr Einfluss auf ein Diskursfeld nachweisen.

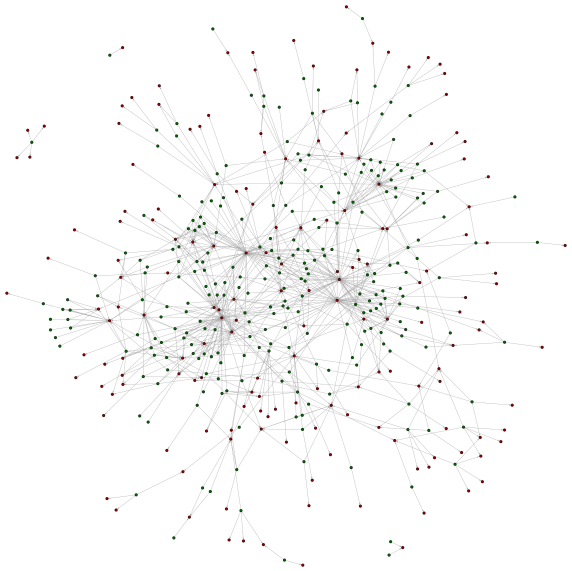


Abbildung 1: Bimodales, gerichtetes Netzwerk der innerhalb Lippes erschienenen Publikationen. Diskursfelder sind rot, Publikationen grün dargestellt. Visualisierung mit dem Fruchtermann-Reingold-Algorithmus.

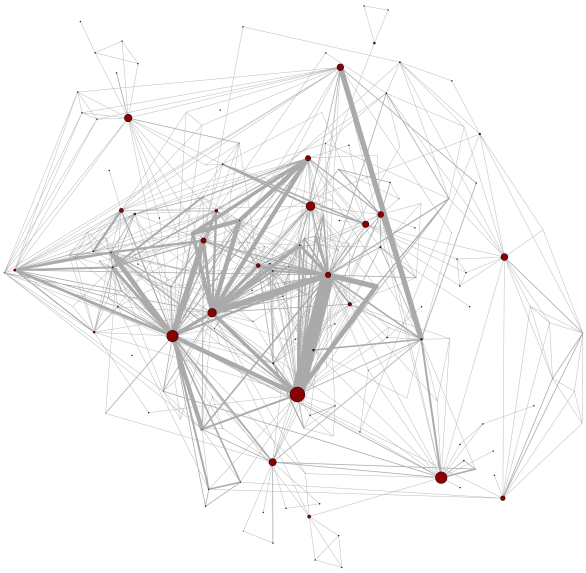


Abbildung 2: Aus Abbildung 1 transformiertes Diskursnetzwerk der innerhalb Lippes erschienenen Publikationen. Die Dicke der Kanten beschreibt dabei die Anzahl der Publikationen, die sich ein Diskursfeld teilen. Visualisierung mit dem Davidson-Harel-Algorithmus.

Netzwerkanalytische Auswertung

Sowohl die quantitative Auswertung mit Methoden der *Sozialen Netzwerkanalyse* als auch die Aufbereitung der Daten erfolgte mit der Skriptsprache *R*; v.a. unter Zuhilfenahme des Pakets *igraph*. Die Verwendung von eigenem Programmcode bietet gegenüber GUI-basierten Lösungen wie *Gephi* oder *Pa-*

jek eine leichtere Nachnutzbarkeit und damit auch eine bessere Überprüfbarkeit der Ergebnisse.

Sowohl die Vielfalt der einzelnen Wissensfelder, als auch die nahezu unüberschaubare Menge historischer Quellen dieser Zeit sollten berücksichtigt werden. Die quantitative Auswertung der Relationen zwischen den Diskursfeldern mittels netzwerkanalytischer Verfahren, wie Berechnung von Betweenness-Zentralität (Jansen 2003: 134f) oder Community-Analyse (Rosvall 2019) legten unter anderem eine moderate Diversität der aufgeklärten Diskurslandschaft Lippes offen. Über den gesamten Untersuchungszeitraum ließ sich sowohl in der Publikations- als auch in der Rezeptionstätigkeit ein breites thematisches Interesse nachweisen, wie es für die Zeit der Aufklärung als charakteristisch beschrieben wird (Stollberg-Rilinger 2000). In Bezug auf die innerhalb des Fürstentums Lippe erschienen Publikationen existierte diese Vielfalt jedoch nicht durchgängig während des betrachteten Zeitraumes zwischen 1796 und 1820. Vielmehr war der innerhalb Lippes geführte Diskurs von wenigen Spezialthemen geprägt, die einen dauerhaften Bezug zu nahezu allen anderen Diskursfeldern aufwiesen. Hier dominierten insbesondere religiöse und pädagogische Themen im öffentlichen Diskurs. Volksaufklärerische Schriften mit einer großen territorialen Selbstreferenzialität nahmen ebenfalls einen wichtigen Platz im lippischen Publikationswesen ein. Politische Aktivität und Publikationswesen waren im Fürstentum weitgehend voneinander entkoppelt.

Methodische Perspektiven

Aus einer landesgeschichtlichen Perspektive ermöglichte die Untersuchung eine Schärfung des Verständnisses zu ideengeschichtlichen Schwerpunkten und Strukturen der Spätaufklärung im Fürstentum Lippe. Damit nimmt sie die oben skizzierte Mesoperspektive auf die Erforschung von Aufklärung als geistesgeschichtliches Phänomen ein. Hinsichtlich der *Historischen Netzwerkforschung* – der Anwendung der *Sozialen Netzwerkanalyse* in den Geschichtswissenschaften – erprobte die Untersuchung einen Ansatz zur Operationalisierung und Auswertung von Netzwerken, die nicht aus Personen, sondern aus abstrakten Ideen bestehen. Die Untersuchung nichtpersonaler Netzwerke findet in den Historischen Disziplinen bislang nur selten statt (Düring 2016: 38). Daher ist es erforderlich, Möglichkeiten ebenso wie Herausforderungen bei der Quellensammlung, Datenmodellierung, Abgrenzung (Laumann 1992) und Auswertung eines Netzwerks, das aus ideengeschichtlichen Entitäten besteht, auszudifferenzieren und zu diskutieren.

Auf diese Weise sind auch unausweichlich zentrale Themen der Tagung betroffen, welche nicht nur für das Teilgebiet der *Sozialen Netzwerkanalyse*, sondern für die *Digital Humanities* im Allgemeinen relevant sind. Die Vorauswahl bestimmter Datenmodelle und Algorithmen wird bei der *Sozialen Netzwerkanalyse* besonders evident: Hier beruht das gesamte Forschungsdesign auf einem Denkparadigma, dem ein streng relationales Welt-, beziehungsweise Geschichtsbild, zu Grunde liegt. Erst die Wahl dieses Paradigmas bedingt den Einsatz bestimmter Methoden und computergestützter Werkzeuge. So ist auch das Thema dieses Posters ein Beispiel dafür, wie diese Vorannahmen einerseits die Selektion von Fragestellungen und bestimmten Quellentypen beeinflussen. Andererseits bietet die *Soziale Netzwerkanalyse* in Verbindung mit einem praxeologischen Ansatz jedoch auch eine ‚Unvoreingenommenheit' gegenüber dem historischen Material. Serielle Textkorpora – wie in diesem Fall Publikationsverzeichnisse – können auf Grundlage formalisierender Vorüberlegungen vollständig ausgewertet werden. Dadurch kann auch einer, der manuellen Reduktion eines Korpus inhärenten, Gefahr von Meistererzählungen vorgebeugt werden.

Fußnoten

1. Eine Veröffentlichung der Forschungsergebnisse, der dabei angefallenen Daten sowie des für deren Aufbereitung und Auswertung entwickelten Programmcodes ist derzeit in Vorbereitung.
2. Das Land Lippe verfügte zwar über eine aufgeklärten Ideen gegenüber sehr aufgeschlossene landesherrliche Regierung, Verlagswesen und kleine Sozietätslandschaft, besaß jedoch kein eigenes geistiges Zentrum in Form einer Universität.
3. Diese sind im *Landesarchiv NRW, Abteilung Ostwestfalen-Lippe* vollständig erhalten.
4. Wegen ihres Umfangs bieten Literaturtitel aus dem 18. Jahrhunderts sehr ausführliche Informationen zum Inhalt des jeweiligen Werkes.

Bibliographie

Arndt, Johannes (1992): *Das Fürstentum Lippe im Zeitalter der Französischen Revolution 1770 – 1820.* Münster: Waxmann.

Behrisch, Lars (2016): *Die Berechnung der Glückseligkeit.* Statistik und Politik in Deutschland und Frankreich im späten Ancien Régime (= Beihefte der Francia 78), Ostfildern: Thorbecke.

Bödeker, Hans Erich (1987): „Aufklärung als Kommunikationsprozeß", in: *Aufklärung* 2: 89–111.

Düring, Marten / Kerschbaumer, Florian (2016): „Quantifizierung und Visualisierung. Anknüpfungspunkte in den Geschichtswissenschaften" in: Düring, Marten / Eumann, Ulrich / Stark, Martin / von Keyerlingk, Linda (eds.): *Handbuch Historische Netzwerkforschung.* Grundlagen und Anwendungen (= Schriften des Kulturwissenschaftlichen Instituts Essen (KWI) zur Methodenforschung 1), Münster: Lit Verlag 31–43.

Füssel, Marian (2015): „Praxeologische Perspektiven in der Frühneuzeitforschung" in: Brendecke, Arndt (ed.): *Praktiken der Frühen Neuzeit*. Akteure, Handlungen, Artefakte (= Frühneuzeit-Impulse 3), Köln / Weimar / Wien: Böhlau 21–33.

Jansen, Dorothea (2003): *Einführung in die Netzwerkanalyse*. Grundlagen, Methoden und Forschungsbeispiele. Opladen: Leske + Budrich.

Landwehr, Achim (2018): *Historische Diskursanalyse.* Frankfurt a. Main: Campus Verlag.

Laumann, Edward O. / Marsden, Peter V. / Prensky, David (1992): „The Boundary Specification Problem in Network Analysis" in: Freeman, Linton C. / White, Douglas R. / Romney, Antone Kimball (eds.): *Research Methods in Social Network Analysis.* New Brunswick / New Jersey: Taylor & Francis 61–88.

Lemercier, Claire (2015a): „Formal network methods in history: why and how?" in: Fertig, Georg (ed.): *Social Networks, Political Institutions, and Rural Societies* (= Rural History in Europe 11), Turnhout : Brepols 281–310.

Dies. (2015b): „Taking time seriously. How do we deal with change in historical networks?“ in: Gamper, Markus / Reschke, Linda / Düring, Marten (eds.): *Knoten und Kanten III. Soziale Netzwerkanalyse in Geschichts- und Politikforschung* (= Sozialtheorie), Bielefeld: transcript 183–212.

Rosvall, Martin / Bergstrom, Carl T. (2008): „Maps of random walks on complex networks reveal community structure“, in: *Proceedings of the National Academy of Sciences 4* 1118–1123.

Stollberg-Rilinger, Barbara (2000): *Europa im Jahrhundert der Aufklärung*, Stuttgart: Reclam.

Dies. (2011): *Die Aufklärung.* Europa im 18. Jahrhundert, Stuttgart: Reclam.

Trappmann, Mark / Hummell, Hans J. / Sodeur, Wolfgang (2011): *Strukturanalyse sozialer Netzwerke*. Konzepte, Modelle, Methoden (Studienskripte zur Soziologie). Wiesbaden: Springer VS.

Wehrmann, Volker (1972): *Die Aufklärung in Lippe*. Ihre Bedeutung für Politik, Schule und Geistesleben (= Lippische Studien 2). Detmold: Landesverband Lippe.

Weißbrodt, Ernst (1914): *Die Meyersche Buchhandlung in Lemgo und Detmold und ihre Vorläufer*. Festschrift zum 250jährigen Bestehen der Firma am 12. Juni 1914. Detmold: Meyer.

Early Stage Digital Medievalist Subcommittee. Vernetzen, entgrenzen, Spielräume schaffen

Busch, Hannah

hannah.busch@huygens.knaw.nl
KNAW Huygens ING, Amsterdam; Universität Utrecht

Gengnagel, Tessa

tessa.gengnagel@uni-koeln.de
Universität zu Köln; Cologne Center for eHumanities

Schulz, Daniela

dschulz@uni-wuppertal.de
Bergische Universität Wuppertal; Herzog August Bibliothek Wolfenbüttel

Obwohl mediävistische Forschung sich bereits seit geraumer Zeit digitaler Methoden bedient und in der Gründungsgeschichte des Humanities Computing, heute als Digital Humanities bekannt, eine herausragende Stellung einnimmt (s. Bleier et al. 2019: 1-12) , spiegelt sich diese real existierende Interdisziplinarität bisher kaum bis gar nicht in den im deutschsprachigen Raum etablierten mediävistischen Studiengängen wider. Dies mag umso mehr verwundern, als die digitale Mediävistik über vergleichsweise klar umgrenzte Fragestellungen, Anwendungsfelder und Vorgehensweisen verfügt; im Gegensatz zu den allgemeiner gehaltenen Digital Humanities, zu denen es einige Lehrstühle und Studiengänge gibt,[1] wenngleich lokal wiederum meist mit fachlicher Einschränkung (s. Trier, Leipzig, Stuttgart, Würzburg, Köln, siehe außerdem Sahle 2016) . Themenfelder, denen sich die digitale Mediävistik dezidiert widmet, sind – um nur einige Beispiele zu nennen – computergestützte Analysen von paläographischen Befunden (DigiPal, KPDZ/ CPDA 1-4), historisch-geographische Informationssysteme im mediävistischen Kontext (Mapping Medieval Conflict) und der Einsatz des Maschinellen Sehens zur Mustererkennung in mediävistischen Bildwerken (Computer Vision Lab Heidelberg).

Zu den Herausforderungen einer solch interdisziplinären und im besten Fall auch innovativen Forschung kommen für den wissenschaftlichen Nachwuchs, wie bereits angedeutet, weitere Herausforderungen hinzu: Neben der Frage der Ausbildung stehen hier vor allen Dingen Fragen nach Karriereperspektiven, Anerkennung alternativer Publikationsformen (die Publikation von Forschungsdaten (vgl. Andorfer 2015) , die Debatte um die Vorzüge einer kumulativen Dissertation versus Monographie) und die Einbettung in bestehende fachliche Infrastrukturen im Vordergrund. Das neu gegründete Subcommittee der *Digital Medievalist* Community, das sich an Early Stage Researcher richtet, hat sich zum Ziel gesetzt, dieser Gruppe eine Plattform zu bieten und den Gesprächsbedarf in einen Dialog mit der größeren Fachgemeinde zu übersetzen.

Digital Medievalist ist eine internationale interessenbasierte virtuelle Forschungsgemeinschaft, die mit einem breiten thematischen Zuschnitt über Disziplingrenzen hinweg Wissenschaftler*innen unterschiedlicher Statusgruppen miteinander vernetzt und verschiedene Wege geht, um gemeinschaftlich Spielräume der einzelnen Fachdisziplinen zu erweitern. Die Gemeinschaft wurde bereits 2003 gegründet, seit 2005 ist sie Herausgeberin des gleichnamigen Open Access Journals. Unabhängig von Standorten bietet die *Digital Medievalist* Community beispielsweise im Rahmen von gemeinsam organisierten Konferenzaktivitäten ein Netzwerk sowohl für etablierte Wissenschaftler*innen als auch für solche am Beginn ihrer wissenschaftlichen Karriere. Digital Medievalist steht allen Interessierten offen, unabhängig bestehender Erfahrungen in den Digital Humanities oder den mediävistischen Disziplinen, von absoluten Neulingen bis hin zu sogenannten Pionieren im Bereich der (digitalen) Mediävistik.

Spielräume der DM Community sind bisher bereits, *in a nutshell:*

- Die *Digital Medievalist* Mailingliste : Mehr als 1.300 Abonnenten (Stand September 2019) nutzen diesen Kanal als Diskussionsplattform, um Rat einzuholen und um Informationen jeglicher Art im Bereich der (digitalen) Mediävistik zu teilen.
- Das *Digital Medievalist* Journal: Verlagsunabhängige (APC freie) Open-Access Fachzeitschrift der Community; die wissenschaftliche Qualität der Artikel wird im Peer-Review-Verfahren gesichert.
- Die *Digital Medievalist* Webseite: Die Onlinepräsenz der Community versammelt alle Informationen über die Community: Wie wird man Mitglied? Wie ist die Organisation aufgebaut, wie lautet die Satzung? Darüber hinaus finden sich hier Ankündigungen sowie wie eine stetig aktualisierte Liste von vergangenen und anstehenden Konferenzen, Kolloquien, Workshops und Sommerschulen mit Relevanz für die (digitale) Mediävistik. Im Webblog werden zukünftig neben CfPs und Veranstaltungshinweisen Projekte und Tools aus der digitalen Mediävistik vorgestellt,

auf aktuelle Veröffentlichungen verwiesen sowie die Reihe " What do Digital Medievalists do? " (Campagnolo 2017) weitergeführt.
- *Digital Medievalist* Zotero Bibliographie : Sammlung einschlägiger Literatur zu allen themenbereichen der digitalen Mediävistik.
- Die *Digital Medievalist* Facebookgruppe mit mehr als 2.500 Mitgliedern sowie ein Twitteraccount @digitalmedieval mit derzeit über 6.000 Followern erweitern den Spielraum der Wissenschaftskommunikation.

Während der Postersession möchten wir die verschiedenen Initiativen der *Digital Medievalist* Community vorstellen und die Vernetzung innerhalb der deutschsprachigen DH vorantreiben. Hierbei möchten wir vor allen Dingen die geplanten Aktivitäten des neugegründeten Postgraduate Subcommittees skizzieren und einen Peer-to-Peer-Austausch fördern. Das Subcommittee hat es sich zum Ziel gesetzt, einerseits bereits bestehende Aktivitäten wie den Blog auf der Webseite oder die Präsenz der Community auf Twitter zu beleben und andererseits ab 2020 neue eigene Aktivitäten in Angriff zu nehmen; hierzu zählen insbesondere die Organisation von gemeinsamen Panels (so etwa auf dem International Medieval Congress in Leeds) und die Produktion von einem Podcast, in dem Nachwuchsforscher zu Wort kommen sollen. Die Erhöhung der Sichtbarkeit der existenten Infrastrukturen soll außerdem Denkanstöße für eine Diskussion um interdisziplinäres Arbeiten, erforderliche Skills und eine Reform der universitären Curricula im mediävistischen Kontext liefern und damit auch in übergreifender Perspektive beispielhaft zu aktuellen Debatten um die Profilierung geisteswissenschaftlicher Disziplinen zwischen Tradition und gegenwärtigen Anforderungen beitragen.

Fußnoten

1. Für ein ausführliches Verzeichnis der DH Studiengänge siehe https://registries.clarin-dariah.eu/courses/

Bibliographie

Andorfer, Peter (2015): „Forschungsdaten in den (digitalen) Geisteswissenschaften: Versuch einer Konkretisierung", Göttingen: GOEDOC (DARIAH-DE working papers 14) http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-7-2

Bleier, Roman / Fischer, Franz / Hiltmann, Torsten / Viehhauser, Gabriel / Vogeler, Georg (2019): „Digitale Mediävistik und der deutschsprachige Raum" in: Das Mittelalter 24. 1 : 1-12 DOI:

Campagnolo, Alberto (2017): „What do digital medievalists do?" https://digitalmedievalist.wordpress.com/2017/08/10/what-do-digital-medievalists-do/

Computer Vision Lab Uni Heidelberg: https://hciweb.iwr.uni-heidelberg.de/compvis/

DigiPal (2011-2014): „Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic" http://www.digipal.eu/

Digital Medievalist Webseite: https://digitalmedievalist.wordpress.com/

Digital Medievalist Journal: https://journal.digitalmedievalist.org/

Digital Medievalist Mailingliste: https://digitalmedievalist.wordpress.com/mailing-list/

Digital Medievalist on Facebook: https://www.facebook.com/groups/49320313760/

Digital Medievalist on Twitter: https://twitter.com/digitalmedieval

Digital Medievalist on Zotero: https://www.zotero.org/groups/2138266/digitalmedievalist

Kodikologie und Paläographie im digitalen Zeitalter 1-4 / Codicology and Palaeography in the Digital Age 1-4 (2009, 2010, 2015, 2017): Herausgegeben vom Institut für Dokumentologie und Editorik, Norderstedt: BoD. https://www.i-d-e.de/publikationen/schriften/

Mapping Medieval Conflict: https://www.i-d-e.de/publikationen/schriften/

Sahle, Patrick (2016) : "Zur Professoralisierung der Digital Humanities" in: DHdBlog 23. März 2016 https://dhd-blog.org/?p=6174

Alle angegeben Links wurden am 4. Januar 2020 geprüft.

Eine Programmierschnittstelle für Metadaten zu DH Lehraktivitäten: Die DH Course Registry API

Schmeer, Hendrik
mail@hendrikschmeer.de
Österreichische Akademie der Wissenschaften, Österreich; CLARIN ERIC

Wissik, Tanja
tanja.wissik@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Die DH Course Registry ist eine von den beiden Forschungsinfrastrukturen CLARIN und DARIAH getragene Initiative um Lehraktivitäten im Bereich der digitalen Geisteswissenschaften auch über Universitätsgrenzen hinweg sichtbar zu machen. Auf der Plattform werden Metadaten über BA, MA und PhD Studiengänge, einzelne Lehrveranstaltungen und Module sowie Summer School im Bereich der Digitalen Geisteswissenschaften gesammelt und veröffentlicht.
Die Daten werden nicht zentral gesammelt, sondern community-basiert von den Lehrenden selber eingegeben. Die Daten sind über bestimmte Filteroptionen wie z.B. Land, Art des Studiengangs usw. durchsuchbar und die Ergebnisse werden auch auf einer Karte visualisiert. Um die Daten aktuell zu halten, gibt es im DH Course Registry einen eigenen Workflow und die spezifische Rolle der „Nationalen Moderatoren". Wenn die Metadaten eines Kurses längere Zeit nicht mehr aktualisiert wurden, wird dieser Kurs nicht mehr online angezeigt. In der Datenbank sind diese Daten aber noch immer gespeichert. Demzufolge birgt die Plattform einen Datenschatz z.B. zur zeitlichen Entwicklung DH bezogen Lehraktivitäten, der bis jetzt der Forschungsgemeinschaft nur auf Nachfrage als

Datenbankauszug zugänglich war. Eine Möglichkeit diese Daten für unterschiedliche Nutzergruppen und Nutzungsszenarien zugänglich zu machen ist die Bereitstellung von standardisierten Programmierschnittstellen – kurz APIs genannt. In den letzten Jahren ist die Nutzung von APIs in DH Projekten angestiegen (vgl. Tasovac et al. 2016). Aber auch die Daten, die über die Plattform bereits abfragbar waren, können durch die API sozusagen „mit nach Hause genommen" und für weitere Forschungsfragen verarbeitet werden (vgl. Cooper, 2010).

Im DARIAH geförderten Projekt „DH Course Registry Sustain" wurde nun durch die Entwicklung und Bereitstellung einer API dieser Datenschatz für die Forscherinnen und Forscher geöffnet. Bevor die API beschrieben wird, soll auf das Datenmodell eingegangen werden, das der Plattform zugrunde liegt.

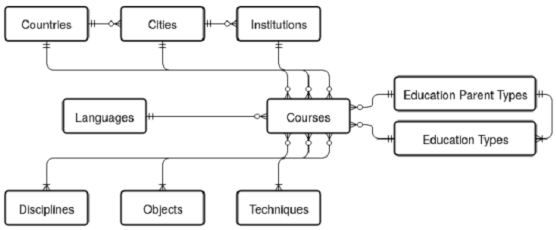


Abbildung 1: Entity-Relationship Diagramm (ERD) (Wissik et al. im Erscheinen)

Die Beziehungen der verschiedenen Objekt-Typen werden in Abbildung 1 wiedergegeben, wie sie durch die Datenbankstruktur der DH Course Registry definiert sind. Alle um "Courses" herum gruppierten Objekt-Typen stellen gleichermaßen kontrollierte Vokabulare dar, nach denen die Datenbank durchsucht werden kann. Die mit den Typen korrespondierenden Tabellen sind ebenfalls durch die API zugänglich, um etwa Filteroptionslisten zu generieren. Auf diese Weise entsteht die Durchsuchbarkeit der Datenbasis für einzelne Universitäten, Länder, Methodiken oder Fachbereiche. Das Datenmodell umfasst darüber hinaus auch noch Informationen zur Zyklizität, Start-Terminen und Präsentationsform (online oder Präsenzkurs) oder der Sprache, um Möglichkeiten zum internationalen Austausch aufzuzeigen. Durch die Erstellungs- und Modifikationsdaten der Einträge, die im Web-UI nicht zugänglich sind, kann die zeitliche Entwicklung des in der DH Course Registry vertretenen Lehrangebots nachvollzogen werden. Eine vollständige OAS 3.0 kompatible Beschreibung der über die API zugänglich gemachten Datenstrukturen im JSON Format ist unter https://app.swaggerhub.com/apis-docs/hashmich/DHCR-API/1.0 verfügbar.

Im Rahmen der Posterpräsentation werden wir die DH Course Registry API vorstellen und einige exemplarische Analysen präsentieren, die durch den API Zugang nun möglich gemacht werden.

Bibliographie

Wissik, Tanja / Edmond, Jennifer /Fischer, Frank / de Jong, Franciska / Stefania Scagliola, Stefania / Andrea Scharnhorst, Andrea / Schmeer, Hendrik / Walter Scholger, Walter / Wessels, Leon (im Erscheinen): Teaching Digital Humanities Around the World: An Infrastructural Approach to a Community-Driven DH Course Registry.

Toma Tasovac, Toma / Adrien Barbaresi, Adrien / Thibault Clèrice, Thibault / Jennifer Edmond, Jennifer / Ermolaev, Natlia et al. (2016): APIs in Digital Humanities: The Infrastructural Turn. Digital Humanities 2016, Jul 2016, Cracovie, Poland. pp.93-96, 2016, Digital Humanities 2016 Conference Abstracts.

Cooper, Doug (2010): When Nice People Won't Share: Shy Data, Web APIs, and Beyond, Second International Conference on Global Interoperability for Language Resources (ICGL 2010).

Ein Spielraum der Digital Humanities: Die Europäische Sommeruniversität „Kulturen & Technologien"

Annisius, Marie
marie.annisius@uni-leipzig.de
Universität Leipzig, Deutschland

Burr, Elisabeth
elisabeth.burr@uni-leipzig.de
Universität Leipzig, Deutschland

Fußbahn, Ulrike
ulrike.fussbahn@uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

Im Dezember 2007 fand an der Universität Leipzig ein von der *Association for Literary and Linguistic Computing* (ALLC) geförderter Workshop "Text Markup & Database Design" statt. Im Wintersemester 2007/2008 boten das King's College London (Großbritannien), die Debreceni Egyetem (Ungarn), die Universität Leipzig (Deutschland) und die Oulun Yliopisto (Finnland) per Videokonferenz gemeinsam ein Seminar *Culture & Technology* für ihre Studierenden an. Im Wintersemester wurde das Seminar zusammen mit acht europäischen Universitäten, nämlich Universidad de Alicante (Spanien), Università Bologna (Italien), Debreceni Egyetem (Ungarn), University College Dublin (Irland), University of Glasgow (Schottland), Universität Leipzig (Deutschland), King's College London (Großbritannien) und Oulun Yliopisto (Finnland) fortgesetzt. Die Erfahrungen waren zwar insgesamt sehr positiv, wirklich befriedigen konnten sie aber schon wegen der fehlenden internationalen Community im Falle des Workshops

und des fehlenden physischen Kontakts im Falle des Videokonferenzseminars nicht.

Deshalb wurde 2009 ein Experiment begonnen, das bis heute andauert und als ein Spielraum der *Digital Humanities* mit Nachjustierungen, Modifizierungen und Erweiterungen gesehen werden kann, nämlich die Europäische Sommeruniversität in Digitalen Geisteswissenschaften „Kulturen & Technologien", die allgemein als ESU bekannt ist, obwohl ihr Akronym eigentlich ESU DH sein müsste.

Die Europäische Sommeruniversität in Digitalen Geisteswissenschaften „Kulturen & Technologien"

Dass die ESU direkt an das Europäische Seminarprogramm anknüpfte, zeigt sich schon in dem Motto „Kulturen & Technologien" bzw. "Culture & Technology". Auch nahm sie von Anfang an eine europäische Perspektive ein was Sprachen und Wissenskulturen betrifft. Englisch musste zwar aus pragmatischen Gründen generell als *lingua franca* fungieren, doch den Tendenzen hin zu einer mehr und mehr monolingualen Wissenschaftskultur sollte die Wertschätzung der europäischen Mehrsprachigkeit und der Vielfalt europäischer Wissenskulturen entgegengesetzt werden. So sollte etwa das einzige Kriterium für den Gebrauch einer bestimmten Sprache die Teilhabe aller sein. Ansprechen wollte sie ursprünglich vor allem ein europäisches Publikum. Schon während der Bewerbungsphase um einen Platz bei der zweiten Sommeruniversität 2010 wurde allerdings deutlich, dass die ESU als eine Institution in Europa gesehen wurde, und eben nicht als eine Institution, die sich der Verbreitung und Entwicklung der *Digital Humanities* in Europa verschrieben hatte.

Seit 2009 haben insgesamt 851 Personen (Vortragende und Workshopleitende eingeschlossen) aller Alters- und Karrierestufen und unterschiedlichster fachlicher Provenienz aus 60 Ländern der ganzen Welt an der ESU teilgenommen. Viele darunter wurden gerade auch durch die mehrsprachige und mehrkulturelle Ausrichtung der ESU zur Teilnahme motiviert. Bei der ESU haben sie nicht nur Werkzeuge und Methoden, die in den *Digital Humanities* eine Rolle spielen, oder deren Anwendung im Rahmen von Projekten kennengelernt, sondern wurden mittels Vorlesungen in die *Digital Humanities* als Epistemologie eingeführt und haben bei Podiumsdiskussionen auch deren sozio- und bildungspolitische Relevanz diskutiert.

Im Unterschied zu anderen Sommerschulen in den *Digital Humanities* setzt die ESU gerade nicht auf quantitatives Wachstum,[1] stattdessen will sie ihren möglichst nur 60 Teilnehmerinnen und Teilnehmern und den etwa 25 Lehrenden ermöglichen, fachliche und freundschaftliche Netzwerke zu schaffen, die die Sommeruniversität über Jahre hinweg überdauern. Diese Netzwerkbildung fördert sie dezidiert durch ein reichhaltiges Rahmenprogramm, bei dem sie darauf setzt, ihrem Titel „Kulturen & Technologien" durch die Wahl der in Leipzig und Umgebung zu besuchenden Orte gerecht zu werden, sowie durch den entspannten Austausch über Kulturen und Sprachen hinweg bei gemeinsamen Mittag- und Abendessen.

Jede Ausgabe der ESU war, für sich selbst genommen, ein Spielfeld mit bestimmten Regeln. Über die Jahre hinweg blieb die ESU aber immer auch ein Spielraum, dem das Nachjustieren, Modifizieren und Erweitern inhärent sind.[2]

Poster

In unserem Poster wollen wir zeigen, was die ESU grundsätzlich charakterisiert, wie sie sich über die Jahre verändert hat, wie es ihr gelungen ist, die einzige *Digital Humanities* Sommerschule in Europa zu werden, die die traditionellen Grenzen zwischen Sprachwissenschaft und Sprachressourcen einerseits und den *Digital Humanities*, so wie sie sich mehrheitlich präsentieren, andererseits zu überwinden, welche Folgen und Bereicherungen dies für beide Seiten bedeutet und was die Zukunft bringen soll.

Fußnoten

1. Vgl. etwa das *Digital Humanities Summer Institute* (DHSI) https://dhsi.org/ oder die *Digital Humanities at Oxford Summer School* (DHOxSS) https://www.dhoxss.net/.
2. Siehe hierzu die Webpräsenz der *Europäischen Sommeruniversität in Digitalen Geisteswissenschaften „Kulturen & Technologien"* http://esu.culintec.de/.

Erweiterung eines Forschungsdatenrepositoriums um ein Modul für die Nachnutzbarkeit und Analyse von Textressourcen

Schneider, Gerlinde
gerlinde.schneider@uni-graz.at
Universität Graz, Österreich

Vasold, Gunter
gunter.vasold@uni-graz.at
Universität Graz, Österreich

Digitale Editionen stellen als digitalisierte und tiefenerschlossene Textressourcen eine wertvolle Quelle zur Nachnutzung innerhalb großflächiger linguistischer und literaturwissenschaftlicher Analysen dar (Rybicki, 2019). Zusätzlich werden innerhalb von digitalen Editionsprojekten selbst immer öfter textanalytische Verfahren eingesetzt.

Das am Zentrum für Informationsmodellierung der Universität Graz entwickelte und betriebene Repositorium GAMS

(Geisteswissenschaftliches Asset Management System)[1] umfasst als Forschungsdateninfrastruktur Daten von mehr als hundert Forschungsprojekten aus verschiedenen Wissenschaftsbereichen. Digitale Editionen und Textsammlungen machen dabei, neben neben digitalen Sammlungen aus dem Kulturerbebereich, den Großteil der im Repositorium vorhandenen Bestände aus.

Um die bereits im Repositorium vorhandenen Textressourcen in geeigneten Formaten nachnutzbar bereitzustellen, beziehungsweise diesen Aspekt im Zuge laufender und zukünftiger Projekte berücksichtigen zu können, wurden während der letzten Monate Adaptierungen an der GAMS-Infrastruktur vorgenommen, die mit diesem Poster erläutert und dargestellt werden sollen.

Technischer Hintergrund

GAMS ist eine registrierte[2], trusted[3] Repositoriumsinfrastruktur, die auf der Free and Open Source Software Fedora Commons[4] basiert. Sie setzt auf eine OAIS-konforme Architektur und verfolgt eine weitgehend XML-basierte Content-Strategie. GAMS ermöglicht seinen Benutzer*innen die Verwaltung und Veröffentlichung von Ressourcen aus Projekten mit permanenter Identifizierung und Anreicherung mit Metadaten. Ein speziell entwickelter Client (*Cirilo*) stellt Funktionalitäten für Massenoperationen an den gespeicherten Objekten zur Verfügung. (Stigler/Steiner 2018)

Objekt Modell

Content Models definieren komplexe digitale Objekte, die dem Fedora-Objektmodell entsprechen. Sie sind speziell auf die Anforderungen, die Forschungsdaten aus unterschiedlichen geisteswissenschaftlichen Bereichen an Langzeitarchivierung und Datendissemination stellen, ausgelegt. Für wissenschaftliche Editionen wird beispielsweise ein speziell entworfenes *TEI Content Model* eingesetzt.

Jedes Modell enthält einen primären Datenstrom, der die Inhaltsdaten des Objekts enthält, zum Beispiel ein TEI-Dokument. Zusätzliche Datenströme können Metadaten (z.B. Dublin Core), weitere Inhaltsdaten oder aus dem primären Datenstrom derivierte Daten enthalten (z. B. aus dem TEI-Dokument extrahierte RDF Daten).

Für die jeweiligen Modelle definierte Services kombinieren und transformieren Datenströme zu Präsentationsinhalten, auf die in verschiedenen Ausgabeformaten über im Content Model definierte Schnittstellen zugegriffen werden kann. Ein häufig verwendetes Format zur Dissemination ist HTML, was die Präsentation der Daten über eine dynamisch erzeugte Webseite ermöglicht.

*Context*e, als spezielle Containerobjekte, ermöglichen es, einzelne Inhaltsobjekte in größere Einheiten zusammenzufassen und zu organisieren. Sie enthalten wiederum eigene Datenströme und Disseminationsmethoden.

Anpassungen für Textressourcen

Zur Verwaltung und Bereitstellung von im GAMS vorliegenden Textressourcen wie auch dezidiert linguistischen Forschungsdaten wurde das bestehende TEI Content Model angepasst und erweitert. Über den Cirilo Client können Objekte als Text- bzw. Sprachressourcen gekennzeichnet werden. So gekennzeichnete Objekte werden dann automatisch mit für das CMDI Framework (Goosen et al., 2015) aufbereiteten, komponentenbasierten Metadaten und einem eigenen Handle Identifier versehen. Solche Daten können dann geharvestet werden und über das Virtual Language Observatory (Van Uytvanck et al., 2012) der CLARIN Infrastruktur[5] als Sprachressource gefunden werden.

Der OAI-Endpoint des Repositoriums wurde dementsprechend angepasst. Auf inhaltlicher Ebene wurde ein XML-basiertes Konfigurationsformat eingeführt, das es erlaubt, auf den Ausgangsdaten operierende Pipelines bzw. Toolchains zu definieren und als Massenoperation zu triggern. Ein Anwendungsfall hierfür ist beispielsweise Preprocessing zur Aufbereitung der Daten für darauf aufbauende Analyseschritte. Per Default wird eine, auf dem an der Österreichischen Akademie der Wissenschaften entwickelten XSL-Tokenizer (Schopper, 2019) basierende Pipeline ausgeführt, was einerseits ein tokenisiertes TEI-Dokument als separaten Datenstrom im Objekt erzeugt, und andererseits die Daten als Plain Text, im von den im Rahmen von CLARIN entwickelten Weblicht Tools verwendeten *Text Corpus Format* (TCF)[6], sowie im von gängigen Corpus Tools verwendeten *Vertical*-Format bereit stellt. Diese Daten können daraufhin direkt mit den genannten Tools verarbeitet und analysiert werden. Wie der Tokenizer ist auch die Pipeline selbst projektbezogen anpassbar und kann aus mehreren Transformationsschritten bestehen, darunter beispielsweise auch die Möglichkeit, die jeweiligen Texte via TreeTagger (Schmid, 1995) zu annotieren.

Die über diese Pipelines erzeugten Datenformate können benutzerdefiniert gekapselt und mit dem primären TEI-Datenstrom als Objekt im Repositorium langzeitarchiviert werden. Durch die Speicherung der Verarbeitungspipeline gemeinsam mit den zu verarbeitenden Daten wird jeder Prozessierungsschritt dokumentiert und nachvollziehbar gemacht, was wesentlich für die Nachnutzung ist.

Für die Aggregation mehrerer TEI Objekte zu einem verarbeitbaren Corpus wurde ein sogenanntes *Corpus Context Model* geschaffen. Diesem Modell entsprechende Objekte können vom Benutzer selbst über den *Cirilo* Client angelegt und mit entsprechenden Textobjekten befüllt werden.

Dieser spezielle Context stellt über die entsprechenden Datenströme Dublin Core wie auch CMDI Metadaten bereit. Die VERTICAL Datenströme der zugeordneten TEI-Objekte werden zu einem Datenstrom aggregiert, welcher bei Bedarf in einem Corpus Management System (Vorzugsweise NoSketch Engine) indiziert und über dieses abgefragt werden kann. Das aggregierte Corpus kann außerdem als ZIP-Datei heruntergeladen werden.

Die beschriebenen Features stehen für sämtliche im Repositorium vorhandenen Textressourcen, also nicht nur für genuin linguistische Daten zur Verfügung. Das bedeutet, dass etwa bestehende, im Repository vorhandene Digitale Editionen mit geringem Aufwand auch für linguistische Analysen verfügbar gemacht werden können.

Fußnoten

1. http://gams.uni-graz.at/
2. https://www.re3data.org/

3. https://www.coretrustseal.org
4. Flexible Extensible Digital Object Repository Architecture, https://duraspace.org/fedora
5. European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/
6. http://weblicht.sfs.uni-tuebingen.de

Bibliographie

Goosen T., et al. 2015. CMDI 1. 2: Improvements in the CLARIN Component Metadata Infrastructure. Selected papers from the CLARIN 2014 Conference, pp. 36-53. https://hdl.handle.net/20.500.11755/91536b93-31cb-4f4a-8125-56f4fe0a1881.

Rybicki, J. (2019). Keynote at the 2019 TEI Conference and members meeting "What is text, really? TEI and beyond".

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schopper, D. (2019). XSLT-Tokenizer (Software), https://github.com/acdh-oeaw/xsl-tokenizer.

Van Uytvanck, D., et al. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 1029-1034. http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf.

Geisteswissenschaftliches Forschungsdatenmanagement in der Lehre – Konzepte, Methoden, Erfahrungen

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities (CCeH);
Universität zu Köln, Data Center for the Humanities (DCH)

Helling, Patrick
patrick.helling@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH);
Universität zu Köln, Institut für Digital Humanities (IDH)

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH);
Universität zu Köln, Institut für Digital Humanities (IDH)

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities (CCeH);
Universität zu Köln, Data Center for the Humanities (DCH)

Rau, Felix
f.rau@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH);
Universität zu Köln, Institut für Linguistik (IfL)

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Universität zu Köln, Data Center for the Humanities (DCH)

Wieners, Jan Gerrit
jan.wieners@uni-koeln.de
Universität zu Köln, Institut für Digital Humanities (IDH)

Abstract

Mit diesem Beitrag möchten wir unsere Erfahrungen in der Vermittlung von Strategien und Ansätzen im Bereich des Forschungsdatenmanagements (FDM), die wir unter Einbeziehung unterschiedlicher Zielgruppen in der Lehre an der Universität zu Köln (UzK) gesammelt haben, vorstellen und mit der DH-Community diskutieren. Wir möchten dabei insbesondere die Frage aufwerfen, ob das Thema FDM als obligatorischer Lehrinhalt verstetigt werden sollte und wenn ja, inwieweit sich das Thema dauerhaft in die Lehre einbringen lässt, damit frühzeitig Sensibilität für Herausforderungen in der Erfassung, Bereitstellung und langfristigen Archivierung von Forschungsdaten aufgebaut werden kann.

Entwicklung des institutionellen Rahmens

Die Beschäftigung mit den Themen FDM und Langzeitarchivierung ist am *Cologne Center for eHumanities* (CCeH), dem Kölner *Data Center for the Humanities* (DCH) sowie dem *Institut für Digital Humanities* (IDH) (2017 hervorgegangen aus dem Institut für Historisch-Kulturwissenschaftliche Informationsverarbeitung und der Abteilung Sprachliche Informationsverarbeitung des Instituts für Linguistik) intensiv mit Forschung, Lehre und Projektpraxis verknüpft.[1] So flossen bereits früh Erfahrungen aus Projekten wie dem bis 2010 von der europäischen Union geförderten Projekt *PLANETS* oder dem *Digitalen Archiv NRW* in Bachelor- und Masterveranstaltungen ein.[2]

Ausgehend von der Annahme, dass es längerfristig zu einer Verbesserung der nachhaltigen Sicherung von wissenschaftlichen Inhalten führt, wenn Forschende bereits zu Beginn ihrer Ausbildung mit dem Problembereich und den verfügbaren Lösungsansätzen vertraut gemacht werden, haben wir am DCH mehrere Lehrveranstaltungen zum Themenbereich FDM konzipiert und im Rahmen der Studiengänge "Informationsverarbeitung" und "Medieninformatik" durchgeführt.

Die Lehrveranstaltungen thematisieren sowohl das Management von Forschungsprimärdaten, wie auch die Nachnutzbarkeit von erzeugten digitalen Artefakten (z.B. Präsentationssysteme, Dateneingabemasken und weitere). Zielgruppe der Lehrveranstaltungen sind vor allem Studierende der (di-

gitalen) Geisteswissenschaften. Darüber hinaus veranstalten wir regelmäßig einen Workshop zum Thema Forschungsdatenmanagement für Promovierende an der a.r.t.e.s. Graduiertenschule der Philosophischen Fakultät, in dem insbesondere praktische FDM-Aspekte während der Promotion im Fokus stehen.[3]

Ziel ist es unter anderem direkte Rückmeldung von den Studierenden/Promovierenden als – wenn auch punktuelle – praktische Anwender von FDM-Angeboten zu erhalten. Der Fokus liegt hierbei besonders auf der Nutzererfahrung. So wird in den Kursen unter anderem diskutiert, wie aufwendig FDM in der Forschungsarbeit ist, und ob es in manchen Fällen als hinderlich oder gar belastend empfunden wird. Im Zuge der Veranstaltungen wird auch der Bekanntheitsgrad typischer FDM-Werkzeuge evaluiert. Diese direkte Rückmeldung aus der Erfahrungswelt angehender Wissenschaftler*innen fließt unmittelbar in die Beratungs- und Schulungskonzepte des DCH ein.

Im Folgenden beschreiben wir zunächst die Lehrinhalte des Workshops und der beiden Übungen, bevor wir abschließend ein kurzes Fazit ziehen, das als Grundlage für eine Diskussion mit der Community dienen soll.

Workshop „Forschungsdatenmanagement“ an der a.r.t.e.s. Graduiertenschule der Philosophischen Fakultät

Um Promovierende frühzeitig mit dem Thema Forschungsdatenmanagement vertraut zu machen, veranstalten wir seit dem Sommersemester 2018 einmal im Jahr einen zweitägigen Workshop an der a.r.t.e.s. Graduiertenschule der Philosophischen Fakultät der UzK. Der Workshop findet aufgeteilt in vier Sessions à vier Stunden/Woche statt, ist auf maximal 15 Promovierende ausgelegt und gehört zum Wahlpflichtbereich.

Im ersten Workshop im Sommersemester 2018 lag der Fokus auf der theoretischen Betrachtung von Forschungsdatenmanagement innerhalb der Geisteswissenschaften. Als Struktur dienten hierzu die acht Säulen des Forschungsdatenmanagements der AG Datenzentren des Verbandes “Digital Humanities im deutschsprachigen Raum” (DHd) [Helling, Moeller und Mathiak 2018].[4]

In der Evaluation des Workshops wurde deutlich, dass sich die Informations- und Servicebedarfe der Promovierenden deutlich von denen von BA-/MA-Studierenden oder Forschenden im universitären Regelbetrieb unterscheiden: Neben Informationen über Servicestrukturen an der UzK wurden auch insbesondere praktische Lösungen für FDM in Bezug auf das eigene Promotionsprojekt nachgefragt.

Entsprechend wurde für den zweiten Workshop im Sommersemester 2019 gemeinsam mit Vertreter*innen des Regionalen Rechenzentrums Köln (RRZK) ein Diensteportfolio für Software- und Hardwareservices an der UzK zusammengestellt, das in einer der Sessions vorgestellt und diskutiert wurde.[5] Der gesamte Workshop wurde deutlich praxisorientierter geplant und durchgeführt: Nach einem verkürzten Theorie- und Einführungsblock wird in einer praktischen Übung der Forschungsdatenlebenszyklus aus einer datengebenden und datennehmenden Perspektive nachvollzogen. Mit einer Rechercheübung wird anhand eines Fragenkatalogs das passende Repositorium für den eigenen Fachbereich gesucht, eigene Forschungsdaten werden mit Metadaten beschrieben und in einer Diskussionsrunde werden die Themen Urheberrecht und Persönlichkeitsrecht und deren Auswirkungen auf das Forschungsdesign behandelt. Als Kursabschluss formulieren die Teilnehmenden einen Datenmanagementplan für ihr Promotionsprojekt nach dem Datenmanagementplan-Template der UzK.[6]

FDM-Übung für Masterstudierende an der Philosophischen Fakultät

Für Studierende im Masterstudiengang “Informationsverarbeitung” bieten wir ebenfalls Veranstaltungen an. Die im Sommersemester 2017 von Simone Kronenwett und Jan G. Wieners durchgeführte Veranstaltung “Forschungsdatenmanagement und Langzeitarchivierung” intendierte, Teilnehmende an verschiedenste Aspekte des FDM und der Langzeitarchivierung heranzuführen und animierte dazu, eigene Forschungsfragen zu entwickeln.[7] So gossen Gastvorträge von domänenrelevanten Fachwissenschaftler*innen aus Einrichtungen wie dem Forschungsdatenzentrum Archäologie & Altertumswissenschaften IANUS oder dem Rheinisch-Westfälischen Wirtschaftsarchiv das Fundament für Ausarbeitungen der Studierenden, die sich beispielsweise mit Fragen nach guten Speicherformaten für Rastergraphiken oder die Sicherheit physischer Speichermedien im Hinblick auf Langzeitverfügbarkeit beschäftigten.[8]

Auch die Übung “Forschungsdatenmanagement”, die wir seit 2019 anbieten, richtet sich an Studierende in den Masterstudiengängen “Informationsverarbeitung” und “Medieninformatik”, hat dabei aber einen stärkeren praktischen Fokus.[9] Anders als bei Kursen, die zukünftige Wissenschaftler*innen ausschließlich als potentielle Datenproduzent*innen adressieren, legen wir den Fokus nicht nur auf FDM-Grundlagen, Awareness und Fortbildung, sondern beschäftigen uns aktiv mit den Problemen aus der Sicht von Forschungsdatenmanager*innen und -kurator*innen. Es wird das Datenmanagement und die -kuration als mögliches Berufsfeld vorgestellt.

Bei der inhaltlichen Gestaltung haben wir uns an Kursen orientiert, die bereits andernorts erfolgreich durchgeführt wurden. Insbesondere der Kurs von Iris Vogel von der Universität Hamburg hat uns in vielen Punkten inspiriert, wie auch schon die bereits erwähnte Veranstaltung von 2017.[10] Ähnlich wie bei der Veranstaltung für Doktorand*innen setzen wir auf eine Mischung von Theorie und Praxis. Etwa die Hälfte der Sitzungen schließt mit einer praktischen Hausaufgabe, für die andere Hälfte soll Literatur vorbereitet werden, die in einem Quizformat diskutiert wird.

Während das Format bei den Studierenden grundsätzlich gut ankam, waren die Themen zum Teil doch sehr weit von ihrem Arbeitsalltag entfernt. Dazu kam, dass die Gruppe sehr heterogen war. Während die linguistisch orientierten Studierenden bereits Forschungsdaten für ihre Projekte genutzt hatten und gerne über die Details sprechen wollten, waren die eher medien- oder literaturwissenschaftlich ausgerichteten Studierenden zum Teil noch gar nicht mit dem Thema Forschungsdaten in Berührung gekommen und dementsprechend skeptisch. In der Zukunft planen wir dies zu adressieren, indem

wir beide Gruppen noch stärker fachspezifisch abholen und beispielsweise digitale Editionen und die dazugehörige Datenmodellierung stärker in den Fokus nehmen.

Seminar „Virtualisierungstechnologien für Forschungssoftware“

In einem weiteren Seminar, das wir seit dem Wintersemester 2018/2019 anbieten, fokussieren wir das spezifische Problem der nachhaltigen Sicherung von Forschungssoftware. Im Gegensatz zu datenbezogenem FDM ist die Frage nach dem Umgang mit Forschungssoftware bislang noch weitgehend unbeantwortet und stellt deshalb eine besondere Herausforderung im Hinblick auf die langfristige Sicherung von Forschungsergebnissen dar. Kontext der Seminare ist das von der DFG geförderte Projekt “SustainLife”, welches die Erprobung eines Konzepts zur nachhaltigen Sicherung lebender Systeme auf Basis des OASIS-Standards TOSCA (*Topology and Orchestration Specification for Cloud Applications*) [OASIS 2013, 2016] zum Gegenstand hat [Barzen et. al 2018, Neuefeind et al. 2018, 2019].[11] Die im Projekt erarbeiteten Methoden und Erfahrungen sollen dazu beitragen zukunftssichere Standards und Nachhaltigkeitsstrategien für den Umgang mit Forschungssoftware zu etablieren. Dies soll unter anderem durch die Modellierung von Applikationstopologien mittels der Open-Source Implementierung OpenTOSCA erreicht werden, die durch den Projektpartner IAAS Stuttgart bereitgestellt wird.[12]

Im Rahmen des Seminars werden zum einen verschiedene Nachhaltigkeitsstrategien beleuchtet, zum anderen wollen wir den im Projekt entwickelten Lösungsansatz auf Grundlage des TOSCA-Standards von Studierenden erproben lassen. Im Mittelpunkt steht dabei die Frage, ob die TOSCA-konforme Modellierung heterogener Softwarelösungen in einer Weise vermittelt werden kann, dass Studierende der (digitalen) Geisteswissenschaften in der Lage sind, eigene Use Cases mit akzeptablem Aufwand zu bearbeiten.

Die Übungen wurden trotz der stark spezialisierten Ausrichtung sehr gut angenommen. In der ersten Übung zeigte sich dabei eine sehr deutliche Lücke zwischen geisteswissenschaftlich-konzeptueller und informationstechnologischer Kompetenz. Während das grundlegende Konzept einer TOSCA-konformen Modellierung und dessen Mehrwert von den Studierenden schnell erfasst und gut nachvollzogen wurde, stellte sich die praktische Umsetzung als deutlich größere Herausforderung heraus. Dies konnte durch Gruppenarbeiten zwar zu Teilen aufgefangen werden, dennoch wurde deutlich, dass die konkrete Umsetzung von Anwendungsmodellen ein erhebliches softwaretechnologisches Vorwissen voraussetzt. In der Folge adressierten wir dies dadurch, dass wir die reine Modellierung von der technischen Umsetzung stärker trennen, so dass die Gruppen sich nach technischen Kompetenzen aufteilen können.

Fazit und Ausblick

Die hier vorgestellten Lehrveranstaltungen fokussieren auf unterschiedliche Weise verschiedene Aspekte des Forschungsdatenmanagements und sprechen unterschiedliche Zielgruppen an. Über die genannten Veranstaltungen hinaus wird das Thema FDM an der UzK auch in weiteren Veranstaltungen thematisiert, wenn auch nicht so ausschließlich. Insgesamt haben wir gute Erfahrungen mit den Veranstaltungen gemacht. Das Feedback zeigt, dass eine starke Verankerung in der Praxis bei den Studierenden gut ankommt.

In unserem Poster werden wir den Aufbau der Kurse genauer beleuchten und den Benefit für die Studierenden diskutieren. Wir laden dazu ein, Erfahrungen auszutauschen, um zu eruieren welche Kompetenzen bereits im Studium zu Forschungsdatenmanagement erworben werden sollten und welche Aspekte für besonders dringlich gehalten werden.

Fußnoten

1. Cologne Center for eHumanities (CCeH) an der Universität zu Köln, Online: https://cceh.uni-koeln.de/ (letzter Zugriff: 12.09.2019); Data Center for the Humanities (DCH) an der Universität zu Köln, Online: https://dch.phil-fak.uni-koeln.de/ (letzter Zugriff: 12.09.2019); Institut für Digital Humanities (IDH) an der Universität zu Köln, Online: https://dh.uni-koeln.de/ (letzter Zugriff: 12.09.2019).
2. Preservation and Long-term Access through Networked Services, Online: (letzter Zugriff: 12.09.2019); Digitales Archiv NRW, Online: (letzter Zugriff: 12.09.2019).
3. a.r.t.e.s. Graduate School der Philosophischen Fakultät der Universität zu Köln, Online: http://artes.phil-fak.uni-koeln.de/ (letzer Zugriff: 12.09.2019); Workshop Forschungsdatenmanagement a.r.t.e.s., Online: https://artes.phil-fak.uni-koeln.de/40285.html (letzter Zugriff: 12.09.2019).
4. Digital Humanities im deutschsprachigen Raum (DHd), Online: https://dig-hum.de/ (letzter Zugriff: 12.09.2019); AG Datenzentren im DHd-Verband, Online: https://dhd-ag-datenzentren.github.io/ (letzter Zugriff: 12.09.2019).
5. Regionales Rechenzentrum (RRZK) der Universität zu Köln, Online: https://rrzk.uni-koeln.de/ (letzter Zugriff: 12.09.2019); Diensteportfolio des RRZK, Online: https://rrzk.uni-koeln.de/datenmanagement.html (letzter Zugriff: 12.09.2019).
6. Datenmanagementplan-Template der Universität zu Köln, Online: https://fdm.uni-koeln.de/27602.html (letzter Zugriff: 12.09.2019).
7. Forschungsdatenmanagement und Langzeitarchivierung - Übung Sommersemester 2017, Online: http://lehre.idh.uni-koeln.de/lehrveranstaltungen/sosem17/forschungsdatenmanagement-und-langzeitarchivierung/ (letzter Zugriff: 12.09.2019).
8. Forschungsdatenzentrum Archäologie & Altertumswissenschaften IANUS, Online: https://www.ianus-fdz.de/ (letzter Zugriff: 12.09.2019); Rheinisch-Westfälische Wirtschaftsarchiv, Online: https://www.rwwa.de/ (letzter Zugriff: 12.09.2019).
9. Forschungsdatenmanagement - Übung Sommersemester 2019, Online: http://lehre.idh.uni-koeln.de/lehrveranstaltungen/sosem19/forschungsdatenmanagement/ (letzter Zugriff: 12.09.2019).
10. Stefan Thiemann, Iris Vogel und Juliane Jacob: Materialien zum Kurs “Forschungsdatenmanagement” an der Universität Hamburg, Online: https://www.hoou.de/materials/forschungsdatenmanagement (letzter Zugriff: 12.09.2019).
11. SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften; Projektnummer 379522012. Online:

https://gepris.dfg.de/gepris/projekt/379522012 (letzter Zugriff: 12.09.2019).
12. Institut für Architektur von Anwendungssystemen, Universität Stuttgart; Online: https://www.iaas.uni-stuttgart.de (letzter Zugriff: 12.09.2019).

Bibliographie

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, S. 471-474, DOI: https://doi.org/10.18716/KUPS.8085.

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, Seiten 251–261, ISSN (Online) 2191-4664, ISSN (Print) 0720-6763, DOI: https://doi.org/10.1515/abitech-2018-3006.

Neuefeind, Claes / Schildkamp, Philip / Mathiak, Brigitte / Marčić, Alexander / Hentschel, Frank / Harzenetter, Lukas / Breitenbücher, Uwe / Barzen, Johanna / Leymann, Frank (2019): "Sustaining the Musical Competitions Database: a TOSCA-based Approach to Application Preservation in the Digital Humanities". To Appear in: Book of Abstracts of the Digital Humanities Conference 2019 (DH2019). Utrecht, Netherlands. 09.07-12-07.2019. Web: https://dev.clariah.nl/files/dh2019/boa/0574.html [letzter Zugriff 5.9.2019]

Neuefeind, Claes / Harzenetter, Lukas / Schildkamp, Philip / Breitenbücher, Uwe / Mathiak, Brigitte / Barzen, Johanna / Leymann, Frank (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 5.9.2019]

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 5.9.2019]

Historische Schulbücher als Spielräume für Digital Humanities? Mapping von unterschiedlichen Metadatenformaten für Bibliotheken und linguistische Analysen

De Luca, Ernesto William
deluca@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Fallucchi, Francesca
fallucchi@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Hertling, Anke
hertling@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Klaes, Jan Sebastian
klaes@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Schmitz, Claudia
schmitz@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Towara, Nadine
towara@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Schulbücher transportieren gesellschaftliche und staatlich sanktionierte Werte und Normen. Als Quellengattung stellen sie einen vielversprechenden Gegenstand für zahlreiche wissenschaftliche Fragestellungen dar. Schulbücher werden in zahlreichen Bibliotheken gesammelt, aber in den seltensten Fällen werden sie systematisch erschlossen und für die Digitalisierung genießen sie in der Regel keine hohe Priorität.

Im Rahmen der digitalen Schulbuchbibliothek GEI-Digital[1] wurden in den letzten 10 Jahren historische deutsche Schulbücher der Fächer Geschichte, Geographie und Politik, so-

wie Realien- und (Erst-)Lesebücher von den Anfängen der Schulbuchproduktion im 17. Jahrhundert bis zum Ende des Ersten Weltkriegs digital zugänglich gemacht (Hertling/Klaes 2018). Digitalisiert und integriert wurden dabei sowohl Schulbücher aus den Beständen der Forschungsbibliothek des Georg-Eckert-Instituts – Leibniz-Institut für internationale Schulbuchforschung (GEI) als auch Schulbücher aus zahlreichen Partner-Bibliotheken im deutschsprachigen Raum. Die externen Schulbuchbestände wurde dem GEI zum Zwecke der Digitalisierung im Rahmen von Kooperationen leihweise überlassen oder als Fremddigitalisate virtuell in die GEI-Digital-Sammlung integriert.

Die bibliothekarische Erschließung des historischen Schulbuch-Korpus folgt den spezifischen Bedürfnissen der Schulbuchforschung. Typisches Kennzeichen von Schulbüchern sind dabei viele Ausgaben und die unterschiedlichen Bände eines Schulbuchs. Neben Angaben zum Verlag und Erscheinungsjahr werden zusätzlich auch Schulfächer und Schulstufen als deskriptive Metadaten erfasst. Sie stehen auf GEI-Digital als MARCXML, Metadata Object Description Schema (MODS) oder Dublin Core (DC) zur Nachnutzung zur Verfügung. Die Erschließung umfasst auch die intellektuelle Verknüpfung der Schulbuchautoren mit Normdateneinträgen in der Gemeinsamen Normdatei (GND), um das Korpus für biographische Forschungsansätze zu öffnen. Darüber hinaus werden die digitalisierten Schulbücher in Form einer Tiefenerschließung in ihrer Struktur erschlossen und die Elemente, wie Titelblätter, Inhaltsverzeichnisse, Abbildungen etc. im Metadata Encoding & Transmission Standard (METS) ausgewiesen.

Die über 1,5 Millionen in GEI-Digital gescannten Seiten wurden zudem einer Optical Character Recognition (OCR) unterzogen und stehen als durchsuchbare Volltexte über eine OAI-PMH-Schnittstelle zur Verfügung. Die Resultate der Volltexterkennung werden im Zuge des Digitalisierungsworkflows im XML-Schema Analyzed Layout and Text Object (ALTO) ausgegeben.

Mit GEI-Digital ist für die Digital Humanities ein einzigartiges Korpus mit über 6.100 digitalisierten Schulbüchern entstanden, dass die gesamte Epoche der deutschen Schulbücher von deren Entstehung bis 1918 mit hoher Vollständigkeit virtuell zusammenführt. Die Digitalisate und Daten werden in zahlreichen Digital-Humanities-Projekten bereits nachgenutzt, z.B. im Projekt „Welt der Kinder“[2] (Heuwing/Weiß 2018 und Nieländer/Weiß 2018), in dem das Korpus mit Topic Modeling-Verfahren untersucht wurde . Das Portal GeoPortOst[3] nutzt u.a. das in GEI-Digital vorhandene Kartenmaterial für Georeferenzierungen.

In einem nächsten Schritt ist geplant, das Korpus in dem von CLARIN betriebenen Virtual Language Observatory (VLO)[4]nachzuweisen, um es für weiterführende und v.a. linguistische Analysen zugänglich zu machen. Voraussetzung für einen Nachweis ist die Repräsentation der digitalisierten Schulbücher in derComponent MetaData Infrastructure (*CMDI*). CMDI stellt ein Framework zur Verfügung, um Profile für Metadaten für die Beschreibung und Benutzung bereitzustellen.

Ausgehend von Metadatenformaten, die sich v.a. an bibliothekarischen Standards orientieren, werden auf dem Poster Anforderungen und Strategien für Mapping-Prozesse als Grundlage für Digital-Humanities-Projekte präsentiert. Im Mittelpunkt stehen dabei die in GEI-Digital gemachten Mapping-Erfahrungen mit den Formaten METS/MODS, TEI, CMDI und Dublin Core (DC) und die Herausforderung ihrer jeweiligen Interoperabilität.

In einem ersten Schritt In einer Machbarkeitsstudie stellte sich im Projekt GEI-Digital ein Mapping von METS zu CMDI als undurchführbar heraus (Fallucchi/De Luca 2019). Ein Mapping von Dublin Core (DC) zu CMDI als CLARIN-Empfehlung ist mit Blick auf die Besonderheiten der Erschließung von Schulbüchern insbesondere mit Blick auf die für die Forschung wichtigen Ausgabebezeichnungen und Bandangaben stark verlustbehaftet. Vor dem Hintergrund werden derzeit alternative Optionen diskutiert, die auf dem Poster aufgezeigt und erörtert werden sollen. Eine Möglichkeit stellt die Anreicherung von Dublin Core-Metadaten und ihre Konvertierung in CMDI dar. Eine weitere Option besteht in der Umwandlung von textbasierten ALTO-Dateien in CMDI.

Fußnoten

1. Projektwebseite GEI-Digital – die digitale Schulbuch-Bibliothek:: http://gei-digital.gei.de/viewer/
2. Projektwebseite Welt der Kinder: http://welt-der-kinder.gei.de/
3. Projektwebseite *GeoPortOst*: Portal für thematische und versteckte Karten zu Ost- und Südosteuropa: http://geoportost.ios-regensburg.de/
4. CLARIN Virtual Language Observatory: https://vlo.clarin.eu/;jsessionid=6E94D6CA11D6A8A1889F2C3E6C6A79FE?0

Bibliographie

Fallucchi, Francesca / De Luca, Ernesto William (2019): “Connecting and Mapping LOD and CMDI Through Knowledge Organization” Springer, Cham, pp. 291-301.

Hertling, Anke / Klaes, Sebastian (2018): Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek, in: Maret Nieländer / Ernesto William De Luca (eds): Digital Humanities in der internationalen Schulbuchforschung - Forschungsinfrastrukturen und Projekte. Göttingen: V&R unipress 21–44

Hertling, Anke / Klaes, Sebastian (2018): »GEI-Digital« als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung, in: Maret Nieländer / Ernesto William De Luca (eds): Digital Humanities in der internationalen Schulbuchforschung - Forschungsinfrastrukturen und Projekte. Göttingen: V&R unipress 45–68

Heuwing, Ben / Weiß, Andreas (2018): Suche und Analyse in großen Textsammlungen: Neue Werkzeuge für die Schulbuchforschung in: Maret Nieländer / Ernesto William De Luca (eds): Digital Humanities in der internationalen Schulbuchforschung - Forschungsinfrastrukturen und Projekte. Göttingen: V&R unipress 145-170

Nieländer, Maret / Weiß, Andreas (2018): »Schönere Daten« – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten, in: Maret Nieländer / Ernesto William De Luca (eds): Digital Humanities in der internationalen Schulbuchforschung - Forschungsinfrastrukturen und Projekte. Göttingen: V&R unipress 91-116

Interdisziplinäres Streitgespräch – Nutzerkommentar-analysen aus ethisch-rechtlicher Perspektive

Brokering, Annalena
A.Brokering@sms.ed.ac.uk
The University of Edinburgh, Law School, Vereinigtes Königreich

Guhr, Svenja
Guhr@linglit.tu-darmstadt.de
TU Darmstadt, Deutschland

Einleitung

Täglich werden auf der ganzen Welt Onlineartikel, Blogbeiträge etc. veröffentlicht, zu denen Leserinnen und Leser (i.F. generisches Femininum) Kommentare verfassen. Aufgrund der hohen Anzahl an Partizipierenden gelten Nutzerbeiträge als besonders authentische Echtzeitrückmeldungen und erlauben einen Zugang zu heterogenen Meinungsäußerungen (Busch, 2017). Auch durch Partizipation auf Social Media Plattformen, in Onlineforen und öffentlichen Chats werden Daten generiert, die wertvolle Informationen über Nutzerverhalten und menschliches Denken beinhalten. Dies gilt umso mehr, als diese Plattformen Orte sind, an denen Menschen miteinander in Verbindung treten, in verschiedenen Formen und Dimensionen Gemeinschaft pflegen, Informationen verbreiten und ihre Meinungen austauschen. Dabei generierte Daten zeichnen sich durch ihren interaktiven, kontemporären und personenbezogenen Charakter aus und ermöglichen folglich Rückschlüsse auf Meinungen, Interessen und Stimmungen in der Bevölkerung. Entsprechend sind sie von besonderem Interesse für private Unternehmen oder öffentliche Institutionen (Schoen, 2002; Holtz-Bacha, 2019: 276). Zunehmend wird auch im akademischen Kontext auf Nutzerdaten zurückgegriffen (u.a. Mohammad, 2016; Aker et al., 2016).

Forschungsgegenstand

Gegenstand des vorzustellenden Projektes ist eine Pilotstudie, in der zwei Masterthesisprojekte in Beziehung zueinander gesetzt wurden. Bei dem ersten Thesisprojekt (Guhr, 2019) handelt es sich um eine im Bereich der Digital Humanities durchgeführte computergestützte Analyse von Leserkommentaren in französischen Onlinemedien. Das zweite Thesisprojekt aus dem Bereich des IT-/Datenschutzrechts betrachtet ethische und rechtliche Erwägungen bei der Analyse von nutzergenerierten Inhalten auf Social Media Plattformen und die Frage, wie deren Berücksichtigung in wissenschaftlichen Datenanalyseprojekten unterstützt werden kann. Infolge der Auseinandersetzung mit dem jeweils anderen Masterthesisprojekt entstand ein interdisziplinäres Streitgespräch zwischen den Autorinnen.

Die Masterthesis (Guhr, 2019) umfasst u.a. verschiedene gemischt qualitativ-quantitative Analysen von Onlinezeitungsartikeln und zugehörigen Leserkommentaren zur französischen Präsidentschaftswahl 2017. Die Daten sind über den Onlineauftritt einer großen französischen Tageszeitung öffentlich zugänglich. 40 ausgewählte Onlineartikel mit den dazugehörigen 3.127 Leserkommentaren wurden zu einem Korpus zusammengestellt. Die extrahierten Leserkommentardaten beinhalteten zusätzlich zu den nutzergenerierten Beitragstexten auch Datum und Uhrzeit der Beitragserstellung sowie die Nicknames und teilweise bürgerliche Namen der Nutzerinnen. Mithilfe von Distant Reading Methoden wurden im Korpus behandelte Themen identifiziert. Anschließend wurde eine automatisierte Sentimentanalyse der Kommentare durchgeführt, um Informationen über die emotionale Einstellung der Nutzerinnen zu Wahlkampfthemen und zum/zur Präsidentschaftskandidat/in herausstellen zu können.

Der zweiten Masterthesis (Brokering, 2019) lag die Frage zugrunde, wie Datenanalyseprojekte im akademischen Kontext rechtskonform und ethischer gestaltet werden können. Am Beispiel von Forschung mit Social Media Daten wurde herausgestellt, wie durch Analysen von nutzergenerierten Inhalten Interessen und Rechte der Nutzerinnen berührt werden. Für die hierdurch aufgeworfenen, neuen ethischen und besonders datenschutzrechtlichen Fragestellungen fehlt es bestehenden inhaltlichen und institutionellen Ansätzen der Forschungsethik noch an befriedigenden Antworten, die eine ethische Praxis von Social Media Datenanalysen gewährleisten. Daraufhin wurde evaluiert, inwiefern das IT-rechtliche Konzept des *Regulation by Design* eine effektivere Implementierung ethischer und rechtlicher Erwägungen in Social Media Datenanalysen unterstützen kann. *Regulation by Design* zielt auf eine proaktive Berücksichtigung regulatorischer Erwägungen bereits im Zeitpunkt des Designs, d.h. der Planung und Entwicklung, von Produkten und Aktivitäten wie auch Forschung. Es findet seine bekannteste Ausprägung im datenschutzrechtlichen Prinzip des *Privacy by Design*.

Interdisziplinäres Streitgespräch

Im Dialog der Autorinnen trafen die Perspektiven der praxisorientierten und der juristischen Forschung aufeinander. Aus letzterer wurde Kritik am Umgang mit persönlichen Informationen geäußert und für eine höhere Sensibilität gegenüber den Interessen der Nutzerinnen und insbesondere datenschutzrechtlichen Erwägungen plädiert. Sobald Social Media Daten eine Identifizierbarkeit der postenden Personen auch nur ermöglichen, z.B. weil sie die Nutzernamen oder auch die IP-Adresse der Nutzerin enthalten, handelt es sich um persönliche Daten und damit finden datenschutzrechtliche Vorgaben wie die europäische DSGVO Anwendung. Diese erfordert typischerweise die Information der betroffenen Nutzerin über die konkrete Verwertung ihrer Daten und die Einwilligung in diese. Die Wirksamkeit einer mittels der AGB des jeweiligen Social Media Anbieters erteilten Einwilligung ist als zweifelhaft zu bewerten, da sie nicht projektspezifisch ist. Angesichts des Umstands, dass die verwendeten Nutzerdaten zu Forschungszwecken umgewidmet werden und originär im Rahmen der privaten Nutzung von Social

Media Diensten entstanden sind, ist in Erwägung zu ziehen, ob über das datenschutzrechtlich erforderliche Mindestmaß hinausgehende Maßnahmen zum Schutz der Interessen der Nutzerinnen ethisch geboten sind. Auch eine Anonymisierung der Nutzerdaten, z.B. durch Entfernen des Nutzernamens kann das Re-Identifikationsrisiko angesichts fortschrittlicher De-Anonymisierungstechniken nur reduzieren. Hier sind weitergehende u.a. auch im Verhältnis zum Grad an Sensibilität der betroffenen Nutzerinhalte angemessene Anonymisierungsmaßnahmen in Erwägung zu ziehen. Gleichzeitig ist zu berücksichtigen, dass Nutzerinnen an ihren originell und kreativ gestalteten Beiträgen auch urheberrechtliche Interessen und Rechte haben, sodass andererseits eine Erkennbarkeit der Autorin durch die Forschenden sicherzustellen sein kann. Die praxisorientierte Forscherin wies demgegenüber auf die schwierige Umsetzbarkeit aufwendiger Maßnahmen zum Schutz der Nutzerinnen angesichts begrenzter finanzieller, technischer und zeitlicher Spielräume in der Forschungspraxis hin sowie auf die Gefahr, dass Forschungsdaten durch Datenschutzmaßnahmen an Wert/Aussagekraft verlieren würden. Als Beispiele nannte sie den Wert von Nutzernamen als potenzielle Informationsquelle hinsichtlich Gender und Nationalität sowie für die kumulative Betrachtung verschiedener Beiträge einer Person. Auch die Erhebung von Datum und Uhrzeit der Beitragserstellung ermögliche eine chronologische Ordnung von Beiträgen. Dabei kritisierte die Juristin, dass bereits aus derartigen Informationen umfangreiche Aktivitätsprofile über einzelne Nutzerinnen erstellt werden könnten, die ggf. in Verbindung mit Nutzungsdaten derselben Nutzerinnen auf weiteren Social Media Plattformen Rückschlüsse auf Tagesabläufe, Vorlieben und Social Media Verhalten einzelner Nutzerinnen erlauben. Im daraus resultierenden Streitgespräch wurde erkennbar, wie schwierig es ist, die jeweiligen Positionen in einer der anderen Forschenden verständlichen Weise zu kommunizieren.

Weiteres Vorgehen im Projekt war es, die rechtlich-ethischen Herausforderungen gemeinsam zu definieren, wobei die Perspektiven beider Forschungsrichtungen Beachtung finden sollten. Auf dieser gemeinsamen Grundlage und unter Berücksichtigung verschiedener Ansätze des *Regulation by Design*-Konzeptes wurden Methoden und Herangehensweisen diskutiert, die eine effektive Berücksichtigung der Herausforderungen in der Forschungspraxis erreichen sollen.

Das Projekt verfolgt damit das Ziel, den Dialog zwischen datenbasierter Forschung und IT-Recht anzuregen und insbesondere das Bewusstsein für Nutzerinteressen und Datenschutzerwägungen unter Forschenden zu erhöhen. Es soll reflektiert werden, wie die Kommunikation zwischen IT-Rechtlerinnen und Datenforschenden unter Berücksichtigung der unterschiedlichen Perspektiven, Interessen und Limitierungen verbessert werden kann. Die gemeinsamen Definitionen der Herausforderungen und die diskutierten Lösungsvorschläge sollen Datenforschenden ermöglichen, ihre Forschungsarbeit ohne größeren Mehraufwand bereits im Stadium der Vorbereitung und Durchführung von Datenanalysen rechtskonform und ethisch sensibel zu gestalten.

Bibliographie

Aker, Ahmet / Paramita, Monica / Kurtic, Emina / Funk, Adam / Barker, Emma (2016): „Automatic label generation for news comment clusters“, in: *Proceedings of the 9th International Natural Language Generation Conference , Edinburgh, UK:* 61–69 https://pdfs.semanticscholar.org/4da7/ac02c56d43312425a854d63e71f89dd288ec.pdf [letzter Zugriff 26. Juni 2019].

Brokering, Annalena (2019): *Drawing from approaches in regulatory theory for the regulation of new technologies and design theory, how can ethical considerations be effectively incorporated into data science activities?.* Masterthesis, The University of Edinburgh, Edinburgh Law School.

Buchanan, Elizabeth / Zimmer, Michael (2016): „Internet Research Ethics“, in: *Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/ethics-internet-research/ [letzter Zugriff 19. August 2019].

Busch, Andreas (2017): Informationsinflation: Herausforderungen an die politische Willensbildung in der digitalen Gesellschaft, in: Gapski, Harald / Oberle, Monika / Staufer, Walter (eds.): *Medienkompetenz. Herausforderungen für Politik, politische Bildung und Medienbildung,* Schriftenreihe der Bundeszentrale für Politische Bildung. Bonn: Bundeszentrale für Politische Bildung 53-62 http://www.bpb.de/lernen/digitale-bildung/medienpaedagogik/medienkompetenz-schriftenreihe/257594/informationsinflation [letzter Zugriff 11. Juni 2019].

Dashtipour, Kia / Poria, Soujanya / Hussain, Amir / Cambria, Erik / Hawalah, Ahmad Y. A. / Gelbukh, Alexander / Zhou, Qiang (2016): „Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques“, in: *Cognitive Computation* (2016) 8: 757-771 https://link.springer.com/article/10.1007/s12559-016-9415-7 [letzter Zugriff 26. Juni 2019].

Golder, Susan P. / Ahmed, Shahd / Norman, Gill / Booth, Andrew (2017): „Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review“, in: *Journal of Medical Internet Research* 19 http://eprints.whiterose.ac.uk/117721/ [letzter Zugriff 26. September 2019].

Guhr, Svenja (2019): *Computergestützte Analyse von französischen Onlinemedien zur Präsidentschaftswahl 2017*, Masterthesis, Georg-August-Universität Göttingen.

Holtz-Bacha, Christina (2019): „Demoskopie - Medien - Politik. Umfragen im Bundestagswahlkampf 2017“, in: Holtz-Bacha, Christina: *Die (Massen-)Medien im Wahlkampf. Die Bundestagswahl 2017.* Wiesbaden: Springer Fachmedien Wiesbaden GmbH 263-280.

Locatelli, Elisabetta (2018): „Ethics of Social Media Research: State of the Debate and Future Challenges“, in: Hunsinger, Jeremy / Klastrup, Lisbeth / Allen, Matthew M. (eds.): *Second International Handbook of Internet Research.* Dordrecht: Springer 1-22.

McKee, Heidi / Porter, James E. (2008): „The Ethics of Digital Writing Research: A Rhetorical Approach“, *College Composition and Communication* 59: 711 http://wrconf08.writing.ucsb.edu/Pdf_Articles/McKee_Article.pdf [letzter Zugriff 26. September 2019].

Mohammad, Saif (2016): „A Practical Guide to Sentiment Annotation: Challenges and Solutions“, in: *Proceedings of the NAACL 2016 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)* 174-179 http://www.aclweb.org/anthology/W16-0429 [letzter Zugriff 16. Mai 2019].

Moreno, Megan A. / Goniu, Natalie / Moreno, Peter S. / Diekema, Douglas (2013): „Ethics of Social Media Research: Common Concerns and Practical Considerations“, in: *Cyberpsychology, Behavior, and Social Networking* 16: 708-713 https://

www.ncbi.nlm.nih.gov/pmc/articles/PMC3942703/ [letzter Zugriff 16. Mai 2019].

Perez Vallejos, Elvira / Koene, Ansgar / Carter, Christopher J. / Hunt, Daniel / Woodard, Christopher / Urquhart, Lachlan / Bergin, Aislinn / Statche, Ramona (2019): „Accessing Online Data for Youth Mental Health Research: Meeting the Ethical Challenges", in: *Philosophy & Technology* 32: 87-110 https://link.springer.com/article/10.1007/s13347-017-0286-y [letzter Zugriff 26. September 2019].

Schoen, Harald (2002): „Wirkung von Wahlprognosen auf Wahlen" in: Berg, Thomas (ed.) (2002): *Moderner Wahlkampf. Blick hinter die Kulissen.* Opladen: Leske und Budrich 171-191.

Williams, Matthew L. / Burnap, Pete / Sloan, Luke (2017): „Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation", in: *Sociology* 51: 1149-1168 https://journals.sagepub.com/doi/pdf/10.1177/0038038517708140 [letzter Zugriff 26. September 2019].

Intermediation der Forschungsinfrastruktur. Ein Rollenmodell für den Umgang mit einer komplexen Infrastrukturlandschaft

Wübbena, Thorsten
wuebbena@protonmail.com
Leibniz-Institut für Europäische Geschichte (IEG), Deutschland

Neumann, Katrin
Neumann@MaxWeberStiftung.de
Max Weber Stiftung (MWS), Deutschland

Cremer, Fabian
cremer@maxweberstiftung.de
Leibniz-Institut für Europäische Geschichte (IEG), Deutschland; Max Weber Stiftung (MWS), Deutschland

Kontext

Die Infrastrukturlandschaft für digital gestützte Forschung in den Geisteswissenschaften verändert sich seit einigen Jahren auf mehreren Ebenen: a) Breite: Der Anteil der Forschungsvorhaben, die Bedarf an digitalen Infrastrukturkomponenten aufweisen, wächst stetig (BMBF 2019); b) Diversität: Die Zahl der entwickelten Werkzeuge sowie deren Anwendungsgebiete nimmt zu (RfII 2016); c) Professionalisierung: Zusammenschlüsse von Infrastrukturanbietern und der Aufbau von skalierbaren Diensten erlauben zuverlässige und verteilte Nutzung, insbesondere mit den anstehenden Entwicklungen der Nationalen Forschungsdateninfrastruktur (RfII 2018). Mit breiterer Nutzung, vielfältiger Anwendung und verteilten Diensten steigt zum einen die Komplexität bei Konzeption, Organisation und Betrieb einer Forschungsinfrastruktur, zum anderen verschiebt sich die Aufgabe der forschungsorientierten Einrichtungen von der Administration eigener Systeme zum Management verteilter Infrastrukturkompetenten in Verbünden, Kooperationen und Konsortien. Dies drückt auch in der Entwicklung der sogenannten „Marketplaces" in übergreifenden Infrastrukturen aus (Kalman et al. 2019). Für das Management einer solch komplexen Infrastrukturlandschaft entwirft dieser Beitrag ein Rollenmodell für eine Vermittlungsposition.

Modell

Das Selbstverständnis der Digital Humanities basiert auf Zusammenarbeit, bei der Spezialisierung und Kooperationsfähigkeit gleichermaßen gefordert sind (The Digital Humanities Manifesto 2.0). Die Notwendigkeit der Zusammenarbeit resultiert aus der Komplexität der Vorhaben, die fachwissenschaftliche, informationswissenschaftliche und -technologische Ansätze verbinden, die Einzelpersonen nicht vereinen können. Als ein Modell für die Operationalisierung der Computing Humanities schlägt Jennifer Edmond den Digital Humanities Intermediary vor, der u.a. die Zusammenarbeit moderiert (Edmond 2005; Edmond 2016). Solch eine vermittelnde Figur für den Umgang mit Komplexität wird bereits im Informationswesen als Lösungsansatz gesehen und ist Teil vieler Wertschöpfungsketten, auch in der Wirtschaft (Rose 1999; Womack 2002).

Jenseits der Teammoderation und der Informationsversorgung bietet sich auch der Bereich der Forschungsinfrastruktur für eine Vermittlungsfigur an, die zwischen den Akteuren der Forschungseinrichtungen und der Infrastrukturanbieter operiert. Dabei ist, anders als in Konstellation aus Geisteswissenschaftler*in und Informatiker*in, keine direkte Moderation notwendig, sondern ein Makeln der Serviceangebote der Infrastruktureinrichtungen mit den Anforderungen der Forschenden. Die „Infrastrukturintermediation" komprimiert nicht nur Informationen oder übersetzt sie, sondern übernimmt die Verantwortung für Konzeption und Betrieb der Forschungsinfrastruktur.

Intention

Die Sichtung, Bewertung und Auswahl der zur Verfügung stehenden Angebote, die Prüfung der Zugangs- und Nutzungsbedingungen, die Implementierung und Organisation innerhalb des Forschungsvorhabens sowie die Abwicklung, Betreuung oder Überführung nach Projektende erfordern zeitliche Ressourcen und spezialisierte Kompetenzen. Die Verlagerung des Aufgabenbereiches auf ein/e Expert*in reduziert den Administrationsaufwand bei den Forschenden und Projektverantwortlichen. Die Definition der sich weiter ausdifferenzierenden Rollen in Forschungsvorhaben bilden die Grundlage für funktionierende kooperative Arbeitsformen (Beispiel Kunstgeschichte: Langmead et al. 2018). Gerade für

die Digital Humanists, die bisher sowohl in der Praxis (Reed 2014) wie in der Konzeption (Tabak 2017) als Vermittlungsfiguren fungieren, vergrößert sich so der Spielraum für die digitalen Methoden und Forschungsansätze. Die Kenntnis und Vermittlung vieler verschiedener Infrastrukturangebote erlaubt Forschungseinrichtungen ihrerseits auf ein größeres Portfolio an Angeboten zurückgreifen zu können und Forschungsvorhaben individualisierter unterstützen zu können ohne eigene Infrastruktur entwickeln oder anpassen zu müssen. Die Dienstanbieter und großen Infrastrukturverbünde erreichen durch die zusätzliche Vermittlungsstelle einen größeren und breiteren Nutzer*innenkreis. Die Intermediation kann gleichzeitig die Diversität in der Infrastrukturlandschaft unterstützen, die Komplexität der Nutzung reduzieren oder sogar Fehler ausgleichen (Beispiel e-Governance: Chaudhuri 2019). Idealerweise kann ein „Infrastrukturintermediär" so auch einen Interessensausgleich zwischen Forschung (Spezialisierung), Infrastruktur (Generalisierung) und Organisation (Skalierung) herbeiführen. Grundsätzlich werden auf diesem Weg Infrastrukturkomponenten nutzbar, die experimentell und leichtgewichtig sind und der Forschung die notwendigen Spielräume eröffnen (van Zundert 2012).

Verortung

Der Beitrag skizziert die Konzepte einer Infrastrukturmediation zweier außeruniversitärer Forschungseinrichtungen. Die Max Weber Stiftung hat mit der Digitalen Redaktion der Publikationsplattform *perspectivia.net* eine zentrale Einheit eingerichtet, deren Angebotsportfolio fast vollständig auf der Vermittlung institutionsfremder und international verteilter Dienste basiert (Cremer/Neumann 2019). Das Leibniz-Institut für Europäische Geschichte hat mit der Einrichtung eines Digital Humanities Lab sowohl institutionelle Ressourcen als auch mit der Kooperation im lokalen Netzwerk *mainzed* regionale Strukturen aufgebaut, die auch das Infrastrukturangebot der Einrichtung verändern und erweitern. Das Konzept der Infrastrukturintermediation ist jedoch auch auf Universitäten sowie diverse Institutionen übertragbar und als Rollenmodell nicht an Personen gebunden. Die entscheidende Voraussetzung ist jedoch eine neutrale Verortung ohne Eigeninteresse und mit einer Äquidistanz zu den Akteur*innen in Forschung und Infrastruktur.

Diskussion

Neben der Auseinandersetzung mit dem Modell der Intermediation soll die Diskussion im Rahmen der Posterpräsentation die Operationalisierung dieses Konzeptes vorantreiben. Dabei bietet die DHd-Tagung die einmalige Gelegenheit, alle im Modell benannten Akteur*innen zu Wort kommen zu lasen. Inwieweit ergeben sich so neue Spielräume für die Forschenden, wie profitieren Einrichtungen von einer zunehmenden Diversität der Infrastrukturlandschaft und wie sichern Infrastrukturen bei zunehmender Skalierung die Nähe zur Forschung? Welche Funktion und Bedeutung haben die benannten Aufgabenbereiche der Intermediation für die NFDI und ihre Konsortien?

Bibliographie

BMBF – Bundesministerium für Bildung und Forschung (2019): „Gesellschaft verstehen – Zukunft gestalten – BMBF". Bonn.

Chaudhuri, Bidisha (2019): „Paradoxes of Intermediation in Aadhaar: Human Making of a Digital Infrastructure". In: South Asia: Journal of South Asian Studies. 42 (3), S. 572–587, doi: https://doi.org/10.1080/00856401.2019.1598671.

Cremer, Fabian; Neumann, Katrin (2019): „Service intermediation as a concept for an institutional publishing department". In: ELPUB 2019 23rd edition of the International Conference on Electronic Publishing. Marseille, France, https://hal.archives-ouvertes.fr/hal-02141898.

Edmond, Jennifer (2016): „Collaboration and Infrastructure". In: Schreibman, Susan; Siemens, Raymond George; Unsworth, John (Hrsg.) A new companion to digital humanities. Chichester, West Sussex, UK: Wiley/Blackwell S. 54–67.

Edmond, Jennifer (2005): „The Role of the Professional Intermediary in Expanding the Humanities Computing Base". In: Literary and Linguistic Computing. 20 (3), S. 367–380, doi: https://doi.org/10.1093/llc/fqi036.

Kálmán, Tibor; Ďurčo, Matej; Fischer, Frank; Larrousse, Nicolas; Leone, Claudio; Mörth, Karlheinz; Thiel, Carsten (2019): „A landscape of data – working with digital resources within and beyond DARIAH". In: International Journal of Digital Humanities. 1 (1), S. 113–131, doi: https://doi.org/10.1007/s42803-019-00008-6.

Langmead, Alison; Berg-Fulton, Tracey; Lombardi, Thomas; Newbury, David; Nygren, Christopher (2018): „A Role-Based Model for Successful Collaboration in Digital Art History". In: International Journal for Digital Art History. 1 (3), doi: https://doi.org/10.11588/dah.2018.3.34297.

Reed, Ashley (2014): „Managing an Established Digital Humanities Project: Principles and Practices from the Twentieth Year of the William Blake Archive". In: Digital Humanities Quarterly. 8 (1), http://www.digitalhumanities.org/dhq/vol/8/1/000174/000174.html.

RfII – Rat für Informationsinfrastrukturen (2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Göttingen (RfII Empfehlungen), http://www.rfii.de/?wpdmdl=1998.

RfII – Rat für Informationsinfrastrukturen (2018): Rat für Informationsinfrastrukturen: In der Breite und forschungsnah: Handlungsfähige Konsortien. Dritter Diskussionsimpuls zur Ausgestaltung einer Nationalen Forschungsdateninfrastruktur (NFDI) für die Wissenschaft in Deutschland. Göttingen (RfII Empfehlungen), http://www.rfii.de/?p=3509.

Rose, Frank (1999): The Economics, Concept, and Design of Information Intermediaries: A Theoretic Approach. Physica-Verlag Heidelberg (Information Age Economy).

Tabak, Edin (2017): „A Hybrid Model for Managing DH Projects". In: Digital Humanities Quarterly. 11 (1), http://digitalhumanities.org/dhq/vol/11/1/000284/000284.html.

The Digital Humanities Manifesto 2.0 Authors (2009): „The Digital Humanities Manifesto 2.0", http://manifesto.humanities.ucla.edu/.

Van Zundert, Joris (2012): „If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities". In: Historical Social Research/Historische Sozialforschung. 37 (3), S. 165–186, https://doi.org/10.12759/hsr.37.2012.3.165-186.

Womack, Ryan (2002): „Information intermediaries and optimal information distribution“. In: Library & Information Science Research. 24 (2), S. 129–155, doi: https://doi.org/10.1016/S0740-8188(02)00109-3.

Keine Panik! Manifest für Softwareentwicklung in Studierendenprojekten

Eschweiler, Mark
mark.eschweiler@uni-koeln.de
Universität zu Köln, Deutschland

Evers, Anna-Maria
aevers1@smail.uni-koeln.de
Universität zu Köln, Deutschland

Kruhl, Dominik
dkruhl@smail.uni-koeln.de
Universität zu Köln, Deutschland

Reuhl, Elisabeth
ereuhl1@uni-koeln.de
Universität zu Köln, Deutschland

Türkoğlu, Enes
enes.tuerkoglu@uni-koeln.de
Universität zu Köln, Deutschland

Die in Forschungsprojekten entwickelten Anwendungen besitzen für die Digital Humanities einen hohen Stellenwert. Sie sind nicht nur Werkzeuge zum Lösen der jeweiligen Fragestellungen, sondern können auch als neue digitale Gestaltungsformen von Theoriebildung begriffen werden (Kleymann 2019). Genauso beinhaltet auch ein DH-Studium die Entwicklung zahlreicher Anwendungen, die dem Erlernen sowohl der programmiertechnischen Konzeption und Umsetzung wie auch der Organisation eines Softwareprojekts dienen.

Softwareentwicklung als integraler Bestandteil der Forschungsaktivitäten der Digital Humanities braucht ein geeignetes Vorgehen zur Organisation und Planung (Druskat et al. 2018). Agile Methoden erfahren nicht nur innerhalb der IT-Branche große Aufmerksamkeit, sondern auch im DH-Bereich (Heyer et al. 2019: 177). Eine Anwendung, die im Rahmen von Forschung oder Lehre entwickelt wird, besitzt jedoch andere Anforderungen als in der freien Wirtschaft. Die Umsetzung von Grundsätzen agiler Softwareentwicklung in Forschung und Lehre weist diverse Problemfelder auf. Scrum etwa wurde im nordamerikanischen Raum und für die freie Marktwirtschaft entwickelt, und orientiert sich mit seinen Prinzipien an der „lean production“, die wiederum sehr erfolgreich in japanischen Unternehmen etabliert wurde (Nonaka und Takeuchi 2008: 215-216). Daraus resultierende kulturelle und situative Barrieren und unterschiedliche Arbeitsrechte erschweren eine adäquate Adaption solch einer Methodologie in den europäischen Raum, und im universitären Kontext kommen weitere Problematiken hinzu.

Selbst zwischen Forschenden und Studierenden sind die Rahmenbedingungen sehr divergent. In studentischen Projekten sind Erkenntnisgewinn und Lernerfolg, neben einer fachgerechten technischen Umsetzung, entscheidend für die abschließende Bewertung und den Erfolg eines Projektes. Der Aufwand wird nicht monetär vergütet und auch ein Scheitern mit Erkenntnisgewinn kann ein Erfolg sein. Ein experimentelles Lehrveranstaltungsformat am Institut für Digital Humanities der Universität zu Köln, in dem Masterstudierende als Scrum Master eine Projektgruppe aus Bachelorstudierenden betreuen, stellte diese Schwierigkeiten der Umsetzung agiler Methodologien im Studium dar.

Welche Aspekte aus den agilen Methodologien sowie Manifesten übertragbar sind und wo neue Wege sinnvoller sind, wurde im Rahmen eines Seminars im Sommersemester 2019 von Masterstudierenden – unterstützt von Prof. Dr. Øyvind Eide – diskutiert und evaluiert. Orientierungspunkte boten die Erarbeitung von Lektüre zur agilen Entwicklung, die Reflexion eigener Vorerfahrungen, sowie Beobachtungen und Befragungen der Bachelorstudierenden-Projekte aus der übergreifenden Lehrveranstaltung. Um die Überlegungen festzuhalten und für nachfolgende Studierende nutzbar zu machen, fiel die Wahl, angelehnt an u.a. das „Manifesto for Agile Software Development“ (Beck et al. 2001), auf das Format eines Manifestes. Als „Manifest für Softwareentwicklung in Studierendenprojekten“ wurden schließlich die formulierten Rahmenbedingungen für die Planung und Umsetzung von Softwareentwicklung innerhalb von Gruppenprojekten im universitären Kontext vorgestellt.

Manifest für Softwareentwicklung in Studierendenprojekten

Softwareprodukt und **Lernprozess** sollten sowohl bei der Entwicklung als auch bei der Dokumentation die gleiche Wichtigkeit erfahren. Der Erfolg studentischer Softwareprojekte ist im Gegensatz zu kommerziellen Softwareprojekten nicht nur von einem funktionierenden Endprodukt abhängig, sondern soll gleichermaßen an in ihrem Umfang angemessenen und **nachhaltigen Lernergebnissen** gemessen werden. Ziel studentischer Softwareprojekte ist demnach ein nachweisbares Ergebnis, welches sich sowohl im Softwareprodukt als auch im dokumentierten Lernprozess zeigt.

Die Organisation studentischer Softwareprojekte erfolgt **selbstorganisiert** durch die Projektgruppe. Die Aufgabenverteilung soll dabei als ein Prozess der **Kompetenzverhandlung** begriffen werden. Verhandelt werden sollen die realistische (Selbst-)Einschätzung der vorhandenen Kenntnisse der teilnehmenden Studierenden, sowie jene für das Projekt relevante Kompetenzen, deren Aneignung noch angestrebt wird.

Eine zutreffende Einschätzung des Arbeitspensums und adäquate Verteilung korrespondierender Aufgaben werden durch **modulare Mikroaufgaben** gewährleistet. Die Definition spezifischer **Meilensteine** und ihr anschließendes Erreichen im Arbeits- und Lernprozess soll durch Modularität vereinfacht werden. Eine modulare Arbeitsweise erlaubt zudem, die Kompetenzverhandlung einzelner Gruppenmitglieder **flexibel** zu gestalten und ein kontinuierliches Arbeitstempo zu gewährleisten. Voraussetzung für die Definition solcher Mi-

kroaufgaben ist es, Abhängigkeiten zu anderen Aufgaben weitestgehend zu vermeiden.

Agilität stellt einen essentiellen Bestandteil des Projektes dar. Sowohl die Organisation als auch die technische Umsetzung müssen es zulassen, dynamisch auf neue Erkenntnisse oder Hindernisse zu reagieren, ohne dabei das Gesamtziel aus den Augen zu verlieren. Dabei gilt: je modularer die Entwicklung abläuft, desto flexibler kann auf Veränderungen reagiert werden.

Agilität ist insbesondere auf Kompetenzen in der **Gruppenkommunikation** angewiesen, welche sich um Transparenz und Zuverlässigkeit bemühen sollte. Hierzu bedarf es auch einer geeigneten gemeinsamen **Kommunikationsplattform** , welche den Austausch zur Organisation und Umsetzung des Projektes in Form **regelmäßiger Treffen** garantiert. Hier soll jedes Mitglied umstandslos den **Status Quo** des Projektes einsehen können. Außerdem muss hier auf die Möglichkeit geachtet werden, zwischenmenschliche Aspekte effizient behandeln und arrangieren zu können. Falls gewünscht, kann über jene Plattform auch dem Dozierenden ein Zugang zwecks Beurteilung erteilt werden.

Ein **lauffähiger und vorzeigbarer Prototyp** sollte möglichst schnell erstellt und dann **kontinuierlich weiterentwickelt** werden. Dabei ist im Sinne der Agilität darauf zu achten, durch eine modulare Arbeitsweise die Aufwandskurven möglichst flach zu halten. Aktuelle Entwicklungen können stets funktionsfähig in das Produkt aufgenommen werden.

Bibliographie

Beck, Kent / Grenning, James / Martin, Robert C. / Beedle, Mike / Highsmith, Jim / Mellor, Steve / van Bennekum, Arie / Hunt, Andrew / Schwaber, Ken / Cockburn, Alistair / Jeffries, Ron / Sutherland, Jeff / Cunningham, Ward / Kern, Jon / Thomas, Dave / Fowler, Martin / Marick, Brian (2001): „Manifesto for Agile Software Development". Agile Alliance. http://agilemanifesto.org/ [letzter Zugriff 27. September 2019].

Heyer, Gerhard / Kahmann, Christian / Kantner, Cathleen (2019): „Generic tools and individual research needs in the Digital Humanities – Can agile development help?", in: Draude, C. / Lange, M. / Sick, B. (eds.): *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V. 175-180.

Kleymann, Rabea (2019): „Prototypen als Proto-Theorie? – Plädoyer einer digitalen Theoriebildung", in: Sahle, Patrick (ed.): *DHd 2019 Digital Humanities: multimedial & multimodal.* Frankfurt am Main 197-200. http://doi.org/10.5281/zenodo.2596095 [letzter Zugriff 27. September 2019].

Nonaka, Ikujiro / Takeuchi, Hirotaka (2008): „A Theory of the Firm's Knowledge-Creation Dynamics", in: Chandler, Alfred / Hagstrom, Peter / Sölvell, Örjan: *The Dynamic Firm. The Role of Technology, Strategy, Organization, and Regions*. Oxford: University Press 214-241.

Schwaber, Ken / Sutherland, Jeff (2017): „Der Scrum Guide™. Der gültige Leitfaden für Scrum: Die Spielregeln." https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-German.pdf [letzter Zugriff 27. September 2019].

Druskat, Stephan / Czmiel, Alexander / Schrade, Torsten (2018): „Research Software Engineering und Digital Humanities. Reflexion, Kartierung, Organisation", DHd 2018, Köln, 27.02.2018. https://dh-rse.github.io/dhd-workshop-2018-presentation/ [letzter Zugriff 27. September 2019].

Kein Spiel(raum): Rechtliche und ethische Rahmenbedingungen geisteswissenschaftlicher Forschung

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

Hannesschläger, Vanessa
Vanessa.Hannesschlaeger@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Internationale Dachorganisationen, europäische Infrastrukturprojekte, regionale Verbände und nationale Initiativen in den Digitalen Geisteswissenschaften fördern (und fordern) den offenen Zugang zu digitalen Methoden, Daten und Werkzeugen, offene und nachnutzbare Formen der Wissenschaftskommunikation und Dissemination sowie einen verantwortungsvollen und integren Umgang innerhalb der wissenschaftlichen Community sowie mit jenen Personen, die als BeiträgerInnen oder sogar Gegenstand unserer Forschung involviert sind (McKee/Porter 2009; Markham/Buchanan 2012).

Von besonderem Interesse für Kulturerbeeinrichtungen und GeisteswissenschaftlerInnen sind Fragen des Urheberrechts sowie der Bereitstellung von und des Zugangs zu digitalisiertem Quellenmaterial (Darling 2012; Galina 2017). In der Europäischen Union besteht ein erkennbarer politischer Impuls, den freien und öffentlichen Zugang zu kulturellem Erbe und zu Forschungsdaten, die an öffentlich finanzierten Einrichtungen gehostet werden, zu erleichtern und entsprechende Digitalisierungsvorhaben zu fördern (Europäisches Parlament 2013 et al). Der Mangel an rechtlicher Harmonisierung und die vielfältigen und oft unklaren nationalen Rechtsvorschriften über die Nutzung und Bereitstellung von Ressourcen durch öffentliche Kulturerbe-, Forschungs- und Bildungseinrichtungen – auch die jüngste europäische Urheberrechtsrichtlinie (Europäisches Parlament 2019), die eine Reihe von zentralen Themen der digitalen Geisteswissenschaften wie Text- und Datamining und die grenzübergreifende Nachnutzung von Ressourcen thematisiert, wird sich in den nationalen Umsetzungen sehr unterschiedlich niederschlagen – lösen jedoch Unsicherheit aus oder bedingen restriktive Regelungen an betroffenen Institutionen. Auch die EU-Datenschutzgrundverordnung (DSGVO) (Europäisches Parlament 2016) sorgte für große Unsicherheit und führte mangels konkreter, praxisbezogener Information und Beratung der Forschungsgemeinschaft zu überstürzten Aktivitäten, die oft über die tatsächlichen ge-

setzlichen Anforderungen hinausgingen und die Forschung erheblich erschwerten.

Die großen Themenbereiche des Urheberrechts (und insbesondere der Lizenzierung) sowie des Datenschutzes (vor allem der Zulässigkeit von Datenverarbeitungen in Forschungskontexten) sind zentral für jede wissenschaftliche Tätigkeit. Um auf die vorherrschende Unsicherheit zu den rechtlichen und ethischen Rahmenbedingungen geisteswissenschaftlicher Forschung zu reagieren, wurde eine Reihe von Arbeitsgruppen und Interessensgemeinschaften ins Leben gerufen, die diese Fragen zu beantworten versuchen und die entsprechenden Kenntnisse in Form von Best Practice-Beispielen, Leitfäden und Workshops zu vermitteln – umso schwieriger, als nationale Gesetzgebungen sich mitunter stark unterscheiden, Fragen der Haftung und Legalität aber keine Spielräume gestatten.

Über diesen rechtlichen Rahmen hinaus sind Fragen der ethischen Forschungspraxis und des wissenschaftlichen Verhaltens für die Geistes- und Sozialwissenschaften von zentraler Bedeutung, insbesondere in einem weitgehend digitalen, internetbasierten Forschungskontext (McKee/Porter 2009; Markham/Buchanan 2012):

> "Different ethical issues become salient as the researcher develops research questions, seeks and gains access to individuals and/or information, manages and protects personally identifiable information, selects analytical tools, and represents the data through dissemination, in published reports, conference presentations, or other venues." (Markham/Buchanan 2012)

Gerade in Bezug auf digitale Formen der Wissenschaftskommunikation und des Wissenstransfers kommt dem verantwortungsbewussten und wertschätzenden Umgang nicht nur mit Daten, sondern auch den unterschiedlichen AkteurInnen im Forschungsprozess zunehmend Bedeutung zu.

Dieses Poster soll den TeilnehmerInnen der DHd2020 einige der Initiativen und Arbeitsgruppen vorstellen, die sich diesem Themenfeld im deutschsprachigen Raum widmen, zum Beispiel die DARIAH-EU Arbeitsgruppe "Ethics and Legality in Digital Arts and Humanities" (ELDAH), das CLARIN-ERIC "Legal and Ethical Issues Committee" (CLIC), die DHd Arbeitsgruppe "Digitales Publizieren" oder auch das "Open Science Network Austria" (OANA). Nicht zuletzt aufgrund personeller Überschneidungen herrscht rege Kommunikation und Kooperation zwischen diesen Gruppen: Ressourcen werden gebündelt, Ergebnisse und Aktivitäten übergreifend angeboten und Synergien genutzt.

Die Präsentation des Posters bei der DHd2020 gibt uns die Möglichkeit, die TeilnehmerInnen über bereits bestehende und in Entwicklung befindliche Angebote und Werkzeuge – wie Schulungsunterlagen der ELDAH Arbeitsgruppe auf https://eldah.hypotheses.org/ oder den in Entwicklung befindlichen *Consent Wizard*, einen Generator für DS-GVO-konforme Einwilligungserklärungen – zu informieren und Fragen und Anliegen der TeilnehmerInnen zu rechtlichen und ethischen Herausforderungen ihrer Arbeit (insbesondere zu Urheberrecht, Lizenzierung und der Verarbeitung personenbezogener Daten im wissenschaftlichen Kontext) direkt zu beantworten sowie die Bedürfnisse der Fachgemeinschaft in diesen Bereichen zu erheben und in weiterer Folge – sei es durch die Entwicklung neuer oder die Anpassung bestehender Angebote – praxisorientiert darauf zu reagieren.

Bibliographie

Darling, Kate (2012): "Contracting About the Future: Copyright and New Media", in: Northwestern Journal of Technology and Intellectual Property 10/7, 485–530. http://scholarlycommons.law.northwestern.edu/njtip/vol10/iss7/3

Europäisches Parlament (2013): Richtlinie 2013/37/EU des Europäischen Parlaments und des Rates vom 26. Juni 2013 über die Weiterverwendung von Informationen des öffentlichen Sektors. https://eur-lex.europa.eu/eli/dir/2013/37/oj

Europäisches Parlament (2016): Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679

Europäisches Parlament (2019): Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt. http://data.europa.eu/eli/dir/2019/790/oj

Galina, Isabel et al. (2017): "Copyright and Creator Rights in DH Projects: A Checklist." https://hcommons.org/deposits/item/hc:15109/

Kamocki, Pawel / Ketzan, Erik / Wildgans, Julia (2018): "Language Resources and Research Under the General Data Protection Regulation". https://www.clarin.eu/sites/default/files/CLIC_White_Paper_3.pdf

Klimpel, Paul / Weitzmann, John H. (2015): „Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften.", DARIAH-DE Working papers Nr. 12. https://irights.info/wp-content/uploads/2015/08/Forschen-in-der-digitalen-Welt-Juristische-Handreichung-Geisteswissenschaften-dw-p-2015-12.pdf

Klimpel Paul (2013): "Free Knowledge Thanks to Creative Commons Licenses. Why a Non-commercial Clause often won't Serve Your Needs." https://www.wikimedia.de/w/images.homepage/1/15/CC-NC_Leitfaden_2013_engl.pdf

Markham, Annette / Buchanan, Elisabeth (2012): "Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee." http://aoir.org/reports/ethics2.pdf

McKee, Heidi / Porter, James E. (2009): „The Ethics of Internet Research: A Rhetorical, Case-based Process."

Zimmermann, Claudia (2018): „Leitfaden für die Erstellung von Open Educational Resources. Informationen und praktische Übungen für Hochschullehrende." https://www.openeducation.at/fileadmin/user_upload/p_oea/OEA-Leitfaden_online_Aufl2.pdf

Keter Shem Ṭov - Prozessualisierung eines Editionsprojekts mit 100 Textzeugen

Molitor, Paul
paul.molitor@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Necker, Gerold
gerold.necker@judaistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Rebiger, Bill
bill.rebiger@judaistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Ritter, Jörg
joerg.ritter@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Der folgende Beitrag stellt das Forschungsvorhaben und erste Ergebnisse des DFG geförderten Projekts "Synoptische Edition des kabbalistischen Traktats *Keter Shem Ṭov* mit englischer Übersetzung, Stellenkommentar und rezeptionsgeschichtlichen Studien" vor[1].

Der im 13. Jahrhundert auf Hebräisch verfasste Traktat *Keter Shem Ṭov* (“Krone des guten Namens”) ist einer der wichtigsten Einführungstexte in die esoterischen Lehren der jüdischen Kabbala. Er wird häufig Abraham ben Axelrad von Köln zugeschrieben, der möglicherweise ein Schüler von El‘azar von Worms (ca. 1176–1238) und Ezra ben Salomo von Gerona (um 1240) war. Der Traktat verbindet die in der spanischen Kabbala klassisch ausgeprägte Symbolik der zehn Sefirot bzw. Manifestationen der Gottheit mit solchen Deutungen des Tetragrammatons, d.h. des vierbuchstabigen („guten“) Gottesnamens, wie sie aus der Literatur der „Deutschen Frommen“, den *Ḥaside Ashkenaz*, bekannt sind. *Keter Shem Ṭov* ist der älteste bekannte Text, der diese beiden mystischen Traditionen vereint. Er wird in verschiedenen Versionen in etwa 100 Handschriften bezeugt, die sich voneinander deutlich in Umfang und Struktur unterscheiden. Die Herkunft dieser Handschriften umfasst ashkenazische, sefardische, italienische, byzantinische und orientalische Provenienzen. Die ältesten handschriftlichen Textzeugen stammen bereits aus der zweiten Hälfte des 13. Jahrhunderts und sind somit nur wenige Jahre nach der Komposition dieses Traktats entstanden. Der Traktat wurde nicht nur in Kreisen der jüdischen Kabbala, sondern auch von christlichen Kabbalisten rezipiert.

Das Ziel des Projekts, das für drei Jahre von der Deutschen Forschungsgemeinschaft (DFG) bewilligt wurde, ist eine kritische Edition der verschiedenen Versionen dieses Traktats in Form einer Spaltensynopse, die sowohl in einer Druckausgabe als auch in einer digitalen und interaktiven Edition online verfügbar gemacht werden soll. Eine englische Übersetzung des Textes, ein detaillierter Stellenkommentar und Studien zur Geschichte seiner Rezeption werden die Edition ergänzen.

Für die Kollationierung und Analyse der Textvarianten kommt das webbasierte Werkzeug LERA[2] zum Einsatz, das im Rahmen des Projekts weiterentwickelt wird. Die ersten transkribierten Textfassungen – in UTF-8 codierte Textdateien – konnten bereits mit dem Werkzeug verglichen werden. Abbildung 1 vermittelt einen ersten Eindruck der Oberfläche von LERA und zeigt beispielhaft den Vergleich von vier verschiedenen Textfassungen.

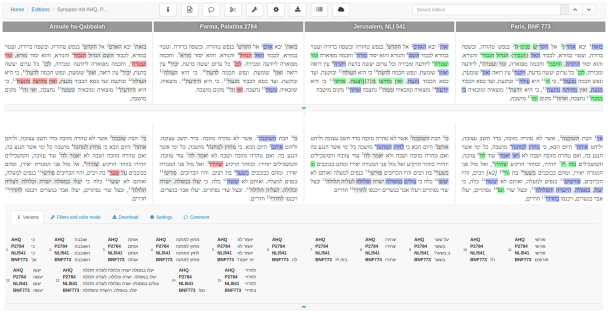


Abbildung 1: Auszug einer mit LERA erstellten Synopse für vier Textfassungen von *Keter Shem Ṭov*. Im Ausschnitt zu sehen sind je zwei alignierte Absätze der vier Fassungen mit den farblich hervorgehobenen Textvarianten. Unter dem zweiten Absatz ist zudem der Variantenapparat eingeblendet.

Ein nächster Arbeitsschritt besteht darin, an das Werk und das Hebräische angepasste Vergleichsfilter zu entwickeln, um spezifische Textvarianten auf Wunsch ausblenden zu können. Beispielsweise enthält *Keter Shem Ṭov* verschiedenartige Bezeichnungen für "Gott", die in ausgeschriebenen oder abgekürzten Formen auftreten und oftmals mehrdeutig sind, sodass ein einfacher Wörterbuchansatz fehlschlägt. Wesentlicher Bestandteil des Textes sind zudem zitierte Bibelverse, die neben ihrer orthografischen Varianz auch in unterschiedlicher Länge wiedergegeben werden. Viele Textfassungen enthalten dabei nur die ersten Wörter; das Wissen um den vollständigen Vers setzt der Verfasser voraus. So entstehen Textvarianten, die vom System gesondert hervorgehoben oder verborgen werden sollen, was auf Basis entsprechender Auszeichnungen realisiert wird. Eine teilautomatisierte Erkennung, die den manuellen Aufwand dafür reduziert, ist derzeit in Entwicklung.

Ein wesentlicher Aspekt des Projektes ist der Umgang mit sehr vielen Textfassungen, der mit voranschreitender Transkribierung weiterer Manuskripte stetig an Bedeutung gewinnt. LERA wurde ursprünglich für die Verarbeitung und Darstellung umfangreicher, aber weniger Textfassungen konzipiert (Bremer et al. 2015), obgleich seitdem für andere Editionsvorhaben angepasst, siehe bspw. (Gründler und Pöckelmann 2018, Roeder 2019). Um den Anforderungen von *Keter Shem Ṭov* gerecht zu werden, sind verschiedene Erweiterungen in Arbeit. So wird die bisherige synoptische Darstellungsform mit je einer Spalte pro Textfassung um eine zeilenweise bzw. partitursynoptische Darstellung ergänzt. Über die Nut-

zeroberfläche soll es möglich sein, einzelne Textfassungen dynamisch ein- und auszublenden, was mit Hilfe der bereits integrierten Übersichtsleiste realisiert werden kann. Es ist eine dynamische Auswahl der Leithandschrift geplant, was neben der Anpassung der Oberfläche auch die Modifizierung des Vergleichsalgorithmus notwendig macht, da dieser die Textfassungen derzeit stets gleichrangig betrachtet. Es wird angestrebt diese interaktive Eingriffsmöglichkeit durch geschickte Vorberechnung möglichst effizient zu gestalten, damit für den Nutzer des Systems keine Wartezeiten entstehen. Ferner ist das Zusammenfassen mehrerer ähnlicher Handschriften in Gruppen angedacht. Der synoptische Vergleich findet dann auf Basis dieser Gruppen statt und wird größere Textunterschiede aufzeigen, während ein Apparat auch die kleinsten Änderungen innerhalb einer Gruppe aufschlüsselt.

Anmerkungen

Diese Arbeiten werden durch die Deutsche Forschungsgemeinschaft (DFG) [Projektnummer 414786977] im Rahmen des Projekts „Synoptische Edition des kabbalistischen Traktats *Keter Shem Ṭov* mit englischer Übersetzung, Stellenkommentar und rezeptionsgeschichtlichen Studien“ unter Leitung von apl. Prof. Dr. Gerold Neckerund Prof. Dr. Paul Molitor gefördert.

Fußnoten

1. Homepage des Projekts:https://kabbalaheditions.org
2. LERA – Locate, Explore, Retrace and Apprehend complex text variants, siehe https://lera.uzi.uni-halle.de

Bibliographie

Abrams, Daniel (2013): *"Kabbalistic Manuscripts and Textual Theory: Methodologies of Textual Scholarship and Editorial Practice in the Study of Jewish Mysticism"*, Jerusalem / Los Angeles: Cherub Press.

Bremer, Thomas / Molitor, Paul / Pöckelmann, Marcus / Ritter, Jörg / Schütz, Susanne (2015): "Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen – Das Fallbeispiel der Histoire philosphique des deux Indes von Guillaume Thomas Raynal" in: *Editio, Internationales Jahrbuch für Editionswissenschaften*, Hrsg. R. v. Nutt-Kofoth, B. Plachta und W. Woesler, Band 29, Heft 1, S. 29–51, de Gruyter.

Gründler, Beatrice / Pöckelmann, Marcus (2018): "Adjusting LERA For The Comparison Of Arabic Manuscripts Of Kalīla wa-Dimna", in: *Book of Abstracts of the 29. International Annual Conference of Digital Humanities, DH2018,* Mexico City, pp. 467–468.

Idel, Moshe (2007): "Ashkenazi Esotericism and Kabbalah in Barcelona", in: *Hispania Judaica Bulletin* 5, S. 69–113.

Jellinek, Adolph (1988): "Kleine Schriften zur Geschichte der Kabbala", Hildesheim: Olms, S. 29–48.

Oron, Michal (2013/14): "Ha-Sifrut ha-parshanit le-ʿeser Sefirot", in: Yoram Ben-Naʻeh et. al. (Hg.), ʾAsupa le-Yosef, FS Joseph Hacker, Jerusalem: Shazar, S. 212–229 (Hebr.).

Roeder , Torsten (2019): "Genesis and Variance: From Letter to Literature", in: *Book of Abstracts of the 19th Annual Conference and Members' Meeting of the Text Encoding Initiative Consortium(TEI)*, Graz, pp. 92–93.

Scholem, Gershom (1933): "Index to the Commentaries on the Ten Sefirot", Qiryat Sefer 10, S. 498–515 (Hebr.).

Korpusbereinigung für größere Textmengen. Eine (kurze) Problematisierung und ein Lösungsansatz für Duplikate

Adelmann, Benedikt
adelmann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Gius, Evelyn
gius@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Korpuserstellung in der Literaturwissenschaft

Eine neue Praxis

Die Zusammenstellung von Primärtexten als Forschungsgrundlage ist in der Literaturwissenschaft Alltagsgeschäft. Die beiden Standardfälle der (nicht-digitalen) Korpusanalyse gelten in Bezug auf die Korpuszusammensetzung als wenig problematisch:

1. Wenn die Forschungsfrage eine spezifische Textgrundlage erfordert, steht das Korpus von vornherein fest (z. B. alle Romane von Thomas Mann für eine Studie zur Repräsentation von Krankheit in Manns Romanwerk).
2. In Korpora kanonischer Texte werden weitergehende Fragen untersucht (z. B. ausgewählte Texte aus dem Realismus zur Untersuchung des Ausdrucks von Idylle im Realismus).

Mit der Verfügbarkeit digitaler Texte wurde die Praxis der Korpuserstellung jedoch erweitert und es wurde offensichtlich, dass sie nicht ohne weiteres in die bestehende literaturwissenschaftliche Disziplinarmatrix (Kuhn 1970) integriert werden kann. Viele der digital verfügbaren Texte sind nämlich weder kanonisch noch repräsentativ, die Qualität einzelner Texte ist aus philologischer Sicht oft fragwürdig und ein Korpus enthält trotz der Vielzahl der verfügbaren digitalen Texte selten die gesamte Population relevanter Texte, sondern nur eine Teilmenge.[1]

Über die philologisch einwandfreie Auswahl von Texten hinaus birgt die Kuration von Korpora weitere Herausforderungen.[2] Die vermutlich größte ist, sich einen Überblick über ein Korpus zu verschaffen, das mehr Texte enthält, als man lesen kann. Bei einer nicht von einem Einzelnen erfassbaren Textmenge kann selbst eine scheinbar einfache Aufgabe wie die Erkennung von Duplikaten unlösbare Probleme bereiten.

Ein exemplarisches Korpus

Der vorgestellte Ansatz wurde für ein Korpus entwickelt, das in einem Forschungsprojekt zur geschlechtsspezifischen Darstellung von Krankheit in literarischen Texten im Rahmen der Forschungskooperation hermA[3] erstellt wurde. Ausgangspunkt war das Kolimo-Korpus, das mehr als 42.000 literarische und nicht-literarische deutsche Texte vor allem von 1880 bis 1930 aus drei großen Repositorien deutscher Texte enthält: dem Deutschen Textarchiv[4], dem TextGrid Repository[5] und Projekt Gutenberg-DE[6] (vgl. Herrmann & Lauer 2017). Wir haben alle Prosatexte von 1870 bis 1920 ausgewählt, die ursprünglich auf Deutsch verfasst waren (vgl. Gius et al., 2019).

Im daraus resultierenden Korpus von mehr als 2.500 Texten mussten Artefakte behandelt werden, die durch unterschiedliche Digitalisierungsstrategien verursacht wurden. Nicht nur die Erhaltung von Sonderzeichen wie das recht häufige lange s (ſ) zwischen oder innerhalb der Repositorien war inkonsistent, sondern auch die Verwendung von Bindestrichen (Wortverbindung, Worttrennung an Zeilenumbruch, andere Bindestriche, Gedankenstriche) oder die Kodierung von Zeilenumbrüchen und Absätzen. Diese Probleme konnten mit einer relativ einfachen Heuristik angegangen werden.

Die Identifizierung von Duplikaten

Ein schwierigeres Problem ist die Frage der Duplikate: Insbesondere bei der Zusammenstellung eines Korpus aus verschiedenen Quellen kann es vorkommen, dass der gleiche Text mehrfach vorhanden ist. In der Regel ist es nicht erwünscht, mehr als eine Instanz desselben Textes im Korpus zu haben, da die Überrepräsentation einzelner Werke bei statistischen Analysen zu verzerrten Ergebnissen führen kann. Daher sollte die Identifizierung von Duplikaten ein wesentlicher Bestandteil der Korpuserstellung sein.

Dabei gibt es zwei Probleme: Erstens wächst die Anzahl der ungeordneten Werkpaare, die alle potenziell Duplikate sein könnten, quadratisch mit der Anzahl der Werke. In unserem Korpus mit gut 2.500 Texten müssten deshalb 3,1 Millionen Werkpaare überprüft werden. Zweitens ist auch für jedes einzelne Textpaar die Feststellung, ob es sich um Duplikate handelt, aufgrund von Metadaten- und Textinkonsistenzen eine nicht-triviale Aufgabe. Ansätze zur *text reuse detection* (z. B. Bär et al. 2012) blenden den kombinatorischen Aspekt oft aus.

Wir haben mit zwei Methoden zur automatischen Duplikatidentifizierung experimentiert. Beide sind Heuristiken für die Suche nach Werkpaaren, die Duplikate sind; sie lösen aber nicht das daran anschließende Problem, zu entscheiden, welche von mehreren Instanzen tatsächlich in das Korpus aufgenommen werden sollen.

Für die Evaluation wurden alle Duplikatkandidaten, die von mindestens einer der beiden Methoden gefunden wurden, manuell auf ihre Richtigkeit überprüft. Wir berichten über den Prozentsatz der automatisch als Duplikate identifizierten Paare, die tatsächlich Duplikate sind (Precision), und den Prozentsatz der tatsächlichen Duplikate, die automatisch identifiziert werden (Recall). Allerdings lässt sich der Recall nur exakt bestimmen, wenn alle 3,1 Millionen Werkpaare manuell untersucht werden. Als Annäherung verwenden wir daher stattdessen den Prozentsatz der bei der manuellen Prüfung identifizierten tatsächlichen Duplikate (Gesamtzahl: 355), die ebenfalls automatisch gefunden werden.[7]

Die erste Methode basiert auf Metadaten und ist deshalb schnell genug, um alle ungeordneten Werkpaare im Korpus zu testen. Für jedes Werkpaar werden Autor*innen- und Titelinformationen verglichen. Was die Autor*innen-Informationen betrifft, so wird die sogenannte Edit-Distanz (Levenshtein 1965) der vollständigen Namen der Autor*innen berechnet; die Edit-Distanz ist die kleinste Anzahl von Zeicheneinfügungen, Zeichenlöschungen und Zeichenersetzungen („Edits“), mit der der erste Autor*innen-Name in den zweiten umgewandelt werden kann.

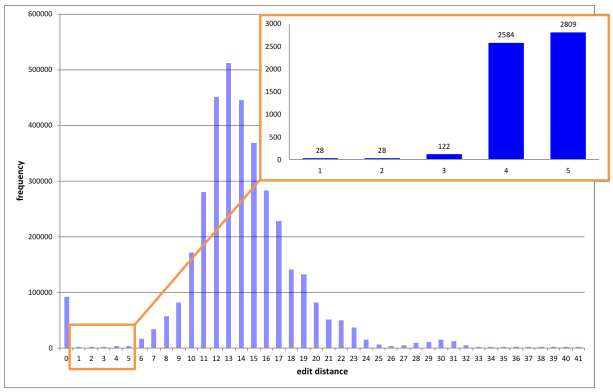


Abbildung 1: Edit-Distanzen Autor*innen-Namen (Schwellenwert bei 2)

Für die Titelinformationen haben wir das Maß leicht modifiziert: Wir verwenden die kleinste Anzahl von Edits, die einen der Titel in einen *Teil* (d. h. Teilzeichenkette) des anderen verwandeln können, mit der zusätzlichen Einschränkung, dass ganze Wörter vollständig abgeglichen werden müssen, um unangemessen kurze Entfernungswerte zu vermeiden (sonst ließe sich beispielsweise „Tot“ nach nur zwei Edits als Teilzeichenkette von „Das Jüngste Gericht“ finden, hätte also Abstand 2).[8] Durch dieses modifizierte Maß sollen hohe Titeldistanzen vermieden werden, wenn Untertitel in nur einem der beiden Titel enthalten sind.[9]

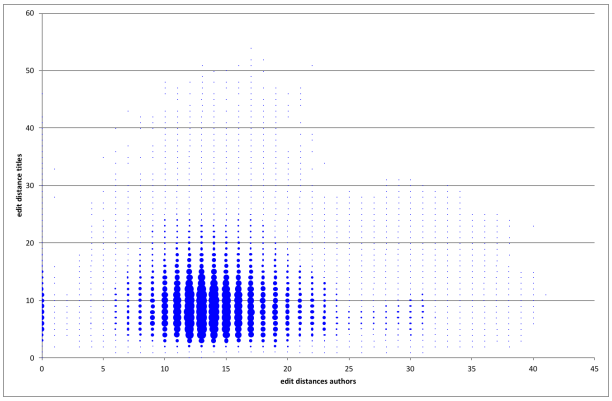


Abbildung 2: Kombinierte Edit-Distanzen für Autor*innen und Titel

Zwei Texte werden als Duplikate betrachtet, wenn der Abstand sowohl beim Autor*innen-Namen als auch beim Titel höchstens so hoch wie der Schwellenwert ist. Bei einem Schwellenwert von 2 wurden 672 Duplikatpaare mit einer Precision von 51,9 % und einem Recall von 98,3 % identifiziert. Unter den sechs nicht identifizierten Duplikaten finden sich beispielsweise zwei Werke von Karl May mit abweichender Behandlung von Sammelband-/Einzelwerktitel („Ardistan und Dschinnistan. 1. Band“ vs. „Der Mir von Dschinnistan“; „Satan und Ischariot III“ vs. „Im Todesthale“) und ein Fall, in dem bei einem der beiden Werke fälschlich der Name des Autors als Titel eingetragen war (Ferdinand von Saar, „Vae victis!“).

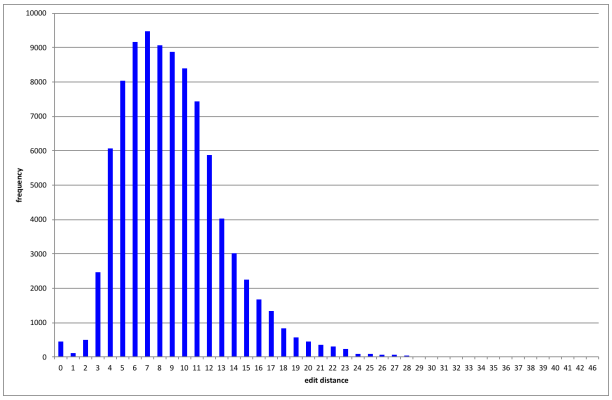


Abbildung 3: Edit-Distanzen in Titeln, wo Autor*innen-Namen max. Edit-Distanz 2 aufweisen

Das zweite Verfahren berechnet die Edit-Distanzen von Volltexten. Da dies eine zeitaufwändige Operation ist, beschränken wir uns auf den Vergleich von Texten, bei denen der Autor*innen-Name eine Edit-Distanz von maximal zwei hat. Manuelle Überprüfungen zeigten, dass diese Schwelle alle Rechtschreibfehler und Varianten in unseren Daten erfasst, verschiedene Autor*innen mit ähnlichen Namen jedoch ausnimmt. Für Volltexte verwenden wir wieder Teilzeichenketten-Edit-Distanzen, da ein Text mehr oder weniger vollständig in einem anderen enthalten sein kann (z. B. bei Anthologien). Zugunsten der Rechenzeiten verwenden wir wortbezogene Distanzen mit Insertionskosten gleich Deletionskosten gleich Substitutionskosten gleich eins.[10]

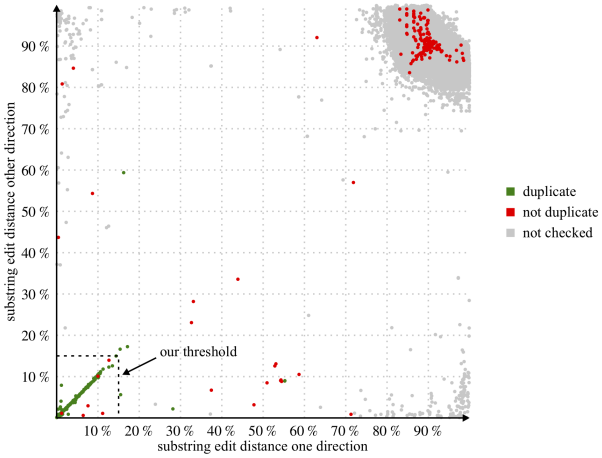


Abbildung 4: Edit-Distanzen der Volltexte

Zwei Texte gelten als Duplikate, wenn die Teilzeichenketten-Edit-Distanzen für beide Richtungen (1 als Teilzeichenkette von 2 oder umgekehrt), dividiert durch die Länge des jeweils als Teilzeichenkette einzubettenden Texts, unter 15 % liegt. Auf diese Weise können wir 307 Duplikate mit einer Precision von 98 % und einem Recall von 84,8 % bestimmen.

678 Paare wurden so durch mindestens ein Verfahren als Duplikate identifiziert (Precision: 52,4 %),[11] 301 durch beide (Precision: 98 %, Recall: 83,1 %). Letzteres ist bedeutsam für praktische Anwendungen, bei denen man mit Methode 1 kostengünstig Werkpaare auswählen möchte, die anschließend mit der teuren Methode 2 getestet werden sollen.

Fazit

Die Digitalisierung hat die Arbeit mit literarischen Korpora erheblich gefördert. Die schiere Menge an Texten in einem Korpus muss sowohl technisch als auch konzeptionell unterstützt werden. Für die Erreichung dieser Ziele ist es umso wichtiger, Qualitätskriterien für die Zusammenstellung von Korpora im Hinblick auf die verfügbaren Daten, d. h. für Texte aus heterogenen Quellen unterschiedlicher Qualität, zu entwickeln und umzusetzen. Zusätzlich zu diesen noch zu entwickelnden Kriterien für die wissenschaftliche Qualitätssicherung können einige pragmatische Entscheidungen die Qualität eines Korpus und seiner Texte in Fällen mit geringer Daten- und insbesondere Metadatenqualität erheblich verbessern.

Der vorgestellte Ansatz zum Umgang mit Duplikaten kann zusammen mit den genannten Vorverarbeitungsschritten ein wichtiger erster Schritt in diesem Prozess sein.

Fußnoten

1. Für verschiedene Typen von Datensammlungen oder Korpora vgl. Schöch (2017).
2. Einige der Probleme und Lösungsmöglichkeiten wurden in Gius et al. (2019) vorgestellt, für einen breiteren Überblick vgl. Lauer & Herrmann (2018).
3. Vgl. https://www.herma.uni-hamburg.de und Gaidys et al. (2017) [alle Links in diesem Beitrag wurden am 15.09.2019 das letzte Mal abgerufen].

4. Vgl. http://www.deutschestextarchiv.de/
5. Vgl. https://textgridrep.org/
6. Vgl. https://www.gutenberg.org/wiki/DE_Hauptseite (Achtung: seit 2018-02 nicht mehr über deutsche IP-Adressen abrufbar).
7. Unter den manuell geprüften Werkpaaren finden sich keine Duplikate innerhalb DTA, zwei innerhalb TextGrid und 46 innerhalb Gutenberg sowie 23 zwischen DTA und TextGrid, 27 zwischen DTA und Gutenberg und 257 zwischen TextGrid und Gutenberg.
8. Formal: Edit-Distanzen auf Wortebene mit folgenden Kosten: Einfügungen, Löschungen: an Anfang und Ende der Wortsequenzen 0, sonst Anzahl Zeichen im eingefügten/gelöschten Wort Substitutionskosten: zeichenbasierte Edit-Distanzen
9. Wir haben bewusst auf spezialisierte Heuristiken für Untertitel, Schreibweisealternativen und andere für dieses Korpus spezifische Titelabweichungsphänomene verzichtet.
10. Ob die Edit-Distanz über einem vorgegebenen Schwellwert liegen würde, lässt sich außerdem in einigen Fällen auch ohne deren tatsächliche Berechnung (und damit erheblich effizienter) feststellen (Tateishi & Kusui 2008), allerdings waren wir für diesen Beitrag an exakten Maßen interessiert. Außerdem ist die dortige Formel auf vollständige Edit-Distanzen und nicht auf Teilzeichenfolgen-Distanzen ausgelegt.
11. Eine Recall-Angabe ist hier nicht sinnvoll, da die manuelle Auswertung sich genau auf diese Menge beschränkt.

Bibliographie

Bär, Daniel / Zesch, Torsten / Gurevych, Iryna (2012): Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012*, S. 167–184.

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. S. 119–123.

Gius, Evelyn / Krüger, Katharina / Sökefeld, Carla (2019): Korpuserstellung als literaturwissenschaftliche Aufgabe. In *DHd2019 Book of Abstracts*.

Herrmann, Berenike / Lauer, Gerhard (2017): Das „Was-bisher-geschah" von KOLIMO. Ein Update zum Korpus der literarischen Moderne. In *DHd 2017 Digitale Nachhaltigkeit Book of Abstracts*. S. 107–111.

Kuhn, T. S. (1970). *The Structure of scientific revolutions*. 2nd ed., enlarged. Chicago: Chicago Univ. Press.

Lauer, Gerhard / Herrmann, Berenike (2018): Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Osnabrücker Beiträge zur Sprachtheorie* 92. 7–35

Levenshtein, Vladimir (1965): Binary codes capable of correcting deletions, insertions, and reversals. Englische Übersetzung in: *Soviet Physics Doklady*, Bd. 10, Nr. 8, S. 707–710, 1966.

Schöch, Christof (2017): Aufbau von Datensammlungen. In: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (Hrsg.): *Digital Humanities: eine Einführung.* Stuttgart: J. B. Metzler Verlag. S. 223–233.

Tateishi, Kenji / Kusui, Dai (2008): Fast Duplicated Documents Detection using Multi-level Prefix-filter. *Proceedings of the Third International Joint Conference on Natural Language Processing*, Bd. II.

Linked Ogham Stones – Semantische Modellierung und prototypische Analyse irischer Ogham-Inschriften

Homburg, Timo
timo.homburg@gmx.de
Hochschule Mainz, Deutschland

Thiery, Florian
rse@fthiery.de
Mainzer Zentrum für Digitalität in den Geistes- und Kulturwissenschaften, Deutschland

Einleitung

Die Ogham-Schrift ist eine in Irland und im westlichen Teil Britanniens (Wales und Schottland) vokommende antike Sammlung von Zeichen, die überwiegend auf (Grab-)steinen zur Dokumentation von Namen der Verstorbenen, ihrer Verwandtschafsbeziehungen oder auch für kleinere Geschichten verwendet wurde. Dabei stellt die Ogham-Systematik kein eigenständiges Alphabet, sondern eine Codierung in z.B. lateinische Alphabete dar.

Der Ogham-Zeichensatz besteht aus 26 Buchstaben. Wörter bzw. Sätze sind jeweils von unten nach oben bzw. links nach rechts zu lesen. Die Zeichen[1] werden neben einem zentralen Mittelstrich durch nach links (Aicme Beithe, ᚁᚂᚃᚄᚅ) bzw. rechts (Aicme hÚatha, ᚐᚑᚒᚓᚔ) abgehende rechtwinklige Striche charakterisiert. Durch schräge oder waagerechte Striche (oft auch Punkte), die den Mittelstrich durchkreuzen, werden Aicme Muine (ᚋᚌᚍᚎᚏ) und Aicme Ailme (ᚇᚈᚉᚊᚋ) dargestellt. Darüber hinaus gibt es sechs ergänzende Forfeda Zeichen (ᚕᚖᚗᚘᚙᚚ), welche im Mittelalter durch Lautveränderungen und Anpassungen an die lateinische Sprache hinzugefügt wurden. Das Ogham-Alphabet ist im Unicode Standard[2] (Range: 1680–169F) enthalten und gewährleistet so die Digitalisierbarkeit der Schrift. Eine der größten öffentlich zugänglichen Sammlungen von Ogham-Steinen befindet sich im Stone Corridor des University College Cork (UCC).

Abbildung 1: Ogham Stone 4 im UCC Stone Corridor [3] (Florian Thiery, CC BY 4.0, https://commons.wikimedia.org/wiki/File:UCC_Stone_4.jpg)

Stand der Forschung

Die Ogham-Schrift wurde in der Vergangenheit von verschiedenen Wissenschaftlern erforscht. Nach der Entdeckung der Ogham-Schrift (Ferguson 1864) und der anschließenden Erstellung des Ogham-Alphabets (Graves et. al. 1878) datierte MacNeill (1929) einige der Ogham-Inschriften. Das wohl kompletteste Standardwerk findet sich in Macálister (1945, 1949). Dieser hat darin das weitverbreitete Nummerierungsschema CIIC etabliert. Neben weiteren Forschungen der Geschichtswissenschaft zum Inhalt der Ogham-Inschriften erstellte Forsyth (1997) ein erstes Korpus von 37 Inschriften. Das Ogham 3D Projekt[4] scannt derzeit (155 Ogham Steine verfügbar) irische Ogham-Steine und stellt diese als Epidoc zur Verfügung.

Die auf den Steinen transkribierten Inschriften können in `formula words` (FW) und `nomenclature words` (NW) unterschieden werden. Beispiele für FW[5] sind MAQI (engl. son, z.B. CIIC 203) MUCOI (engl. tribe/sept, z.B. CIIC 197), ANM (engl. name, z.B. CIIC 206), AVI (engl. descendant, z.B. CIIC 40), CELI (engl. follower/devotee, z.B. CIIC 215) und KOI (engl. here is, z.B. CIIC 48).

Die Nomenklatur der irischen Personennamen enthüllt Details der frühgälischen Gesellschaft. Beispiele für solche NW[6] sind CUNA (engl. wolf/hound, z.B. CIIC 154) oder CATTU (engl. battle, z.B. CIIC 58). Andere Namen weisen auf einen göttlichen Vorfahren hin. Der Gott Lugh (LUC) oder das Wort ERC (engl. heaven/cow) kommt in vielen Namen wie LUGADDON (vgl. CIIC 4) oder ERCAVICCAS (vgl. CIIC 196) vor. Elemente, die physikalische Eigenschaften beschreiben, sind ebenfalls üblich. Zum Beispiel DALAGNI (engl. one who is blind, CIIC 119) oder DERCMASOC (engl. one with an elegant eye, CIIC 46).

Ansatz

Wir stellen die Ogham-Steine, deren Inhalte, die Beziehungen der auf Steinen vermerkten Personen, ihre Stammeszugehörigkeiten und weitere Metadaten als Linked Data bereit und ermöglichen somit deren Verarbeitung durch eine Reihe von Wissenschafts-Communities. Durch die Verwendung von Vokabularen wie Wikidata (Vrandečić et. al. 2014), FOAF (Brickley 2007) und Lemon (McCrae 2012) gewährleisten wir die Erstellung eines semantischen Wörterbuchs für Ogham, welches wir dynamisch aus Textquellen mittels Natural Language Processing Verfahren der Keyword Extraktion extrahieren. Die für uns relevanten Keywords haben wir aus der Literatur gesammelt und in unserem Repository veröffentlicht.[7] Die Erfassung der Ogham-Steine als Linked Data Ressourcen erlaubt es, durch Verknüpfung von Wissen und dessen Anreicherung folgende Forschungsfragen anzugehen:

- Klassifikation von Steinen (Familienhierarchie, Namensbeschreibung etc.)
- Visualisierung von Zusammenhängen (Verwandtschaftsbeziehungen, Stammesgrenzen) aus Linked Data generierten Karten
- Formale Erfassung und maschinenlesbare Kodierung von Ogham-Zeichen nach dem Vorbild von PaleoCodage (Homburg 2019)

Als Datenbasis für die Analysen stützen wir uns auf eine Wikidata-Retrodigitalisierung des CIIC Corpus von Macálister (1945, 1949), Epidoc-Daten des Ogham in 3D Projekts, sowie auf die Celtic Inscribed Stones Project (CISP[8]) Datenbank, die uns dankenswerterweise von Dr. Kris Lockyear zur Verfügung gestellt wurde. Des Weiteren pflegen wir aktiv fehlende und passende Elemente in Wikidata ein, um so später die Daten der Research Community im Sinne des SPARQL Unicorn (Thiery and Trognitz 2019a, 2019b) bereitzustellen. Der Sourcecode unserer App steht quelloffen auf GitHub zur Verfügung (Homburg & Thiery 2019).

Ergebnis

Erste Extraktionsergebnisse zeigen zunächst die Clusterung der irischen CIIC Oghamsteine (323 Steine) in Abbildung 2. In den Abbildungen 3-5 ist ersichtlich, dass sich Inschriften

zu FW und NW räumlich unterscheiden (z.B. Abbildungen 3 und 4 zu MAQI und CUNA). Für das CISP Datenset (2504), für das wir aktuell noch keine Geokoordinaten besitzen, sind in Tabelle 1 die erste Analyseergebnisse aufgeführt. Sie zeigen eine ähnliche Verteilung der Keywords wie im CIIC Datenset. Das Poster stellt diese und weitere Ergebnisse unserer ersten räumlichen Analysen dar. Weitere Daten können unserem Webviewer[9] entnommen werden.

Tabelle 1: Klassifikation der Steine aus CISP

Keyword	Vorkommen im CISP Dataset
MAQI (son)	710 (28%)
AVI (grandson)	72 (3%)
CELI (fellow)	7 (0,2%)
CUNA (dog)	174 (7%)
CATTU (cattle)	94 (4%)
NIOTA (nephew)	20 (0,8%)
BRAN (raven)	34 (1,3%)
BROCI (badger)	2 (0,08%)
ERC (cow)	38 (1,5%)
IVA (tree)	18 (0,8%)
VIR (man of battle)	38 (1,5%)
LUG (god)	73 (3%)

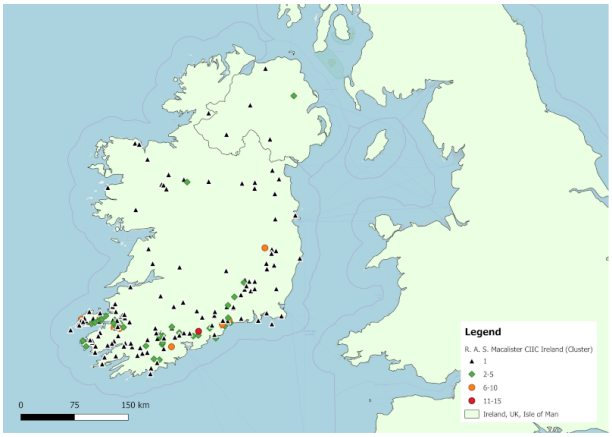


Abbildung 2: Cluster von Oghamsteinen in Irland nach Fundorten

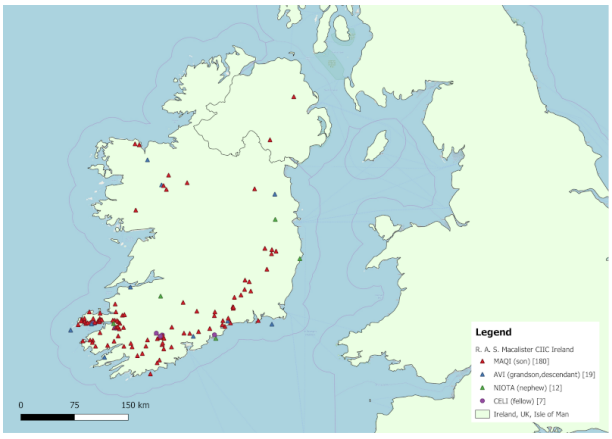


Abbildung 3: Oghamsteine mit Verwandschaftsbeziehungen

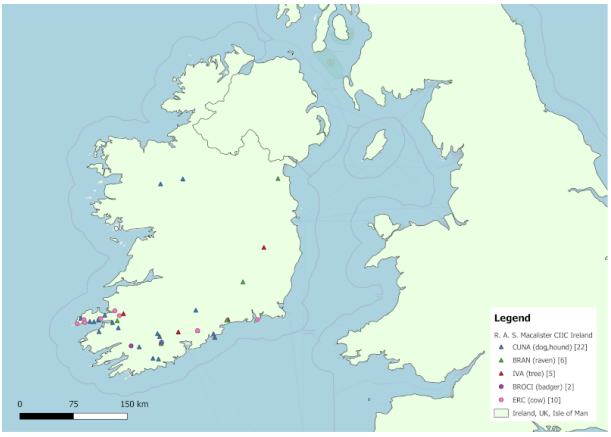


Abbildung 4: Oghamsteine mit Referenzen zu Tieren oder der Natur

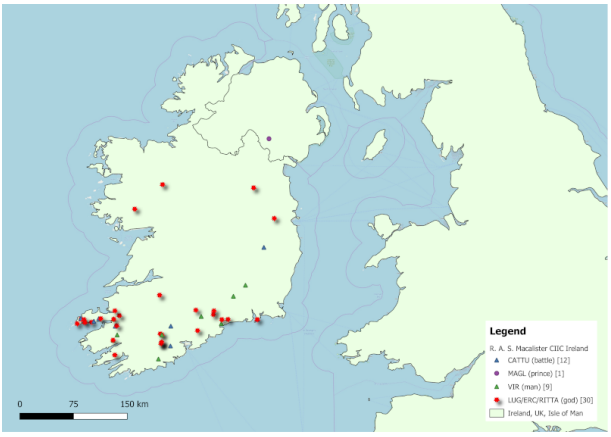


Abbildung 5: Oghamsteine mit Referenzen zu Menschen und Schlachten

Future Work

In Zukunft wird die Datenbasis mit erneuerten Daten des Ogham 3D Projektes und weiteren Quellen angereichert, sowie die Bereitstellung von Ogham-Daten in Wikidata forciert[10]. Hierbei wird uns Sophie Charlotte Schmidt weiterhin in unserem Ogi-Ogham Projekt unterstützen. Mit einem größeren Korpus werden die Analysen noch bessere Ergebnisse hervorbringen. Zudem ist eine Erweiterung der OliA Ontologien (Chiarcos and Sukhareva 2015) für Ogham angedacht, sowie die Publikation der Daten als Linked Data in Web.

Fußnoten

1. https://commons.wikimedia.org/wiki/Category:Ogham_letters
2. https://www.unicode.org/charts/PDF/U1680.pdf
3. https://www.ucc.ie/en/discover/visit/centre/stone-corridor/
4. https://ogham.celt.dias.ie/
5. https://ogham.celt.dias.ie/menu.php?lang=en&menuitem=40

6. https://en.m.wikipedia.org/wiki/Ogham_inscription#Nomenclature
7. https://github.com/ogi-ogham/oghamextractor/blob/master/words/words.csv
8. https://www.ucl.ac.uk/archaeology/cisp/database/
9. https://ogi-ogham.github.io/oghamextractor
10. Ein Zwischenstand hier: https://ogham.link

Bibliographie

Brickley, Dan/ Libby, Miller (2007): "FOAF vocabulary specification 0.91.".

Chiarcos, Christian / Sukhareva, Maria (2015): Olia–ontologies of linguistic annotation." in: *Semantic Web* 6, no. 4: 379-386.

Ferguson, Samuel (1864): Account of Ogham Inscriptions in the Cave at Rathcroghan, County of Roscommon. in: *Proceedings of the Royal Irish Academy (1836-1869)* 9: 160-170.

Forsyth, Katherine Stuart (1997): *"The ogham inscriptions of Scotland: An edited corpus."*: 2160-2160.

Graves, Charles / Limerick, C. (1876): "The ogham alphabet." *Hermathena* 2, no. 4: 443-472.

Homburg, T. (2019): *"PaleoCodage - A machine-readable way to describe cuneiform characters paleographically"*. DH2019 Book Of Abstracts - Short Papers.

Homburg, Timo/ Thiery Florian (2019): "OGI Ogham oghamextractor". https://github.com/ogi-ogham/oghamextractor.

Macalister, R. A. S. (1945): *"Corpus Inscriptionum Insularum Celticarum Vol. I."*, Dublin: Stationery Office.

Macalister, R. A. S. (1949): *"Corpus Inscriptionum Insularum Celticarum Vol. II.*, Dublin: Stationery Office.

MacNeill, Eoin (1929): *"Archaisms in the ogham inscriptions." Proceedings of the Royal Irish Academy. Section C: Archaeology, Celtic Studies, History, Linguistics, Literature* 39: 33-53.

McCrae, John / Montial-Ponsoda, Elena / Cimiano, Philipp (2012): "Integrating WordNet and Wiktionary with lemon." in *Linked Data in Linguistics*, pp. 25-34. Springer, Berlin, Heidelberg.

Thiery, Florian / Trognitz, Martina (2019a): "SPARQL Unicorn Github Repository", https://github.com/sparqlunicorn.

Thiery, Florian / Trognitz, Martina (2019b): Wikidata: A SPARQL(ing) Unicorn?", CAA: Computer Applications & Quantiative Methods in Archaeology - Book of Abstracts, Bournemouth .

Vrandečić, Denny / Markus Krötzsch. (2014): "Wikidata: a free collaborative knowledge base." .

Merkmale registrieren oder textuelle Phänomene identifizieren? Zur Vereinbarkeit von automatischer und manueller Textsortenanalyse

Thielert, Frauke
frauke.thielert@upb.de
Universität Paderborn, Deutschland

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

Schuster, Britt-Marie
brittms@mail.upb.de
Universität Paderborn, Deutschland

Georgi, Christopher
christopher.georgi@upb.de
Universität Paderborn, Deutschland

Kategorien wie „Textsorte“, „(kommunikative) Gattung“ oder „Genre“ gehören zu einem disziplinenübergreifenden Bestand und werden entsprechend in Sprach-, Literatur-, Kultur- sowie den Sozialwissenschaften verwendet. Allgemein lässt sich die Frage stellen, ob und inwieweit die genannten Kategorien in die Digital Humanities eingehen und inwieweit sie methodologisch reflektiert werden.

Textanalysen in der Text- und Korpuslinguistik

Die Auseinandersetzung mit einer Kategorie wie „Textsorte“ kann auf eine jahrzehntelange Fachgeschichte, besonders in der Sprachwissenschaft, zurückblicken, in der zwar keine Einigkeit hinsichtlich des Verständnisses des Konzepts „Textsorte“ erzielt worden ist, jedoch deutlich geworden ist, dass in diesem Zusammenhang die Unterschiede zwischen Texten unter Einbezug unterschiedlichster Textebenen zu modellieren sind. Textsorten zeigen sich – grosso modo – nicht nur anhand von musterhaften Ausprägungen auf textgrammatischer, -semantischer und -pragmatischer Ebene, sondern berühren auch die Materialität, Kodalität (Nutzung unterschiedlicher Zeichenressourcen) und ggf. eine spezifische Ortsgebunden-

heit, die Lokalität. Heutige Textsortenmodelle sind Mehrebenen-Modelle, was mit der Annahme verknüpft ist, dass die einzelnen Ebenen in einem wechselseitigen Abhängigkeitsverhältnis stehen (vgl. Adamzik [2] 2016; Heinemann/Heinemann 2002). Um eine „Textsorte" oder ein „Textmuster" zu erfassen, ist eine umfassende Nutzung des linguistischen Beschreibungsinstrumentariums erforderlich. Die Attraktivität von Kategorien wie „Textsorte" ist v.a. darin gesehen worden, dass sie Einblick in den ‚kommunikativen Haushalt', also in spezifische Ordnungsleistungen einer Gesellschaft und Kultur ermöglichen (vgl. Fix 2006). I.d.R. wird die Ausprägung von Textmustern auf rekurrente Aufgaben und deren Lösung zurückgeführt, die wiederum einen Einblick in gesellschaftliche Relevanzen bieten. Gerade der in den letzten zwei Jahrzehnten geführte (text)linguistische Diskurs hat zudem erbracht, dass zunächst als dem Text äußerlich gedachte Faktoren wie Kontext, einschließlich der Beziehung zwischen Textproduzent und -rezipient, nichts dem Text Äußerliches sind, sondern durch den Text hergestellt werden. Zudem ist eine Kategorie wie „Stil" die etwa auch Dialogizität oder Perspektivität umfasst, verstärkt als Textsortenstil verstanden worden, der sich aus der Sichtung aller Textebenen im Zusammenspiel ergibt (vgl. Sandig [2] 2006). Eine wichtige Neuorientierung in der textlinguistischen Betrachtung stellen Modelle dar, die konsequent von der textlichen Oberfläche ausgehend, ohne sich allerdings auf Syntax und Lexik zu beschränken, thematische, situative und funktionale Hinweise und damit zentrale Textdimensionen erschließen (vgl. Hausendorf et al. 2017, historisch: Schuster 2019).

Mehrebenen-Modelle zur Beschreibung von Textsorten sind fast ausnahmslos Produkt von Annahmen, die ebenso aus Sprach- und Kommunikationstheorien wie aus einzelnen Textexemplaren hergeleitet werden. Diese werden zumeist nur an geringen Textmengen überprüft. Da wichtige Untersuchungsebenen ‚vorgegeben' sind, verfährt die Methode top down. Wie korpuslinguistische Untersuchungen mit kulturanalytischen Interesse – also nicht im engeren Sinne textlinguistische Studien – deutlich gemacht haben, ließen sich einige auch in der Textlinguistik für wichtig erachtete Ausdrucksmuster durch die Berechnung von Kollokationen, n-Grammen auf Wort und Phrasenebene oder Keywords ermitteln (vgl. Bubenhofer/Scharloth 2016). Dabei handelt es sich um Bottom-Up-Verfahren, die zu neuen Hypothesen und Annahmen führen können.

Innerhalb der Diskussionen um Textsortenklassifikation und Texttypologie ist deutlich geworden, dass „Textsorten" keine starren Entitäten sind; sie sind nicht vollständig festgelegt und erlauben Veränderungen. Aus dieser Variabilität ergibt sich das generelle Potential zum Wandel von Textsorten, der durch die Nutzung und Grenzen von Spielräumen bestimmt wird. Die entsprechenden Konventionalisierungsprozesse sind jedoch bisher kaum betrachtet worden.

Textanalysen in den Digital Humanities

Den bisher skizzierten Textauffassungen stehen Zugriffe auf die Kategorie „Text" gegenüber, die in den Digital Humanities bevorzugt werden. Grundsätzlich scheint die Kategorie „Textsorte" eine Hilfskategorie zu sein, mit der größere Datenmengen (z.B. Referenzkorpora) geordnet werden. Fragen der Textstrukturiertheit werden im Zusammenhang mit dem Text-Encoding z.B. in digitalen Editionen aufgeworfen (vgl. z.B. TEI-P5 Guidelines 2019), wobei die Ergebnisse nur selten Niederschlag in quantitativen Analysen finden. Texte werden zudem für das Training von Methoden ganz unterschiedlicher Anwendungen (z.B. Sentiment-Analysis, Stilometrie oder Topic Modelling) verwendet. Der Text(sorten)begriff bleibt dabei unterspezifiziert, indem „Text" mit Dokumenten, Sätzen oder Mengen sinntragender Struktureinheiten gleichgesetzt (vgl. z.B. Ravi/Ravi 2015: 16; de Rose et al. 1997: 6) oder nach Alltagsverständnis differenziert wird (vgl. z.B. Medhat et al. 2014: 1096). Einschlägige Kategorien der DH sind daneben die des (Gattungs)Stils, Autorenstils oder Registers. Dabei deckt sich das Stilverständnis nicht mit dem holistischen Verständnis von „Stil" als einer alle Textebenen durchwirkenden Kategorie, mit der sozialer Sinn erzeugt wird. Das Text- und insbesondere auch das Stil- und Registerverständnis der DH ist wesentlich an Merkmalen orientiert, wie dies etwa in der folgenden Äußerung zum Tragen kommt, die hinsichtlich des Verständnisses hochaggregierter geisteswissenschaftlicher Kategorien in den DH charakteristisch ist: „Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively." (Hermann et al. 2015: 44).

Merkmale bei Untersuchungen zu Textgattungen und Diskursen sind etwa Frequenzen von Inhalts- und Funktionswörtern, der Variantenreichtum des Wortschatzes, Satzlängen, n-Gramme oder mit Parsern ermittelte syntaktische Strukturen; die Auswahl wird in der Regel nicht begründet und scheint durch ihre Operationalisierbarkeit selbst gerechtfertigt. Exemplarisch hierfür steht das Topic-Modeling (Fankhauser et al. 2016, Viehhauser 2017). Dabei wird Text im Sinne des „bag-of-words"-Ansatzes als „Behältnis" von Wörtern verstanden, wobei die grammatikalische Struktur und selbst die Wortfolge unberücksichtigt bleiben (vgl. Blei et al. 2003). Bibers (1988) und Bibers/Finegans (2014) multidimensionale Analysen (Schöch/Pielström 2014: S. 2f.), die sich am Genre- und Registerbegriff orientieren, fassen eine Vielzahl von Merkmalen zu Merkmalbündeln zusammen, berücksichtigen jedoch kaum die Funktionalität bzw. pragmatische Dimension von Texten. Auffällig ist, dass in diesen und anderen Studien Merkmalen wie der Satzlänge oder Komplexität von Sätzen eine Bedeutsamkeit für Stil, Register oder Genre zugeschrieben wird, die in qualitativen Studien randständig ist. Dass „formal features" auch durchaus auf interpretierbaren Kategorien basieren, rückt ebenfalls wenig ins Bild. Zusammenfassend darf behauptet werden, dass bei Text- und Stilklassifizierungen in den DH Merkmale der Textoberfläche bevorzugt behandelt werden.

Unvereinbare Traditionen? Ein Fallbeispiel

Man kann mit Blick auf diese unterschiedlichen Forschungstraditionen, die hier bewusst pointiert gegenübergestellt wurden, konstatieren, dass herkömmliche qualitativ-linguistische Studien, obgleich sie stark mit dem Begriff „Muster" operieren, sich bisher kaum für statistische Signifikanzen u.ä. interessiert haben, während wiederum stilo- und textometrische Studien mit einem „unterkomplexen Textbegriff" arbeiten und nach Bubenhofer/Scharloth es bisher versäumt haben, „Texte als komplexes Gewebe zu operationalisieren" (2015: 13). Grundsätzlich gilt: Während merkmalsorientierte Zugänge auf der textlichen bzw. sprachlichen Oberfläche operieren, gehen phä-

nomenorientierte Modelle von textlichen Dimensionen (z.B. der Beziehungsdimension) aus, die in ihrer Relevanz für die textliche Kommunikation erkannt worden sind und auf ihre sprachliche Gestaltung hin befragt werden. Zwar mehren sich in den letzten Jahren die Versuche, im Sinne der „mixed methods" quantitative und qualitative Methoden miteinander zu verbinden, jedoch ist im Hinblick auf den Text- und Textsortenbegriff bisher nicht deutlich, ob sich diese komplementär zueinander verhalten oder zu möglicherweise sich widersprechenden Befunden führen.

In unserem Beitrag möchten wir ein Mehrebenen-Modell vorstellen, das in dem DFG-Projekt: „Die Evolution von komplexen Textmustern: Entwicklung und Anwendung eines korpuslinguistischen Analyseverfahrens zur Erfassung der Mehrdimensionalität des Textmusterwandels" entstanden ist. Es verbindet unterschiedliche Zugriffe auf die Kategorie „Text" und bezieht quantitative und qualitative Text(sorten)analyse spiralförmig aufeinander. Am Beispiel der Verwendung personaldeiktischer Ausdrücke (*ich – du – wir – ihr*) und entsprechender Possessiva sowie Indefinitpronomen wie *man* , die in unterschiedlichen historischen Textgruppen leicht identifizierbar sind, möchten wir auf Basis eines Pilotkorpus von Zeitungstextsorten des Zeitraums 1830 bis 1930 sowie mehrerer Vergleichskorpora aus dem Deutschen Textarchiv (DTA) zeigen:

1. welche Texteigenschaften (allein) durch die automatische, korpusbasierte Textanalyse, insbesondere durch die Nutzung von Part-of-Speech- und Lemma-Informationen, auch in Bezug auf verschiedene Binnentextsorten, zutage treten und hinsichtlich welcher Forschungsfragen dies aufschlussreich ist. So werden durch diachrone Längsschnittuntersuchungen Frequenz, Signifikanz und Typizität entsprechender Ausdrücke, letzteres insbesondere durch Bezugnahme auf Vergleichskorpora, jedoch auch eine hohe Varianz der Ausdrücke sichtbar. Eine derartige Zugriffsweise erlaubt, ergänzt durch POS-sensitive Suchen, einen Einblick in Konstanz und Wandel von Verfasserreferenz und Rezipientenansprache. Sie bieten durch ihre Irritationsmomente einen Ansatzpunkt, um Hypothesen zu Zeiträumen, die für Wandelphänomene interessant sind, zu bilden. Sie dienen ferner zum Abgleich mit auf schmalen Korpora generierten Ergebnissen (vgl. Lefévre 2017: 150), die durch eine solche Zugangsweise relativiert werden. So zeigt sich – gemessen an der vorliegenden Forschungsliteratur und an Vergleichskorpora – ein erstaunlicher Anstieg von *ich* -, *du* -, -*wir* und *ihr* -Verwendungen.
2. was durch eine flankierende manuelle Annotation mit einem vordefinierten Tagset ins Blickfeld rückt. Es wird deutlich, dass die personaldeiktischen Verwendungen sich nicht gleichmäßig über alle Textsorten verteilen, sondern sich besonderen Textsorten wie dem Erfahrungs- und Erlebnisbericht verdanken. Ferner wird deutlich, dass sich relativ von Textkotext und -kontext bestimmte Lesarten (z.B. das Verfasserkollektive oder Rezipienten umschließende, inklusive *wir*) herausbilden, die weiterführende Analysen zu sprachlicher Inklusion und Exklusion erlauben und damit die Beziehungsdimension von Texten erschließen sowie die Beantwortung von Fragestellungen zu Funktionalität und Sprachhandlungsprofilen der vorliegenden Textsorten ermöglichen.

Somit stehen einerseits die Wandelbarkeit der Verteilung von sprachlichen Einheiten vor dem weiten Horizont von Textgruppen, andererseits die Funktionalität von sprachlichen Einheiten für die Konstitution bestimmter Textsorten im Vordergrund. Sowohl die unterschiedliche Verteilung von personaldeiktischen Formen als auch die spezifische Funktionalität von sprachlichen Einheiten, wie wir diskutieren möchten, ist nicht selbsterklärend, sondern gleichermaßen von Forschungshypothesen und -interessen abhängig. Abschließend möchten wir deshalb Überlegungen zu den folgenden Fragen bieten: Ist die „Bricolage" (Bubenhofer/Dreesen 2018) aus Ansätzen und Methoden sehr unterschiedlicher Forschungstraditionen überhaupt sinnvoll? Lassen sich komplexe, kontextbasierte deiktische Kategorien messen, aber auch: Lassen sich damit verknüpfte Handlungsmuster überhaupt operationalisieren und in einem Tagset darstellen?

Bibliographie

Adamzik, Kerstin (2016): *Textlinguistik. Grundlagen, Kontroversen, Perspektiven.* 2., völlig neu bearbeitete, aktualisierte und erweiterte Neuauflage. Berlin, Boston: De Gruyter.

Biber, Douglas / Finegan, Edward (1994): *Multi-Dimensional Analyses of Authors' Styles:* Some Case Studies from the Eigtheenth Century. Oxford: Oxford University Press.

Biber, Douglas (1988): *Variation across Speech and Writing* . Cambridge: Cambridge University Press.

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent Dirichlet Allocation", in: *Journal of Machine Learning Research* 3: 993–1022.

Bubenhofer, Noah / Dreesen, Philipp (2018): "Linguistik als antifragile Disziplin? Optionen in der digitalen Transformation", in: *Digital Classics Online* 4: 63–75.

Bubenhofer, Noah / Scharloth, Joachim (2016): "Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik", in: *Sprache – Kultur – Kommunikation / Language – Culture – Communication* . Ein internationales Handbuch zu Linguistik als Kulturwissenschaft / An International Handbook of Linguistics as a Cultural Discipline. Bd. 43. Berlin, Boston: De Gruyter: 924–933.

Bubenhofer, Noah / Scharloth, Joachim (2015): "Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate", in: *Zeitschrift für Germanistische Linguistik* 43: 1–26.

DeRose, Steven J. / Durand David G. / Mylonas, Elli / Renear, Allen H. (1997): "What is text, really?", in: *SIGDOC Asterisk Journal of Computer Documentation* 21.3: 1–24.

Fankhauser, Peter/ Knappen, Jörg / Teich, Elke (2016): "Topical Diversification Over Time In The Royal Society Corpus" in: Eder, Maciej / Rybick, Jan (eds.): *Digital Humanities* , 11.–16. Juli 2016, Krakow: Conference Abstracts.

Fix, Ulla (2006): "Was heißt Texte kulturell verstehen? Ein- und Zuordnungsprozesse beim Verstehen von Texten als kulturellen Entitäten", in: Blühdorn, Hardarik / Breindl, Eva / Waßner, Ulrich Hermann (eds.): *Text – Verstehen.* Grammatik und darüber hinaus. Berlin / Boston: De Gruyter: 254–276.

Hausendorf, Heiko / Kesselheim, Wolfgang / Kato, Hiloko / Breitenholz, Martina (2017). *Textkommunikation* : ein textlinguistischer Neuansatz zur Theorie und Empirie der Kommunikation mit und durch Schrift. Berlin / Boston: De Gruyter.

Heinemann, Wolfgang / Heinemann, Margot (2002): *Grundlagen der Textlinguistik* . Interaktion – Text – Diskurs. Berlin / Boston: De Gruyter.

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

Hermanns, Fritz (2009): "Linguistische Hermeneutik. Überlegungen zur überfälligen Einrichtung eines in der Linguistik bislang fehlenden Teilfaches", in: Felder, Ekkehard (eds.): *Sprache* . Heidelberg: Springer: 179–214.

Jannidis, Fotis (2019): "Digitale Geisteswissenschaften – Offene Fragen, schöne Aussichten", in: *ZMK* 10: 63–70.

Lefévre, Michel (2017): "Von der "Berlinischen Privilegierten Zeitung" zur "Königlich Privilegierten Berlinischen Zeitung". Entwicklungstendenzen in der Äußerungsstruktur, Textgestaltung und Syntax", in: Pfefferkorn, Oliver / Riecke, Jörg / Schuster, Britt-Marie (eds.): *Die Zeitung als Medium* . Berlin / Boston: De Gruyter: 149–163.

Medhat, Walaa / Hassan, Ahmed / Korashy, Hoda (2014): "Sentiment analysis algorithms and applications. A survey", in: *Ain Shams Engineering Journal* 5: 1093–1113.

Ravi, Kumar / Ravi, Vadlamani (2015): "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", in: *Knowledge-Based Systems* 89: 14–46.

Sandig, Barbara (2006): *Textstilistik der deutschen Sprache* . 2. völlig neu bearbeitete und erweiterte Auflage. Berlin / Boston: De Gruyter.

Schöch, Christof / Steffen Pielström (2014): *Für eine computergestützte literarische Gattungsstilistik, Jahrestagung der Digital Humanities im deutschsprachigen Raum.* http://dig-hum.de/sites/dig-hum.de/files/Schoch-Pielstrom_2014_Gattungsstilistik.pdf. [letzter Zugriff 23. September 2019]

Schuster, Britt-Marie (2019): "Sprachgeschichte als Geschichte von Texten", in: Bär Joachim / Lobenstein-Reichmann, Anja /Riecke, Jörg (eds.): *Handbuch Sprache in der Geschichte* . Berlin / Boston: De Gruyter: 219–240.

TEI Consortium (2019): TEI P5. Guidelines for Electronic Text Encoding and Interchange. Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative, now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium. Version 3.6.0 (16. Juli 2019)

Viehhauser, Gabriel (2017): "Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende." In: Zeitschrift für digitale Geisteswissenschaften. 2017. PDF Format ohne Paginierung. DOI: 10.17175/2017_003.

Metadaten-basierte Visualisierungen im Stilometrie-Paket „Stylo“

Pielström, Steffen

pielstroem@biozentrum.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Maciej, Eder

maciejeder@gmail.com
Pedagogical University of Kraków, Polen

Die Programmbiblothek *Stylo* (Eder et al. 2016) für die Programmiersprache *R* bietet ein breites Spektrum an Funktionen für die stilometrische Analyse von Textcorpora, darunter Clusteranalyse auf der Basis von Wort- und NGramm-Frequenzen, Textklassifikation und die Identifikation distinktiver Merkmale für eine bestimmte Textgruppe. Die Funktionen nehmen dabei ganze Ordner nicht vorverarbeiteter Textdateien und geben umfangreiche Analyseergebnisse zurück, in den meisten Fällen inklusive fertiger Visualisierungen. Zusätzlich können viele der wichtigen *High-Level*-Funktionen auch über ein graphisches *Userinterface* bedient werden, womit die Basisfunktionalitäten von *Stylo*, obwohl es sich um ein Programmbibliothek handelt, auch ohne Programmierkenntnisse genutzt werden können. Nicht zuletzt bedingt durch den Komfort und die Einsteigerfreundlichkeit dieser Zugänge ist die *Stylo*-Bibliothek eines der populärsten Werkzeuge für die stilometrische Forschung in den Digital Humanities.

Dabei ist *Stylo* ursprünglich aus einer Sammlung von Skripten und Funktionen entstanden, die von den Entwicklern selbst für ihre Forschung gebraucht wurden. Die schrittweise Weiterentwicklung und Funktionserweiterung spiegelt in vielen Fällen die Bedürfnisse und Forschungsinteressen des Entwicklerteams wieder, und auch die Art und Weise, wie bestimmte Probleme in *Stylo* gelöst werden, ist nicht zuletzt durch die Arbeitsgewohnheiten der Entwickler bestimmt.

Ein Aspekt, der immer wieder zu Nachfragen von Usern geführt hat, ist der Umgang mit Metadaten in der durch die Community wohl am häufigsten genutzte *High-Level*-Funktion *stylo()*. Diese Funktion nimmt ein Corpus in Form eines Ordners mit Textdateien und erzeugt daraus wahlweise eine Clusteranalyse in Form eines Baumdiagramms, oder eine Hauptkomponentenanalyse, dargestellt als *Scatterplot*, um die Ähnlichkeitsbeziehungen der Texte untereinander darzustellen. Texte, die aufgrund von Vorwissen einer bestimmten Gruppe zugeordnet werden, erscheinen in der Visualisierung in der gleichen Farbe. So werden zum Beispiel bei einem klassischen Autorenschaftsproblem alle Texte, von denen vorher bekannt ist, daß sie von der gleichen Autorin/vom gleichen Autor stammen, in der gleichen Farbe dargestellt. Dadurch lässt sich in der Graphik schnell erkennen, wie gut Texte einer Gruppe tatsächlich nach stilometrischen Kriterien zusammen clustern.

Die Informationen über die Gruppenzugehörigkeit eines Textes entnimmt *Stylo* traditionell dem Dateinamen. Dafür muss jede Textdatei nach der Konvention *Gruppe_Dokument.Endung* benannt sein. Das Drama "Hamlet" von Shakespeare wird also zum Beispiel mit dem Dateinamen Shakespeare_Hamlet.txt versehen, wenn alle Stücke von Shakespeare in der gleichen Farbe erscheinen sollen.

Bislang war die systematische Benennung der Textdateien der einzige Weg, solche Information zur Gruppenzugehörigkeit an die Funktion zu übermitteln. Von Nutzerweite wurde immer wieder der Wunsch nach zusätzlichen Möglichkeiten geäußert, Metadaten zur Gruppenzugehörigkeit der Texte an die Funktion zu übergeben.

In den neueren *Stylo*-Versionen haben wir nun eine flexiblere Möglichkeit implementiert. Die Funktion *stylo()* vefügt nun über einen Parameter *metadata,* dem die Information zur Gruppierung der Texte in Form einer Gruppierungsvariable übergeben werden kann. Im einfachsten Fall ist das ein Vektor, dessen Länge der Anzahl der Texte im Corpus entspricht, und der für jeden Text ein Gruppenlabel liefert.

```
authornames <- c("Goethe", "Goethe", "Goe-
the", "Rodan", "Rodan", ...)
stylo(metadata = authornames)
```

Die Funktion akzeptiert sowohl Faktor als auch einen Vektor von Strings als Gruppierungsvariable. Die andere Möglichkeit ist, die Information zur Gruppenzugehörigkeit der Texte in einer CSV-Datei zu hinterlegen und dem Parameter den Dataipfad als String zu übergeben. Die betreffende CSV-Datei enthält eine Spalte mit der Überschrift "filename", die alle Dateinamen des Corpus in alphabetischer Reihenfolge enthält, und mindestens eine weitere Spalte mit Gruppenlabels. Um die Spalte mit der gewünschten Gruppierungsvariable auszuwählen wird der Titel der gewünschten Spalte an den Funktionsparamter *grouping.column* übergeben.

```
stylo(metadata = "metadata.csv", groupin-
g.column = "author")
```

Der Default-Wert ist "author". Wenn dem Paramter *grouping.column* kein Wert zugewiesen wird, muss die Datei eine Spalte mit dem Default-Wert "author" als Überschrift enthalten.

Dieser zusätzliche Parameter in der *stylo()*-Funktion erlaubt nun flexibel mit der Gruppenzugehörigkeit der Texte zu experimentieren, ohne daß dafür die Textdateien umbenannt werden müssen. Das Poster wird diese neuen Funktionalitäten vorstellen und durch Codebeispiele und Visualisierungen erläutern.

Bibliographie

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): „Stylometry with R: a package for computational text analysis“, in *R Journal*, **8**(1): 107-121, url: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Modeling disciplinary structure with uniform manifold approximation and projection

Noichl, Maximilian
noichlmax@hotmail.co.uk
Universität Wien, Österreich

Philosophy seems have been slower in the adoption of digital methods for investigative purposes than other humanities. But the expanded view which digital methods allow does make sense when broad questions of intellectual structure and change, particularly in philosophy of science (Pence, Ramsay, 2018) or history of philosophy (Betti et al. 2019) are concerned.Assessments of disciplinary structures, which are commonly found in philosophical texts tend to focus on the personal and argumentative relations between small numbers of prominent actors, from which views about the disciplines as a whole are deduced.

To supplement such detailed accounts, this project proposes the use of Uniform Manifold *Approximation and Projection* (McInnes et al. 2018) for the mapping and clustering of disciplines. Benchmarks and comparisons to similar methods are produced, and the method is applied to medium-large to large samples of papers from the disciplines of philosophy and history. In particular, it tries to answer questions about the (often stereotypical) relations between specific subfields, the relation between gender and academic subfield, and the prominence of different subfields in public debate.

Preliminary results are already available for philosophy. As the establishment of disciplinary boundaries of philosophy is a notoriously hard problem, the adopted sampling strategy was as expansive as possible. PhilPapers, an expansive index of recent philosophy, has compiled a list of 1349 journals of philosophy, which links, at the time of writing to 1782816 indexed articles. The list was downloaded and compiled into a Web of Science query, to get the citation data for individual papers. Journals were removed from the query if they were clearly from the core of another discipline (e.g. Experimental Psychology, Historia) and had more than a thousand entries in the result of the query, as those journals could be expected to strongly change the results of further analysis. Only records were sampled that were cited at least four times, which resulted in a total sample of n=75942 articles, with 1194451 citations events to 159647 unique sources.

The citation-data of the articles was treated like the words in a standard computational text classification problem: For the map-layout, citation-vectors were reduced with singular value decomposition (SVD) to remove noise. This data was in turn transformed with uniform manifold approximation and projection (UMAP) into a two dimensional map, which will not only be presented as a poster, but is made available in interactive online form, (https://homepage.univie.ac.at/noichlm94/full/zoom_final/index.html) of which the graphic below ought to give an idea. Each dot of the mapping represents a paper, which is positioned relative to all other papers according to similarities in the sources it cites. The clustering was produced using hDBSCAN5 on a 30-dimensional UMAP-embedding of the data. Clusters can be interpreted as groups of scholarly co-engagement. The clusters were identified using the most frequent words from the abstracts of the papers, the most cited sources, and the most common venues of publication.

For philosophy, the preliminary results suggest that (a) the divide into analytical and continental philosophy is generally overstated. While continental philosophy, in contrast with multiple accounts (West, 1996; Glendinning, 2006), does form a distinct cluster at an appropriate level of detail, and therefore is quite coherent in terms of scholarly co-engagement, analytical philosophy fails to do so, validating recent scholarship (Preston, 2004; Glock, 2008) (b) Using a two-sided binomial test on a small-cluster-solution, 135 of 170 clusters show significant ($p < 0.0003$) deviations from a 1:1 gender-ratio. Of those 135, 2 clusters were dominated by female, 133 by male authors. (c) There is a clear relation between certain academic fields and the amount of public attention they generate. In particular medical ethics are frequently discussed by the public.

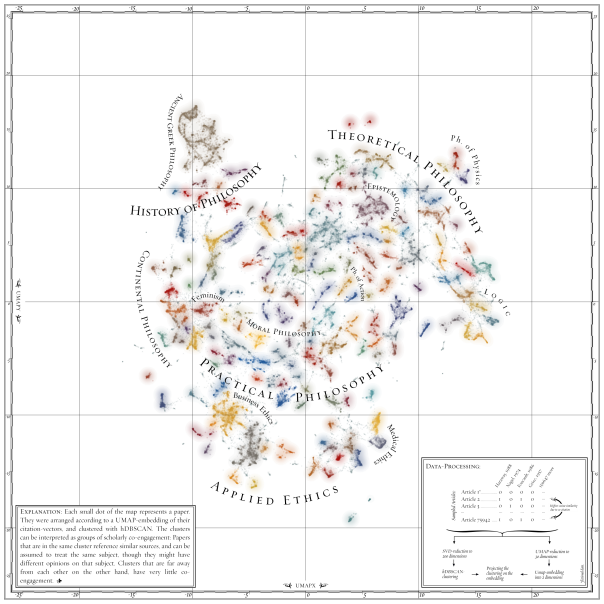


Figure 1: Map-layout.

Bibliography

Betti, A. / Van Den Berg, H. / Oortwijn, Y. / Treijtel, C. (2019): History of Philosophy in Ones and Zeros. In Curtis M. & Fischer E. (eds.), Methodological Advances in Experimental Philosophy. Bloomsbury: London.

Healy, K. (2013): *A Co-Citation Network for Philosophy* (accessed Feb 28, 2019).

Hobbs, V. (2014): Accounting for the Great Divide: Features of Clarity in Analytic Philosophy Journal Articles. *English for Academic Purposes*, 15, 27–36.

Glendinning, S. (2006): *The Idea of Continental Philosophy: A Philosophical Chronicle;* Edinburgh University Press: Edinburgh.

Glock, H.-J. (2008): *What Is Analytic Philosophy?;* Cambridge, MA.

Kreuzman, H. (2001): A Co-Citation Analysis of Representative Authors in Philosophy: Examining the Relationship between Epistemologists and Philosophers of Science. *Scientometrics*, 50 (3), 525–539.

McInnes, L. / Healy, J. / Melville, J. (2018): UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*:1802.03426.

McInnes, L. / Healy, J. / Astels, S. (2017): Hdbscan: Hierarchical Density Based Clustering. *The Journal of Open Source Software*, 2 (11), 205.

Pence, C. / Ramsey, G. (2018): How to Do Digital Philosophy of Science. *Philosophy of Science* 85 (5): 930–41.

Preston, A. (2004): Prolegomena to Any Future History of Analytic Philosophy. *Metaphilosophy*, 35 (4), 445–465.

Weingart, S. B. (2015): Finding the History and Philosophy of Science. *Erkenntnis*, 80 (1), 201–213.

West, D. (1996): *An Introduction to Continental Philosophy;* Polity Press: Cambridge, MA.

Modellierung von Annahmen als Basis für Rekonstruktionen von Architektur

Albers, Laura

laura.albers@kunstgeschichte.uni-muenchen.de
LMU München, Deutschland

Große, Peggy

peggy.grosse@hs-mainz.de
HS Mainz, Deutschland

Die Rekonstruktion von historischen Bauten und ihren Ausstattungen stellen einen wichtigen Beitrag für kunst- und architekturhistorische, soziokulturelle und denkmalpflegerische Fragestellungen dar. Die anschauliche Form einer Rekonstruktion ist das analoge oder digitale dreidimensionale Modell. Damit gewinnt das unter Umständen nicht mehr vorhandene oder veränderte Bauwerk seine äußere Form ‚zurück'. Gleichermaßen kann unter Rekonstruktion eine Ansammlung von Wissen zu einem Objekt gemeint sein, deren Teile zueinander ins Verhältnis gesetzt werden und eine netzwerkartig aufgebaute Kontextualisierung ermöglichen. Diese Teile können textuell oder bildlich vorliegen, in Form von Quellenmaterial oder von wissenschaftlichen Analysen.

Die Grundlage für jede Rekonstruktion ist entweder vorliegendes, sich direkt auf das Bauwerk beziehende Quellenmaterial (Fotografien oder Zeichnungen, mündliche oder schriftliche Überlieferungen) oder wissenschaftliche Annahmen und Zuweisungen, Rückschlüsse und Schlussfolgerungen. Letztere können sich wiederum auf Quellenmaterial oder wissenschaftliche Analysen berufen, welche sich nicht direkt auf das Bauwerk beziehen.

In vielen Fällen ist die Quellenlage für Rekonstruktionen dürftig, weshalb es sich oftmals um hypothetische Ergebnisse handelt. Dadurch wird die Argumentation, wie es zu einem bestimmten Ergebnis kommt, umso wichtiger. Diese zu dokumentieren stellt die Basis für die weitere wissenschaftliche Forschung und Rezeption anhand des entsprechenden Modells dar.

Es stellt sich die Frage, wie die Belegbarkeit und Nachvollziehbarkeit digital richtig abgebildet werden können, um die Informationen zur Rekonstruktion im Sinne des Semantic Web angemessen zu beschreiben. Anhand von zwei Fallbeispielen werden Lösungen vorgestellt, wie man die Herkunft von Informationen bzw. ihren wissenschaftlichen Bearbeitungskontext dokumentieren kann. Beide Projekte arbeiten mit der virtuellen Forschungsumgebung WissKI (http://wiss-ki.eu/), in der die Erfassung und semantische Erschließung der Daten auf Grundlage von CIDOC CRM (als Standard anerkannte Ontologie für das kulturelle Erbe, ISO 21127, http://www.cidoc-crm.org/) erfolgt.

Architekturgebundene Ausstattung als Medium von Herrschaftsanspruch am Beispiel zweier Deutschordensresidenzen

Das erste Projekt befasst sich im Rahmen einer Dissertation an der Ludwig-Maximilian-Universität München (Betreuer Prof. Dr. Stephan Hoppe) mit architekturgebundener Ausstattung zweier Deutschordensresidenzen als politische Sprache und Medium von Herrschaftsanspruch: die Residenz Ellingen und das Schloss Mergentheim. Unter architekturgebundener Ausstattung werden Wand- und Deckenmalerei, Stuckarbeit und Bauplastik, Fußböden und alle weiteren mit der Architektur fest verbundenen Ausstattungsgegenstände begriffen, denen im Kontext des Herrschaftsausdrucks ein Sinngehalt beigemessen werden kann.

Der Deutsche Orden ist als mittelalterlicher Ritterorden entstanden und war in der Frühen Neuzeit ein wichtiger politischer, zwischen religiösen Zielen und weltlichen Interessen situierter Akteur des Heiligen Römischen Reiches. Der Landkomtur der größten Ordensprovinz Franken hatte seinen Sitz im mittelfränkischen Ellingen. (Konter 2011: 16) Die Ordensleitung in Person des Hoch- und Deutschmeisters residierte seit Mitte des 16. Jahrhunderts im repräsentativen Ordensschloss in Mergentheim. (Trentin-Meyer 2004: 42) Dieser Ort befindet sich ebenfalls in der Fränkischen Ordensprovinz, stellt jedoch ein eigenständiges, dem Hoch- und Deutschmeister unterstelltes Kammergut, das „Meistertum Mergentheim", dar. (Konter 2011: 16) Der ständige Interessenkonflikt über die territoriale Herrschaft, die stetig unabhängiger werdende Ordensprovinz Franken und die traditionsgemäß verpflichtende Wahrung der Hierarchie innerhalb des Ordens erforderten ein ständiges Neupositionieren und Reagieren auf den jeweils anderen. Gleichzeitig musste sich der Deutsche Orden als politischer Akteur im Reich bewegen und sich trotz innerer Konflikte geschlossen behaupten. Zu Beginn des 18. Jahrhunderts erfahren die süddeutschen Territorien des Deutschen Ordens einen relativen Aufschwung, der sich sowohl in Ellingen als auch in Mergentheim in Bautätigkeiten im Sinne von Neubauten und Erweiterungen äußert. Ende des 18. Jahrhunderts löst sich der Landkomtursitz in Ellingen auf und wird Mergentheim unterstellt. Kurz darauf folgt die Auflösung des gesamten Ordens. (Konter 2011: 15) Das 18. Jahrhundert stellt somit eine von Aufschwung und Niedergang geprägte Zeitspanne des Deutschen Ordens dar. Die zentrale Forschungsfrage thematisiert, inwiefern sich diese internen und externen Dynamiken über die künstlerischen Ausstattungsmerkmale in den Residenzen ausdrücken.

Für die Beantwortung dieser Frage ist die Rekonstruktion unterschiedlicher Zeitpunkte bzw. Zustände notwendig, die nur durch ihren Bezug entweder auf Quellen oder auf fundierte Annahmen nachvollzogen werden kann. Erst durch die Differenzierung der ursprünglichen Ausstattung und späteren Veränderungen (meistens durch unterschiedliche Auftraggeber) können die Ausstattungsmerkmale kontextualisiert werden. Der Ausgangspunkt sind die erhaltenen materiellen Zeugnisse (Architekturobjekte, Ausstattungen und Quellen). Doch basiert die Erschließung von Objekten und ihren vergangenen Zuständen zum Großteil auf wissenschaftlichen Annahmen. Damit diese nachvollziehbar sind, muss neben der inhaltlichen Aussage auch dokumentiert werden, von wem sie wann in welchem Kontext getroffen wurden. Für das Datenmodell ist demnach die Ausrichtung auf die wissenschaftliche Arbeit als Zuschreibung zentral.

Auf der Ebene der semantischen Datenmodellierung mit CIDOC CRM wird nahegelegt, den einzelnen Gegenstand möglichst genau zu beschreiben, um die Spezifizität des Sachverhalts treffend anzusprechen und diesen von anderen besser abzugrenzen. Je tiefer die Klasse demnach in der Hierarchie Ontologie steht, desto zahlreicher und spezifischer werden die zur Verfügung stehenden Eigenschaften. Für den Vorgang der wissenschaftlichen Tätigkeit bietet sich die Abbildung durch ein Zuschreibungsereignis (E13 Attribute Assignment) an. Eine solche Zuschreibung kann in Bezug auf jeden beliebigen Themengegenstand ausgeführt werden, weshalb die Klasse E13 Attribute Assignment über die Property P141 assigned mit jeder Klasse in CIDOC CRM verknüpft werden kann.

Zuschreibungen können Aussagen über das Bauwerk (oder dessen Ausstattung) treffen und sich dabei auf bestimmte Quellen und andere Zuschreibungen stützen. Quellen werden als Instanzen der Klasse E73 Information Object, als immaterielles Objekt, abgebildet. Diese Klasse erlaubt eine weitere Auslegung dessen, was als Quelle gelten kann, und ermöglicht so das Abbilden von vielfältigem Quellenmaterial.

Die Architektur wird als materieller Gegenstand begriffen, den es in seiner Erscheinung zu erfassen gilt. Die hierfür geeignete Klasse ist E22 Human-Made Object, da sie jene physischen Objekte („physical objects") umschreibt, die absichtlich von Menschenhand geschaffen wurden. Ihre Oberklasse E24 Physical Human-Made Thing beschreibt im abstrakteren Sinne „all persistent physical items", die durch Menschenhand geschaffen wurden. Architektur und ihre Teile werden durch die Klasse E22 Human-Made Object treffender beschrieben.

Für die Modellierung der Ausstattung wurde die Klasse E25 Human-Made Feature gewählt, da diese als Unterklasse von E26 Physical Feature die Eigenschaft hat, wesentlich mit physischen Objekten in Verbindung zu stehen („physically attached in an integral way to particular physical objects"). Hierdurch wird die Auslegung erlaubt, dass ein solches Feature nicht von seinem Trägerobjekt gelöst werden kann, bzw. ein Teil des übergeordneten Objekts (Architektur als E22 Human-made Object) dieses gänzlich 'trägt'. Der Fall eines Freskos veranschaulicht diese Verbindung besonders gut. Dass jedoch die Klasse E25 Human-Made Feature gewählt wurde, hängt mit der beabsichtigten Schaffung durch Menschenhand zusammen, die kein Kriterium eines E26 Physical Feature ist, jedoch in unserem Zusammenhang eine elementare, zu untersuchende Eigenschaft darstellt.

Diese variablen Verknüpfungen zwischen den zu untersuchenden Objekten, den Ausstattungen, den Quellen und den jeweiligen Zuschreibungen bilden ein komplexes, ineinandergreifendes Wissensnetzwerk, das den interpretierenden Ansätzen eine zentrale Position einräumt.

Umgang mit Quellen in 3D-Rekonstruktionen

Digitale 3D-Rekonstruktionen nicht mehr vorhandener Bauten bzw. früherer Zustände von historischen Gebäuden werden seit langer Zeit in unterschiedlichen Kontexten, z.B. in Ausstellungen oder zum Zweck der wissenschaftlichen Forschung eingesetzt. (Messemer 2016) Deren Rekonstruktionen beruhen auf den hypothetischen Annahmen, die vom Bearbei-

ter mithilfe von unterschiedlichen Quellen abgeleitet werden müssen: Fotografien, Entwurfs- und Bauzeichnungen, zeitgenössischen Beschreibungen uvm. (Bruseker/Guillem/Carboni 2015: 33) Die Fokussierung auf das 3D-Modell als Endprodukt führt häufig zur nicht oder nur unzureichenden Dokumentation der Quellen. Das 3D-Modell ist damit schnell der Kritik fehlender Wissenschaftlichkeit ausgesetzt. (Kuroczyński 2018: 162)

Die Erfassung der dem 3D-Modell zugrundeliegenden Quellen sollte möglichst nachvollziehbar und transparent sein. Eine Möglichkeit wäre die Quellen direkt mit dem 3D-Modell zu verbinden. Der vorliegende Vorschlag stellt jedoch nicht das Modell, sondern den Prozess der Rekonstruktion in den Mittelpunkt. Diese Aktivität wird als Instanz der Klasse E7 Activity aufgefasst. Durch die Beschreibung als Aktivität wird auch der kreativ-interpretative Anteil des Bearbeiters deutlich. Das 3D-Modell (Instanz der Klasse E73 Information Object) wird in der jeweiligen Aktivität implizit, als ihr Produkt angesprochen. Dadurch wird die Quelle nicht direkt mit dem 3D-Modell verknüpft, sondern der Realität entsprechend in Beziehung zu der Aktivität gesetzt, die zu diesem Modell führt. Informationen zu den jeweiligen Quellen (Instanzen der Klasse E31 Document) werden in einem eigenen Formular erfasst. Dieses Datenmodell bietet die Möglichkeit nachvollziehbar zu beschreiben, wann das 3D-Modell quellenbasiert ist und wann es ein nur auf Hypothesen basierendes, fiktives Produkt ist.

Fazit

Auf Unschärfen und Unsicherheiten von Informationen und Aussagen reagieren beide Projekte, indem sie neben den verwendeten Quellen auch die wissenschaftliche Auseinandersetzung bzw. die interpretative Tätigkeit des Modelleurs dokumentieren. Dieser Aspekt, der in der traditionellen geisteswissenschaftlichen, linearen Textarbeit oftmals 'nur' in den Fußnoten Platz findet, rückt als elementarer Bestandteil wissenschaftlicher Arbeit weiter ins Zentrum und erfährt durch die netzwerkartige Datenstruktur eine gleichrangige Behandlung im Kontext von (visuellen) Rekonstruktionen von Wissensbeständen.

Mit der geschaffenen Struktur könnte man nach Diskussion geeigneter Kriterien für die qualitative und quantitative Auswertung Aussagen zur Plausibilität von Quellen, Zuschreibungen und quellenbasierten 3D-Modellen treffen, die automatisiert abgefragt werden können.

Bibliographie

Bruseker, G. / Guillem, A. / Carboni, N. (2015): „Semantically Documenting Virtual Reconstruction: Building a Path to Knowledge Provenance", in: *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci II-5/W3: 33-40.*

CIDOC CRM Version 6.2.7, siehe: http://www.cidoc-crm.org/Version/version-6.2.7 (zuletzt aufgerufen am 13.12.2019)

Konter, Erich (2011): Deutsche Residenzen. Zur Sozialgeschichte der Repräsentation herrschaftlicher Ansprüche. Berlin: Edition Sigma.

Kuroczyński, Piotr (2018): „Neuer Forschungsraum für die Kunstgeschichte. Virtuelle Forschungsumgebungen für 3D-Rekonstruktionen", in *: Computing Art Reader. Einführung in die digitale Kunstgeschichte (=Computing in art and architecture 1). Heidelberg, 160-180.*

Messemer, Heike (2016): „The beginnings of digital visualization of historical architecture in the academic field", in: *Hoppe, Stephan / Breitling, Stefan (eds.): Virtual Palaces. Lost palaces and their afterlife. Virtual reconstruction between science and media. München, 21-54.*

Trentin-Meyer, Maike (ed.) (2004): Deutscher Orden 1190-2000. Ein Führer durch das Deutschordensmuseum in Bad Mergentheim. Baunach: Spurbuchverlag.

Normdaten der Faktenanker für Qualität im semantischen Retrieval. Der Ausbau der Gemeinsamen Normdatei (GND) im Projekt GND für Kulturdaten (GND4C).

Rosenkötter, Martha
rosenkoe@fotomarburg.de
Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg, Marburg

Fischer, Barbara
b.k.fischer@dnb.de
Deutsche Nationalbibliothek Arbeitsstelle für Standardisierung, Leipzig

Mit zunehmender Präsenz von Museen, Archiven, Forschungs- und anderen Kulturgut verwahrender Einrichtungen im Internet, steigt auch die Nachfrage nach verlässlichen Möglichkeiten spartenübergreifender Vernetzung. Sei es zur Verdichtung von Informationsgehalten oder zur Anregung neuer, interdisziplinärer Diskurse zu Sammlungs- und Forschungsobjekten. Um eine semantisch korrekte Verknüpfung und Auffindbarkeit von Informationen spartenübergreifend zu garantieren sind gemeinsam verwendete Normdaten unverzichtbar. Um diese Indentifizierbarkeit zu garantieren, sind Normdaten (siehe Abbildung 1) eindeutig, persistent und begriffsnormierend. So werden Fakten zu Bestands- oder Forschungsdaten zum Anker für ein verlässliches, semantisches Retrieval in Kulturportalen wie der Deutschen Digitalen Bibliothek oder Europeana, aber auch in den Datenbanken von Institutionen. Darum gehört der Einsatz von Normdaten und kontrollierten Vokabularen für eine verbesserte Auffindbarkeit, Vernetzung und Nachnutzbarkeit von Bestands- oder Forschungsdaten längst zur digitalen Dokumentation und somit unweigerlich in die Arbeitsbereiche der Forschungs- und Kultureinrichtungen.

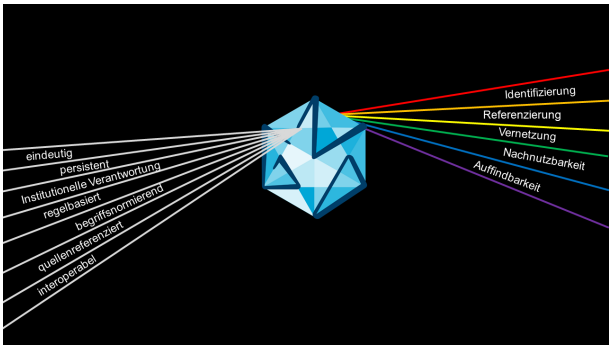


Abbildung 1: Definition und Leistungsspektrum der GND, Credit: Martha Rosenkötter, CC-BY-SA

Doch Faktenlagen können überworfen und durch neue wissenschaftliche Erkenntnisse ersetzt werden. Wer kümmert sich um deren quellenreferenzierte Anpassung innerhalb einer Normdatei bei der Fülle an unterschiedlichen, fachspezifischen Nutzerkreisen und übernimmt die Verantwortung für die neu eingebrachten Inhalte? Welche Eigenschaften sind überhaupt zwingend notwendig um eine Normdatei zu erstellen und welche sollten zusätzlich über die reine Identifikation eines Begriffes oder Objektes hinaus angeben werden?

Eine dieser Normdatein ist die als allgemeiner Datenhub anerkannte Gemeinsame Normdatei (GND) der Deutschen Nationalbibliothek und ihrer Partner in der GND-Kooperative. Historisch bedingt ist sie ausgelegt auf die Bedarfe der Bibliotheken und deckt somit bislang nicht im erforderlichen Maße die Anforderungen der Forschungs- und Kultureinrichtungen ab. Um diesem Problem entgegen zu treten, hat sich die Deutsche Nationalbibliothek in ihrem GND Entwicklungsprogramm 2017-2021 (Kett 2017: 2) zum Ziel genommen, die GND als Rückgrat eines maschinenlesbaren, semantischen Netzes der Kultur und Wissenschaft auszubauen. Für die GND bedeutet dies, sich für Blickwinkel aller Kultursparten zu öffnen und Elemente aufzunehmen, für die sie bisher als bibliothekarisches Werkzeug nicht gemacht war, die aber in anderen Sparten benötigt werden.

Das DFG-geförderten Projekt *GND für Kulturdaten (GND4C)*[1] treibt diese Entwicklung voran. Im Rahmen des Projekts wurden die im Fokus befindlichen Entitäten[2] (Sachbegriffe, Personen, Bauwerke, Orte) auf die Anforderungen aus Sicht der Museen und anderer Kultursparten in Fallbeispielen analysiert. Die gewonnenen Erkenntnisse fließen in die Weiterentwicklung des GND-Datenmodells und in die Verbesserung von Schnittstellen und Werkzeugen zur Unterstützung nicht-bibliothekarischer Anwendungskontexte. Kultureinrichtungen sollen in die Lage versetzt werden nicht nur Normdaten über Schnittstellen anbinden, sondern selbst Normdatenverbindungen, als selbstverständlichen Teil ihres Arbeitsalltags, erstellen zu können. Doch dazu braucht es mehr als nur einen technischen Grundstock an Werkzeugen, um dem spartenübergreifenden Anspruch gerecht zu werden. Es braucht solide Organisationsstrukturen, ein weites Netzwerk zur Informationsvermittlung und die Bereitschaft die Verantwortung (Kett u.a. 2019: 86) für diese Daten zu übernehmen. Nur so kann garantiert werden, dass der Ausbau der GND zu einem Erfolgsprojekt wird, dass verstetigt werden kann. Gerade mit Blick auf die wachsende Selbstverständlichkeit der Verwendung digitaler Technologien in den Geisteswissenschaften ist der souveräne Umgang mit Normdaten und ihrer Ergänzung von großer Bedeutung für den gesamten Bereich der Digital Humanities.

In einer Posterpräsentation möchten wir den aktuellen Stand des Projektes anhand von drei Postern vorstellen. Ausgangspunkt wird eine allgemeine Definition von Normdaten im Projektkontext, sowie deren Leistungsspektrum für Anwenderkreise sein. Einen Überblick über die bereits eingeflossenen Anforderungen aus den Communities an das zu erweiternde Datenmodell der GND soll anhand von Kernaussagen sowie Pluseigenschaften am Beispiel von Personen und Bauwerken erläutert werden. Das letztes Poster skizziert die Anpassungsprozesse durch neue Anforderung an das Datenmodell.

Fußnoten

1. Weitere Informationen zum Projekt: https://wiki.dnb.de/pages/viewpage.action?pageId=134055796, [letzter Zugriff: 25.09.2019].
2. Als Entität (auch Informationsobjekt genannt, englisch „entity“) wird in der Datenmodellierung ein eindeutig zu bestimmendes Objekt bezeichnet, über das Informationen gespeichert oder verarbeitet werden sollen.

Bibliographie

Kett, Jürgen / GND-Kooperative (2017): *Initiative für Normdaten und Vernetzung: GND-Entwicklungsprogramm 2017-2021 (Stand 06/2017)*, Deutsche Nationalbibliothek: 2 https://wiki.dnb.de/download/attachments/132749726/GND_Entwicklungsprogramm-m17-21_2017-06.pdf, [letzter Zugriff: 25.09.2019].

Kett, Jürgen u.a. (2019): Das Projekt "GND für Kulturdaten" (GND4C), in: *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, 6, Heft 4, 2019: 59–97 10.5282/o-bib/2019H4S59-97.

Opaque – digitale Arbeitsumgebung für die Humanities

Schlicht, Helene
helene.schlicht@uni-bielefeld.de
Univerität Bielefeld - SFB 1288 Praktiken des Vergleichens, Deutschland

Jentsch, Patrick
p.jentsch@uni-bielefeld.de
Univerität Bielefeld - SFB 1288 Praktiken des Vergleichens, Deutschland

Porada, Stephan
sporada@uni-bielefeld.de
Univerität Bielefeld - SFB 1288 Praktiken des Vergleichens, Deutschland

Opaque – digitale Arbeitsumgebung für die Humanities

Im Rahmen der DHd 2020 Spielräume möchten wir unsere in aktiver Entwicklung befindliche Webanwendung Opaque vorstellen. Anspruch ist es, Opaque als Arbeitsumgebung für DH-Projekte zu etablieren. Die Entwicklung der Webanwendung, deren Funktionen sukzessive erweitert werden sollen, wird im Rahmen des DFG-geförderten Sonderforschungsbereichs (SFB) 1288 "Praktiken des Vergleichens" im Teilprojekt (TP) INF "Dateninfrastruktur und Digital Humanities" durchgeführt.

Das TP INF betreut das Forschungsdatenmanagement des SFB und unterstützt dessen Wissenschaftler*innen darüber hinaus bei der Planung, Konzeptionierung und Durchführung von Forschungsprojekten unter Zuhilfenahme digitaler Methoden. Diese beiden Bereiche sollen in Opaque synthetisiert werden. Aufbauend auf den Erfahrungen der Kooperationen entwickeln wir Opaque zur Bündelung und Automatisierung der erprobten Workflows und Best Practices. Eine besondere Schwierigkeit ist hierbei die Heterogenität und Komplexität von Forschungsdaten in den Geisteswissenschaften. Um dieser Schwierigkeit zu begegnen orientiert sich unsere Etablierung von Best Practices an den verschiedenen Stadien des Data Life Cycle, bestehend aus Planung/Beratung, Sammlung, Datenorganisation, Datenanalyse, Dissemination und Nachnutzung, und hat zum Ziel, für alle diese Stadien Best Practices zu entwickeln oder implementieren und so den Forscher*innen verfügbar zu machen. Die einzelnen in Opaque verfügbaren Funktionen werden durch etablierte Open Source-Lösungen realisiert, die durch die modulare Konstruktion der Webanwendung nicht nur gut erweitert sondern auch beständig auf dem neuesten Stand gehalten werden können, sowie reproduzierbare Routinen gewährleisten. Der Fokus auf Nachnutzung bestehender Software ermöglicht es uns, ein breites Spektrum an Funktionalitäten in Opaque zu integrieren.

Opaque: Die Webanwendung

Opaque bündelt verschiedene Werkzeuge und Services, die Geisteswissenschaftler*innen Methoden der DH an die Hand geben und somit deren verschiedene individuelle Forschungsprozesse unterstützen können. Mittels Opaque können Forschende digitalisiert vorliegende Quellen einer *Optical Character Recognition* (OCR) unterziehen. Die daraus resultierenden Textdateien können anschließend als Datengrundlage zum *Natural Language Processing* (NLP) weiterverwendet werden. Die Texte werden hierbei automatisiert verschiedenen linguistischen Annotationen unterzogen. Die via NLP prozessierten Daten können in der Webanwendung anschließend als Corpora zusammengefasst und mittels eines *Information Retrieval System* durch komplexe Suchanfragen analysiert werden. Der Funktionsumfang der Webanwendung wird zudem anhand der Bedarfe der Forschenden sukzessive erweitert.

Die Funktionsschwerpunkte von Opaque unterscheiden sich von anderen deutschen DH-Softwareentwicklungen. Hervorzuheben sind *TextGrid*, *FuD* und *CQPweb*, die einen ähnlichen Anspruch als virtuelle Forschungsumgebung verfolgen. Im Unterschied zu Opaque legen *TextGrid* und *FuD* ihre Schwerpunkte auf händische Datenaufbereitung und nachhaltige Speicherung via integrierter Publikationsplattformen, wohingegen *CQPweb* ein Werkzeug zur Korpusanalyse darstellt, dessen Query Processor in Opaque übernommen wurde. Opaque soll demgegenüber keine Publikationsplattform integrieren, sondern eine automatisierte Aufbereitung und Informationsanreicherung von Forschungsdaten mit anschließender Analyse ermöglichen. Die aufbereiteten Daten und Analyseergebnisse können mittels Exportfunktionen anhand gängiger Standards in offene Dateiformate exportiert und anschließend auf eigens gewählten Publikationsplattformen veröffentlicht werden. Die bereits in Opaque integrierten und beständig auf dem neuesten Stand gehaltenen Funktionen im Bereich des NLP und der OCR grenzen die Plattform von den genannten bestehenden Lösungen ab.[1, 2]

Da Opaque plattformunabhängig konzipiert ist, können die verschiedenen Funktionen von den Wissenschaftler*innen auf beliebigen Endgeräten ohne vorangehende Einrichtung genutzt werden. Alle Funktionen wie z.B. OCR werden innerhalb der Cloud-Infrastruktur ausgeführt, so dass Nutzer*innen selbst keine leistungsfähigen Endgeräte benötigen.

Nutzerorientiertes Design

Die in Opaque implementierten Funktionen und Workflows orientieren sich an den aus unserer Zusammenarbeit im SFB hervorgegangenen Erfahrungen, etablierter Best Practices sowie Vorgaben und Standards des Forschungsdatenmanagements.

Dies führt nicht nur zu besseren Ergebnissen für die Forscher*innen, sondern auch zu einer besseren Datenorganisation mittels anerkannter Standards.

Durch eine Gegenüberstellung soll auf dem Poster anhand der verschiedenen Stadien des Data Life Cycle veranschaulicht werden, wie sich Arbeitsprozesse und -schritte durch die Einführung von Opaque verändert haben. Prägnante Beispiele für diese Gegenüberstellung sind Datensammlung und Datenanalyse. Mit Hilfe der Webanwendung können Forscher*innen eigene Quellen und Texte einem OCR-Prozess unterziehen und die Ergebnisse zeitnah selbstständig hinsichtlich der Güte der Texterkennung evaluieren. Diese Automatisierung der Prozesse in Verbindung mit der intuitiven Bedienoberfläche tragen zu einer erhöhten Autonomie der Forschenden bei. Gleichzeitig macht die Echtzeitverfolgung der Jobstatus die Prozessabläufe transparent und nachvollziehbar. Gespräche, die vorher technischer und organisatorischer Natur waren, können nun gezielter für inhaltliche Diskussionen und Planung der Forschung genutzt werden.

Bezüglich der Qualität der Eingabedateien (z.B. Scans) offerieren wir Hinweise zur bestmöglichen Digitalisierung von Ausgangsmaterialien und orientieren uns an gängigen Standards zur Speicherung und Veröffentlichung von Forschungsdaten (z.B. FAIR), um deren Nachnutzung zu gewährleisten. Dies schließt neben den Forschungsdaten auch die Nachhaltung und Bereitstellung von für den Forschungsprozess genutzter Software in den jeweils genutzten Versionen mit ein, um die Reproduzierbarkeit von Forschungsergebnissen sicherzustellen.

Implementierung

Die Umsetzung beruht auf *Free Open Source Software* und Python. Auf dem Poster werden die Vorteile von Linux Containern in einem skalierbaren Docker-Rechencluster, wie z.B. eine einfache Verwaltung verschiedener Softwareversionen – insbesondere wichtig um Forschungsdaten reproduzieren zu können –, vorgestellt und die einzelnen im Folgenden aufgeführten Module der Plattform näher beleuchtet.

- **Webanwendung**: Die Webanwendung dient als Schnittstelle zwischen Nutzer*innen und Recheninfrastruktur. Hier können Datenaufbereitungen in Form von Jobs gestartet und in Echtzeit verfolgt werden, dabei werden die Jobs automatisch auf das zugrundeliegende Rechencluster verteilt. Das Webinterface bietet außerdem die Möglichkeit über ein *Information Retrieval System* Auswertungen durchzuführen.
- **Daemon**: Agiert im Hintergrund, um die von den Nutzer*innen durch die Webanwendung abgesetzten Befehle und Services umzusetzen bzw. zu verwalten.
- **Datenbank**: Die Datenbank speichert alle Metadaten, die während der Nutzung der Webanwendung anfallen. Als Datenbanksystem wird *PostgreSQL* benutzt.
- **Netzwerkspeicher**: Speichert die von den Nutzer*innen hochgeladenen Dateien sowie die daraus generierten Resultate. Die Netzwerkspeicherlösung garantiert den Servern des Cloud-Rechenclusters gleichermaßen Zugriff auf die zu bearbeitenden Dateien.
- **Services**: OCR und NLP-Dienste werden mittels der state of the art Software *Tesseract OCR* und *spaCy* realisiert. Die Korpusanalyse erfolgt durch eine Anbindung an den *CQP query processor* der IMS Open Corpus Workbench. Jede Ausführung eines Dienstes ist mit einem Job assoziiert, der in einem eigens dafür erstellten Container bearbeitet wird.

Ein zusätzliches Hands-On von Opaque soll zu einem Erfahrungsaustausch einladen.

Fußnoten

1. Eintrag des offiziellen DARIAH-Wikis schildert, dass gängige Funktionen wie ein Lemmatisierer nicht mehr nachinstalliert werden können.
2. Der Abschlussbericht des Projekts TextGrid aus dem Jahr 2012 schildert die Implementierung einer OCR Funktion mittels OCRopus, welche in den aktuellen Versionen nicht mehr zu finden ist.

Orte in narrativen biographischen Interviews: automatische Methoden und manuelle Analysen

Ruppenhofer, Josef
ruppenhofer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Flinz, Carolina
carolina.flinz@unimi.it
Universität Mailand

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

Das sogenannte Israelkorpus ist ein Korpus gesprochener Sprache, das von Anne Betten und MitarbeiterInnen in den Jahren 1989 bis 2012 erstellt wurde und aus 274 Aufnahmen narrativer autobiographischer Interviews mit Emigranten aus deutschsprachigen Regionen Mitteleuropas besteht, die vorwiegend in den 1930er Jahren zur Auswanderung gezwungen wurden. Es besteht aus drei Subkorpora, die unter anderem in der Datenbank für Gesprochenes Deutsch (DGD), einem Korpusmanagementsystem des Leibniz-Instituts für Deutsche Sprache, abrufbar und recherchierbar sind: IS (Emigrantendeutsch in Israel), ISW (Emigrantendeutsch in Israel: Wiener in Jerusalem) und ISZ (Zweite Generation deutschsprachiger Migranten in Israel).

In unserem Beitrag untersuchen und vergleichen wir automatische und manuelle Zugänge zu Ortsnennungen in den biographischen Interviews des ISW-Korpus. Orte spielen im Israelkorpus eine besondere Rolle. Einerseits dienen sie als geographische Bestimmungen vor, während und nach der Emigration. Andererseits haben sie auch eine Funktion innerhalb der Erinnerungsarbeit während der Interviews, die sehr stark mit der emotionalen Dimension verbunden ist. Im Rahmen des Projekts *Orte und Erinnerung. Eine Kartographie des Israelkorpus* soll die emotionale Funktion von Ortsnennungen auf dem ganzen Israel-Korpus untersucht werden. Unsere Arbeit gehört daher in die Reihe neuerer Arbeiten, die mit korpuslinguistischen Methoden das gesamte Korpus untersuchen (Flinz 2019; Flinz/Brambilla 2019), während bisherige grammatische, syntaktisch-stilistische oder dialoglinguistische Untersuchungen am Israel-Korpus sich auf wenige qualitativ untersuchte Interviews beschränkten.

Um die Untersuchung der Ortsnennungen im Gesamtkorpus zu ermöglichen, müssen wir das relevante Ortskonzept so operationalisieren, dass computerlinguistische und NLP-Werkzeuge Ortsnennungen mit hoher Präzision und einer hohen Trefferquote (EN recall) auffinden und den menschlichen Analystinnen zur Valdierung und Interpretation vorlegen können. In unserer Pilotstudie evaluieren wir hier zunächst, wie

gut sich verfügbare Werkzeuge und Ressourcen 'out of the box' dafür eignen. Zu diesem Zweck haben wir das relevante Ortskonzept händisch durch Expertenannotation auf alle Transkripte des ISW-Teilkorpus angewendet. Parallel dazu haben wir die Daten einerseits mithilfe eines state-of-the art Named-Entity-Recognizers (Akbik et al 2018) annotiert, der unter anderem auch Orte ('Locations') auszeichnet, und andererseits programmatisch alle Wörter, die von GermaNet (Hamp & Feldweg 1997) dem Wortfeld ORT zugeordnet werden, als Ort annotiert.

In der Auswertung der parallelen Annotationen zeigen wir, dass das Ortskonzept mit guter Übereinstimmung händisch annotiert werden kann (kappa > 0.9), es aber weder durch die Annotationen des NER-Systems noch durch die Annotationen auf der Grundlage von GermaNet adäquat erfasst wird. Die NER-Annotationen decken nur Eigennamen ab (z.B. *Wien*, *Israel*), während für die Auswertung auch Bezeichnungen durch Appellativa (z.B. *Lager*, *Ausland*, *Grenze*) sehr wichtig sind. Die Erkennung von Ortsnennungen in Form von Appellativa mithilfe von GermaNet ist für unsere Forschungsfragen ebenfalls nicht ausreichend. Einerseits werden viele Konzepte, deren Bedeutung eine Ortsfacette (im Sinne von Cruse und Croft 2004) besitzen, von GermaNet nicht mit einer eigenen Ortsbedeutung ausgewiesen. Ein zentrales Beispiel hierfür ist *Schule*, welches in GermaNet den Wortfeldern GRUPPE, KOGNITION, GESCHEHEN und ARTEFAKT zugeordnet wird, aber nicht dem Wortfeld ORT. Andererseits besitzen viele relevante Wörter wie im Fall von *Schule* auch andere Bedeutungsfacetten und die räumliche ist im konkreten Kontext nicht unbedingt wichtig (z.B. im Fall der Erwähnung einer 'Schule' der Musikgeschichte).

Der Vergleich der händischen und automatischen Annotationen legt nahe, dass wir für die automatische Annotation aller Orte in den Israel-Korpora eine auf das Problem zugeschnittene Lösung brauchen. Die Annotationen eines auf unseren Expertenannotationen trainierten Systems werden wir mit den oben genannten out-of-the-box Annotationen vergleichen. Als weiteren Ansatz werden wir automtisch alle Wörter/Konzepte identifizieren, die in der Wikidata-Ressource (Vrandečić und Krötzsch 2014) einen Link zu OpenStreetMap besitzen und damit implizit als geographisch lokalisierbar ausgewiesen werden. Beispielsweise wird dadurch das Konzept *Schule* erfasst, das von GermaNet nicht als ORT ausgewiesen wird.

Neben praktischen Erkenntnissen hat uns der Vergleich der manuellen und automatischen Annotationen auch auf theoretische Fragen und Möglichkeiten aufmerksam gemacht, die wir vorher nicht Betracht gezogen hatten. So erwägen wir, die von uns im weiteren Arbeitsverlauf manuell korrigierten NER-Annotationen als separate Sicht auf die Daten in die korpuslinguistische Analyse der Beziehung zwischen Orten und Emotionen mit einzubeziehen.

Bibliographie

Akbik, Alan / Blythe, Duncan / Vollgraf, Roland (2018): "Contextual string embeddings for sequence labeling." *Proceedings of the 27th International Conference on Computational Linguistics*.

Brambilla, Marina / Flinz, Carolina (2019): Orte und entgegengesetzte Emotionen (Liebe und Hass) im Korpus ISW. In: Studi Germanici. Im Druck.

Croft, William, / Cruse, D. Alan (2004): *Cognitive linguistics*. Cambridge University Press.

Flinz, Carolina (2019): "Multiword units and N-Grams naming FEAR in the Israel-Corpus". Corpas Pastor, G. / Mitkov, R. Computational and Corpus-Based Phraseology. Springer Verlag. Lecture Notes in Computer Science. 86—98.

Hamp, Birgit / Feldweg, Helmut (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.

Vrandečić, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase". *Commun. ACM* 57, 10 (September 2014). 78-85.

Prosopographische Interoperabilität – Stand der Dinge

Vogeler, Georg

georg.vogeler@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz, Österreich; Austrian Center for Digital Humanities, Österreichische Akademie der Wissenschaften

Schlögl, Matthias

matthias.schloegl@oeaw.ac.at
Austrian Center for Digital Humanities, Österreichische Akademie der Wissenschaften

Vasold, Gunter

gunter.vasold@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz, Österreich

Prosopographical Interoperability – State of the Art

In einem wegweisenden Artikel zu Prosopographie aus dem Jahr 1971 schreibt Lawrence Stone (Stone, 1971) "The method employed is to establish a universe to be studied, and then to ask a set of uniform questions...". Dieses *Erstellen des Universums* ist einer der Reize der Prosopographie, stellt das Feld aber auch vor besondere Herausforderungen. Ein *Universum* besteht gemeinhin aus Millionen von Objekten und noch mehr Relationen zwischen diesen Objekten. Projekte, die sich mit prosopographischer Forschung beschäftigen, können oder würden deshalb von weiterverwendbaren Daten besonders profitieren. Ihr Daten *universum* dreht sich nicht nur um für ihre Forschung zentrale Daten (die oft neu erstellt oder überprüft werden), sondern auch um angebundene Daten: Geburtsorte von Personen, Orte in denen sich Institutionen befinden, Lehrer von Personen in der Kerngruppe etc.

Eine logische Konsequenz daraus ist es, zumindest für periphere Daten der eigenen Prosopographie LOD-Daten nachzunutzen. Um dies ohne großen Aufwand tun zu können, bräuchte es kompatible Ontologien/Datenmodelle. In den letzten Jahren wurden mehrere Versuche unternommen, für das Feld der Prosopographie einheitliche Datenmodelle vor-

zuschlagen (Fokkens and ter Braake, 2018; Tuominen et al., 2017), ohne dass es schon zu einem Konsens gekommen wäre.

Das Poster wird über den Stand einer Initiative berichten, die seit vergangenem Jahr an der Definition einer RESTful API arbeitet, welche die Veröffentlichung von maschinenlesbaren prosopographischen Daten so erleichtern soll, dass typische Anfragen performant und für Softwareentwickler einfach zu realisieren sind. Dieses "International Prosopography Interoperability Format" (IPIF) hat als Kern die Definition in einer RESTful API, die in OpenAPI beschrieben ist.[1]

Das dort vorgeschlagene Datenmodell deckt zum einen die Notwendigkeit ab, zwischen Person, Quelle und Quelleninterpretation zu unterscheiden ("Factoid"-Modell, Bradley & Short 2005), zum anderen vereinfacht es den Zugriff auf Informationen über eine Person in klassischen Benutzungssenzarien der Prosopographie. In einem "Statement" über eine Person können verbale Beschreibungen oder Quellenzitate ebenso wie strukturierte Informationen enthalten sein. Um die prosopographische Benutzung zu erleichtern, lassen sich die strukturierten Informationen im Modell von IPIF als datierbare Ereignisse verstehen, wenn sie mit einer Property "date" versehen sind. Sie können aber für reine Identifikationszwecke auch einfache Eigenschaften (Name, Geschlecht) abbilden. Properties "relatesToPerson" und "isMemberOf" bedienen ein drittes zentrales Szenario der Nutzung von prosopographischen Daten, nämlich Beziehungen zu anderen Personen. Schließlich sind Ortsangaben zu einer Person mit der Property "place" möglich. Mit diesen Angaben ermöglicht IPIF Anzeigen wie die des DARIAH-Cosmotools[2] und Ego-Netzwerke wie z.B. in der Deutschen Biographie[3].

Proof of concepts

Von Beginn an war eine Grundidee des Unterfangens, die Praxistauglichkeit des Datenmodells und der API Definition möglichst früh zu testen. Zu diesem Zweck wurden seit 2018 mehrere "Proof of Concept" Applikationen erstellt. Das Poster wird den aktuellen Entwicklungsstand dieser Proof of Concept Applikationen darstellen.

APIS (Austrian Prosopographical Information System)

APIS ist ein Web-basiertes System zur Arbeit an prosopographischen Daten (Schlögl/Lejtovicz 2018). Es bietet Webformulare, aber auch API-Schnittstellen zu den Daten. Aufbauend auf die schon vorhandenen APIs wurde ein Renderer erstellt, der vorhandene Daten in das IPIF Format überführt. Zusätzlich wurden als parallele API die IPIF endpoints implementiert und somit compliance level 1 laut API Definition (GET requests inklusive Filter) erreicht.

API Wrapper

Eines der Anwendungsszenarien von IPIF ist die Suche schon vorhandener Identifier über mehrere Referenz Ressourcen hinweg (vgl. Vogeler et al 2020 forthcoming). Um die Anwendbarkeit von IPIF für dieses Szenario zu testen, wurde eine einfache Applikation erstellt, die als Middle-Layer zwischen der IPIF API auf der einen Seite und dem Wikidata SPARQL Endpoint und der Lobid GND API auf der anderen Seite fungiert. Die Applikation übersetzt dabei Anfragen an die API in eine wikidata kompatible SPARQL query bzw. einen Lobid kompatiblen GET request und überführt die Antworten in das IPIF format. Die Applikation macht sich für diese Übersetzung die Django-Templating Engine zu Nutze und ist damit auch für andere APIs einfach konfigurierbar.

Papilotte

Auf der in OpenAPI veröffentlichten Spezifikation aufbauend lässt sich über Frameworks wie das von Zalando als Open Source Software bereitgestellte "Connexion" schnell ein Stand-Alone-Server in Python schreiben, der über flexible Konnektoren den Zugriff auf unterschiedlichste interne oder externe Datenquellen ermöglicht und diese IPIF-konform bereitstellt. Ein solcher Server wurde mit Beispieldaten aus dem Monasterium.net-Projekt erstellt.[4]

Papi-Cosmotool

Das DARIAH-Cosmotool bietet eine prototypische Oberfläche für eine prosopographische Datenbank an. Es enthält eine biographische ("Zeitleiste"), eine textuelle ("Ereignis-Detail") und ein geographische ("Kartendarstellung") Ansicht. Die Datenanzeige wird mit einem Quellenverweis ergänzt. Als Test für die Verwendbarkeit der API-Definition wird von Sebastian Stoff am Zentrum für Informationsmodellierung eine JavaScript basierte Anwendung erarbeitet, die vergleichbare Funktionalitäten bietet.[5]

JSON-LD

Schließlich arbeiten wir an einer Integration des JSON-Outputs der API-Definition in das Semantic Web. Dafür soll eine context.json-Datei bereitgestellt werden, die in den Resultsets der API-Anfragen gültige RDF-Aussagen identifiziert.[6]

Fußnoten

1. https://github.com/GVogeler/prosopogrAPhI
2. https://cosmotool.de.dariah.eu/
3. z.B. https://www.deutsche-biographie.de/graph?id=sfz53095
4. https://github.com/gvasold/papilotte, mit einer Beispielinstallation: https://ginko.uni-graz.at/illurk/api/ui/
5. http://glossa.uni-graz.at/gamsdev/stoffse/erla/mapp/map/, Code von Sebastian Stoff.
6. Ein erster Entwurf ist unter https://github.com/GVogeler/prosopogrAPhI/blob/master/context.json einsehbar.

Bibliographie

Bradley, John / Short, Harold (2005): "Texts into databases. The Evolving field of New-style Prosopography." in: *LLC* 20, suppl. 1: 3-24.

Fokkens, A. / ter Braake, S. (2018): Connecting People Across Borders: a Repository for Biographical Data Models, in:

Proceedings of the Second Conference on Biographical Data in a Digital World 2017. Linz, Austria, November 6-7, 2017. CEUR Workshop Proceedings: 83–92.

Stone, Lawrence (1971): "Prosopography." in: *Daedalus* 100: 46–79.

Schlögl, Matthias / Katalin Lejtovicz . (2018). "A Prosopographical Information System (APIS)." In: Antske Fokkens, ter Braake, Serge, Sluijter, Ronald, Arthur, Paul, and 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.

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* , ed. by Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt, Budapest: CEUR (CEUR Workshop Proceedings 2119): 59-66.

Vogeler, Georg / Vasold, Gunter / Schlögl, Matthias (2020 forthcoming): Data exchange in practice: Towards a prosopographical API. In: BD2019, Workshop on Biographical Data in Occasion of the RANLP 2019, Varna, ed. by Antske Fokkens et al.

Requirements on the Punctuation Reconstruction for the Translation of Post-modern Poetry

Meyer-Sickendiek, Burkhard
bumesi@zedat.fu-berlin.de
Freie Universität Berlin, Deutschland

Baumann, Timo
baumann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Hussein, Hussein
hussein@zedat.fu-berlin.de
Freie Universität Berlin, Deutschland

Punctuation is an important and cohesive device in all kinds of written discourse. Standard marks used to separate words, phrases, clauses and sentences for the purpose of cohesion. Already [2][5][1] pointed out that through punctuation marks, one can signal different information structures in written language. Regarding the translation of texts, we use such marks to identify the ends of sentences, closely related sentences or clauses, etc. This is why missing punctuation burdens the translations and forces the translator to go over the text several times to understand its meaning [10]. Understanding the uses and functions of punctuation marks, therefore, is extremely important for translators, as their purpose is to clarify the meaning of a particular construction within a text. On the other hand, modern poetry often disregarded such punctuations. Ever since Italian Futurism around 1900 spoke of the 'parole in libertà', i.e. the liberation of words from grammatical and syntactic limitations, modern poetry has hardly used punctuation. This lack of punctuation makes analysis, but also translation, more difficult. The only way to reconstruct this punctuation is by listening to the poems, i.e. by subsequently identifying sentence boundaries. However, this lack of punctuation can be found very often in modern and post-modern poetry, so the challenge is to recognize the phrase boundaries. We contribute in the paper an application towards the problem of identifying left-out punctuation in post-modern poetry, by proving that only a very simple type of punctuation - the semicolon - is needed to improve machine translation. This simple punctuation refers to phrase boundaries, the so-called "grammetrical units", which Donald Wesling defined in his study "The Scissors of Meter" [11]. Such units must be identified in order to improve machine translation.

The need for adding left-out punctuation becomes in case of creating machine translations obvious with regards to the poem "bitte verlassen sie diesen raum" (english: please leave this room) written by the German poet Nicolai Kobus [6] (Text A):

bitte verlassen sie diesen raum
so wie sie ihn vorfinden möchten
danke möchten sie diesen raum
vorfinden wie sie ihn verlassen
haben bitte räumen sie alles so
vorgefundene als wären sie
verlassen worden danke sie
möchten doch nicht daß man
sie so verlassen im raum vor
findet bitte seien sie für einen so
verlassen vorgefundenen raum
dankbar [...]

The challenge for the interpretation of this poem lies in the adequate identification of the line endings. These endings can only be identified correctly by listening to the poet's reading, which is possible because we got the audio version on the *lyrikline* [7] (the world's largest corpus of spoken (post-) modern poetry which also features translations for many of the poems) webpage. This is the reason, why the manual translation, made by Catherine Hales, is able to translate these endings in a correct manner (Text B):

please leave this room
in the state in which you would like
to find it thank you would you like
to find this room in the state in which
you have left it please clear out
everything thus found as though you
had been left thank you you would not
like somebody to find you left
abandoned in the room now
would you please be grateful for
a room a space found in such
an abandoned state (...)

In the human translation or the target poem, made by Hales, there is just a little difference. This difference is caused by the missing punctuation. And it can basically be explained

by the fact that Hales has chosen a different line arrangement. In terms of content, however, her translation is reproduced correctly. Since there is no specific translation system trained with poem data with/without punctuation (small amounts of training data), we used a Google machine translation (GMT) system [3]. When we compare this (human) translation with the GMT system, we recognize the difficulty of recognizing the sentence boundaries within the poem without punctuation (Text C):

please leave this room
as they would like to find him
Thank you for wanting this room
find out how to leave him
please have everything clear
found as if they were
Thank you
you do not want that one
So leave them in the room
please find one for you
leave found space (...)

Obviously, this machine translation (MT) becomes much better if we add the full punctuation marks to the source text, when listening to the audio of the poem (Text D):

please leave this room
as you would like him to find
Thank you. Do you want this room
find how they leave him
to have? Please clear everything up
found as if they were
been left. thank you
Do not want that one
So leave them in the room
please, please be for one
leave found space
grateful. (...)

Punctuation is an essential aspect of poetry translations, as it is for discourse analysis in general [8]. Punctuation "gives a semantic indication of the relationship between sentences and clauses, which may vary according to languages", as well as to translations [4].

A first step towards solving the problem of translation unpunctuated texts is the correct localization of the missing punctuation within such sentences and clauses. In the Google translation, which was completely without punctuation, we see that Google system translated every single line anew (Text C), ignoring the line-arrangement and the "enjambments", when one phrase continues beyond the line, or continues from the previous line. This explains the translation error in the third line: Reading the line as a full sentence disregarding its character as an enjambment, the translation produces a full sentence (Thank you for wanting this room), which does not fit to the original (... danke. möchten sie diesen raum ...). However, this translation error will be improved if we add the missing punctuation to the machine translation, which could be identified as Text D.

It is hard to translate automatically without having information about the sentence boundaries and the punctuation as a discourse unit for meaning demarcation. But to what extent punctuation information has to be recovered for the translation of post-modern poetry? Which kind of information do we need to improve machine translation? Do the questions have to be distinguished from the statements? Or is the simple marking of phrase boundaries already sufficient? To answer these questions, we analysed unpunctuated German poems. There are 234 german-speaking poets on the *lyrikline* webpage reading a total of 2591 poems. A total of 733 German poems are translated to English which are used in this work. There are 98 German poems which do not contain any punctuation information. We analysed 120 poems in this work with a maximal punctuation information ratio of 0.05%. This process yields a total of 2924 lines out of which only 28 (0.009%) with punctuation information.

The philological scholar of our project annotated the punctuation information manually by using text and audio information in the 120 poems, focusing on the intonation of poets reading their poems. In order to clarify the question which type of punctuation has to be added, we inserted two kinds of punctuation in the source text. In a first step, we focused on six different punctuation marks: full stop (.), comma (,), semicolon (;), colon (:), exclamation mark (!), and question mark (?). In a second step, we simplified this insertion by reducing these six marks to a single semicolon.

The human reference translations are compared with the automatic translation of GMT system without/with consideration of punctuation information. The experiment consists of three tasks based on the GMT system:

- Task 1: Standard translations of original poems (without punctuation).
- Task 2: Translations with one level of punctuation information: replacement of all manually annotated punctuation information by one level of punctuation (;).
- Task 3: Translations with six punctuation information: consideration of the six manually annotated punctuation information (.,;:!?).

The translation enhancement should be observable from improved translation quality scores. The results are calculated by bilingual evaluation understudy (BLEU) [9] score, which used for evaluating the quality of text by translation. The BLEU score of tasks 1, 2, and 3 are 0.256, 0.275, and 0.280, respectively. The results indicate that we need just one type of punctuation - semicolon - to improve the scoring for automatic translations of post-modern poetry.

Every generic translation system is trained with data in which segments are defined by end points. It is astonishing that even the addition of a semicolon to segmental boundaries is sufficient to improve machine translation. This also explains the central problem: machine translation does not fail because of mixing up questions and statements, but because of mixing up segmental units and enjambements.

In our future work, we plan to train a specific system on translating unpunctuated poetry in order to compare the results with manual translations. The fact that we add punctuation signs on the basis of oral representations of the poems is acceptable when it comes to audio poems, in which the oral representation is an essential part of the poem as a piece of art, closely connected to the written form.

Bibliography

[1] Baker, M. (1994): In Other Words: A Course Book on Translation. London, New York: Routledge.

[2] Halliday, M. A. K. (1985): An Introduction to Functional Grammar. London: Edward Arnold.

[3] Han, S.: Free Google Translate API for Python. Available on https://pypi.org/project/googletrans/. Last accessed at 15. August 2019.

[4] Hosseini-Maasoum, S. M. / Mahdiyan, M. (2012): Punctuation in Translation: The Unseen Side of the Coin. Mediterranean journal of social sciences, 3(11):25–32.

[5] Kirkman, J. (2006): Punctuation Matters: Advice on Punctuation for Scientific and Technical Writing. Routledge study guides. Routledge.

[6] Kobus, N. (2006): Hard cover: Gedichte. Ardey Verlag, Münster.

[7] Lyrikline Literaturwerkstatt Berlin: Lyrikline: listen to the poet. Available on www.lyrikline.org. Last accessed at 03. September 2019.

[8] Newmark, P. (1988): A Textbook of Translation. Prentice Hall.

[9] Papineni, K. / Roukos, S. / Ward, T. / Zhu, W-J (2002): BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[10] Shiyab, S. M. (2017): Translation: Concepts and Critical Issues. Garant Publishers.

[11] Wesling, D. (1996): The Scissors of Meter: Grammetrics and Reading. University of Michigan Press.

„Romantik" im aktuellen parteipolitischen Diskurs auf Twitter

Duan, Tinghui

tinghui.duan@uni-jena.de
DFG Graduierten Kolleg "Modell Romantik"; Jena University Language & Information Engineering Lab (JULIE Lab)

Buechel, Sven

sven.buechel@uni-jena.de
Jena University Language & Information Engineering Lab (JULIE Lab)

Hahn, Udo

udo.hahn@uni-jena.de
Jena University Language & Information Engineering Lab (JULIE Lab)

Die vorliegende Arbeit versucht, im Rahmen einer empirisch fundierten Diskursanalyse von Texten sozialer Medien eine Brücke zwischen qualitativ-hermeneutischer Kulturwissenschaft (hier: Literatur- und Politikwissenschaft) und quantitativ-komputationeller digitaler Geisteswissenschaft zu bauen und beide Methodenlinien synergetisch miteinander zu verschränken. In diesem erweiterten Abstract beschreiben wir einen neuen Datensatz von Twitter-Beiträgen deutscher Parlamentarier des 19. Deutschen Bundestags als Datengrundlage der Diskursanalyse und erste Teilergebnisse, die aus der Analyse dieses Datensatzes resultieren. Ein Fixpunkt dieses Vorgehens ist das historisch markierte Epochenkonstrukt der Romantik in seiner literarischen und sozialen Ausformung (Lebensform, Wertekanon usw.) und seine (Wieder-)Aufnahme bzw. Adaption im aktuellen parteipolitischen Diskurs in Deutschland.

Ausgangspunkt unserer Arbeiten waren Beobachtungen, die einen Bezug zwischen rechtspopulistischen Parteien und Symbolen der deutschen Romantik nahelegten. Während der AfD-Politiker Björn Höcke von seinem Parteikollegen beispielsweise als „romantischer Nationalist" bezeichnet wurde, trug sein Parteigenosse Andreas Wild bei einem Auftritt im Bundestag eine blaue Kornblume an seinem Revers. Diese Blume, ein zentrales Symbol der Romantik, wurde in den 1930er Jahren sogar zu einem Erkennungszeichen der illegalen Nationalsozialisten in Österreich. Die semantische Doppelbesetzung der blauen Kornblume eröffnet folglich rechtspopulistischen Politikern einen diskursiven Spielraum, sich einerseits implizit an den Nationalsozialismus anzulehnen, andererseits diese Identifikation in der Öffentlichkeit nicht eindeutig zum Ausdruck bringen zu müssen.

Um diese Einzelbeobachtungen systematischer einordnen und die Hypothese von der auffälligen Verwendung von Konzepten der Romantik-Epoche im Diskursverhalten einer rechtspopulistischen Partei einer strengeren Prüfung unterziehen zu können, entwickelten wir ein Korpus von Twitter-Beiträgen aller Abgeordneten des (aktuellen) 19. Bundestags (es kann damit als Ergänzung der Redenkorpora des Bundestags von Barbaresi (2018) bzw. Blätte & Blessing (2018) betrachtet werden, die aber auch frühere Legislaturperioden umfassen). Dieses Korpus sollte Grundlage für eine computerlinguistische Diskursanalyse zur Prüfung der Hypothese sein (einen ähnlichen Studienansatz zur Überprüfung sprachlich markierter Stereotypen zwischen politischen Parteien beschreiben Sylwester & Purver (2015)).

Korpus: Für unsere Untersuchung haben wir DeBAC (*De*utscher *B*undestags*A*bgeordnete-*C*orpus), das nach unserem Kenntnisstand erste Twitter-Korpus deutscher Bundestags-abgeordneter für die laufende 19. Legislaturperiode, aufgebaut. Es umfasst zum Zeitpunkt der Abfassung dieses Abstracts (Januar 2020) 887.008 Tweets von 478 Parlamentariern über einen Zeitraum vom 21.11.2008 bis 2.1.2020; dieses Korpus wird fortlaufend aktualisiert. Es umfasst *alle* im Bundestag vertretenen Parteien sowie parteilose Abgeordnete.

Da dieser Datensatz natürlich nicht nur für Fragestellungen im Romantik-Kontext, sondern für die deutschsprachige politische Diskursanalyse generell wertvoll sein kann, stellen wir es der Fachöffentlichkeit zur Verfügung (https://github.com/JULIELab/DeBAC). Aus rechtlichen Gründen distribuieren wir dabei nur die Tweet-IDs und dazugehörigen Meta-Daten (u.a. Autor, Erstellungszeitpunkt und Parteizugehörigkeit), während die Rohtexte über ein ebenfalls mitgeliefertes Skript heruntergeladen werden können.

Analytik: Im ersten Anlauf suchten wir nach Stichwörtern, die Romantik-Konzepte indizieren. Hierzu wurde eine explorative Umfrage unter mehreren Literaturwissenschaftlern (allesamt Mitglieder des Graduiertenkollegs „Modell Romantik" an der Friedrich-Schiller-Universität Jena)[1] durchgeführt, um gebräuchliche lexikalische Signale für diese Epoche zu bestimmen. Dabei stellte sich heraus, dass nicht nur direkte Lexikalisierungen wie „*Romantik*", „*Romantiker*", „*romantisch*" romantikrelevant sind, sondern auch solche wie „*Gemeinschaft*",

„*Wesen*“, „*Glauben*“, „*Heimat*“ (man denke an Friedrich Schlegels *Über den Republikanismus*, Novalis' *Glauben und Liebe* usw.). Das Suchergebnis wurde sowohl quantitativ analysiert als auch qualitativ interpretiert. Die folgende Tabelle zeigt die Häufigkeiten von Tweets mit diesen Stichwörtern und ihre Zuordnung zu Parteien:

Tabelle 1: Häufigkeit der Stichwörter mit Romantikbezug, gruppiert nach Parteien im Bundestag. Tweets der insgesamt vier fraktionslosen Abgeordneten (mit sehr niedrigen Belegzahlen) sind zur Übersichtlichkeit nicht aufgeführt

Suchwort (Regulärer Ausdruck)	CDU/CSU	SPD	AfD	FDP	LINKE	GRÜNE	S
/[Rr]omantik/	29	14	7	11	20	15	96
/[Rr]omantisch/	9	10	7	13	2	3	44
/[Rr]omantisier/	1	2	2	5	0	4	14
/[Gg]lauben/	375	298	350	252	198	277	1750
/[Gg]emeinschaft/	424	399	104	234	260	343	1764
/[Ww]esen/	925	844	504	700	688	835	4496
/[Hh]eimat/	1504	941	562	312	314	639	4272
Insgesamt	3267	2508	1536	1527	1478	2116	12436

Die Tabelle zeigt, dass die direkten Lexikalisierungen „ *Romantik*“, „ *Romantiker*“ und „ *romantisch*“ vergleichsweise selten vorkommen und wenn, dann verweisen sie meist auf eine Lesart im Sinne von „ *realitätsfern*“, z.B.:

> #Grüne und #Linke wollen, dass #Karlsruhe die Patenschaft für ein Seenotrettungsschiff einer Nichtregierungsorganisation (NGO) im Mittelmeer übernimmt. Eine romantische, realitätsferne Weltsicht. (https://twitter.com/MarcBernhardAfD/status/1062048613923201026)

Dagegen kommen indirektere Lexeme wie „ *Gemeinschaft*“ und „ *Heimat*“ weitaus häufiger vor und werden im Sinne eines abgrenzenden und ausschließenden Charakters eingesetzt, z.B.:

> *Feste, Feiern, Schwimmbäder: Der Verlust öffentlicher Orte und von Gemeinschaftserlebnissen. Nicht alle haben private Pools*. https://t.co/jZsxnmFjCP(https://twitter.com/Renner_AfD/status/1155441711105134592)

> #Bayern gibt Unsummen für illegale Migranten aus. Geld, das vielen älteren Menschen fehlt, die Jahrzehnte für unsere Heimat und unsere Gesellschaft hart gearbeitet haben. Schützen Sie unser Sozialsystem gegen Armutseinwanderung und geben wir den Rentnern mehr. #AfD zur #LtwBayern https://t.co/0imAQg3oCj(https://twitter.com/ProfMaier/status/1044102746411073536)

Diese überwiegend qualitative inhaltsanalytische Vorgehensweise haben wir anschließend durch eine einfache quantitative Untersuchung im Rahmen einer automatischen Emotionsanalyse ergänzt (s.a. entsprechende Vorarbeiten von Hellrich et al. (2019) bzw. Buechel et al. (2017). Hierzu haben wir sämtliche Tweets unseres Korpus mithilfe des Software-Werkzeugs JEmAS (Buechel & Hahn 2016) analysiert und ihnen so einen emotionalen Stimmungswert anhand der darin vorkommenden Lexeme zugewiesen.

Dieses Verfahren liefert für relativ häufige Wörter intuitiv plausible Ergebnisse. Das Lexem „*Heimat*“, das in insgesamt 4.325 Tweets vorkommt, wird etwa von CDU und CSU am positivsten verwendet und von Der Linken am wenigsten (aber immer noch) positiv. Demgegenüber mussten wir feststellen, dass für unsere Ausgangsforschungsfrage zentrale Begriffe („*Romantik*“, „*romantisch*“, „*romantisieren*“) in unserem derzeitigen Korpus zu selten vorkommen, um damit auf Grundlage von reinen Worthäufigkeiten zuverlässige Daten erheben zu können. Eine sinnvolle Erweiterung unserer bisherigen Arbeiten besteht daher in der Anwendung fortgeschrittenerer komputationaler Modelle zur Emotionserkennung, die etwa auf Deep Learning (Nay 2016) oder Topic Modeling (Nguyen et al., 2015) beruhen. Unsere Studie ist damit dem weiteren Kontext der Meinungsklima- und Emotionsanalytik im Umfeld parlamentarischer politischer Akteure zuzuordnen (vgl. a. Abercrombie & Batista-Navarro 2018, Green & Larasati 2018, Blätte 2018, van der Zwaan et al. 2016, Rheault et al. 2016, Nguyen et al. 2015, Zirn 2014, Lietz et al. 2014), ein aktueller Schwerpunkt im zur Zeit stark expandierenden Bereich *Computational Social Science*.

Danksagung. Tinghui Duan ist Doktorand des Graduiertenkollegs „Modell Romantik“, das von der DFG unter Fördernummer GRK 2041 gefördert wird; Sven Buechel ist Mitarbeiter eines unter der Förderlinie „Big Data in der makrooökonomischen Analyse“ (Fachlos 2; GZ 23305/003#002) geförderten Projekts des Bundesministeriums für Wirtschaft; Udo Hahn ist PI in beiden Projekten. Die Autoren bedanken sich bei den zwei anonymen Gutachtern für Ihre kritische Anmerkungen und bei Christof Schöch für seine verständnisvolle Kommunikation.

Fußnoten

1. http://modellromantik.uni-jena.de/

Bibliographie

Abercrombie, Gavin / Batista-Navarro, Riza T. (2018): "Identifying opinion-topics and polarity of parliamentary debate motions", in: *WASSA 2018 – Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2018* 280-285.

Barbaresi, Adrien (2018): "A corpus of German political speeches from the 21st century", in: *LREC 2018 – Proceedings of the 11* th *International Conference on Language Resources and Evaluation* 792-797.

Blätte, Andreas (2018): "Zum Verwechseln ähnlich? Eine Klassifikationsanalyse parlamentarischen Diskursverhaltens auf Basis des PolMine-Plenarprotokollkorpus", in: *Computational Social Science. Die Analyse von Big Data, Nomos* 139-162.

Blätte, Andreas / Blessing, André (2018): "The GermaParl corpus of parliamentary protocols", in: *LREC 2018 – Proceedings of the 11* th *International Conference on Language Resources and Evaluation* 810-816.

Buechel, Sven / Hahn, Udo (2016): "Emotion analysis as a regression problem: dimensional models and their implications on emotion representation and metrical evaluation", in: *ECAI 2016 – Proceedings of the 22* nd *European Conference on Artificial Intelligence* 1114-1122.

Buechel, Sven / Hellrich, Johannes / Hahn, Udo (2017): "The course of emotion in three centuries of German text: a methodological framework", in: *dh 2017 – Digital Humanities*

2017: Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO).

Green, Nathan / Larasati, Septina (2018): "The first 100 days: a corpus of political agendas on Twitter", in: *LREC 2018 – Proceedings of the 11* th *International Conference on Language Resources and Evaluation* 2785-2789.

Hellrich, Johannes / Buechel, Sven / Hahn, Udo (2019): "Modeling word emotion in historical language: quantity beats supposed stability in seed word selection", in: *LaTeCH-CLfL 2019 – Proceedings of the 3* rd *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature @ NAACL-HLT 2019* 1-11.

Lietz, Haiko / Wagner, Claudia / Bleier, Arnim / Strohmaier, Markus (2014): "When politicians talk: assessing online conversational practices of political parties on Twitter", in: *ICWSM 2014 – Proceedings of the 8* th *International AAAI Conference on Weblogs and Social Media* 285-294.

Nay, John J. (2016): "gov2vec: learning distributed representations of institutions and their legal text", in: *NLP + CSS 2016 – Proceedings of the [1* st *] Workshop on Natural Language Processing and Computational Social Science @ EMNLP 2016* 49-54.

Nguyen, Viet-An / Boyd-Graber, Jordan / Resnik, Philip / Miler, Kristina (2015): "Tea Party in the House: a hierarchical ideal point topic model and its application to Republican legislators in the 112 th Congress", in: *ACL-IJCNLP 2015 – Proceedings of the 53* rd *Annual Meeting of the Association for Computational Linguistics & 7* th *International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* 1438-1448.

Rheault, Ludovic / Beelen, Kaspar / Cochrane, Christopher / Hirst, Graeme (2016): "Measuring emotion in parliamentary debates with automated textual analysis", in: *PLoS ONE*, 11, e0168843.

Sylwester, Karolina / Purver, Matthew (2015): "Twitter language use reflects psychological differences between Democrats and Republicans", in: *PLoS ONE*, 10, e0137422.

van der Zwaan, Janneke M. / Marx, Maarten / Kamps, Jaap (2016): "Validating cross-perspective topic modeling for extracting political parties' positions from parliamentary proceedings", in: *ECAI 2016 – Proceedings of the 22* nd *European Conference on Artificial Intelligence* 28-36.

Zirn, Cäcilia (2014): "Analyzing positions and topics in political discussions of the German Bundestag", in: *Proceedings of the Student Research Workshop @ ACL 2014* 26-33.

Routinen, Ressourcen und Tools der digitalen Texterforschung. Ein einfacher Einstieg

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

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Deutschland

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

Traditionelle und digitale Arbeitsweisen

Die Anwendung computergestützter Verfahren in den Geistes- und Kulturwissenschaften prägt seit geraumer Zeit die Entwicklung unterschiedlicher Fachdisziplinen (vgl. Thaller 2012). Neue Methoden bahnen sich ihren Weg in den Methodenkanon ganz unterschiedlicher Domänen (vgl. Sahle 2015). Wie aber kann man Lehrenden – mit den unterschiedlichen Ansprüchen universitär Dozierender oder Lehrender an Schulen – einen möglichst niedrigschwelligen, aber dennoch wissenschaftlich seriösen Zugang zu dem Repertoire digitaler Methoden der Texterforschung eröffnen, das zum Spektrum der Digital Humanities zählt? Wie kann man sowohl Begeisterung wie kritische Kompetenz im konkreten Umgang mit Verfahren der digitalen Textanalyse so vermitteln, dass die Alltagspraxis des Lehrens und Forschens davon profitiert? Man muss nicht immer gleich einen theoretischen „Paradigmenwechsel" ausrufen, sondern kann das „neue" Feld besser zunächst im „hands-on"-Modus erschließbar machen. Durch einen niedrigschwelligen Disseminationsansatz entsteht die Möglichkeit, dass alte Fragen und neue Methoden sinnvoll aufeinander bezogen werden können (vgl. etwa Horstmann / Kleymann 2019).

Das im November 2017 an der Universität Hamburg gestartete DFG-Projekt forTEXT (https://fortext.net) entwickelt vor diesem Hintergrund Strategien zur Dissemination digitaler Verfahren für die Arbeit mit Texten (vgl. Horstmann / Jacke / Meister 2018). In den auf der projekteigenen Webseite als Open-Access-Publikationen bereitgestellten zitierfähigen Besprechungen von Routinen, Ressourcen und Tools werden sämtliche Phasen eines literaturwissenschaftlichen Forschungsprojekts abgedeckt. Das Projekt leistet damit die Übersetzungsarbeit zwischen literaturwissenschaftlichen Fragestellungen und technischem Know-how, die für die Vermittlung digital gestützten Arbeitens an traditionellere Geisteswissenschaftlerïnnen notwendig ist.

Routinen

In der Rubrik Routinen stellen wir einführende Einträge zu digitalen *Methoden* der Textdigitalisierung, -annotation, -analyse, -visualisierung, -präsentation etc. zur Verfügung, in denen neben Definition, Diskussion und technischen Hintergründen stets auch die literaturwissenschaftliche Tradition der jeweiligen Methode betont wird. In *Lerneinheiten* zum Selberlernen werden Nutzerïnnen schrittweise an die Umsetzung der vorgestellten Methode in Kombination mit der Anwendung eines konkreten Tools (vgl. Abschnitt 4) und ausgewählter Ressourcen (vgl. Abschnitt 3) herangeführt. Die

Lehrmodule bieten ebenfalls in Verbindung mit konkreten Ressourcen und Tools die Möglichkeit, das bereitgestellte Material in die eigene universitäre Lehrveranstaltung zu integrieren. Es werden zudem Unterrichtsmaterialien für den schulischen Unterricht erarbeitet, die durch eine noch erhöhte Komplexitätsreduktion Routinen der digitalen Literaturerforschung zugänglich machen und dezidiert an fachliche und KMK-Lernziele anknüpfen.

Ressourcen

Ausgewählte und etablierte deutschsprachige *Textsammlungen*, die sinnvoll mit den besprochenen Routinen der digitalen Literaturwissenschaft kombiniert werden können, stellen wir nicht nur vor, sondern ordnen und bewerten diese entsprechend ihrer thematischen Schwerpunkte. Die einzelnen Einträge folgen dabei einem wiedererkennbaren Schema, sodass insgesamt eine schnelle und bedarfsgerechte Orientierung ermöglicht wird. In der Kategorie Ressourcen bieten wir außerdem Tutorial- *Videos*, die digitale Methoden anhand ausgewählter Tools Schritt für Schritt als Screencasts erklären und Video-Fallstudien, die literaturwissenschaftliche Fragestellungen beispielhaft mithilfe digitaler Tools bearbeiten und vorstellen. Außerdem enthält die Ressourcen-Kategorie auf literaturwissenschaftlichen Theorien basierende *Tagsets* und ein umfangreiches *Glossar* mit Erläuterungen zu Standardbegriffen der DH.

Tools

Für jede vorgestellte Methode stellen wir mindestens ein Tool vor, das für die praktische Umsetzung dieser Methode eingesetzt werden kann. Die Tools werden bedarfsgerecht hinsichtlich ihrer Funktionalität, Anwendungsfreundlichkeit, Nutzerbetreuung, Datensicherheit, Nutzungsbedingungen und des Grads ihrer Etablierung im wissenschaftlichen Diskurs befragt. Die Tooleinträge folgen – wie auch die einzelnen Beitragsformate in den Kategorien Routinen und Ressourcen – einem wiedererkennbaren Schema, in dem konkrete Fragen aus Nutzerïnnenperspektive gestellt und beantwortet werden.

CATMA 6

Mit der Entwicklung der sechsten Version von CATMA (https://catma.de) hat forTEXT im Oktober 2019 neue Funktionen, eine projektzentrierte Arbeitsstruktur und ein vollständig überarbeitetes, intuitiver nutzbares Interface des webbasierten, kollaborativ nutzbaren Annotations- und Analysetools (derzeit weltweit gut 13.000 Accounts[1]) zur Verfügung gestellt. Das Tool integriert sich durch seine nutzerïnnenfreundliche Zugänglichkeit und die Konzentration auf die Methode der manuellen Annotation sowie der Analyse und Visualisierung von Text- wie Annotationsdaten in das forTEXT-Disseminationmodell und orientiert sich an den Bedarfen textwissenschaftlicher Fachwissenschaften.

Nicht-digitale und digitale Dissemination

Das Projekt wird durch umfangreiche Maßnahmen der nicht-digitalen Dissemination seiner Inhalte begleitet. Einerseits bieten die Projektmitarbeiterïnnen bedarfsgerechte Workshops und Vorträge für Forschungsgruppen oder Veranstaltungsreihen an Universitäten und auf Konferenzen an. Darüber hinaus werden schulinterne Workshops durchgeführt, die auf die z. T. sehr unterschiedliche technische Infrastruktur vor Ort eingehen und sich in der inhaltlichen Ausrichtung ebenfalls eng an der spezifischen Bedarfslage der Teilnehmerïnnen orientieren.

Die umfangreiche Social-Media-Strategie von forTEXT (vgl. Horstmann / Schumacher 2019) ist ein essentieller Teil des gesamten Dissiminationsprogramms: Auf Twitter, Youtube, Facebook und Pinterest treten wir in unterschiedlichen Modi mit diverse Zielgruppen in Kontakt und führen diese in die digitale Arbeit mit Texten ein. So tritt forTEXT nicht nur an neue Nutzerïnnengruppen heran, sondern integriert sich auch selbst im fachwissenschaftlichen/DH-Diskurs.

Individualisiertes Empfehlungssystem

Im Januar 2020 wird ein digitales Empfehlungssystem implementiert, das im Frage-Antwort-Schema die Projekte der Nutzerïnnen so klassifiziert, dass die automatische Generierung individualisierter Empfehlungen von Routinen, Ressourcen und Tools zur Bearbeitung der jeweiligen Fragestellung möglich sein wird. Das Empfehlungssystem wird somit dafür sorgen, dass die einzelnen Bereiche von forTEXT einerseits zusammengefasst, andererseits aber auch bedarfsorientiert und effektiv durch sie navigiert werden kann. Das System macht damit insbesondere Nutzerïnnen ohne vorherige DH-Erfahrung den Einstieg in digitale Methoden zur Unterstützung ihrer Projekte individuell möglich.

Fußnoten

1. Von den derzeit 13.033 Accounts wurden 3030 nur einmalig benutzt und 1876 waren Guest-Accounts, sodass man von 8127 Nutzerïnnen ausgehen kann (Stand: Dez. 2019).

Bibliographie

Horstmann, Jan / Jacke Janina / Meister, Jan Christoph (2018): „Digital vs. Humanities. Didaktische Aufbereitung digitaler Methoden für die klassischen Geisteswissenschaften im Projekt forTEXT“, in: *Kritik der digitalen Vernunft. DHd 2018 Köln. Konferenzabstracts*, 386–391. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [Zugriff: 26. August 2019].

Horstmann, Jan / Schumacher, Mareike (2019): „Social Media, YouTube und Co: Multimediale, multimodale und multicodierte Dissemination von Forschungsmethoden in for-

TEXT", in: Sahle, Patrick (ed.): *DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 207–211. DOI: 10.5281/zenodo.2596095 .

Horstmann, Jan / Kleymann, Rabea (2019): „Alte Fragen, neue Methoden – Philologische und digitale Verfahren im Dialog. Ein Beitrag zum Forschungsdiskurs um Entsagung und Ironie bei Goethe", in: *Zeitschrift für digitale Geisteswissenschaften* DOI: 10.17175/2019_007 .

Sahle, Patrick (2015): „Digital Humanities? Gibt's doch gar nicht!", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities.* Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1. DOI: 10.17175/sb001_004 .

Thaller, Manfred (2012): „Controversies around the digital humanities: an agenda", in: *Historical Social Research*, *37*(3): 7–23.

Schmankerl Time Machine. Rechnerisch-explorative Zugänge zur Gastronomie in München

Schulz, Julian
julian.schulz@lmu.de
IT-Gruppe Geisteswissenschaften, Ludwig-Maximilians-Universität München, Deutschland

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Institut für Kunstgeschichte, Ludwig-Maximilians-Universität München, Deutschland

Cakir, Osman
osman.cakir@campus.lmu.de
IT-Gruppe Geisteswissenschaften, Ludwig-Maximilians-Universität München, Deutschland

Kohl, Linus
linus@munichresearch.com
Munich Research

Reißer, Alexandra
areisser92@googlemail.com
Institut für Kunstgeschichte, Ludwig-Maximilians-Universität München, Deutschland

Die Web-Applikation *Schmankerl Time Machine*[1] wurde im Rahmen des Hackathons für offene Kulturdaten, „Coding da Vinci Süd 2019" (Bergmann, 2019), von einem interdisziplinären Team aus Informatikern, Statistikern und Geisteswissenschaftlern entwickelt (Deck, 2018). Das Projekt basiert auf den digitalisierten Speisekarten Münchner Restaurants, die die *Monacensia* der Stadtbibliothek München für den Hackathon zur Verfügung gestellt hatte. Am Ende der sechswöchigen Sprintphase konnte der Prototyp reüssieren und wurde von der Jury mit dem Preis in der Kategorie „Most Technical" bedacht (Lehr, 2019). Seitdem lädt die *Schmankerl Time Machine* zu einem lukullischen Streifzug durch die traditionsreiche Münchner Wirtshausgeschichte der vergangenen 150 Jahre ein. Einen ähnlichen Weg schlägt die Plattform „What's on the Menu?" ein, die auf dem Speisekartenbestand der *New York Public Library* basiert und die überwiegend US-amerikanische Gastronomie zwischen 1851 und 2008 abbildet.[2] Andere interessante Bestände harren dagegen noch ihrer Digitalisierung aus.[3]

Die Applikation besitzt bereits jetzt ein großes Potenzial für eine breite Öffentlichkeit (Guyton, 2019; Kotteder, 2019) und zeigt damit exemplarisch, wie die *Digital Humanities* über den wissenschaftlichen Raum hinaus zu einer Beschäftigung mit kulturgeschichtlichen Daten anregen können. Das einzureichende Poster möchte die Idee hinter der Applikation, ihre technische Umsetzung und Funktionalität gleichermaßen wie die Nachhaltigkeitsstrategie sowie künftige Entwicklungsmöglichkeiten präsentieren.

Daten und Datenaufbereitung

375 Speisekarten mit 1.020 Seiten aus den Jahren 1855 bis 2008 wurden inklusive *Metadaten* durch die *Monacensia* bereitgestellt. Sie entstammen 144 Münchner Gaststätten, Restaurants, Cafés, Bars, Festzelten und -hallen, die regional größtenteils in den Stadtbezirken Altstadt-Lehel und Ludwigvorstadt-Isarvorstadt zu verorten sind. Aufgrund unterschiedlicher Schriftarten wurde in *Transkribus* ein komplementärer Ansatz mit *Handwritten Text Recognition* (*HTR*) und *Optical Character Recognition* (*OCR*) verfolgt. Anschließend erfolgte eine manuelle Fehlerüberprüfung. Zusätzlich wurde ein *Tagset* entworfen, um die konsistente Annotation von Mengenangaben und Preisen, Gerichten und deren Zusammensetzung zu gewährleisten. Für die kollaborative Projektarbeit, insbesondere die Datenorganisation und -analyse, wurde die Lehr- und Forschungsinfrastruktur *Digital Humanities Virtual Laboratory* (*DHVLab*) eingesetzt, die seit 2016 an der IT-Gruppe Geisteswissenschaften der Ludwig-Maximilians-Universität München entwickelt wird (Klinke, 2018: 29–32).

Technische Umsetzung und Funktionalitäten

Folgende Frage stand im Vordergrund: Wie kann die enorme Vielfalt der Speisekarten auch von einer technisch wenig versierten Zielgruppe auf möglichst unterschiedliche Art und Weise exploriert werden? Um dies zu erreichen, wurde eine interaktive, responsive Web-Applikation mit der Open-Source-Umgebung *R* und den auf *R* basierenden Paketen *Shiny* und *Tidyverse* entwickelt, die auf Clientseite ergänzt wird durch *HTML5*, *JavaScript* und das *Frontend-CSS-Framework Bootstrap*. Eine Lokalität kann entweder über ein *Dropdown*-Menü oder eine dynamische Karte (basierend auf *Leaflet* und *LocationIQ*) ausgewählt werden (Abbildung 1). Zu jeder Lokalität werden weiterführende Informationen angeboten. Sofern di-

gital vorhanden, wird auf alte Ansichten der Restaurants aus dem Münchner Stadtarchiv verlinkt.

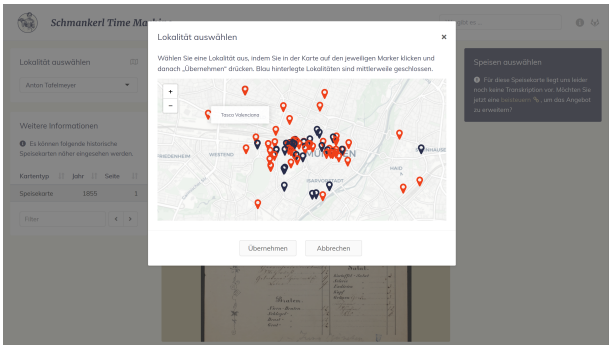


Abbildung 1: Auswahl einer Lokalität über eine dynamische Karte. Bildnachweis: *Schmankerl Time Machine*; lizenziert unter CC BY-SA 4.0.

Jede zu einer Lokalität gehörende Speisekarte kann beliebig gezoomt und verschoben werden. Zudem ist jede Annotation, und damit auch jedes Gericht, direkt anwählbar (Abbildung 2). Besonders „exquisite" Speisen werden algorithmisch über das 0,65-Quantil ausfindig gemacht – und sogar komplette Menüs zusammengestellt; wobei nicht nur die Präferenzen der jeweiligen Nutzerin oder des jeweiligen Nutzers berücksichtigt werden, sondern auch ihr oder sein Budget (Abbildung 3). Ein virtueller Warenkorb unterstützt die Exploration des Fundus weiterhin: Durch die Verknüpfung mit der Rezeptdatenbank des Webportals *Chefkoch.de* können ausgewählte Gerichte nachgekocht werden; Zutatenliste inklusive.

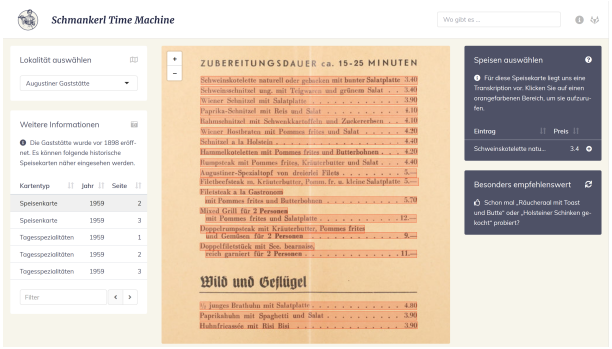


Abbildung 2: Speisekarte der Augustiner Gaststätte von 1959. Bildnachweis: *Schmankerl Time Machine*; lizenziert unter CC BY-SA 4.0.

Neben diesem spielerischen Zugang zu den Speisekarten kann die *Schmankerl Time Machine* als Ausgangspunkt für wissenschaftliche und gesellschaftliche Fragestellungen dienen:

- In welchem Jahr findet sich erstmals ein bestimmtes Gericht? Wie stellt sich die Preisentwicklung dar?
- Welche Strategien verfolgten die Restaurants, um ihre Kunden zu einer profitablen Speisenauswahl zu animieren?
- Finden sich in der Beschreibung der Gerichte Hinweise auf ein sich veränderndes Ernährungsbewusstsein?

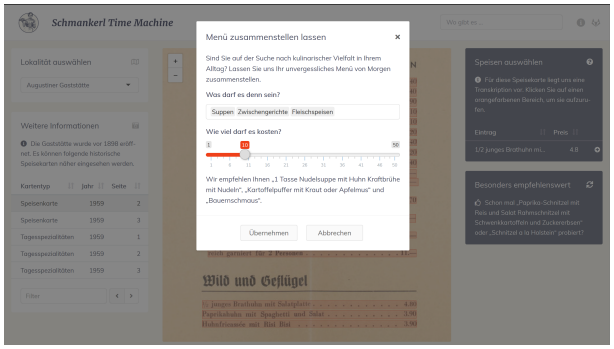


Abbildung 3: Nutzerpräferenzen-basierte Menüempfehlung. Bildnachweis: *Schmankerl Time Machine*; lizenziert unter CC BY-SA 4.0.

Diese Beispiele zeigen, wie vielfältig sich die Beschäftigung mit den hier erstmals dargebotenen Speisekarten gestalten kann (Roth und Rauchhaus, 2018). Um einen möglichst niederschwelligen Einstieg zu gewährleisten, werden *Jupyter Notebooks* in *Python 3* zur Verfügung gestellt, die die Daten importieren, bereinigen und exemplarische Statistiken beinhalten. Hierfür werden gängige Bibliotheken im Bereich *Data Science* verwendet (etwa *pandas*, *NumPy* und *Matplotlib*).

Nachhaltigkeitskonzept

Gemäß den *FAIR*-Prinzipien (*Findable, Accessible, Interoperable, Re-usable*) wurde ein umfassendes Nachhaltigkeitskonzept verfolgt. Der Quelltext der Applikation sowie die Skripte sind auf *GitLab* verfügbar.[4] Die Abbildungen der Speisekarten sowie die im Projekt entstandenen Forschungsdaten stehen im Repositorium der Ludwig-Maximilians-Universität München (*Open Data LMU*) unter einer offenen Lizenz (*CC BY-SA 4.0*) dauerhaft und mittels *DOI* eindeutig referenzierbar zur Nachnutzung bereit.[5] Die Beschreibung des Projekts erfolgt im Metadatenschema *DataCite* unter Verwendung eines Best-Practice-Guides, der eine standardisierte Anreicherung der Metadaten unterstützt.[6] Dies ermöglicht die Einbindung der Projektdaten in übergeordnete Forschungsdateninfrastrukturen (z. B. *GeRDI*) und damit ihre leichtere Auffindbarkeit.

Ausblick

Bei der *Schmankerl Time Machine* handelt es sich um einen fortgeschrittenen Prototyp. Um sein Potenzial sowohl für wissenschaftliche Fragestellungen als auch für eine interessierte Öffentlichkeit zu vergrößern, ist die Integration weiterer Speisekartensammlungen und damit eine wesentliche Erweiterung des Datenbestands vorgesehen. Damit einhergehend wird darauf abgezielt, die Annotation der Karten – unter Einbezug der „Crowd"[7] – fortzuführen und um weitere Analysekategorien zu erweitern. In Kooperation mit der *Monacensia* ist zu diesem Zweck auch ein *Edithaton* geplant, bei dem Studierende der Ludwig-Maximilians-Universität München u. a. praktische Kenntnisse im Umgang mit *Transkribus* erhalten.

Ein Beispiel, welche Forschungsfragen dadurch eröffnet werden können, stellt das Projekt „Menu Journeys" dar, das Studierende der *Berkeley School of Information* 2015 angestoßen haben.[8] In interaktiven Grafiken wird auf Basis des Spei-

sekartenbestands der *New York Public Library* anschaulich dargestellt, wie sich etwa der durchschnittliche Preis eines Gerichts über die Jahrzehnte hinweg und in Relation zur Inflationsrate entwickelt hat. Analysen dieser Art wären auch für die Münchner Gastronomiegeschichte begrüßenswert. Die hier vorgestellte Web-Applikation bietet einen Ausgangspunkt für künftige Entwicklungen in diese Richtung.

Fußnoten

1. https://dhvlab.gwi.uni-muenchen.de/schmankerltimemachine/ (27.09.2019).
2. http://menus.nypl.org/ (27.09.2019).
3. Größere Bestände mit geographischem Schwerpunkt auf München befinden sich im Münchner Stadtarchiv sowie der Stadtbibliothek München.
4. https://gitlab.com/cds19-team/cds19 (27.09.2019).
5. https://doi.org/10.5282/ubm/data.146 (27.09.2019).
6. Der *DataCite*-Best-Practice-Guide wurde im Rahmen des Projekts „eHumanities – interdisziplinär“ in Kooperation mit dem Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften entwickelt: https://doi.org/10.5281/zenodo.3559800.
7. Eine Mitarbeit am Projekt „cds19“ ist nach Registrierung möglich: https://transkribus.eu/r/read/projects/ (27.09.2019).
8. http://people.ischool.berkeley.edu/~carlos/menujourneys/ (18.12.2019).

Bibliographie

Bergmann, Claudia (2019): „Kultur-Hackathon Coding da Vinci Süd. Sterbende Jesuiten, visualisierte Theaterdaten und Wirtshausgeschichte zum Nachkochen“, auf: *Wikimedia Blog* vom 27.05.2019, URL: https://blog.wikimedia.de/2019/05/27/kultur-hackathon-coding-da-vinci-sued-sterbende-jesuiten-visualisierte-theaterdaten-und-wirtshausgeschichte-zum-nachkochen/.

Deck, Klaus-Georg (2018): „Digital Humanities. Eine Herausforderung an die Informatik und an die Geisteswissenschaften“, in: Huber, Martin / Krämer, Sybille (Hrsg.): *Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden*. Sonderband der Zeitschrift für digitale Geisteswissenschaften 3. DOI: 10.17175/sb003_002.

Guyton, Patrick (2019): „Zeitreise durchs kulinarische München“, in: *Der Tagesspiegel* vom 26.07.2019, URL: https://www.tagesspiegel.de/gesellschaft/panorama/historische-speisekarten-zeitreise-durchs-kulinarische-muenchen/24843382.html.

Klinke, Harald (2018): „Datenanalyse in der Digitalen Kunstgeschichte. Neue Methoden in Forschung und Lehre und der Einsatz des DHVLab in der Lehre“, in: Ders. (Hrsg.): *#DigiCampus. Digitale Forschung und Lehre in den Geisteswissenschaften*. München: Universitätsbibliothek der Ludwig-Maximilians-Universität, 19–34, DOI: https://doi.org/10.5282/ubm/epub.42415.

Kotteder, Franz (2019): „Hat’s geschmeckt?“, in: *Süddeutsche Zeitung* vom 09.07.2019, URL: http://sz.de/1.4516782.

Lehr, Andrea (2019): „Hochkarätig. CDV Süd punktet mit Projekten auf hohem Niveau“. *Pressemitteilung* vom 21.05.2019, URL: https://codingdavinci.de/news/2019/05/21/cdvs-preisverleihung_2.html.

Roth, Tobias / Rauchhaus, Moritz (Hrsg.) (2018): *Wohl bekam’s! In hundert Menus durch die Weltgeschichte*. Berlin: Verlag Das Kulturelle Gedächtnis.

Science Data Center für Literatur

Ulrich, Mona
mona.ulrich@dla-marbach.de
Deutsches Literaturarchiv Marbach

Hess, Jan
jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach

Kamzelak, Roland
roland.kamzelak@dla-marbach.de
Deutsches Literaturarchiv Marbach

Kramski, Heinz Werner
heinz.werner.kramski@dla-marbach.de
Deutsches Literaturarchiv Marbach

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

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Institut für maschinelle Sprachverarbeitung der Universität Stuttgart

Schlesinger, Claus-Michael
claus-michael.schlesinger@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft / Digital Humanities der Universität Stuttgart

Viehhauser, Gabriel
gabriel.viehhauser-mery@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft / Digital Humanities der Universität Stuttgart

Schembera, Björn
schembera@hlrs.de
Höchstleistungsrechenzentrum Stuttgart

Bönisch, Thomas
boenisch@hlrs.de
Höchstleistungsrechenzentrum Stuttgart

Kaminski, Andreas
kaminski@hlrs.de
Höchstleistungsrechenzentrum Stuttgart

Die Digitalisierung verändert die Bedingungen für die Produktion, Distribution und Rezeption und damit auch für die Erforschung von Literatur. In den Digital Humanities stehen dabei bislang insbesondere die neuen Möglichkeiten der digitalen Auswertung (Distant/Scalable Reading) und die Digitalisierung vorhandener Druckbestände im Zentrum der Aufmerksamkeit. Die veränderten medialen Bedingungen führen jedoch nicht nur zu einer Übersetzung von gedruckten Texten in digitale Objekte, sondern bringen selbst produktiv neue literarische Formen und Gattungen hervor, für die computergestützte Elemente konstitutiv sind. Hierzu zählen etwa literarische Hypertexte, Blog-Formate, computergestützte kollektive und kollaborative Projekte, literarische Tweets und Twitter-Bots, Texte und Textgeneratoren, die auf computerlinguistische Methoden setzen, schließlich auch frühere Formen computergestützter Literaturproduktion wie der Poesieautomat von Hans Magnus Enzensberger oder die *Stochastischen Texte* von Theo Lutz. (Rettberg 2019, Suter 2012, Tomaszek 2011, Lutz 1959). Hinzu kommen im Bereich Literaturforschung und -archive zunehmend digitale Vor- und Nachlässe, die eine Vielzahl von unterschiedlichen Datenträgern und Datenformaten beinhalten.

Das jüngst ins Leben gerufene interdisziplinäre *Science Data Center für Literatur (SDC4Lit)* hat sich das Ziel gesetzt, die Anforderungen, die Digitale Literatur an ihre Archivierung, Erforschung und Vermittlung stellt, systematisch zu reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus Literatur langfristig umzusetzen.

Für die Archivierung, Analyse und Vermittlung von Digitaler Literatur wird eine Forschungsplattform entwickelt. Da eine solche Plattform nur in der interdisziplinären Zusammenarbeit zu bewerkstelligen ist, sind im Projekt Partner mit unterschiedlichen Expertisen in den einzelnen Teilbereichen vereint, nämlich das Deutsche Literaturarchiv Marbach, das Höchstleistungsrechenzentrum Stuttgart, sowie das Institut für Maschinelle Sprachverarbeitung und die Abteilung Digital Humanities der Universität Stuttgart.

Die born-digital Bestände des Deutschen Literaturarchivs bestehen zum einen aus digitalen Nachlässen und zum anderen aus archivierten netzliterarischen Werken. Der umfangreichste digitale Nachlass am Deutschen Literatuarchiv ist von Friedrich Kittler und umfasst 1,5 Millionen Dateien. Zur deutschsprachigen Netzliteratur können weitaus weniger Objekte gezählt werden. Netzliteratur ist durch Verlinkungen und Multimedialität geprägt. Das erschwert die Definition von Objektgrenzen und führt zu nichtlinearen Objektstrukturen, die in der Rezeption nichtlineare Handlungen ermöglichen

Zum einem scheinen sich diese Texte also zur Anwendung computergestützter und computerlinguistischer Methoden besonders anzubieten, da sie genuin in elektronischer Form vorliegen. Zum anderen bringt gerade diese Form für ihre Archivierung und Bereitstellung eine Reihe von besonderen Anforderungen mit sich.

Digitale Nachlässe sind aufgrund großer Mengen an Daten ohne computergestützte Methoden kaum erschließbar und zugänglich zu machen. Um auf diese wachsende Herausforderung in Archiven und Bibliotheken einzugehen, soll der Einsatz von Methoden der Digital Humanities für die inhaltliche Erschließung textbasierter born-digital Bestände erprobt werden. Wenn digitale Nachlässe bereits obsolete Dateiformate enthalten, sind diese nicht ohne vorherige Formatmigration für aktuelle computergestützte Analysen zugänglich.

Auch literarische Webseiten sind von den hochfrequenten Erneuerungszyklen digitaler Technik betroffen. Weiterentwicklungen der Betriebssysteme, der Browser, des HTML-Standards und gängiger Webtechnologien können zu fehlerhafter Darstellung oder fehlenden Funktionen einer Webseite führen. Um ein Werk der Netzliteratur dokumentieren zu können, sind daher neue Formen der Modellierung von Texten, die über eine bloß lineare Form hinausgehen, gefragt.

Diese und weitere Bestände sollen mit modernen digitalen Methoden erschlossen, erforscht und vermittelt werden können. Im Zentrum stehen daher der Aufbau verteilter langzeitverfügbarer Repositories für Digitale Literatur inklusive Forschungsdaten und die Entwicklung der SDC4Lit-Forschungsplattform. Die Repositories werden vom Projekt und seinen Kooperationspartnern regelmäßig erweitert und bilden den zentralen Speicher für das Harvesting von Netzliteratur und weiteren Formen elektronischer Literatur im künftigen Betrieb des SDC. Die Forschungsplattform bietet die Möglichkeit zum computergestützten Arbeiten mit den Beständen der Repositories.

Bereits entwickelte oder in der Entwicklung befindliche Ansätze zur Archivierung und Bereitstellung von WARC-Archiven (Lin et al. 2017), Textkorpora (Fischer et al. 2019) und Analysefunktionen (Hinrichs et al. 2010) sowie strukturierte Reflexionen eigener Strategien (Kramski, von Bülow 2011) weisen auf eine modulare und integrierte Lösung bei der Bereitstellung von Daten und Services. Die entsprechend geplante modulare Architektur der bereitgestellten Services ermöglicht eine nachhaltige Integration von Repositories und Analysemethoden sowie die Möglichkeit zur späteren bedarfsorientierten Einbindung von Korpora und Analysewerkzeugen.

Für die Entwicklung des Repositories und der Forschungsplattform ist der Kontakt zu an der Herstellung, Verbreitung, Erforschung und Vermittlung von elektronischer Literatur beteiligten Communities ein entscheidendes Element. Diese Beteiligung wird über einen mehrköpfigen Beirat und Outreach-Maßnahmen wie Workshops, Seminare und die Arbeit mit Fokusgruppen erreicht. Eine wichtige Aufgabe des Projekts ist in diesem Zusammenhang die Modellierung von Formen digitaler Literatur, die zunächst beispielorientiert im Umgang mit einem bereits vorhandenen Corpus digitaler Literatur erfolgt.[1] Daraus entstehen sowohl technische als auch gattungspoetologische Herausforderungen, etwa bei der Begriffsbildung (digitale vs. elektronische Literatur), bei der medienbezogenen Abgrenzung von digitaler und nicht-digitaler und post-digitaler Literatur, und schließlich in Bezug auf gattungspoetologische und literaturgeschichtliche Fragen zur elektronischen Literatur seit den 1950er Jahren mit einem Fokus auf den deutschsprachigen Raum und mit Blick auf internationale Entwicklungen in Literatur und Literaturforschung. (Block, 2004; Gould, 2012; Rettberg, 2019; Seiça, 2015)

Neben digitalen Objekten und entsprechenden Metadaten wird auch ein Repository der anfallenden Forschungsdaten nachvollziehbar und nachhaltig gespeichert. Zu den Forschungsdaten zählen erstens die bei der Arbeit des SDC an-

fallenden Forschungsdaten, insbesondere solche, die für das Anbieten von Diensten auf der Plattform notwendig sind, etwa mittels Machine Learning errechnete Datenmodelle für an das Corpus angepasste computerlinguistische Analysewerkzeuge (Eigennamenerkenner, Parser, Topic Models etc.). Zweitens soll das Repository die Möglichkeit bieten, die von Nutzer*innen der Forschungsplattform generierten Forschungsdaten strukturiert zu speichern und für die weitere Forschung zur Verfügung zu stellen, etwa Annotationen oder ergänzte Metadaten zu einzelnen Objekten oder zu Objektklassen.

Die Sammlung, Bereitstellung, Erforschung und Vermittlung von Literatur im medialen Wandel ist eine Aufgabe, die Forschung und Archive gleichermaßen betrifft. SDC4Lit verfolgt deshalb das Ziel, diese Aufgabe und die entsprechenden Unteraufgaben interdisziplinär zu bearbeiten.

Fußnoten

1. Deutsches Literaturarchiv Marbach: Literatur im Netz, http://literatur-im-netz.dla-marbach.de/, Zugriff 20.9.2019.

Bibliographie

Block, Friedrich W. (2004): *p0es1s. Ästhetik digitaler Poesie = The aesthetics of digital poetry*. Erscheint anlässlich der Ausstellung "p0es1s. Digitale Poesie" im Kulturforum Potsdamer Platz, Berlin, 13. Februar bis 4. April 2004. Ostfildern: Hatje Cantz.

Gould, Amanda Starling (2012): „A Bibliographic Overview of Electronic Literature“. In: *Electronic Literature Directory* o.V.

Hinrichs, Erhard W., Marie Hinrichs und Thomas Zastrow (2010): *WebLicht: Web-Based LRT Services for German*, Proceedings of the ACL 2010 System Demonstrations, S. 25–29.

Kramski, Heinz Werner, Ulrich von Bülow (2011): *„Es füllt sich der Speicher mit köstlicher Habe“ – Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach*, in: Caroline Y. Robertson von Trotha. Robert Hauser (Hg.), Neues Erbe : Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung, Karlsruhe: KIT Scientific Publishing, S. 141-162.

Rettberg, Scott (2019): *Electronic literature*. Cambridge, UK: Polity Press.

Seiça, Álvaro (2015): *Um Feixe Luminoso: Uma Leitura da Coleção de Literatura Electrônica Portuguesa*. Florianópolis: Universidade Federal de Santa Catarina.

Suter, Beat (2012): *Von Theo Lutz zur Netzliteratur. Die Entwicklung der deutschsprachigen elektronischen Literatur*, <https://netzliteratur.net/suter/Geschichte_der_deutschsprachigen_Netzliteratur.pdf>, Zugriff am 31.12.2019.

Tomaszek, Patricia (2011): German Net Literature: In the Exile of Invisibility, <http://elmcip.net/critical-writing/german-net-literature-exile-invisibility>, Zugriff am 19.9.2019.

SoNAR (IDH): Datenschnittstellen für historische Netzwerkanalyse

Bludau, Mark-Jan
bludau@fh-potsdam.de
Fachhochschule Potsdam

Dörk, Marian
doerk@fh-potsdam.de
Fachhochschule Potsdam

Fangerau, Heiner
heiner.fangerau@hhu.de
Heinrich-Heine-Universität Düsseldorf

Halling, Thorsten
thorsten.halling@hhu.de
Heinrich-Heine-Universität Düsseldorf

Leitner, Elena
elena.leitner@dfki.de
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Menzel, Sina
sina.menzel1@hu-berlin.de
Humboldt Universität Berlin

Müller, Gerhard
gerhard.mueller@sbb.spk-berlin.de
Staatsbibliothek Berlin

Petras, Vivien
vivien.petras@ibi.hu-berlin.de
Humboldt Universität Berlin

Rehm, Georg
georg.rehm@dfki.de
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Neudecker, Clemens
clemens.neudecker@europeana-newspapers.eu
Staatsbibliothek Berlin

Zellhoefer, David
david.zellhoefer@sbb.spk-berlin.de
Staatsbibliothek Berlin

Moreno Schneider, Julian
Julian.Moreno_Schneider@dfki.de
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Die Beziehungen zwischen Menschen etwa in Familien, Organisationen oder Märkten bilden das Gewebe sozialer Ordnungen. Beziehungen konstituieren Möglichkeiten und Zwänge; sie beeinflussen den Zugang zu sozialem Kapital und damit Handlungs- und Wahloptionen (Lin 2001). Die Analyse dieser Beziehungen ist wesentlich für das Verstehen und Erklären von sozialen Phänomenen. Mit der Sozialen Netzwerkanalyse (SNA) entwickelte vor allem die sozialwissenschaftliche Forschung auf der Grundlage der Graphentheorie geeignete Methoden und anschlussfähige empirische Theorien zur Beschreibung und Erklärung dieser Strukturen (Jansen 2007). Die methodischen und theoretischen Ansätze der SNA etwa für die Untersuchung von sozialen Positionen (Kadushin 2012), finden zunehmend auch in Verbindung mit historischen Fragestellungen Anwendung (Bauerfeld/Clemens 2014, Düring et al. 2016). Forschung in diesem Bereich ist aber mit zwei grundlegenden Herausforderungen konfrontiert: Zum einen ist Erhebung und Aufbereitung von Daten für Analysen aus den dezentral, teilweise verstreut überlieferten Archiv- und Bibliotheksbeständen aufwendig. Zum anderen ist die Nutzung der einmal erhobenen Daten für neue Forschungsfragen oder auch nur die Überprüfung der Ergebnisse quantitativer historischer Analysen bisher vor allem von persönlichen Faktoren wie der Kenntnis über Datenbezug, -format und -auswahl sowie technische Verfahren abhängig.

Im Ergebnis der Digitalisierungsprojekte unserer Kultur- und Wissenschaftseinrichtungen stehen inzwischen signifikant große, vielfältig repräsentative Datenkorpora bereit. Durch stete Innovation und Standardisierung in der Aufbereitung digitaler Bestände – beispielhaft genannt seien Optical Character Recognition (OCR) und Named Entity Recognition (NER) – gewinnen diese Daten auch das Interesse einer noch jungen quantitativen Perspektive auf historische Phänomene. Doch trotz der erheblichen Potenziale beruhen bisherige Angebote in erster Linie auf den Logistik- und Nutzungskonzepten für analoge Bestände: So erfolgt die Datennutzung und -generierung über Kataloge, Discovery-Systeme oder digitale Sammlungen einzelner Einrichtungen für die ebenso konventionelle Beschäftigung mit Einzelobjekten. Einrichtungsübergreifende Aggregatoren wie die Deutsche Digitale Bibliothek (DDB) optimieren zwar den zeit- und ortsunabhängigen Zugang, aber die quantitative Verwertung der Daten bleibt hinter den Möglichkeiten zurück.

An dieser Stelle setzt das Projekt SoNAR (IDH), Interfaces to Data for Historical Social Network Analysis and Research, an. In diesem anwendungsbezogenen Forschungs- und Entwicklungsprojekt werden systematisch forschungsorientiert das Aufbereiten, Bereitstellen und Analysieren von Massendaten für den Aufbau einer Forschungstechnologie für die Historische Netzwerkanalyse (HNA), die als ein Zweig der SNA historische Fragestellungen untersucht, erprobt. Ausgangspunkte für das Datenmaterial sind:

» Kalliope, KPE (Archiv- und archivähnliche Bestände wie Nachlässe und Autographen),
» Zeitschriftendatenbank, ZDB (fortlaufende Sammelwerke wie Zeitungen und Zeitschriften),
» Gemeinsame Normdatei, GND (Entitäten wie Personen, Körperschaften und Orte) sowie
» exemplarische Brief- (Edition Berliner Intellektuelle) und Zeitungsvolltexte (ZEFYS).

Das entstandene und stetig expandierende referenzielle System dieser verteilten Datenangebote bietet der Wissenschaft die Chance, mit statistischen und visuellen Mitteln einen breiten, tiefen Einblick in Genese und Konstellation vergangener sozialer Beziehungen zu gewinnen. Einzelne wissenschaftliche Arbeiten zeigen sehr überzeugend, aber notgedrungen in reduzierter und abstrahierter Form den Wert quantitativer Methoden anhand von Korrespondenzen aus Archivbeständen, wie sie in KPE erfasst sind, und belegen das enorme Erkenntnispotenzial für die historische Forschung (z.B. Mücke und Schnalke 2009, Boschung et al. 2002, Dauser 2008, Fangerau 2010, 2013). Die Titeldaten der ZDB flankieren Aussagen über soziale Netze (KPE, GND) mit Aussagen über Produktions- und Distributionskonstellationen (z.B. Verlag, Herausgeber, Verbreitung, Sprache). Durch das Aufbereiten von Entitäten in Volltexten von Briefen oder Zeitungsartikeln ist es möglich, die formalisierten Aussagen der Metadaten von KPE und ZDB substanziell zu erweitern.

SoNAR (IDH) soll für den breiten, fächerübergreifenden Bedarf Einzellösungen durch ein standardisiertes Angebot ersetzen und so Hürden für die Arbeit mit Methoden der HNA signifikant reduzieren. Im Ergebnis dieses Vorhabens werden die Leistungsfähigkeit bestehender Frameworks und Werkzeuge in einer Prozesskette zur Datenaufbereitung und -bereitstellung sowie die Chancen neuer Visualisierungs- und Interfacekonzepte für eine Forschungsumgebung demonstriert. Mit einem Implementierungs- und Betriebskonzept werden geeignete Ansätze und Konditionen für Aufbau und Betrieb der Forschungstechnologie aufgezeigt. Dieses Vorhaben knüpft an Konzepte der Infometrie, vor allem der Biblio- und der Szientometrie an, wobei jedoch weniger Fragen nach Trends (Tunger 2009), Impact (Hirsch 2005), Wachstum oder Marktwert (Haustein und Tunger 2013, Umstätter 2004) im Vordergrund stehen, sondern z.B. Figurationen (Elias 1970) oder die räumliche und zeitliche Evolution von sozialen Beziehungen und Kontexten (z.B. Themen). Es wird dabei der Umstand berücksichtigt, dass die Ausgangsdaten nicht für nur ein Forschungsthema erhoben sind, sondern vielfältig nutzbar gemacht werden können. Daher gilt es aber auch, belastbare Aussagen über den Umgang mit fehlenden oder fehlerhaften Daten zu treffen. Die Forschungsumgebung wird durch wissenschaftshistorische Fallstudien begleitet, die im Projekt zu abstrakteren Forschungsdesigns ausgearbeitet werden und so die Potenziale der Technologie für fachwissenschaftliche Fragestellungen demonstrieren.

Erstmals soll ein standardisiertes Instrumentarium zur Verfügung stehen, um mit großen aufbereiteten Datenmengen und einer Forschungsumgebung etwa komplexe, multimodale sozio-historische Kontexte zu untersuchen und Erkenntnisse nach wissenschaftlichen Kriterien in Forschungsprozesse zu integrieren.

Das Poster stellt Konzeption und die einzelnen Teilziele des Projekts vor.

Bibliographie

Auer, Sören (2018): Towards an Open Research Knowledge Graph. Zenodo. (http://doi.org/10.5281/zenodo.1157185)

Bauerfeld, Daniel / Clemens, Lukas (2014): Gesellschaftliche Umbrüche und religiöse Netzwerke. In: Bauerfeld, Daniel/Clemens, Lukas (Hg.) (2014): Gesellschaftliche Umbrüche und religiöse Netzwerke. Analysen von der Antike bis zur Gegenwart. Bielefeld, 2014

Boschung, Urs et al. (Hg.)(2002): Repertorium zu Albrecht von Hallers Korrespondenz 1724-1777. Basel, 2002 (Studia Halleriana ; VII/1)

Dauser, Regina (2008): Informationskultur und Beziehungswissen. Das Korrespondenznetz Hans Fuggers. Tübingen.

Düring, Marten et al. (Hg.)(2016): Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen. Münster.

Elias, Norbert (1970): Was ist Soziologie? (Gesammelte Schriften in 19 Bänden, 5. Berlin. 2009)

Fangerau, Heiner (2010): Spinning the Scientific Web. Jacques Loeb (1859-1924) und sein Programm einer internationalen biomedizinischen Grundlagenforschung. Berlin.

Fangerau, Heiner (2013): Evolution of knowledge from a network perspective. Recognition as a selective factor in the history of science. In: Fangerau, Heiner et al. (Hg.): Classification and Evolution in Biology, Linguistics and the History of Science. Concepts, Methods, Visualization. Stuttgart, 11-32

Haustein, Stefanie / Tunger, Dirk (2013): Sziento- und bibliometrische Verfahren. In: Grundlagen der Praktischen Information und Dokumentation. Berlin, 479-492

Hirsch, Jorge E. (2005): An index to quantify an individuals scientific research output. In: Proceedings of the National Academy of Science of the United States of America. 102, 46. 16569-16572

Isenberg, Petra et al. (2008): Grounded Evaluation of information visualizations. In: ACM DL. BELIV '08 Proceedings of the 2008 Workshop on Beyond time and errors: novel evaluation methods for Information Visualization, 56–63.

Jansen, Dorothea / Wald, Andreas (2007): Netzwerktheorien. In: Benz, Arthur et al. (Hg.): Handbuch Governance. Theoretische Grundlagen und empirische Anwendungsfelder. Wiesbaden, 188-199

Kadushin, Charles (2012): Understanding Social Networks. Theories, Concepts, and Findings. Oxford.

Kromrey, Helmut (2002): Empirische Sozialforschung. Opladen, 2002

Lin, Nan (2001/2011): Social Capital. A Theory of Social Structure and Action. Cambridge.

Luhmann, Niklas (1987): Soziale Systeme. Grundriß einer allgemeinen Theorie. Frankfurt am Main.

Moretti, Franco (2009): Abstrakte Kurven, Karten, Stammbäume. Abstrakte Modelle für die Literaturgeschichte. Frankfurt am Main.

Munzer, Tamara (2009): A Nested Model for Visualization Design and Validation. In: IEEE Transactions on Visualization and Computer Graphics (TVCG). 15, 6. 921-928

Mücke, Marion / Schnalke, Thomas (2009): Briefnetz Leopoldina. Die Korrespondenz der Deutschen Akademie der Naturforscher um 1750. Berlin, 2009

Umstätter, Walter (2004): Szientometrische Verfahren. In: Grundlagen der Information und Dokumentation. Berlin, 237-243

Tunger, Dirk (2009): Bibliometrische Verfahren und Methoden als Beitrag zu Trendbeobachtung und Trenderkennung in den Naturwissenschaften. Jülich.

Spielräume des digitalen Publizierens nutzen: Das Online Journal „Entangled Religions“ als ‚Research Hub’

Heinig, Julia

Julia.Heinig@ruhr-uni-bochum.de
Ruhr-Universität Bochum, Deutschland

Elwert, Frederik

Frederik.Elwert@rub.de
Ruhr-Universität Bochum, Deutschland

Entangled Religions ist ein Open Access Journal, das seit 2014 mit dem Themenschwerpunkt Religionskontakte im eurasischen Raum fortlaufend erscheint. Die Fallstudien beziehen sich dabei immer auf einen geographischen Ort oder Raum, eine spezifische Zeit sowie auf zwei oder mehr religiöse Traditionen, die miteinander in Kontakt treten. Durch den Einbezug analytischer Konzepte (*tertia comparationis*) durch die Autor*innen wird zudem die Möglichkeit geschaffen, einzelne Fallstudien miteinander in Bezug zu setzen und vergleichbar zu machen.

Um diese vergleichenden Aspekte auch für Leser*innen und Nutzer*innen zugänglich zu machen, wird das Journal derzeit zu einer innovativen Forschungsplattform ausgebaut. Erstens entstehen neue Zugriffsmöglichkeiten, Visualisierungen und Filterfunktionen für die journaleigenen Inhalte; zweitens werden die Inhalte von *Entangled Religions* durch Einbindung externer Ressourcen und Datenbanken angereichert. Wir verstehen dabei die Zukunft des wissenschaftlichen digitalen Publizierens nicht mehr als ein Nebeneinander von abgeschlossenen Publikationsorganen, sondern als ein Netzwerk digitaler Ressourcen, in dem der Artikel als Ganzes seine Bedeutung behält, Teile davon aber je nach Forschungsfrage dynamisch mit anderen wissenschaftlichen Texten sowie Forschungs- und Metadaten kombiniert werden können.

Obwohl die technischen Innovationen dies inzwischen erlauben, tendieren digitale Publikationen noch immer dazu, die Beschaffenheit von Printpublikationen zu kopieren (Degkwitz 2013, 83; Kohle 2017, 200), anstatt die Spielräume der digitalen Umgebung zu nutzen. Dies wirkt sich unter anderem auf die verfügbare Journal Management Software aus: beispielsweise sieht *Open Journal Systems* (OJS) noch immer das PDF als bevorzugtes Veröffentlichungsformat vor. Die Archiv-Ordnung der Artikel in Heften und Jahrgängen orientiert sich hierbei ebenfalls am gedruckten Gegenstück. Im Gegensatz dazu sehen wir das Archiv von *Entangled Religions* als wachsenden Wissensspeicher, in dem Leser*innen über Grenzen einzelner Hefte hinaus Verbindungen herstellen können.

Das „Fehlen von Technologien, die Texte und Kontexte zur Geltung bringen können“ (Söllner 2017, 10) führt unter anderem dazu, dass Online-Journale, die über die Funktionalitäten von OJS hinausgehen wollen, meist ein auf sie zugeschnittenes und in sich geschlossenes Journal Management System schaffen (vgl. bspw. *Arcadia* oder die *Zeitschrift für digitale Geisteswissenschaft*). Bei der Weiterentwicklung von *Entangled Religions* soll im Gegensatz dazu mit Open Encyclopedia System (OES) ein bestehendes Open-Source-System nachgenutzt werden, das ursprünglich für Nachschlagewerke entwickelt wurde und nun am Beispiel von *Entangled Religions* zum ersten Mal für eine wissenschaftliche Zeitschrift genutzt wird. Das Endprodukt wird am Ende des Projektes ebenfalls zur Nachnutzung für andere Online-Journals zur Verfügung stehen. Der modulbasierte Aufbau von OES kommt dabei einer dynamischen Anpassung an unterschiedliche Anforderungen zugute. Hierbei soll an einem konkreten Beispiel aus der Religionsforschung ein Tool entwickelt werden, mit dem wissenschaftliche Texte nicht mehr isoliert für sich stehen, sondern durch Anreicherung mit Metadaten und Schlagworten – auch auf Paragraphenebene (Schwaderer u. a. 2016) – offen für vielfältige Verknüpfungen werden. Das zu beobachtende gesteigerte Interesse an OES im deutschspracheigen Raum (beispielsweise am CeDiS und CeMoG Berlin sowie am SFB 948 in Freiburg) hat bereits zu einer Nutzung von OES in mehreren voneinander unabhängigen Projekten geführt. Eine Herausforderung für die kommenden Jahre ist der Aufbau einer gleichermaßen aktiven Entwicklercommunity, die die Weiterentwicklung stärker dezentralisiert. Die geplante Veröffentlichung des Codes als Open Source wird hierfür die Ausgangsbasis schaffen.

Ganz konkret planen wir die Umstellung des Online Journals zu einem ‚Research Hub‘ bis Mitte 2021 in den folgenden Schritten:

- Das geplante Journal Management System wird ermöglichen, die Journalinhalte nicht nur auf der Artikelebene mit Metadaten zu versehen, sondern auch einzelne Absätze zu verschlagworten (im Fall von *Entangled Religions* nach Ort, Zeit, Religion, theoretischem Konzept). Eine neue, facettierte Suche wird darauf aufbauend zulassen, gezielter relevante Auszüge eines Artikels zu entdecken und ausgehend von einer relevanten Passage eines Artikels über eine Empfehlungsfunktion ähnliche Passagen aus anderen Artikel zu finden.
- Die Verschlagwortung ist ebenfalls Grundlage für vielfältige Visualisierungsmöglichkeiten der Inhalte von *Entangled Religions*, beispielsweise auf einer digitalen Karte, Timeline oder Keywordcloud.
- Durch Anbindung externer Datenbanken und Bibliographien, wie RelBib und JSTOR, können über die journaleigenen Inhalte hinaus weitere relevanten Ressourcen entdeckt und aufgerufen werden. Neben der Nutzung standardisierter Schlagworte (z.B. GND) sollen verschiedene Recommender-Ansätze ausprobiert werden. Gemeinsam mit RelBib wird eine Schnittstelle für die dort genutzte Software VuFind implementiert, die es erlaubt die bislang nur innerhalb der Plattform verfügbare Ähnlichkeitssuche auch aus *Entangled Religions* heraus aufzurufen. Zusätzlich wird die Nutzung von Techniken des maschinellen Lernens für das Information Retrieval evaluiert. So können etwa über Topic Models ähnliche Artikel identifiziert werden, die nicht mit dem gleichen Schlagwortsystem ausgezeichnet werden. Hier ist die Einbindung von JSTOR Text Analyzer (Snyder 2017) angedacht, um passende Artikel aus JSTOR vorzuschlagen.

Dementsprechend kommt die neue Wissensplattform erstens Forscher*innen und Leser*innen zugute, die mit Hilfe neuer Suchfunktionen relevante und ihnen unbekannte Forschungsliteratur finden sowie neue Zusammenhänge erschließen können. Die Plattform hilft Forscher*innen zudem “to find related content in unfamiliar disciplines or subject areas. In doing so, it breaks them out of the disciplinary or citation-based siloes that they’d been working in […]” (Humphreys 2018). Zweitens profitieren Autor*innen von einer potentiell höheren Sichtbarkeit, da die Suche nach einzelnen Absätzen der Artikel eine neue Einstiegsmöglichkeit in wissenschaftliche Arbeit bietet.

Bibliographie

Degkwitz, Andreas (2013): “What Will Future Publications Be Like?”, in Hobohm Hans-Christoph (ed.): *Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten.* Glückstadt: Verlag Werner Hülsbusch 81-92.

Humphreys, Alex (2018): “Sprechen Sie Textanalysierer? Creating a Multilingual Text Analyzer“. JSTOR Labs Blog. https://labs.jstor.org/blog/#!sprechen_sie_textanalysierer?-creating_a_multilingual_text_analyzer_.

Kohle, Hubertus (2017): “Digitales Publizieren“, in Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds): *Digital Humanities*. Stuttgart: J.B. Metzler 199–205 https://doi.org/10.1007/978-3-476-05446-3_13.

Schwaderer, Christian / Stäcker, Thomas / Walkowski, Niels Oliver / Baillot, Anne / Ernst, Thomas / Baum, Constanze / Chen, Esther / Steyer, Timo / Kaden, Ben / Kleineberg, Michael (2016): “Workingpaper ‘Digitales Publizieren‘“, in *DHd Workingpaper* https://doi.org/10.15499/dhd-wp.001.

Snyder, Ron (2017): “Under the Hood of Text Analyzer“. JSTOR Labs Blog. https://labs.jstor.org/blog/#!under_the_hood_of_text_analyzer.

Söllner, Konstanze (2017): “1a. Warum und für wen Open Access?“, in Söllner, Konstanze / Mittermaier, Bernhard (eds.): *Praxishandbuch Open Access*. Berlin, Boston: De Gruyter 3–11 https://doi.org/10.1515/9783110494068-001.

Spielräume zwischen Yakshis und Dibias: Vladimir Propps Morphologie des Märchens im ontologiegestützten interkulturellen Vergleich

Pannach, Franziska
franziska.pannach@stud.uni-goettingen.de
Georg-August Universität Göttingen, Deutschland

Krishnan, Aravind
aravindhp1999@gmail.com
College of Engineering, Trivandrum, Indien

Vladimir Propps Theorie *Morphology of the Folktale* (Propp 1968) definiert 31 invariante Funktionen, Unterfunktionen und sieben Klassen von Charakteren, um die narrative Struktur der russischen Zaubermärchen zu beschreiben. Seit seiner ersten Veröffentlichung im Jahr 1928 wurde der Ansatz von Propp auf verschiedene Volkserzählungen mit unterschiedlichen kulturellen Hintergründen angewendet (z.B. Azuonye 1990, Okodo 2012 oder Harun und Jamaludin 2016).

Wir haben eine Ontologie erstellt, die die Theorie von Propp modelliert, indem sie narrative Funktionen als Klassen und Relationen implementiert. Ein besonderer Schwerpunkt liegt auf den von Propp definierten Einschränkungen, welche Dramatis Personae eine bestimmte Funktion erfüllen kann. So kann die Funktion *XI Departure* nur durch die Figur der Heldin oder des Helden ausgeführt werden, bricht ein Charakter, der nach Propp zu einer andere Klasse von Dramatis Personae gehört von einem Ort auf, z.B. der Igbo Medizinmann *Dibia* (Helfer) oder die schöne, aber böse *Yakshi* (Gegenspielerin) im indischen Märchen, so greift diese Funktion nicht.

Diese Restriktion definiert Propp sehr streng, sie wurde jedoch unserer Kenntnis nach in keinem Projekt zuvor in einer Ontologie als *range* und *domain* einer zu der Funktion gehörigen Object Property definiert (z.B. Peinado 2004). Die Ontologie wurde auf Grundlage von Noy und McGuiness‘ (2001) Prinzipien erstellt, inklusive einer deskriptionslogischen Grundlage und einer Reihe von Kompetenzfragen, die ein ontologie-gesteuertes System beantworten können sollte. Außerdem wurden zwei ergänzende Ontologien (Koleva 2011 und Declerck 2017) importiert, damit weitere interessante Fragestellungen, wie zum Beispiel die Verbindung von Motifverwendung und Propp’schen Funktionen, oder das Vorkommen von Familienrelationen in Propp’s Dramatis Personae, untersucht werden können.

Abbildung 1 zeigt beispielweise eine Visualisierung der Frage „Wer ist der Held oder die Heldin in den untersuchten Märchen?“

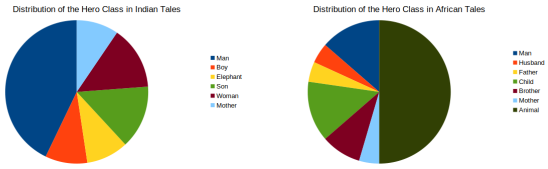


Abbildung 1: Verteilung der Held*Innen-Figuren in indischen und afrikanischen Märchen auf andere Subklassen der Ontologie

Wir haben in diesem Projekt untersucht, wie eine Ontologie die traditionelle geisteswissenschaftliche Forschung dabei unterstützen kann, zu untersuchen, wie gut Propps Theorie für Volksmärchen außerhalb der russisch-europäischen Volkskultur geeignet ist. Wie viel Spielraum bleibt für die Auslegung der Propp’schen Funktionen um sie auf narrative Strukturen von Märchen aus anderen Kulturen anzuwenden?

Zu diesem Zweck wurde ein ontologie-gesteuertes Querysystem mit einem Apache Jena Fuseki[1] Backend implementiert.[2] Um das Browsing in der Ontologie zu ermöglichen, stellen wir gleichzeitig eine institutionelle Webprotégé-Instanz (Tudorache et. al 2008) zur Verfügung.

Analoge Analysen, die auf Propp’s Theorie basieren, können mit wenig Einarbeitung in die Ontologie eingepflegt und so in den Kontext anderer Analysen gestellt werden.

Um festzustellen, wie gut sowohl das an Propp angelegte Annotationsschema als auch das Abfragesystem funktionieren, haben wir zwanzig hauptsächlich sub-saharische und fünfzehn südindische (Kerala) Märchen und Volkserzählungen annotiert.

Wir evaluieren das System, indem wir zwei Fallstudien über die Repräsentation von Charakteren und die Verwendung propp‘scher Funktionen in afrikanischen und indischen Geschichten untersuchen. Unsere Ergebnisse stehen im Einklang mit der traditionellen analogen geisteswissenschaftlichen Forschung, z.B. Reuster-Jahn’s (2002) Arbeiten zum fehlenden guten Ausgang in afrikanischen Märchen. Insbesondere die Funktionen des *Dénouement*, die Propp für die Ausgänge des Russischen Zaubermärchens definiert, kommen in Märchen der untersuchten Kulturkreise kaum vor. Viel eher bestätigen die Daten, dass die Ausgänge in afrikanischen Märchen anderer Natur sind. In unseren Daten zeigt sich, dass sowohl die sub-saharischen als auch die Märchen aus dem indischen Kulturkreis eher auf die Wiederherstellung des Status-Quo ausgerichtet sind, die Bekämpfung der Not oder Mangelsituation stehen im Fordergrund, ebenso die Rückkehr der Held*Innen und Opfer nach Hause oder an einen sicheren Ort. Die Propp‘schen Anfangsfunktionen, insbesondere die *Absentation*, die *Interdiktion* oder *das Verbot* und der *Bruch des Verbots*, sind jedoch sehr prominent in allen Märchen vertreten.

Propp’s Theorie hat nicht den Anspruch, für Märchen aller Kulturkreise gleich gut adaptierbar zu sein, was sich beispielsweise an den beschriebenen Mängeln seiner Auswahl von Endfunktionen ablesen lässt.

Eine weitere Schwierigkeit besteht in der individuellen Auslegung propp‘scher Funktionen durch die jeweiligen Annotator*Innen. So verwendet Azuonye (1990) die Funktion *Transformation* um eine moralische Transformation des Opfers und der Gesellschaft im Märchen *Obaraedo* zu codieren. Okodo

(2012) folgt dieser sehr freien Auslegung bei seiner Analyse desselben Märchens nicht. Unser System kann hier durch die Verwendung von Querverweisen zwischen den einzelnen Analysen und dem Einsatz von rdfs:comments auf individuelle Auslegungen eingehen.

Dieses Projekt zeigt, wie sorgfältig modellierte Ontologien traditionelle literaturwissenschaftliche bzw. folkloristische Theorien darstellen und zugänglich machen können. Außerdem wollen wir zeigen wie sie als Wissensbasis für die vergleichende Folkloreforschung genutzt werden können.

Das Poster stellt die Designprinzipien der Ontologie und des darauf basierenden Ontologie-Query-Systems dar und visualisiert die Ergebnisse der Auswertungen.

Fußnoten

1. https://jena.apache.org/documentation/fuseki2/
2. https://teaching.gcdh.de/ontology/index

Bibliographie

Azuonye, Chukwuma (1990): „Morphology of the Igbo folktale: Ethnographic, historiographic and aesthetic implications“, in: *Folklore*, 101(1): 36-46.

Declerck, Thierry / Kostová, Antónia / Lisa Schäfer (2017): „Towards a linked data access to folktales classified by Thompson's motifs and Aarne-Thompson-Uther's types“, in: *Proceedings of Digital Humanities 2017*. ADHO, 8 2017.

Harun, Harryizman / Jamaludin, Zulikha (2016): „Structural Classification as Preservation Means of Malaysian Folktales“, in: *Proceedings of the International Soft Science Conference (ISSC'16)*.

Koleva, Nikolina (2011): *Ontology-based iterative detection of characters and their recognition in folktales*, Bachelorarbeit, Universität des Saarlands.

Noy, Natalya / McGuinness, Deborah L. (2001): „Ontology development 101: A guide to creating your first ontology.“ *KSL-01-05*, Knowledge Systems Laboratory, Stanford University.

Okodo, Ikechukwu (2012): „Obaraedo: Conformity to Proppian morphology.“ *AFRREV IJAH: An International Journal of Arts and Humanities*, 1(2): 100-111.

Peinado, Federico / Gervás, Pablo / Díaz-Agudo, Belén (2004): „A description logic ontology for fairy tale generation“, in: *Proceedings of the Workshop on Language Resources for Linguistic Creativity*, LREC, 4: 56-61.

Propp, Vladimir (1968): *Morphology of the Folktale*, volume 10. Austin: University of Texas Press.

Reuster-Jahn, Uta (2002): „Gute und schlechte Ausgänge in europäischen Märchen und in den Volkserzählungen der Mwera in Tansania“, in: *Märchenspiegel*, 13(2): 17-18.

Tudorache, Tania / Vendetti, Jennifer / Noy, Natalya (2008): „Web-protege: A lightweight OWL ontology editor for the web“, in: *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008)*.

Stilometrische Untersuchung von Figurenreden in realistischen Erzähltexten

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Einführung

Das Poster stellt ein Korpus deutschsprachiger Erzählungen des 19. Jahrhunderts vor, in dem Figurenreden und ihre jeweiligen Sprecher annotiert und extrahiert wurden. Sie dienen als Basis für stilistische Auswertungen mit dem etablierten Abstandsmaß Delta. Es stellt sich die Frage, ob sich der Autorenstil in den jeweiligen Figurenreden niederschlägt, sich also Figuren desselben Autors zusammengruppieren, oder ob Figurentypen dominanter sind, sich gleiche Figurentypen also werkübergreifend stilistisch ähneln. Erste Ergebnisse hiervon werden als Grafiken präsentiert.

Verwandte Forschung

Stilometrische Verfahren gehen v.a. auf John Burrows zurück. Sein entwickeltes Abstandsmaß *Delta* (Burrows 2002) gilt als Standardverfahren in der Stilometrie und es existieren zahlreiche Studien und Verbesserungsvorschläge (z.B. Smith/Aldridge 2011, Büttner et al. 2017). Für die einfache informatische Anwendung wurde es durch das R-Package *stylo* (Eder/Rybicki/Kestemont 2016) zugänglich gemacht. Die ersten quantitativen Untersuchungen des Figurenstils liefert ebenfalls erstmals Burrows (1987) in der anglistischen Literatur. Allerdings führt die unterschiedlich große Menge an Reden pro Figur zu disparatem Analysematerial. Um das Problem unterschiedlich langer Texte zu umgehen, nutzt Hoover (2017) Textauszüge bzw. zufällige Textanordnung in seiner Studie zur intratextuellen Stilvariation. Stilometrische Analysen erfreuen sich auch in der heutigen Forschung noch hoher Beliebtheit (so z.B. Bonch-Osmolovskaya/Skorinkin 2019, auf Dramentexte Galleron 2019).

Korpus: Annotation und Datenaufbereitung

Das Korpus setzt sich aus acht realistischen Erzähltexten zwischen 1848 und 1871 zusammen, da dieser Zeitraum allgemein als Kernzeit des Realismus anerkannt ist (Aust 2006, Plumpe 2007). Um Vergleiche zu ermöglichen, enthält das Korpus zusätzlich drei Erzähltexte von vor 1848. Die Korpu-

sauswahl beruht auf einem mehrschrittigen Prozess: Mit der Längenbegrenzung von 8.000-20.000 Wörtern wurde darauf geachtet, dass die Erzählungen einerseits lang genug sind, um stilometrische Verfahren anwenden zu können und andererseits kurz genug, um die manuelle Annotation in einem angemessenen zeitlichen Rahmen durchzuführen. Außerdem wurde darauf geachtet, sowohl kanonisierte als auch gänzlich unbekannte Texte zu integrieren, weibliche Autoren ins Korpus aufzunehmen und die Erstpublikationsorgane zu variieren. Wie in der damaligen Zeit üblich, wurde ein Großteil der Erzählungen in Zeitschriften, Almanachen oder Taschenbüchern veröffentlicht. Diese waren auf ganz verschiedene Leserschichten ausgerichtet, so dass eine Variation hier alle Stilniveaus erfassen sollte. Die Korpustexte sind die folgenden elf Erzählungen:

Titel	Autor	Jahr	Wortanzahl
Der Gefangene	Malsburg, Otto von der	1822	9.108
Die Doppelgängerin	Ungern-Sternberg, Alexander von	1834	8.094
Die Judenbuche	Droste-Hülshoff, Annette von	1842	16.191
Der arme Spielmann	Grillparzer, Franz	1848	15.132
Das Erdbeerimareili	Gotthelf, Jeremias	1850	15.720
Bergmilch	Stifter, Adalbert	1853	9.727
Phosphorus Hollunder	François, Louise von	1857	14.082
Die schwarze Galeere	Raabe, Wilhelm	1861	15.585
Die zwölf Apostel	Marlitt, Eugenie	1865	20.288
Eine Malerarbeit	Storm, Theodor	1867	9.392
Der Leuchtturm von Livorno	Eckstein, Ernst	1871	8.992

Tabelle 1: Im Korpus enthaltene Erzählungen.

Da einige der Texte noch nicht erschlossen waren, wurden sie vor der Annotation OCR-korrigiert. Für die Annotation wurde der im Zuge des Redewiedergabe-Projekts (Brunner et al. 2018) entstandene STWR-View des Annotationstools ATHEN (Krug et al. 2018) verwendet. Bei der Annotation wurden sämtliche direkten Figurenreden manuell annotiert und ihrem jeweiligen Sprecher zugeordnet (zur automatischen Zuordnung von Sprechern: Krug et al. 2016). So konnte die gesamte direkte Redemenge einzelner Figuren extrahiert werden. In direkte Reden einer Figur A eingelagerte Reden einer Figur B wurden dabei nur der Figur B als zugehörig annotiert. Auf diese Weise wurde sichergestellt, dass Figuren ausschließlich ihre eigenen Reden zugeordnet wurden (diese Problematik ist besonders relevant bei Binnenerzählungen). Zusätzlich wurden ausschließlich Figuren in die Auswertung integriert, deren gesamte Redemenge 200 Wörter übersteigt, um stilometrische Verfahren wirksam anwenden zu können. Diese Grenze ist für stilometrische Verfahren noch immer vergleichsweise niedrig. Eder (2015) hat evaluiert, dass korpusabhängig mindestens 2500-5000 Wortformen nötig sind, damit Auswertungen mit Delta zu guten Ergebnissen führen. Aufgrund des Korpus dieser Studie kann dieser Mindestwert allerdings nicht eingehalten werden.

Auswertung

Die folgenden Grafiken zeigen den Output des R-package *stylo* (Eder/Rybicki/Kestemont 2016) erstens der 100 häufigsten gesprochenen Wörter der Figuren und zweitens der 1000 häufigsten. Es wurde kein Sampling durchgeführt und ebenfalls kein Culling, ein Feature musste folglich nicht in einer bestimmten Anzahl Texte vorhanden sein, um in die Auswertung einbezogen zu werden. Als Abstandsmaß wurde klassisch Burrows' Delta gewählt, die Outputgrafiken sind Cluster-Analysen, die stilistisch ähnliche Figuren zueinander gliedern. Zu beachten ist die oben erwähnte Mindestmenge von 200 Wörtern pro Figur. Das führt dazu, dass bei der Analyse der 1000 häufigsten Wörtern von einigen Figuren alle gesprochenen Wörter in der Auswertung enthalten sind.

Auswertung mit 100 häufigsten Wörtern:

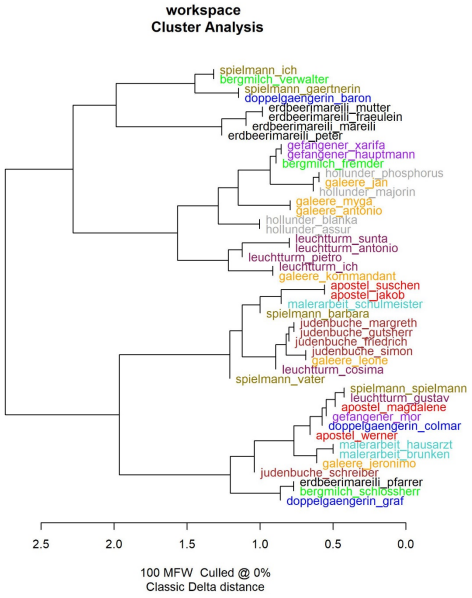


Abbildung 1: Auswertung mit 100 häufigsten Wörtern.

Auswertung mit 1000 häufigsten Wörtern:

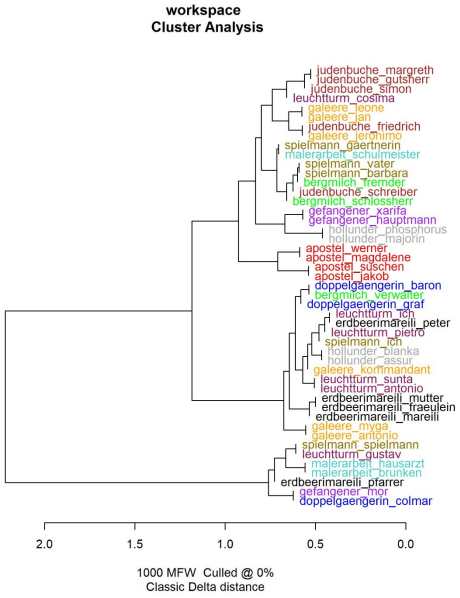


Abbildung 2: Auswertung mit 1000 häufigsten Wörtern.

In beiden Auswertungen ist zu erkennen, dass sich häufig Figuren desselben Autors zueinander gliedern. Besonders

beim Mundartdichter Gotthelf (*Erdbeerimareili*) ist das sehr verständlich. Dennoch gibt es Abweichungen. Besonders bei den 1000 häufigsten Wörtern gruppieren sich auf dem untersten Ast die Figuren mit der größten Redemenge zusammen. Dies sind häufig Binnenerzähler, die in ihrem Redestil häufig schnell die Funktion und den Stil von Erzählerrede einnehmen (Bockwinkel 2016). Um zu untersuchen, ob das Clustering nur der insgesamt größeren Redemenge dieser Figuren geschuldet ist, wurde in mehreren Analysen Sampling durchgeführt. Die hier nicht abgebildeten Auswertungen bestätigen das Ergebnis, wenngleich der Abstand der Binnenerzähler zu den übrigen Figuren geringer wird. Außerdem nimmt Colmar aus der *Doppelgängerin* einen größeren Abstand zu den übrigen Binnenerzählern ein. Auch das ist nachvollziehbar, da Colmar im Gegensatz zu ihnen nur über einen kleinen Teil der Erzählung als Binnenerzähler fungiert und sonst wie eine „normale“ Figur agiert. Figurentypen gliedern sich in dieser ersten Vorstudie dagegen nicht zusammen. Figurenpaare, die sich in gegenseitiger Liebe befinden (wie Magdalene-Werner, Xarifa-Mor, Myga-Jan, Cosima-Antonio) gruppieren sich nur teilweise als Paar und gar nicht als Figurengruppe. Weitere Schlussfolgerungen, dass sich beispielsweise gleiches Geschlecht, Figuren aus ähnlichen Subgenres (Abenteuer/Liebe) oder Erzählungen aus einer bestimmten Epoche gruppieren, können in dieser ersten Vorstudie ebenfalls noch nicht gezogen werden. Gleichfalls kann diese Studie aber auch noch nicht als Beweis fungieren, dass sich deren Stil nicht ähnelt.

Ausblick

Im weiteren Verlauf der Arbeit müssen die Maße verfeinert und sollen andere Abstandsmaße getestet, Variablen geändert und Ergebnisse evaluiert werden. Die Problematik der Kürze der Texte könnte durch eine Optimierung des Verfahrens verringert werden. So könnten eine Kombination aus Wortform-*Grammar-Tags* und besonders gut zur Autorschaftsattribution geeigneter Wörter Verbesserungen bringen (Dimpel 2019). Eine Integration von Gedanken- und Schriftzitaten ist ebenfalls denkbar. Interessant wäre auch die Berücksichtigung von indirekter Rede, da hier ebenfalls die Figurenstimme stark ist. In einem Schritt weg von der Stilometrie sollen in späteren Tests darüber hinaus Topic Modeling und Sentimentanalyse durchgeführt werden, um die Figurenreden auch auf diesen Ebenen zu vergleichen.

Bibliographie

Aust, Hugo (2006): *Realismus*. Lehrbuch Germanistik, Stuttgart: Metzler.

Bockwinkel, Peggy (2018): "Wie anders ist Figurenrede? Die Rolle der direkten Rede in quantitativen Erzähltextanalysen", in: Bockwinkel, Peggy / Nickel, Beatrice / Viehhauser, Gabriel (eds.): *Digital Humanities. Perspektiven der Praxis.* Berlin: Frank&Timme 117-148.

Bonch-Osmolovskaya, Anastasia / Skorinkin, Daniil (2019): "The Complexity of Character-building: Speech, Portraits, Interactions in Leo Tolstoy's 'War and Peace'", in: *Conference Abstracts of DH2019 Utrecht*.

Brunner, Annelen / Engelberg, Stefan / Jannidis, Fotis / Tu, Ngoc Duyen Tanja / Weimer, Lukas (2018): "Projektvorstellung – Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse", in: *Konferenzabstracts der DHd2018 Köln* 458-460.

Burrows, John (1987): *Computation into Criticism*. A Study of Jane Austen's Novels and an Experiment in Method, Oxford: Clarendon Press.

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

Büttner, Andreas / Dimpel, Michael / Evert, Stefan / Jannidis, Fotis / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (2017): "'Delta' in der stilometrischen Autorschaftsattribution", in: *Zeitschrift für digitale Geisteswissenschaft* 2.

Dimpel, Friedrich Michael / Zeppezauer-Wachauer, Katharina / Schlager, Daniel (2019): "Der Streit um die Birne. Autorschafts-Attributionstest mit Burrows' *Delta* und dessen Optimierung für Kurztexte am Beispiel der ‚Halben Birne' des Konrad von Würzburg", in: *Das Mittelalter* 24/1: 71-90.

Eder, Maciej (2015): "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Digital Scholarship in the Humanities* 30/2: 167-182.

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

Galleron, Ioana (2019): "Stylometric Analyses of Character Speeches in French Plays", in: *Conference Abstracts of DH2019 Utrecht*.

Krug, Markus / Jannidis, Fotis / Reger, Isabella / Macharowsky, Luisa / Weimer, Lukas / Puppe, Frank (2016): "Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts. Methoden zur Bestimmung des Sprechers und des Angesprochenen", in: *Konferenzabstracts der DHd2016 Leipzig* 181-186.

Krug, Markus / Tu, Ngoc Duyen Tanja / Weimer, Lukas / Reger, Isabella / Konle, Leonard / Jannidis, Fotis / Puppe, Frank (2018): "Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment", in: *Konferenzabstracts der DHd2018 Köln* 19-21.

Plumpe, Gerhard (2007): "Realismus", in: Müller, Jan-Dirk / Braungart, Georg / Fricke, Harald / Grubmüller, Klaus / Vollhardt, Friedrich / Weimar, Klaus (eds.): *Reallexikon der deutschen Literaturwissenschaft* 3. Berlin / New York: De Gruyter 221-224.

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

StreamReaderSD 0.2 – Eine prototypische Webanwendung für das Lesen von Texten als Zeichenstrom

Drach, Sviatoslav
s.drach@uni-koeln.de
Universität zu Köln, Deutschland

Einleitung

Lesen auf Papier und am Bildschirm sind nicht gleich. Studien besagen, dass digitales Lesen zu flüchtigem, weniger konzentrierten Lesen führe (Stavanger Declaration 2019). Was kann getan werden, um dieses Problem zu lösen? Ist das Lesen von Text als Zeichenstrom eine mögliche Lösung?

Es gibt bislang keine Anwendung, in der man (längere) Texte als einen einzeiligen Strom von Zeichen rezipieren würde. Es existieren einige (auf den ersten Blick sehr ähnliche) Tools für Speed Reading, aber hier geht es nicht um Schnelllesen. Das hier vorgestellte Projekt ist explorativ. Es basiert auf konzeptionellen Überlegungen, die Patrick Sahle (2020) in einem Vortrag auf dieser Tagung vorstellt.

Text als Zeichenstrom lesen

Unser trainiertes Leseverhalten folgt u.a. den Augenbewegungsprinzipien der Sakkaden (Springen zwischen Textelementen) und der Fixationen (Informationsaufnahme bei fixierten Punkten) (Landau 2016: 15). Dies scheint auf den ersten Blick dem Lesen von Text als Zeichenstrom zu widersprechen, weil die ständige Bewegung des Textes die Sakkaden und Fixationen erschwert. Auch die typographischen Prinzipien, die gute Lesbarkeit ausmachen, basieren ganz auf dem zweidimensionalen Schriftraum und lassen sich nicht einfach auf einen laufenden Strom übertragen (Hegewald et al. 2011: 27).

Trotz aller Aspekte, die gegen das Lesen von Text als Zeichenstrom sprechen, kann ein Experiment interessant sein. Zum einen gibt die Bewegung dem Text seinen Fluss zurück, den er in der gesprochenen Sprache noch hatte und der eine sorgfältige Rezeption verlangt, um dem Ablauf der Informationsübermittlung zu folgen. Zum anderen entstehen durch die Befreiung des Darstellungsraumes der "Seite" neue Möglichkeiten für die Medialisierung von Texten. Denn die Textpräsentation kann jenseits der einfachen laufender Zeile nun den Raum nutzen, um beispielsweise Bilder, Fußnoten oder mehrere Zeilen darzustellen.

Eine prototypische Anwendung

Die Web-App StreamReader$_{SD}$ bietet eine digitale Umgebung für das Lesen von Text als Zeichenstrom. Dabei handelt es sich um eine "Proof of Concept"-Lösung, die eine grundsätzliche Realisierbarkeit auslotet. Aus technischer Sicht beschränkt sie sich zunächst auf gängige Webtechnologien wie HTML, CSS und JavaScript.

Abbildung 1. StreamReader$_{SD}$ (Screenshot; vergrößert und beschriftet. Quelle: http://dev.cceh.uni-koeln.de/sr-sd)

Zu den allgemeinen Funktionen der Anwendung gehört die Möglichkeit, einen Text entweder aus einem Repositorium auszuwählen (2) oder eigene Texte hinzuzufügen (3). Allgemeine Einstellungsoptionen betreffen Layoutmerkmale wie Breite und Hintergrundfarbe des Anzeigebereichs, Schriftfamilie, Schriftfarbe oder Schriftgröße (4). Beim Lesen kann ein Lesezeichen gesetzt werden (5). Eine Suchfunktion ermöglicht das Finden von Wörtern (6). Statistische Berechnungen (8) veranschaulichen die Anzahl von Zeichen und Wörtern im Text sowie die eingeschätzte Lesezeit bei maximaler und minimaler Lesegeschwindigkeit. Die Steuerung der Anwendung erfolgt mit Hilfe von Bedientasten (9a) sowie einem Geschwindigkeitsschieber (9b). Die Anwendung kann ebenso mit Hilfe der Tastatur gesteuert werden. Eine kurze Anleitung dazu befindet sich unter dem Hilfe-Button (7). Mit dem Home-Button (1) kann die Anwendung neu gestartet werden.

Wenn ein Text über eine Seitenzählung verfügt, dann wird eine seitenorientierte Navigation (10a) generiert und die Nummer der aktuellen Seite angezeigt (10b). Andernfalls basiert die allgemeine Navigationsleiste auf einem hinsichtlich der Zeichenmenge prozentualen Verständnis der aktuellen Position. Ist ein Text durch Kapitel oder andere Einheiten strukturiert, dann werden zwei Typen von Inhaltsverzeichnissen generiert: Eine konventionelle Liste mit Links zu den jeweiligen Kapiteln und Unterkapiteln (11), sowie eine visuelle Navigationsleiste, die proportional die Größe der Kapiteln und Unterkapiteln abbildet und ebenfalls zum gezielten Ansteuern von Textstellen genutzt werden kann (12).

Zu den eher textspezifischen Funktionen der Anwendung gehören weiterhin der Umgang mit Fußnoten und Illustrationen. Beide laufen zunächst mit dem Zeichenstrom ein, können dann aber pausiert (Illustrationen) oder als unabhängig laufende Schrift gelesen werden, während der Haupttext angehalten wird (Fußnoten).

Es besteht die Möglichkeit, Schrift in laufenden Zeilen versetzt beziehungsweise mehrere Zeilen parallel darzustellen. Dies ist zum Beispiel sinnvoll, wenn ein Text komplexe Nebensatzstrukturen enthält, die durch versetzte Darstellung besser wahrgenommen werden können oder wenn unterschiedliche Übersetzungen desselben Werkes parallel gelesen werden sollen.

Ausblick

Effekte des StreamReader$_{SD}$ auf das Leseverhalten, die Textrezeption und die Informationsaufnahme sind bisher nicht untersucht worden. Dies würde eigene Studien erfordern, steht aber aktuell nicht im Fokus der Arbeiten. Bei der prototypischen Anwendung geht es derzeit um ein Ausloten der Mög-

lichkeiten, die Entwicklung von ersten Funktionen und das Testen von Darstellungsoptionen.

Die im StreamReader$_{SD}$ aktuell hinterlegten Texten demonstrieren Leseszenarien für unterschiedliche textuelle Situationen. Jenseits der einfachen laufenden Zeile werden Lösungen für den Umgang mit Illustrationen oder Fußnoten angeboten, Effekte eines Zeilenversatzes als neues Lese-Strukturelement beobachtbar gemacht und "mehrspurige" Texte realisiert. Dadurch kann ausgelotet werden, welche neuen Möglichkeiten für die Darstellung von bestimmten Textsorten entstehen, die vielleicht für die Präsentation und Rezeption nützlich sein könnten.

Für die Zukunft wäre es denkbar, weitere Beispieltexte, die andere Probleme aufwerfen, aufzunehmen und dazu weitere Funktionalitäten zu implementieren. Die visuelle Gestaltung der Anwendung wäre noch zu professionalisieren. Nutzungsstudien mit Lesern könnten interessante Einblicke in Stärken und Schwächen des Ansatzes erlauben.

Bibliographie

Filek, Jan / Uebele, Andreas (2013): *Read/ability. Typografie und Lesbarkeit*, Sulgen: Niggli.

Hegewald, Falk / Tritschler, Johannes / Hien, Katharina / Rümpler, Steffen (2011): *Typografische Animation für Studium und Praxis*, Berlin, Heidelberg: Springer.

Korthaus, Claudia (2016): *Grundkurs Typografie und Layout. Für Ausbildung und Praxis*, 5. Auflage, Bonn: Rheinwerk Verlag.

Landau, Angelina (2016): *Wie das Gehirn liest. Die neuronalen Prozesse beim Lesen*, Marburger Schriften zur Lehrerbildung, Bd. 12, Marburg: Tectum Verlag.

Myrberg, Caroline / Wiberg, Ninna (2015): *Screen vs. paper: what is the difference for reading and learning?* Insights: the UKSG Journal 28, 49–54 <https://doi.org/10.1629/uksg.236>.

Sahle, Patrick (2020): *Wie wir lesen könnten. StreamReaderPS 0.1*. Book of Abstracts zur DHd2020, Paderborn.

[Members and relevant stakeholders of the EU funded COST research Action E-READ]: *COST E-READ* Stavanger Declaration *Concerning the Future of Reading* (2019). < http://ereadcost.eu/wp-content/uploads/2019/01/StavangerDeclaration.pdf >.

Topic Modeling der Hugo-Schuchardt-Korrespondenz – Möglichkeiten und Grenzen

Saric, Sanja
sanja.saric@uni-graz.at
Universität Graz, Österreich

Scholger, Martina
martina.scholger@uni-graz.at
Universität Graz, Österreich

Einleitung

Digitale Analyseverfahren verändern immer intensiver die Forschungsweise der GeisteswissenschaftlerInnen und mit dem wachsenden Spielraum der Methoden wächst auch die Anzahl an Fragen, die sich vor allem an den Grad der Genauigkeit und wissenschaftliche Relevanz dieser Methoden richtet. Das Topic Modeling gewinnt als eine Methode für automatische Erkennung von versteckten thematischen Strukturen in großen Textmengen (Blei 2012: 8) immer mehr an Beliebtheit, erweckt aber auch Unsicherheiten. Daher beschäftigt sich diese Arbeit mit den Möglichkeiten und Problemen des Topic Modeling am Beispiel von Briefen und stellt unter anderem die Fragen, 1) wie Topic Modeling in der Analyse von Briefkorpora eingesetzt werden kann und 2) wie die Qualität der Ergebnisse dieses Prozesses beeinflusst werden kann.

Forschungsmaterial

Das Forschungsmaterial besteht aus Briefen des Grazer Sprachwissenschaftlers Hugo Schuchardt (1842-1927). Die umfangreiche und mehrsprachige Korrespondenz dieses schon seinerzeit sehr geschätzten Wissenschaftlers ist seit 2007 Teil des Digitalisierung-Projektes *Hugo Schuchardt Archiv* (Hurch 2019). Für die Topic-Modeling-Analyse werden 2261 Briefdateien im TEI-Format in Betracht gezogen, da die restlichen zurzeit noch in keinem entsprechenden Format vorhanden sind. Der Vorteil einer solchen Methode ist es aber, dass das gleiche Modell jederzeit auf eine erweiterte Menge an Daten anwendbar ist. Eine Besonderheit dieses Korpus ist, dass Schuchardt in mehreren Sprachen korrespondiert hat, von denen hier elf repräsentiert sind (Abbildung 1). Daher wird das Modell für einzelne Sprachen separat angewendet. Dies ist insofern eine Herausforderung, weil 1) Vorgänge den jeweiligen Sprachen angepasst werden müssen (wie etwa die Lemmatisierung), 2) der Textumfang bei vielen Sprachen nicht ausreichend ist und daher nicht auf alle Sprachen effektiv angewendet werden kann und 3) die verschiedenen Ergebnisse pro Sprache verglichen werden sollten. Ein weiteres Problem für das Topic Modeling ist die große Diskrepanz in den Textlängen der einzelnen Dateien (Abbildung 2), da die Korrespondenz auch kürzere Formen wie Postkarten und Telegramme beinhaltet. So enthalten etwa die kürzesten deutschsprachigen Dateien etwa drei Tokens, die längste jedoch 3947. Dies ist aber ein Zustand, den viele Briefkorpora in der Realität begegnen, da wir als ForscherInnen selten einem ‚idealen' Korpus gegenüberstehen. Die Auseinandersetzung mit solchen Problemen ist ein fester Bestandteil unserer Arbeit.

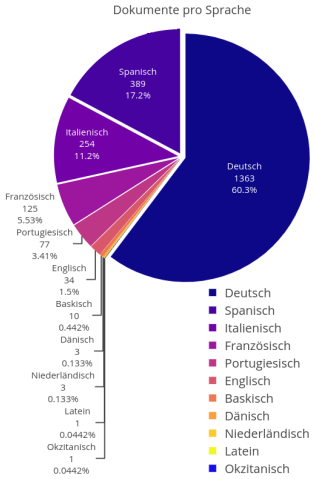


Abbildung 1: Anteil der einzelnen Sprachen im Briefkorpus

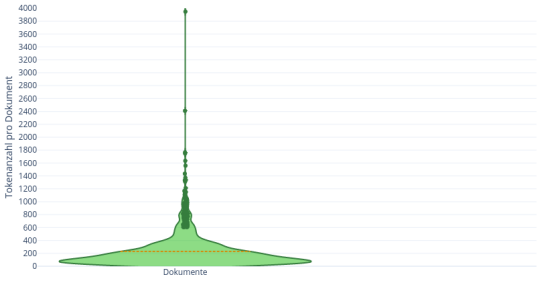


Abbildung 2: Menge der deutschsprachigen Briefdateien nach ihrer Anzahl der Tokens

Methode

Für die Beantwortung der Forschungsfragen war zuerst die Literaturrecherche nötig, und zwar erstens zum Topic Modeling, zweitens zur Textsorte Brief und drittens zu dieser Korrespondenz. Um eine genauere Vorstellung zum Forschungsstand des Topic Modeling zu bekommen, wurden wissenschaftliche Aufsätze und Anwendungsbeispiele in Betracht gezogen, wie etwa Blei 2010, Jagarlamudi/Daumé 2010, Boyd-Graber/Blei 2012, Riddell 2015, Vulić et al. 2015, Bock et al. 2016, Andorfer 2017, Fechner/Weiß 2017, Schöch 2017, Murakami et al. 2017 und Arora et al. 2018. Zudem wird am genannten Korpus Topic Modeling mit Hilfe der Programmiersprache *Python* (Python Software Foundation 2001-2019), der Software MALLET (McCallum 2002-2019) und der Anweisungen der Jupyter-Notebooks von DARIAH-DE (DARIAH-DE 2019) vollzogen. Darüber hinaus werden verschiedene Tools zur Vorverarbeitung evaluiert – z. B. *spaCy* (Explosion AI 2019) und DTA::CAB (Berlin-Brandenburgische Akademie der Wissenschaften 2011-2018) für die Lemmatisierung – sowie verschiedene Tools und Parameter für die Topic-Modellierung – z. B. *Topics Explorer* (DARIAH-DE 2018) – und die daraus resultierenden Ergebnisse und Erfahrungen verglichen.

Ergebnisse

Obwohl es sich um ein laufendes Projekt handelt, gibt es bereits einige relevante Ergebnisse und Schlussfolgerungen.

1) Die Vorverarbeitung stellt einen wichtigen Schritt in der Topic-Modellierung dar und beeinflusst die Ergebnisse. Dabei spielen nicht nur die eingesetzten Tools eine Rolle, sondern auch die gewählte Vorgehensweise.

2) Die Lemmatisierung, auf die beim Topic Modeling oft verzichtet wird, ermöglicht mehr semantische Differenz in den Topics.

3) Der unterschiedliche Textumfang von einzelnen VerfasserInnen kann zu falschen Ergebnissen führen, wenn die Topics pro VerfasserIn analysiert werden.

4) Entscheidungen über Parameter wie Optimierungsintervall, Topic- und Iterations-Anzahl können die Ergebnisse beeinträchtigen und müssen immer projektspezifisch getestet werden, bis ein sinnvolles Resultat vorliegt. Das ‚Sinnvolle' zu erkennen ist eine Herausforderung, die fachwissenschaftliches Verständnis verlangt.

Die Inkonsistenz der Topics und manchmal verwirrende Ergebnisse zeigen, dass die naive Anwendung eines Topic-Modeling-Tools nicht immer befriedigend sein kann. Intensivere Beschäftigung mit den einzelnen Schritten und Ergebnissen kann sich jedoch positiv auf den Erfolg der Analyse auswirken. Die weitere Arbeit wird zeigen, ob und welchen Mehrwert Topic Modeling bei der Analyse der Schuchardt-Korrespondenz leisten kann, die durch *close reading* nicht erreicht werden können.

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* 2. http://zfdg.de/2017_002 [letzter Zugriff 27. September 2019].

Arora, Sanjeev / Ge, Rong; Halpern, Yoni / Mimno, David / Moitra, Ankur / Sontag, David / Wu, Yichen / Zhu, Michael (2018): "Learning topic models - provably and efficiently", in: *Communications of the ACM* 61 / 4: 85–93. 10.1145/3186262.

Berlin-Brandenburgische Akademie der Wissenschaften (ed.) (2011-2018): *Das DTA-Basisformat*. http://www.deutschestextarchiv.de/doku/basisformat/ [letzter Zugriff 27. September 2019].

Blei, David M. (2010): "Introduction to Probabilistic Topic Models", in: *Semantic Scholar*. https://pdfs.semanticscholar.org/5f10/38ad42ed8a4428e395c96d57f83d201ef3b3.pdf [letzter Zugriff 27. September 2019].

Blei, David M. (2012): "Topic Modeling and Digital Humanities", in: *Journal of Digital Humanities* 2 / 1: 8–11. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ [letzter Zugriff 27. September 2019].

Bock, Sina / Du, Keli / Huber, Michael / Pernes, Stefan / Pielström, Steffen (2016): *Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften. Bericht über den Stand der Forschung.* (= DARIAH-DE working papers 18). Göttingen: GOEDOC – Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen. http://webdoc.sub.gwdg.de/pub/

mon/dariah-de/dwp-2016-18.pdf [letzter Zugriff 27. September 2019].

Boyd-Graber, Jordan / Blei, David (2012): *Multilingual Topic Models for Unaligned Text*. http://arxiv.org/pdf/1205.2657v1 [letzter Zugriff 27. September 2019].

DARIAH-DE (2018): *Topics Explorer*. V. 2.0.1. https://github.com/DARIAH-DE/TopicsExplorer [letzter Zugriff 27. September 2019].

DARIAH-DE (2019): *DARIAH Topics. Easy Topic Modeling in Python*. V. 2.0.1. https://github.com/DARIAH-DE/Topics [letzter Zugriff 27. September 2019].

Explosion AI (2019): *spaCy*. V. 2.1.6. https://github.com/explosion/spaCy [letzter Zugriff 27. September 2019].

Fechner, Martin / Weiß, Andreas (2017): "Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts", in: *Zeitschrift für digitale Geisteswissenschaften* 2. http://zfdg.de/2017_005 [letzter Zugriff 27. September 2019].

Hurch, Bernhard (2019): "Hugo Schuchardt Archiv". Institut für Romanistik, Karl-Franzens-Universität Graz (ed.). https://schuchardt.uni-graz.at [letzter Zugriff 27. September 2019].

Jagarlamudi, Jagadeesh / Daumé, Hal (2010): "Extracting Multilingual Topics from Unaligned Comparable Corpora", in: Gurrin, Cathal (ed.): *Advances in information retrieval. Proceedings* 444–456. (= Lecture notes in computer science 5993). Berlin / Heidelberg / New York: Springer.

McCallum, Andrew Kachites (2002-2019): *MALLET. A Machine Learning for Language Toolkit*. V. 2.0.8. http://mallet.cs.umass.edu [letzter Zugriff 27. September 2019].

Murakami, Akira / Thompson, Paul / Hunston, Susan / Vajn, Dominik (2017): "'What is this corpus about?': using topic modelling to explore a specialised corpus", in: *Corpora* 12 / 2: 243–277. https://www.euppublishing.com/doi/10.3366/cor.2017.0118 [letzter Zugriff 27. September 2019].

Python Software Foundation (2001-2019): *Python*. V. 3.7.4. https://github.com/python [letzter Zugriff 27. September 2019].

Riddell, Allen (2015): *Text Analysis with Topic Models for the Humanities and Social Sciences — Text Analysis with Topic Models for the Humanities and Social Sciences*. DARIAH-DE Initiative (ed.). https://liferay.de.dariah.eu/tatom/ [letzter Zugriff 27. September 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 27. September 2019].

Vulić, Ivan / Smet, Wim de / Tang, Jie / Moens, Marie-Francine (2015): "Probabilistic topic modeling in multilingual settings. An overview of its methodology and applications", in: *Information Processing & Management* 51 / 1: 111–147. https://www.sciencedirect.com/science/article/pii/S0306457314000739 [letzter Zugriff 27. September 2019].

Wiener Ballette <Tanz Musik="mei" Bild="jpg" Text="tei" Bewegung="?"/>

Vera, Grund

vgrund@mail.uni-paderborn.de
Universität Paderborn

Henner, Drewes

henner.drewes@folkwang-uni.de
Folkwang Universität der Künste Essen

Die Auseinandersetzung mit der ephemeren Kunstform Tanz und insbesondere mit der Aufführungsform Ballett bedeutet zwangsläufig den Umgang mit „Interpretationsspielräumen". Durch die Mediatisierung von Tanz vor der Möglichkeit von Videoaufzeichnungen als Wort- oder Bilddokumente und selbst bei der Übertragung in Tanznotationen bestehen Leerstellen, die nur durch Ausprobieren oder durch eine experimentierende Herangehensweise ausgefüllt werden können. Interpretation ist daher ein unvermeidlicher Faktor im Umgang mit Tanz, re-staging eine probate Methode der historischen Tanzwissenschaft. Durch moderne Bewegungsästhetik vorgeprägte Körper sind dafür ein zusätzlicher unbewusst interpretierender Faktor.

Für die digitale Annäherung an Bewegung wurde unter Bezugnahme auf frühere Versuche (Schiphorst 1997) und weiter führende Konzepte (Calvert 2005) eine experimentelle Software entwickelt (MovEngine), die die Visualisierung von strukturierten Bewegungsinstruktionen ermöglicht und einen zunächst ‚neutralen Körper' bereitstellt (Drewes 2014). Unter Anwendung bewegungsanalytischer Prinzipien von Notationssystemen (Eshkol-Wachman Movement Notation und Kinetographie Laban) löst diese den kontinuierlichen Bewegungsfluss in diskrete Raum, Zeit und Körper beschreibende Elemente auf, und passt sie in eine digitale Kodierung ein. Insbesondere von der Eshkol-Wachman Movement Notation übernimmt sie ein System von simultanen bzw. aufeinander folgenden Bewegungsinstruktionen, welche für jedes beteiligte Gelenk separate kreisförmige Bewegungsverläufe geometrisch definieren (Eshkol / Wachman 1958). Die einzelnen Bewegungsinstruktion lassen sich wiederum in eine bestimmte Anzahl zeitlicher und räumlich-geometrischer Parameter aufsplitten. Die Stärke dieses Ansatzes in Bezug auf die Leerstellen der historischen Überlieferung besteht darin, dass es möglich ist, Elemente in dieser hierarchischen Struktur nicht vollständig zu definieren bzw. für bestimmte Parameter eine gewisse Bandbreite zuzulassen und somit Interpretationsspielräume zu konkretisieren.

Der Umgang mit dem interdisziplinären Ballett wirft neben den zuvor beschriebenen Unsicherheiten durch die Kombination von Tanz und Musik zusätzliche Fragen auf. Zugleich transportiert Musik jedoch auch Informationen für den Tanz wie standardisierte Tanzsätze, tonmalerische Elemente, semantisierte musikalische Modelle, aber auch Stimmungen, die

affirmativ oder kontrastierend binäre Entscheidungsmöglichkeiten für die Interpretation anbieten.

Dem Projekt „Vienese Ballet: Encoding Music–Image–Dance“ liegt der Umgang mit den zuvor beschriebenen Leerstellen und Spielräumen zugrunde: Als Anwendungsbeispiel dienen hier Ballette, die in der mittleren Hälfte des 18. Jahrhunderts im Zuge der ‚Französisierung‘ der habsburgischen Kultur als diplomatische Maßnahme an den Wiener Hoftheatern aufgeführt wurden. Aufgrund der politischen Implikation war die Dokumentation der Aufführungen eine notwendige Maßnahme, der eine für das Ballette im 18. Jahrhundert ungewöhnlich dichte Quellenüberlieferung zu verdanken ist.

Zeitungsberichte enthalten szenische Beschreibungen ebenso wie die Theaterchroniken, die der Tänzer und Choreograph Philipp Gumpenhuber von 1758 bis 1763 aufzeichnete. Eine Sammlung von ca. 180 Ballettmusiken erhielt sich im Schwarzenberg Archiv Česky Krumlov; zusätzlich legte der Theaterdirektor Giacomo Durazzo eine Sammlung von Bildern an, die Szenen oder ganze Ballette ‚dokumentieren‘. Die Bildquellen, bei denen es sich um künstlerische Umsetzungen handelt, bieten zwar Informationen über die Ästhetik der Aufführungen, der Tänzer*innenkörper sowie der Bewegungsästhetik, sind jedoch bereits Interpretationen, die ebenso in Bezug auf den Umgang mit dem Raum zum Spielraum werden. In Bezug auf die Ästhetik der Aufführungen, der Tänzer*innenkörper sowie die Bewegungsästhetik bieten sie jedoch Informationen, die aus der Analyse mit Hilfe der MovEngine Software deutlich werden können: Die Animierung einzelner Figuren bzw. Figurengruppen hilft bei der Analyse der kinetischen Struktur der dargestellten Gesten.

Mit dem Projekt wird eine digitale Aufarbeitung und Kombination der überlieferten Materialien aus der Mitte des 18. Jahrhunderts angestrebt, in der die unterschiedlichen Quellen zum Bühnentanz – szenische Beschreibung, Musik, Bild in Verbindung gebracht und in ihrem Verhältnis zueinander analysiert werden können. Zusätzlich soll eine Tanzbibliothek entstehen, in der die in Traktaten von Raoul-Auger Feuillet in großer Detailgenauigkeit in Tanznotation mit Musik überlieferten historischen Tanzformen für die Herleitung zu Analogien zu den Balletten bereitgestellt werden. Die Forscherin Gisela Reber übertrug die Tänze aus der Feuillet-Notation in Kinetographie Laban (Reber 1986 und weiteres Material im Tanzarchiv der Folkwang Universität, Essen). Mit Hilfe von MovEngine kann aus den direkt überlieferten Quellen zum Repertoire und den in Tanznotation festgehaltenen stilistischen Informationen eine Materialsammlung für Gestik und Bewegungsästhetik generiert und eine ‚Tanzbibliothek‘ für historische Tänze angelegt werden.

Ausgehend von den bisherigen Arbeiten an MovEngine soll diese Software im Rahmen des Projekts weiter entwickelt werden, sowie auf dieser Basis ein Codierungsstandard und eine Software für die Verarbeitung von Tanz- und Bewegungsdaten entstehen, die sich für die Kombination mit dem Musikcodierungsstandard MEI eignen. Durch die Edirom-Technologie, die bereits für Musikeditionen die Darstellung unterschiedlicher Quellenarten ermöglicht, können die unterschiedlichen digitalen Übertragungen der Materialien zusammengebracht werden. Ermöglicht werden soll dadurch ein flexibles virtuelles, re-enactement, das den digitalen Umgang mit der ephemeren Kunstform Tanz erlaubt, das die Ebenen Bewegung, Musik und Bühne gleichermaßen berücksichtigt, um sich dadurch den „Spielräumen“ anzunähern.

Bibliographie

Brown, Bruce (1991): *Gluck and the French Theatre in Vienna.* Oxford: Oxford University Press.

Calvert, Tom / Wilke, Lars / Ryman, Rhonda / Fox, Ilene (2005): „Application of Computers to Dance“, in: *IEEE Computer Graphics and Applications* 25, no. 2: 6-12.

Drewes, Henner (2014): „MovEngine – Movement Values Visualized“, in: Jeschke, Claudia / Haitzinger, Nicole (eds.): *Tanz & Archiv Forschungsreisen. Mobile Notate.* München: epodium 22-33.

Drewes, Henner (2003): *Transformationen – Bewegung in Notation und digitaler Verarbeitung.* Essen: Die Blaue Eule.

Eshkol, Noa / Wachman, Abraham (1958): *Movement Notation.* London: Weidenfeld and Nicolson.

Reber, Gisela (1986): *Die Schrittformen und Armführungen nach Le Maître a Danser 1725 von Pierre Rameau übertragen in Kinetographie Laban*. Essen: unveröffentlicht.

Schiphorst, Thecla (1997): „Merce Cunningham: Making Dances with the Computer“, in: Vaughan, David (ed.): *Merce Cunningham: Creative Elements in choreography and dance.* Amsterdam: Harwood Academic Publishers (Choreography and Dance) 79-97.

Zwischen geisteswissenschaftlicher Offenheit und informatischer Explikation: Motivsuche als Herausforderung bei der Arbeit mit digitalen Ressourcen

Rastinger, Nina Claudia
ninaclaudia.rastinger@oeaw.ac.at
Austrian Academy of Sciences, Österreich

Resch, Claudia
claudia.resch@oeaw.ac.at
Austrian Academy of Sciences, Österreich

Theoretischer Hintergrund

Literarische Motive – in ihrer Gestalt als *„a theme, character, or verbal pattern which recurs in literature“* (Beckson & Ganz: 1960) – stellen seit Langem einen Untersuchungsgegenstand der Literaturwissenschaft dar. Als *„anthropologische Grundsituationen, die zwar historisch variiert werden, aber*

in ihrem Kern konstant bleiben" (Nünning 2013: 542), ziehen sie sich durch die Literaturgeschichte und damit durch den geisteswissenschaftlichen Forschungsbereich. Hiervon angeregte Erkenntnisinteressen umfassen sowohl Fragen danach, wie bestimmte Motive in ausgewählten Werken auftauchen (u. a. Ester et al. 2017, Nölle 2017), als auch danach, wie sich dieses Auftauchen diachron verhält: Wie beständig bzw. flüchtig ist das jeweilige literarische Motiv (Freedman 1971: 126) und inwiefern koppelt es sich an bestimmte Perioden, Textgattungen oder Autor*innen (von Wilpert 2013: 534)? Daran, welcher Aspekt eines Motivs dabei im Wandel der Zeit bestehen bleibt, kann man Wilpert (2013: 533–534) zufolge zudem zwischen verschiedenen Motivarten differenzieren: Bei konstanten Situationen, wie jener des *heimkehrenden Sohnes*, handelt es sich um Situationsmotive, während konstant bleibende Charaktere, wie der *Menschenfeind*, Typus-Motive konstituieren.

Forschungsstand

Obwohl die Motivforschung, wie oben beschrieben, ein vielseitiges Forschungsfeld aufspannt, wurde dieses im Rahmen der Digital Humanities bisher kaum beachtet: Im deutschsprachigen Raum existieren derzeit keine digitalen Korpora, in welchen eine Motivannotation vorgenommen wurde, und auch im nicht-deutschsprachigen Raum scheinen sich derartige Bestrebungen primär auf die Textgattung der Volksmärchen zu beschränken (u. a. Karsdorp et al. 2012, Garcia-Fernandez et al. 2014). Nur in wenigen Ausnahmefällen, wie in der Mittelhochdeutschen Begriffsdatenbank (Springeth 2005), liegt eine semantische Annotation vor, die Themen oder andere motivähnliche Aspekte mit Mark-Up versieht. Begründung findet dieser Umstand der fehlenden Motivannotation vor allem in der vagen Natur des zu Annotierenden. So erfahren Motive nach Best (2008: 349) schließlich erst *„im sprachl. Kunstwerk ihre individuelle Ausformung [und sind] somit erst durch Abstraktion faßbar"* und nicht an spezifische sprachliche Ausdrücke gebunden. Die Annotation von literarischen Motiven müsste insofern (noch) manuell vorgenommen werden, wodurch ein hoher Zeitaufwand entstünde sowie eine starke Subjektivität gegeben wäre.

Forschungsvorhaben

Dennoch macht die Überführung der Motivforschung in den Bereich der Digital Humanities Sinn: Korpusbasierte Untersuchungen ermöglichen die systematische Erforschung sowie quantitative Auswertung umfangreicher Textmengen und gewähren somit – vor allem im Hinblick auf die diachron angelegte Motivgeschichte und die Häufigkeiten spezifischer Motive (Freedman 1971: 126) – neue Erkenntnisse. Da es sich bei motivannotierten Korpora zum jetzigen Zeitpunkt jedoch noch um ein Desiderat handelt, muss hierfür ein anderer Zugang gewählt werden und zwar jener der Motivsuche – über folgende Frage: Mithilfe welcher Suchstrategien können Motive in nicht motivisch annotierten Korpora ausfindig gemacht werden?

Für ebendiese Herausforderung versucht der hier geplante Beitrag mögliche Lösungsansätze aufzuzeigen, wobei der Fokus auf ausgewählten Typus-Motiven, wie dem *weisen Salomon* oder dem *armen Sünder*, liegt. Anhand dieser sollen exemplarisch diverse Suchstrategien entwickelt werden, mithilfe welcher literarische Motive in digitalen Korpora, wie dem Austrian Baroque Corpus (ABaC:us) oder dem Deutschen Textarchiv (DTA), identifiziert werden können. Die Potentiale der verwendeten digitalen Ressourcen für die Motivsuche werden dabei genauso diskutiert wie ihre auftauchenden Limitationen. Immer mitzubedenken gilt es etwa die bereits angesprochene wechselnde Gestalt von literarischen Motiven: Da es sich bei ihnen um abstrakte Konzepte handelt, welche auf der Textoberfläche eine Vielfalt an sprachlichen Formen annehmen können, sind sie nur schwer über wenige ausgewählte „key words" aufspürbar. Hierzu trägt auch ihre Komplexität bei: *„Motive zeigen Personen und Sachen nicht isoliert, sondern in einem Zusammenhang"* (Frenzel 1978: 29) und bestehen damit immer aus mehreren inhaltlichen Komponenten – wie im Falle von Typus-Motiven, bei welchen Charakteren gewisse Eigenschaften zugeschrieben werden.

Diesen Problemen versucht das vorliegende Forschungsvorhaben mit vielfältigen Mitteln beizukommen. So wird beispielsweise aus der in vielen Korpora bereits vorhandenen linguistischen Basisannotation, bestehend aus einer Lemmatisierung und einer Wortartenzuordnung (POS-Tagging), Nutzen gezogen: Eigenschaften, wie Weisheit, können etwa als attributive Adjektive operationalisiert und über Abstandsoperatoren nahe eines interessierenden Charakters lokalisiert werden. Zudem können über Open-Source-Anwendungen wie AntConc oder Voyant Tools Kookkurrenzanalysen durchgeführt werden, um Charaktere und Eigenschaften zueinander in Bezug zu setzen und typische Zuschreibungen und Formulierungen sichtbar zu machen. Passend hierzu wird ebenfalls das von Huijnen und Lonij (2016) für die thematische Suche entwickelte Programm „Keyword Generator" auf seinen Ertrag für die Motivsuche hin befragt werden: Kann die Generierung motivspezifischer Suchwörter anhand bereits erkannter Textpassagen bei der Identifikation weiterer Belegstellen behilflich sein? Diese Frage soll genauso zu beantworten versucht werden wie jene nach der Wahl der adäquaten Suchtermini, für deren Ermittlung sowohl Synonym- als auch Motivdatenbanken zum Einsatz kommen werden. Damit soll der geplante Beitrag letztendlich anhand konkreter Beispiele aus der literaturwissenschaftlichen Praxis verschiedenste Suchstrategien sowie deren Vor- und Nachteile aufzeigen, um die Motivsuche für zukünftige Nutzer*innen zu erleichtern und dadurch zur vermehrten digitalen Motivforschung anzuregen.

Bibliographie

Anthony, Laurence (2019): *AntConc (Version 3.5.8)*. Tokyo: Waseda University https://www.laurenceanthony.net/software [letzter Zugriff 25. September 2019].

Berlin-Brandenburgische Akademie der Wissenschaften (eds.) (2019): *Deutsches Textarchiv.* Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. http://www.deutschestextarchiv.de/ [letzter Zugriff 25. September 2019].

Beckson, Karl / Ganz, Arthur (1960): *A Reader's Guide to Literary Terms*. New York: The Noonday Press.

Best, Otto F. (2008): *Handbuch literarischer Fachbegriffe.* Definitionen und Beispiele. Frankfurt am Main: Fischer.

Ester, Hans / Mariacher, Barbara / Tax, Evelyne (eds.) (2017): *Abschied als literarisches Motiv in der deutschsprachigen Literatur.* Festschrift zu Ehren des 75. Geburtstages von Jattie Enklaar. Würzburg: Königshausen & Neumann.

Freedman, William (1971): „The Literary Motif: A Definition and Evaluation", in: *NOVEL: A Forum on Fiction* 4 (2): 123–131.

Frenzel, Elisabeth (1978): *Stoff-, Motiv- und Symbolforschung*. 4., durchges. u. erg. Aufl. Stuttgart: Metzler.

Garcia-Fernandez, Anne / Ligozat, Anne-Laure / Vilnat, Anne (2014): „Construction and Annotation of a French Folkstale Corpus", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* 2430–2435.

Huijnen, Pim / Lonij, Juliette (2016). „From Keyword Search To Discourse Mining – The Meaning Of Scientific Management In Dutch Vocabulary, 1900-1940", in: *Digital Humanities 2016: Conference Abstracts* 569–570.

Karsdorp, Folgert / van Kranenburg, Peter / Meder, Theo / Trieschnigg, Dolf / van den Bosch, Antal (2012): *In search of an appropriate abstraction level for motif annotations*. http://dolf.trieschnigg.nl/papers/CMN.2012.karsdorp.pdf [letzter Zugriff 25. September 2019].

Nölle, Volker (2017): *Der heimliche Blick: Motiv und Modell – eine Matrix innovativer Perspektiven*. Würzburg: Königshausen & Neumann.

Nünning, Ansgar (ed.) (2013): *Metzler Lexikon Literatur- und Kulturtheorie. Ansätze – Personen – Grundbegriffe*. Fünfte, aktualisierte und erweiterte Auflage. Stuttgart: J. B. Metzler.

Resch, Claudia / Czeitschner, Ulrike (eds.) (2015): *ABaC:us – Austrian Baroque Corpus*. http://acdh.oeaw.ac.at/abacus/ [letzter Zugriff 25. September 2019].

Sinclair, Stéfan / Rockwell, Geoffrey (2019): *Voyant Tools (Version 2.4)*. https://voyant-tools.org/ [letzter Zugriff 25. September 2019].

Springeth, Margarete (2005): „Auf der Suche nach Begriffen und Motiven. Die Mittelhochdeutsche Begriffsdatenbank (MHDBDB) an der Universität Salzburg", in: Schubert, Martin J. (ed.): *Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion* (= Beihefte zu Editio 23). Tübingen: Niemeyer 317–323.

Von Wilpert, Gero (2013): *Sachwörterbuch der Literatur.* Sonderausgabe der 8. verbesserten und erweiterten Auflage 2001. Stuttgart: Alfred Kröner.

Index der Autorinnen und Autoren

Adelmann, Benedikt 331
Albers, Laura 342
Albrecht, Jens 284
Andorfer, Peter 271
Andresen, Melanie 219
Annisius, Marie 315
Ardesi, Denise 286
Babl, Florian 39
Baierer, Konstantin 244
Baillot, Anne 21, 31
Balbach, Nico 235
Bauer, Marlene 296
Baumann, Timo 350
Begerow, Anke 219
Bell, Peter 98
Bernhard, Bermeitinger 158
Bernhart, Toni 77
Böhm, Astrid 286
Bludau, Mark-Jan 114, 360
Blumtritt, Jonathan 18, 318
Bönisch, Thomas 359
Boenig, Matthias 244
Bohl, Benjamin W. 29
Bonsergent, Lou-Ann 31
Brandl, Stephanie 21
Braud, Camille 31
Breitenbücher, Uwe 36
Bürgermeister, Martina 277
Brüggemann, Viktoria 114
Brokering, Annalena 323
Brunner, Annelen 190
Buechel, Sven 352
Burghardt, Manuel 215, 232
Burr, Elisabeth 315
Busch, Anna 279
Busch, Hannah 313
Cakir, Osman 356
Capelle, Irmlind 227
Christlein, Vincent 230
Clados, Christiane 123
Cremer, Fabian 27, 271, 325
Dahnke, Michael 43
De Luca, Ernesto William 321
Demmer, Dennis 184
Dinger, Patrick 141
Dogunke, Swantje 27, 174
Donig, Simon 158
Drach, Sviatoslav 367
Dörk, Marian 114, 360
Du, Keli 104
Duan, Tinghui 352
Dziudzia, Corinna 116
Eckl, Markus 300
Effinger, Maria 101
Eggert, Lisa 200
Eide, Øyvind 46
Eisterhues, Marcel 39
Elwert, Frederik 27, 362
Engl, Elisabeth 244
Eschweiler, Mark 327
Evers, Anna-Maria 327
Faßhauer, Vera 292
Fallucchi, Francesca 321
Fangerau, Heiner 360
Feldmann, Alina 208
Fischer, Barbara 344
Fischer, Frank 167, 278
Flanders, Julia 13
Flüh, Marie 15, 162, 167, 354
Flinz, Carolina 347
Franken, Lina 74, 219
Freyberg, Linda 148
Fricke-Steyer, Henrike 306
Fußbahn, Ulrike 315
Fuchs, Alexandra 291
Gaidys, Uta 219
Gassner, Sebastian 300
Geiger, Bernhard 291
Geiger, Jonathan 48
Geißler, Nils 301
Gengnagel, Tessa 313
Georgi, Christopher 337
Gervais, Ludovic 31
Gius, Evelyn 90, 331
Güntner, Lydia 280
Gradl, Tobias 274
Große, Peggy 342
Görz, Günther 272
Guhr, Svenja 267, 323
Guy, Louisa 21
Haaf, Susanne 337
Habler, Florian 296
Hadersbeck, Maximilian 39, 202
Hahn, Udo 352
Hall, Mark 116
Halling, Thorsten 360
Handschuh, Siegfried 158
Hannesschläger, Vanessa 328
Hartelt, Alexander 171
Hartmann, Volker 244
Harzenetter, Lukas 36
Heßbrüggen-Walter, Stefan 182
Hechtl, Angelika 279
Heinig, Julia 362
Heinisch, Barbara 211
Helling, Patrick 18, 318
Henner, Drewes 371
Henrich, Andreas 274
Hermes, Jürgen 184
Herrmann, J. Berenike 144
Hertling, Anke 321
Hess, Jan 358
Heuberger, Hannes 296
Hiltmann, Torsten 135
Hinkelmanns, Peter 131
Hinrichsen, Lena 230
Hobisch, Elisabeth 291
Hodel, Tobias 84
Holzer, Matthias 277
Homburg, Timo 334
Horstmann, Jan 15, 154, 167, 354
Howanitz, Gernot 24

Hussein, Hussein ... 350
Iffland, Joachim ... 288
Ihden, Sarah ... 240
Jacke, Janina ... 154
Jannidis, Fotis ... 94, 190
Jegan, Robin ... 274
Jentsch, Patrick ... 345
Jung, Kerstin ... 33, 358
Kaminski, Andreas ... 359
Kamzelak, Roland ... 358
Kepper, Johannes ... 260
Ketschik, Nora ... 52
Klaes, Jan Sebastian ... 321
Klaffki, Lisa ... 306
Klinke, Harald ... 184
Klug, Helmut W. ... 205, 286
Klusik-Eckert, Jacqueline ... 276
Koch, Gertraud ... 219
Kohl, Linus ... 356
Koller, Barbara ... 208
Koncar, Philipp ... 291
Konle, Leonard ... 94
Kraft, Anneli ... 289
Kramski, Heinz Werner ... 358
Krautter, Benjamin ... 52, 127, 177
Kröber, Cindy ... 87
Kremer, Gerhard ... 33
Krishnan, Aravind ... 364
Krug, Markus ... 151
Kruhl, Dominik ... 327
Kuhn, Jonas ... 223, 358
Kunze, Kristina ... 307
Landes, Florian ... 39, 43
Landes, Lisa ... 141
Landkammer, Miriam ... 131
Lassner, David ... 21
Laurioux, Bruno ... 286
Lehmann, Robert ... 284
Leinen, Peter ... 94
Leitner, Elena ... 360
Lemaire, Marina ... 18
Leymann, Frank ... 36
Liebl, Bernhard ... 232
Liedtke, Clemens ... 107
Limbach, Saskia ... 229
Lindinger, Matthias ... 39
Liu, Alan ... 13
Lordick, Harald ... 27
Lüschow, Andreas ... 80
Luhmann, Jan ... 215
Maciej, Eder ... 340
Maier, Andreas ... 230
Maria, Christoforaki ... 158
Mathiak, Brigitte ... 36, 318
Meister, Jan Christoph ... 15, 167
Menzel, Sina ... 360
Mertgens, Andreas ... 187
Messemer, Heike ... 87, 123
Messerli, Thomas ... 144
Meyer-Sickendiek, Burkhard ... 350
Müller, Gerhard ... 360
Müller, Melissa ... 200
Münster, Sander ... 87
Molitor, Paul ... 330
More, Jacqueline ... 291
Moreno Schneider, Julian ... 361
Murr, Sandra ... 52
Nasarek, Robert ... 43
Necker, Gerold ... 330
Neudecker, Clemens ... 244, 360
Neuefeind, Claes ... 36, 318
Neumann, Katrin ... 325
Nicka, Isabella ... 131
Niebes, Kai Michael ... 46
Niewöhner, Laura Maria ... 31
Noichl, Maximilian ... 341
Nussmüller, Antonia ... 277
Offert, Fabian ... 98
Ortloff, Anna-Marie ... 280
Ott, Katrin ... 27
Pagel, Janis ... 52, 177, 194
Pannach, Franziska ... 364
Pöckelmann, Marcus ... 330
Petras, Vivien ... 360
Petris, Marco ... 15, 354
Pfeiffer, Jasmin ... 48
Pichler, Alois ... 202
Pichler, Axel ... 223
Pielström, Steffen ... 340
Pilsl, Florian ... 296
Plaksin, Anna ... 119
Poirier, Corentin ... 286
Porada, Stephan ... 345
Probst, Nora ... 187
Puppe, Frank ... 151, 171, 235
Radisch, Erik ... 24
Rastinger, Nina ... 304
Rastinger, Nina Claudia ... 372
Rau, Felix ... 318
Raunig, Elisabeth ... 286
Rebiger, Bill ... 330
Rehm, Georg ... 360
Reißer, Alexandra ... 356
Reiners, Stefan ... 266
Reinert, Matthias ... 111
Reinthal, Angela ... 303
Reiter, Nils ... 52, 177, 194, 223
Resch, Claudia ... 304, 372
Rettinghaus, Klaus ... 138
Reuhl, Elisabeth ... 327
Reul, Christian ... 43, 235
Röhrer, Ines ... 39
Richts, Kristina ... 227
Risse, Benjamin ... 135
Ritter, Jörg ... 330
Roeder, Torsten ... 27, 138
Roller, Ramona ... 250
Rosenkötter, Martha ... 344
Ruppenhofer, Josef ... 347
Sahle, Patrick ... 255
Saric, Sanja ... 291, 369
Schöch, Christof ... v
Schembera, Björn ... 358
Scheuermann, Leif ... 277
Schildkamp, Philip ... 36, 318
Schilke, Elena ... 227
Schlesinger, Claus-Michael ... 358
Schlögl, Matthias ... 348

Schlicht, Helene ... 345
Schmeer, Hendrik ... 314
Schmidt, David ... 151
Schmidt, Thomas ... 280, 296, 347
Schmitz, Claudia ... 321
Schmunk, Stefan ... 18
Schneider, Gerlinde ... 316
Schneider, Philipp ... 310
Schneider, Stefanie ... 356
Scholger, Martina ... 291, 369
Scholger, Walter ... 328
Scholz, Martin ... 276
Schrott, Maximilian ... 111
Schubert, Zoe ... 46
Schulte, Judith ... 309
Schulz, Daniela ... 313
Schulz, Julian ... 356
Schumacher, Mareike ... 15, 162, 167
Schuster, Britt-Marie ... 337
Schwandt, Silke ... 31
Schweitzer, Frank ... 250
Schwembacher, Manuel ... 131
Seidl, Chiara ... 272
Seltmann, Melanie E.-H. ... 265
Seuret, Mathias ... 229
Sobriel, Nicole ... 101
Sonnberger, Jakob ... 277
Söring, Sibylle ... 27
Stadler, Peter ... 29
Stefan, Schulte ... 18
Steiner, Christian ... 205, 286
Steyer, Timo ... 271
Still, Sebastian ... 39, 202
Strothotte, Adrian ... 31
Thiele, Sebastian ... 135
Thielert, Frauke ... 337
Thiering, Martin ... 272
Thiery, Florian ... 334
Thomas, Clement ... 31
Tiepmar, Jochen ... 215
Towara, Nadine ... 321
Trautmann, Marjam ... 303
Trilcke, Peer ... 167, 279
Türkoğlu, Enes ... 46, 187, 327
Tscheu, Amelie ... 303
Tu, Ngoc Duyen Tanja ... 190
Tumanov, Rostislav ... 207
Ullrich, Sabine ... 39, 202
Ulrich, Mona ... 358
Vasold, Gunter ... 316, 348
Veit, Joachim ... 258
Vera, Grund ... 371
Viehhauser, Gabriel ... 208, 223, 358
Viglianti, Raffaele ... 29
Vogel, Andreas ... 279
Vogeler, Georg ... 348
Wagner, Sarah ... 238
Wübbena, Thorsten ... 27, 325
Wehner, Maximilian ... 43
Weichselbaumer, Nikolaus ... 229
Weimer, Lukas ... 190, 365
Wick, Christoph ... 171
Wiedmer, Nathalie ... 194
Wieners, Jan Gerrit ... 46, 318
Willand, Marcus ... 177
Windl, Maximiliane ... 280
Wissik, Tanja ... 314
Wolff, Christian ... 296
Wuttke, Ulrike ... 18, 247
Zellhoefer, David ... 361
Zeppezauer-Wachauer, Katharina ... 131
Zinsmeister, Heike ... 219

DHd2020

SPIELRÄUME

DIGITAL HUMANITIES ZWISCHEN MODELLIERUNG UND INTERPRETATION