K

KRITIK DER
DIGITALEN VERNUNFT

DHd 2018 Köln

26.02. – 02.03.2018

Konferenzabstracts

DHd digital humanities im deutschsprachigen raum

DCH Data Center for the Humanities Kölner Datenzentrum für die Geisteswissenschaften

Universität zu Köln

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

DHd 2018

Kritik der digitalen Vernunft

Konferenzabstracts

Universität zu Köln
26. Februar bis 2. März 2018

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: Georg Vogeler
Korrektur der Auszeichnungen:
Dana Persch, Claes Neuefeind, Patrick Helling
Konvertierung TEI nach PDF: Claes Neuefeind
https://github.com/GVogeler/DHd2018
unter Verwendung der Konversionsskripte von Karin Dalziel
https://github.com/karindalziel/TEI-to-PDF
und der bearbeiteten Version von Aramís Concepción Durán
https://github.com/aramiscd/dhd2016-boa.git
Konferenz-Logo: Anja Neuefeind

Online verfügbar: http://dhd2018.uni-koeln.de/

ISBN 978-3-946275-02-2

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

Vorwort

Das „Book of Abstracts“ war in der jungen Tradition der DHd zunächst eher ein Nebenprodukt des Bewerbungsprozesses. Die Professionalisierung unserer Disziplin hat nun dazu geführt, dass für die DHd 2018 in Köln die Ansprüche an die Qualität der Beiträge gestiegen sind. Damit geht einher, dass die Vorschläge immer mehr ihre guten Ideen mit den üblichen Argumentationsmustern und äußeren Formen wissenschaftlichen Arbeitens entwickeln, also die Standards einer wissenschaftlichen Publikation erfüllen. Das „Book of Abstracts“ wird so klassischen „Conference Proceedings“ immer ähnlicher. Das ist eine zu begrüßende Entwicklung. Sie trägt hoffentlich auch dazu bei, die auf der DHd 2018 vorgestellten Forschungsergebnisse im weiteren wissenschaftlichen Diskurs aufzugreifen und zu referenzieren.

Für die Qualität der Beiträge sind natürlich die Autoren verantwortlich. Ihrem Bemühen, auch dem der vielen, die wir leider ablehnen mussten, gebührt das erste Dankeschön. Dazu beigetragen haben dann die vielen Gutachterinnen und Gutachter, die es ertragen mussten, von mir immer wieder daran erinnert zu werden, doch bitte ihre Expertise in den Dienst der Tagung zu stellen. Auch Ihnen/euch allen: Danke! Es blieben dann trotz dieser Expertise immer noch einige Nüsse zu knacken, die wir im Programmkomitee so sachlich, konstruktiv und wohlwollend diskutieren konnten, wie ich mir es für eine Tagung nur wünschen kann: Danke an Mareike König als meine Stellvertreterin, Andreas Münzmay, Claudine Moulin, Christof Schöch, Anne Baillot, Peter Gietz, Walter Scholger, Lisa Dieckmann, Andreas Henrich, Petra Gehring, Lars Wieneke und als Vertreter der lokalen Organisation Andreas Witt. Ebenso ein Danke an die übrigen lokalen Organisatorinnen und Organisatoren für die produktive Zusammenarbeit bei der Vorbereitung des Tagungsprogramms.

Die Konferenz – und damit auch diese Publikation – wäre natürlich nicht möglich, wenn es nicht Einrichtungen gäbe, die sie finanziell unterstützen. Sie sind auf der Rückseite des Titelblatts aufgelistet. Ich hoffe, das Ergebnis macht sie zufrieden. Wir bedanken uns auf jeden Fall für die Unterstützung!

Der Status einer wissenschaftlich nutzbaren Publikation hat auch etwas mit einer gründlichen Redaktion der Texte zu tun. Diese haben die Kölner Patrick Helling, Claes Neuefeind und Dana Persch geleistet. Sie haben die Vielfalt der TEI-Kodierungen, die der DHconvalidator aus der Vielfalt der eingereichten Dateiformatierungen erzeugte, angeglichen und daraus ein PDF erzeugt. Für die technische Abwicklung konnten Sie auf die Vorarbeiten der Leipziger Kolleginnen und Kollegen aufbauen. Ihrer aller Arbeit ist in diesem Band und in einem github-Repository (https://github.com/GVogeler/DHd2018) dokumentiert. Ihnen gebührt deshalb ganz ausdrücklicher Dank!

Graz, Februar 2018
Georg Vogeler
Vorsitzender des Programmkomitees

Inhaltsverzeichnis

Keynotes

Workshops

Panels

Vorträge

Poster

Anhang

Keynotes

Der ‚Stachel des Digitalen' - ein Anreiz zur Selbstreflexion in den Geisteswissenschaften?

Krämer, Sybille
sybkram@zedat.fu-berlin.de
Freie Universität Berlin, Deutschland, Institut für Philosophie

Geht es um eine Kritik an der digitalen Vernunft? Oder kann die ‚digitale Vernunft' ihrerseits eine kritische Perspektive eröffnen, insofern ‚der Stachel' ihrer Praktiken das Selbstverständnis von Geisteswissenschaften herausfordert? Die leitende Idee ist, dass ein Nachdenken über die ‚strukturentdeckenden ', über ‚datengetriebene' algorithmische Forschungsverfahren der Digital Humanities die Geistes- und Kulturwissenschaften anregen kann (oder anregen sollte) zu einer Metareflexion, durch welche auch die Verfahrensweisen ‚herkömmlicher' geistes- und kulturwissenschaftlicher Forschungsarbeit neu beleuchtet werden. Der dabei eingenommene methodische Gesichtspunkt ist ein praxeologischer: Was eine Wissenschaft ist, zeigt sich im Insgesamt ihres Forschungs-, Lehr- und Vermittlungshandelns.

Alle Geistes- und Kulturwissenschaften zielen darauf, etwas das Texten, Bildern, Artefakten implizit ist, explizit zu machen – ob nun durch traditionelle Interpretation oder algorithmische Datenanalyse. Doch bereits diese Unterscheidung von ‚Interpretation' und ‚Datenanalyse' hinkt, denn es gibt weder rohe Daten noch material- und texturunabhängige Interpretationen. Doch wenn das so ist: Warum sollte eine maschinelle – im Idealfall statistisch-empirische - Auswertung großer Datenbestände die herkömmlichen geisteswissenschaftlichen Verfahrensweisen in neuem Licht erscheinen lassen?

Kritik der digitalen Vernunft

Sperberg-McQueen, C. Michael
cmsmcq@blackmesatech.com
Black Mesa Technologies LLC, New Mexico

Die Organisatoren der Tagung stellen den Teilnehmern die Frage: "Gibt es im Umgang mit digitalen Medien, in der Modellierung, Operationalisierung und Formalisierung der Arbeit mit Computern implizite, stillschweigend akzeptierte Agenden, die einer Reflexion durch einen „Intellectual Criticism" bedürfen?" Wie sähe ein solcher Intellectual Criticism aus? Worauf könnte er basieren?

In den drei großen Kritiken der reinen Vernunft, der praktischen Vernunft, und der Urteilskraft hat Kant eine 'kopernikanische Revolution' in der Philosophie mit dem Postulat eingeleitet, unsere Erkenntnis richte sich nicht nach den Dingen, sondern die Dinge richten sich nach unserer Erkenntnis: d.h. nach den apriorischen Formen der Anschauung und nach den vorgegebenen Begriffen des Verstandes (die Kategorient).

Gibt es apriorische Formen, die den zu bearbeitenden Stoff der digitalen Vernunft bestimmen, ähnlich wie Zeit und Raum die menschliche Anschauung bestimmen? Gibt es vorgegebene Begriffe, die aller digitalen Vernunftarbeit zu Grunde liegen? Wie verhält es sich im digitalen Raum mit der Eigenverantwortung und der Autonomie, die nach Kant das Wesen der Aufklärung und der Freiheit ausmachen?

Workshops

Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

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

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Würzburg, Deutschland

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

Puppe, Frank
puppe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Goals of the workshop

The workshop presents ATHEN [1] (**A**nnotation and **T**ext **H**ighlighting **En**vironment), an extensible desktop-based annotation environment which supports more than just regular annotation. Besides being a general purpose annotation environment, ATHEN supports indexing and querying support of your data as well as the ability to automatically preprocess your data with Meta information. It is especially suited for those who want to extend existing general purpose annotation tools by implementing their own custom features, which cannot be fulfilled by other available annotation environments. On the according gitlab, we provide online tutorials, which demonstrate the use of specific features of ATHEN.

Related Work

We compare ATHEN to three web-based and four desktop applications in 12 categories by adapting most criteria defined by Neves and Leser (Neves / Leser 2012) to compare different annotation tools:

1. Availability and up-to-dateness of the documentation
2. Active development at the present time
3. Source code for download
4. Complexity of system requirements
5. Interoperability by supporting certain formats
6. Support of different annotation layers
7. Support of NLP-preprocessing to speed up manual annotation
8. Support of visualization
9. Support of self-learning systems to speed up manual annotation
10. Support of querying annotated data
11. Possibility to do an inter-annotator-agreement, this is important for projects, in which more than one annotator labels the same documents
12. Extensibility

We explicitly do not want to compare subjective features like usability or how the annotations are presented.

Annotator / Criterion	ATHEN	CATMA	Web-Anno	BRAT	UAM	MMAX2	Know-tator	WordFreak
Documentation	✓	✓	✓	✓	✓	✓	✓	Homepage
Active development	✓	✓	✓	✓	✓	×	×	×
Open Source	✓	✓	✓	✓	×	✓	✓	✓
System requirements	Java	Webbrowser	Java, Web-browser	Python	Win, Mac	Java	Java, Protégé	Java
Supported formats	.txt, .xmi, .xml (incl. page .xml, .tei)	.tei, .txt	.CoNLL, .tcf, .teicph, .xmi	.ann, .txt,	.txt, .xml	.txt, .xml	.txt, .xml	.ace, .ptb, .muc, .txt
Supported annotation layer	Customizable	Customizable	Customizable	Entities, Relations, Events	Customizable	Markables, Attributes and Relations in between	Ontology-based	Depending on plugin
Preprocessing	UIMA Analysis Engines	using heuréCLEA	×	Sentences, tokens	POS-Tagging, Syntactic parsing	Tokenization	×	Depending on plugin
Visualization	Social networks	✓	×	×	×	×	×	×
Automation	RUTA annotation support	Machine Learning	Machine Learning	×	Query Based	×	×	×
Query language for corpus analysis	Apache Lucene Support	Query	×	×	Search and statistics	Query	×	×
Inter-annotator agreement	Interactive side by side	×	Curation View	Side by side	×	×	✓	×
Extensibility	Runtime/ Code	Code	Code + Tutorial	Code	×	Code	Code	Code/ Plugin

Table 1: Comparison of three web-based and four desktop applications with ATHEN in twelve categories. No tool excels in every category.

All of the listed tools have an accessible documentation, either web-based or as a PDF, available for download. Besides UAM (O'Donnell 2008), every other application is listed as open source, so at least extensions based on code level can be made. WebAnno (Yimam et al. 2013) is the only

application having a tutorial supporting a new developer to make changes in their project. ATHEN stands out in the sense that extensions to its UI can be made at runtime, therefore easing the process of adding functionality to it. WebAnno supports the largest number of formats and comes with a machine learning based automatic annotation, however lacks integrated NLP-preprocessing. CATMA (Meister 2017) is the only project that has a very good visualization component and also supports TEI-XML (Wittern et al 2009), the unspoken standard of text processing. Being a standalone web application, CATMA itself does not support NLP-preprocessing. ATHEN comes with the support of the execution of UIMA analysis engines, accessible from web repositories or a local repository, giving the user a chance to integrate her custom-made annotators. Four tools, ATHEN, UAM, MMAX2 (Müller / Strube 2006) and CATMA feature an integrated query language which helps to analyze existing corpora. Most tools allow the annotation of user-defined annotation schemas earning therefore the title “generic annotation tool”. Alongside UAM, ATHEN supports the annotation based on queries, while UAM defines its own language, ATHEN supports the annotation using Apache UIMA Ruta (Kluegl et al. 2016) rules. Three of the listed tools, MMAX2, Knowtator (Ogren 2006) and WordFreak (Morton / LaCivita 2003) are currently no longer in active development.

Brief technological description of ATHEN

ATHEN is a Java-based desktop application with the vision to be extensible. Therefore, it makes use of the flexible plugin architecture of the eclipse **R**ich- **C**lient- **P**latform (RCP) [2]. Internally, it is built around Apache UIMA, which means incoming data is automatically converted into the UIMA specific Common Analysis Subject (CAS) architecture. Working with UIMA allows the integration of standalone analysis engines, which can be used to preprocess data and speed up manual annotation. The use of Apache Lucene enables ATHEN to create an index comprising documents, as well as their annotations, which results in queries that can answer questions based on text and meta information in real time. With the ability to execute Apache UIMA Ruta one can even create queries of far higher complexity. On top of that, ATHEN features OWL-Support, which allows the definition of an ontological annotation schema in a machine-readable format. Using Apache UIMA internally allows ATHEN to even address more complicated input. Currently ATHEN supports the annotation of image regions, based on user defined polygons.

Program of the workshop

The program is split into four sections:

1. Introduction to ATHEN and distinguishing from other existing annotation environments.
2. Working with ATHEN, which contains the definition of scenarios and annotating sample documents.
3. Utility of ATHEN (beyond regular annotation), which addresses the following topics:
   1. Defining an annotation schema using OWL
   2. Preprocessing texts based on Apache UIMA Analysis Engines
   3. Creating and executing queries based on Apache Lucene
   4. Annotating images with ATHEN
4. Extending ATHENs functionalities and adapting it to your needs (developer specific)

The first section is a presentation which shows the main differences between the existing annotation tools. The second section defines an ordinary annotation scenario and it is used to introduce the participants to the general-purpose annotation view of ATHEN. Afterwards, for tasks to which ATHEN has special support (annotating character references and their coreferences, annotating direct speech and their speaker) an introduction to the special purpose views of ATHEN is given.

The third panel introduces the participants to the functionality of ATHEN beyond regular text annotation. It starts with the definition of an OWL ontology (and its utilization for texts). This is centered on relation detection of character references, as well as an attribution of those references.

To speed up manual annotation it is helpful to have it preprocessed with existing tools. The task definition is then changed from pure annotation to an application with a consecutive correction of the output of the automatic engines. In this context, Nappi, a submodule of ATHEN is presented and it is shown how to define, execute and integrate custom analysis engines.

The next part is dedicated towards extracting knowledge from annotated data, for this purpose, an Apache Lucene Index is created using ATHEN and is queried in a live fashion. This feature allows rapid insight into an existing corpus and enables the user to answer their own hypothesis.

The tutorial continues with the presentation of how images can be annotated with polygon-based

annotations to show, that ATHEN is not only limited to textual resources.

The last part is directed towards Java developers who are interested in developing their own annotation component.

Each section starts with a set of slides which introduce the features in focus and presents the participants with one or more tasks that can be fulfilled by using ATHEN.

Requirements

The participants need their own laptops with an active internet connection. The number of participants is limited to 15 to 20. The last section requires knowledge of Java. Data which is necessary for the tutorials will be hosted on our own server and will be made accessible for download.

Research projects

ATHEN is mainly developed in the context of the project Kallimachos at the University of Wuerzburg. Its main purpose was to support the annotation process of DROC (**D**eutscher **RO**man **C**orpus). Currently automatic creation of literary interaction networks, automatic genre detection of novels and sentiment analysis in literary novels are in the focus of interest.

An extension to ATHEN was made in the project "Redewiedergabe" to manually annotate different forms of speech, thought and writing representation (STWR). These annotations will then be used to train an automatic recognizer for STWR.

Fußnoten

1. https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen
2. A web version of ATHEN with its major features is available at: https://webathen.informatik.uni-wuerzburg.de

Bibliographie

Kluegl, Peter / Toepfer, Martin / Beck, Philip-Daniel / Fette, Georg / Puppe, Frank (2016): "UIMA Ruta: Rapid Development of Rule-based Information Extraction Applications", in: *Natural Language Engineering 22.1* 1-41.

Meister, Jan Christoph / Gius, Evelyn / Jacke, Janina / Petris, Marco: *CATMA 5.0.* http://catma.de/ [Accessed September 22, 2017].

Morton, Thomas / LaCivita, Jeremy (2003): "WordFreak: an open tool for linguistic annotation", in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4* 17-18.

Müller, Christoph / Strube, Michael (2006): "Multi-level annotation of linguistic data with MMAX", in *: Corpus technology and language pedagogy: New resources, new tools, new methods 3* 197-214.

Neves, Mariana / Leser, Ulf (2012): "A survey on annotation tools for the biomedical literature", in: *Briefings in Bioinformatics 15.2* 327-340.

O'Donnell, Mick (2008): "Demonstration of the UAM CorpusTool for text and image annotation", in: *Proceedings of the 46th annual meeting of the Association for computational linguistics on human language technologies: Demo session* 13-16.

Ogren, Philip V. (2006): "Knowtator: a Protégé plug-in for annotated corpus construction", in: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations* 273-275.

Stenetorp, Pontus / Pyysalo, Sampo / Topić, Goran / Ohta, Tomoko / Ananiadou, Sophia / Tsujii, Jun'ichi (2012): "BRAT: a Web-based Tool for NLP-Assisted Text Annotation", in : *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* 102-107.

Wittern, Christian / Ciula, Arianna / Tuohy, Conal (2009): "The making of TEI P5", in: *Literary and Linguistic Computing 24.3* 281-296.

Yimam, Seid Muhie / Gurevych, Iryna / de Castilho, Richard Eckart / Biemann, Chris (2013): "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations", in: *Proceedings of ACL-2013, demo session, Sofia, Bulgaria* 1-6 .

Audio Mining für die Geistes- und Kulturwissenschaften: Nutzungsszenarien und Herausforderungen

Köhler, Joachim
joachim.koehler@iais.fraunhofer.de
Fraunhofer IAIS

Leh, Almut
almut.leh@fernuni-hagen.de
FernUniversität in Hagen

Himmelmann, Nikolaus
sprachwissenschaft@uni-koeln.de
Universität zu Köln, Deutschland

Rau, Felix
f.rau@uni-koeln.de
Universität zu Köln, Deutschland

Workshopleiterinnen und -leiter

- Nikolaus P. Himmelmann, Universität zu Köln,
- Joachim Köhler, Fraunhofer IAIS,
- Almuth Leh, FernUniversität in Hagen,
- Felix Rau, Universität zu Köln,

Ort

Der Workshop findet in Raum S16 des Seminargebäudes (Gebäude 106) der Universität zu Köln statt.

Programm

Zum Workshop gibt es Beiträge, die primär aus Entwicklerperspektive das Thema angehen, und solche, die eher eine Anwenderperspektive einnehmen, wobei das keine scharfe Trennung ist.

Aus Entwicklerperspektive berichten Alexandre Arkhipov vom Hamburger Zentrum für Sprachkorpora, Christoph Draxler vom Bayerisches Schallarchiv (LMU), Jens Gorisch vom IDS Mannheim, Joachim Köhler vom Fraunhofer IAIS sowie Burkhard Meyer-Sickendiek und Hussein Hussein von der FU Berlin über Audiomining Werkzeuge, die bei Ihnen entwickelt wurden und werden.

Aus der Anwenderperspektive stellen Thomas Beckers vom WDR, Thorsten Dresing von der Firma *audiotranskription* (Marburg), Almuth Leh von der FernUniversität Hagen sowie Anna-Maria Götz, Annabelle Petschow & Ruth Rosenberger von der Stiftung Haus der Geschichte der Bundesrepublik Deutschland (Bonn) Herausforderungen für die automatischen Spracherkennung in unterschiedlichen Anwendungsbereichen dar.

Ein detailliertes Programm ist Ende Januar über die Webseite des KA3 Projekts zugänglich. Im Anschluss an die Beiträge gibt es jeweils Gelegenheit zu Fragen und kurzer Diskussion. Am Schluss des Workshops gibt es eine kurze Abschlussdiskussion. Workshopsprache ist Deutsch, Beiträge können aber auch auf Englisch präsentiert werden.

Weitere Details zum Thema

Moderne intelligente Analyse- und Auswertungsmethoden für Audiodaten ermöglichen effektivere Arbeitsweisen und neue Fragestellungen in den Geistes- und Kulturwissenschaften. Dieser Workshop soll Fortschritte, Herausforderungen und Perspektiven im Audio Mining präsentieren und Forscherinnen und Forscher aus dem Bereich der Sprachtechnologie mit (potentiellen) Nutzerinnen und Nutzern aus den Geistes- und Kulturwissenschaften zusammenbringen.

Audiovisuelle (multimodale) Primärdaten spielen eine zunehmend größer werdende Rolle in den Geistes- und Kulturwissenschaften. Einfacher und kostengünstiger werdende Aufnahme- und Speicherungstechniken für Audio- und Videodaten ermöglichen die Erhebung von immer umfangreicheren Datensets. Die wissenschaftliche Analyse und Auswertung dieser Daten basiert aber weiterhin überwiegend auf Transkripten, also konvertiert ins schriftsprachliche Medium. Dabei gehen wesentliche Merkmale dieses Datentyps (Sprechmelodie, Stimmqualitäten, Gestik, Mimik etc.) verloren. Selbst wenn die Primärdaten als audiovisuelle Daten analysiert werden geschieht dies meist durch arbeitsintensive manuelle Segmentierung und Annotation.

Von zunehmender Bedeutung ist auch die Nachnutzung audiovisueller Daten, namentlich qualitativer Interviews, wie sie in fast allen Geistes- und Kulturwissenschaften eine Rolle spielen. AV-Daten stellen schon allein aufgrund ihres großen Umfangs besondere Herausforderungen dar für den flexiblen und systematischen Zugriff wie auch eine zweckmäßige langfristige Speicherung. Es bedarf einer Arbeitsumgebung, in der flexibel online auf den gesamten Datensatz zugegriffen werden kann und je nach Bedarf Unterkorpora definiert werden können. Für die einschlägigen Archive, Medienzentren und Dokumentationsstellen ermöglicht das Audio Mining eine nutzerspezifische Recherche in der Gesamtdatenmenge bei direktem Zugriff auf das Audiosignal und sekundengenauer Trefferanzeige. Dadurch wird die Ermittlung und Bereitstellung relevanter Daten für Sekundäranalysen erheblich verbessert.

Der Einsatz von intelligenten Analyse- und Auswertungsmethoden für multimodale Daten gewinnt vor diesem Hintergrund in den letzten Jahren stark an Bedeutung. Fortschritte im Bereich

des maschinellen Lernens ermöglichen stärker automatisierte Arten der Datenanalyse und führen zu einer weniger zeit- und arbeitsintensiven Aufbereitung. Diese Analysemethoden operieren darüber hinaus auf den audiovisuellen Primärdaten und nicht auf textuellen Derivaten.

Im Vergleich zu Textdaten sind die Möglichkeiten der Analyse und Auswertung von audiovisuelle Daten noch weniger verbreitet. Besonders bei der Sprachanalyse von Audio- und Videoaufzeichnungen gibt es aber in jüngster Zeit bemerkenswerte Fortschritte. So lassen sich inzwischen umfangreiche Sprachkorpora automatisch analysieren. Die Sprachaufnahmen können mittels statistischer Verfahren in kleinere Segmente unterteilt werden, Sprecher erkannt und Sprache in Text umgewandelt werden. Diese Techniken werden unter dem Begriff Audio Mining zusammengefasst und ermöglichen den exakten Zugriff auf einzelnen Begriffe, Abschnitte und Ereignisse in Audiodaten.

Technologien und Anwendungen des Audio Minings sind daher für verschiedenste Bereiche der Geistes- und Kulturwissenschaften von großem Interesse. So lassen sich zum Beispiel Sprachaufnahmen automatisch vorsegmentieren, so dass diese anschließend mit gängigen Tools (z.B. ELAN und EXMARaLDA) weiter verarbeitet werden können. Diese Anwendung von Audio Mining ist besonders für die Forschung in Linguistik und Gesprächsanalyse sowie anderen Bereichen, die sich mit Interaktion und Sprache beschäftigen, interessant. Für Oral History, Ethnologie und andere vor allem an inhaltlichen Aspekten der Sprachdaten interessierte Fachgebiete lassen sich die hier typischerweise sehr langen Interviews automatisch segmentieren und transkribieren. Biographische, narrative Interviews sind das Produkt eines kommunikativen Geschehens und damit hoch komplexe Quellen mit vielfältigen Ansatzpunkten für die Interpretation. Aus forschungspraktischen Gründen wird die Analyse von Interviewdaten bisher meist auf das schriftsprachliche Medium reduziert. Das Audio Mining eröffnet hier ganz neue Dimensionen und Fragestellungen. Während die Fallzahlen bei konventionellen Analysemethoden meist bei um die 30 Interviews liegen, können mit technischer Unterstützung viel größere Fallzahlen bearbeitet und somit unter vielfältigen vergleichenden Fragestellungen auch quantitativ ausgewertet werden. Gleichzeitig bieten die Werkzeuge auch der qualitativen Analyse neue Dimensionen, indem sowohl sprachliche wie nicht-sprachliche Aspekte der Kommunikation differenziert erfasst und somit für Forschungsfragen zugänglich gemacht werden können. Um diese Potentiale realisieren zu können, besteht allerdings Forschungsbedarf.

Für die Sprachtechnologien stellen die Audiodaten vieler Fachrichtungen der Digital Humanities interessante Herausforderungen dar. Denn während die Spracherkennung bei Sprachaufnahmen aus dem Nachrichtenbereich (optimale Aufnahmebedingungen und Aufnahmetechnik, artikulierte Hochsprache professioneller Sprecher) inzwischen gute Ergebnisse liefert, stellen zum Beispiel Zeitzeugeninterviews noch eine erhebliche Herausforderung dar. Ursachen sind die oft schlechte Qualität der Audiodaten bedingt durch Modalitäten der Aufzeichnung (Rauschen, Übersteuerung, geringe Lautstärke durch schlechte Platzierung des Mikrophons), Alterungsprozesse der Magnetbänder bis zur Digitalisierung und Fehlentscheidungen bei der Wahl des Audioformats sowie die Charakteristiken der Spontansprache (undeutliche Aussprache, schnelles Sprechen und Dialektfärbung). Bei den Aufnahmen aus Linguistik und Interaktionsforschung handelt es sich oft um Daten aus wenig erforschten Sprachen, zu denen es kaum Ressourcen und wenig andere linguistische Information gibt. Aber selbst Daten aus gut erforschten Sprachen wie dem Deutschen können für die Analyseverfahren anspruchsvoll sein. Datensets bestehen in beiden Fällen häufig aus natürlicher, spontaner Sprache mit mehreren Gesprächsteilnehmern und entsprechenden Sprechüberlappungen. Werkzeuge und Verfahren, die an Daten aus geplanter Sprache trainiert wurden, stoßen bei diesen Aufnahmen schnell an ihre Grenzen. Da diese Daten nicht unter Studiobedingungen erhoben werden können, ist darüber hinaus die Aufnahmequalität zumeist deutlich geringer als bei Aufnahmen aus Hörfunk und Fernsehen. Die im Vergleich zu Daten auf dem Rundfunk recht kleinen Datensets erschweren dabei die Anwendung von intelligenten Analyse- und Auswertungsverfahren noch weiter. So bestehen eine Vielzahl von Forschungsthemen im Bereich der Sprachanalyse, wie beispielsweise die Detektion von überlappenden Sprachsegmenten, eine robuste Sprechersegmentierung von kurzen Dialogsequenzen, Diarisierung von Aufnahmen für die Analyse von Turn-Takings, die Erkennung von Sprechern, Erkennung von Dialekten, robuste Transkription von Sprachdaten hinsichtlich Hintergrundgeräusche und Raumhall, Code-Switching in gesprochener Sprache, Erkennung von Pausen und gegebenenfalls Satzzeichen.

Weitere Informationen zu den Ausrichtern des Workshops

Der Workshop wird von dem an der Universität zu Köln angesiedelten BMBF-Zentrum »Kölner

Zentrum Analyse und Archivierung von AV-Daten« (KA³) ausgerichtet. Ziele des über drei Jahre geförderten Projektes sind die Erforschung und Entwicklung von Werkzeugen und Services zur akustischen Analyse von AV-Daten, die Einrichtung einer Nutzerplattform für die Analyse und Archivierung von AV-Daten sowie die Erprobung und Anwendung der Methoden in ausgewählten Pilotprojekten im Bereich der Oral History/Biografieforschung und der Linguistik/Interkulturelle Kommunikation. Besondere Aufmerksamkeit gilt den miteinander zusammenhängenden Problemen der interaktionsbezogenen Strukturierung und der effizienten Bereitstellung und Archivierung von audiovisuellen Daten. Im Rahmen des Verbundprojektes wird ein fach- und standortübergreifendes Zentrum für die Analyse und Archivierung audiovisueller Daten mit den drei Komponenten Analyse, Archivierung/Publikation und Schulung/Beratung aufgebaut, das die im Projekt entwickelten Werkzeuge und Services interessierten Forscherinnen und Forschern zur Verfügung stellt. Das geplante Zentrum ist institutionell eingegliedert in das umfassender angelegte Kölner Data Center for the Humanities (DCH), das informationstechnologische Unterstützung für alle Datentypen in den Geisteswissenschaften anbietet und erforscht. Das Projekt KA³ ist eine Kooperation des Instituts für Linguistik (N. P. Himmelmann), dem Regionalen Rechenzentrum (U. Lang) und dem Data Center for the Humanities (A. Witt und B. Mathiak) der Universität zu Köln, dem Fraunhofer- Institut für Intelligente Analyse- und Informationssysteme, Sankt Augustin (J. Köhler), dem Institut für Geschichte und Biographie der FernUniversität Hagen (A. Leh) und dem Max-Planck-Institut für Psycholinguistik (P. Trilsbeek).

Nikolaus P. Himmelmann lehrt Allgemeine Sprachwissenschaft an der Universität zu Köln. Zentrales Forschungsgebiet ist sprachliche Diversität und was diese uns über menschliche Kognition und Gesellschaftlichkeit lehrt. Er hat entscheidend zur Entwicklung des Konzepts digitaler Sprachdokumentationen und Spracharchive beigetragen.

Joachim Köhler leitet den Bereich Spracherkennung am Fraunhofer Institut IAIS. Zentrale Arbeitsgebiete sind maschinelles Lernen, Mustererkennung, Deep Learning, Information Retrieval, Medieninformationssysteme, Metadaten sowie Linked Data.

Almuth Leh leitet das Archiv ‚Deutsches Gedächtnis' an der FernUniversität Hagen und forscht zu deutscher Mentalitätsgeschichte im 20. Jahrhundert mit Einzelarbeiten zum Naturschutz in Nordrhein-Westfalen, Gewerkschafterinnen, Soldaten im Zweiten Weltkrieg, Wehrmachtsjustiz, Universitätsgeschichte. Weitere Interessensgebiete sind forschungsethische und methodische Fragen der Oral History und die Archivierung lebensgeschichtlicher Interviews.

Felix Rau ist Projektmitarbeiter im Projekt »Kölner Zentrum für Analyse und Archivierung audiovisueller Daten« (KA³) und koordiniert die Faschspezifische Arbeitsgruppe »Linguistische Feldforschung, Ethnologie, Sprachtypologie« (Leitung: Himmelmann) in CLARIN-D. Er ist Feldlinguist mit einem Schwerpunkt auf den Munda-Zweig der austroasiatischen Sprachen und ist ein Archivmanager am Language Archive Cologne.

Automatic Text Recognition: Mit Transkribus Texterkennung trainieren und anwenden

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

Strauß, Tobias
tobias.strauss@uni-rostock.de
Universität Rostock CITLab

Diem, Markus
diem@caa.tuwien.ac.at
Vienna University of Technology, Computer Vision Lab

Die Aufbereitung und Erkennung von handschriftlichen Dokumenten oder von speziellen Druckschriften ist sowohl für Menschen als auch für Computeralgorithmen eine (technische) Herausforderung. Die Bearbeitung von schriftlichem, insbesondere handschriftlichem Material aber auch früher Drucke wird bislang von spezialisierten Experten durchgeführt, um technisch und qualitativ hochstehende Resultate aus historischen Dokumenten zu erhalten. Zur Erstellung hochwertiger Editionen sind hilfswissenschaftliche Kenntnisse (Paläographie, Editorik), historisches Kontextwissen und technisches Know-how gefragt.

Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Documents) werden unterschiedliche Aufgaben der Automatisierung (weiter-)entwickelt, um qualitativ gute Ergebnisse mit optimalem Ressourceneinsatz zu erhalten. Ein speziell dafür entwickeltes Tool ist die Software Transkribus und die Transkribus Weboberfläche (öffentliche Vorstellung im November 2017). Beide Ansätze verkoppeln auf unterschiedliche Weise die Arbeit von Expertinnen und maschinelle Erkennleistung. Software und Webservice sind frei verfügbar unter www.transkribus.eu. Im Workshop wird Transkribus vorgestellt und kann durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden.

Transkribus unterstützt alle Prozesse vom Import der Bilder über die Identifikation der Textblöcke und Zeilen, die zu einer detaillierten Verlinkung zwischen Text und Bild führt, sowie die Transkription und Annotation der Handschrift bis zum Export der gewonnen Daten in standardisierten Formaten.

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Dokumente können entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt werden oder die Transkription erfolgt händisch und kann danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung großer Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit Metadaten (Identifikation von Personen, Orten und Sachwörtern) ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu den meisten herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen. Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) implementiert. Schließlich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Zielpublikum

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

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout- und Handschriftenerkennung, lässt sich nur durch die enge Zusammenarbeit zwischen GeisteswissenschaftlerINNEn und ComputerspezialistINNen mit unterschiedlichen Voraussetzungen und Ansprüchen an Datenqualität und Herstellung von Transkriptionen erreichen. Die Algorithmen werden somit nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in grösseren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die ComputerwissenschaftlerINNEN sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird.

Die Speicherung der Dokumente erfolgt in der Cloud, gehostet auf Servern der Universität Innsbruck. Die importierten Daten bleiben auch während der Bearbeitung unverändert im Dateisystem liegen und werden ergänzt durch METS und PAGE XML. Alle bearbeiteten Dokumente und Da-

ten bleiben somit in den unterschiedlichen Bearbeitungsstadien nicht nur lokal verfügbar, sondern können für andere Transkribusnutzerinnen und -nutzer freigegeben werden. Dank elaboriertem *user-management* ist die Zuteilung von Rollen möglich.

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

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

Ein- und Ausblicke im Workshop

Der Workshop richtet sich sowohl an GeisteswissenschaftlerINNEN als auch an ComputerwissenschaftlerINNEN, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Zwei zentrale Forschungsaspekte aus READ werden im Rahmen des Workshops durch Experten vorgestellt:

Einerseits das technische Verfahren des Automatic Text Recognition mit rekurrenten neuronalen Netzen (Leifert et al. 2016). Dabei wird kurz in die Trainings- und Auswertungsmechanismen mit neuronalen Netzen eingeführt und Möglichkeiten der Auswertung demonstriert.

Andererseits wird die Erkennung von komplexen Layouts, insbesondere Tabellen, erklärt und neueste technische Lösungen vorgestellt.

Programm/Ablauf des Workshops

- *Begrüßung und Informationen zum Projekt READ* (Tobias Hodel, Zürich): 20‘
  Überblick über Ziele und Fortschritte im Rahmen des von der EU geförderten Projekts.
- *Machine Learning und automatisierte Text Erkennung* (Tobias Strauß, Rostock): 30‘
  Einführung und Erklärung zum Einsatz neuronaler Netze bei der Texterkennung
- *Einführung in Transkribus* 30‘
  Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen.
- *Selbstständiges Arbeiten der Teilnehmenden mit Transkribus:* 90‘
- Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (falls gewünscht mit eigenen Dokumenten) selbst ausgetestet werden.
- *Layout Analyse: Tabellen und andere schwierige Formen* (Markus Diem, Wien): 30‘
  Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist es, auch komplexe Strukturen korrekt als Layout zu erkennen, um die automatisierte Texterkennung überhaupt zu ermöglichen. Tabellen gehören in dem Bereich zu den schwierigsten Formen der Texterkennung.
- *Diskussion über Vor- und Nachteile der Software:* 30‘
  Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software und Webtools (usability, Umfang und Leistung der Automatisierungen etc.).
- Nach Interesse der Teilnehmenden werden während des Workshops Kurzinputs zu folgenden Themen angeboten:
  1. Matching von Text und Bild (bspw. aus bestehenden Transkriptionen),
  2. Transkribus Learn (e-Learningumgebung),
  3. Crowdsourcing-Infrastruktur,
  4. ScanTent und DocScan (Fotografieren eigener Dokumente mit Android App).

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Dokumente können entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt werden oder die Transkription erfolgt händisch und kann danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung großer Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit Metadaten (Identifikation von Personen, Orten und Sachwörtern) ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu den meisten herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen. Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) implementiert. Schließlich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Zielpublikum

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

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout- und Handschriftenerkennung, lässt sich nur durch die enge Zusammenarbeit zwischen GeisteswissenschaftlerINNEn und ComputerspezialistINNen mit unterschiedlichen Voraussetzungen und Ansprüchen an Datenqualität und Herstellung von Transkriptionen erreichen. Die Algorithmen werden somit nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in grösseren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die ComputerwissenschaftlerINNEN sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird.

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

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

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

Ein- und Ausblicke im Workshop

Der Workshop richtet sich sowohl an GeisteswissenschaftlerINNEN als auch an ComputerwissenschaftlerINNEN, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Zwei zentrale Forschungsaspekte aus READ werden im Rahmen des Workshops durch Experten vorgestellt:

Einerseits das technische Verfahren des Automatic Text Recognition mit rekurrenten neuronalen Netzen (Leifert et al. 2016). Dabei wird kurz in die Trainings- und Auswertungsmechanismen mit neuronalen Netzen eingeführt und Möglichkeiten der Auswertung demonstriert.

Andererseits wird die Erkennung von komplexen Layouts, insbesondere Tabellen, erklärt und neueste technische Lösungen vorgestellt.

Programm/Ablauf des Workshops

Begrüßung und Informationen zum Projekt READ (Tobias Hodel, Zürich): 20‘

Überblick über Ziele und Fortschritte im Rahmen des von der EU geförderten Projekts. *Machine Learning und automatisierte Text Erkennung* (Tobias Strauß, Rostock): 30‘

Einführung und Erklärung zum Einsatz neuronaler Netze bei der Texterkennung *Einführung in Transkribus* 30‘

Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen. *Selbstständiges Arbeiten der Teilnehmenden mit Transkribus:* 90‘

Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (falls gewünscht mit eigenen Dokumenten) selbst ausgetestet werden.

Layout Analyse: Tabellen und andere schwierige Formen (Markus Diem, Wien): 30‘

Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist es, auch komplexe Strukturen korrekt als Lay-

out zu erkennen, um die automatisierte Texterkennung überhaupt zu ermöglichen. Tabellen gehören in dem Bereich zu den schwierigsten Formen der Texterkennung.

Diskussion über Vor- und Nachteile der Software: 30‘

Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software und Webtools (usability, Umfang und Leistung der Automatisierungen etc.).

Nach Interesse der Teilnehmenden werden während des Workshops Kurzinputs zu folgenden Themen angeboten:

- Matching von Text und Bild (bspw. aus bestehenden Transkriptionen),
- Transkribus Learn (e-Learningumgebung),
- Crowdsourcing-Infrastruktur,
- ScanTent und DocScan (Fotografieren eigener Dokumente mit Android App).

Während des gesamten Workshops stehen drei wissenschaftliche Mitarbeitende des Projekts für Fragen und Auskünfte zur Verfügung. **Tobias Hodel (nimmt bereits im Vorfeld gerne Dokumente oder Projektideen an, damit sich die Veranstalter bereits vor dem Workshop Gedanken zu möglichen technischen Umsetzungen machen können.**

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

Zahl der möglichen Teilnehmerinnen und Teilnehmer: 30-40 Personen (auch abhängig von der Raumgrösse).

Benötigte technische Ausstattung: Allgemein: Beamer, evtl. Whiteboard.

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

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

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

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

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

Tobias Strauß, Institut für Mathematik, Ulmenstraße 69, Universität Rostock, 18051 Rostock, Deutschland; tobias.strauss@uni-rostock.de; (Deep Learning, Information Retrieval und Natural Language Processing).

Bibliographie

Leifert, G., Strauß, T., Grüning, T., Wustlich, W., Labahn, R., 2016. Cells in Multidimensional Recurrent Neural Networks. Journal of Machine Learning Research 17, 1-37.

Leifert, G., Strauß, T., Grüning, T., Wustlich, W., Labahn, R., 2016. Cells in Multidimensional Recurrent Neural Networks. Journal of Machine Learning Research 17, 1-37.

CorpusExplorer v2.0 - Seminartauglich in einem halben Tag

Rüdiger, Jan Oliver
jan.ruediger@uni-kassel.de
Universität Kassel, Deutschland

Grundzüge:

Die Fähigkeit zur Kritik setzt voraus, dass die notwendigen Fähigkeiten zur Durchführung vorhanden sind. Es bedarf jedoch des Mutes sich zu entschließen, durchzuhalten und nicht bequem zu werden und somit die eigenen Fähigkeiten/Methoden kontinuierlich zu verbessern. Methoden wie Sie in den Digital Humanities und speziell in der Korpuslinguistik zum Einsatz kommen, lassen sich nur verbessern, wenn man selbst tätig wird, hinterfragt, ausprobiert und gemeinsam diskutiert. Im Rahmen dieses Workshops wird der CorpusExplorer v2.0 vorgestellt (OpenSource), der unterschiedlichste Methoden aus dem Bereich der Forschung holt und diese für die universitäre Lehre bereitstellt. Studenten sollen mit dieser Software ermutigt werden, eigene kleine Forschungsprojekte zu realisieren (es wurden bereits Seminararbeiten, Bachlor-/Masterarbeiten sowie (laufende) Dissertationsprojekte mittels CorpusExplorer umgesetzt).

Dies ist nicht selbstverständlich, so weisen bereits (Bubenhofer 2011) „Oft bedingen korpuslinguistische Arbeiten einen großen Aufwand, sowohl für Lernende als auch die Betreuenden, der im Rahmen eines Studiums nicht geleistet wer-

den kann.“ oder (Dipper 2011) „Bei der Arbeit mit ‚echten‘ Daten, [...] werden die Computerlinguistik- Studenten früh mit Problemen wie dem Daten-Encoding oder der Datengröße konfrontiert [...]“ auf elementare Probleme zu Seminar-/Projektstart hin. Außerdem ist es in der Regel notwendig, dass unterschiedliche Programme kombiniert werden, um ein (visuelles) Ergebnis zu erzielen.

Der CorpusExplorer v2.0 beseitigt viele dieser (Einstiegs-)Hürden. Unterschiedlichste Programme und Methoden werden unter einer benutzerfreundlichen Programmoberfläche kombiniert, die zudem vielfältige Visualisierung/Weiterverarbeitungsmöglichkeiten zur Verfügung stellt (wie auch in den Video-Tutorials von (Rüdiger 2017) bereits gezeigt wurde). Im Vergleich zu AntConc, TXM und anderen verbreiteten Tools wird schnell klar, wie stark sich der CorpusExplorer an Forschung -und- Lehre orientiert.

Einen Einblick in die Workshopinhalte:

Korpora erstellen

Der CorpusExplorer automatisiert den gesamten Erstellungsprozess. Die Möglichkeiten Korpusmaterial zu akquirieren sind vielfältig – PDF, eBooks, X/HML, Tweets, Blogs, uvm. – lassen sich einlesen, Text-/Metadaten werden getrennt, der Text bereinigt (z. B. von unerwünschten HTML/XML-Tags), abschließend erfolgt eine Annotation (z. B. durch den TreeTagger, Stanford POS oder TnT). Vom Rohtext zum analysefertigen Korpus ist die Nutzer*in nur wenige Mausklicks entfernt. Dies bietet verschiedene Möglichkeiten, Korpora werden entweder zentral durch Dozent*in/Tutor*in bereitgestellt oder selbst von den Student*innen aufgebaut. Außerdem besteht die Möglichkeit, Korpusmaterial im Seminarverlauf gemeinsam zu pflegen, zu erweitern und auszurollen.

Auswertungen

Im CorpusExplorer stehen über 45 Analysemodule zur Verfügung. Bei einigen davon überschneidet sich die Datengrundlage – es wird erprobbar, wie Visualisierungen in den Rezeptionsprozess eingreifen. Am Beispiel der Kookkurrenzanalyse wird dies besonders deutlich, optisch unattraktiv weil mächtig und funktional umfangreich – die Tabellen-Darstellung zeigt alle Daten auf einmal (filter-/gruppier-/sortierbar). Als WordCloud (auf Basis der TagCloud-Visualisierung von (Jänicke et al. 2015)) werden Bedeutungen/Schnittmengen einzelner Begriffe schnell sichtbar. In der Graphen-Darstellung lassen sich Zusammenhänge explorativ erkunden. Kombiniert mit der Auswertung von N-Grammen lassen sich Sprachgebrauchsmuster schnell identifizieren (welche Teile eines N-Gramms sind signifikante Kookkurrenzpartner?).

Nicht nur Text-Daten lassen sich auswerten, sondern auch Meta-Daten, diese werden während des Erstellprozesses automatisch mit erfasst oder lassen sich nachträglich manuell erweitern/ändern. Im Seminar können unterschiedliche Ansätze/Haltungen auf diese Weise erprobt werden – z. B. ob und wie sich die Korpuszusammensetzung auf das Analyseergebnis auswirkt (Bsp.: ausgewogenes Korpus, pragmatischer Ansatz oder „more data is better data“).

Das Konzept der Schnappschüsse erlaubt die Filterung/Zusammenstellung individueller Teilkorpora. Korpusmaterial und Analysen sind durch Schnappschüsse voneinander isoliert. Neu hinzukommendes Material verwirft keine bisherigen Ergebnisse. Aus vielen Analysen lassen sich Schnappschüsse erstellen, um konkreten Forschungsfragen nachzugehen – Es können individuelle Filter auf Text-, Meta- und Korpus-Ebene getroffen werden oder der CorpusExplorer kann anhand von Vorgaben mehrere Teilkorpora auf einmal erstellen (z. B. jeder Autor/Verlag isoliert – Datums/Zeitabschnitte, etc.). Letztlich lassen sich Schnappschüsse durch Mengenoperatoren kombinieren.

Schnappschüsse erlauben es, das gegenwärtige Forschungsinteresse zu lenken und gezielt durch große Korpusmengen zu navigieren. Gegenwärtig stehen sich corpus-driven und corpus-based Ansätze sowie close- und distant-reading gegenüber. Da Ansätze beider Denkrichtungen in diesem Programm vereint sind, erlaubt der CorpusExplorer nicht nur einen schnellen Perspektivwechsel sondern schafft neue Möglichkeiten des Arbeitens.

Transparenz & Anbindung an andere Programme / Programmiersprachen

Ein wesentliches Teil des OpenSource-Gedankens im CorpusExplorers ist: Transparenz – Nicht nur, dass sich viele Formate verarbeiten lassen, auch der Prozess lässt sich ohne viel Aufwand verifizieren (Generator zur Erzeugung von Dummy-Korpora für die Prozessvalidierung) und alle Analyseergebnisse lassen sich exportieren (inkl. einer Konvertierung des CorpusExplorer-Formats in andere XML/JSON-Formate). Aus dieser, anfänglich als Möglichkeit gedachten Funktion, entstanden zwei Möglichkeiten für fortgeschrittene

Nutzer*innen. Das HTML5-Labor erlaubt es, Daten/Analysen direkt im CorpusExplorer mittels HTML5, JavaScript und CSS zu visualisieren (Syntax-Editor inkl. HTML5 Rendering-Engine auf Basis von Chromium stehen bereit). Die zweite Möglichkeit erlaubt den kompletten Verzicht auf die Programmoberfläche. Mittels R, Kommandozeile oder anderer Programmiersprachen kann auf den CorpusExplorer zugegriffen werden. Der CorpusExplorer ist keine Ultima Ratio – er kann aber helfen, einen schnelleren Einstieg in die Korpuslinguistik zu finden und erleichtert selbst in fortgeschrittenen Szenarien die Arbeit.

Wie im Seminar einbinden

Ein didaktisches Konzept wird im Rahmen dieses Workshops nicht vorgestellt. Aus der Praxiserfahrung heraus werden einige Beispiele erfolgen, wie man diese Software bereits im Bachelor-Studium einsetzen kann. Zum Beispiel wie die von (Bubenhofer 2011) angesprochene Problematik „[…] die Studierenden überhaupt dazu zu motivieren, korpuslinguistisch, also empirisch zu arbeiten.“ durchbrochen werden kann. Dies gelingt im Wesentlichen dadurch, dass die Studenten bereits in der ersten Seminarsitzung aktiv arbeiten können und erkennen, welchen Mehrwert und Spaß empirisches Arbeiten bietet.

Workshopverlauf

Im Workshop wechseln sich vier Sozialformen in unterschiedlicher Reihung ab. In den Vortrags-/Frontal-Sequenzen werden die wesentlichen Funktionen und deren Hintergründe erklärt – z. B. wie werden N-Gramme ausgezählt? In Live-Demonstrationen wird die konkrete Anwendung gezeigt – Was muss eingestellt / ausgewählt / angeklickt werden, um ein Ergebnis zu erreichen? Dabei können die Teilnehmer*innen das Gezeigte an ihrem Rechner nachverfolgen. Außerdem erfolgt eine eigene Erprobungsphase – hier probieren die Teilnehmer*innen selbständig eigene Parameter aus und durchforsten das Korpusmaterial auf eigene Faust. Abschließend wird alles innerhalb der Workshopgruppe diskutiert. Für den Workshop werden unterschiedliche Korpustypen zur Verfügung gestellt, eigenes Korpusmaterial kann ggf. in der Erprobungsphase getestet werden.

Voraussetzungen

Im Idealfall würde ein PC-Poolraum mit Windows-Rechnern ab Windows 7 inkl. Beamer zur Verfügung gestellt, auf dem der CorpusExplorer bereits vorinstalliert ist. Ggf. könnten die Teilnehmer*innen die Software auch selbst installieren (für die Installation sind keine Administratoren-Rechte notwendig). Falls kein geeigneter Poolraum zur Verfügung stehen sollte, wäre ein Seminarraum mit Internetzugang, Beamer und ausreichenden Steckdosen für die Teilnehmer*innen notwendig. Die Teilnehmer*innen könnten ihren eigenen Windows-Rechner mitbringen. Teilnehmer*innen mit Linux/MacOS Notebook sollten sich zuvor melden, damit eine Virtualisierungslösung bereitgestellt werden kann.

Bibliographie

Bubenhofer, N. (2011): Korpuslinguistik in der linguistischen Lehre: Erfolg und Misserfolge. In: Journals for Language Technology and Computational Linguistics 26 (1), S. 141–156.

Dipper, S. (2011): Digitale Korpora in der Lehre: Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik. In: Journals for Language Technology and Computational Linguistics 26 (1), S. 81–95.

Jänicke, S.; Blumenstein, J.; Rücker, M.; Zeckzer D. & Scheuermann, G. (2015): Tagpies.

Rüdiger, J. (2017): Korpushermeneutische Analyse politischer Reden mittels CorpusExplorer. In: 10plus1 - Living Linguistics (3), S. 11–21.

Digitale Bildrepositorien – wirkliche Arbeitserleichterung oder zeitraubend?

Friedrichs, Kristina
kristina.friedrichs@tu-dresden.de
Universität Würzburg, Deutschland

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

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

Maiwald, Ferdinand
ferdinand.maiwald@tu-dresden.de
TU Dresden, Deutschland

Bruschke, Jonas
jonas.bruschke@uni-wuerzburg.de
Universität Würzburg, Deutschland

Barthel, Kristina
kristina.barthel@tu-dresden.de
TU Dresden, Deutschland

Diese Erfahrung teilen sicher viele: Bei der Suche nach einer bestimmten Abbildung, zum Beispiel nach einem Foto eines Gebäudes, konsultiert man verschiedene Online-Plattformen, doch entweder werden gar keine Treffer ausgegeben oder aber man erhält eine unüberschaubare Flut an Ergebnissen, die leider oftmals nicht relevant für das eigene Suchanliegen sind. So kann die gewünschte Arbeitserleichterung durch digitale Mittel leider fast schon hinderlich wirken.

Der Workshop setzt drei Schwerpunkte: Erstens werden im Rahmen einer Podiumsdiskussion die Experten der Bildrepositorien, so von Prometheus, der Deutschen Fotothek und der UB Heidelberg, mit einem tiefgreifenden Austausch in das Programm starten. Daraufhin sollen die Teilnehmer für den Umgang mit Bildrepositorien sensibilisiert, Suchstrategien in den Fokus gestellt und Kompetenzen in diesem Bereich vertieft werden. Hierbei werden vor allem diejenigen, die mit solchen Quellenbeständen arbeiten, mit ihren Erfahrungen und Wünschen zu Wort kommen. Schließlich soll ein komplementärer Lösungsansatz vorgestellt und kritisch durch die Teilnehmer geprüft werden.

Der Workshop richtet sich vorrangig an Kunst- und Architekturhistoriker, Nutzer und Anbieter von Bild- und Fotoarchiven sowie digital humanists im Bereich Bild.

Problemaufriss

Gerade vor dem Hintergrund, dass immer mehr historische Bestände an Bild- und Planmaterial digitalisiert und in Online-Repositorien der Öffentlichkeit zur Verfügung gestellt werden, stellt sich das Problem, dass sich die Suche vor allem nach ortsbezogenen Bildquellen stetig verkompliziert (Bauer 2015). Bei der Arbeit mit diesem Material ist eine hohe Kompetenz im Umgang mit Schlagwörtern erforderlich; es muss ergo die Erschließung des dargestellten Gegenstandes anhand von Wörtern sowie gegebenenfalls ein sinnvoller Einsatz von etwaig vorhandenen Filtern beherrscht werden – andernfalls wird eine solche Recherche unter Umständen ineffizient und wenig zielführend bleiben (Beaudoin und Brady 2011; Kohle 2013: 15-62; Kamposiori 2012).

Im Rahmen des Workshops soll ein erweiterter Ansatz, der die Erschließung umfangreicher historischer Bildbestände mit örtlicher Gebundenheit unterstützen kann, vorgestellt und getestet werden. Hiernach bildet die schlagwortgebundene Suche nur einen ergänzenden Aspekt, während die Quellen deutlich in den Vordergrund treten. Die Idee basiert auf einer Verortung der Fotografien innerhalb eines vierdimensionalen Modells, so dass die topographische Lage innerhalb einer Stadt den Aufhängungspunkt für die dazugehörigen Quellen bildet und die Zeit als zusätzliche Komponente hinzutritt, um Veränderungen gerecht zu werden.

Forschungsstand & Projektskizze

Eine automatische Bildersuche auf Grundlage von bestimmten Bildmerkmalen erweitert hierbei die schlagwortgebundene Suche (Jégou et al., 2010). Möglicherweise können auch Bildersuchen mit neuronalen Netzwerken klassische Suchanfragen unterstützen (Wan et al., 2014). Mit manueller Verortung beziehungsweise an einzelnen Objekten existieren bereits dreidimensionale und vierdimensionale Ansätze (Schindler & Dellaert 2012, Agarwal et al. 2009, Bitelli et al. 2017).

Der hinzutretende Ansatz ist hingegen der Versuch, historische Bilder aus einer Datenbank möglichst automatisch zu verorten. Nach einer Auswahl von geeignetem Bildmaterial werden die Bilder mittels eines zweistufigen Verfahrens in einem dreidimensionalen Modell lokalisiert. Zuerst werden die Fotografien über spezifische Merkmale relativ zueinander orientiert, so dass Gebäudeteile, die mehreren Bildern gemein sind, perspektivisch korreliert werden können. Anschließend erfolgt die absolute Verortung über aktuelle 3D-Modelle in einem Stadtmodell (Vietze et al. 2017). Für Gebäude mit genügend Bildmaterial ist es sogar möglich, historische dreidimensionale Modelle mittels Structure-from-Motion zu erstellen, so dass in mehreren Zeitschnitten eine tatsächliche Vierdimensionalität erreicht wird.

Aus dem beschriebenen Modell heraus soll die eigentliche Mediensuche geschehen. Es existieren bereits einige Lösungen, die auf Basis einer räumlichen Suche funktionieren, zum Beispiel www.historypin.org oder www.phillyhistory.org. Diese greifen auf Google Maps oder OpenStreetMap und damit auf aktuelles Kartenmaterial zurück. Dadurch ist nur eine zweidimensionale räumliche Suche innerhalb der jetzigen Bausitua-

tion möglich. Zudem richtet sich diese Anwendung nicht an Fachleute, die sie im Sinne eines wissenschaftlichen Medienrepositoriums benutzen. Über diesen Stand soll hiermit hinausgegangen werden.

Durch einen Abgleich mit einem dreidimensionalen Modell der Stadt können Bilder aus Medienrepositorien direkt Objekten im Stadtbild zugeordnet werden, was einerseits ein objektbasiertes Durchsuchen des gesamten Repositoriums innerhalb eines 4D-Bowsers ermöglicht. Andererseits erleichtert ein historisches 3D-Stadtmodell darüber hinaus die Orientierung im damaligen Stadtbild, unter anderem durch die Abbildung der gesamten Situation und die klare Markierung des Standpunkts des Fotografen (Schindler & Dellaert 2012). Erweitert durch geeignete Analysewerkzeuge kann die Arbeit von Kunsthistorikern und Bildwissenschaftlern unterstützt werden (Bruschke et. al 2017).

Aufbauend auf diesen Überlegungen wird derzeit ein Prototyp erarbeitet. Es handelt sich dabei um eine Webanwendung, die in einem modernen, HTML5/WebGL-fähigen Browser läuft. Einen Großteil des Seitenaufbaus nimmt der 3D-Viewport ein, in dem ein (derzeit noch aktuelles) 3D-Stadtmodell zusammen mit den bereits verorteten Bildern angezeigt wird und in dem navigiert werden kann. Über eine Suchleiste können die Bilder mittels Schlagwörtern eingegrenzt werden. Alternativ ist dies über eine facettierte Suche für ausgewählte Metadaten möglich. Eine weitere Filterung erfolgt über die Auswahl der Gebäude im 3D-Viewport. Neben den 3D-Repräsentation im Viewport werden die gefunden Bilder auch als klassische Ergebnisliste mit Thumbnails präsentiert. Für jedes Bild kann eine Detailansicht geöffnet werden, um alle Metadaten und das Bild im Detail zu betrachten. Diese Anwendung soll hiermit zur Diskussion gestellt werden.

Programmplanung

Der Workshop wird sich nicht allein auf die kritische Betrachtung des vorgeschlagenen Ansatzes konzentrieren, sondern ebenso Suchstrategien innerhalb von Online-Plattformen vermitteln und Defizite sowie gelungene Lösungen seitens bestehender Repositorien eruieren. Daher wird es ein mehrteiliges Programm geben.

Begonnen wird mit entsprechenden Impulsvorträgen, die in den Problemkreis einführen. Es schließt sich eine Podiumsdiskussion mit Vertretern namhafter Bildrepositorien an; so werden Prometheus mit Prof. Holger Simon, die Deutsche Fotothek mit Marc Rohrmüller und die UB Heidelberg mit Dr. Maria Effinger vertreten sein.

In einem nächsten Durchgang soll ein Vergleich zwischen bereits existierenden Repositorien durch die Nutzer gezogen werden, in dem Charakteristika, Vorteile und ggf. Verbesserungspotential eruiert werden. Auf diese Weise können verschiedene Suchstrategien vermittelt und erprobt werden, mit denen in vorhandenen Repositorien effizient Resultate erzielt werden können.

Daraufhin soll die seitens der ausrichtenden Gruppe die Testversion des 4D-Browsers als Plattform vorgestellt werden, um den erreichten Stand zu präsentieren und eine Einweisung in die Funktionsweise zu geben. Schließlich werden die Teilnehmer aktiv involviert, indem sie verschiedene Aufgaben in kleinen Gruppen von fünf bis sechs Personen bearbeiten, die der tatsächlichen kunsthistorischen Tätigkeit entlehnt sind.

Zu Beginn soll die Funktion zum Einbinden von Quellen kritisch getestet werden, sowohl im Hinblick auf räumliche als auch auf zeitliche Aspekte. Welche Features sind für Architekturhistoriker besonders nutzbringend? Wo werden Probleme gesehen?

Daraufhin wird die eigentliche Suche erprobt, indem einerseits eine von Metadaten dominierte Suche mit einer rein ortsbasierten Suche verglichen wird. Wie zufriedenstellend sind die Resultate? Wo ergänzen sich beide sinnvoll?

Als nächster Fall sollen Baustrukturen verglichen werden, etwa der Vorkriegszustand eines Bauwerks mit der Fassung der späteren Rekonstruktion. Welche Funktionen sind hier besonders wünschenswert, etwa Überblenden, Markierungen, Lineale?

Eine vierte Aufgabe ist den digital humanities entlehnt und wird sich mit der Anzahl der Bilder beschäftigen, und mittels eines numerisches Verfahren nach der ikonischen Strahlkraft eines Bauwerks oder einer besonderen Ansicht fragen. Welche Ansicht hat besondere Aufmerksamkeit erfahren? Wird dieses Tool auch zur Erfassung von eher komplexen Auswertungsaufgaben dieser Art als geeignet empfunden?

Der Workshop wird im Rahmen einer gemeinsamen Präsentation der Erfahrungen zur Ergebnissicherung vornehmen.

Ausblick

Der Workshop soll zur Sensibilisierung von Nutzern und auch Entwicklern von Bildrepositorien beitragen, indem sich beide Gruppen aktiv in die Veranstaltung einbringen können und sollen. So wird neben einer Kompetenzvermittlung auch die Entwicklung neuer Suchstrategien im Fokus stehen, um die Nutzbarkeit von Medienrepositorien zu erleichtern beziehungsweise einen Dialog

über zukünftige Möglichkeiten des Einsatzes solcher Plattformen anzusteuern. Somit werden sowohl bestehende Angebote als auch der ergänzende geotemporale Ansatz vorangebracht.

Infos

Max. 30 Teilnehmer. Bitte bringen Sie Ihre eigenen Laptops oder Tablets mit, möglichst mit Login in eduroam.

Bibliographie

Bauer, Elke (2015): „Analoge Bildarchive auf dem Weg ins digitale Zeitalter“ in: Irmgard Christa Becker (ed.): *Digitalisierung im Archiv - Neue Wege der Bereitstellung des Archivguts*. Marburg: Archivschule Marburg: 61–74.

Beaudoin, Joan / Brady, Jessica (2011): “Finding Visual Information: A Study of Image Resources Used by Archaeologists, Architects, Art Historians, and Artists” in: *School of Library and Information Science Faculty Research Publications*.

Kamposiori, Christina (2012): “Digital Infrastructure for Art Historical Research: thinking about user needs” in: Stuart Dunn (Hg.): *EVA 2012. London, UK, 10 - 12 July 2012*. Swindon: British Computer Soc: 245–253.

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

Vietze, Theresa / Schneider, Danilo / Maiwald, Ferdinand (2017): „Untersuchung der Eignung photogrammetrischer Methoden zur Erzeugung von 3D-Punktwolken aus historischen Bilddatenbeständen“ in: *37. Wissenschaftlich-Technische Jahrestagung der DGPF in Würzburg*, 26/2017.

Brusche, Jonas / Niebling, Florian / Maiwald, Ferdinand / Friedrichs, Kristina / Wacker, Markus / Latoschik, Marc Erich (2017) : “ Towards Browsing Repositories of Spatially Oriented Historic Photographic Images in 3D Web Environments” in: *Proceedings of the 22nd International Conference on 3D Web Technology*. ACM, New York, Article 18.

Schindler, Grant / Dellaert, Frank (2012): “4D cities. Analyzing, visualizing, and interacting with historical urban photo collections” in: *Journal of Multimedia* 7,2: 124-131.

Agarwal, Sameer / Snavely, Noah/ Simon, Ian / Seitz, Steven M. / Szeliski, Richard (2009): “Building Rome in a day” in: *International Conference on Computer Vision*.

Bitelli, Gabriele / Dellapasqua, M. / Girelli, V. A. / Sbaraglia, S. / Tini, M. A. (2017): “Historical photogrammetry and terrestrial laser scanning for the virtual 3D reconstruction of destroyed structures: a case study in Italy” in: *GEOMATICS & RESTORATION, Florence, Italy*: 113-119.

Jégou, Hervé / Douze, Mattijs / Schmid, Cordelia (2010): “Improving bag-of-features for large scale image search” in: *International journal of computer vision*, *87*(3), 316-336.

Wan, Ji / Wang, Dayong / Hoi, Steven Chu Hong / Wu, Pengcheng / Zhu, Jianke / Zhang, Yongdong / Li, Jintao (2014, November): “Deep learning for content-based image retrieval” in: *Proceedings of the 22nd ACM international conference on Multimedia*, 157-166.

Digitale Sammlungserschließung mit WissKI und CIDOC CRM

Scholz, Martin
martin.scholz@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Wagner, Sarah
s.wagner@gnm.de
Germanisches Nationalmuseum, Deutschland

Die systematische Erfassung und wissenschaftliche Erschließung einer Sammlung sind grundlegende Voraussetzungen, um ihr wissenschaftliches Potential sichtbar zu machen. Häufig aber fehlen Software-Lösungen und Know-How für eine flächendeckende Digitalisierung und Online-Präsenz.

Dieser Workshop führt anhand praktischer Beispiele in die digitale Sammlungsarbeit mit WissKI und in die Modellierung mit dem CIDOC Conceptual Reference Model (CRM) ein. Durch die praktische Arbeit lernen die Teilnehmer die im Projekt „Objekte im Netz" bereitgestellte Modellierung sowie die Konfiguration der Virtuellen Forschungsumgebung (VFU) für universitäre Sammlungen kennen.

Erschließung und Digitalisierung von Sammlungen

Neben Museen beherbergen auch Universitäten einen großen Schatz an Sammlungen, die der Wissenschaftsrat 2011 „als wertvolle Infrastruktur

für [...] Forschung" mit „beachtliche[m] wissenschaftliche[n] Potential" identifiziert hat. [1] Allein in Deutschland existieren rund 1000 Sammlungen an über 80 Universitäten. [2] Zwar sind darunter auch renommierte Sammlungen, doch leidet das Gros an unzureichender Erschließung, Sichtbarkeit, Betreuung, Pflege oder Unterbringung. [3] Auch bei der Digitalisierung gibt es enormen Aufholbedarf: Lediglich ein Drittel der Sammlungen sind digital zugänglich. Grund dafür sind u.a. auch das Fehlen von Software-Lösungen und Know-how für eine flächendeckende Digitalisierung und Online-Präsenz.

Seit einigen Jahren gibt es vermehrt Anstrengungen, universitäre Sammlungen aus ihrem Dornröschenschlaf zu wecken und sie zu einer wichtigen Ergänzung objektgebundener Forschung und Lehre weiter zu entwickeln. Dies drückt sich unter anderem in deutschlandweiten Förderprogrammen aus, wie etwa der „Allianz für universitäre Sammlungen" des Bundesministeriums für Bildung und Forschung. Das darin geförderte Projekt „Objekte im Netz" [4] konzentriert sich auf die Digitalisierung universitärer Sammlungen und entwickelt in einer Kooperation zwischen der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) und dem Germanischen Nationalmuseum Nürnberg (GNM) eine gemeinsame Erschließungs- und Digitalisierungsstrategie für die Sammlungen der FAU, um die wissenschaftliche Nutzbarkeit der reichhaltigen Bestände zu verbessern.

Im Fokus stehen jedoch nicht einzelne Sammlungen oder Fachbereiche, sondern die Bereitstellung von Software-Werkzeugen und Lösungswegen, um die digitale Erschließung und Verfügbarkeit an breiter Front voranzutreiben. Die über 20 Sammlungen der FAU bilden dabei eine äußerst heterogene Entwicklungs- und Testlandschaft, um Lösungen zu erarbeiten, die über die FAU hinaus anwendbar sind. Die nötige Generizität der Ansätze und die Nachhaltigkeit sind daher zentrale Herausforderungen, wobei bei letzterem die langfristige Interpretierbarkeit der Daten im Blickpunkt des Projekts steht. Daneben müssen die meist knappen personellen und finanziellen Mittel berücksichtigt werden.

Als besonders geeignet zur Umsetzung der Ziele erscheinen auf technischer Seite Lösungen, die unter freien Lizenzen (Open Source) zur Verfügung stehen und die Ideen des Semantic Web implementieren: Flexible Wissensnetze mit klar definierter Semantik, die weltweit – und damit auch sammlungsübergreifend – verknüpft werden können. Das Projekt erweitert daher die virtuelle Forschungs- und Dokumentationsumgebung WissKI [5] zu einem Werkzeug für die digitale Sammlungserschließung und stellt auf verschiedene Sammlungsbereiche abgestimmte Konfigurationen der Software sowie Leitfäden zur Verfügung. Für die standardisierte semantische Auszeichnung der Daten kommt das CIDOC Conceptual Reference Model (CRM) zum Einsatz.

Aufgrund einer erfolgreichen Pilotstudie kann das Projekt bereits auf erste Ergebnisse verweisen. Im Rahmen des Workshops wird die bereits publizierte generische Konfiguration vorgestellt und von den Teilnehmern angewandt.

Die Werkzeuge - WissKI und CIDOC CRM

Semantische Technologien - im Speziellen Semantic Web und Linked Open Data - erfreuen sich zunehmender Beliebtheit in den Digital Humanities. Für objektbasierte Forschung bieten die flexible, netzwerkartige Grundstruktur des Resource Description Framework (RDF) und darauf aufbauende Formate ein adäquates Mittel zur Repräsentation, Verwaltung und Publikation von (Meta-)Daten. Zahlreiche VFUs unterstützen das Erstellen von komplexen Wissensnetzen und deren Export in Tripelformaten. Wichtige Normdateien und Thesauri stehen als Linked Open Data zur Verfügung. Ontologien wie das CIDOC CRM bilden das semantische Rückgrat dieses Ansatzes und garantieren ein Mindestmaß an Interoperabilität und Datenaustausch, das über das klassische Verlinken von Web-Dokumenten hinausgeht.

Wenngleich die Nutzung semantischer Technologien zunimmt, stellt der praktische Einsatz unerfahrene oder wenig technikaffine Nutzer meist vor große Herausforderungen. Dies gilt weniger für die Beherrschung bestimmter Formate und Werkzeuge als vielmehr für die semantische Modellierung der Daten, d.h. die Erstellung von und den richtigen Umgang mit Ontologien. Da hierbei die Bedeutung der Daten formalisiert niedergelegt wird, ist mitunter ein gehöriges Maß an Wissen über einen Anwendungs-/Fachbereich erforderlich, um Modellierungsfehler zu vermeiden und so eine spätere korrekte Interpretation zu gewährleisten. Insbesondere das CIDOC CRM [6], das eine Top-Level-Ontologie für die Dokumentation kulturellen Erbes darstellt, steht immer wieder in der Kritik, für Einsteiger zu komplex zu sein.

Die virtuelle Forschungs- und Dokumentationsumgebung WissKI nimmt sich dieser Herausforderung an. Die browserbasierte Software ist das Produkt aus zwei DFG-geförderten Projekten und entstand aus Anforderungen an die kooperative Forschung in Museen bzw. im Bereich des Kul-

turerbes und seiner Dokumentation im digitalen Medium. Zentraler Fokus von WissKI ist das vernetzte Arbeiten auf Basis semantischer Tiefenerschließung von Forschungsdaten. Eine Schlüsselrolle kommt hierbei dem CIDOC CRM zu, das um projektspezifische Anwendungsontologien erweitert werden kann.

Aus Nutzersicht ist das System an die tradierten Formen der Datenakquise und -präsentation angelehnt. Die Daten werden jedoch semantisch aufbereitet und nativ als RDF mitsamt Ontologie-Konstrukten gespeichert. Dem Nutzer werden so die Vorteile von Linked Open Data und Semantic Web zugänglich, ohne dass dieser sich mit technischen und ontologischen Details auseinandersetzen muss. Kern dieses Ansatzes ist eine Abbildung zwischen den tradierten, meist datensatz-basierten, tabellarischen Darstellungen und der graphbasierten Wissensrepräsentation, die die ontologiegestützte, formale Semantik der verwendeten Datenfelder beinhaltet. Diese Abbildung wird von einem inhaltlichen Administrator festgelegt und ist für die Nutzer standardmäßig nicht sichtbar. Die formale Semantik muss also nicht verstanden werden, um das System effektiv zu nutzen. Abbildungen oder Teile davon können zwischen verschiedenen Systemen wiederverwendet und erweitert werden, so dass sich Best-Practice-Modellierungen herausbilden.

Die Open-Source-Lizenzierung aller in diesem Workshop verwendeten Werkzeuge und Standards ist ein wichtiger Aspekt. Die kostenfreie Nutzung trägt zum einen der häufig angespannten finanziellen Situation universitärer Sammlungen Rechnung und ist zum anderen Bestandteil des partizipativen Konzepts: Anwender können die Materialien nutzen, sie an ihre Bedürfnisse anpassen und wiederum der Community zur Verfügung stellen.

Zielgruppe sowie Inhalt und Ziele des Workshops

Der Workshop richtet sich an alle, die mit Sammlungsobjekten oder mit Objekten des kulturellen Erbes im Allgemeinen arbeiten und diese digital dokumentieren oder erschließen. Auch spricht der Workshop interessierte Wissenschaftler an, die Objekte standardisiert dokumentieren und ihre Metadaten semantisch aufbereiten möchten. Es werden von den Teilnehmern keine Vorkenntnisse für die VFU WissKI oder das CIDOC CRM vorausgesetzt.

Der Workshop zeigt anhand praktischer Beispiele, wie Erfassungsschemata und -modi aus der universitären Sammlungslandschaft mithilfe der Referenzontologie CIDOC CRM und der VFU WissKI auf Objekte universitärer Sammlungen bzw. des kulturellen Erbes im Allgemeinen umgesetzt werden können.

Während des Workshops arbeiten die Teilnehmer mit ihrem eigenen WissKI-System, wahlweise einzeln oder in Kleingruppen. Dabei stehen weniger die informationstechnischen Details der Werkzeuge im Vordergrund. Vielmehr werden die nötigen Schritte bis zum effektiv einsetzbaren System vermittelt und durchgeführt. Angefangen bei der Installation und einigen grundlegenden Funktionalitäten, binden die Teilnehmer die vom Projekt „Objekte im Netz" angebotene Konfiguration zur Sammlungserschließung in WissKI ein und erhalten somit ein einsetzbares System mit standardisierten Eingabe- und Anzeigemöglichkeiten. Darauf aufbauend werden Möglichkeiten der einfachen Anpassung der semantischen Modellierung aufgezeigt und selbständig geübt. Das Erfassen von (selbst mitgebrachten) Datensätzen rundet die praktische Einführung ab.

Neben einer allgemeinen Einführung in das Arbeiten mit WissKI und der semantischen Dokumentation von Daten sind die Teilnehmer nach dem Workshop in der Lage, einfache Erfassungsmasken zu modellieren, Daten mit WissKI zu erfassen und zu recherchieren.

Kurzbiographien

Martin Scholz ist einer der Hauptentwickler der Virtuellen Forschungsumgebung WissKI. Er studierte Informatik und Sinologie an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Nach seinem Diplom 2008 arbeitete er für die Arbeitsgemeinschaft Digital Humanities der FAU für das DFG-geförderte Projekt "Wissenschaftliche KommunikationsInfrastruktur" (WissKI). Seit 2017 engagiert er sich für die Digitalisierung der Sammlungen der FAU im Rahmen des BMBF-geförderten Projekts "Objekte im Netz". Seine Forschungsinteressen liegen in den Digital Humanities, insbesondere in den Bereichen Wissensrepräsentation, Semantic Web und Verarbeitung natürlicher Sprache.

Martin Scholz
Friedrich-Alexander-Universität
Erlangen-Nürnberg
Referat H2 – Zentralkustodie
Hugenottenplatz 1a, 91054 Erlangen
martin.scholz@fau.de

Sarah Wagner ist Kunsthistorikerin und arbeitet seit 2012 in der Abteilung für Kulturinformatik am Germanischen Nationalmuseum Nürnberg. Sie studierte Kunstgeschichte und Museumsarbeit in Bamberg, Erlangen und Leiden (NL) und

betreut seit 2014 verschiedene Forschungsprojekte, die mit WissKI arbeiten. Aktuell ist sie für das BMBF-geförderte Kooperationsprojekt “Objekte im Netz” tätig und vertritt dort die Seite des Museums. Ihre Forschungsschwerpunkte liegen in der frühneuzeitlichen Sammlungspraxis und der semantischen Wissensmodellierung.

Sarah Wagner
Germanisches Nationalmuseum Nürnberg
Kornmarkt 1, 90402 Nürnberg
s.wagner@gnm.de

Fußnoten

1. Vgl. „Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen”, Verfügbar unter: https://www.wissenschaftsrat.de/download/archiv/10464-11.pdf) [letzter Zugriff: 10. Januar 2018]
2. Kennzahlen zu den folgenden Aussagen sind verfügbar unter: https://portal.wissenschaftliche-sammlungen.de/kennzahlen / [letzter Zugriff: 10. Januar 2018]
3. siehe Fußnote 1
4. Das Projekt wird vom Bundesministerium für Bildung und Forschung von 2017 bis 2020 im Rahmen der Förderlinie „Vernetzen - Erschließen - Forschen. Allianz für universitäre Sammlungen” gefördert. Mehr Informationen unter URL: http://objekte-im-netz.fau.de/ [letzter Zugriff: 10. Januar 2018]
5. WissKI (=„Wissenschaftliche Kommunikations-Infrastruktur, URL: [letzter Zugriff: 10. Januar 2018]) basiert auf dem Open-Source Content Management System Drupal (URL: [letzter Zugriff: 10. Januar 2018]) und wurde in Zusammenarbeit zwischen dem Germanischen Nationalmuseum, Nürnberg, dem Zoologischen Forschungsmuseum Alexander Koenig, Bonn und der Friedrich-Alexander- Universität Erlangen-Nürnberg entwickelt.
6. Das CIDOC CRM wurde vom International Committee for Documentation als Teil des International Council of Museums (ICOM) als formale Referenzontologie erarbeitet und ist seit 2006 als ISO Norm (ISO 21127) anerkannt. In der „Erlangen CRM“ (URL: http://erlangen-crm.org/ [letzter Zugriff: 10. Januar 2018]) auf Basis der Web Ontology Language (OWL) liegt eine maschinenlesbare Version vor. Weitere Informationen zum CIDOC CRM unter URL: http://cidoc-crm.org/ [letzter Zugriff: 10. Januar 2018].

Bibliographie

Definition of the CIDOC Conceptual Reference Model: Version 5.0.4., autor. durch die CIDOC CMR Special Interest Group (SIG), 2011. http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf) [letzter Zugriff 25.09.2017].

Görz, Günther: „WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung” in: Kunstgeschichte, Open Peer Reviewed Journal, urn:nbn:de:bvb:355-kuge-167-7 [letzter Zugriff 10.01.2018].

Hohmann, Georg (2011): „Die Anwendung von Ontologien zur Wissensrepräsentation und -kommunikation im Bereich des Kulturellen Erbes” in: Schomburg, Silke u.a. (eds.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Köln: Hochschulbibliothekszentrum NRW 33-39.

Hohmann, Georg / Schiemann, Bernhard (2013): „An Ontology-Based Communication System for Cultural Heritage. Approach and Progress of the WissKI Project” in: Hans Bock u.a. (eds.): Scientific Computing and Cultural Heritage. Berlin: Springer 127-135.

Hohmann, Georg/Fichtner, Mark. Chancen und Herausforderungen in der praktischen Anwendung von Ontologien für das Kulturerbe. In: Digitales Kulturerbe. Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis. Vol. Kulturelle Überlieferung – digital. Karlsruhe 2015. S. 115-128.

Wissenschaftsrat (2011): *Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen*. Berlin https://www.wissenschaftsrat.de/download/archiv/10464-11.pdf) [letzter Zugriff 25.09.2017].

Embedded Humanities

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

Kestemont, Mike
mike.kestemont@gmail.com
Universität Antwerpen, Belgien

The Use of Distributional Models in the Digital Humanities

In recent years, the Digital Humanities have witnessed the steadily growing popularity of models from distributional semantics which can be used to model the meaning of documents and words in large digital text collections. Well-known examples of influential distributional models include Latent Dirichlet Allocation for topic modelling (Blei et al.) or Word2vec for estimating word vectors (Mikolov et al. 2013). Such distributional models have recently gained much prominence in the fields of Natural Language Processing and, more recently, Deep Representation Learning (Manning 2016). Humanities data is typically sparse and distributional models help scholars obtain smoother estimations of them. Whereas, for instance, words are conventionally encoded as binary 'one-hot vectors' in digital text analysis, embedding techniques from distributional semantics allow scholars to obtain dense, yet rich representations of vocabularies. These embedded representations are known to capture all sorts of valuable relationships between data points, although embedding techniques are typically trained using unsupervised objectives and require relatively little parameter tuning from scholars. Inspiring applications of this emergent technology in DH have ranged from more technical work in cultural studies at large (Bamman et al. 2014), case studies in literary history (Mimno 2012; Schoech 2017) or valuable DH-oriented web apps, such as ShiCo (Martinez-Ortiz et al. 2016). The availability of high-quality implementations in the public domain, in software suites as gensim, word2vec, or mallet etc. has greatly added these methods' popularity.

In spite of their huge potential for Digital Humanities, multiple aspects of their application still remain untapped. Unsupervised models such as Word2vec, for instance, are notoriously hard to evaluate directly – often researchers have to resort to indirect evaluations in this respect. This renders it intriguing to which extent the output of distributional models should play a decisive role in hermeneutical debates or controversies in the Humanities. With other techniques for Distant Reading, distributional models moreover share the drawback that they typically only yield a *single reading* for a particular corpus so that for example the polysemy of a word isn't rendered adequately. Interesting progress into representing the complex variability of meaning has been achieved, for example on the level of diachronic word embeddings, where convincing attempts have been made to allow for semantic shifts in an individual word's meaning (Hamilton et al. 2016). Likewise, critical studies have revealed how tightly distributional models reproduce cultural biases with respect to gender and race (Bolukbasi et al. 2016), which calls for a debate about the ethical aspects of the matter. Likewise, it deserves emphasis how distributional models depend on large datasets and typically yield poor estimates for more restrictive data collections. This might help explain why word embeddings so far have not that many applications in fields like stylometry, that mostly work with relatively small corpora.

The DARIAH working group on Text and Data Analytics (@dariahtdawg), in collaboration with the FWO-sponsored scientific community Digital Humanities Flanders (DHuF) proposes to collocate a one-day workshop with the 2018 DHd conference in Cologne. The workshop aims to bring together ca. 10-12 practitioners from the Digital Humanities to present and discuss recent advances in the field, through 30-minute presentations on focused case studies, including work-in-progress or theoretical contributions. Additionally, the workshop aims to reach an audience of non-presenting participants who take an active interest in distributional models and who are planning to apply distributional models to their own data in the near future. We aim to bring together a diverse group of both junior and senior stakeholders in this nascent subfield of DH. The goal of the workshop is to identify the state of the art in the field, identify common challenges and share recommendations for a best practice. Special attention will be given to the (both hermeneutic and quantitative) evaluation of distributional models in the context of Humanities research, which remains a challenging issue. The workshop is open to scholars from all backgrounds with an interest in semantic representation learning and encourages submissions that deal with under-researched resource-scarce and/or historic languages. Abstracts (between 250 and 300 words, not including references) can be submitted to mike.kestemont@uantwerp.be. The workshop also explicitly welcomes submissions presenting previously published research which is of interest to the DH community (although this work should not overlap strongly with work presented at the main conference).

Topics which seem of special interest to the DH community nowadays include, but are not limited to:

- the general use of distributional semantics in DH (such as topic modelling and word embed-

dings), but also more specific case studies, including work in progress;
- the diachronic study of cultural phenomena via distributed methods;
- the evaluation of distributional models, both from an empiric and hermeneutic perspective;
- modelling the role and behaviour of (individual) readers or reading communities;
- …

In terms of technical requirements, the workshop would need a beamer to project from a laptop.

As keynote speaker, we have found David Bamman (University of California, Berkeley) willing to join our workshop and give a plenary lecture. Bamman is an authority in the field and will certainly increase the attractiveness of the workshop to potential participants.

Convenors

Fotis Jannidis (University of Würzburg, Germany)

Fotis.jannidis@uni-wuerzburg.de, www.jannidis.de

Fotis holds the chair for literary computing in the department of German studies at the University of Würzburg. In the last years, the main focus of his work is the computational analysis of larger collections of literature, especially narrative texts. He is interested in developing new research methods for this new subfield of literary studies, but also in new applications for established methods and also in a better understanding, why successful algorithms in this field work.

Mike Kestemont (University of Antwerp, Belgium)

mike.kestemont@uantwerp.be,
www.mike-kestemont.org

Mike is a tenure track research professor in the department of Literature the University of Antwerp. He specializes in computational text analysis for the Humanities, in particular stylometry or computational stylistics. He has published on the topic of authorship attribution in various fields, such as Classics or medieval European literature. Mike actively engages in the debate surrounding the Digital Humanities and attempts to merge methods from Artificial Intelligence with traditional scholarship in the Humanities. He recently took up an interest in so-called 'deep' representation learning using neural networks.

Bibliographie

Bamman, D./ Underwood, T./ Smith, N. A. (2014): „A Bayesian Mixed Effects Model of Literary Character“, in *Proceedings of ACL*.

Blei, David (2012): „Probabilistic Topic Models“, in *Communications of the ACM* 55, 77-84.

Bolukbasi, T./ Chang, K. W./ Zou, J./ Saligrama, V./ Kalai, A. (2016): „Quantifying and Reducing Stereotypes in Word Embeddings“, in *arXiv:1606.06121*.

Chang, J./ Gerrish, S./ Sang , C./ Boyd-Graber, J. L./ Blei, D. M. (2009): „Reading Tea Leaves: How Humans Interpret Topic Models“, in Proceedings of NIPS, 288-296.

Hamilton, W. L./ Leskovec, J./ Jurafsky, D. (2016): „Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change“, in *Proceedings of ACL*.

Manning, Christopher D. (2016): „Computational Linguistics and Deep Learning“, in *Computational Linguistics* 41, 701-707.

Martinez-Ortiz, C./ Huijnen, P./ Kenter, T./ Verheul, J./ Wevers, M./ van Eijnatten, J. (2016): „ShiCo: A Visualization Tool for Shifting Concepts Through Time“, in *Digital Humanities Benelux: Book af Abstracts* , s.p.

Mikolov, T./ Sutskever, I./ Chen, K./ Corrado, G. S./ Dean, J. (2013): „Distributed representations of words and phrases and their compositionality“, in *Proceedings of NIPS 2013*.

Mimno, David (2012): „Computational Historiography: Data Mining in a Century of Classics Journals“, in *ACM Journal of Computing in Cultural Heritage* 5.

Schoech, Christof (2017): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“, *Digital Humanities Quarterly* 11, s.p.

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

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Ketschik, Nora
nora.ketschik@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universtität Stuttgart, Deutschland

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

Schulz, Sarah
sarah.schulz@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Einleitung

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

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

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

Entitätenreferenzen

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

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

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

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

Textkorpus

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

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

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

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

Ablauf

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

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

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

Wir verzichten bewusst auf eine graphische Benutzerschnittstelle (cf. 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 Teilnehmerinnen und Teilnehmer bekommen einen zusammenhängenden Überblick von der manuellen Annotation ausgewählter Texte über die Feinjustierung der Lernverfahren bis zur Evaluation der Ergebnisse. Die vorgestellte Vorgehensweise für den gesamten Ablauf ist grundsätzlich auf ähnliche Projekte übertragbar.

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

Zeitplan

- Im Vorfeld der Veranstaltung: Installationsanweisungen und Online-Support

Dauer in Minuten (ca.)

- 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 & Preisverleihung

Beitragende (Kontaktdaten und Forschungsinteressen)

Der Workshop wird ausgerichtet von Mitarbeiterinnen und Mitarbeitern 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.

Nils Reiter

Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Die Forschungsinteressen von Nils Reiter liegen generell in der Anwendung computerlinguistischer Methoden auf Fragen aus den Geistes- und Sozialwissenschaften. Insbesondere die Operationalisierung literarischer Forschungsfragen und die adäquate Interpretation von Ergebnissen ist dabei ein Schwerpunkt, neben der regelgeleiteten Annotation und damit zusammenhängenden Fragen.

Nora Ketschik

Institut für Literaturwissenschaft
Keplerstraße 17

70174 Stuttgart
Nora Ketschik ist Promotionsstudentin in der Abteilung für Germanistische Mediävistik. Im Rahmen des CRETA-Projekts nimmt sie Analysen narratologischer Kategorien (u.a. Figur, Raum) an ausgewählten mittelhochdeutschen Romanen vor und setzt sich dabei mit der Verwendung computergestützter Methoden für literaturwissenschaftliche Analysezwecke auseinander.

Gerhard Kremer

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.

Sarah Schulz

Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart
Sarah Schulz beschäftigt sich überwiegend mit der automatischen Verarbeitung von Texten, die syntaktischen oder lexikalischen Eigenschaften aufweisen und damit vom *Standard* abweichen. Sie hat einen Hintergrund in sowohl Computerlinguistik als auch Theater-und Medienwissenschaften und Germanistik.

Zahl der möglichen Teilnehmerinnen und Teilnehmer

Zwischen 15 und 25.

Benötigte technische Ausstattung

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

Fußnoten

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

Bibliographie

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

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

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

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

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

Modellathon „Digitale 3D-Rekonstruktion"

Münster, Sander
sander.muenster@tu-dresden.de
Medienzentrum/TU Dresden, Deutschland

Christen, Jonas
jonas.christen@zhdk.ch
Zürcher Hochschule der Künste, Schweiz

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

Die Arbeitsgruppe „Digitale Rekonstruktion“ ging aus der 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (25.-28.03.2014, Universität Passau) hervor. Die Arbeitsgruppe versammelt Kolleginnen und Kollegen, die sich dem Thema digitale Rekonstruktion aus dem Blickwinkel der Architektur, Archäologie, Bau- und Kunstgeschichte sowie Computergraphik und Informatik verschrieben haben. Die Arbeitsgruppe bietet eine Plattform für einen Austausch und eine feste Etablierung der digitalen Rekonstruktion im Dienste einer Erfassung, Erforschung und Vermittlung kultureller und geschichtlicher Inhalte innerhalb der Digital Humanities. Vorrangiges Ziel der Arbeitsgruppe ist es, die Akteure im deutschsprachigen Raum zusammenzubringen, um sich den Fragen der Begriffsklärung und der Arbeitsmethodik sowie der Dokumentation und Langzeitarchivierung von digitalen Rekonstruktionsprojekten zu widmen. Der Arbeitsgemeinschaft gehören ca. 50 Personen aus 29 Einrichtungen im deutschsprachigen Raum sowie 4 assoziierten Mitgliedseinrichtungen im europäischen Raum an. In bisher 4 Beiträgen wurden bei vergangenen DHd-Jahrestagungen durch die Arbeitsgemeinschaft „Allgemeine Standards, Methodik und Dokumentation“ (Kuroczyński et al., 2014) und „aktuelle Herausforderungen“ (Kuroczyński et al., 2015) sowie Transformationsprozesse „vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell“ (Kuroczyński et al., 2016) und nicht zuletzt aktuelle Projekte und Forschungsschwerpunkte (Münster et al., 2017) beleuchtet. Während damit vor allem theoretische Perspektiven digitaler Rekonstruktion als Forschungs- und Vermittlungsmethode aufgezeigt wurden, soll auf der DHd 2018 im Rahmen eines an das Format des Hackathons (Wikipedia, 2017) angelehnten Modellathons praktisches Handwerkszeug digitaler 3D-Rekonstruktion in wissenschaftlichen Kontexten vermittelt und erprobt werden. Der Modellathon beinhaltet einen ½ tägigen Workshop zur Vermittlung von Grundlagen der 3D-Rekonstruktionen und deren filmischer oder bildlicher Inszenierung. Daran anschliessend sollen im Rahmen eines die DHd-Jahrestagung begleitenden Wettbewerbs durch studentische Arbeitsteams 3D-Rekonstruktionen eines noch zu benennenden Objekts erstellt werden. Die Ergebnisse werden im abschliessenden DHd-Plenum vorgestellt und durch das Publikum sowie eine Expertenjury bewertet sowie prämiert.

Der Modellathon

Workshop

Der Modellathon besteht aus einem ½ tägigen, der DHd-Jahrestagung (26.2.-2.3.2018) vorgelagerten Workshop, bei welchem sowohl den studentischen Arbeitsteams als auch interessierten Besuchern eine Einführung in den Gegenstand sowie Grundlagen wissenschaftlicher 3D-Rekonstruktion geboten werden. Hierfür sind folgende Inhalte vorgesehen, welche durch Mitglieder der AG vermittelt werden:

Zeitplanung	Inhalt	Personen
14.00-15.30 Uhr (90 min.)	Grundlagen wissenschaftlicher 3D-Rekonstruktion: • Wissenschaftliches Arbeiten und Forschungseinbettung • Quellengestützte Modellierung	Sander Münster
16.00-17.30 Uhr (90 min.)	"Digitale 3D-Rekonstruktion richtig skaliert – Drei Beispiele mit unterschiedlichem Detailgrad" „Tipps um die Rekonstruktion in Photoshop in Szene zu setzen“	Jonas Christen

Wettbewerb

An den Workshop anschließend wird ein die DHd-Jahrestagung begleitender Wettbewerb zur 3D-Rekonstruktion eines zum Workshop-Auftakt zu benennenden historischen Objekts durchgeführt. Zur individuellen Bearbeitung der Projekte durch die studentischen Teams steht während der gesamten Jahrestagung ein durch Mitglieder der AG „Digitale Rekonstruktion“ betreuter Arbeitsraum zur Verfügung, welcher während der Keynotes geschlossen wird. Die individuelle Arbeit am Projekt wird durch Kurztutorien von jeweils ca. 15-20 min. Dauer begleitet, in welchen praxisrelevante Techniken bspw. zu Interdisziplinarität, Modellierungspraktiken, interaktiven Web-Modellen u. ä. vermittelt werden. Ergebnisse

des Wettbewerbs können Bilder, Filme oder interaktive Präsentationen der durch die studentischen Teams erstellten 3D-rekonstruierten Modelle sein. Diese werden durch die Teams im DHd-Plenum in Form eines „Elevator-Pitchs“ (je Team max. 2 Minuten Vorstellung + 2 Minuten für Fragen) vorgestellt und durch das Publikum sowie eine Expertenjury anhand der Kriterien Wissenschaftlichkeit, Inszenierung, Qualität und Vermittlung bewertet.

Zeitplanung	Inhalt	Personen
27.2. 9.00-10.30 Uhr (90 min.)	Start des Modellathons: • Vorstellung der Aufgabenstellung und der Bewertungskriterien • Gegenstand & Material	Wird zum Wettbewerbsstart bekanntgegeben.
26.2.-2.3.2018	Arbeit am Projekt Begleitende Tutorien bspw. zu Interdisziplinarität, Modellierungspraktiken, Präsentationstechniken etc.	
DHd-Plenum	Vorstellung der Arbeit in Form eines Elevator-Pitch (max. 2 Minuten Vorstellung + 2 Minuten für Fragen) im Plenum Bewertung durch (a) Publikum und (b) Expertenjury	

Rahmenbedingungen

Preise

Als Sachpreis wird eine Maxon Cinema 4D Studio Lizenz (Wert ca. 3000 EUR, gestiftet durch die Maxon Computer GmbH) vergeben.

Teilnahmevoraussetzungen

Workshop
Eine Teilnahme am Workshop steht allen Interessierten offen. Teilnahmevoraussetzungen sind ein eigener Laptop mit einem gängigen 3D-Modellierungspaket (bspw. Autodesk 3DStudio Max oder Maya, Maxon Cinema 4D, Blender, Mc Neel Rhino) sowie Grundkenntnisse der 3D-Modellierung. Teilnehmer des Workshops erhalten einen Überblick zu Problemstellungen und Ansätzen wissenschaftlicher 3D-Rekonstruktion sowie Visualisierung und können diese anhand einer Beispielaufgabe praktisch erproben.

Wettbewerb
Der darüber hinaus stattfindende Wettbewerb steht studentischen Arbeitsteams sowie Einzelstudenten offen. Bis zum 10. Dezember konnten sich Studierende von außerhalb Kölns für eine durch den DHd-Vorstand gestiftete Förderung der Teilnahme an Workshop und Wettbewerb in Höhe von 250 EUR je Person bewerben. An 10 Bewerber wurden dabei 10 Stipendien vergeben, weitere 3 Studierende sind ohne Förderung zum Wettbewerb eingeladen.

Zeitplanung

30. September 2017	Veröffentlichung eines an die Teilnehmer gerichteten Calls
10. Dezember 2017	Ende der Bewerbungsfrist für die geförderte Teilnehmer
20. Dezember 2017	Bekanntgabe der nominierten 10 geförderten Teilnehmer
26. Februar 2018	½ tägiger Workshop zur DHd-Jahrestagung, anschliessend konferenzbegleitender Wettbewerb
2. März 2018	Präsentation und Prämierung der Ergebnisse im DHd-Plenum

Jury

Die Jury setzt sich aus Mitgliedern des DHd-Vorstands, der AGDR sowie ggf. Vertretern der noch zu benennenden inhaltlichen Schirmherren zusammen. Aufgaben der Jury sind (a) die Vergabe der Teilnahmeförderungen sowie (b) die Bewertung der im DHd-Plenum vorgestellten Wettbewerbsergebnisse aus Expertensicht. Die Jury setzt sich aus Vertretern des DHd-Vorstandes, der AG Digitale Rekonstruktion sowie der inhaltlichen Schirmherren zusammen.

Ausrichter

Dr. Sander Münster
Technische Universität Dresden
Media Center
D-01062 Dresden
http://mz.tu-dresden.de
Phone: +49-(0)351-463-32530
Fax: +49- (0) 351 463-35606
eMail: sander.muenster@tu-dresden.de

Sander Münster studierte Geschichts-, Erziehungs- sowie Wirtschaftswissenschaften an der TU Dresden und promovierte 2014 im Bereich Bildungstechnologie zum Thema „Interdisziplinäre Kooperation bei der Erstellung virtueller geschichtswissenschaftlicher 3D-Rekonstruktionen“ und wurde 2016 an der TU Dresden zum Young Investigator ernannt. Seine Forschungsschwerpunkte stellen interdisziplinäre Kooperation sowie Teamwork und Arbeitsabläufe innerhalb von 3D-Rekonstruktionsprojekten dar. Er leitet den Bereich Mediendesign und -produktion am Medienzentrum der TU Dresden und ist seit 2015 Koordinator der vom BMBF geförderten eHumanities Nachwuchsforschergruppe HistStadt4D. Daneben arbeitet er seit mehr als einem Jahrzehnt im Bereich 3D-Grafik mit Schwerpunkt wissenschaftliche Visualisierung und war in dieser Funktion an zahlreichen 3D-Rekonstruktionsprojekten historischer Bauwerke beteiligt.

Jonas Christen
Pfingstweidstrasse 96, Postfach
CH-8031 Zürich
Telefon +41 43 446 32 80
Mobile +41 78 717 83 85 (Di, Fr)

Jonas Christen ist wissenschaftlicher Mitarbeiter an der Zürcher Hochschule der Künste im Bereich Design. Er gehört der Forschungsgruppe Knowledge Visualization an und ist langjähriges Mitglied der AG Digitale Rekonstruktion.

Dr.-Ing. Mieke Pfarr-Harfst
Technische Universität Darmstadt
Fachbereich Architektur
DDU - Digital Design Unit - Digitales Gestalten
El-Lissitzky-Str.1
64287 Darmstadt
fon: +49 6151 16-22482 | fax: +49 6151 16-22480
mail: pfarr@dg.tu-darmstadt.de | http://www.dg.tu-darmstadt.de

Mieke Pfarr-Harfst leitet derzeit den Forschungsschwerpunkt "Digitale Rekonstruktionen" am Fachgebiet Digitales Gestalten an der TU Darmstadt. In ihrer Promotion „Dokumentationssystem für Digitale Rekonstruktionen“ und ihrer aktuellen Forschungsarbeit setzt sich Frau Pfarr-Harfst mit den digitalen dreidimensionalen Gebäudemodellen als innovative Methode zur Erforschung des kulturellen Erbes auseinander.

Erwartete Teilnehmer

Ca. 15

Benötigte Ausstattung

Für Workshop und Wettbewerb wird um die Bereitstellung eines (idealerweise abschließbaren) Raumes mit ca. 15-20 Plätzen für die gesamte DHd-Konferenzdauer gebeten. Benötigt werden ein Datenprojektor und ca. 15 Steckdosenplätze.

Bibliographie

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

Kuroczyński, P./ Hauck, O./ Hoppe, S./ Münster, S. / Pfarr-Harfst, M. (2016): Der Modelle Tugend 2.0 – Vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell. (Panel). 3. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2016). Leipzig.

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

Münster, S./ Kuroczyński, P. / Pfarr-Harfst, M. (2017): Projekte und Aktivitäten im Kontext digitaler 3D-Rekonstruktion im deutschsprachigen Raum. 4. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2017). Bern.

Wikipedia (2017): Hackathon [Online]. Available: https://de.wikipedia.org/wiki/Hackathon, 10.9.2017.

Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Deutschland

Rau, Felix
f.rau@uni-koeln.de
Universität zu Köln, Deutschland

Beschreibung

Digitale Spracharchive sind ein integraler Bestandteil der Forschungsdateninfrastruktur und haben den spezifischen Auftrag audiovisuelle Sprachdaten und Dokumente zu sichern und auf deren Basis Wissensgenerierung zu ermöglichen und zu unterstützen. Ein Spracharchiv ist in diesem Sinn eine Plattform, die zwischen Produzenten und Konsumenten von Primärdaten vermittelt, so dass diese direkt oder indirekt interagieren können. Den datenproduzierenden Forschern ermöglicht das Archiv, Audio- und Videoaufnahmen menschlicher Kommunikation zu archivieren und idealerweise web-basiert zugänglich zu machen. Auf der anderen Seite werden Forscher, Sprachgemeinschaften und die weitere Öffentlichkeit in die Lage versetzt, diese Daten aufzufinden, zu betrachten, herunterzuladen und weiterzuverwenden und auf dieser Grundlage neues Wissen zu generieren. Um diesen Austausch zu unterstützen, haben die verschiedenen Spracharchive komplexe Webplattformen entwickelt.

Wie alle Forschungsdatenarchive profitieren Repositorien für audiovisuelle Daten von einem voranschreitenden Standardisierungsprozess. Das OAIS Referenzmodell hat die Grundlage für ein gemeinsames Beschreibungsvokabular gelegt. Das Data Seal of Approval/CoreTrustSeal entwickelt sich zum de-facto Standard für die Zertifizierung von Forschungsdatenrepositorien in den digitalen Geisteswissenschaften. Auf dem Gebiet der Sprachressourcen wirken die Infrastrukturinitiative CLARIN und der vom BMBF geförderte deutsche Partner CLARIN-D als starke integrative Kraft und haben mit der Etablierung von Standards, die in allen Aspekten des Datenlebenszyklus zur Anwendung kommen, gemeinsame Lösungen geschaffen. Forschungsdatenrepositorien für Sprachdaten sind herausgefordert, gültige Standards zu implementieren und gleichzeitig attraktive Dienste anzubieten, die die Spezifika der Datentypen und die Bedürfnisse der jeweiligen Nutzergruppen berücksichtigen, um eine erfolgreiche Nachnutzung von Forschungsdaten zu befördern.

Der Workshop soll Archivbetreibern, Datenkuratoren, Datenproduzenten und Datenkonsumenten die Möglichkeit zum Austausch über Angebote, Bedarfe und zentrale Weiterentwicklungen geben. Der Workshop richtet sich an Betreiber von Forschungsdatenrepositorien mit audiovisuellen (Sprach-)Daten sowie Mitarbeiter von Institutionen, die Forschungsdatenmanagement für Forschung mit audiovisuellen Daten anbieten. Ebenso sind auch WissenschaftlerInnen als aktive oder potentielle Nutzer von Forschungsdatenrepositorien angesprochen.

Ablauf

Der Workshop wird eingeleitet durch eine 15-minütige Einführung in den Themenkomplex und eine Vorstellung der Beitragenden aus verschiedenen Institutionen, die sich mit den Problemen und Lösungen rund um die Archivierung und Bereitstellung von AV-Daten auseinandersetzen.

Der Hauptteil des Workshops gliedert sich in drei Sektionen, die unterschiedliche Aspekte der Nutzerunterstützung im Datenlebenszyklus behandeln. In jeder Sektion wird es zwei aufeinanderfolgende 15-minütige Impulsvorträge geben, die diesen Themenkomplex jeweils aus der Sicht eines der beitragenden Forschungsdatenrepositorien beleuchten und Antworten darauf geben, mit welchen Maßnahmen, Policies oder technischen Implementierungen sie den jeweiligen Aspekt adressieren und wie sie diese an ihre Nutzerschaft kommunizieren. Jede Sektion endet jeweils in einer 30-minütigen Diskussion.

Der Workshop schließt mit einer kurzen Zusammenfassung und Abschlussdiskussion.

Sektion 1: Wie komme ich an die Daten?

Die Bereitstellung der Daten gehört neben einer nachhaltigen Sicherungsstrategie zu den grundlegendsten Aufgaben eines Datenarchivs. Dennoch ist die Implementierung und Vermittlung der dazugehörigen Prozesse alles andere als tri-

vial: Erschließungsmechanismen müssen je nach Fachdomäne und Zielgruppe unterschiedlichen Anforderungen genügen, die Vergabe von Persitent Identifiern (PIDs) garantieren eine stabile Adressierung und werten die Zitierfähigkeit der Bestände auf, OAI-PMH und andere technische Schnittstellen ermöglichen eine weitergehende Auffindbarkeit in externen Portalen und Integration mit externen Diensten. Schließlich wird in dieser Frage auch der sensible Aspekt der Authentifizierung, Autorisierung und Lizenzierung berührt: Welche Bedingungen sind an den Zugriff der Daten geknüpft? Welche Verwertungsrechte werden dem Nutzer eingeräumt? Welche konkreten Schritte muss ein Nutzer unternehmen, um Zugriff zu erlangen und wie wird der Prozess kommuniziert?

Sektion 2: Wie kommen die Daten ins Archiv?

Archive haben verschiedene Workflows für die Einreichung und Integration neuer Daten. Zum Teil werden vollständig technisch geleitete Verfahren angeboten, die es dem Produzenten ermöglichen, weitgehend autark Daten über das Archiv bereitzustellen („self-archiving"). Andere Institutionen optieren bewusst für eine intensive Begleitung und Prüfung durch einen digitalen Archivar oder Kurator. Ebenso können die Herangehensweisen an die Veränderlichkeit von Datenbeständen weit auseinandergehen. Während manche Archive ausschließlich oder bevorzugt abgeschlossene Datensammlungen integrieren, begünstigen andere in ihren Implementierung eine laufende und teils feingranulare Aktualisierung und Erweiterung von Sammlungen („living archive").

Der Wirkungsbereich eines digitalen Archivs geht idealerweise weit über den eigentlichen Einreichungsprozess hinaus. Qualitätssicherung beginnt mit einer frühen Beratung, Begleitung/Schulung und der Etablierung von Workflows für Arbeitsgruppen, die Daten in einem Repositorium archiven wollen. Datenarchive sind nicht selten auch an der Entwicklung von Software beteiligt, die in der Korpuserstellung und -kuratierung zum Einsatz kommt (z.B. Annotations-Tools, Metadaten-Editoren).

Sektion 3: Was kann ich mit den Daten machen?

Archive können „Mehrwert-Dienste" anbieten, die die Daten in einer fachspezifischen Art und Weise darstellen, kontextualisieren und vernetzen, oder sogar die Auswertung und Weiterverarbeitung unterstützen. In welcher Weise unterstützt das Archiv den wissenschaftlichen Nutzer des Archivs?

Vorträge

Dynamische Forschungsdatenrepositiorien für die Geisteswissenschaften

Der Vortrag präsentiert den Ansatz eines dynamischen Forschungsdatenrepositoriums für die Geisteswissenschaften, der im BMBF-Zentrumsprojekt „Kölner Zentrum Analyse und Archivierung von AV-Daten" (KA^3) implementiert wird. Im Mittelpunkt steht die Vereinbarkeit von Skalierbarkeit und Berücksichtigung fachspezifischer Anforderungen.

Andreas Witt (andreas.witt@uni-koeln.de), Institut für Digital Humanities, Universität zu Köln und *Michael Lönhardt* (loenhardt@uni-koeln.de), Diensteentwicklung, Regionales Rechenzentrum, Universität zu Köln

Eines für alle - Alles für einen

Der Beitrag beleuchtet die besonderen Anforderungen, die an ressourcentyp- bzw. fachspezifische Repositorien gestellt werden. Am Beispiel des HZSK-Repositoriums am Hamburger Zentrum für Sprachkorpora wird das Spannungsfeld zwischen dem Ziel einer möglichst breiten Nutzbarkeit und der Anpassung an spezifische Nutzergruppen diskutiert.

Hanna Hedeland (hanna.hedeland@uni-hamburg.de) Hamburger Zentrum für Sprachkorpora (HZSK), CLARIN-D, Universität Hamburg und *Timm Lehmberg* (timm.lehmberg@uni-hamburg.de), INEL, Institut für Finnougristik/Uralistik, Universität Hamburg

Muss es immer Fedora sein? Die Repositoriumslösung der Sprachbank von Finnland (Kielipankki)

Vorgestellt wird ein alternativer Ansatz der Datenbereitstellug, der ohne Repositoriumssoftware wie Fedora oder DSpace auskommt. Das Zusammenspiel zwischen verschiedenen Zugangsformen, PID-Verwaltung, Metadatenverwaltung, Versionierung und Zugangskontrolle wird erläutert, wobei besonders auf den Download-Service, die PID- und Metadatenverwaltung näher eingegangen wird.

Martin Matthiesen (martin.matthiesen@csc.fi), The Language Bank of Finland, CSC - IT Center for Science

Virtuelles An-die-Hand-nehmen: Qualitätssicherung für linguistische und kulturelle Datensammlungen

ELAR betreut weltweit intensiv Linguisten mit unterschiedlichsten linguistischen und digitalen Vorkenntnissen, um diese in die Lage zu versetzen, digitale Daten selbst zu kurieren, zu archivieren und langfristig für andere nutzbar zu machen.

Vera Ferreira (vf4@soas.ac.uk), *Sophie Salffner* (ss123@soas.ac.uk) und *Mandana Seyfeddinipur* (ms123@soas.ac.uk), Endangered Languages Archive, SOAS University of London

Visualisierung zeitalignierter Audio-Annotationen mit IIIF

Der Vortrag gibt Einblick in ein im Rahmen des Projekts KA³ entwickeltes Softwaresystem zur Visualisierung zeitalignierter Audio-Annotationen. Neben einer Live Demonstration werden die konzeptionellen Anpassungen und technischen Entwicklungen vorgestellt, die nötig waren, um die REST Standardisierungsbemühungen des International Image Interoperability Framework (IIIF Image API, IIIF Presentation API, IIIF Search API) für den Bereich zeitalignierter Audio-Annotationen nutzbar zu machen.

Jochen Graf (jochen.graf@uni-koeln.de), Mitarbeiter im Projekt KA³, Regionales Rechenzentrum, Abteilung Diensteentwicklung, Universität zu Köln

Erschließung audiovisueller Daten im AGD am Beispiel des FOLK-Korpus

Im AGD werden Varietäten- und Gesprächskorpora im Deutschen archiviert und bereitgestellt. Wir konzentrieren uns in diesem Beitrag darauf, wie Audios, Videos und Transkripte durch die DGD nutzbar gemacht werden.

Jan Gorisch (gorisch@ids-mannheim.de) & *Thomas Schmidt* (thomas.schmidt@ids-mannheim.de), Programmbereich Mündliche Korpora, Institut für Deutsche Sprache, Mannheim

Liste der Beitragenden

Dr. Vera Ferreira, *Endangered Languages Archive, SOAS, University of London*

Vera is a trained linguist with a background in language documentation and field research. Her main research interests lie in European endangered languages and in the connection between documentary data and language revitalisation. She is head of CIDLeS (Interdisciplinary Centre for Social and Language Documentation) and digital archivist at Endangered Languages Archive (SOAS University of London). As the digital archivist, Vera provides advice and training on all aspects of data management, metadata preparation and digital archiving.

Dr. Jan Gorisch, *Institut für Deutsche Sprache, Mannheim, Programmbereich Mündliche Korpora*

Jan Gorisch ist Mitarbeiter des Archivs für Gesprochenes Deutsch (AGD) und wirkt bei Datenübernahmen und deren Kuration mit. Er arbeitet an der Automatisierung des Workflows am AGD und entwickelt Tools zur Integration von Sprachtechnologie bei der Erschließung der Sprachdaten.

Jochen Graf, M.A., *Regionales Rechenzentrum, Diensteentwicklung, Universität zu Köln*

Jochen Graf hat einen langjährigen Hintergrund in Digital Humanities Projekten und ist Entwickler im BMBF-Zentrumsprojekt „Kölner Zentrum Analyse und Archivierung von AV-Daten“ (KA³) am Regionalen Rechenzentrum an der Universität zu Köln.

Hanna Hedeland, M.A., *Hamburger Zentrum für Sprachkorpora, CLARIN-D, Universität Hamburg*

Hanna Hedeland koordiniert als Geschäftsführerin des Hamburger Zentrums für Sprachkorpora (HZSK) die Arbeit der dort angesiedelten Infrastrukturprojekte sowie die Kooperationen mit externen Forschungsprojekten. Im Rahmen des Projekts CLARIN beschäftigt sie sich mit der Entwicklung von Standards und Best Practices sowie entsprechenden Technologien und Workflows für die Aufbereitung und Bereitstellung (mehrsprachiger) gesprochener Daten.

Timm Lehmberg, M.A., *Institut für Finnougristik/Uralistik, Langzeitprojekt INEL/Hamburger Zentrum für Sprachkorpora (HZSK), Universität Hamburg*

Timm Lehmberg ist technischer Koordinator des Langzeitvorhabens INEL („Grammatiken, Korpora und Sprachtechnologie für indigene nordeurasische Sprachen“), das im Rahmen des gemeinsam von Bund und Ländern finanzierten Akademieprogramms durchgeführt wird so-

wie Mitwirkender am Hamburger Zentrum für Sprachkorpora (HZSK).

Dipl.-Inf. Michael Lönhardt, *Regionales Rechenzentrum, Diensteentwicklung, Universität zu Köln*

Michael Lönhardt ist Leiter der Diensteentwicklung am Regionalen Rechenzentrum der Universität zu Köln.

Martin Matthiesen, M.A., *Senior Application Specialist, CSC - IT Center for Science, Finnland*

Martin Matthiesen ist Administrator der Sprachbank von Finnland (www.kielipankki.fi/languagebank/) und beschäftigt sich mit der nachhaltigen Bereitstellung von multimodalen Forschungsdaten für die Sprachwissenschaft und angrenzenden Disziplinen. Dies umfasst Versionierung, Metadaten, Persistente Identifikatoren (PIDs) und Zugangsverwaltung für nicht frei zugängliche Daten. Die Sprachbank ist ein gemeinsamer Service von CSC - IT Center for Science (www.csc.fi) und der Universität Helsinki (www.helsinki.fi/en) für FIN-CLARIN (www.kielipankki.fi/organization/).

Dr. Sophie Salffner, *Endangered Languages Archive, SOAS, University of London*

Sophie Salffner ist Digitale Archivarin am Endangered Languages Archive an der SOAS University of London. Sie arbeitet u.a. in der Aus- und Weiterbildung der BenutzerInnen des Archivs und schult WissenschaftlerInnen im wissenschaftlichen Datenmanagement, in der Aufbereitung von Metadaten und im Archivieren in digitalen Archiven.

Dr. Thomas Schmidt, *Institut für Deutsche Sprache, Mannheim, Programmbereich Mündliche Korpora*

Thomas Schmidt leitet das Archiv für Gesprochenes Deutsch (AGD) und den Aufbau des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) am IDS. Mit EXMARaLDA und der Datenbank für Gesprochenes Deutsch (DGD) beschäftigt er sich außerdem mit der Entwicklung von Technologie für die Arbeit mit und den Zugriff auf audiovisuelle Daten gesprochener Sprache.

Dr. Mandana Seyfeddinipur, *Endangered Languages Archive, SOAS University of London*

Mandana Seyfeddinipur ist Leiterin des Endangered Languages Archive (ELAR) an der SOAS University of London.

Prof. Dr. Andreas Witt, *Institut für Digital Humanities/Data Centre for the Humanities, Universität zu Köln*

Andreas Witt ist geschäftsführender Direktor des Instituts für Digital Humanities an der Universität zu Köln. Er beschäftigt sich mit digitalen Forschungsinfrastrukturen für linguistische Ressourcen, insbesondere mit der Standardisierung von Datenformaten und mit ethischen und juristischen Aspekten beim Umgang mit Forschungsdaten.

Ausrichter des Workshops

Der Workshop wird von dem an der Universität zu Köln angesiedelten BMBF-Zentrum „Kölner Zentrum Analyse und Archivierung von AV-Daten" (KA^3) ausgerichtet. Das Zentrum wird am Standort Köln vom Institut für Liguistik (IfL), dem Regionalen Rechenzentrum der Universität zu Köln (RRZK) und dem Data Center for the Humanties (DCH) getragen.

Rechtsfragen in DH-Projekten: Alles, was man wissen muss

Hannesschläger, Vanessa
Vanessa.Hannesschlaeger@oeaw.ac.at
ACDH-ÖAW, Österreichische Akademie der Wissenschaften, Österreich

Kamocki, Pawel
pawel.kamocki@gmail.com
L'Université Paris Descartes, Frankreich

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

Mit dem Eintritt der Geisteswissenschaften in den digitalen Raum öffnet sich für Forschende auch ein neuer Rechtsraum mit Anforderungen und Problemstellungen, die uns bislang nicht oder nur marginal betroffen haben.

Dieser Workshop soll die gängigsten Fragen beantworten, die sich aus unterschiedlichen Rechtsbereichen bei der Realisierung von Digitalisierungsvorhaben und digitalen Forschungsprojekten ergeben. Besonders eingegangen wird auf jüngste Entwicklungen und Neuerungen in den Legislaturen speziell der deutschsprachigen Länder, die für die DH von besonderer Relevanz sind (etwa die *UrhWissG*-Novelle in Deutschland oder die *EU-Datenschutz-Grundverordnung*, die jeweils 2018 in Kraft treten).

Ziel dieses Workshops ist es, die Teilnehmenden für die rechtlichen Aspekte des digitalen Arbeitens zu sensibilisieren und ihnen im Speziellen einen Überblick über jene Rechtsbereiche zu

verschaffen, mit denen wir im Rahmen unserer Forschungstätigkeiten konfrontiert werden.

Dabei wird ein interaktives Format gewählt, bei dem die Teilnehmer*innen die Gelegenheit finden, ihre konkreten Fragen aus ihrer Anwendungspraxis im Rahmen der thematischen Blöcke sowie am Ende des Workshops in einer offenen Diskussion einzubringen und mit den anwesenden Expert*innen und Kolleg*innen zu erörtern.

Urheberrecht

Das **Urheberrecht** schützt sämtliche materiellen und immateriellen Rechte von Urheber*innen an deren Werken. Ob ein Werk veröffentlicht bzw. erschienen ist oder nicht hat auf das Entstehen von Urheberrecht keinen Einfluss. Nach dem Ableben der Schaffenden geht das Urheberrecht auf die Erb*innen über und erlischt 70 Jahre nach Ableben der Schaffenden (gerechnet ab dem 1.1. des Folgejahres).

Sämtliche **Werknutzungsrechte** können von Urheber*innen bzw. deren Erb*innen verkauft oder anders veräußert werden. Das Urheberpersönlichkeitsrecht (also das Recht, durch die Schaffung eines Werkes dessen - zu nennende*r - Urheber*in zu werden) ist nicht veräußerbar. Archive und andere Institutionen, in deren Eigentum urheberrechtlich geschütztes Material vor oder nach Ableben durch Schenkung oder Verkauf durch die Urheber*innen oder deren Erb*innen übergeht, besitzen häufig nur das Material, nicht aber das (ererbte) Urheber- oder anderweitig erlangte Werknutzungsrecht an dem von ihnen Aufbewahrten.

Am 1. März 2018 tritt in Deutschland das **Urheberrechts-Wissensgesellschafts-Gesetz** (UrhWissG) in Kraft. Dieses novelliert die bestehenden Schranken des Urheberrechts für Unterricht und Forschung (§52a UrhG) und für Bibliotheken bzw. öffentliche Gedächtnisinstitutionen (52b UrhG) und ersetzt diese durch eine Reihe neuer Bestimmungen betreffend Forschung, Lehre, Bibliotheken und Archive (§60a - 60h UrhG). Insbesondere sind auch die Paragraphen 60d (enthält eine breite Ausnahme für data-mining zu nicht-kommerziellen Forschungszwecken) und 60c (für andere nicht-kommerzielle Forschung) von besonderem Interesse für die Digitalen Geisteswissenschaften. Dabei ist festzuhalten, dass die in diesen Paragraphen festgelegten Schranken durch anderslautende Verträge **nicht** ausgesetzt werden können. Allerdings ist eine “angemessene Vergütung” an die zuständigen Verwertungsgesellschaften (wie z.B. VG Wort) zu entrichten. Die Verhandlungen über eine solche „angemessene Vergütung“ sind, wie die Praxis zeigt, langwierig und kompliziert – ein Grund mehr, umgehend diesbezüglich tätig zu werden.

Vom Urheberrecht erfasst sind neben dem Digitalisierungsbereich beispielsweise auch die folgenden (häufigen) Vorhaben:

1. Faksimilierung (Abdruck in Büchern)
2. digitale Faksimilierung (Wiedergabe von Scans auf Webseiten): auch bei der Verwendung z.B. von Fotos oder anderen Einzelbildern, etwa für Startseiten, ist das Urheberrecht zu berücksichtigen!
3. Edition (analog oder digital, auch ohne Beigabe von Faksimiles)
4. Corpuserstellung (auch ohne Zugriff auf den Volltext)

“Copyright” - nationales Recht & der digitale Raum

Der Begriff “copyright” führt oft zu Missverständnissen, da er ein rechtliches Konzept bezeichnet, das es in Österreich bzw. im deutschsprachigen / europäischen Raum nicht gibt. “Copyright”, das dominierende Rechtskonzept im angelsächsischen Raum, stellt das Recht zur Verwertung und Verbreitung eines Werks in den Vordergrund, während im Mittelpunkt des Urheberrechts die geistige Schöpfung und ihre Besitzer*innen stehen. Eine Möglichkeit, diese unterschiedlichen Rechtskonzepte bis zu einem gewissen Grad in Einklang zu bringen, stellen offene Lizenzen dar (siehe unten).

Datenschutzrecht

Forschungsprojekte können unter Umständen mit vom Datenschutzgesetz umfassten Material arbeiten und müssen in diesem Fall die Gesetzeslage berücksichtigen. Das gilt dann, wenn mit personenbezogenen Daten (jegliche Angaben über natürliche oder juristische Personen, deren Identität bestimmt oder bestimmbar ist) und insbesondere mit sensiblen Daten (personenbezogene Daten über natürliche, Personen, über ihre ethnische Herkunft, politische Meinung, Gewerkschaftszugehörigkeit, religiöse oder philosophische Überzeugung, Gesundheit, phänotypische Merkmale oder ihr Sexualleben) gearbeitet werden soll. Das Datenschutzrecht betrifft in erster Linie lebende, natürliche Personen. Auch sensible Informationen über Verstorbene können jedoch rechtlich relevant sein.

Im Mai 2018 tritt die EU-weite Datenschutz-Grundverordnung in Kraft (DS-GVO), die

eine Reihe von Änderungen und Vorgaben enthält sowie eine Novellierung der nationalen Datenschutzgesetze angestoßen hat, die überblicksartig vermittelt werden.

Persönlichkeitsrechte

Das Urheberrechtsgesetz umfasst einige Persönlichkeitsrechte, die es zu beachten gilt und die nach Ableben u.U. jene Aspekte betreffen, die vor Ableben vom Datenschutzrecht abgedeckt wurden. So sind vertrauliche Aufzeichnungen (Briefe, Tagebücher, etc.) auch nach Ableben unabhängig von der Urheberschaft nicht frei zur Veröffentlichung, solange Angehörige oder Adressaten berechtigtes Interesse daran haben, dass das nicht geschieht. Dasselbe gilt für Bilder (vor Ableben spricht man dabei in Österreich auch vom “Recht am eigenen Bild”). Für die Verwendung von Lichtbildern sind daher neben den Reproduktionsrechten (von den Urheber*innen) auch die Einwilligung der Abgebildeten bzw. deren Nachkommen einzuholen.

Lizenzierung

So es die rechtliche Ausgangslage zulässt, ist es empfehlenswert und gute wissenschaftliche Praxis, digitale Forschungsdaten- und Ergebnisse mit möglichst offenen Lizenzen (z.B. Creative-Commons-Lizenzen) zu versehen, um ihre Wieder- und Weiterverwendbarkeit zu gewährleisten. Die Wahl der richtigen Lizenz ist dabei nicht immer einfach, da die Möglichkeiten zu verschiedenen Graden der Offenheit vor allem vom rechtlichen Status des Ausgangsmaterials abhängig sind. Mittlerweile gibt es allerdings schon einige Tools, die bei der Wahl der richtigen Lizenz Unterstützung bieten.

Impressumspflicht

Alle Medieninhaber*innen sind laut Mediengesetz bzw. e-Commerce-Gesetz (für Webseiten) dazu verpflichtet, gewisse Informationen in Form eines Impressums offen zu legen. Dazu gehören im Kontext digitaler wissenschaftlicher Projekte: Angaben zu den Medieninhaber*innen selbst (bei wissenschaftlichen Projekten zumeist die Institution als juristische Person, also Universität, Akademie, etc.), zu den Herausgeber*innen (jene Personen, die die grundlegende Richtung des Mediums bestimmen) und zur grundlegenden Richtung. Umso diffiziler wird es, sobald eine kommerzielle Nutzung der Webseiten gegeben ist, zum Beispiel durch die Vergabe von kostenpflichtigen Zugängen zu Restriktionen unterworfenem Material.

Diensteanbieter

Als Bereitsteller öffentlich zugänglicher digitaler Ressourcen und elektronischer Dienste werden auch Bildungs- und Kulturerbeeinrichtungen zu "Diensteanbietern" im Sinne der E-Commerce-Gesetzgebung. Dabei bestehen wesentliche Unterschiede bezüglich der Pflichten der Institutionen, abhängig von den zur Verfügung gestellten Diensten und Inhalten: Während Content Provider lediglich den Zugang zu selbst generierte Inhalten ermöglichen, wird man durch die Möglichkeit, Benutzer*inneneingaben zu speichern - wie das bereits durch eine Kommentar- oder Gästebuchfunktion der Fall ist, geschweige denn bei Crowdsourcing Projekten - zum Host Provider. Wer aber haftet für fehlerhafte Informationen oder missbräuchliche Nutzung der Ressourcen?

Ablauf

1. Block (14:00-15:30)

Begrüßung & Einführung
Einführung Urheberrecht

- Internationaler Kontext (**Urheber**-Recht vs. **Copy**-Right)
- Rechte der Urheber*innen
- Verwertungsrechte
- Freie Werknutzungen (Gesetzliche Ausnahmen vom Urheberrecht)

Vertiefung Bildrechte

- Digitalisate: Lichtbildwerk, Lichtbild oder Vervielfältigung?
- Persönlichkeitsrechte (Recht am eigenen Bild)
- Freiheit des Straßenbildes

Vertiefung Lizenzierung

- Offene Lizenzen für Forschung und Bildung
- Lizenzierungswerkzeuge

2. Block (16:00-17:30)

Datenschutzrecht

- Begrifflichkeiten
- Dauer
- Datenschutz bei personenbezogenen Quellen

- Datenschutz für BenutzerInnen von Online-Angeboten

E-Commerce Gesetz

- Diensteanbieter: Definitionen und Pflichten
- Impressum und Offenlegung

Fragen und Diskussion

Vortragende

Vanessa Hannesschläger

… studierte Germanistik in Wien. Sie ist wissenschaftliche Mitarbeiterin des Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften (ACDH-ÖAW) und dort für Rechts- und Lizenzierungsfragen zuständig. Schon im Rahmen eines vorangegangenen Forschungsprojekts am Literaturarchiv der Österreichischen Nationalbibliothek beschäftigte sie sich mit praktischen Fragen des Urheberrechts. Mitarbeit in CLARIN (CLARIN PLUS) und DARIAH (WG Thesaurus Maintenance, ELDAH). Zu ihren Forschungsinteressen gehören digitales Edieren, Text- und Datenmodellierung, das Archiv im digitalen Kontext, Vermittlungsstrategien in den DH sowie digitale Infrastrukturen.

Pawel Kamocki

… verfügt sowohl im Bereich des Rechts als auch im Bereich der Sprachwissenschaften über breites Fachwissen; derzeit ist er wissenschaftlicher Mitarbeiter am Institut für Deutsche Sprache in Mannheim und Lehr- und Forschungsassistent an der Descartes Universität in Paris und promoviert zu den rechtlichen Fragestellungen der Open Science. Er ist Mitglied des CLARIN Legal Issues Committee und arbeitete als rechtlicher Berater in zahlreichen anderen Projekten und Arbeitsgruppen (z.B. EUDAT, RDA, OpenMinTeD). Neben Urheberrecht und Datenschutz gilt sein Interesse auch den Sprachwissenschaften (insb. rechtliche Fachsprache).

Walter Scholger

… studierte Geschichte und Angewandte Kulturwissenschaften in Graz und Maynooth und ist administrativer Leiter des Zentrums für Informationsmodellierung - Austrian Centre for Digital Humanities an der Universität Graz. In Projekten, internationalen Workshops und universitärer Lehre widmet er sich rechtlichen Aspekten des digitalen Kulturerbes und Fragen offener digitaler Publikationsformen.

Er ist Mitglied in facheinschlägigen Arbeitsgruppen der Digital Humanities Dachverbände und internationaler Projekte (ADHO, DHd, ICARUS, DARIAH) zu rechtlichen Aspekten, digitalen Publikationen und Lehre im Bereich der Digital Humanities.

Bibliographie

Amini, Seyavash / Blechl, Guido / Losehand, Joachim (2015): FAQs zu Creative-Commons-Lizenzen unter besonderer Berücksichtigung der Wissenschaft. https://phaidra.univie.ac.at/view/o:408042

Bergauer, Christian / Jahnel, Dietmar (2017): Das neue Datenschutzrecht DSG-VO und DSG (2018), Wien: Jan Sramek Verlag

Bundeszentrale für politische Bildung (2013): Urheberrecht und Copyright. https://www.bpb.de/gesellschaft/medien/urheberrecht/169971/urheberrecht-und-copyright

Datenschutzbehörde der Republik Österreich: Gesetze zum Datenschutzrecht. https://www.dsb.gv.at/gesetze-in-osterreich

Deutscher Bundestag (2017): Entwurf eines Gesetzes zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissensgesellschafts-Gesetz – UrhWissG). https://www.bundestag.de/blob/507608/be49b0e7039f039593112136e262b55f/gesetzentwurf-data.pdf

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 (2014): Creative Commons and Language Resources: General Issues and What's New in CC 4.0. CLARIN Legal Issues Committee: White Paper Series. http://clarin-d.de/images/legal/CLIC_white_paper_1.pdf

Klimpel, Paul / Weitzmann, John H. (2015): Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften. DARIAH-DE Working papers Nr. 12. Göttingen: DARIAH-DE. https://irights.info/wp-content/uploads/2015/08/Forschen-in-der-digitalen-Welt-Juristische-Handreichung-Geisteswissenschaften-dwp-2015-12.pdf

Klimpel, Paul (2013): Free Knowledge Thanks to Creative Commons Licenses. Why a Non-commercial Clause often won't Serve Your Needs. Wikimedia Deutschland/iRights.info/CC

DE. https://www.wikimedia.de/w/images.homepage/1/15/CC-NC_Leitfaden_2013_engl.pdf

Kucsko, Guido / Zemann, Adolf (2017). CC0 1.0 Universal - Beurteilung der Verzichtserklärung und der Lizenzerteilung im Rahmen der Fallback-Klausel nach österreichischem Recht. https://phaidra.univie.ac.at/view/o:528411

Saferinternet.at (2013): Urheberrecht. 24 Fragen und Antworten. https://www.saferinternet.at/fileadmin/files/Materialien_2013/Ratgeber_Urheberrecht.pdf

Reisewege in Raum und Zeit

Aschauer, Anna
aschauer@ieg-mainz.de
Institut für Europäische Geschichte, Deutschland

Büchler, Marco
buechler@ieg-mainz.de
Institut für Europäische Geschichte, Deutschland

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Otto-Friedrich-Universität Bamberg

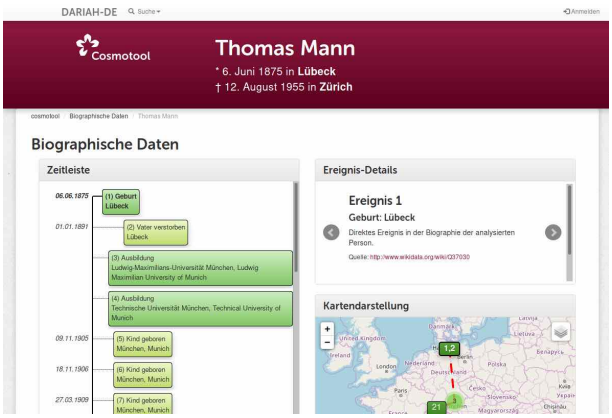
Henrich, Andreas
andreas.henrich@uni-bamberg.de
Otto-Friedrich-Universität Bamberg

Reisewege gut und genau zu kennen, ist für die Geschichtsforschung von enormer Bedeutung. Nicht nur, dass so persönliche Kontakte reflektiert werden können, so zeigen die Reisewege vielmehr auf, wie sich Güter oder Ideen verbreitet haben können. Weiterhin zeigen Reisewege auch auf, über welches Ortswissen eine bestimmte Person verfügt hat. So sei bspw. auf die aktuelle Diskussion rund um Shakespeare hingewiesen. Auf der einen Seite wird über stilometrische Analysen gezeigt, dass Shakespeare wirklich der Autor seiner Werke sein soll. Auf der anderen Seite wird in Romeo und Julia Verona so detailreich dargestellt, dass eine gute Ortskenntnis notwendig gewesen sein muss. Da Shakespeare nach jetzigem Kenntnisstand niemals in Italien war, nährt dies die Ansichten der Gegenseite, dass er nicht der Autor gewesen sein kann. Ferner gibt es nach dieser Ansicht auch stilometrische Signale, welche auf eine Frau hinweisen. Zusätzlich wird diese Sichtweise dahingehend unterstützt, dass sich die Charaktere in Shakespeares Werken mit den Namen von Familienmitgliedern von Emilia Lanier aus Bassano, Italien, auffällig gleichen. Das größte Problem in der gesamten Diskussion ist, dass es eine sieben-jährige Lücke in Shakespeares Lebenslauf gibt, in welcher völlig unklar ist, wo er war und mit wem er in Kontakt stand. Das Füllen dieser biographischen Lücke würde, wahrscheinlich, wichtige Impulse in der vorangestellten Diskussion geben. Ferner zeigt es im Allgemeinen wie wichtig die Vollständigkeit von Biographien nicht nur für Historiker, sondern auch für andere Disziplinen in den Geisteswissenschaften ist

Neben der dargestellten Vollständigkeit besteht auch die Notwendigkeit der Konzeptualisierung. Hierbei werden mehrere Objekte und Personen zu Gruppen formiert. Derartige Konzeptualisierungen können als latente Variablen verwendet werden, um auch Objekte und Personen zu finden, welche unter einem anderen Namen gefunden werden. Durch den Einsatz digitaler Werkzeuge ist es möglich die Eigenschaften, die einer Gruppe inhärent sind, zusammenzufassen. Diese Aggregation zu Gruppenprofilen ermöglicht historisch Arbeitenden, ihre Hypothesen zu bilden oder zu stärken. Es würde ermöglichen, beispielsweise, etablierte Thesen in der klassischen Geschichtswissenschaft bezüglich der Korrelation zwischen Konfessionalität, bzw. Beruf und Mobilität zu prüfen.

Grundlage

Die Partner der Otto-Friedrich-Universität Bamberg und des Leibniz-Institutes für Europäische Geschichte arbeiten im Rahmen von DARIAH-DE an einem Werkzeug welches die Basis für diesen Workshop darstellt: Das CosmoTool steht interessierten Nutzern als Prototyp unter https://cosmotool.de.dariah.eu zur Verfügung. Neben der Darstellung derzeit implementierter Funktionalität zielt der Workshop wesentlich auf die Gewinnung von Anforderungen und Interessen aus der Fachcommunity zur Weiterentwicklung des CosmoTools.



Motiviert wurde das CosmoTool aus dem Bedürfnis, biographische Datenquellen unterschiedlicher Art und Strukturiertheit zu verarbeiten, analysieren und in so genannte biographische Profile zusammenzuführen. Durch die Integration verschiedener Perspektiven auf historische Personen soll neben der Transnationalität und -kontextualität einzelner Biographien insbesondere auch die aggregierte Betrachtung von Personengruppen ermöglicht werden und Unterschiede beispielsweise in der Mobilität unterschiedlicher Berufsgruppen evaluiert werden können (Panter, Paulmann 2015). Durch den Vergleich biographischer Profile untereinander oder mit benutzerdefinierten Gruppen sollen inhaltliche Vergleiche, Zusammenhänge oder Abweichungen offenbart werden, die hypothesengenerierend auf die qualitative Forschung einwirken können.

Aus der infrastrukturellen Perspektive von DARIAH-DE bildet das CosmoTool eine Brücke zwischen der konkreten fachwissenschaftlichen Forschung und den generischen Diensten von DARIAH-DE (Gradl, Henrich 2016a). So bedient sich das CosmoTool zur Umsetzung inhaltlicher Funktionalität insbesondere bei dem DARIAH-DE Data Modeling Environment (DME), welches die Modellierung und Explizierung von Daten z. B. auch durch die Anwendung von Methoden des Natural Language Processings (NLP) kapselt. Einmal modellierte Daten stehen einer Vielzahl möglicher Anwendungsfälle zur Verfügung - von denen einer im CosmoTool besteht.

Konzeptualisierung

Das CosmoTool ermöglicht eine Selektion und Filterung großer Datenmengen. Eine grundlegende Hypothese ist, dass ein Mehrwert durch die automatisierte Erstellung biographischer Profile entsteht. Die Gruppenprofile können anhand der von FachwissenschaftlerInnen erstellten Wortfelder aggregiert werden. So können Aggregationskriterien wie beispielsweise Berufsfelder oder Konfessions-/Religionsgruppen beschrieben werden.

	Katholiken	Protestanten	Juden	Muslime
Berufsbezeichnung	Papst, Priester, Pater/Patres, ...	Pastor, Kantor*, Diakonisse	Rabbi, Rabbiner, Rav, Chazan, Zaddik, Sofer ...	Imam, Hoca/Hodscha, Alim/Ulema, ...
Studienorte	Italien (Rom, Ragusa, Pavia) ...	Jena, Wittenberg, Halle, ...	Jeshivot/Yeshiva/Jeschiwa (sehr spezifisch) in Prag, ...	Mekka, Istanbul/Konstantinopel, Madrasa/Medresa, Moschee
Schulart	Jesuitenkolleg/Jesuitenschule, Priesterseminar, Klosterschule, Domschule	Akademisches Gymnasium/Gymnasium Academicum, Gelehrtenschule *, Pädagogium. ...	Jüdische Freischule; Jacobsonschule; Rabinatsschule; Cheder (=Elementarschule)	Devsirme (Knabenlese)
Tätigkeitsorte	Erzbistümer: Mainz, Hamburg-Bremen, Köln, Salzburg, ...		Polen, Litauen	Marokko, Tunis, Polen, Krim, Kaukasus, Balkan, ...

Abbildung 1: Auszug aus der Tabelle, die Konfessions-, bzw. religionsspezifische Wörter zur Identifikation eines (vormodernen) Katholiken, Protestanten, Juden oder eines Muslims in modernen (deutschsprachigen) biographischen Lexika, ermöglicht.

Thematische Schwerpunkte

Der Workshop ist auf einen halben Tag konzipiert. Gegenstand des Workshops ist die geographische Darstellung von Reisewegen sowie deren technischen Anforderungen und wissenschaftlichen Implikationen für die Digital Humanities. Der Fokus des Workshops besteht neben der Präsentation des aktuellen Entwicklungsstandes insbesondere in der Gewinnung von Rückmeldungen und Anforderungen für die weitere Konzeptualisierung und Entwicklung des Werkzeugs.

Aus wissenschaftlicher Sicht unterteilt sich der Workshop in zwei Impulsreferate mit je einem Beitrag aus den Geschichtswissenschaften sowie der Informatik. Dem schließt sich eine Gruppenarbeit zu vorgegebenen Themen an.

Impulsvortrag 1 - Informatik und Infrastruktur: Analyse und Kombination forschungsspezifischer Methoden auf dem infrastrukturellen Fundament von DARIAH-DE

Das CosmoTool basiert auf verschiedenen Komponenten der Föderationsarchitektur von DARIAH-DE. Auf Basis der DARIAH-DE Collection Registry und ihrer Schnittstellen werden biographische Datenquellen und deren Zugriffsmöglichkeiten verzeichnet, beschrieben und im CosmoTool nachgenutzt. Mit Hilfe des DARIAH-DE Data Modeling Environment (DME) können strukturierte und unstrukturierte Daten modelliert und anzuwendende Werkzeuge, wie Methoden des Natural Language Processings (NLP) konfiguriert werden (Gradl, Henrich 2016b). Das CosmoTool nutzt die Funktionalität des DME beispielsweise, um auf Basis unstrukturierter Texte biographische Zusammenhänge erkennen und extrahieren zu können.

Das CosmoTool verknüpft die generische Perspektive einer Forschungsinfrastruktur, die stets auch Aspekte der Nachhaltigkeit und Nachnutzbarkeit fokussiert, mit spezifischen historischen Forschungsfragen. Durch das Zusammenwirken der verschiedenen Komponenten kann die Funktionalität des CosmoTools durch Experten unterschiedlicher fachwissenschaftlicher Kontexte um neue analytische Verfahren erweitert werden. Einmal verwendete Verfahren z. B. zur Disambiguierung von Namen oder Zeitangaben können für artverwandte Datenquellen nachgenutzt und angepasst werden.

Impulsvortrag 2: Reise in der Frühen Neuzeit: Rekonstruktion der Bewegung konfessioneller Minderheiten

In der Geschichtsschreibung wird der Fokus oft auf religiöse/kulturelle Minderheiten gelegt, weg vom Blickwinkel der dominierenden Gruppen, da diesen Minderheiten hohe politische, ökonomische und kulturelle Einflüsse zugeschrieben werden. Solche religiösen Gruppen werden oft zu Medien zwischen zwei Kulturen und führen Transfers (u. A. Reisen) durch. In bestimmten Fällen bilden diese Gruppen, wenn sie von Fremden umgegeben sind, ein starkes nationales bzw. religiöses Bewusstsein, assimilieren sich aber nicht und bilden Diasporen heraus. Um die Diasporen zu rekonstruieren ist es hilfreich Reisewege der Gruppen oder einzelnen Reisenden zu verfolgen, da sie Einblicke in das Netzwerk der Vertrauten erlauben. Dabei sind Reisewege in der Frühen Neuzeit eine besondere Herausforderung: z. B. Reise/Ausreisebeschränkung, (aus moderner Sicht) fehlende Infrastruktur wie befestigte Straßen oder fehlende Sicherheit, was die Verlässlichkeit auf eigene Netzwerke noch wichtiger macht.

Die Reisetätigkeiten dieser Gruppen werden anhand der Beispiele der Lutheraner aus deutschsprachigen Gebieten in Russland des 18. Jahrhunderts erläutert: Ausgehend aus ihrer “Beheimatung”, über ihrer Reisewege bis hin zu ihren Tätigkeitsorten.

Arbeitsgruppen:

Den Impulsvorträgen schließen sich Gruppenarbeiten zu jeweils einer der folgenden drei Themen an:

Vollständigkeit: Für Historiker ist die Vollständigkeit ein wichtiges Kriterium. Um Reisewege zu verfolgen, ist es zumindest theoretisch notwendig, eine vollständige Datengrundlage vorzufinden. Jedoch ist dies praktisch nur in den seltensten Fällen realisierbar. Deshalb soll sich eine Arbeitsgruppe mit der Fragestellung beschäftigen, ob und vor allem wie mit unvollständigen Daten umgegangen werden kann und ob sich dennoch ein wissenschaftlicher Mehrwert für Historiker ergibt.

Technische Machbarkeit / Repräsentationsformen: Eine weitere Arbeitsgruppe bearbeitet das Thema der technischen Machbarkeit. Der Fokus liegt hierbei auf der Frage, wie entsprechende Daten vorgehalten und gespeichert werden sollen, so dass bspw. auch mehrere Datenbanken miteinander verknüpft werden können.

Rechtliche und ethische Aspekte: Werkzeuge, wie das CosmoTool, entfalten erst dann ihre volle Leistungsfähigkeit, wenn über Personen möglichst viel oder gar alle relevanten Reisewege enthalten sind. Auf der anderen Seite werden so für Forschungszwecke historische Personen völlig transparent dargestellt, ohne dass überhaupt klar ist, ob ebenje Personen einer derartigen Darstellung überhaupt zustimmen würden. Die Kernfrage, die sich demnach stellt, ist, wo die Forschung auch im Interesse der entsprechenden Personen enden muss (vgl. auch Berendt et al 2015: *Is it research or is it spying? – Rethinking ethics in Big Data AI and other fields of knowledge science*).

Zielpublikum

Das Zielpublikum sollte sich aus interessierten Forschern aus den Digital Humanities zusammensetzen. Neben potenziellen Anwendern mit ihren spezifischen Forschungsfragen und Anforderungen, könnte der Workshop insbesondere auch für Sammlungsinhaber und Entwickler artverwandter Werkzeuge von Interesse sein. Die Organisatoren würden den Preconference-Workshop auch separat bewerben, um interessierte Historiker auf das Event aufmerksam zu machen. Der Workshop ist für 15-20 Teilnehmer ausgelegt.

Geplantes Programm

Der Workshop besteht im Wesentlichen aus den drei Teilen Einführung, Impulsreferate sowie einer Gruppenarbeit mit den Workshop-Teilnehmern. Der Workshop gliedert sich im Detail wie folgt:

- 09:00 - 09:30 Welcome und Einführung in das Workshopthema (Aschauer, Büchler, Gradl, Henrich)
- 09:30 - 10:00 Impulsvortrag Informatik (Gradl, Henrich): *Informatik und Infrastruktur: Analyse und Kombination forschungsspezifischer Methoden auf dem infrastrukturellen Fundament von DARIAH-DE*
- 10:00 - 10:30 Impulsvortrag Geisteswissenschaften (Aschauer, Büchler): Reise in der Frühen Neuzeit: Rekonstruktion der Bewegung konfessioneller Minderheiten
- 10:30 - 11:00 Kaffeepause
- 11:30 - 12:30 Arbeiten in Arbeitsgruppe
- 12:30 - 13:00 Diskussionsrunde mit dem Publikum (Aschauer, Büchler, Gradl, Henrich)

Kontaktdaten und Forschungsinteressen der Beitragenden

In alphabetischer Reihenfolge:

Anna Aschauer, Leibniz-Institut für Europäische Geschichte, Mainz. Interessen: Pietismusforschung, Migration der religiösen Minderheiten in der Frühen Neuzeit, Digital Humanities. E-Mail: aschauer@ieg-mainz.de

Marco Büchler, Leibniz-Institut für Europäische Geschichte, Mainz. Interessen: Digital Humanities, Text Mining, Information Retrieval, Big Humanities Data. E-Mail: buechler@ieg-mainz.de

Tobias Gradl, Otto-Friedrich-Universität Bamberg. Interessen: Forschungsdaten und Forschungsdatenmanagement, Digital Humanities, Datenintegration. E-Mail: tobias.gradl@uni-bamberg.de

Andreas Henrich, Otto-Friedrich-Universität Bamberg. Interessen: Datenbanken, Information-Retrieval, Digital Humanities, Multimediale Systeme, Softwareentwicklung. E-Mail:

Ausstattung

Es wird keine zusätzliche Ausstattung neben der üblichen Präsentationstechnik benötigt.

Bibliographie

Berendt, B., Büchler, M., Rockwell, G. (2015) 'Is it research or is it spying? – Rethinking ethics in Big Data AI and other fields of knowledge science', German Journal on Artificial Intelligence. Springer.

Gradl, T.; Henrich, A.(2016a): Die DARIAH-DE-Föderationsarchitektur: Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen, Bibliothek Forschung und Praxis. Band 40, Heft 2, Seiten 222-228, ISSN (Online) 1865-7648, ISSN (Print) 0341-4183, DOI: https://doi.org/10.1515/bfp-2016-0027

Gradl, T.; Henrich, Andreas (2016b): „Data Integration for the Arts and Humanities: A Language Theoretical Concept". In: Fuhr, Norbert et al. (Hg.): Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings. Cham: Springer International Publishing, S. 281–293

Panter, S.; Paulmann, Johannes und Margit Szöllösi-Janze (2015): "Mobility and Biography. Methodological Challenges and Perspectives", in: Mobility and Biography, hrsg. von Sarah Panter (=Jahrbuch für Europäische Geschichte / European History Yearbook 16), Berlin: De Gruyter Oldenbourg, S. 1-14

Research Software Engineering und Digital Humanities. Reflexion, Kartierung, Organisation.

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

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

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

Software ist in vielen Fällen ein integraler Bestandteil von Forschungsaktivität, so auch in den Digital Humanities zur Bearbeitung von geistes- und kulturwissenschaftlichen Forschungsfragen mit digitalen Methoden. Im weiteren Sinn verstanden reicht das Anwendungsfeld für geisteswissenschaftliche Forschungssoftware von der täglichen Arbeit mit Webbrowsern und Textverarbeitungsprogrammen bis hin zum Einsatz spezialisierter Softwarelösungen, virtueller Forschungsumgebungen oder auch webbasierter Publikations- und Analyseinstrumente für geisteswissenschaftliche Forschungsdaten. Im engeren Sinn verstanden ist Forschungssoftware in den Digital Humanities als Summe spezifischer Komponenten aufzufassen, die unter entwicklerischer Durchdringung der jeweiligen geisteswissenschaftlichen Wissens- und Anwendungsdomäne konzipiert und implementiert werden.

Zum aktuellen Zeitpunkt finden die Entwicklungsprozesse für Forschungssoftware in den Digital Humanities häufig noch isoliert, unreflektiert, undokumentiert, nicht an gängigen Industriestandards ausgerichtet und insbesondere nicht innerhalb eines organisierten, disziplinspezifischen Rahmens statt. Aus dieser Situation heraus ergibt sich ein eminentes Nachhaltigkeitsproblem für geisteswissenschaftliche Forschungssoftware. Dies verwundert umso mehr, denn während das Bewusstsein für eine nachhaltige Erschließung kultureller Objekte durch die Entwicklung und den Einsatz entsprechender systemneutraler Datenformate und -standards inzwischen als hoch eingeschätzt werden kann, spielt die Ebene der Softwareentwicklung in den Nachhaltigkeitsdiskussionen der Digital Humanities kaum eine Rolle. Während andere Disziplinen wie beispielsweise die Natur- und Ingenieurswissenschaften, aber auch die informatisch geprägten Lebenswissenschaften diese Problematik erkannt und erste Gegenmaßnahmen ergriffen haben, gilt in den Digital Humanities immer noch das schon vor Jahren geprägte Mantra „Data ages like wine, software ages like fish“. [1] Hinzu kommt, dass DH-Softwareentwickler_innen häufig noch als digitale Geisteswissenschaftler_innen 'zweiter Klasse' wahrgenommen werden und deutlich schlechtere Möglichkeiten zur Entfaltung ihrer akademischen Karriere haben – obwohl sie die 'Ermöglichenden' bzw. zentrale Beitragende zum jeweiligen wissenschaftlichen Erkenntnisgewinn sind (vgl. Brett 2017, S. 22).

Der 2016 von Simon Hettrick vorgelegte Bericht des *Software Sustainability Institute* kommt in Bezug auf die generellen Voraussetzungen zur Steigerung der Nachhaltigkeit von Forschungssoftware zu folgendem Schluss: „Many researchers know how to code, but few understand the wider set of skills that are needed to develop reliable, reproducible and reusable software. […] software engineering should be incorporated […] at the very start of a research career“. (Hettrick 2016, S. 14).

Der hier vorgeschlagene Workshop versteht sich als eine erste Maßnahme, in den Digital Humanities im deutschsprachigen Raum einen kritischen Reflexionsprozess zum Thema 'Nachhaltige Softwareentwicklung' anzustoßen und durch die Etablierung einer gemeinsamen Diskussionsplattform ein stärkeres Bewusstsein für diesen zentralen, aber vernachlässigten Baustein guter Digital Humanities Forschung zu wecken.

Der Workshop bildet den Arbeitsauftakt für die ins Auge gefasste Beantragung einer AG *Research Software Engineering* innerhalb des DHd-Verbandes (AG DH-RSE). Ziel ist es, im Rahmen des Workshops möglichst vielen Wissenschaftler_innen sowohl aus universitären wie auch außeruniversitären Kontexten, die als digitale Geisteswissenschaftler_innen im Feld der geisteswissenschaftlichen Softwareentwicklung tätig sind, ein gemeinsames Forum für die Arbeit an einem 'Manifest' für eine gute und nachhaltige Implementierungspraxis zu geben. Dieses Manifest versteht sich gleichzeitig als Gründungsdokument der AG DH-RSE. Eine im Vorfeld dieser Einreichung getätigte informelle Erhebung hat einen großen Bedarf für ein solches Format ermittelt. Die Gruppe von Interessent_innen und potentiellen Teilnehmer_innen eines solchen Workshops sowie der angestrebten AG besteht bereits jetzt aus 20 Personen. Die Ausrichtung in Form eines Workshops auf der DHd-Konferenz soll das Einzugsgebiet und die Teilnahmemöglichkeit für weitere interessierte DH-Softwareentwickler_innen vergrößern und öffnen, wobei als Zielgruppe insbesondere solche Entwickler_innen aus dem inner- wie außeruniversitären Kontext ins Auge gefasst werden, die über eine entsprechende Projekterfahrung verfügen.

In Anlehnung an die im anglo-amerikanischen Raum bereits etablierte *UK Research Software Engineer Association* (http://rse.ac.uk/who/) und den *Workshops on Sustainable Software for Science: Practice and Experiences* (WSSSPE, http://wssspe.researchcomputing.org.uk) wird der DHd-Workshop verschiedene Formate integrieren. Hierzu gehören neben einer *Keynote* auch *Lightning Talks* und *Breakout Groups*, die sich im Verlauf des Workshops spezifischen Unterthemen

widmen. Eine Zusammenfassung zum Ende des Workshops fokussiert auf den erreichten Stand des gemeinsamen Manifests sowie auf die konkreten weiteren Schritte zur Implementierung der AG DH-RSE. Der Workshop gliedert sich wie folgt:

1. Impuls: Warum eine AG DH-RSE?
2. Keynote
3. Impuls: Gegenwärtige *best practices* im Bereich nachhaltiger Softwareentwicklung
4. *Breakout Groups*
   - a) Ausbildung, Training, Software Carpentry, Standards
   - b) Struktur, Infrastruktur, Workflows für die AG DH-RSE
   - c) Definitionen: Sustainability, Software, Dokumentation, etc.
   - d) Manifest der AG DH-RSE; mit agiler Integration von Gruppen a), b), c)
5. *Wrap-Up* und Online-Veröffentlichung der Workshop-Ergebnisse sowie einer ersten Version des Manifests

Im Fazit möchte der Workshop eine kritische Reflexion der gegenwärtigen Prozesse und Praktiken im Bereich der DH-Softwareentwicklung anstoßen, eine erste Kartierung dieses wissenschaftlichen Tätigkeitsfeldes und seiner Akteure erstellen und eine Organisationsform für eine koordinierte Weiterentwicklung dieses wichtigen Teilbereiches der Digital Humanities finden. Als Startplattform für den Workshop steht eine GitHub-Website zur Verfügung, an der im Rahmen des Workshops kollaborativ gearbeitet wird: https://dh-rse.github.io/dhd-workshop-2018/ .

Kontaktdaten und Forschungsinteressen der Beitragenden

Torsten Schrade
Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz
06131/577 119

Torsten Schrade ist Leiter der Digitalen Akademie der Akademie der Wissenschaften und der Literatur | Mainz und Professor für Digital Humanities an der Hochschule Mainz. Zu seinen Forschungsinteressen zählen Methoden, Verfahren und Prozesse zur Steigerung der Nachhaltigkeit und Qualität geisteswissenschaftlicher Forschungssoftware, insbesondere aus dem Umfeld der agilen Softwareentwicklung. Weitere Schwerpunkte liegen im Forschungsdatenmanagement, dem Einsatz von Webtechnologien für geisteswissenschaftliche Forschungsapplikationen sowie der Anwendung von *Semantic Web* und *Linked Open Data* Technologien zur Erschließung neuer Analyse- und Nachnutzungspotentiale für geistes- und kulturwissenschaftliche Forschungsdaten.

Stephan Druskat
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Unter den Linden 6
10099 Berlin
030/2093 9726

Stephan Druskat arbeitet als Research Software Engineer in der Linguistik und den Digital Humanities. Seine Forschungsinteressen umfassen die technische Nachhaltigkeit von Forschungssoftware, ihre Messung und Dokumentation, die Anwendung informatischer Methoden auf die Entwicklung von Forschungssoftware und Community- und Öffentlichkeitsaspekte dieses Feldes. Weitere Schwerpunkte liegen auf Forschungsdatenmanagement, insbesondere die Langzeitverfügbarkeit und Nachnutzbarkeit von Sprachdaten. Er ist aktives Mitglied von de-RSE, der Interessenvertretung der Research Software Engineers in Deutschland und Teil von WSSSPE, einer internationalen Arbeitsgruppe zur Nachhaltigkeit von Forschungssoftware.

Alexander Czmiel
Berlin-Brandenburgische Akademie der Wissenschaften
TELOTA - The electronic life of the Academy
Jägerstr. 22/23
10117 Berlin
030/20370276
czmiel@bbaw.de

Alexander Czmiel ist seit 2005 wissenschaftlicher Mitarbeiter bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften und dort für die fachliche Beratung und Umsetzung von Projekten im Bereich der "Digital Humanities" verantwortlich. Seine Forschungsinteressen konzentrieren sich vor allem auf den Bereich der Digitalen Editionen und dort auf Konzepte, Workflows, Standardisierungsmöglichkeiten und Nutzeroberflächen. In diesem Kontext treibt ihn sein einiger Zeit die Nachhaltigkeit der Funktionalitätsschicht Digitaler Editionen um. Seit kurzem beschäftigt er sich zudem mit den Möglichkeiten von Augmented und Virtual Reality in den Digital Humanities.

Fußnoten

1. Dies wird bspw. am Programm der letzten internationalen RSE-Konferenz (September 2017, Manchester, http://rse.ac.uk/conf2017/) und der fachlichen Provenienz der Vortragenden deutlich. S.a. die Verteilung der Disziplinen in

Hettrick, S. et. al (2014): "UK Research Software Survey 2014" *[Data set]*, Zenodo, http://doi.org/10.5281/zenodo.14809 [letzter Zugriff am 25.09.2017].

Bibliographie

Brett, Alys / Croucher, Michael et.al. (2017): "Research Software Engineers: State of the Nation Report 2017", Zenodo http://doi.org/10.5281/zenodo.495360 [letzter Zugriff am 25.09.2017].

Czmiel, Alexander (2017): "Dokumentation, Werkzeugkasten, Pakete - Nachhaltigkeit von Daten und Funktionalität Digitaler Editionen", Einreichung für die DHd-Konferenz 2018, Köln.

Druskat, Stephan / Vertan, Cristina (2017). "Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora", Einreichung für die DHd-Konferenz 2018, Köln.

Druskat, Stephan / N. Chue Hong et.al. (2017): "Proceedings of the Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE5.1)", figshare https://doi.org/10.6084/m9.figshare.c.3869782 [letzter Zugriff am 25.09.2017].

Faniel, Ixchel (2015): "Data Management and Curation in 21st Century Archives", 21. September 2015 http://hangingtogether.org/?p=5375 [letzter Zugriff am 25.09.2017].

Fowler, Martin (2001): "Manifesto for Agile Software Development", http://agilemanifesto.org/iso/de/manifesto.html [letzter Zugriff am 25.09.2017].

Hettrick, Simon (2016): "Research Software Sustainability: Report on a Knowledge Exchange Workshop", Edinburgh, The Software Sustainability Institute http://repository.jisc.ac.uk/6332/1/Research_Software_Sustainability_Report_on_KE_Workshop_Feb_2016_FINAL.pdf [letzter Zugriff am 25.09.2017].

Hong, N. Chue et al. (2010): "Software Preservation Benefits Framework. Software Sustainability Institute Technical Report", Edinburgh, The Software Sutainability Institute http://www.research.ed.ac.uk/portal/files/1219870/SoftwarePreservationBenefitsFramework.pdf [letzter Zugriff am 25.09.2017].

Katz, Daniel S. et al. (2016): "Report on the Fourth Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE4)", The Computing Research Repository (CoRR), abs/1705.02607 https://arxiv.org/abs/1705.02607 [letzter Zugriff am 25.09.2017].

Komus, Ayelt et.al. (2015): "Studie Status Quo Agile", Hochschule Koblenz, BPM Labor http://www.status-quo-agile.de/ [letzter Zugriff am 25.09.2017].

Schrade, Torsten (2017): "Nachhaltige Softwareentwicklung in den Digital Humanities. Konzepte und Methoden", in: Konferenzband der DHd2017, Bern 2017, S. 168-171, http://www.dhd2017.ch/wp-content/uploads/2017/03/Abstractband_def3_März.pdf und https://digicademy.github.io/2017-dhd-sustainable-software/ [letzter Zugriff am 25.09.2017].

Suche und Visualisierung von Annotationen historischer Korpora mit ANNIS. Kritik der korpuslinguistischen Analysemethoden in einem erweiterten Nutzungskontext

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

Krause, Thomas
krauseto@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Guescini, Rolf
rolf.guescini@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Kühnlenz, Frank
frank.kuehnlenz@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

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

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

Historische Korpora (Gippert und Gehrke 2015; Claridge 2008; Rissanen 2008) dienen in vielen geisteswissenschaftlichen Disziplinen als Analysegrundlage und können sehr unterschiedlich aufbereitet sein. Mit korpusbasierten Studien können qualitative und quantitative Analysen, die für die Überprüfung von Hypothesen über ein bestimmtes Phänomen notwendig sind, durchgeführt werden. Dem gegenüber steht methodisch die korpusgetriebene Studie, die das Korpus selbst nutzt, um Hypothesen über ein Phänomen zu generieren (vgl. McEnery und Hardie 2012; Lüdeling und Zeldes 2007). Neben diesen zwei Studientypen können mit Hilfe von Korpora auch einzelne Belege und Kontexte für die Beantwortung verschiedenster Forschungsfragen ermittelt werden.

Eine andere methodische Unterscheidung wird mit dem *close reading* und dem *distant reading* gemacht (vgl. Moretti 2016; Federico 2015; Gooding et al. 2013; Simanowski 2011). Wobei *close reading* hermeneutische, nicht zwingend digitale Methoden umfassen und eine methodische Nähe zu den korpusinformierten Belegstudien sowie zu qualitativen korpusbasierten Studien aufweisen kann. *Distant reading* ist hingegen vergleichbar mit überwiegend quantitativen, korpusgetriebenen Studien. Ein zusätzlicher Aspekt dieser Methoden ist auch die Visualisierung der Daten für *distant reading* oder des Textes für *close reading.* Verschiedene Visualisierungen der Annotationen werden für die korpusbasierten, -getriebenen und -informierten Studien eingesetzt und können so verschiedene Analysen unterstützen oder auch erst ermöglichen.

In den digitalen Geisteswissenschaften müssen daher für die jeweiligen Methoden und Forschungsdaten Analyse- und Visualisierungswerkzeuge entwickelt werden, die es den Forscherinnen und Forschern ermöglichen, für ihren jeweiligen Forschungskontext aus einem breiten methodischen Spektrum wählen zu können (vgl. für einen Überblick z.B. Kupietz und Geyken 2016). Ein solches Werkzeug ist ANNIS (Krause und Zeldes 2016), das Such- und Visualisierungstool für Annotationen, das wir in unserem Workshop den Forscherinnen und Forschern aus den Digital Humanities vorstellen möchten. ANNIS erlaubt das Durchsuchen von Korpora, die unterschiedliche Arten von Annotationen, die möglicherweise durch unterschiedliche Forschergruppen unter verschiedenen Gesichtspunkten annotiert worden, in einem Korpus vereinen. Diese Flexibilität erlaubt es, annotierte Phänomene in der Suche zu kombinieren und damit komplexere Strukturen zu finden.

Neben der Unterstützung der vielfältigen Analysemethoden ist eine weitere Herausforderung für die Analysewerkzeuge, dass historische Korpora je nach Forschungskontext und -frage unterschiedlich erstellt und aufbereitet werden (Lüdeling 2011). Dies zeigt sich unter anderen in den vielfältigen Transkriptions- und Normalisierungsverfahren (vgl. z.B. Odebrecht et al. 2016; Krasselt et al. 2015; Archer et al. 2015; Bollmann et al. 2012; Jurish 2010) und Annotationsguidelines (für z.B. Annotation von Wortarten für historisches Deutsch Coniglio et al. 2016; Dipper et al. 2013) sowie verschiedenen Formaten (z.B. Romary et al. 2015; Schmidt und Wörner 2009; Burnard und Baumann 2008; Wittenburg et al. 2006; Dipper 2005), die allein für die Erstellung von historischen Korpora eingesetzt werden.

Damit historische Korpora mit verschiedenen Methoden analysiert werden können, muss deren Wiederverwendung ermöglicht werden. Die Wiederverwendung von historischen Korpora wird durch u.a. deren freie Veröffentlichung und umfassende Dokumentation möglich (Odebrecht 2014; Borgmann 2012; Büttner et al. 2011). Weiterhin erhöht eine Wiederverwendung ihre Sichtbarkeit und stellt eine Chance zur engeren Vernetzung und Zusammenarbeit in den digitalen Geisteswissenschaften dar. So können auch historische Korpora in unterschiedlichen Wiederverwendungsszenarien gedacht werden (vgl. Simons und Bird 2008) und als empirische Grundlage für die verschiedenen Analysemethoden dienen.

Dieser Workshop möchte ausgehend von diesen Themenkomplex mit den Teilnehmerinnen und Teilnehmern folgende Fragen diskutieren: Wie können Analysewerkzeuge den Forscherinnen und Forschern vielfältige Analysemethoden und Visualisierungsmethoden für verschiedene historische Korpora ermöglichen? Wie kann ANNIS die verschiedenen Analysemethoden bislang unterstützen? Wie kann es gelingen, auch die Vielfältigkeit der Forschungsdaten als solche zu berücksichtigen und deren Wiederverwendung zu ermöglichen? Wie können Werkzeuge spezifisch genug entwickelt werden, um genaue und für den Forschungskontext und die Forschungsdaten angepasste Analysen zu ermöglichen?

Der Workshop hat das Ziel, anhand mehrerer historischer Korpora des Deutschen das generische Such- und Visualisierungstool ANNIS (Krause und Zeldes 2016) für den Einsatz in den Digital Humanities zu diskutieren und anzuwenden, da es bislang überwiegend für korpusbasierte und korpusgetriebene Studien sowie für das Auffinden von sprachlichen Belegen eingesetzt wird.

ANNIS wird seit 2009 als ein generisches webbasiertes Such- und Visualisierungstool für verschiedene Korpustypen und Annotationskonzepte in verschiedenen Kooperationen mit der Humboldt-Universität zu Berlin und der George-

town University und in mehreren Projekten entwickelt. Der Quellcode von ANNIS ist frei zugänglich veröffentlicht und bietet gleichzeitig eine Desktop- sowie Server-Installation. In ANNIS können Korpora mit Token-, Spannen-, Baum- und Pointingannotationen unabhängig von den einzelnen, jeweils korpusspezifischen Annotationsguidelines in ANNIS analysiert werden. ANNIS bietet weiterhin den Korpuserstellerinnen und -erstellern annotations- oder fachspezifische Visualisierungen für Korpora. Mit einer wiederum generischen und mächtigen Anfragesprache (ANNIS Query Language – AQL) können alle Korpora in ANNIS nach Annotationen und Kombinationen von Annotationen durchsucht werden. Weiterhin können die Suchergebnisse für bspw. weitere statistische Auswertungen exportiert werden. Jedes Korpus, jede Suchanfrage und jeder Beleg kann über einen permanenten Link stabil referenziert werden. Mit dem Konverterframework Pepper (Zipser und Romary 2010) werden Korpora, die in verschiedenen Formaten vorliegen können, in das ANNIS-Format überführt.

Repositorien wie das LAUDATIO-Repository (Odebrecht et al. 2015) ermöglichen einen Open Access Zugang zu verschiedensten historischen Korpora und stellen eine umfassende Korpusdokumentation (Odebrecht 2014) zur Verfügung, die eine Erschließung dieser heterogenen Datengrundlage unabhängig von den Korpuserstellerinnen und -erstellern ermöglicht. Damit wird eine Voraussetzung für die Wiederverwendung der historischen Korpora erfüllt. Für den Workshop werden aus LAUDATIO beispielhaft die Korpora „Referenzkorpus Altdeutsch“ (Donhauser 2015) und „RIDGES Herbology Korpus“ (Odebrecht et al. 2016) verwendet.

Das Referenzkorpus Altdeutsch ist ein historisches Mehrebenenkorpus der ganzen Sprachperiode des Althochdeutschen mit ca. 650.000 Wörtern (von den ersten Überlieferungen bis Mitte des 11. Jahrhunderts). Als Grundlage für die diplomatischen Transkription sind Editionen der jeweiligen Handschriften, die mit weiteren Annotationen zur Textstruktur sowie mit komplexen Wortartenannotation (Dipper et al. 2013), Annotation zu Flexionsklassen und Lemmatisierung versehen sind. Das RIDGES Korpus ist ein tief annotiertes Korpus mit Auszügen aus gedruckten Kräuterbüchern aus der Zeit zwischen 1487 und 1910, anhand derer die Entwicklung der deutschen Wissenschaftssprache auf vielen Ebenen untersucht wird. Die Drucke sind diplomatisch transkribiert (wo möglich, nach vorher digitalisierten oder durch OCR-Verfahren erstellte Vorlagen, vgl. Springmann und Lüdeling 2017). Die Daten sind mehrfach normalisiert und auf vielen Ebenen annotiert (unter anderem mit Wortart, Lemma, Informationen zu Kompositionstypen (Perlitz 2014), Dependenzsyntax, Informationen zur graphischen Struktur nach den TEI Guidelines). Dabei werden automatische und manuelle Annotationsverfahren und Prüfverfahren genutzt.

Um die eingangs formulierten Fragen adressieren zu können, wird der Workshop zwei Schwerpunkte enthalten. Der erste Schwerpunkt wird die Einführung in die Funktionen und Suchanfragesprache von ANNIS sowie die damit verbundene Vorstellung der zwei historischen Beispielkorpora umfassen. Wir wollen den Teilnehmerinnen und Teilnehmern die verschiedenen Analyse- und Visualisierungsmöglichkeiten online und hands-on vorstellen. Über die Vorstellung zweier historischer Korpora mit dem generischen ANNIS können bereits die Herausforderungen der heterogenen Datengrundlage in den digitalen Geisteswissenschaften für Analysetools herausgearbeitet werden.

Der zweite Schwerpunkt soll Raum für eine Diskussion mit den Teilnehmerinnen und Teilnehmern sowie auch die Möglichkeit geben, weitere Korpora in ANNIS – geleitet von den Forschungsinteressen der Teilnehmerinnen und Teilnehmern – zu durchsuchen. Mit diesem Workshop wollen wir uns gemeinsam mit den Teilnehmerinnen und Teilnehmern kritisch mit den Anforderungen an ein Analysetool für verschiedene Methoden zur Analyse und Visualisierung von historischen Korpora auseinandersetzen und prüfen, in wie weit ANNIS bereits einige dieser Anforderungen erfüllen kann. So wollen wir ANNIS in einem neuen Forschungskontext der Digital Humanities diskutieren und dabei neue Nutzerszenarien für die weitere Entwicklung erarbeiten.

Zeitplan:

- 1,5 Stunden Online Hands-on-Einführung in das Such- und Visualisierungstool, der Anfragesprache mit dem Referenzkorpus Altdeutsch und RIDGES Korpus
- 1,5 Stunden Teilnehmergeleitete Anfragen, weitere Korpora und Diskussion der Anforderungen

Technische Anforderungen:

- Für den Workshop werden ein Raum mit Beamer und Zugang zu eduroam, ggf. einzelne WLAN-Zugänge für die Teilnehmerinnen und Teilnehmer benötigt.
- Alle Teilnehmerinnen und Teilnehmer benötigen eigenes Notebook.

Teilnehmeranzahl:

- max. 30

Bibliographie

Moretti, Franco (2016): *Distant Reading*. Konstanz: Konstanz University Press.

Romary, Laurent / Zeldes, Amir / Zipser, Florian (2015): "<tiger2/>. Serialising the ISO SynAF syntactic object model", in: *Language Resources and Evaluation* 49 (1), S. 1–18. 10.1007/s10579-014-9288-x.

Archer, Dawn / Kytö, Merja / Baron, Alistair / Rayson, Paul (2015): "Guidelines for normalising Early Modern English corpora. Decisions and justifications", in: *ICAME Journal* (39), 5–24.

Bollmann, Marcel / Dipper, Stefanie / Krasselt, Julia / Petran, Florian (2012): "Manual and semi-automatic normalization of historical spelling. Case studies from Early New High German", in: *Proceedings of KONVENS 2012* 342–350. http://www.oegai.at/konvens2012/proceedings/51_bollmann12w/ [letzter Zugriff am 24.08.2016].

Borgmann, Christine L. (2012): "The conundrum of sharing research data", in: *Journal of the American Society for Information Science and Technology* 63 (6), 1059–1087. 10.2139/ssrn.1869155.

Burnard, Lou / Baumann, Sid (eds.) (2008): *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Oxford. http://www.tei-c.org/Guidelines/P5/ [zuletzt geprüft am 11.11.2015].

Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (2011): "Research Data Management", in: Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (eds.): *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen, 13–23.

Claridge, Claudia (2008): "Historical Corpora", in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook*, Bd. 1. 2. Berlin: De Gruyter (1), 242–259.

Coniglio, Marco / Donhauser, Karin / Schlachter, Eva / Rasskazova, Oxana / Odebrecht, Carolin / Wirth, Matthias / Miltenberger, Anke (2016): *Historisches Predigtenkorpus zum Nachfeld (HIPKON Version 1.0).* Technischer Bericht, Humboldt-Universität zu Berlin. 10.18452/13681.

Dipper, Stefanie (2005): "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation", in: *Proceedings of Berliner XML Tage 2005 (BXML 2005). Berlin,* 39–50.

Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): "HiTS. Ein Tagset für historische Sprachstufen des Deutschen“, in: Zinsmeister, Heike / Heid, Ulrich / Beck, Kathrin (eds.): Das Stuttgart-Tübingen Wortarten-Tagset. Stand und Perspektiven", in: *Journal for Language Technology and Computational Linguistics, 28(1),* 85–137.

Donhauser, Karin (2015): "Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten", in: Gippert, Jost / Gehrke, Ralf (eds.): *Historical Corpora.* Tübingen: Narr (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 5), 35–49.

Federico, Annette (2015): *Engagements with Literature. Engagements with Close Reading.* Florence: Routledge.

Gippert, Jost / Gehrke, Ralf (eds.) (2015): *Historical Corpora*. Tübingen: Narr (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 5).

Gooding, Paul / Terras, Melissa / Warwick, Claire (2013): “The myth of the new. Mass digitization, distant reading, and the future of the book”, in: *Literary and Linguistic Computing* 28 (4), 629–639. 10.1093/llc/fqt051.

Jurish, Bryan (2010): “More than Words: Using Token Context to Improve Canonicalization of Historical German”, in: *Journal for Language Technology and Computational Linguistics* 25 (1), 23–40.

Krasselt, Julia / Bollmann, Marcel / Dipper, Stefanie / Petran, Florian (2015): *Guidelines für die Normalisierung historischer deutscher Texte*. Bochumer Linguistische Arbeitsberichte, 15. urn:nbn:de:hebis:30:3-419680.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3. “A new architecture for generic corpus query and visualization”, in: *Digital Scholarship in the Humanities* 31 (1), 118–139. 10.1093/llc/fqu057.

Kupietz, Marc / Geyken, Alexander (eds.) (2016): “Corpus Linguistic Software Tools”, in: *Journal for Language Technology and Computational Linguistics* 31(1).

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

Lüdeling, Anke / Zeldes, Amir (2007): “Three Views on Corpora. Corpus Linguistics, Literary Computing, and Computational Linguistics”, in: *Jahrbuch für Computerphilologie* (9), 149–178.

McEnery, Tony / Hardie, Andrew (2012): *Corpus Linguistics. Method, Theory and Practice*. Cambride [u.a.]: Cambridge University Press (Cambridge Textbooks in Linguistics).

Odebrecht, Carolin (2014): "Modeling Linguistic Research Data for a Repository for Historical Corpora", in: *Digital Humanities Conference Abstracts*. Lausanne 284–285.

Odebrecht, Carolin / Belz, Malte / Zeldes, Amir / Lüdeling, Anke / Krause, Thomas

(2017): „RIDGES Herbology. Designing a Diachronic Multi-Layer Corpus”, in: *Language Resources and Evaluation 51 (2)* First Online 2016, 695-725 . 10.1007/s10579-016-9374-3.

Odebrecht, Carolin / Krause, Thomas / Lüdeling, Anke (2015): "Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository", *37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, DGfS-CL Poster Session, Leipzig. http://conference.uni-leipzig.de/dgfs2015/fileadmin/zusatzdokumente/dgfs-tagung-2015-final.pdf [letzter Zugriff am 17.08.2017].

Perlitz, Laura (2014): *Konkurrenz zwischen Wortbildung und Syntax. Historische Entwicklung von Benennung.* Bachelorarbeit. Humboldt-Universität zu Berlin, Berlin. 10.18452/14232

Rissanen, Matti (2008): "Corpus Linguistics and Historical Linguistics", in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook*. 2 Bände. Berlin: De Gruyter (1), 53–68.

Schmidt, Thomas / Wörner, Kai (2009): "EXMARaLDA. Creating, analysing and sharing spoken language corpora for pragmatic research", in: *Pragmatics* 19 (4), 565–582.

Simanowski, Roberto (2011): *Digital Art and Meaning. Reading Kinetic Poetry, Text Machines, Mapping Art, and Interactive Installations* (Electronic Mediations). Minnesota: University of Minnesota Press.

Simons, Gary / Bird, Steven (2008): "Toward a global infrastructure for the sustainability of language resources", in: *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*. Cebu City, 87–100.

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 [letzter Zugriff am 12.09.2017].

Wittenburg, Peter / Brugmann, Hennie / Russel, Albert / Klassmann, Alex / Sloetjes, Han (2006): "ELAN. A Professional Framework for Multimodality Research", in: *Proceeedings of LREC. Language Resources and Evaluation Conference*. Genoa 1556–1559. http://www.lrec-conf.org/proceedings/lrec2006/ [letzter Zugriff am 23.12.2016].

Zipser, Florian / Romary, Laurent (2010): "A model oriented approach to the mapping of annotation formats using standards", in: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. http://hal.archives-ouvertes.fr/inria-00527799/en/ [letzter Zugriff am 12.11.2014].

Wikidata: Nutzungsmöglichkeiten und Anwendungsbeispiele für den Bereich Digital Cultural Heritage

Müller-Birn, Claudia
clmb@inf.fu-berlin.de
Freie Universität Berlin, Deutschland

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

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

Wübbena, Thorsten
wuebbena@kunst.uni-frankfurt.de
Goethe-Universität Frankfurt

Einführung

Das im Jahr 2012 gegründete Projekt Wikidata ist ein Schwesterprojekt der Wikipedia, welches strukturierte Daten verwaltet, die in Wikipedia oder in anderen Wikimedia-Projekten verwendet werden können. Der Wikidata-Ansatz basiert auf einer Reihe von Designentscheidungen, wie der offenen Bearbeitung, der Pluralität (Daten aus unterschiedlichen Quellen mit unterschiedlicher Aussage sind erlaubt) und der Mehrsprachigkeit (Vrandečić & Krötzsch, 2014).

Die strukturierten Daten sind nach konkreten Entitäten (Items) organisiert, die mit Aussagen (Statements) beschrieben werden. Jede Entität hat ein sprachspezifisches Label. Eine Entität, beispielsweise Q12418 besitzt das Label „Mona Lisa“. Sogenannte Eigenschaften (Properties) erlauben dann die Beschreibung der Entitäten mit Aussagen. Q12418 (Subjekt) ist eine Instanz „instance-of” (P31 Property) der Entität „Painting“ (Q3305213 Objekt) und sagt aus, dass es sich hier um ein Gemälde handelt. Mit dieser Aussagestruktur bestehend aus Subjekt-Prädikat-Objekt bietet Wikidata maximale Flexibilität, da Wi-

kidata erlaubt, neue Eigenschaften zu erstellen und Aussagen über sie zu machen. Durch die zusätzliche (eigentlich verpflichtende) Angabe von Referenzen kann Wikidata die Funktion einer sekundären, verifizierbaren Wissensbasis übernehmen. Durch die Angabe von Referenzen werden in Wikidata nicht in erster Linie Tatsachen über die Welt gespeichert, sondern vielmehr Verlinkungen zu Quellen erstellt. Die Umstände, dass derzeit noch häufig de facto Referenzen fehlen und dass grundsätzlich Personen Falscheingaben machen werden können, sind kritisch zu berücksichtigen, jedoch nicht Gegenstand des Workshops.

Im September 2017 umfasste Wikidata Informationen von über 29 Millionen Entitäten, die mit Hilfe von über 175.000 registrierten Nutzern erstellt wurden. Ein wesentlicher Mehrwert von Wikidata liegt in der Nutzung der strukturierten Daten, denn diese können von Menschen und Maschinen gleichermaßen „verstanden“ werden (Müller-Birn et al., 2015). Das wurde dadurch ermöglicht, dass das bestehende Faktenwissen in terminologische Wissen überführt wurde (Erxleben et al., 2014). Somit können einerseits die Daten über ein User Interface von Menschen bearbeitet sowie abgefragt und andererseits (unter anderem) über den Wikidata SPARQL Query Service (https://query.wikidata.org/), von Softwareanwendungen verwendet werden. Darüber hinaus bieten Anwendungen wie SQID die Möglichkeit, das terminologische Wissen zu explorieren. Wikidata findet mittlerweile in vielfältigen Bereichen Anwendung: wie beispielsweise in interaktiven Abfrageansichten zu bestimmten Themen (z. B. das Oscar-Portal der Zeitung FAZ online), aber auch zur Unterstützung von Q&A-Systeme, wie beispielsweise Apples Siri Suchmaschine.

Wikidata im Bereich Digital Cultural Heritage

Solche oder ähnliche Anwendungen sind aber ebenfalls im Bereich der Geistes- und Kulturwissenschaften denkbar, wie die Beiträge zu der diesjährigen WikidataCon (https://www.wikidata.org/wiki/Wikidata:WikidataCon_2017) zeigen. Die wachsende Vernetzung zwischen Datenrepositorien und der zunehmende Einsatz entsprechender Datenmodelle (z.B. Graphen-Datenbanken, LOD) sowie Normdaten im Cultural Heritage-Bereich machen Wikidata zu einer interessanten Infrastruktur, die auch in diesem Gebiet eingesetzt werden könnte.

Kulturhistorische Datenbanken können beispielsweise die in ihnen enthaltenen Entitäten (historische Personen, Objekte, Orte etc.) auf Wikidata referenzieren, neue Zusammenhänge in Form von Aussagen an Wikidata zurückspielen und bestehende Aussagen aus Wikidata verwenden. Im Fall von Personen-Datensätzen enthält Wikidata beispielsweise Bezeichner, wie z.B. VIAF, GND oder der ULAN ID (Union List of Artist Names: Identifier von der Getty Union List von Künstlernamen). Wikidata stellt hier bereits ein internationales Bindeglied dar, da die überwiegend national (z.B. durch nationale Bibliotheken wie die DNB, BNF oder Library of Congress) oder disziplinär (z.B. ULAN für die Kunstgeschichte) organisierten Normdatenrepositorien nicht alle Bezeichner abdecken bzw. auf diese verweisen. Im Bereich der Werke (Bauwerke, Kunstwerke und Artefakte aller Art) könnte diese übergreifende Rolle von Wikidata noch bedeutender werden. Anders als bei Personen existieren hier noch immer wenige Normdatenrepositorien. Zudem ist die Zahl der Artefakte wesentlich höher und ihre Definition viel weniger eindeutig (es kann von einer Zahl von mehreren -zig Millionen „relevanter“ Kunstwerke weltweit ausgegangen werden; die Zahl der für die verschiedenen kulturhistorischen Disziplinen relevanten Artefakte ist um ein Vielfaches höher). Es ist unwahrscheinlich, dass die betreffenden GLAM-Institutionen, d.h. kulturellen Institutionen und Gedächtnisorganisationen, hier in absehbarer Zeit ein ausreichendes und vor allem standardisiertes Datenmaterial bereitstellen können. Auch dort wo sich einschlägige Institutionen der Aufgabe angenommen haben, bleibt das entweder auf eine nationale Ebene beschränkt (z.B. mit den Datenbanken Merimee oder Joconde in Frankreich, oder dem RKD in den Niederlanden), oder droht unausgewogen und fragmentarisch zu bleiben (z.B. CONA, das Cultural Objects Name Authority des Getty Research Institute). Das Deutsche Dokumentationszentrum für Kunstgeschichte, Foto Marburg, hat zwar vielfach die Bedeutung von Werknormdaten unterstrichen (Locher/Warnke 2013), jedoch bislang keinen Vorschlag für deren Bereitstellung gemacht. Nach heutigem Ermessen kann wohl auch nicht davon ausgegangen werden, dass es möglich oder sinnvoll ist, ein vollständiges Referenzrepositorium aller Bau- und Kunstwerke anzustreben. Selbst die Europeana (mit der Europeana ID) als zusammenfassender Aggregator ist hier noch lückenhaft und bietet vor allem keine Möglichkeit, weiteres Material direkt hinzuzufügen.

Wikidata kann hier eine bestehende Lücke schließen und die Funktion eines Drehkreuzes (Hub) erhalten, von dem aus auf weitere digitale

Repräsentationen der Werke wie auch auf zugehörige Wissensbestände verwiesen wird. Freilich muss hierzu der Bestand in Wikidata – parallel zu den genannten Repositorien – ständig weiter ausgebaut werden. Projekte wie beispielsweise die „Sum of all paintings“ treiben diese Aufgabe systematisch voran (Poulter, 2017).

Für die praktische Arbeit im Bereich des Kulturerbes und der Kulturgeschichte bietet Wikidata nicht zuletzt den Vorteil der freien Zugänglichkeit und der verwendeten offenen technischen Standards. Selbst kleinste und sehr spezialisierte Projekte können so fehlende Gegenstände in Wikidata selbst ergänzen und sich damit zugleich in den globalen Wissensbestand einschreiben. Wohin sich Wikidata entwickeln wird, ist nicht absehbar. Es werden sich in jedem Fall viele weiteren Anwendungsgebiete herausbilden und ein Anwendungsfall könnte eine offene Metadatenbank für das Weltkulturerbe sein. Die bereits erwähnte Einschränkung der Manipulierbarkeit ist hierbei einerseits kritisch zu sehen, jedoch steht zu erwarten, dass das Zusammenspiel zahlreicher Statements einschließlich zugehöriger Referenzquellen sowie die Referenzierung auf andere Normdaten einen sich selbst bestätigenden oder für automatisierte Überprüfungen immer besser geeigneten Datenbestand erzeugen wird.

Ausrichtung des Workshops

Das Ziel des Workshops liegt darin, eine fundierte Einführung in Wikidata zu geben (Weitere Informationen zum Ablauf und Materialien unter https://blogs.fu-berlin.de/dhd2018_wikidata-workshop/ verfügbar). Im Rahmen dieser Einführung wird den Teilnehmer_innen der Semantic Web Technologie-Stack (Triple, RDF, OWL, etc.) vermittelt und Wikidata in bestehende andere Initiativen (z.B. DBpedia) eingeordnet. In einem praxisorientierten Tutorial wird eine Einführung in Wikidatas Abfragesprache SPARQL gegeben. Die Teilnehmer_innen sollten nach dem Workshop ein grundlegendes Verständnis zu den Möglichkeiten des Einsatzes von Wikidata in ihrer Forschungspraxis erlangt und erste Möglichkeiten zum selbstständigen Einsatz kennengelernt haben. Dies wird sichergestellt, indem anhand mehrerer Fallbeispiele aus dem Bereich der kunsthistorischen Bilddatenbanken – unter anderem der Historischen Glasdias-Sammlung des Instituts für Kunst- und Bildgeschichte der HU Berlin – praktische Ansätze zur Nutzung von Wikidata in der geisteswissenschaftlichen Forschung aufgezeigt werden. Dabei werden Probleme diskutiert und exemplarische Lösungen durchgespielt werden. Hierzu gehören Fragen, wie vorhandene Datensätze zu Kulturgütern (automatisiert) aufgefunden, wie neue Entitäten möglichst effektiv ohne die Erzeugung von Dubletten angelegt und bestehende Aussagen abgefragt werden können. Zu den exemplarischen Lösungen gehören unter anderem ein Web-Service für die Ermittlung von Wikidata-IDs unter Verwendung der Google-Bildersuche und ein Skript zur Extraktion von Daten unter Nutzung der GND-Nummern. Nachfolgend werden die bisher geplanten Beispielanwendungen kurz eingeführt.

Beispielanwendung 1: Die Datenbank der digitalisierten Glasdias des Instituts für Kunst- und Bildgeschichte der Humboldt-Universität in Berlin. (Georg Schelbert, Humboldt-Universität zu Berlin; Martin Raspe, MPI f. Kunstgeschichte in Rom)

Fotografien oder Dias, die Kunstwerke zeigen, werden normalerweise erfasst, indem – neben einigen Metadaten zum Bild – das abgebildete Kunstwerk beschrieben wird. Dies erfolgt oft nach bestimmten Standards, die dann Pflichtfelder und Terminologien vorgeben (Harpring 2010). Allerdings bedeutet die Beschreibung der auf Dias befindlichen Kunstwerke einen im Grunde überflüssigen Aufwand, da es sich zumeist um bekannte oder sogar berühmte Kunstwerke handelt. Bei diesen Kunstwerken ist anzunehmen, dass diese bereits beschrieben wurden. Daher wurde entschieden, die Fotografien oder Dias mit Normdaten-Identifikatoren aus Wikidata zu versehen. So muss z.B. zumindest ein Teil der Beschriftung entziffert oder das Bildmotiv erkannt werden, um dann in Wikidata gesucht werden zu können.

Um diesen Prozess zu beschleunigen bzw. zuverlässiger zu machen (indem zu eingegebenen Namen der Objekte zugleich Bilder zur Überprüfung der Übereinstimmungen angezeigt werden) wurde prototypisch ein Webservice entwickelt, der (mit einer entsprechend konfigurierten Google „custom search engine“) Bilder sucht. Er sucht dabei mit Vorrang bei Wikipedia, und ermittelt dazu die passenden Wikidata-IDs. Dieser Service soll im Rahmen des Workshops als Ausgangspunkt für praktische Arbeiten und weitere Diskussionen dienen.

Darüber hinaus stellt sich die Frage, wie Wissen, das aus dem Umgang mit den Bildern – etwa auf der Basis der Beschriftungen – entsteht, jedoch noch nicht Wikidata enthalten ist, in Wikidata eingespielt werden kann. Hierfür werden Möglichkeiten (potentielle Workflows und Ansätze) im Rahmen des Workshops diskutiert.

Beispielanwendung 2: Die ConedaKOR-Datenbank im Kunsthistorischen Institut in Frankfurt (Thorsten Wübbena, DFK Paris / Kunstgeschichtliches Institut der Univ. Frankfurt)

Aufgrund ihres Kenntnisstands sind Studierende in einer frühen Phase des Kunstgeschichtsstudiums häufig nicht in der Lage, über die textuelle Suche in einem Bilddatenbanksystem zu einem befriedigenden Ergebnis zu kommen. Sie wissen schlichtweg nicht, wonach sie suchen müssen bzw. welche Suchbegriffe sie verwenden müssen, um das entsprechende Resultat zu erhalten. Beim Einsatz von ConedaKOR hilft hier sicher die Option, die Daten visuell zu explorieren. So reicht es zum Beispiel, die aufbewahrende Institution zu kennen, um zu einem bestimmten Bildwerk zu gelangen. Mittels Relationen werden die übergeordneten Entitäten durch die diversen Räume bis hin zu den Werken sichtbar. Diese Möglichkeit stellt verständlich keine spezifische Besonderheit dar, macht aber deutlich, dass nur eine hohe Datendichte dieses Vorgehen wirklich sinnvoll macht. Um nun die eingeschränkten personellen Ressourcen im Frankfurter Kunstgeschichtlichen Institut zu unterstützen, wurde eine semi-automatisierte Lösung im Zusammenspiel mit dem Datenbestand aus Wikidata konzipiert. Da in der Frankfurter ConedaKOR-Installation die GND-ID für Personen und Orte abgelegt wird und genau diese Identifier auch in Wikidata vorhanden sind, werden die Relationen zwischen den – mit einer GND-ID versehenen – Entitäten in Wikidata mit denen in der Frankfurter KOR-Installation verglichen. Jede dort nicht vorhandene Relation wird eingetragen und mit einem entsprechenden Label („Aussage aus Wikidata übernommen“) versehen. Neben diesem Verfahren wird mit einer speziell entwickelten Web-Extension noch eine weitere Möglichkeit der systemübergreifenden Nutzung von Wikidata vorgestellt. Hierbei wird zum einen auf jeder Website, die eine Wikidata-ID enthält im eigenen Datenbanksystem nachgesehen, ob Bildmaterial zum Objekt vorhanden ist, welches sich die Nutzer_innen dann direkt anzeigen lassen können. Ebenso kann auf diesem Weg neues Bildmaterial in das eigene System hochgeladen werden. Damit ist die Nutzung der Datenbasis in Wikidata möglich, fehlende Objekte können direkt dort ergänzt werden und die rechtlich häufig schwierige Abbildungssituation wird weiterhin im System der Nutzer_innen geklärt. Auch die Übernahme von Wikidata-Inhalten in die eigene Datenbank ist auf diesem Weg möglich.

Beitragende

Claudia Müller-Birn: Prof. Dr. rer. nat. Claudia Müller-Birn erforscht Fragestellungen im Bereich der Computer-Supported Cooperative Work und Social Computing. Ihr Ziel ist es, basierend auf einem besseren Verständnis der bestehenden Wissensprozesse neuartige Interaktionskonzepte zu entwickeln. Daher verbindet sie die Bereiche der Datenanalyse von Online Communities (z.B. Wikidata) mit dem Design von Kollaborationssoftware (z.B. Annotationssoftware neonion) eng miteinander. Ein zentraler Anwendungsbereich ihrer Forschung ist der Bereich der Scientific Collaboration, in welchen sie unter anderem durch Einsatz von Linked Data für eine nachhaltigere Nutzung von Forschungsdaten eintritt.

Georg Schelbert: Dr. phil. Georg Schelbert interessiert sich als Architektur- und Kunsthistoriker insbesondere für die Kunstgeschichte der Stadt Rom in der frühen Neuzeit, sowie für ihre Dokumentation in Karten und Veduten (Urban iconography). Ferner beschäftigt er sich als Leiter der Mediathek des Instituts für Kunst- und Bildgeschichte mit der Geschichte kunsthistorischer Bildmedien (Fotografien, Diapositive). Innerhalb der Digital Humanities beschäftigt er sich schwerpunktmäßig mit kulturhistorischen Bild- und Forschungsdatenbanken (Datenmodelle, linked data, Wissensrepräsentation, open heritage, Chronotopographie).

Martin Raspe: Dr. phil. Martin Raspe arbeitet generell über italienische Kunst und Architektur sowie niederländische Kunst der frühen Neuzeit. Außerdem beschäftigt er sich mit der Kunsttopographie der Stadt Rom von der Antike bis zur Gegenwart. In den Digital Humanities interessieren ihn Bild- und Forschungsdatenbanken (Datenmodelle, Graphdatenbanken, Wissensrepräsentation), Verfahren und Tools.

Thorsten Wübbena: Thorsten Wübbena M.A. arbeitet über Themen der bildenden Kunst und deren mediale Rezeption (insbesondere in Musikvideos). Er arbeitet an Projekten zum Einsatz von Informationstechnologie in der kunstgeschichtlichen Forschung. Innerhalb der Digitalen Kunstgeschichte interessieren ihn insbesondere kulturhistorische Bild- und Forschungsdatenbanken (Datenmodelle, Wissensrepräsentation).

Bibliographie

Erxleben, Fredo / Günther, Michael / Krötzsch, Markus / Mendez, Julian / Vrandeǎić, Denny (2014): *“Introducing Wikidata to the Linked Data Web”,* in: *Proceedings of the 13th Interna-*

tional Semantic Web Conference. New York: Springer-Verlag 50-65.

Harpring, Patricia (2010): *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*, Los Angeles: Getty Publications.

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

Locher, Hubert / Warnke, Martin (2014): Ergebnisse des Round Table „Kritische Massen – Zur Anschlussfähigkeit digitaler Bildbestände an die aktuelle Kunsthistorische Forschung“ , in: *Kunstgeschichte. Open Peer Reviewed Journal* 07 Okt 2014 http://www.kunstgeschichte-ejournal.net/412/ [letzter Zugriff 25. September 2017]

Ohlig, Jens / Schelbert, Georg (2017): „Datenpartnerschaften mit Wikidata – Projekt Durchblick“, *Wikimedia DE-Blog*, 21. Aug. 2017 https://blog.wikimedia.de/2017/08/21/datenpartnerschaften-mit-wikidata-projekt-durchblick/ [letzter Zugriff 25. September 2017]

Müller-Birn, Claudia / Karran, Benjamin / Lehmann, Janette / Luczak-Rösch, Markus (2015): “ *Peer-production system or collaborative ontology development effort: what is Wikidata?”* in *Proceedings of OpenSym 2015 - Conference on Open Collaboration.* San Francisco: ACM.

Patton, Glenn E. (2010): *Funktionale Anforderungen an Normdaten: Ein konzeptionelles Modell* (IFLA Working Group on Functional Requirements and Numbering of Authority Records – FRANAR) Berlin, New York: De Gruyter.

Poulter, Martin (2017): “Wikidata – the new hub for cultural heritage”, *Wikimedia UK-Blog*, 20. Jan. 2017 https://blog.wikimedia.org.uk/2017/01/wikidata-the-new-hub-for-cultural-heritage/ [letzter Zugriff 25. September 2017]

Vrandečić, Denny / Krötzsch, Markus (2014): “Wikidata: a free collaborative knowledge base”, in: Commun. ACM 57(10), 78–85.

Woitas, Kathi: „Bibliografische Daten, Normdaten und Metadaten im Semantic Web – Konzepte der Bibliografischen Kontrolle im Wandel“, in: *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft*, 10.05.2013 (urn:nbn:de:kobv:11-100209272 [letzter Zugriff 25. September 2017]

Workshop eComparatio: Textvergleich und digitaler Apparat

Schubert, Charlotte
schubert@uni-leipzig.de
Universität Leipzig\Historisches Seminar, Deutschland

Kahl, Hannes
hannes.kahl@uni-leipzig.de
Universität Leipzig\Historisches Seminar, Deutschland

Meins, Friedrich
friedrich_meins@uni-leipzig.de
Universität Leipzig\Historisches Seminar, Deutschland

Bräckel, Oliver
oliver.braeckel@uni-leipzig.de
Universität Leipzig\Historisches Seminar, Deutschland

I. Motivation

In den textbasierenden Geisteswissenschaften ist eine optimale Beschaffenheit der zu Grunde liegenden Texte eine wesentliche Bedingung für wissenschaftliches Arbeiten. Wie sehr sich die Bedeutung eines Textes schon durch scheinbar minimale Unterschiede wie die der Interpunktion verschieben kann, hat sich einer breiteren Öffentlichkeit zuletzt in der Diskussion über einen Punkt in einer Abschrift der amerikanischen Unabhängigkeitserklärung gezeigt (The Atlantic 7/2014).

Während die historische Diskussion über digitale Editionen vielfach mit dokumentarischen Editionstypen bzw. der Frage nach dem Dokumentcharakter der Grundlagen einer Edition (Manuskripte etc.) und der Frage nach der Essenz des Textbegriffes beschäftigt ist (P. Sahle 2013), ergibt sich insbesondere für diejenigen historischen und altertumswissenschaftlichen Disziplinen, die weiterhin auf die klassische Form kritischer Texteditionen angewiesen sind, ein anderes Problem: Da sie traditionell in hohem Maße mit Texten beschäftigt sind, die notwendig als Interpretationen und Rekonstruktionen gelten müssen (West 1973,32), ist hier vor allem die Frage nach

der Beschaffenheit und Begründung der zum Teil massiven editorischen Eingriffe nicht nur in die Lesart, sondern auch in den Umfang und die Zuschreibung von Texten, etwa im Falle sogenannter „Fragmente“, von vorrangiger Bedeutung. Der Workshop soll in diese allgemeine Problematik einführen und sich im Hinblick auf das Thema der Konferenz dem Bereich Kritik der digitalen Geisteswissenschaften (traditionelle Fächer und DH) zuordnen. Den Teilnehmenden soll dies praktisch anhand der Anwendung der Software eCOMPARATIO vermittelt werden. eCOMPARATIO bietet eine einfache Möglichkeit der Kollationierung verschiedener Varianten eines Textes und ermöglicht eine digitale, auf dem automatischen Textvergleich beruhende Form des kritischen Apparates.

Dazu sind einzelne Use Cases ausgearbeitet worden (anhand der Fragmentsammlung der Vorsokratiker von Diels/Kranz [griechisch], der Res Gestae des Augustus [lateinisch], des Genfer Gelöbnisses [deutsch], der Gettysburg Address von Abraham Lincoln [englisch]), anhand derer den Teilnehmern die Funktionalitäten demonstriert werden.

II. Die Software

Im Projekt eCOMPARATIO, das von 2014 bis 2016 von der DFG gefördert wurde und in Leipzig am Lehrstuhl für Alte Geschichte in Kooperation mit dem Center of E-Humanities in History and Social Sciences (ICE) am Max-Weber-Kolleg für kultur- und sozialwissenschaftliche Studien durchgeführt wurde, ist vor dem Hintergrund dieser Problematik ein einfach zu bedienendes Tool für den Vergleich prinzipiell beliebig vieler und beliebig langer digitalisierter Editionen vorrangig griechischer und lateinischer, technisch gesehen aber auch sämtlicher anderer in einem UNICODE-Format zugänglicher Texte entstanden. Der Textvergleich arbeitet auf der Basis der Identifikation von Ungleichheiten. Hierbei reicht die Spanne der programmiertechnisch unterscheidbaren Differenzen von Ungleichheiten innerhalb von Wörtern und Buchstabenfolgen bis hin zu vertauschten Passagen. Im Unterschied zu anderen Vergleichsprogrammen benötigen Anwender weder eine Installation von Python, eine Levenshtein Bibliothek oder ein Java-Plugin, sondern erhalten eine für die Anwender ohne Vorkenntnisse sofort im Browser nutzbare und mit Copy-and-Paste einfach zu bedienende Oberfläche. Auch Fallbeispiele, eine Textdokumentation sowie Videoanleitungen stehen zur Verfügung. In zwei derzeit (2016-17) von der DFG und der Andrew W. Mellon Foundation geförderten Projekten an der Universität Leipzig in Kooperation mit Christopher Blackwell von der Furman University in Greenville, SC/ USA, erfolgt eine Einbindung des Vergleichstools in ein Interface (zu dem Protokoll Canonical Text Services (CTS) als Teil der CITE Architecture: http://cite-architecture.github.io/cts/), dessen Ziel nicht nur die Bereitstellung möglichst vieler digitalisierter Editionen einzelner Texte, sondern damit auch erweiterte Möglichkeiten des Textvergleiches, der Suche nach Parallelstellen und weiterer Formen des sog. „Text-Mining“ ist.

Die Texteingabemaske

Die eigenständige Version der Vergleichssoftware (unter http://ecomparatio.net/~khk/instanzen/ecompp/ ist ein Beispiel frei zum Gebrauch bereitgestellt) ermöglicht weiterhin eine browserbasierte oder auch offline verwendbare Anwendung für Texte nach dem Copy-and-Paste-Prinzip: Hier können beliebig viele Versionen eines Textes eingefügt werden.

Die Darstellung

Das Tool ermöglicht verschiedene Ausgaben des Vergleiches:

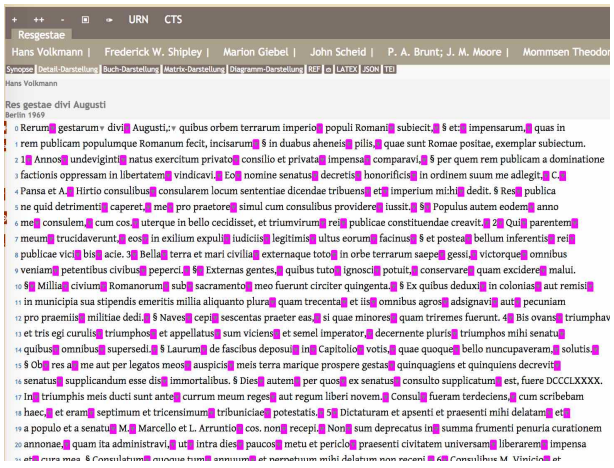


Abbildung 1, Detail-Vergleich

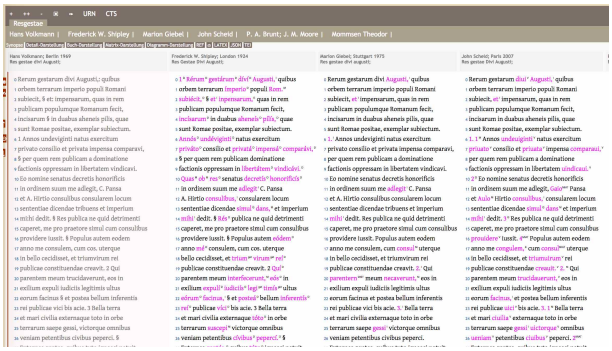


Abbildung 2, Synopse

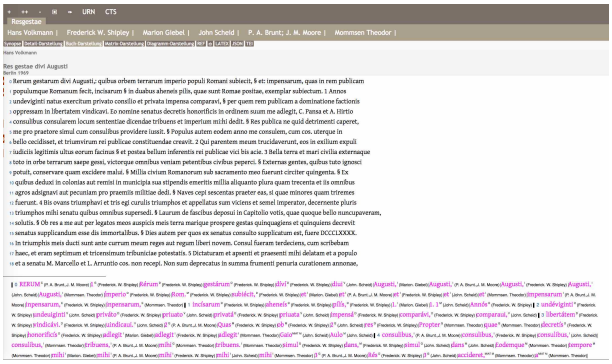


Abbildung 3, Buch-Darstellung mit digitalem Apparat zum Textvergleich

Ausgabe/Export der Ergebnisse

Die Ausgabe der Ergebnisse kann zur weiteren Integration in den Arbeitsprozess als TEI XML oder LaTeX Code erfolgen. Will man die Vergleichsdaten abfragen, so steht ein JSON Interface zur Verfügung, über das weitere Software angebunden werden kann.

III. Ziele und Zielgruppe

Ziel des Workshops ist eine Einführung in die Anwendung der eCOMPARATIO Vergleichssoftware, auch in Abgrenzung zu bereits verfügbarer Software (collateX, Juxta) mit ähnlichen Anwendungsbereichen. Dazu soll zunächst eine Einführung in die Relevanz der Frage nach scheinbar marginalen textlichen Unterschieden anhand prägnanter Beispiele aus verschiedenen historischen Epochen und in verschiedenen wissenschaftlichen Diskurssprachen gegeben werden (anhand griechischer, lateinischer, englischer und deutscher Texte).

Darüber hinaus soll den Teilnehmenden die Möglichkeit gegeben werden, auch Texte aus den eigenen Disziplinen oder beliebige im Internet verfügbare Versionen von Texten selbst miteinander zu vergleichen und sich so mit den Funktionen der Software vertraut zu machen.

Zielgruppe des Workshops sind prinzipiell alle Interessierten aus dem Bereich der textbasierten Geisteswissenschaften. Dabei sind diejenigen, die, wie oben angesprochen, vor allem mit der Vielzahl digitalisierter Editionen arbeiten, ebenso angesprochen wie solche, die selbst an der Erstellung von Editionen etc. arbeiten, und für die eCOMPARATIO ein hilfreiches Mittel bei der Sichtung und Kollationierung von Textzeugnissen sein kann.

Besondere technische Kenntnisse sind nicht erforderlich, da sich eCOMPARATIO bewusst an Anwender richtet, die entweder selbst an einer kritischen Edition arbeiten oder auf der Grundlage mehrerer kritischer Editionen wissenschaftliche Fragestellungen verfolgen.

IV. Ablauf und Teilnehmerzahl

Im Workshop soll die browsergestützte Version zur Anwendung kommen, ein Download der Software wird aber auch möglich sein. Neben der bereits beschriebenen Einführung in die wissenschaftliche Grundproblematik erfolgt zunächst eine kurze Präsentation eigener Ergebnisse im Zuge der Forschung mit eCOMPARATIO.

Hauptsächlich soll im praktischen Teil den Teilnehmenden die Möglichkeit gegeben werden, auf ihren eigenen Rechnern in der browsergestützten Version selbst Vergleiche der für sie relevanten Texte vorzunehmen, die mitgebracht werden sollten oder vor Ort aus Onlinedatenbanken heruntergeladen werden können.

Während des Workshops soll dann von Seiten der Organisatoren auf eventuelle individuelle Probleme und Schwierigkeiten eingegangen werden. Die Organisatoren des Workshops erhoffen sich hiervon einen eigenen Erkenntnisgewinn auch im Hinblick auf die Anwendungsmöglichkeiten und notwendige Ergänzungen im Hinblick auf die Arbeit gerade mit nicht-lateinischen oder nicht-griechischen Texten.

V. Ablauf

Die TeilnehmerInnen erhalten vorab eine detaillierte Anleitung (Handbuch eCOMPARATIO).

Im eigentlichen Workshop werden die jeweiligen Arbeitsschritte von einem der Organisatoren live vorgeführt (dafür wird ein leistungsstarker Beamer benötigt). Die konkreten Inhalte orientie-

ren sich dabei an den bisher von den Organisatoren ausgearbeiteten Use Cases (s.o.), die von den Organisatoren präsentiert werden.

Während des Workshops werden wir bei auftretenden Fragen und Problemen den Teilnehmenden helfend zur Seite stehen, da sie auch die Möglichkeit haben sollen, anhand eigener Texte zu arbeiten. Um eine möglichst gute Betreuung der TeilnehmerInnen gewährleisten zu können, sollte die Teilnehmerzahl 25-30 nicht überschreiten.

VI. Organisatoren

Charlotte Schubert ist Althistorikerin, hat zu Themen der Mentalitätsgeschichte, Medizin- und Wissenschaftsgeschichte sowie zu verschiedenen Bereichen der griechischen Geschichte gearbeitet; seit 2006 verantwortliche Koordinatorin in verschiedenen DH-Projekten, die vom BMBF (eAQUA, eXChange), der DFG (eCOMPARATIO, CTS, Etablierung eines Open Access Online eJournals: Digital Classics Online), der VolkswagenStiftung (Digital Plato) und der Andrew W. Mellon Foundation (CTS) gefördert wurden und werden.

Hannes Kahl ist Informatiker, Berufserfahrung aus diversen DH-Projekten, Entwickler von eCOMPARATIO, arbeitet an der Weiterentwicklung von CTS und an einer Dissertation zu dem Thema „Form und Formalisierung – mit Anwendung innerhalb automatischer Ermittlung von Buchstabenwerten aus digitalen Abbildungen griechischer Lettern innerhalb wissenschaftlicher Editionen sowie deren digitaler Formatierung“ (Betreuer: Prof. Ch. Schubert, Alte Geschichte/Universität Leipzig/ Prof. O. Arnold, Informatik/FH Erfurt).

Friedrich Meins ist Althistoriker und hat eine Dissertation zum Thema „Literarische Kritik, rhetorische Theorie und historische Methode bei Dionysios von Halikarnassos“ geschrieben. Im Bereich der eHumanities hat er in den Projekten „eAQUA“ und „eAQUA Dissemination“ an der Uni Leipzig sowie im Projekt „eCOMPARATIO“ am ICE der Universität Erfurt mitgearbeitet. Derzeit arbeitet er im von der Andrew W. Mellon Foundation geförderten Kooperationsprojekt der Universität Leipzig und der Furman University (Greenville, SC/ USA) „Annotating and Editing With Canonical Text Services (CTS)“.

Oliver Bräckel ist Althistoriker und arbeitet an einer Dissertation zu Politischen Flüchtlingen im Römischen Reich. Im Bereich der eHumanities hat er im Projekt „eCOMPARATIO“ am ICE der Universität Erfurt sowie im Projekt eXChange an der Uni Leipzig mitgearbeitet. Derzeit arbeitet er im von der Andrew W. Mellon Foundation geförderten Kooperationsprojekt der Universität Leipzig und der Furman University (Greenville, SC/ USA) „Annotating and Editing With Canonical Text Services (CTS)“.

Bibliographie

A. Olheiser, Have We Been Reading the Declaration of Independence All Wrong?, https://www.theatlantic.com/entertainment/archive/2014/07/typo-could-mean-weve-been-reading-the-declaration-of-independence-all-wrong/373915/ (letzter Zugang 7.7.2017).

P. Sahle, Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels, Norderstedt 2013, 3 Bde. (Schriften des Instituts für Dokumentologie und Editorik Bd. 7).

M.L. West, Textual Criticism and Editorial Technique, Stuttgart 1973.

Links:

The Atlantic: https://www.theatlantic.com/entertain ment/archive/2014/07/typo-could-mean-weve-been-reading-the-declaration-of-independence-all-wrong/373915/ (letzter Zugang 7.7.2017)

http://www.eaqua.net

http://www.ecomparatio.net/

http://www.ecomparatio.net/~khk/instanzen/ecompp/ (letzter Zugang17.9.2017)

http://digital-plato.org

http://digital-classics-online.eu

http://cite-architecture.github.io/cts/

Zur Zukunft der Digitalen Briefedition – kooperative Lösungen im kulturwissenschaftlichen Forschungsdaten-management

Strobel, Jochen
strobel@staff.uni-marburg.de
Philipps-Universität Marburg, Deutschland

Bürger, Thomas
buerger@slub-dresden.de
Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden

Ausgehend von thesenförmigen Impulsreferaten (**jeweils 10 Min., die PPT-Folien werden nach der Tagung publiziert**) sollen editionstheoretische und -technische Fragen diskutiert werden.

Block 1 Impulse (90 Minuten) 14:00 – 15:30 Uhr

1. Begrüßung und Einführung (Thomas Bürger und Jochen Strobel)
2. Offenheit und institutionelle Schließung (Patrick Sahle)
3. Kommentierung – ein Auslaufmodell? (Anne Bohnenkamp)
4. Versionierung/Zitation (Joachim Veit)
5. Hemmnisse und Katalysatoren digitaler Brief-Infrastrukturen (Thomas Stäcker)
6. Schnelle Wege zu den Briefen (Stefan Dumont)
7. DARIAH-Services für Briefeditionen (Mirjam Blümm)
8. Akteure und Rollen (Jochen Strobel)

Block 2 Diskussion, Fazit, Ausblick (90 Minuten) 16:00 – 17:30 Uhr

Diskussion der Impulsreferate und der fachlichen und förderpolitischen Schlussfolgerungen (Moderation: Thomas Bürger, Jochen Strobel)

Block 1: 14:00 – 15:30 Uhr

1. Einführung (Thomas Bürger, Jochen Strobel)

Ob es nun zutrifft, dass die Edition zu den Kerngeschäften der Geisteswissenschaften gehört und sogar die Königsdisziplin der Digital Humanities sei, bleibe dahingestellt – nutzerseitig handelt es sich bei den Forschungsergebnissen dieser Disziplin um diejenigen mit der längsten Halbwertszeit. Die Aussicht, dass gut gemachte Editionen zu den unverzichtbaren Grundlagen der wissenschaftlichen Praxis gehören, setzt die an Editionen Beteiligten aber auch unter Druck. Editionen sind langwierig und teuer, sie müssen im fachlichen und technischen Sinne zuverlässige Daten vorhalten, sie müssen dauerhaft verfügbar und doch auf der Höhe der Zeit sein. Dazu bedarf es fachlich und technisch versierter Bearbeiter. Tagungen und Publikationen zur Digitalen Editionverdeutlichen das Bedürfnis nach Orientierung, Standardisierung, Weiterentwicklung und transparenterer Vernetzung. Eine pragmatische Ausrichtung schließt diese Ziele ein:

- die Orientierung der Usability an Forschungsfragen der Nutzer*innen, insbesondere eine Optimierung der Durchsuch- und Findbarkeit und eine gute Lesbarkeit der Texte
- die zeitgemäße Einbettung der Edition in stabile virtuelle Forschungsumgebungen
- die Anschlussfähigkeit in Infrastrukturen
- Sparten und Institutionen übergreifende Projektvernetzung.

Die Briefedition ist ein prominentes Paradigma: In analoger Form ist sie besonders zeitraubend und kostspielig. Als Zeugnis historischer Netzwerke und Kommunikations-strukturen und wegen der vergleichsweise fortgeschrittenen Standardisierung bietet sie sich als Avantgarde digitaler Projektvernetzung und Daten-Aggregierung geradezu an.

Patrick Sahle hat für Editionen plausibel zwischen der Repräsentation der eigentlichen Daten („Inhalt“) und der Präsentation („Form“) unterschieden. Unumstrittene Geltung besitzen Kriterien, wie sie auch durch Roland Kamzelak benannt wurden, also z. B. XML-Standard, Open Source-Software und Open Access, Nachnutzbarkeit, persistente URL, Langzeitarchivierung. Ein faktischer Standard besteht mit den von der DFG herausgegebenen Förderkriterien für Wissenschaftliche Editionen. Komplementär hierzu und um ein Vielfaches differenzierter bietet der vom IDE bereitgestellte Kriterienkatalog für die Besprechung digitaler Editionen Anhaltspunkte. Mit RIDE besteht ein Rezensionsjournal für Editionen. Thomas Stäcker, Thomas Burch u.a. verweisen auf innovative multiple Analyse- und Darstellungsmöglichkeiten (Distant Reading, Mapping the Republic of Letters), die neue Fragestellungen und Vermittlungsformen erlauben und zu neuen Verständigungen (z.B. über Remediatisierung, über innovative Vermittlungswege der wissenschaftlichen und kulturellen Überlieferung, das Potential graphorientierter Datemodelle) anregen.

2. Offenheit und institutionelle Schließung (Patrick Sahle)

Die Digitale Edition verheißt ewige Unabgeschlossenheit und technische wie fachliche Offenheit – was aber bedeutet dies: Paradies oder Inferno? Bei allen Vorzügen der Korrigier- und Ergänzbarkeit durch das Editorenteam oder auch durch User: editorische Daten müssen zuverläs-

sig sein – wir wollen nicht ewig auf Baustellen leben. Was bedeutet dies im Hinblick auf mögliche Schließungsregeln für ein offenes Medium (s. u.: Zitation/Versionierung)? Was folgt etwa in puncto Nachnutzbarkeit hieraus? Ist zu unterscheiden zwischen gesicherten Forschungsdaten und solchen, die als Forschungsergebnisse zwar intersubjektiver Nachprüfbarkeit unterliegen, in ihrer Thesenhaftigkeit wie auch Kontextabhängigkeit aber zugleich diskussionswürdig und revidierbar sind, wie etwa Interpretamente?

In der Praxis sind es institutionelle und finanzielle Zwänge, die die Begrenzung und Schließung von an sich offenen Projekten erforderlich machen. Hieraus ergeben sich grundlegende Fragen nach der gebündelten Kuratierung abgeschlossener Projekte.

3. Kommentierung – ein Auslaufmodell? (Anne Bohnenkamp)

Ist der „Stellenkommentar" überflüssig, ein in der Editionsphilologie zwischen Redundanz- und Interpretationsverdikt immer schon umstrittenes Element, das gleichwohl der Profilierung der Editor*innen wie vor allem der Usability diente? Ist über Tagging und Verlinkung hinaus ein editionsspezifisches kommentierendes Angebot sinnvoll – und wie sollte es sich zu den zuhauf existenten externen Informationsressourcen technisch und sachlich verhalten? Ist die Kommentierung vielleicht sogar jenseits der scheinbar mechanischen und objektiven digitalen Repräsentation jener letzte Ort, an dem das Fachwissen und die interpretatorische Leistung die Unverzichtbarkeit des Editors als eigentlichem Experten und bestem Kenner der Materie belegt?

4. Versionierung/Zitation (Joachim Veit)

Die Herstellung von beliebig vielen, zitationsfähigen Versionen eines edierten Texts oder Textabschnitts ist technisch kein Problem, die Zuweisung persistenter Identifier ebenso wenig. Wie sollten Forschungsdaten jenseits der Text(präsentations)ebene zitiert werden? Was sind die (Haupt-?)Bestandteile der Editionen, wo liegen ihre Grenzen? Nach der verbindlichen Veröffentlichung eines Textes müssen editorische Veränderungen als unterschiedliche Versionen markiert werden. Hierzu dienlich ist ein Versionierungssystem. Präzise Zitierfähigkeit ist ein unverzichtbares Merkmal von Editionen – doch wie bleibt sie unter digitalen Auspizien praktikabel? Hierzu zählt z. B. die nicht triviale Frage nach der Länge des Zitatnachweises, der ganz oder teilweise zugleich Link in einer Online-Publikation ist.

5. Hemmnisse und Katalysatoren digitaler Brief-Infrastrukturen (Thomas Stäcker)

Es besteht seit langem Einigkeit darüber, dass die Verwendung normierter Metadaten sowie die Codierung gemäß den Regeln des TEI-Konsortiums eine Garantie für Qualität und Nachhaltigkeit bietet. GND, VIAF – inzwischen verwenden Digitale Editionen bibliothekarische Normdaten zur Referenzierung von Personen, Orten u.a. Die Standardisierung von Werktiteln schreitet eher langsam voran. In der Praxis ist aber ebenso bekannt, dass von Projekt zu Projekt kleinere oder größere Abweichungen zu verzeichnen sind. Welche Folgerungen sind aus dieser Inkongruenz zu ziehen, wie ist mit einer unterschiedlichen Nutzung von Standards umzugehen?

Mit dem Verlassen der Zweidimensionalität des gedruckten Textes findet nicht nur durch das Verschwinden des Substrates ein Verlust an Stabilität statt, sondern es treten durchaus andere stabile Konstituentien an dessen Stelle. Welche Entwicklungen sind zu erwarten bzw. zu organisieren?

6. Schnelle Wege zu den Briefen (Stefan Dumont)

CorrespSearch will verbesserte Voraussetzungen für die Vernetzung von Briefeditionen schaffen. Damit sollen Briefmetadaten über das bisherige Basis-Set hinaus verarbeitet und auch die Erstellung von digitalen Briefverzeichnissen aus gedruckten Publikationen unterstützt werden. Über Schnittstellen zu anderen Diensten ist der Webservice in die existierende und zu entwickelnde Infrastrukturlandschaft tiefer einzubetten. Wie lassen sich nutzer- und forschungsfreundlich die Zugänge zu Briefen und insbesondere zu den Volltexten vereinfachen? In welchem Verhältnis steht CorrepSearch zu den vielen anderen Angeboten (Deutsche Digitale Bibliothek, Europeana, Kalliope u.a.)?

7. DARIAH-Services für Briefeditionen (Mirjam Blümm)

DARIAH.DE ist eine Informations- und Technikinfrastruktur für Lehre, Forschung, Forschungsdaten und technische Werkzeuge. Welchen Stand hat die Infrastruktur zur Einbindung digitaler Editionen erreicht und welche kollaborativen Strukturen und Ziele sollten angestrebt werden?

8. Akteure und Rollen (Jochen Strobel)

Die Digital Humanities fördern kollaboratives Arbeiten und generell eine Diversifizierung und Verflüssigung von Rollen und Verantwortlichkeiten. Zu den Rollen einer digitalen wissenschaftlichen Autor- und Beiträgerschaft zählen u.a. Hauptherausgeber*in, Kurator*in, Programmierer*in, Kodierer*in, Tagger*in, wissenschaftliche Hilfskraft, Crowdsourcer*in. Einige dieser und weitere Rollen sind so neu nicht, manche dürften vom bisherigen Urheberrecht nicht hinreichend abgedeckt sein. Editionen sind seit langem kollaborative Projekte, die Rollen der Beteiligten sind dementsprechend vielfach ausdifferenziert. Welche Folgerungen ergeben sich für die digitale Edition einerseits und für die akademische Kar-

riereplanung zwischen Forschung und Informationsinfrastruktur andererseits?

Block 2: 16:00 – 17:30 Uhr

9. Diskussion, Fazit, Ausblick (Moderation: Thomas Bürger, Jochen Strobel)

In der Schlussdiskussion sollen die Impulsreferate diskutiert und ein Fazit gezogen werden. Dabei sind auch forschungspolitische Schlussfolgerungen zu thematisieren. Editionen und insbesondere Briefeditionen sind in der Regel langfristige Projekte und deshalb vor allem an Akademien angesiedelt. Durch die sich verbreiternde Retrodigitalisierung und verfügbare digitale Werkzeuge können Projektlaufzeiten perspektivisch weiter verkürzt und Projektfortschritte transparent in Forschung und Lehre eingebunden werden. Wie groß ist der Bedarf an digitalen Editionen und wie können sie Teil einer Nationalen Forschungsdateninfrastruktur werden? Welche drängenden Fragen sind zeitnah in weiteren Workshops zu besprechen?

Bibliographie

Roland Kamzelak: Empfehlungen zum Umgang mit Editionen im Digitalen Zeitalter, in editio 26 (2012), S. 202–209.

Patrick Sahle: Digitale Editionsformen. 3 Bände. Norderstedt 2013.

http://www.mww-forschung.de/blog/blogdetail/wie-sieht-die-digitale-edition-der-zukunft-aus-herr-kamzelak/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail (30.8.2017)

http://dhd-wp.hab.de/?q=ag-text#abschnitt4 (30.8.2017)

http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf (30.8.2017)

https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ (30.8.2017)

http://ride.i-d-e.de/ (30.08.2017)

Panels

Abgrenzung oder Entgrenzung? Zum Spannungsverhältnis zwischen Historischen Hilfswissenschaften und Digital Humanities

Schulz, Daniela
dschulz@uni-wuppertal.de
Bergische Universität Wuppertal

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich

„Als Grundwissenschaft erwirken die Digital Humanities das so elementar wichtige *Nutzenkönnen* digitaler Methoden und Daten, wie die Paläographie uns das *Lesenkönnen* unserer Quellen sicherstellt". (Rehbein 2015) Wie Malte Rehbein hier andeutet, scheint es hinsichtlich des Stellenwertes, der den Historischen Hilfs- oder Grundwissenschaften (HGW) wie auch den Digital Humanities (DH) eigentlich beigemessen werden sollte, durchaus Ähnlichkeiten zu geben. „Sollte"! Denn sowohl bei den HGW als auch bei den DH ist ihr Status als eigenständiger wissenschaftlicher Zweig nicht gänzlich unumstritten. Beide werden aktuell häufig als reine Zulieferer-Wissenschaften oder Dienstleister gegenüber der „richtigen" Forschung wahrgenommen und ihr eigener wissenschaftlicher Wert in Zweifel gezogen.

Die zum Kanon der traditionellen HGW gehörenden, teils sehr unterschiedlichen Teildisziplinen – neben der bereits genannten Paläographie zählen unter anderem auch Kodikologie, Epigraphik, Heraldik, Sphragistik oder Diplomatik dazu – arbeiten allesamt quellennah und betreiben damit wertvolle Grundlagenforschung. Ob der Breite dieses Kanons fällt es nicht ganz leicht, die HGW in Patrick Sahles „3-Sphären-Modell zur Kartierung der Digital Humanities als Schnittmenge, Brücke und eigenständigem Bereich zwischen (ausgewählten) traditionellen Disziplinen" (Sahle 2015) zu verorten. In vielen Aspekten scheinen sie den DH im Hinblick auf Interdisziplinarität, Methoden, Stellenwert etc. jedoch sogar näher zu stehen als die Geschichtswissenschaft, unter die sie im Allgemeinen subsumiert werden. Ja, es gab sogar eine Phase, in der die Historische Fachinformatik als neue Teildisziplin der HGW galt. [1]

Während Professuren mit einer DH-Ausrichtung oder Denomination auf dem Vormarsch zu sein scheinen – in seinem Beitrag „Zur Professoralisierung der Digital Humanities" zählt Sahle mittlerweile 53 Ausschreibungen (Stand: Januar 2018) im deutschsprachigen Raum mit allerdings äußerst diversen Ausrichtungen (Sahle 2016) – ist in den letzten Jahren die Zahl der Universitätsstandorte, die HGW im Programm haben, zunehmend kleiner geworden, so dass diese heute mit zu den strukturprekären Disziplinen gehören. [2] (Arbeitsstelle Kleine Fächer) Diese Situation war Ende 2015 Anlass für die Formulierung des Positionspapieres „Quellenkritik im digitalen Zeitalter. Die Historischen Grundwissenschaften als zentrale Kompetenz der Geschichtswissenschaft und benachbarter Fächer" von Eva Schlotheuber und Frank Bösch (Schlotheuber / Bösch 2015), welches eine breite Diskussion auf „H-Soz-Kult" in Gang setzte, bei der (teils beiläufig) auch immer wieder das Verhältnis von HGW und DH thematisiert wurde. Trotz aller Differenzen, die bei diesem – mitunter durchaus kontrovers geführten – Austausch zutage kamen, herrschte hinsichtlich eines Aspektes mehrheitlich Einigkeit: Der Wegfall von Professuren, Studiengängen und Lehrveranstaltungen, die das notwendige methodische, grundwissenschaftliche Rüstzeug an heutige und künftige Generationen von Studierenden weitergeben, resultiert in einem Mangel an entsprechenden Fachkompetenzen. In einer Zeit, in der im Zuge zunehmender Digitalisierung historische Quellen in großer Zahl allgemein und jederzeit verfügbar geworden sind, führt dies zu der grotesken Situation, dass das Auffinden von Quellen und der Zugriff auf sie heute zwar deutlich einfacher geworden ist, die Mittel, mit diesen adäquat umzugehen, vielen Personen aber nicht mehr (oder noch nicht) zur Verfügung stehen. Dass der Zugang zu Datenbanken, die beispielsweise bei der Datierung, Verortung oder Einordnung von Einbänden, Wasserzeichen, Initialen etc. helfen, die Arbeit der Grundwissenschaftler|innen und auch anderer Forschenden heute erleichtert und ökonomisiert, wird von den Nutzer|inne|n niemand bestreiten. Das Gros des heutigen in den HGW beschäftigten Lehrpersonals ist allerdings selbst oft nicht ausreichend geschult, um die notwendigen Kenntnisse im Umgang mit diesen Ressourcen an die Studierenden zu vermitteln.

Dennoch fordern aktuelle Ausschreibungen von Bewerber|innen häufig ausgeprägte grundwissenschaftliche Kompetenzen und gleichzeitig Kenntnisse im Bereich der DH. Die wenigsten Absolvent|inn|en deutscher Hochschulen können

diesem Profil heute wirklich gerecht werden. Personen, die im Rahmen ihres Studiums noch eine tiefergehende Ausbildung im erstgenannten Bereich genossen haben, stehen oft vor der Problematik, dass sie sich ihr Wissen im Bereich der DH mühevoll im Selbststudium oder in den zahlreich angebotenen Summer Schools erarbeiten müssen.

Neben diesen praktischen Problemen der Zugänglichkeit zu entsprechenden Weiterbildungsangeboten, differieren aber auch die grundlegenden Auffassungen darüber, welche (technischen) Kompetenzen überhaupt notwendig sind, um das „digital“ Dargebotene hinsichtlich seiner Wissenschaftlichkeit und Vollständigkeit hinreichend bewerten zu können. [3] Auch herrscht weiterhin Uneinigkeit darüber, wie diese Kompetenzen (an Studierende und Lehrende) überhaupt vermittelt werden können. Sind die Grundwissenschaften in der Pflicht, ihre Vermittlungskonzepte auf die veränderte Situation anzupassen? (Vogeler 2015) Definitiv! „[S]ind digitale Techniken und Methoden nicht nur eine Chance, sondern vielleicht auch die einzige Möglichkeit für eine sinnvolle Weiterentwicklung der Hilfswissenschaften“? (Hiltmann 2015) Wahrscheinlich ja! Doch wie kann diese Weiterentwicklung ganz konkret aussehen? (Wie) Können die hergebrachten grundwissenschaftlichen Kompetenzen erhalten, und dabei gleichzeitig neue Kompetenzen aufgebaut werden, die für das heutige und zukünftige wissenschaftliche Arbeiten benötigt werden? Wie müsste eine Neuausrichtung grundwissenschaftlicher Curricula, die stärker digitale Methoden in die Lehre integrieren, aussehen? Welche Umsetzungsversuche gibt es hier bereits? Welche Kompetenzen werden gebraucht und wo kann Kompetenzaufbau (auch für Graduierte) stattfinden? Welche Maßnahmen sind hierfür notwendig? Welche konkreten Maßnahmen wurden in den vergangen zwei bis drei Jahren vielleicht auch schon in Gang gesetzt, um den Bereich der HGW zu erhalten bzw. zu neuem Leben zu erwecken? [4] Wie kann das unzweifelhaft vorhandene Potenzial bestmöglich genutzt werden, um die grundwissenschaftliche Forschung zu befördern? Aber auch: Wo liegen vielleicht grundlegende Probleme in der Kollaboration von HGW und DH?

Bei der Aushandlung des Stellenwertes bzw. der Verortung der DH ging es bisher meist um das Verhältnis zwischen angewandter Informatik und traditionellen Geisteswissenschaften. Der Dialog zwischen HGW und DHs erscheint aber besonders für einen fruchtbaren Austausch geeignet, da beide „Disziplinen“ teils mit sehr ähnlichen Problemen zu kämpfen haben und aus sich heraus schon interdisziplinär arbeiten. Andererseits stellen aber die DH gerade aus Sicht einiger Grundwissenschaftler|innen vermehrt ein Feindbild dar, da sie vermeintlich die Existenz des eigenen Faches bedroht sehen und/oder sich den neuen Herausforderungen nicht gewachsen fühlen.

Das Panel, welches diesmal Vertreter|inn|en der traditionellen HGW *UND* Digital Humanists als dezidierte „Grenzgänger“ an einen Tisch bringt, soll also zum einen dazu dienen, Vorurteile abzubauen und die bereits begonnenen Diskussionen am Leben zu halten, zu konkretisieren und weiterzuführen, zum anderen auch Gelegenheit bieten, beispielsweise Projekte mit grundwissenschaftlicher Ausrichtung sichtbar zu machen und damit gleichzeitig deren Ansätze, Methoden und Umsetzung zur allgemeinen Diskussion zu stellen.

Die folgenden Personen (in alphabetischer Reihenfolge) haben ihre Teilnahme am Panel zugesagt:

Jun.-Prof. Dr. Étienne Doublier, Juniorprofessur für Historische Hilfswissenschaften (Bergische Universität Wuppertal)

Jun.-Prof. Dr. Torsten Hiltmann, Juniorprofessur für die Geschichte des Hoch- und Spätmittelalters / Historische Hilfswissenschaften (Westfälische Wilhelms-Universität Münster)

Prof. Dr. Andrea Stieldorf, Institut für Geschichtswissenschaft, Abteilung für Historische Hilfswissenschaften und Archivkunde, Rheinische Friedrich-Wilhelms-Universität Bonn

Prof. Dr. Georg Vogeler, Zentrum für Informationsmodellierung in den Geisteswissenschaften, Karl-Franzens-Universität Graz

Der geplante Ablauf ist wie folgt:

Nach einer knappen Einleitung in die Thematik des Panels und der Motivation zu dessen Organisation, erhalten die Diskutant|inn|en zunächst Gelegenheit zu einer individuellen Stellungnahme. Der Verlauf der anschließenden Diskussion wird nicht – wie sonst üblich – durch vorgegebene Fragen seitens der Moderatorin vorgegeben, sondern „von außen“ bestimmt, so dass dieser für alle Beteiligten nicht vorhersehbar ist. Mit „von außen“ ist hier zum einen das Plenum der Anwesenden gemeint, unten denen sich hoffentlich auch zahlreiche Vertreter der Studierendenschaft und des wissenschaftlichen Nachwuchses befinden, zum anderen aber gerade auch Personen, die selbst nicht an der Veranstaltung teilnehmen können. Hierzu wurde Ende 2017 unter anderem über Twitter ein Aufruf gestartet, (unter # **dhdp3a**) Fragen und Diskussionsthemen einzureichen, die dann vor Ort besprochen werden können. Dieses Vorgehen soll sicherstellen, dass 1) in der Diskussion tatsächlich jene Themen

aufgegriffen werden, die von allgemeinem Interesse sind, 2) nicht bereits vielfach geführte Debatten lediglich repliziert werden, und 3) ein wirklicher Dialog zwischen „Betroffenen“ und anderen Interessierten stattfinden kann, da diesmal nicht exklusiv auf professoraler Ebene diskutiert wird.

Fußnoten

1. http://www.hgw.geschichte.uni-muenchen.de/ueber_uns/faecher/fachinformatik/index.html
2. Die innerdeutsche Verteilung gestaltet sich entsprechend recht übersichtlich:

1. Ruprecht-Karls-Universität Heidelberg (Früheres Mittelalter und historische Grundwissenschaften)
2. Eberhard-Karls-Universität Tübingen (Geschichtliche Landeskunde und Historische Hilfswissenschaften)
3. Otto-Friedrich-Universität Bamberg (Historische Grundwissenschaften)
4. Friedrich-Alexander-Universität Erlangen-Nürnberg (Mittelalterliche Geschichte und Historische Hilfswissenschaften)
5. Ludwig-Maximilians-Universität München (Historische Grundwissenschaften und Historische Medienkunde)
6. Universität Passau (Mittelalterliche Geschichte und historische Hilfswissenschaften)
7. Universität Regensburg (Historische Hilfswissenschaften)
8. Julius-Maximilians-Universität Würzburg (Mittelalterliche Geschichte und historische Hilfswissenschaften)
9. Christian-Albrechts-Universität zu Kiel (Mittelalterliche Geschichte und historische Hilfswissenschaften)
10. Ruhr-Universität Bochum (Historische Hilfswissenschaften)
11. Rheinische Friedrich-Wilhelms-Universität Bonn (Historische Hilfswissenschaften und Archivkunde)
12. Universität zu Köln (Historische Hilfswissenschaften)

3. Das Gleiche gilt auch für die schon zahlreich verfügbaren Tools zu deren vereinfachten Be- oder Verarbeitung.
4. Exemplarisch können hier – neben der noch aktuellen Ausschreibung einer (wenn auch befristeten) W2 Professur für „Historische Grundwissenschaften unter besonderer Berücksichtigung der Digital Humanities“ – auch der Zusammenschluss historisch arbeitender Wissenschaftler|inne|n zur „Arbeitsgemeinschaft Historische Grundwissenschaften“ (AHiG, https://www.ahigw.de/) genannt werden, sowie die Gründung des „Netzwerk Historische Grundwissenschaften“ (NHG, https://www.ahigw.de/nachwuchsnetzwerk/). Bei letzterem handelt es sich um einen Zusammenschluss von Nachwuchswissenschaftler|inn|en verschiedener Disziplinen und Qualifikationsstufen, die neben der Organisation einer jährlichen Konferenz mit grundwissenschaftlicher Ausrichtung, auch auf anderen Wegen versuchen, sich produktiv in die aktuellen Diskussionen einzuschalten und Entwicklungen voran zu treiben – so z.B. auch mit der Organisation dieses Panels.

Bibliographie

Arbeitsstelle Kleine Fächer, Fachstandort der Historischen Hilfswissenschaften. URL: http://www.kleinefaecher.de/historische-hilfswissenschaften/ [letzter Zugriff 24.09.2017]

Hiltmann, Torsten (2015): „Hilfswissenschaften in Zeiten der Digitalisierung“, in: H-Soz-Kult, 14.12.2015. URL: [letzter Zugriff 24.09.2017]

Rehbein, Malte (2015): „Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht“, in: H-Soz-Kult, 27.11.2015. URL: www.hsozkult.de/debate/id/diskussionen-2905 [letzter Zugriff 24.09.2017]

Sahle, Patrick (2015): „Digital Humanities? Gibt’s doch gar nicht! in: Grenzen und Möglichkeiten der Digital Humanities (Sonderband der Zeitschrift für digitale Geisteswissenschaften 1). DOI: http://dx.doi.org/10.17175/sb001_004 [letzter Zugriff 13.01.2017]

Sahle, Patrick (2016): „Zur Professoralisierung der Digital Humanities“, in: DHd-Blog, 23. März 2016. URL: http://dhd-blog.org/?p=6174 [letzter Zugriff 24.09.2017]

Vogeler, Georg (2015): „Digitale Quellenkritik in der Forschungspraxis“, in: H-Soz-Kult, 28.11.2015. URL: www.hsozkult.de/debate/id/diskussionen-2893 [letzter Zugriff 24.09.2017]

Schlotheuber, Eva / Bösch, Frank (2015): „Quellenkritik im digitalen Zeitalter. Die Historischen Grundwissenschaften als zentrale Kompetenz der Geschichtswissenschaft und benachbarter Fächer“, in: H-Soz-Kult, 15.11.2015. URL: [letzter Zugriff 24.09.2017]

Alles ist im Fluss - Ressourcen und Rezensionen in den Digital Humanities.

Neuber, Frederike
neuber.frederike@gmail.com
Institut für Dokumentologie und Editorik e.V.; Berlin-Brandenburgische Akademie der Wissenschaften

Henny-Krahmer, Ulrike
ulrike.henny@uni-wuerzburg.de
Institut für Dokumentologie und Editorik e.V.; Universität Würzburg

Sahle, Patrick
sahle@uni-koeln.de
Institut für Dokumentologie und Editorik e.V.; Universität zu Köln/CCeH

Fischer, Franz
franz.fischer@uni-koeln.de
Institut für Dokumentologie und Editorik e.V.; Universität zu Köln/CCeH

Eine Kritik der digitalen Vernunft muss auch eine Kritik der digitalen Ressourcen umfassen. [1] Eine traditionelle Form der Kritik ist die wissenschaftliche Besprechung oder Rezension. Die kritische Instanz eines Rezensionswesens fehlt den Digital Humanities bisher fast gänzlich. Digitale Ressourcen wie etwa wissenschaftliche Editionen, Textkorpora, Bilddatenbanken oder auch Software werden selten umfassend rezensiert. [2] Das Gleiche gilt für das "Primärmaterial" dieser Ressourcen, also für Datensätze.

Im Gegensatz zu traditionellen Forschungsergebnissen der Geisteswissenschaften sind digitale Ressourcen nicht statisch, sondern wandelbar, oft prozesshaft und nicht abgeschlossen. Die Kritik der Ressourcen muss die besonderen Bedingungen, Eigenschaften und Folgephänomene digitaler Daten berücksichtigen und eine eigene Form finden. [3] Wie sich die Ressourcen wandeln, so muss sich auch die Kritik wandeln, denn panta rhei - "alles fließt", sagt Heraklit. [4]

Vor diesem Hintergrund widmet sich das Panel u. a. folgenden Fragen: Wie können traditionelle Rezensionsorgane die Besprechung digitaler Ressourcen fördern? Brauchen die Digital Humanities ein eigenes Rezensionswesen, um den Besonderheiten digitaler Ressourcen gerecht zu werden? Wenn ja, wie muss die Rezension als ‚Momentaufnahme' einer digitalen Ressource methodisch und technisch konzipiert sein, um nicht der ‚Schnelllebigkeit' der digitalen Welt zu erliegen? Welche Herausforderungen stellen Publikationen von Daten oder Algorithmen an die RezensentInnen? Inwiefern steigen mit dem Zuwachs an technischen Möglichkeiten auch die Ansprüche an digitale Ressourcen und wie lassen sich diese als Standards und Evaluationskriterien verhandeln, dokumentieren und weiterentwickeln? Wie kann man die Digitalität des Rezensionswesens nutzen, um die Prozesshaftigkeit der zu rezensierenden Objekte zu berücksichtigen, z. B. auch durch dynamischere Formen der Kritik jenseits der traditionellen Rezension?

Das Panel wird mit einer Moderation und sechs Impulsbeiträgen einen Überblick über das Themenfeld "Ressourcen und Rezensionen in den Digital Humanities" geben. Die ReferentInnen berichten von Erfahrungen aus der herausgeberischen Praxis, diskutieren Problemfelder und setzen theoretische Impulse, um im Anschluss rasch in eine offene Diskussion mit dem Publikum überzugehen.

Rüdiger Hohls: [5] Die Rezension analoger und digitaler Medien bei H-Soz-Kult

Die Tradition der wissenschaftlichen Rezensionen wird vor allem in den Geistes-, Kunst-, Kultur-, Politik- und Sozialwissenschaften gepflegt, überwiegend Fächer, die unter einer Überproduktion von Texten leiden und deshalb für eine Sichtung und für ein „Marketing" über Rezensionen besonders empfänglich sind.

Zu den Pflichten einer Rezension zählt laut Georg Jäger (2001) die Berichtspflicht über Zielsetzung, Gliederung, Argumentationslinien und Ergebnisse, kritische Reflexion des methodischen Vorgehens und der ausgewählten Quellen, weiterhin die Wertung und Einbettung in den Kontext der einschlägigen Fachdiskussion. Zur Kür bei digital vertriebenen Rezensionen zählt darüber hinaus ein „feuilletonistischer Ton". Außerdem sei es notwendig, die „Ethik wissenschaftlicher Kommunikation" neu zu justieren. Viele dieser Aspekte gelten auch für die Rezension digitaler Medien, und natürlich haben wir bei H-Soz-Kult unsere formatspezifischen Hinweise für Re-

zensentInnen regelmäßig an neue technische Entwicklungen angepasst.

H-Soz-Kult hat seit 1996 über 15.200 Buchrezensionen, mehr als 220 Ausstellungsrezensionen, knapp 160 Rezensionen zu Publikationen auf digitalen Trägermedien und lediglich 33 sogenannte Webrezensionen veröffentlicht. Die Zahlen sprechen nicht nur für die Relevanz der Genres bei H-Soz-Kult, sondern unterstreichen auch die Bedeutung des „Buchs“ in den Geschichtswissenschaften, egal ob analog oder digital publiziert. Es stellt sich aber u. a. die Frage, warum es trotz aller Bemühungen bei H-Soz-Kult nur vergleichsweise wenige Besprechungen von fachwissenschaftlichen digitalen Ressourcen gibt.

Frederike Neuber: [6] RIDE und die Herausforderungen der Digitalität

Seit 2014 gibt das IDE die digitale Rezensionszeitschrift RIDE (*A review journal for digital editions and resources*) heraus. Bis Dezember 2017 wurden in sieben Ausgaben 30 wissenschaftliche digitale Editionen und 10 digitale Textsammlungen rezensiert. Zum methodischen Rahmenprogramm RIDEs zählen Kriterienkataloge für die Besprechung des jeweiligen Ressourcentyps (Henny u. Neuber 2017, Sahle 2014), die Erhebung von Daten in einem Questionnaire und ein externes Peer Reviewing der Beiträge.

Ein “markantes Kennzeichen digitaler Texte ist deren Veränderbarkeit und prinzipielle Offenheit” (cf. Working Paper der DHd-Arbeitsgruppe “Digitales Publizieren” 2016), das gilt auch für die Rezensionsobjekte in RIDE: Etwa kann sich schon während des Rezensierens und/oder nach der Publikation einer Rezension die Datengrundlage einer besprochenen Ressource verändern; im Zuge eines Relaunches kann ein User Interface ersetzt werden. Um derlei Fällen vorzubeugen empfehlen die RIDE- *Reviewing Guidelines*[7] etwa die Erstellung von Screenshots und die Archivierung von Websites. Damit ist der Status Quo der Ressourcen zum jeweiligen Rezensionszeitpunkt zumindest teilweise dokumentiert.

Wie die Rezensionsobjekte, so sind auch die digitalen Rezensionstexte grundsätzlich offener und veränderlicher als das bei gedruckten Rezensionen der Fall ist. Vereinzelt wurden in RIDE bereits nachträgliche Änderungen (z. B. auf Wunsch der BetreiberInnen der rezensierten Ressource) in Rezensionstexte integriert, was in der XML/TEI-Version der jeweiligen Ressource dokumentiert ist.

Grundsätzlich steht die “Aktualität” digitaler Rezensionen von digitalen Ressourcen trotz den in RIDE bestehenden und oben geschilderten Dokumentationsverfahren auf wackeligen Beinen. Es stellt sich daher die Frage, ob eine Rezension mehr sein kann bzw. muss als die “Momentaufnahme” eines bestimmten Entwicklungsstandes einer digitalen Ressource. Und wenn sich eine Ressource nach der Rezension signifikant weiterentwickelt, liegt es dann in der Verantwortung der RIDE-HerausgeberInnen, eine erneute Rezension anzustoßen? Oder sollte man den RessourcenbetreiberInnen selbst die Möglichkeit geben, in RIDE über Updates zu berichten, wenn Rezension und aktueller Ressourcenstand zu sehr divergieren?

Anne Baillot: [8] Daten als Rezensionsobjekte

In den Geisteswissenschaften werden Forschungsfragen verfolgt und in Publikationen beantwortet. Diese sind traditionell Gegenstand von Rezensionen. Im Digitalen kommt mit den Forschungsdaten eine “neue” Publikationsform hinzu, die in das Rezensionswesen einbezogen werden muss. Der Status des "Primärmaterials" ist im digitalen Rezensionswesen zu klären und kann sich unter Umständen am Modell der Naturwissenschaften orientieren. Data Papers und Data Journals sind neue Formen der Dokumentation und Publikation, die sich Struktur, Kohärenz und Vollständigkeit von Daten widmen. Damit rückt auch Datenmodellierung als eine zu evaluierende wissenschaftliche Tätigkeit mehr in den Vordergrund.

Während naturwissenschaftliche Zeitschriften, die Data Papers veröffentlichen oder rezensieren, solche Datensätze in speziellen, selten öffentlich zugänglichen Repositorien verfügbar machen, entwickelt sich in den Geisteswissenschaften die Tendenz, die Primärdaten eher zusammen mit ihren Auswertungen oder auf allgemeinen Plattformen offen und langzeitverfügbar zugänglich zu machen. Wie stabil ist diese offene Bereitstellung auf lange Sicht? Wie komplex wird dadurch bei einer immer wieder aktualisierten digitalen Ressource die Bezugnahme auf jene Version, die in einer Rezension begutachtet wird? Darüber hinaus stellt sich die grundsätzliche Frage: Stoßen Rezensionen von Data Papers und den Daten selbst in einem Kontext der Informationsflut, in der kaum die Auswertungen wahrgenommen werden, überhaupt auf Interesse? Welche Funktion hätten sie dann in einem geisteswissenschaftlichen Rezensionswesen in der digitalen Welt?

Werden sie zur Rettung oder zum endgültigen Sturz des Rezensionswesens in die Bedeutungslosigkeit beitragen?

Christof Schöch: [9] Im Spannungsfeld von Projektzielen und Best Practice-Anforderungen

Bei der Rezension digitaler Ressourcen treffen zwei Perspektiven aufeinander: Einerseits ist das der Rahmen der jeweils spezifischen Projektziele, die für die Erstellung einer Ressource handlungsleitend waren. Hierzu gehören auch praktische Aspekte der Machbarkeit wie Zeitrahmen, Budget und Personal. Nicht immer ist diese Perspektive für die Rezensierenden transparent. Andererseits ist das die Perspektive der theoretischen Anforderungen und der mehr oder weniger gut etablierten Best Practices, die in Form von Qualitätskriterien an eine digitale Ressource herangetragen werden können. Auch hier sind die entsprechenden Maßstäbe nicht immer geteilt und explizit.

Dieses Spannungsfeld und die unter Umständen asymmetrische Verfügbarkeit von Informationen wirft die Frage auf, wie die Rezension einer digitalen Ressource sowohl fair als auch anspruchsvoll sein kann, wenn unterschiedliche Maßstäbe zugleich anzulegen sind. Es stellt sich auch die Frage, inwiefern die Rezension digitaler Ressourcen hier grundsätzlich anders funktioniert als die Rezension von Monographien oder Sammelbänden. Inwieweit sollten bei der Rezension digitaler Ressourcen die äußeren Bedingungen ihrer Erstellung berücksichtigt werden? Braucht es womöglich verschiedene Ebenen von Anforderungen, wenn z. B. Ergebnisse der Individualforschung mit denjenigen großer Projekte verglichen werden? Inwiefern muss beim Anlegen von Maßstäben berücksichtigt werden, ob eine digitale Ressource weitestgehend abgeschlossen oder noch in der Bearbeitung ist?

Jürgen Hermes: [10] Die Rezension als Prozess

Ein Leitsatz, der die Open Source-Softwareentwicklung stark geprägt hat, und der sich auch im Paradigma der agilen Entwicklung wiederfindet, lautet “Release early, release often” (Raymond 2001).

Ist dieses Paradigma auch auf andere Arten digitaler Publikationen wie digitale Monographien, Aufsätze, Editionen, Text- und Datensammlungen anwendbar? Eine Veröffentlichung auf digitalen Plattformen bietet völlig neue Möglichkeiten des Austauschs zwischen ErstellerInnen und KonsumentInnen von digitalen Ressourcen − Errata können kurz nach ihrer Entdeckung korrigiert, Verbesserungsvorschläge zum Zeitpunkt ihrer Einreichung aufgenommen werden. Das Zusammenspiel zwischen Veröffentlichung und kritischer Würdigung wird so granularer, als dies beim althergebrachten Zusammenspiel zwischen Monographie und Rezension der Fall war.

Derartige, sich im Fluss befindliche Ressourcen schaffen aber dort Probleme, wo auf verlässliche Quellen verwiesen werden soll. Ebenso wird die etablierte Kultur des wissenschaftlichen Diskurses in Frage gestellt, der auf der Grundlage fixierter Objekte stattfindet. Derlei Aspekte müssen ernst genommen, können aber womöglich durch Entwicklungen in den Bereichen der Kontrolle, Referenzierung und Kommentierung von Text- und Daten-Versionen aufgefangen werden. Des Vorteils, wissenschaftliche Daten, Methoden und Ergebnisse zeitnaher, barrierefreier und intensiver zu diskutieren, sollten sich auch GeisteswissenschaftlerInnen nicht berauben lassen.

Patrick Sahle: [11] Prinzipien der Digitalität

Unsere digitale Informationsumwelt ist, wie die vorhergehende analoge Welt, von wenigen Grundprinzipien geprägt, die zunächst technischer Natur sind, dann aber Folgen in allen möglichen Bereichen erzeugen: Technik, Methode, Inhalte, Form, Soziologie, Politik von Wissen, Wissensproduktion und Wissenspräsentation in der Forschung. Im Bereich digitaler Ressourcen und Rezensionen lassen sich hier vielfältige Phänomene durch verallgemeinernde Schlagworte andeuten: Multimedialisierung, Vernetzung, Aufhebung von Mengenbegrenzungen, Trennung von Daten, Auswertung und Präsentation, Generativität und Prozesshaftigkeit von Publikationen, Dynamik und Interaktion, Zunahme der Komplexität, Kollaborativität, Interdisziplinarität, unmittelbare Diskursivität etc. Damit sind neue Herausforderungen markiert, deren Lösungen jetzt noch nicht absehbar oder erst thesenhaft formulierbar sind. Zu nennen wären hier unmittelbare Sichtbarkeit, “Unfassbarkeit” von Ressourcen (hinsichtlich Inhalt, Status, Version, Verantwortung), Probleme der Kreditierung von Leistungen, Interdisziplinarität als Kompetenzproblem, unformierte Wissenschaftskommunikation, Halbwertszeit von Kritik, eingefrorene ver-

sus dynamische Kritikdiskurse, Standardisierung versus Innovation etc. Auch die digitale Rezension digitaler Ressourcen steht damit vor ganz anderen Aufgaben als die eingeführte Form der Besprechung.

Vielleicht kann man es aber auch so zusammenfassen: Die Rezension in der Buchkultur ist die Grabrede auf ein in den Brunnen gefallenes Kind − die Rezension in einer digitalen Kultur ist ein Verbesserungsvorschlag, der mit seiner Annahme hinfällig wird. Wie so vieles wird in der digitalen Welt auch die Kritik komplexer, die ihre neue Form und Position noch nicht gefunden hat. Vielleicht muss sie aber auch grundsätzlich anders gedacht werden, denn als Fortsetzung des traditionellen Objekts “Rezension”?

Fußnoten

1. Unter “digitale Ressource” verstehen wir publizierte digitale Objekte, in Anlehnung an die Definitionen der DNB im Papier “Digitale Publikation” (DNB 2017).
2. Werden digitale Ressourcen rezensiert, dann meist aus der Fachperspektive. Einige Beispiele für verstreute Rezensionen aus DH-Perspektive sind Schöch (2017); Schelbert (2017); Assmann und Sahle (2008) mit Rezension zur Rezension (Just 2009); erste Ansätze für systematische Kritik im Sinne gebündelter Aktivitäten zur Bewertung bestimmter Typen digitaler Ressourcen (wie Projekte, Editionen, Textsammlungen, Datensätze), mit Begleitmaterialien und eingebunden in einen kritischen Diskurs, finden sich im DH Commons Journal, dem jTEI und RIDE.
3. Vgl. zu diesem Problemfeld u.a. die Sektion “Evaluating Digital Scholarship” in Profession, herausgegeben von Schreibman et al. 2011; das Working Paper der DHd-Arbeitsgruppe “Digitales Publizieren” 2016; der Band “Closing the Evaluation Gap” des JDH von Cohen / Fragaszy Trojano 2012; zu digitalen Editionen insbesondere Pierazzo 2014, Yates 2008 und Henny, erscheint demnächst.
4. Ein Phänomen dieses Wandels sind diverse Kataloge für Best Practices und Bewertungskriterien digitaler Ressourcen (AHA 2015, Jannidis 1999, MLA 2012, Presner 2012, Rockwell 2012, Sahle et al. 2014, Henny und Neuber 2017).
5. Humboldt-Universität zu Berlin, Redaktionsmitglied von Clio online bzw. H-Soz-Kult (Projektleitung).
6. Berlin-Brandenburgische Akademie der Wissenschaften, Mit-Herausgeberin des Rezensionsjournals RIDE.
7. Siehe Institut für Dokumentologie und Editorik e.V. (2014 ff.), http://ride.i-d-e.de/reviewers/guidelines/ .
8. Le Mans Université, Managing Editor des Journals der Text Encoding Initiative und Redaktionsmitglied der Zeitschrift des französischsprachigen DH-Verbandes Humanistica sowie bei DHCommons.
9. Universität Trier, Redaktionsmitglied bei Romanistik.de.
10. Universität zu Köln, Institut für Digital Humanities.
11. Universität zu Köln, CCeH, Mit-Herausgeber des Rezensionsjournals RIDE.

Bibliographie

American Historical Association (AHA, ed., 2015): “Guidelines for the Evaluation of Digital Scholarship in History”. https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-evaluation-of-digital-scholarship-in-history [letzter Zugriff 14. Januar 2018]

Assmann, Bernhard / Sahle, Patrick (2008): *Digital ist besser. Die Monumenta Germaniae Historica mit den dMGH auf dem Weg in die Zukunft – eine Momentaufnahme.* Norderstedt: Books on Demand.

Cohen, Daniel J. / Fragaszy Trojano, Joan (eds., 2012): “Closing the Evaluation Gap”, in: *Journal of the Digital Humanities* 1, 4. http://journalofdigitalhumanities.org/1-4/closing-the-evaluation-gap/ [letzter Zugriff 14. Januar 2018]

Deutsche Nationalbibliothek (ed., 2017): Definition des Begriffs “Digitale Publikation” und aktuelle Verwendung der Terminologie in der Deutschen Nationalbibliothek. Ergänzende Ausführungen im Rahmen der Diskussion “Zum Sammelauftrag der Deutschen Nationalbibliothek”. http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/definitionDigitalePublikation.pdf [letzter Zugriff 14. Januar 2018]

Digital Humanities im deutschsprachigen Raum (DHd, ed., 2016): “Digitales Publizieren”. Wolfenbüttel: HAB. Stand: 01.03.2016. DOI: 10.15499/dhd-wp.001 http://dx.doi.org/10.15499/dhd-wp.001 [letzter Zugriff 14. Januar 2018].

Henny, Ulrike / Neuber, Frederike / unter Mitarbeit von den Mitgliedern des IDE (2017):. *Criteria for Reviewing Digital Text Collections*, version 1.0. https://www.i-d-e.de/publikationen/weitere-

schriften/criteria-text-collections-version-1-0/ [letzter Zugriff 14. Januar 2018].

Henny, Ulrike [erscheint demnächst]: "Reviewing von digitalen Editionen im Kontext der Evaluation digitaler Forschungsergebnisse." In: *Zeitschrift für digitale Geisteswissenschaften* (ZfdG). *Sonderband 2: Digitale Metamorphose. Digital Humanities und Editionswissenschaft.* Hrsg. von Roland S. Kamzelak und Timo Steyer.

Hohls, Rüdiger (eds., 1996ff.): *H-Soz-Kult. Kommunikation und Fachinformation für die Geisteswissenschaften. Angeboten von Clio-online - Historisches Fachinformationssystem e.V. Berlin.* http://www.hsozkult.de/ [letzter Zugriff 14. Januar 2018].

Institut für Dokumentologie und Editorik e.V. (eds. 2014ff.): *RIDE. A Review Journal for Digital Editions and Resources*. Köln. http://ride.i-d-e.de/ [letzter Zugriff 14. Januar 2018].

Jäger, Georg (2001): "Von Pflicht und Kür im Rezensionswesen." In: *IASLonline Diskussionsforum. Wissenschaftliche Kommunikation in der Kontroverse*. Bayreuth / München, http://www.iasl.uni-muenchen.de/discuss/lisforen/jaerezen.html [letzter Zugriff 14. Januar 2018].

Jannidis, Fotis (1999): "Bewertungskriterien für elektronische Editionen", in: *IASL Diskussionsforum online*. http://iasl.uni-muenchen.de/discuss/lisforen/jannidis.htm [letzter Zugriff 14. Januar 2018].

Just, Thomas (2009): "Rezension von: Digital ist besser", in: *sehepunkte* 9, Nr. 3 http://www.sehepunkte.de/2009/03/14673.html [letzter Zugriff 14. Januar 2018].

Modern Language Association (MLA, ed., 2012): "Guidelines for Evaluating Work in Digital Humanities and Digital Media" https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media [letzter Zugriff 14. Januar 2018].

Pierazzo, Elena (2014): "Trusting Digital Editions? Peer Review and Evaluation of Digital Scholarship", in: dies. *Digital Scholarly Editing: Theories, Models and Methods*, 182-205. Hal Id: hal-01182162. http://hal.univ-grenoble-alpes.fr/hal-01182162 [letzter Zugriff 14. Januar 2018].

Presner, Todd (2012): "How to Evaluate Digital Scholarship", in: *Journal of Digital Humanities* 1, 4 (2012).

Raymond, Eric S. (2001): *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol: O'Reilly, http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ [letzter Zugriff 14. Januar 2018].

Rockwell, Geoffrey (2012): "Short Guide To Evaluation Of Digital Work", in: *Journal of Digital Humanities* 1, 4 (2012).

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 14. Januar 2018].

Schelbert, Georg (2017): "Digital kann mehr! Das neue Graphikportal", in: *blog.arthistoricum.net* https://blog.arthistoricum.net/beitrag/2017/11/23/digital-kann-mehr-das-neue-graphikportal/ [letzter Zugriff 14. Januar 2018].

Schöch, Christof (2017):. "Poston / Niles, eds., Folger Digital Texts", in: *Variants* 11, 2014, pp. 16-20. DOI: http://doi.org/10.5281/zenodo.13745 [letzter Zugriff 14. Januar 2018].

Schreibman, Susan / Mandell, Laura / Olsen, Stephen (eds., 2011): "Evaluating Digital Scholarship", in: *Profession.* DOI: 10.1632/prof.2011.2011.1.123

Yates, Kimberly (2008): "Creating a Prize for the Best Digital Editions / Online Archives.", in: *Scroll – Essays on the Design of Electronic Texts* 1, 1.

Computergestützte Film- und Videoanalyse

Burghardt, Manuel
manuel.burghardt@ur.de
Universität Regensburg

Heftberger, Adelheid
adelheidh@gmail.com
Brandenburgisches Zentrum für Medienwissenschaften, Potsdam

Müller-Birn, Claudia
clmb@inf.fu-berlin.de
Freie Universität Berlin

Pause, Johannes
johannes.pause@hotmail.de
Universität Luxemburg

Walkowski, Niels-Oliver
walkowski@nowalkowski.de
Berlin-Brandenburgische Akademie der Wissenschaften

Zeppelzauer, Matthias
Matthias.Zeppelzauer@fhstp.ac.at
Fachhochschule St. Pölten

Konzeption und Zielsetzung

Das Thema Bild- und Bewegtbildanalyse gewinnt auch in der – bis dato stark auf Text fokussierten – Digital Humanities-Community immer mehr an Bedeutung. Vor diesem Hintergrund wurde Anfang des Jahres eine dedizierte DHd-Arbeitsgruppe „Film und Video“ [1] gegründet. Nachdem die AG im Juli 2017 bereits ein erstes Symposium zum Thema „Film rechnen – Computergestützte Methoden in der Filmanalyse“ [2] in Regensburg abgehalten hat, soll als nächster Schritt ein entsprechendes Panel auf der DHd 2018 in Köln ausgerichtet werden, um in einem größeren Kreis die Grenzen und Möglichkeiten computergestützter Analyseverfahren für Film und Video zu diskutieren.

Wesentliche Herausforderungen bei der computergestützten Analyse von Film und Video ergeben sich dabei vor allem durch die komplexe mediale Struktur von Bewegtbildern, die sich in einer Vielzahl technischer Formate und einer enormen Bandbreite von Analyseansätzen und -perspektiven niederschlagen. Anders als bspw. für den Bereich Text, gibt es für die Filmanalyse bislang nur wenige etablierte Verfahren und Softwarelösungen zur Erschließung und Verarbeitung von Film- und Videodaten. Wenngleich etablierte Felder aus der Informatik – wie etwa *computer vision* oder *audio processing* – vielversprechende Ansätze für die computergestützte Analyse von Film und Video bereitstellen, so erfordern diese mitunter erhebliches informatisches Know-How.

Trotz dieser schwierigen Gemengelage ist – zumindest in Einzelprojekten – im Kontext computergestützter Forschung bereits zu einzelnen filmwissenschaftlichen Fragestellungen und Aspekten wie Schnittrhythmus, Farbstrukturen, Rollenverteilungen, Plotstrukturen, Filmsprache oder der globalen Zirkulation von Filmen geforscht worden. Im Rahmen des Panels sollen nach einer kurzen Vorstellung der entsprechenden DHd-AG in insgesamt fünf Impulsreferaten exemplarische Ansätze zur computerbasierten Analyse von Film und Video vorgestellt und mit dem Plenum diskutiert werden. Im Anschluss an die Panelvorträge findet dann eine allgemeine Diskussionsrunde mit den Vortragenden und dem Plenum statt, in der ggf. weitere digitale Analyseansätze ergänzt und zur Diskussion gestellt werden können. Weiterhin sollen strategische Ziele der AG „Film und Video“ vorgestellt und diskutiert werden. Durch diesen Austausch soll die Möglichkeit methodischer Überschneidungen mit anderen Forschungsfeldern evaluiert und diskutiert werden, um so einen Beitrag zur methodischen Profilierung computergestützter Analyseverfahren und zum Methodentransfer zwischen verschiedenen Forschungsfeldern insgesamt zu leisten.

Ablauf (Gesamtdauer 90 Minuten):

Teil 1) Einführung in das Thema computergestützte Film- und Videoanalyse und Kurzvorstellung der DHd-AG „Film und Video“ (10 Minuten)

Teil 2) Panelvorträge in Form von Impulsreferaten (Gesamtdauer 50 Minuten)

Teil 3) Moderierte Diskussion zu den folgenden Themen (30 Minuten):

- Grenzen und Möglichkeiten der in den Impulsreferaten gezeigten Ansätze
- Weitere Ansätze / Desiderate in der computergestützten Film- und Videoanalyse
- Strategische Ziele der DHd-AG „Film und Video“

Panelvorträge

Computergestützte Analyse von Filmsprache: Zwei Fallstudien

Manuel Burghardt, Universität Regensburg

Für die computergestützte Analyse von Filmen bieten sich auf der visuellen Ebene eine ganze Reihe von quantitativ erfassbaren Parametern an, etwa die Analyse von Einstellungslängen und -frequenzen, Farben, Personen oder Objekten. Weiterhin bietet die Filmsprache – die in Form von Untertiteln für eine Vielzahl von Filmen bereits als textuelle Transkription vorliegt – einen niederschwelligen Analysezugang, für den viele bestehende Tools und Methoden aus der computergestützten Sprach- und Literaturanalyse angewendet werden können (Burghardt & Wolff, 2016; Burghardt et al., 2016). Im Rahmen des Panelvortrags sollen überblicksartig zwei Ansätze zur Analyse von Film und Serien auf Basis der jeweiligen Sprachdaten vorgestellt werden. Die erste Fallstudie illustriert dabei Möglichkeiten der automatischen Berechnung von filmischen Strukturen, indem das Konzept von Marcus (1970)

mathematischer Dramenanalyse auf den Bereich von TV-Serien übertragen wird. Eine zweite Fallstudie zeigt auf, wie mithilfe computergestützter Methoden ein exploratives Tool für die Analyse von Filmen umgesetzt werden kann, welches es erlaubt einen Film nach bestimmten Figurennamen und Schlüsselwörtern zu durchsuchen. Das Ergebnis dieser Suche wird in einer interaktiven Visualisierung der Trefferszenen zur weiteren Analyse dargestellt.

Dem Film auf der Spur – Die computergestützte Auffindbarkeit von Bild, Clip, Stil

Adelheid Heftberger, Brandenburgisches Zentrum für Medienwissenschaften, Potsdam

Filmschaffende haben sich von Anfang bei ihrer Kollegenschaft sowie audiovisuellen Archiven bedient. Einerseits handelt es sich dabei um die Integration von Ausschnitten in eigene Filme (Heftberger, 2016), andererseits um die Übernahme stilistischer Verfahren. Auch aus einer Archiv-Perspektive ist es interessant, die Spuren von filmischer Überlieferung zu verfolgen. Solche Untersuchungen unterliegen bislang nach wie vor den Grenzen der menschlichen Recherchekapazität, der Verfügbarkeit von Quellen und auch Rollenzuschreibungen der beteiligten Institutionen (Heftberger, 2014). Im Panelvortrag sollen aus filmgeschichtlicher Perspektive Desiderate an die computergestützte Analyse formuliert werden, deren Umsetzung erst zögerlich erfolgt. Anhand einer Fallstudie soll das Potential aufgezeigt werden, dass sich aus der multimodalen Analyse für die Filmgeschichte und v.a. die Überlieferungs- und Rezeptionsgeschichte ergeben könnte.

Wikidata für die Zugänglichmachung von Forschungsdaten in den Filmwissenschaften

Claudia Müller-Birn, Freie Universität Berlin

Wissenschaftliche Publikationen in den Geisteswissenschaften haben oft nicht die Reichweite und Nachhaltigkeit, die für HerausgeberInnen und AutorInnen wünschenswert wäre. Die Ablieferung von Faktenwissen an eine breiter rezipierte und maschinenlesbare Wissensdatenbank wie *Wikidata* könnte ein Ansatz sein, diese Reichweite und Nachhaltigkeit zu erzeugen. Dafür müssen allerdings nutzerzentrierte Softwarelösungen entwickelt werden, die es erlauben, dieses Faktenwissen auch niederschwellig und korrekt zu übertragen (Breitenfeld et al., 2017). Am Beispiel der multilingualen Open Access Zeitschrift „Apparatus – Film, Media and Digital Cultures in Central and Eastern Europe“ soll demonstriert werden, wie ein solcher Workflow zur Übertragung von Faktenwissen aussehen könnte. Einerseits werden Einträge in *Wikidata* mit Referenzen (Bezug auf die jeweilige Publikation) versehen, andererseits werden bestehende Einträge auch mit neuen Informationen angereichert oder sogar neu erstellt, sofern sie nicht vorhanden sind. Der Panelbeitrag zeigt das Potential für die bessere Zugänglichmachung von Forschungsdaten in den Filmwissenschaften (Müller-Birn et al., 2017).

Licht und Schatten. Methodische Herangehensweisen an die Analyse von Farbstrukturen in Filmen

Niels-Oliver Walkowski, Berlin-Brandenburgische Akademie der Wissenschaften
Johannes Pause, Universität Luxemburg

Obgleich Film als stark visuell geprägtes Medium in weiten Teilen vom Spiel der Farben, ihrer Inszenierung und Dramaturgisierung lebt, stellte die Untersuchung von Farbstrukturen Wendy Everetts (2007) zufolge innerhalb der Filmtheorie vor 10 Jahren immer noch so etwas wie „the last great wilderness, the one remaining area yet to be explored, mapped, and charted“ dar. Diese Situation hat sich bisher nicht signifikant verändert. Allerdings macht es die Entwicklung digitaler Methoden in den Filmwissenschaften während der seitdem vergangenen Zeit möglich, der Problematik neue Impulse zu verleihen. Eine Reihe von Projekten der letzten Jahre versucht genau dies zu tun. Zu nennen sind hier unter anderem Brodbecks (2011) *Cinemetrics*, Bakers (2015) *Spectrum* sowie die Projekte *MovieBarcodes* (Anonymus, 2016), *MovieAnalyzer* (Burghardt et al., 2016), das *ACTION Toolkit* (Casey, 2014), das Projekt *FilmColors* (Flückiger, 2015) sowie die Arbeit von Pause und Walkowski (2017). Mit Ausnahme der letzten drei Beiträge entstammen diese Arbeiten anderen Forschungsfeldern als den Filmwissenschaften. Diese Situation sowie der Umstand, dass die computergestützte Analyse von Farbe in einem filmwissenschaftlichen Kontext ein so junges Forschungsfeld ist, lassen eine methodische Systematisierung und Diskussion verschiedener Ansätze der Farbanalyse sinnvoll erscheinen. Der Panelbeitrag wird eine solche Übersicht geben. Des Weiteren werden unterschiedliche Strategien vorgestellt, mit denen versucht wird eine Interpretation und Vermittlung der Analyseergebnisse zu unterstützen.

Automatisierte Verfahren für Analyse, Retrieval und Annotation von Video und Film

Matthias Zeppelzauer, Fachhochschule St. Pölten, Österreich

Die manuelle Annotation von Videodaten ist üblicherweise eine zeitaufwendige Tätigkeit, die aufgrund der subjektiven Bewertungen unterschiedlicher AnnotatorInnen leicht zu Inkonsistenzen führen kann, insbesondere wenn die zu annotierenden Konzepte großen semantischen Interpretationsspielraum bieten. Automatische Annotationsmethoden basierend auf automatischen Bild- und Videoanalysemethoden aus dem Bereich des maschinellen Sehens (*computer vision*) bieten eine vielversprechende Alternative, um den Annotationsvorgang einerseits zu beschleunigen und andererseits zu formalisieren und zu objektivieren. In der Vergangenheit wurden zunächst automatisierte Methoden für die Erkennung von Schnitten und Einstellungsübergängen entwickelt. Darauf aufbauend wurden dann Lösungen präsentiert für die automatische Segmentierung von Videos und Filmen in einzelne semantisch kohärente Szenen. Methoden aus dem Bereich der inhaltsbasierten Bildsuche können eingesetzt werden, um nach visuellen Mustern, z.B. häufig auftretende Motive, Objekte oder Personen zu suchen. Methoden der Bewegungsanalyse (*motion tracking*) können eingesetzt werden, um Kamerabewegungen oder Objektbewegungen automatisiert zu erfassen. Ergänzend zur Bildanalyse kann die akustische Analyse der Tonspur weitere interessante Einsichten in den Rhythmus und die Montage von Filmen geben. Im interdisziplinären Projekt *Digital Formalism* wurde ein breites Spektrum automatisierter Verfahren erfolgreich für die Analyse und die Annotation von Video- und Filmdaten eingesetzt (vgl. Mitrovic et al., 2010 / 2011; Zaharieva et al. 2010; Zeppelzauer et al. 2011a / 2011b). In diesem Impulsreferat soll das Potenzial automatischer inhaltsbasierter Analyseverfahren an kurzen Beispielen aus der Praxis illustriert werden.

Fußnoten

1. Mehr Informationen unter http://dig-hum.de/ag-film-und-video
2. Mehr Informationen unter https://dhregensburg.wordpress.com/2017/07/03/symposium-film-rechnen-computergestuetzte-methoden-in-der-filmanalyse-2/

Bibliographie

Anonymus (2016): *MovieBarcodes*. http://moviebarcode.tumblr.com/

Baker, **Dillon** (2015): *Spectrum*. http://dillonbaker.com/spectrum/

Breitenfeld, A. / Mackeprang, M. / Hong, M.-T. / Müller-Birn, C. (2017) **: "**Enabling Structured Data Generation by Nontechnical Experts". In: Burghardt, M., Wimmer, R., Wolff, C. & Womser-Hacker, C. (Hrsg.), *Mensch und Computer 2017 - Tagungsband.* Regensburg: Gesellschaft für Informatik e.V.. (S. 181-192).

Brodbeck, Frederic (2011): *Film Data Visualization*. http://cinemetrics.fredericbrodbeck.de/

Burghardt, Manuel / Wolff, Christian (2016): „Digital Humanities in Bewegung: Ansätze für die computergestützte Filmanalyse", in *Book of Abstracts der DHd-Konferenz*.

Burghardt, Manuel / Kao, Michael / Wolff, Christian (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)*.

Casey, Michael (2014): "ACTION: Audiovisual Cinematic Toolbox for Interactive Object-based Media Navigation". http://aum.dartmouth.edu/~action/index.html

Everett, Wendy (2007): "Mapping Colour. An Introduction to the Theories and Practices of Colour", in *Questions of Colour in Cinema. From Paintbrush to Pixel.* Hg. von Wendy Everett. Bern, S. 7-38.

Flückiger, Barbara (2015): „FilmColors". https://filmcolors.org/

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

Heftberger, Adelheid (2014): "Film archives and digital humanities – an impossible match? New job descriptions and the challenges of the digital era", in *MedieKultur – Journal of media and communication research* 30, 57, S. 135-153.

Marcus, Solomon (1970): *Mathematische Poetik*. Frankfurt am Main: Athenäum Verlag.

Mitrovic, Dalibor / Hartlieb, Stefan / Zeppelzauer, Matthias / Zaharieva, Maia (2010): „Scene Segmentation in Artistic Archive Documentaries", in *HCI in Work and Learning, Life and Leisure*, LNCS, vol. 6389, S. 400-410.

Mitrovic, Dalibor / Zeppelzauer, Matthias / Zaharieva, Maia / Breiteneder, Christian (2011): „Retrieval of VisualComposition in Film", in *Proceedings of the 12th International Workshop on*

Image Analysis for Multimedia Interactive Services, April 13-15, Delft, The Netherlands.

Müller-Birn, Claudia / Heftberger, Adelheid / Höper, Jackob / Walkowski / Niels-Oliver Sharing (2017): “Factual Knowledge from Research in Film and Media Studies using Wikidata”, presented at FORCE2017 Research Communication and e-¬Scholarship Conference, Berlin, Germany.

Pause, Johannes / Walkowski, Niels-Oliver (2017): “Dead and Beautiful: The Analysis of Colors by Means of Contrasts in Neo-Zombie Movies”, in *Book of Abstracts der DHd-Konferenz*.

Zaharieva, Maia / Zeppelzauer, Matthias / Breiteneder, Christian / Mitrovic, Dalibor (2010): “Camera Take Reconstruction”, in *Proceedings of IEEE Multimedia Modeling Conference*, Jan 6-8, 2010, Chongqing, China, S. 379-388.

Zeppelzauer, Matthias / Mitrovic, Dalibor / Breiteneder, Christian (2011a): “Cross-Modal Analysis of Audio Visual Film Montage”, in *Proceedings of 20th International Conference on Computer Communications and Networks*, Maui, USA.

Zeppelzauer, Matthias / Zaharieva, Maia / Mitrovic, Dalibor / Breiteneder, Christian (2011b): “Retrieval of Motion Composition in Film”, in *Digital Creativity*, 22(4), 219-234.

Der ferne Blick. Bildkorpora und Computer Vision in den Geistes- und Kulturwissenschaften - Stand - Visionen - Implikationen

Donig, Simon
simon.donig@uni-passau.de
Universität Passau, Deutschland

Handschuh, Siegfried
Siegfried.Handschuh@uni-passau.de
Universität Passau, Deutschland

Radisch, Erik
erik.radisch@uni-passau.de
Universität Passau, Deutschland

Rehbein, Malte
malte.rehbein@uni-passau.de
Universität Passau, Deutschland

Hastik, Canan
hastik@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Kohle, Hubertus
hubertus.kohle@gmail.com
Ludwig-Maximilians-Universität München, Deutschland

Ommer, Björn
ommer@uni-heidelberg.de
Ruprecht-Karls-Universität Heidelberg, Deutschland

Ausgangslage

Getrieben von einer fortschreitenden Computerisierung ihrer wissenschaftlichen Methodik und einer wachsenden Verfügbarkeit von digitalen Bildwerken ist in den letzten Dekaden in zahlreichen Disziplinen der Geistes- und Kulturwissenschaften ein erneuertes Interesse am Bild als Forschungsgegenstand und Erkenntnisinstrument zu beobachten.

So begreift sich etwa die (digitale) Kunstgeschichte in der Erweiterung explizit als digitale Bildwissenschaft (Kohle 2013), hat die Geschichtswissenschaft als “Visual History” das lange vernachlässigte Medium Bild als Quellengattung neben dem Text wiederentdeckt (Paul 2014), erfindet sich die Architekturgeschichte nicht zuletzt mit digitalen Instrumenten als Disziplin neu und haben sich junge Fächer wie die Material-Culture-Studies oder die Designgeschichte etabliert.

Alle diese Disziplinen begegnen sich in den Digital Humanities als einem geteilten Wissensraum, dessen Erkundung ganz wesentlich von den Möglichkeiten der Informatik mitbestimmt wird.

Die automatisierte Erschließung großer und größter Bildkorpora im Sinne eines effektiven Information Retrieval prägt dabei nach Jeffrey Schnapp die erste Welle der Digitalen Revolution (Schnapp 2011). Die zweite Welle - so wiederum Schnapp - müsse nun “qualitativ, interpretativ, erfahrungsbewusst, intuitiv erfassbar und schöpferisch” sein, um digitale Instrumente in den Dienst der Kernkompetenzen der Geistes- und Kulturwissenschaften stellen zu können: Achtsamkeit gegenüber Komplexität, Medienspezifik, histori-

schem Kontext, analytischer Tiefe, Kritik und Interpretation.

Ziele des Panels

Mit dem hier vorgeschlagenen Panel verorten wir den Stand der digitalen Bildforschung in dieser idealtypischen Abfolge. Forschende aus unterschiedlichen Disziplinen beleuchten den State-of-the-art im Bereich automatisierter Verfahren der Computer Vision sowie gegenwärtige Wege und Visionen ihrer zukünftigen Anwendung in den Geistes- und Kulturwissenschaften. Wir erkunden Herausforderungen, die neue Werkzeuge wie die automatisierte Bildanalyse an Instrumentenkritik und Methodentransparenz stellen und tragen zur Konkretisierung von Best Practices in diesem Bereich bei.

Gegenstand

Verfahren, Praktiken und Projekte

Von Merkmalserkennung (Pattern Matching) bis hin zu Ansätzen aus dem Bereich des maschinellen Lernens wie etwa Deep Learning haben Verfahren der Computer Vision in den letzten Jahren das Potential entwickelt, sich als disruptive Technologie in den Geistes- und Kulturwissenschaften zu erweisen. Dieser Umstand verdankt sich unter anderem raschen Fortschritten bei der verfügbaren Rechenleistung, aber auch einem zunehmend vereinfachten Zugang zu geeigneter Software sowie einer wachsenden Verfügbarkeit digitaler Bilddaten.

Im Rahmen des Panels fragen wir zunächst nach den sich rasch entwickelnden Instrumenten und werfen einen Blick auf Bereiche und Projekte in denen diese zur Anwendung kommen. Wir wollen erkunden, welche Möglichkeiten computerbasierte automatisierte Verfahren etwa der Objekt- und Merkmalserkennung für die Digitalen Geisteswissenschaften eröffnen (zu einer einführenden Diskussion transdisziplinärer Potentiale von Computer Vision und Kunstgeschichte vgl. (Bell & Ommer (2015) sowie dies. (2016)).

Zwischen datengetriebenen und interpretativ hermeneutischen Zugängen

Wir explorieren, ob und wie die anfangs postulierte Dichotomie von datengetriebenen und interpretativ-hermeneutischen Zugängen zur Generierung von Wissen (Drucker 2011) fortbesteht oder einem differenzierteren Modell Raum gegeben hat. Inwieweit lässt sich der am „Close Viewing“ geschulte Methodenapparat direkt auf Verfahren eines „Distant-Viewing“ (oder „Distant-Watching“ im Fall von Bewegtbildern) übertragen, oder inwieweit wird hier eine Wissenschaftstheorie des Digitalen noch zu entwickeln sein? Dies schließt die Frage einer adäquaten Instrumentenkritik ebenso ein, wie die nach Selektionsverfahren, Korpusbildung oder den Auswirkungen der Freigabepraxis von Bildmaterial durch Kulturinstinstitutionen (GLAM).

Die semantische Lücke, Annotation, Multimodalität

Die Analyse von “Big Image Data” auch im Bereich der Forschung wirft die Frage auf, wie Forschungsansätze in den letzten Jahren versucht haben, die “semantische Lücke” (Semantic Gap) im Umgang mit diesem Datenmaterial zu schließen.

Welche Rolle kommt etwa Normdaten (GND), Thesauri (ICONCLASS, AAT oder ULAN) oder Ontologien, die spezifisch für die Dokumentation kultureller Artefakte entwickelt wurden (CIDOC-CRM) bzw. Domäneontologien wie Neoclassica (Donig, Christoforaki, Handschuh 2016) zu?

Welche Rolle könnte alternativ die semantische Annotation (Oren, Möller, Scerri, Handschuh, & Sintek 2006) durch Zugänge etwa aus dem Bereich der Gamification und des Crowdsourcing wie in ARTIGO (Wieser, Bry, Bérard, Lagrange 2013) bieten?

Kann die Informatik hier mit neuen Verfahren zur Verknüpfung verschiedener digitaler Analyseansätze in multimodalen Artefakten – also etwa die Untersuchung der Verbindung von Bild und Text oder Bewegtbild und Ton – digitales Sehen ergänzen und verbessern? (Bruni, Tran, Baroni 2014), (Hiippala 2013).

Methodentransparenz & Best Practices

In der Tat wirft ein digitales Sehen, analog zu kritischen Debatten um “Close” und “Distant Reading” (Bonfiglioli & Nanni 2015) grundlegende epistemologische Fragen nach den Gemeinsamkeiten oder Unterschieden zwischen computergestützter Bildanalyse und in spezifischen Praktiken des Sehens geschulten Menschen auf. Erweitert etwa der ferne Blick nur die Geschwindigkeit und Menge dessen, was wir wahrnehmen können, oder verändert er unsere Wahrnehmung selbst?

Wir erkunden wie digitale Instrumente beschaffen sein müssen, um den Erfordernissen der Geistes- und Kulturwissenschaften nach Methoden-

transparenz zu genügen. (Für eine grundsätzliche Diskussion der Konsequenzen des Designs von Klassifizierungs- und Rankingmechanismen sowie verschiedene Formen von Opacity siehe Burell 2016). Wie kann etwa der Black Box Charakter von Neuronalen Netzen sinnvoll akkommodiert werden? Nicht zuletzt stellt sich die Frage nach dem Zusammenspiel von Instrument und Forschungspraxis. Reproduziert etwa ein (in sich vollständig transparenter) Klassifikator, der aber auf der Basis eines bestimmten Korpus trainiert worden ist, nicht letztlich spezifische kulturelle Konzepte und konfirmiert damit einen bestehenden Kanon?

Wir wollen uns diesen Fragen nicht verschließen, aber sie gleichermaßen rekontextualisieren wie über ihre forschungspraktischen Implikationen nachdenken. Wir werden dabei gerade mit jenen, die an der Weiterentwicklung dieser Instrumente arbeiten, diskutieren, wie Technologien gebaut und eingesetzt werden können, um ein emanzipatorisches und exploratives Potential zu entfalten.

In diesem Sinn wollen wir abschließend Ideen für Best Practices im Sinne eines methodisch angeleiteten Umgangs mit großen Bildkorpora zusammentragen, indem wir etwa nach der Rolle und Modellierbarkeit von Kontextualisierung oder der Rolle von Ontologien und Normdaten dabei fragen.

Impulsvorträge

Siegfried Handschuh, Universität Passau: Bild, mutimodaler Verbund und Automatisierung

Bildanalyse stellt eine besondere Herausforderung an die Fähigkeit von computerbasierten Verfahren Wissen zu generieren. Dass Bilder vollkommen ohne einen Bezug zu anderen Modi – hier besonders Text – vorkommen ist eher eine Ausnahme als die Regel. Der multimodale Verbund von Bild und Text kann so durch Verfahren aus dem Bereich des maschinellen Lernens ausgewertet werden, um Wissen zu erzeugen, das einen breiteren Kontext berücksichtigt. Das Impulsstatement thematisiert diesen Zusammenhang insbesondere eine Methodik für die Repräsentation, Annotation und (teil)automatisierte Entdeckung multimodaler Wissensbestände in großen digitalen Materialkorpora und fragt wie Techniken der Computer Vision gewinnbringend mit Verfahren aus dem Bereich der Verarbeitung Natürlicher Sprache zusammengebracht werden können.

Canan Hastik, Technische Universität Darmstadt: Wie viel Semantik ist nötig und wie viel Automatisierung ist möglich?

Computerbasierte Verfahren sind ein wichtiges Instrument, um die digitale Bildforschung effizienter zu gestalten und zu verstetigen. Es wäre jedoch ein großer Fehler, Verfahren des maschinellen Lernens lediglich dafür einzusetzen, bestehende theoretische und historische Kategorien zu definieren und die Automatisierung für Digital Humanities Experten voranzutreiben. Die Herausforderung besteht darin, den DH Forscher bei der Kuration und Entwicklung von aussagekräftigen, repräsentativen und authentischen Forschungskorpora zu unterstützen. Ontologien sind dabei elementar für die Semantifizierung von großen bildbasierten Korpora und unterstützen zudem die kollaborative Entwicklung und die Verknüpfung von Wissensbeständen. Automatisierte Verfahren gilt es insbesondere dafür einzusetzen, zeitgenössische Kultur neu zu sehen und zu interpretieren.

Hubertus Kohle, Ludwig-Maximilians-Universität München: Ähnlichkeitsbestimmung in der digitalen Kunstgeschichte

Das Statement widmet sich aus einer eher geisteswissenschaftlichen Perspektive heraus der digital gestützten Ähnlichkeitsbestimmung. Ähnlichkeit wird hier verstanden als Erkenntnismotor, der insbesondere in der Kunstgeschichte seinen Ort hat, die schon immer mit dem vergleichenden Sehen eine Methode des Erkenntnisgewinns besaß. Abzuwägen wird sein, ob hierfür eher eine Metadatenanalyse oder die direkte Bildadressierung oder eine Mischung aus beidem zielführend ist.

Björn Ommer, HCI/IWR Ruprecht-Karls-Universität Heidelberg:

Aktuelle Entwicklungen im Bereich des Maschinellen Lernens und der Computer Vision haben Algorithmen hervorgebracht, die Visual Retrieval mit einer zuvor ungeahnten Performanz ermöglichen. Dadurch ergeben sich für die digitalen Bildwissenschaften ganz neue Möglichkeiten große Korpora zu erschließen und zu analysieren. Allerdings bringen diese neuen informatischen Verfahren auch ihre ganz eigenen Beschränkungen mit sich, die insbesondere bei ihrer

Rekontextualisierung in den Geisteswissenschaften zutage treten. Dieser Vortrag wird das Potential einer interdisziplinären Kooperation von Informationswissenschaften und den Geisteswissenschaften diskutieren, grundlegende Beschränkungen beleuchten und mögliche Auswege präsentieren.

Malte Rehbein, Universität Passau: Moderation

Bibliographie

Bell, Peter / Ommer, Björn: Training Argus, Ansätze zum automatischen Sehen in der Kunstgeschichte, in: *Kunstchronik* 68, Nr. 8 (2015): 414–20.

Bell, Peter / Ommer, Björn: Digital Connoisseur? How Computer Vision Supports Art History, in: Aggujaro, A. & Albl, S. (eds.): Il metodo del conoscitore approcci, limiti, prospettive - Connoisseurship nel XXI secolo. Artemide, Roma 2016: 187-200.

Burrell, Jenna: How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, in: *Big Data & Society* 3, Nr. 1 (2016): 1-12. doi:10.1177/2053951715622512.

Bonfiglioli, Rudi / Nanni, Federico: From Close to Distant and Back: How to Read with the Help of Machines, in: Gadducci, Fabio / Tavosanis, Mirko (eds.): *History and Philosophy of Computing*. IFIP Advances in Information and Communication Technology. Springer, Cham, 2015: 87–100. doi:10.1007/978-3-319-47286-7_6.

Bruni, Elia / Tran, Nam Khanh / Baroni, Marco: Multimodal Distributional Semantics, in: *Journal of Artificial Intelligence Research* 49, Nr. 1 (2014): 1–47.

Donig, Simon / Christoforaki, Maria / Handschuh, Siegfried: Neoclassica - A Multilingual Domain Ontology. Representing Material Culture from the Era of Classicism in the Semantic Web, in: Bozic, B. / Mendel-Gleason, G. / Debruyne, C. / O'Sullivan, D. (eds.): *Computational History and Data-Driven Humanities. CHDDH 2016*. Cham: Springer, 2016: 41–53. https://link.springer.com/chapter/10.1007/978-3-319-46224-0_5.

Drucker, Johanna: Humanities approaches to graphical display, in: *Digital Humanities Quarterly* 5, Nr. 1 (2011): 1–21.

Hastik, Canan / Steinmetz, Arnd / Thull, Bernhard: Using CIDOC CRM for Audiovisual Art, in: Smite, R. / Manovich, L. / Smits, R. (eds.): Data Drift. Archiving Media and Data Art in the 21st Century. Riga: RIXC 2015: 59-71.

Hiippala, Tuomo: The Interface between Rhetoric and Layout in Multimodal Artefacts, in: *Literary and Linguistic Computing* 28, Nr. 3 (2013): 461–71. doi:10.1093/llc/fqs064.

Kohle, Hubertus: *Digitale Bildwissenschaft*. Boizenberg: Hülsbusch, Werner, 2013.

Manovich, Lev: Data Science and Digital Art History. In: International Journal for Digital Art History, n. 1, 2015: 12-35. doi:hp://dx.doi.org/10.11588/dah.2015.1.21631., 30.

Oren, Eyal / Möller, Knud / Scerri, Simon / Handschuh, Siegfried / Sintek, Michael: What are Semantic Annotations. Relatório Técnico. DERI Galway 9, 62 (2006).

Paul, Gerhard: „Visual History", in: *Docupedia-Zeitgeschichte* (13. März 2014). doi:http://dx.doi.org/10.14765/zzf.dok.2.558.v3.

Wieser, Christoph / Bry, François / Bérard, Alexandre / Lagrange Richard: ARTigo: building an artwork search engine with games and higher-order latent semantic analysis, in: *First AAAI Conference on Human Computation and Crowdsourcing*, 2013. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7634.

Die Summe geisteswissenschaftlicher Methoden? Fachspezifisches Datenmanagement als Voraussetzung zukunftsorientierten Forschens

Moeller, Katrin
katrin.moeller@geschichte.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland

Ďurčo, Matej
matej.durco@oeaw.ac.at
Austrian Center for Digital Humanities, Österreichische Akademie der Wissenschaften, Österreich

Ebert, Barbara
barbara.ebert@rfii.de
Leiterin des Rats für Informationsinfrastrukturen (RfII), Geschäftsstelle Göttingen

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

Rosenthaler, Lukas
lukas.rosenthaler@unibas.ch
and Service Center for the Humanities DaSCH, Universität Basel (DHLab) und Schweizerische Akademie der Geistes- und Sozialwissenschaften, Schweiz

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

Wuttke, Ulrike
Ulrike.Wuttke@gmx.net
Stellvertretende Vorsitzende der AG Datenzentren des DHd, Fachbereich Informationswissenschaften, Fachhochschule Potsdam, Deutschland

Wettlaufer, Jörg
jwettla@gwdg.de
Göttingen Centre for Digital Humanities, Georg-August Universität Göttingen, Deutschland

Zusammenfassung des Vorhabens

Die AG Datenzentren des DHd möchte ein Panel mit insgesamt fünf Diskussionspartnern und zwei Moderatoren veranstalten, um die Herausbildung fachbezogenen Datenmanagements im deutschsprachigen Raum in den digitalen Geisteswissenschaften weiter zu fördern sowie den Stand und die Perspektiven dieses Forschungsbereichs zu diskutieren. Das Panel reiht sich damit in den fachlich übergreifenden Organisationsprozess zum Datenmanagement ein. Im Mittelpunkt der Diskussion soll hier die Veränderung der Datenkultur in der Geisteswissenschaft stehen und damit die Frage, wie eine Konsolidierung, Vernetzung und fachliche Abrundung von Diensten die Anreize zur wissenschaftlichen Nutzung des Datenmanagements erhöhen können. Dies soll mit zwei Schwerpunktsetzungen erfolgen:

A. Zunächst möchten wir anhand eines Überblicks zu bestehenden Angeboten von Datenzentren informieren und aufzeigen, welche fachbezogenen Leistungen momentan oder in naher Zukunft von Datenzentren in einem Netzwerk verteilter DH-Strukturen eingebracht werden. Dieser Block macht die Schwerpunktsetzungen, Rollen und herausgehobene fachliche Dienste einzelner Datenzentren sichtbar, die über allgemeine Anforderungen an ein Datenzentrum hinausragen. Daran anschließend kann über die mögliche Ausgestaltung kooperativer Strukturen und die Zusammenarbeit innerhalb eines Netzwerkes der DH-Datenzentren diskutiert werden.

B. Weiterhin fragt das Panel grundlegend nach zentralen Aufgabenstellungen im Bereich von Standards aus den einzelnen Fachgebieten der Geisteswissenschaften: Welche fachlichen Standards müssen Datenzentren über die zuvor gezeigten Ansätze hinaus kooperativ entwickeln, um passfähige Daten für eine möglichst große Breite von Nachnutzungen interdisziplinär abzusichern? Wo liegen Möglichkeiten und wo Grenzen von fachlicher Standardisierung und Normalisierung über die einzelnen Fachdisziplinen der Geisteswissenschaft hinweg? Hierzu sollen Erkenntnisse und Erfahrungen aus abgeschlossenen bzw. weitgehend realisierten Projekten dazu dienen, abstrahierende Anforderungen aus fachwissenschaftlicher Perspektive zu benennen, die zugleich einen Mehrwert der Realisierung besitzen. Insgesamt geht es darum, die Notwendigkeit eines geisteswissenschaftlichen sowie fachspezifischen Angebots nachdrücklich unter Beweis zu stellen, zu etablierende Standards zu benennen und Desiderate zu erkennen.

Forschungsstand

Der digitale Wandel, das Entstehen, Archivieren, Erschließen und Nachnutzen großer Datenbestände verändert wissenschaftliche Forschung in ihren Grundlagen, Arbeitsweisen und Methodiken fundamental (OECD 2007; HRK 2014; RfII 2016; Allianz 2010; DFG 2015). Die zahlreichen Aspekte dieses Wandels beschäftigen Fach-Communities und Fachverbände in umfassender Weise mit Fragen des Datenmanagements, der Lizenzen, Rechte und des Datenschutzes, der technischen und inhaltlichen Langzeitarchivierung sowie Nachnutzung von Daten. Die intensiven Diskussionen um das Datenmanagement sind verbunden mit einem grundlegenden Nachdenken über künftige Aufgaben, Strukturen und Workflows bestehender und sich neu formieren-

der Institutionen oder Akteure und führen zu grundsätzlichen Reflexionen über die notwendige Schaffung verteilter nationaler infrastruktureller Angebote (RfII 2016: 39f.; RfII 2017). In der jüngeren Vergangenheit sind bevorzugt technische Aspekte des Datenmanagements diskutiert worden, wie sie etwa die interoperationale Metadatenhaltung (Bibliotheksparadigma), Probleme der Speicherformate, Speichermedien und Dokumentation in der technischen Langzeitarchivierung über einen Rahmen von zehn Jahren hinaus (Archivparadigma) betreffen (Nestor 2016). Solche Lösungen und Angebote können oft in übergreifenden technischen Strategien gefunden werden, die ein zentrales Aufgabengebiet von Bibliotheken und Archiven umfassen. Wie in den Arbeitsbereichen der digitalen Wissenschaft allgegenwärtig zu beobachten ist, lässt sich dabei eine sukzessive, aber dennoch intensive Spezialisierung und Differenzierung digitaler Erfordernisse und Techniken beobachten. War beispielsweise noch das erste große deutschsprachige Überblickswerk der nestor-Arbeitsgruppe zur Langzeitarchivierung fachübergreifend technisch geprägt (Neuroth 2010), offenbaren die neueren Arbeitsergebnisse eine deutliche Hinwendung zur fachbezogenen Darstellung (Neuroth 2012). Allerdings blieben auch in dieser Hinsicht die Parameter der technischen Langzeitarchivierung aus der Sicht der Archive und Bibliotheken bestimmend.

Gleiches lässt sich mit Blick auf die Wissenschaft erkennen, wenn man etwa auf die sich konstituierenden Fachgruppen im Rahmen des Verbandes der Digital Humanities im deutschsprachigen Raum [1] oder andere traditionelle Fachverbände mit ihren sich ausdifferenzierenden digitalen Arbeitsweisen/Methoden schaut. [2] Intensiv diskutiert wurden solche Themen der fachbezogenen Binnendifferenzierung auch in der Auftaktveranstaltung zum Förderprogramm "'Mixed Methods' in den Geisteswissenschaften?" der VW-Stiftung im Frühsommer 2017, die eine Vielzahl deutschsprachiger Forschungsprojekte im Bereich der Digital Humanities versammelte. [3]

Um nachhaltig wirksam zu sein, braucht Forschungsdatenmanagement eine enge Anbindung an die Wissenschaft mit ihren spezifischen Problemen und fachlichen Erfordernissen (RfII 2016: 60; RfII 2017: 1f., AG DZ 2017: 3). Nicht in erster Linie der technische Service zur Langzeitarchivierung standardisierter Daten, sondern die fachbezogene Datenkuration und die Tiefenerschließung auf der Basis von Standards sichert die Güte und Qualität von Forschungsdaten und ihre Nachnutzbarkeit für spezifische wissenschaftliche Forschungsprojekte. Qualitätsgerechtes Forschungsdatenmanagement - das auch im analogen Zeitalter den Kern jeder Forschung ausmachte - sichert den wissenschaftlichen Mehrwert von Daten, schafft vielfältige Nutzungsanreize und innerwissenschaftliche Akzeptanz (RfII 2016: 38-40; AG DZ 2017: 2). Im gleichen Maße wie daher die übergreifende Basis technischer Lösungen zur Metadatenhaltung, Lizenzierung und Langzeitarchivierung entwickelt werden muss, sind fachbezogene Standards der Datenkuration, der kontrollierten Vokabularien, der fachbasierten Quellenkritik und Annotation sowie spezialisierte Werkzeuge der fachlichen Datenverarbeitung zur Erzeugung nachhaltiger Datenstrukturen notwendig. Während erstere eher in den Verantwortungsbereich der Bibliotheken und Archive fallen, sind letztere das zentrale Aufgabenfeld der Wissenschaft mit ihren spezifischen Fachgebieten.

Hinzu treten die zahlreichen Herausforderungen des geisteswissenschaftlichen Datenmanagements im Unterschied zu natur- und sozialwissenschaftlichen Projekten. Geisteswissenschaftliche Daten bilden in vielen Anwendungsbereichen komplexe Textkorpora mit hochspezialisierten Formen der Transkription, Annotation und der Unterscheidung verschiedener Ebenen von Quellenbegriff, Lemmatisierung, Normalisierung, Standardisierung und Begriffserklärung. Ähnliche komplexe Formen der Erschließung gelten ebenso in eher auf Objekte und Gegenstände bezogene Geisteswissenschaften. Ansätze der einen Diziplin sind (bis heute) nicht ohne weiteres auf eine andere übertragbar bzw. nicht fachübergreifend bekannt. Gleichzeitig ist es Anliegen von Datenzentren übergreifende Standards bzw. Austauschformate zu definieren und zu etablieren, welche die Transparenz und Verschneidung von Daten überhaupt erst einmal ermöglichen.

Ebenso muss auf der Ebene der Datenzentren intensiv über Kooperationsformen und Vorgehensweisen für neue heterogene Datenbestände mit einem Mix aus Quellen, Daten, Methoden, Medien und Rechten nachgedacht werden (AG DZ 2017: 3).

Fragestellung und Aufbau des Panels

Die DH ist angetreten, um computergestützte Ansätze und Methoden quer über die geisteswissenschaftlichen Fachdisziplinen zu entwickeln. Was jedoch macht die Summe geisteswissenschaftlicher Bedürfnisse aus, um Daten tatsächlich interdisziplinär und langfristig nachnutzbar zu machen? Die "Kritik der digitalen Vernunft" besteht für ein solches geisteswissenschaftliches

Unterfangen darin, gemeinsame Anforderungen und Standards der Datenhaltung aus bestehenden Projekten und Erfahrungen zu formulieren, zusammenzutragen, zu systematisieren und zu abstrahieren. Zudem bedürfen solche Standards der Akzeptanz durch Forschende, sollen sie Wirksamkeit entfalten. Standards müssen daher einen klaren Mehrwert für wissenschaftliche Arbeit bieten.

1. Viele Datenzentren haben mittlerweile ihre reguläre fachliche Arbeit aufgenommen, projektbasiert Erfahrungen gesammelt und neben der regulären Arbeit oft spezifische Schwerpunkte gebildet. In Form einer Präsentation der wichtigsten Akteure, Dienste und Angebote möchten wir über das bestehende Leistungsangebot innerhalb der AG Datenzentren des DHd-Verbandes informieren, wobei bevorzugt die Schwerpunktsetzungen im Bereich der Standards thematisiert werden. Auf diese Weise können sich Tagungsteilnehmer einen Überblick zum Dienstleistungsspektrum von Datenzentren verschaffen, was angesichts der insgesamt unübersichtlichen (weil sich dynamisch entwickelnden) Forschungslandschaft einen erheblichen Vorteil darstellt. Diese ca. 15-20 Minuten dauernde Präsentation beruht auf einer Umfrage unter den im Verband organisierten Einrichtungen und Akteuren, die von der Organisatorin des Panels in Zusammenarbeit mit der AG Datenzentren vorstrukturiert und präsentiert wird. Bei Bedarf können weitere zentrale fachwissenschaftliche Angebote in den Geisteswissenschaften einbezogen werden. Dieser Leistungskatalog ist einerseits Grundlage der Diskussion mit dem Publikum und andererseits für die Außendarstellung der AG Datenzentren der DHd förderlich.

2. Nach dieser Präsentation leitet die Moderation in den Diskussionsteil über. Eingeladen werden Vertreter und Vertreterinnen geisteswissenschaftlicher Datenzentren mit unterschiedlichen Schwerpunktsetzungen und Funktionen, die Impulsreferate zu notwendigen Anforderungen fachwissenschaftlicher Standards im Datenmanagement zu geben. Diese Statements werden vor der Tagung allen Diskussionspartnern zur Verfügung gestellt, um eine angeregte Diskussion über fachliche Belange und Standards für das Datenmanagement der Geisteswissenschaften zu ermöglichen und auch Kritik zu formulieren. Die Panelorganisatorin und Moderatoren sichten die Beiträge zuvor redaktionell-moderierend. Hier sollen vor allem Möglichkeiten und Grenzen von bestehenden innerfachlichen Standards (z.B. existierende Ansätze wie GND, XML, TEI, Transkriptionsregeln; neue Entwicklungen wie z.B. "Ontologie historischer Berufe” etc.) benannt und diskutiert werden. An Einzelbeispielen soll kontrovers diskutiert werden, welche Standards die Geisteswissenschaft und das Datenmanagement tatsächlich voranbringen. Sind es eher technische Aspekte und Rahmenbedingungen (Tools, Formate, Annotationswerkzeuge), die formale Aspekte des Datenmanagements vereinheitlichen oder thematisch übergreifende Normansätze, Ontologien oder kontrollierte Vokabulare, die häufig aber sehr voraussetzungsvoll und arbeitsintensiv sind? Zielgerichtet sollen so Ansätze, Möglichkeiten und Grenzen diskutiert werden. Fachwissenschaftliche Aspekte sollen zusätzlich durch die Öffnung der Diskussion für das Publikum einbezogen werden. Gleichzeitig kann das Publikum Hindernisse und Desiderate des Forschungsdatenmanagements formulieren. Insgesamt sollen zudem mögliche Kooperationen und Organisationsstrukturen zwischen verschiedenen Datenzentren angerissen werden (Wer darf Standards entwickeln? Wer muss sie nutzen?). Nach Möglichkeiten werden dazu Beiträge des Publikums bzw. die Use Cases der Diskussion aufgegriffen. Diese Diskussion soll die fachliche Positionierung und Ausdifferenzierung der Datenzentren schärfen.

Es ist die Aufgabe der Moderation, entsprechend der Schwerpunktsetzungen in der Diskussion abstrahierende Aussagen zur Ansetzung von Standards zu finden, die aber zuvor innerhalb der AG und auch der eingeladenen Diskutanten bereits andiskutiert wurden, um ein konzentriertes und fachbezogen-klares Fazit der Diskussion zu ermöglichen.

Zusammensetzung des Panels

Organisation und Präsentation:

Dr. Katrin Moeller, Historisches Datenzentrum Sachsen-Anhalt (Hist-Data), Martin-Luther-Universität Halle-Wittenberg, Deutschland

Moderation:

Dr. Ulrike Wuttke, Stellvertretende Vorsitzende der AG Datenzentren des DHd, Fachbereich Informationswissenschaften, Fachhochschule Potsdam, Deutschland

Dr. Jörg Wettlaufer, Göttingen Centre for Digital Humanities, Georg-August Universität Göttingen, Deutschland

Datenzentren:

Marina Lemaire, M.A., Servicezentrum eSciences, Universität Trier, Deutschland

DI Matej Ďurčo, Austrian Center for Digital Humanities, Österreichische Akademie der Wissenschaften, Österreich

Dr. Barbara Ebert, Leiterin des Rats für Informationsinfrastrukturen (RfII), Geschäftsstelle Göttingen

Prof. Dr. Lukas Rosenthaler, Data and Service Center for the Humanities DaSCH, Universität Basel (DHLab) und Schweizerische Akademie der Geistes- und Sozialwissenschaften, Schweiz
apl. Prof. Dr. Patrick Sahle, Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Fußnoten

1. Siehe die Arbeitsgruppen des Verbands DHd: https://dig-hum.de/dhd-ags .
2. Vergleiche hier etwa das CfP und die daraus resultierenden Ergebnisse der Tagung: Quellen und Methoden der Geschichtswissenschaft im digitalen Zeitalter. Neue Zugänge für eine etablierte Disziplin, DIGIMET 2017, 25./26.09.2017 in Berlin, URL: http://welt-der-kinder.gei.de/wp-content/uploads/2017/05/CfP-DH-Abschlusstagung-WdK-02.05.2017.pdf.
3. Dabei nahmen zahlreiche Vertreter der DH-Community in den Geisteswissenschaften teil, die auch über das eigentliche Förderprogramm "Mixed Methods" in den Geisteswissenschaften Projekte durchführen: https://www.volkswagenstiftung.de/fileadmin/downloads/publikationen/20161221_Mixed_Methods_Volkswagen Stiftung_Bewilligungen.pdf.

Bibliographie

AG Datenzentren des DHd (Hg.) (2017): Stellungnahme der DHd AG Datenzentren und des DHd-Verbands zur Nationalen Forschungsdateninfrastruktur (NFDI), URL: *http://dig-hum.de/sites/dig-hum.de/files/DHd_NFDI_Stellungnahme_2017-07-31.pdf* [letzter Zugriff 22. September 2017].

Allianz der deutschen Wissenschaftsorganisationen (Hg.) (2010): Grundsätze zum Umgang mit Forschungsdaten, URL: *http://www.allianzinitiative.de/fileadmin/user_upload/www.allianzinitiative.de/Grundsaetze_Forschungsdaten_2010.pdf* [letzter Zugriff 22. September 2017].

DFG (Hg.) (2015): DFG-Leitlinien zum Umgang mit Forschungsdaten, Bonn 2015, *http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/index.html* [letzter Zugriff 22. September 2017].

Nestor (Hg.) (2016): Standardisierung, 2016, URL: *https://wiki.dnb.de/display/NESTOR/Standardisierung* [letzter Zugriff 22. September 2017].

Neuroth, Heike / Oßwald, Achim / Scheffel, Regine / Strathmann, Stefan / Huth, Karsten (Hg.): nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3, Göttingen 2010, URL: *http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf* [letzter Zugriff 22. September 2017].

Neuroth, Heike / Strathmann, Stefan / Oßwald, Achim / Scheffel, Regine / Klump, Jens / Ludwig, Jens (Hg.) (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Boizenburg / Göttingen.

OECD (Hg.) (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding, Paris 2007, URL: *http://www.oecd.org/sti/scitech/38500813.pdf* [letzter Zugriff 22. September 2017]. **OECD (Hg.)** (2016):, "Research Ethics and New Forms of Data for Social and Economic Research", *OECD Science, Technology and Industry Policy Papers*, No. 34, OECD Publishing, Paris, URL: *http://dx.doi.org/10.1787/5jln7vnpxs32-en* [letzter Zugriff 22. September 2017] .

Hochschulrektorenkonferenz (HRK) (Hg.) (2014): Management von Forschungsdaten als strategische Aufgabe der Hochschulleitungen, 14. Mai 2014, URL: *https://www.hrk.de/positionen/beschluss/detail/management-von-forschungsdaten-eine-zentrale-strategische-herausforderung-fuer-hochschulleitungen/* [letzter Zugriff 22. September 2017].

Rat für Informationsinfrastrukturen (RfII) (Hg.) (2016): Leistung aus Vielfalt, Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, URL: *http://www.rfii.de/de/category/dokumente/* [letzter Zugriff 22. September 2017].

Rat für Informationsinfrastrukturen (RfII) (Hg.) (2017): Schritt für Schritt - oder: Was bringt wer mit? Ein Diskussionsimpuls zur Zielstellung und Voraussetzungen für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI), URL: *http://www.rfii.de/de/category/dokumente/* [letzter Zugriff 22. September 2017].

Gute Forschungsdaten, bessere Forschung: wie Forschung durch Forschungsdaten-management unterstützt wird

Mache, Beata
mache@sub.uni-goettingen.de
Staats- und Universtitätsbibliothek Göttingen

Trippel, Thorsten
thorsten.trippel@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

Effinger, Maria
effinger@ub.uni-heidelberg.de
Universitätsbibliothek Heidelberg

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Otto-Friedrich-Universität Bamberg

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

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

Horstmann, Wolfram
horstmann@sub.uni-goettingen.de
Staats- und Universtitätsbibliothek Göttingen

Müller, Lydia
lydia@informatik.uni-leipzig.de
Universität Leipzig

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

Teich, Elke
e.teich@mx.uni-saarland.de
Universität des Saarlandes

Kurzzusammenfassung

In diesem Panel geht es um die Förderung der geisteswissenschaftlichen Forschung durch eine planvolle Erhebung, Archivierung, Veröffentlichung und die dadurch ermöglichte Nachnutzung von Forschungsdaten, die sowohl zur Qualitätssicherung in der Forschung beitragen als auch nicht zuletzt neue Fragestellungen erlauben. Aus unterschiedlichen Perspektiven soll in dem Panel beleuchtet werden, welchen Mehrwert das Datenmanagement für die Forschung in den digitalen Geisteswissenschaften hat, wie man diesen Mehrwert erreicht und auch die Veröffentlichung der Forschungsdaten als ein selbstverständliches Element der Dissemination der Forschungsergebnisse etabliert und wie man gleichzeitig den Aufwand für die Forschung abschätzen kann.

Ziele des Panels

Vor dem Hintergrund der notwendigen Diskussion zum Management, zur Interoperabilität, Sicherung, Publikation und Zitation aber auch zur Auswahl und Kassation von Forschungsdaten wird das Panel Experten aus der Forschung, Entwicklung, aus wissenschaftsgeleiteten Forschungsinfrastrukturen sowie dem Bibliotheks- und Universitäts-Verlagswesen zusammenbringen. Im Fokus steht die kritische oder auch kontroverse Auseinandersetzung, welche Aufgaben Forschende selbst übernehmen können und wollen, welche Aufgaben bei digitalen Infrastrukturen anzusiedeln sind, welche gemeinsamen Strategien aller Mitwirkenden nötig sind, wo eine Parallelisierung von Entwicklungen vermieden werden muss und wo eine Vielfalt von Ansätzen wünschenswert wäre.

Dazu werden die Teilnehmenden des Panels anhand von vier Leitfragen aus ihrer jeweiligen Perspektive kurz Stellung zum Thema nehmen.

Leitfragen

- Welche spezifischen Anforderungen aus der geisteswissenschaftlichen Praxis beeinflussen das Datenmanagement?
- Wie beeinflusst die Datenveröffentlichung herkömmliche Publikationen und welche Da-

tenmanagementoptionen sind für die Publikationen von Daten relevant?
- Wieweit können Forschungsdatenmanagementpläne die Forschenden dabei unterstützen, schon vom Moment der Formulierung der Forschungsfrage und der Beantragung eines Projektes an für die sichere forschungsbegleitende und forschungsabschließende Veröffentlichung der gesammelten, erhobenen und erstellten Daten – möglichst den FAIR-Prinzipien konform – zu sorgen?
- Wie bestimmen wir die technischen Kriterien für die Sicherung und Interoperabilität der heterogenen Forschungsdaten, damit sie gesichert, gefunden und nachgenutzt werden können? Reichen die Kriterien zur Abschätzung des Aufwands aus?

Hintergrund

Geisteswissenschaftliche Forschung basiert darauf, Ideen und Konzepte anderer Forschender zu betrachten, zu kommentieren und weiter zu entwickeln. Im Rahmen des Paradigmenwechsels (siehe etwa Berry, 2011; Baum and Stäcker 2015 ; Thiel, 2012), der sich gerade auch in den digitalen Geisteswissenschaften durch teilautomatische Analysen (siehe bereits Busa, R., 1951), Visualisierungen und Verknüpfungen unterschiedlicher Datentypen manifestiert, ist die Nachnutzung von Daten, Informationen und Konzepten von Anfang an ein integraler Bestandteil der Forschung gewesen (siehe etwa den Impact von Busa beschrieben von Winter, 1999; die Daten von Busa unter http://www.corpusthomisticum.org arbeiten von Antonio Zampolli, 1973; etc.)

Die Veröffentlichung von Forschungsdaten selbst ist kein Novum - in Anhängen, in Tabellen, in Abbildungen wurden “Daten” analog als Merkmale der Wissenschaftlichkeit einer Publikation veröffentlicht, und selbst Busa (ebd.) referenziert existierende Konkordanzen und Indizes, die allerdings noch nicht elektronisch vorlagen. Die elektronische Publikation und Weitergabe ist auch nicht neu: Bereits in den 1990er Jahren fanden sich elektronische Beilagen auf CD-ROM in Verlagspublikationen und auch bei Hausarbeiten; Webseiten enthielten schon früh Forschungsdaten [1], die – zumindest während der Projektlaufzeit – diese auch für andere Forschende verfügbar machten.

Die Publikation über öffentlich zugängliche Datenrepositorien ist dagegen recht neu, d.h. in zentralen Einrichtungen, die auch über Projektlaufzeiten hinaus Forschungsdaten vorhalten können, dafür aber eine Übergabe durch Datenbereitsteller nach bestimmten Qualitäts- und Beschreibungsrichtlinien verlangen. Dabei erlauben Repositorien, extrem große Mengen von (vernetzten) Daten zu veröffentlichen: alle erhobenen Daten können publiziert werden. Darunter können auch Daten sein, die für die eigene Forschungsfrage letztendlich nicht relevant waren, aber den Forschungsweg und die Ergebnisse nachvollziehbar machen, also sogenannte “null-results”. Innerhalb von Enhanced Publications können sie referenziert werden und z.B. mittels Linked Data als dynamisch vernetzende Daten aufeinander verweisen. Ihre Nachnutzung erlaubt es, Analysen, die in der Vergangenheit nur als Gedankenexperimente möglich waren, praktisch durchzuführen.

Auch deswegen setzen Forschungsförderer zunehmend voraus, dass ein Konzept zur Nachnutzung der Daten, auf denen Forschungsergebnisse beruhen, zusammen mit Projektanträgen eingereicht wird, und dass zudem bereits existierende Daten und Werkzeuge gegebenenfalls nachgenutzt werden (siehe Senat der DFG, 2015; Deutsche Forschungsgemeinschaft, 2013; Deutsche Forschungsgemeinschaft, 2009; Deutsche Forschungsgemeinschaft, undatiert; H2020 Programme, 2013; Allianz der deutschen Wissenschaftsorganisationen, 2010). Gleichzeitig wird durch die Attribution der Datenaufbereitung die Leistung derjenigen sichtbar, die die Daten bereitgestellt haben, so dass die Veröffentlichung “infrastrukturbezogene[r] wissenschaftliche[r] Leistungen” und Forschungsdaten “in Qualifikationsverfahren ergänzend anerkannt werden” können (vgl. Wissenschaftsrat, 2011). Die Bereitstellung von Forschungsdaten folgt dabei den sogenannten FAIR-Prinzipien [2], ist Teil der sich entwickelnden Open-Science-Landschaft und wird etwa von wissenschaftsgeleiteten Forschungsinfrastrukturen wie CLARIN-D (siehe Hinrichs und Trippel, 2017) und DARIAH-DE (Gradl und Henrich, 2016) unterstützt.

Auch wenn der Mehrwert des Forschungsdatenmanagements in diesem Bereich dokumentiert ist, als Anforderung formuliert wird und positiv konnotiert ist (siehe Bargheer u.a. 2017), erscheint es einigen Forschenden trotzdem eher als aufgezwungene, lästige Notwendigkeit denn als Mehrwert für die Forschung in den DH. Auch deshalb sollen in dem Panel die Vertreter der Forschungsinfrastrukturen und die WissenschaftlerInnen reflektieren, ob die bisher an die Forschenden herangetragenen Angebote ausreichen, um ihre Anforderungen zu erfüllen und ihr Vertrauen zu gewinnen, und, was die Forschungsinfrastuktu-

ren tun können, um hier weitere Fortschritte zu erzielen.

Panel Format

Das Panel besteht aus drei Teilen:

- Kurze Impulsvorträge von max. 5 Minuten von den Panel-Teilnehmenden unter Einnahme ihrer jeweiligen Perspektive
- Diskussion der Leitfragen zwischen den Panelteilnehmenden moderiert aus der Anwenderperspektive
- Öffnung der Diskussion für die Zuhörenden durch die Moderation

Panel-Teilnehmende:

Moderatorin

Elke Teich (Saarbrücken): Als Sprecherin eines interdisziplinären SFB verfügt Prof. Teich über eine hervorragende Kenntnis der Erfordernisse und Hindernisse im Forschungsdatenmanagement in großen Forschungsverbünden. Sie wird das Panel aus der spezifischen Perspektive der Geisteswissenschaften moderieren.

Teilnehmende auf dem Panel und Impulsvorträge

- Lydia Müller (Mind Research Repository, Universität Leipzig): In den Kognitionswissenschaften ist die Verwaltung von Daten im Zusammenhang mit der Publikation von Artikeln zusammen mit den zugrundeliegenden Daten teilweise gelebte Praxis. Lydia Müller ist als Maintainerin des Mind Research Repository (http://openscience.uni-leipzig.de/) an einer Schlüsselstelle, da dieses Verzeichnis die Integration von Daten und Publikationen ermöglicht. Vor dem Hintergrund der dort gehosteten Daten und Publikationen kennt sie die Fallstricke bei der Integration von Daten und Publikationen und weiß, wie die Daten nachhaltig und gesichert abgelegt werden können und hat Erfahrungen damit, auf diese Daten zu verweisen. Sie steht damit für einen Aspekt der FAIR-Prinzipien für Forschungsdaten im Rahmen des Forschungsdatenzyklus in den Geisteswissenschaften: der Aufbewahrung und Zugänglichmachung von Daten.
- Susanne Haaf (BBAW) ist am Aufbau und Betrieb des Deutschen Textarchivs (DTA) beteiligt und hat die Spezifikation des standardisierten DTA-Basisformats maßgeblich mit beeinflusst, das Eingang in die einschlägigen DFG-Richtlinien gefunden hat. Damit war sie an der Erstellung von geisteswissenschaftlichen Forschungsdaten selbst beteiligt und entwickelte aktiv Standards und Richtlinien, die für große Datenmengen eingesetzt wurden. Daneben vermittelte sie die zugrundeliegenden Workflows, Richtlinien und Technologien im Rahmen von Workshops und Lehrveranstaltungen innerhalb der DH-Nutzercommunity. Das DTA vertritt mehr als 2000 Nutzende, die zur Weiterentwicklung seines bislang einzigartigen Korpus interoperabler Forschungsdaten des historischen Deutschen für die Geisteswissenschaften beitragen.
- Torsten Schrade (Leiter der Digitalen Akademie der Akademie der Wissenschaften und der Literatur Mainz, Professor für Digital Humanities an der Hochschule Mainz): Torsten Schrade ist ausgewiesener Experte für Forschungsdatenmanagement und Webtechnologien für die geisteswissenschaftliche Grundlagenforschung, für historische Fachinformationssysteme, webbasierte Arbeitsumgebungen, digitale Editionen, Webservices für geisteswissenschaftliche Fachdaten und Semantic Web Anwendungen.
- Thorsten Trippel: Im Rahmen von CLARIN-D und dem europäischen Projekt Parthenos beschäftigt sich Thorsten Trippel mit der Erstellung und Umsetzung von Datenmanagementpläne für die geisteswissenschaftliche Forschung. Zusammen mit Antragstellern entwickelt er dazu Datenmanagementpläne, wie sie von Drittmittelgebern zunehmend gefordert sind, und begleitet die Projekte über alle Phasen der geisteswissenschaftlichen Forschung - vom Erstellen oder Auffinden von Forschungsdaten, über die Aufbereitung, Analyse, Bereitstellung bis zur Archivierung, um die FAIR-Prinzipien umzusetzen.
- Tobias Gradl: Im Rahmen seiner Forschung hat sich Tobias Gradl innerhalb von DARIAH mit der nachhaltigen Aufbereitung von Forschungsdaten beschäftigt, besonders mit Daten, die nicht standardkonform aufbereitet wurden und im Rahmen spezifischer Forschungsfragestellungen anfallen. Daher bearbeitet er Modelle, um Daten in einer nachhaltig erschlossenen Form bereitzustellen.
- Maria Effinger (UB Heidelberg): Im Rahmen des DFG-Projekts OA-Monos, des Universitätsverlags "Heidelberg University Publishing" und der Fachinformationsdienste "arthistoricum.net" und, "Proyplaeum" hat Maria Effinger XML-basiertes Publizieren implementiert und in der Universitätsbibliothek in Heidel-

berg umgesetzt. Die Veröffentlichungsmöglichkeiten über universitätseigene Verlage nehmen für offene Publikationen und damit verbundene Daten eine wichtige Funktion ein, indem sie in einer engen Zusammenarbeit mit den Forschenden die prinzipielle Nachvollziehbarkeit und Überprüfbarkeit wissenschaftlicher Ergebnisse sichern. Diesen Aspekt wird Frau Effinger im Panel vertreten.

Fußnoten

1. Zum Beispiel gab es schon im Jahr 2000 auf der Webseite http://wwwhomes.uni-bielefeld.de/gibbon/EGA/ Sprachdokumentationsdaten der Sprache Ega in Côte d'Ivoire. Diese Seite kann als typisches Beispiel für Projektwebseiten mit Daten angesehen werden, deren Nachhaltigkeit durch das Engagement des betreibenden Einzelwissenschaftlers bestimmt wird.
2. Findable, Accessible, Interoperable, und Reusable, siehe h ttps://www.force11.org/group/fairgroup/fairprinciples

Bibliographie

Allianz der deutschen Wissenschaftsorganisationen (2010): "Grundsätze zum Umgang mit Forschungsdaten". http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze.html [letzter Zugriff 15. Januar 2017].

Bargheer, Margo, / Zeki, Mustafa Dogan / Horstmann, Wolfram / Mertens, Mike / Rapp, Andrea (2017): "Unlocking the digital potential of scholarly monographs in 21st century research," in: *Liber Quaterly 27*(2).

Baum, Constanze / Stäcker, Thomas (2015): "Methoden – Theorien – Projekte," in: *Zeitschrift für digitale Geisteswissenschaften, 1*(Sonderband: Grenzen und Möglichkeiten der Digital Humanities), 4-12. doi:DOI 10.17175/sb001_023

Berry, David M (2011): "The Computational Turn: Thinking about the Digital Humanities", in: *Cultural Machine, 12*, 1-22. http://www.culturemachine.net/index.php/cm/article/view/440

Busa, Roberto (1951): *Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum*. Milano: Fratelli Bocca.

Deutsche Forschungsgemeinschaft (2013): "Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“". *Sicherung guter wissenschaftlicher Praxis*. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf [letzter Zugriff 15. Januar 2017]

Deutsche Forschungsgemeinschaft (DFG): "Umgang mit Forschungsdaten: DFG-Leitlinien zum Umgang mit Forschungsdaten". http://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/antragstellung/nachnutzung_forschungsdaten/ [letzter Zugriff 15. Januar 2017]

Deutsche Forschungsgemeinschaft (DFG) Ausschuss für wissenschaftliche Bibliotheken und Informationssysteme (Unterausschuss für Informationsmanagement) (2009)" *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Bonn: Deutsche Forschungsgemeinschaft. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf [letzter Zugriff 15. Januar 2017]

Geyken, Alexander (2011): " *Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv*," in: Perspektiven einer corpusbasierten historischen Linguistik. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“, Berlin, Germany.

Geyken, Alexander / Haaf, Susanne / Boenig, Matthias / Thomas, Christian / Wiegand, Frank (seit 2007): Deutsches Text Archiv (DTA). http://www.deutschestextarchiv.de/ [letzter Zugriff 15. Januar 2017]

Gradl, Tobias / Henrich, Andreas (2016): "Die DARIAH-DE-Föderationsarchitektur – Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen," in: *Bibliothek - Forschung und Praxis, 40*(2), 222-228. doi:10.1515/bfp-2016-0027

Hinrichs, Erhard / Trippel, Thorsten (2017): "CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften," in: *Bibliothek - Forschung und Praxis, 1*(41).

Senat der Deutschen Forschungsgemeinschaft (DFG) (2015): "Leitlinien zum Umgang mit Forschungsdaten. https://www.wissenschaftsrat.de/download/archiv/10466-11.pdf [letzter Zugriff 15. Januar 2017].

Thiel, Thomas (2012): "Digital Humanities: Eine empirische Wende für die Geisteswissenschaften?", in: *Frankfurter Allgemeine Zeitung*, 24.07.2012. http://www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html [letzter Zugriff 15. Januar 2017].

Winter, Thomas Nelson (1999): "Roberto Busa, S.J., and the Invention of the

Machine-Generated Concordance," in: *Faculty Publications, Classics and Religious Studies Department*. http://digitalcommons.unl.edu/classicsfacpub/70/ [letzter Zugriff 15. Januar 2017].

Wissenschaftsrat (2011): *Übergreifende Empfehlungen zu Informationsinfrastrukturen* Vol. Drs. 10466-11. https://www.wissenschaftsrat.de/download/archiv/10466-11.pdf [letzter Zugriff 15. Januar 2017].

Zampolli, Antonio (1973): "Humanities Computing in Italy," in: *Computers and the Humanities, VII*(6), 343-360. doi:10.1007/BF02395110

musilonline - integral lösen. Dialogfeld Digitale Edition

Bosse, Anke
anke.bosse@aau.at
Robert-Musil-Institut, AAU Klagenfurt

Fanta, Walter
walter.fanta@aau.at
Robert-Musil-Institut, AAU Klagenfurt

Godler, Katharina
katharina.godler@aau.at
Robert-Musil-Institut, AAU Klagenfurt

Brüning, Gerrit
Bruening@em.uni-frankfurt.de
Goethe-Universität Frankfurt / Freies Deutsches Hochstift

Boelderl, Artur
artur.boelderl@aau.at
Institut für Germanistik, AAU Klagenfurt

Ausgangslage

Digitale Editionen haben sich bereits als geeignete Publikationsform für die Präsentation von umfangreichen Textbeständen im Bereich des kulturellen Erbes etabliert. Für ein so umfangreiches und komplexes textgenetisches Korpus wie Robert Musils literarischen Nachlass und die daran entwickelte Datenstruktur der Klagenfurter Ausgabe stehen jedoch keine fertigen Modelle zur Verfügung. Im Rahmen des Panels soll einerseits diskutiert werden, welche Kriterien eine digitale Edition erfüllen muss, um eine Grundlage zur Erforschung von Robert Musils Gesamtwerk zu erstellen und andererseits, ob die derzeit geltenden Standards zur Langzeitarchivierung, interoperablen Repräsentation und Online-Kommentierung von digitalen Textkorpora noch zeitgemäß sind.

Gegenstand

Der umfangreiche literarische Nachlass des österreichischen Schriftstellers Robert Musil umfasst 12.000 Manuskriptseiten und wird bereits seit 1985 digital ediert. Die wichtigsten bisherigen Publikationsetappen markieren die CD-ROM-Ausgabe *Robert Musil: Der literarische Nachlass* (Hg. F. Aspetsberger, K. Eibl, A. Frisé, Rowohlt 1992) und die DVD-Edition Robert Musil: *Klagenfurter Ausgabe. Kommentierte Edition sämtlicher Werke, Briefe und nachgelassener Schriften. Mit Transkriptionen und Faksimiles aller Handschriften* (Hg. W. Fanta, K. Amann, K. Corino, Robert-Musil-Institut/Kärntner Literaturarchiv, AAU Klagenfurt 2009). Um für das Editionsvorhaben, das mittlerweile mehrere Computer-Generationen, System- und Formatwechsel sowie Veränderungen der editorischen Vorgaben und Richtlinien erlebt hat, eine befriedigende und zukunftssichere Lösung zu entwickeln, entsteht am Robert-Musil-Institut/Kärntner Literaturarchiv seit 2016 eine Hybrid-Edition. Sie soll einerseits die Bedürfnisse der LeserInnen und der wissenschaftlichen UserInnen befriedigen, andererseits auch für nachhaltige Verfügbarkeit der Daten sorgen. Die TeilnehmerInnen des vorliegenden Panels präsentieren die einzelnen Lösungsansätze aus theoretischer und methodologischer Sicht in exemplarischer Weise und stellen sie als Best-Practice-Modelle im Bereich des digitalen Edierens zur Diskussion:

a) Die **Hybrid-Edition** setzt sich aus der *Musil-Gesamtausgabe* in 12 Bänden (Salzburg, Jung und Jung, 2016-2022) und dem Internetportal *musilonline* (Prototyp seit 2016: www.musilonline.at) zusammen. Die Buchausgabe enthält einen Lesetext für die literarische Lektüre in leserfreundlicher Ausstattung. *musilonline* wird unter *Musil-Text* die digitale Version des Lesetexts, unter *Archiv* das gesamte textgenetische Dossier (Faksimiles, XML/TEI-Dateien) und unter *Kommentar* neben dem textkritischen bzw. textgenetischen auch einen interdiskursiven Kommentar bieten. Im Kurzvortrag wird die Hybrid-Edition in Hinblick auf ihre medienhistorische Bedeutung, intermediale Funktion, literaturdidaktische Vermittlungsleistung und Leser/User-Orientierung erläutert. (Anke Bosse, Robert-Mu-

sil-Institut/Kärntner Literaturarchiv, AAU Klagenfurt)

b) Das künftige **Interface** für die literaturwissenschaftliche, insbes. textgenetische Forschung auf *musilonline* soll sich aus folgenden Komponenten zusammensetzen:

- *Suchmaschinen* (zur Treffergenerierung im edierten Musil-Textkorpus sowie im XML/TEI-ausgezeichneten Text- und Metadatenbereich)
- *Navigation entlang hypertextueller Verknüpfungen* zwischen ediertem Text, Faksimiles, XML/TEI-Dateien und Kommentarbereich
- *Bild-Browser* (zum Studium der Originalmanuskripte)
- *Textdarstellung* (zur visuellen Inszenierung der Textstufen und -schichten am Manuskript, aus XML/TEI-Dateien generierte HTML-Lösungen). Diese Funktionen werden derzeit auf der Grundlage von zwei zentralen Manuskriptmappen aus Musils Nachlass entwickelt.

Im Kurzvortrag werden noch keine fertigen Lösungen vorgestellt, sondern es erfolgt ein kritischer Aufriss der Problemlage in Folge der komplexen Struktur von Musils Manuskripten und die Präsentation eines Grundkonzepts an Hand von exemplarischen Ausschnitten aus dem Manuskriptbestand. (Walter Fanta, Robert-Musil-Institut/Kärntner Literaturarchiv, AAU Klagenfurt)

c) Die fachgerechte **Speicherung** des gesamten Text- und Metadatenbestands zu Robert Musils autor-autorisierten und nachgelassenen Schriften **via XML/TEI** ist vorgesehen und bereits begonnen worden. Die Manuskripte des Musil-Nachlasses stellen eine besondere Herausforderung für die Textauszeichnung dar, weil sie äußerst komplexe Varianzbeziehungen auf der Ebene der Makrovarianz zwischen den Entwurfsfassungen und der Ebene der Mikrovarianz – Korrekturschichten – enthalten. Es stellt sich heraus, dass die Struktur der großen philosophischen und literarischen Fragmente der Moderne (Nietzsche, Wittgenstein, Musil, Bachmann) den Rahmen sprengt, der von XML/TEI (Baumstruktur) vorgegeben ist. Im Kurzvortrag erfolgt ein Problembericht zu Auszeichnungsschwierigkeiten mit XML/TEI. Im Rahmen des österreichischen Kompetenzzentrums für Digitale Edition (KONDE), sowie mit Hilfe der TEI Guidelines und DariahTeach wurden bereits Lösungen gefunden. In Hinblick auf die Datenkonservierung und –interoperabilität muss aber diskutiert werden, ob XML/TEI für die Langzeitarchivierung des Musil-Nachlasses geeignet ist. (Katharina Godler, Robert-Musil-Institut/Kärntner Literaturarchiv, AAU Klagenfurt)

d) Die **Migration** des gesamten Textdaten-Korpus erfolgt aus dem Flatfile des Formats FolioViews der Klagenfurter Ausgabe in das Zielformat XML/TEI. Dabei werden Scripts entwickelt, welche die im Flatfile enthaltenen Kodierungen (Formatierungen und Sprungverknüpfungen von FolioViews) sowie die diakritischen Zeichen der 1992 publizierten Transkription soweit wie möglich automatisch im automatischen Austausch in sachadäquate XML/TEI-Auszeichnungen umsetzen. Der Kurzvortrag erläutert den erreichten Stand und die Probleme dieser Migrationsprozesse; sie bestehen kurz gesagt im Alter der Nachlass-Transkription (entstanden 1984-1990), in der chaotischen Struktur der FolioViews-Infobase mit insgesamt 735.000 Einträgen und ca. 250.000 Verknüpfungen, zahlreichen Redundanzen, Inkonsistenzen, Fehlern und Ergänzungsbedarf. Für die drei Hauptbereiche gedruckte Quellen, Nachlassmanuskripte, Metadaten müssen jeweils eigene Lösungen gefunden werden. (Gerrit Brüning, Goethe-Universität Frankfurt / Freies Deutsches Hochstift)

e) Der **interdiskursive Online-Kommentar** auf *musilonline* wird 2018-2022 in einem vom Österreichischen Wissenschaftsfonds (FWF) geförderten Projekt am Robert-Musil-Institut / AAU Klagenfurt entwickelt. Die Grundidee besteht darin, über die herkömmliche (textkritische) Erläuterungsfunktion von Kommentaren hinaus zu wirken, auf der Bedeutungsebene anzusetzen und die Bedeutungsvielfalt in Musils Texten dadurch zu bewahren, dass Interpretamente im Musil-Textkorpus identifiziert und mit den Deutungen der bisherigen Interpretationsliteratur verknüpft werden. Aus der Sicht der Digital Humanities stellt sich die Herausforderung, für die Online-Inszenierung der Diskurse um Musils Texte eine digitale Struktur zu finden, die den heterogenen, teils widersprüchlichen Anforderungen und Erwartungen unterschiedlicher Usergruppen Genüge tut. Im Kurzvortrag wird das Vorhaben mit Fokus auf den Online-Kommentar als Desiderat der Literaturvermittlung exemplarisch skizziert. (Artur Boelderl, Institut für Germanistik, AAU Klagenfurt)

Das Panel integriert höchst unterschiedliche Bereiche und Aspekte der Digital Humanities, die sich im Projekt *musilonline* verknüpft finden. Der Stoßrichtung der Tagung, eine *Kritik der digitalen Vernunft* zu formulieren, tragen die geplanten Diskussionsbeiträge in besonderem Maße Rechnung, da nicht in allen Belangen schon mit fertigen Lösungen aufgewartet wird, sondern das Erkennen von Problemen, die Verbesserung etablierter Strukturen, die Suche nach Alternativen und die Erfindung neuer Konzepte im Vordergrund stehen. Die Präsentationen und Diskussionen des Panels reflektieren u.a. auch die neuen editionswissenschaftlichen Forschungsan-

sätze und Best Practices, die im Rahmen des österreichischen Kompetenznetzwerks digitale Edition (KONDE) seit 2017 entwickelt werden.

Bibliographie

Burghart, Marjorie (2017): *Creating a Scholarly Digital Edition with the Text Encoding Initiative.* Demm: https://www.digitalmanuscripts.eu/digital-editing-of-medieval-texts-a-textbook/ (letzer Zugriff 12. Jänner 2018)

Burnard, Lou (2014): *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Encyclopédie Numérique. Marseille: OpenEdition Press: http://books.openedition.org/oep/426 (Letzter Zugriff 12. Jänner 2018)

Fanta, Walter (2016): *"Editionsgeschichte."*,in: **Nübel, Birgit**/ Wolf, Norbert Christian: *Robert-Musil-Handbuch*. Berlin: Walter de Gruyter, S. 799-810.

Fanta, Walter (2016): *"Nachlass."*,in: **Nübel, Birgit/ Wolf, Norbert Christian (eds):** *Robert-Musil-Handbuch*. Berlin: Walter de Gruyter, S. 470-497.

Fanta, Walter (2010): *"Robert Musil – Klagenfurter Ausgabe."*,in: *editio*, S. 117-148.

Fanta, Walter (2011): *"Zur Immortalität elektronischer Korpora am Beispiel der Musil-Edition."*,in: **Braungart, Georg / Gendolla, Peter / Jannidis, Fotis** (eds): *Jahrbuch für Computerphilologie online:* http://computerphilologie.tu-darmstadt.de/jg09/fanta.html (letzter Zugriff 12. Jänner 2018)

Fanta, Walter (2008): *"Das Zögern vor dem letzten Schritt. Zur digitalen Edition von Robert Musils „Mann ohne Eigenschaften"*, in: **Golz, Jochen / Koltes, Manfred** (eds.): *Autoren und Redaktoren als Editoren - Tagung der Arbeitsgemeinschaft für germanistische Edition an der Klassik Stiftung Weimar*, Tübingen: Max Niemeyer Verlag, S. 342-352

Pierazzo, Elena (2015): *Digital Scholarly Editing: Theories, Models and Methods.* Farnham, Surrey; Burlington, VT: Ashgate.

Sperberg-McQueen, C.M. / Burnard, Lou (2017): *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative Consortium.

„Storied Collections“? Ein kritischer Blick auf die Arbeit an digitalen (Musik)-Editionen

Stadler, Peter
stadler@weber-gesamtausgabe.de
Universität Paderborn

Kepper, Johannes
kepper@edirom.de
Universität Paderborn

Capelle, Irmlind
irmlind.capelle@uni-paderborn.de
Universität Paderborn

Oberhoff, Andreas
oberhoff@upb.de
Universität Paderborn

Der Aufbau von digitalen (Musik)-Editionen und den entsprechenden Online-Publikationen hat sich im wissenschaftlichen Umfeld etabliert und damit Fakten und Muster geschaffen, die nicht notwendigerweise mit den zeitlich parallel entwickelten Theorien kongruent gehen. Das mag zum einen daran liegen, dass die zeitgenössische Theoriebildung selbst nicht einheitlich ist – man vergleiche nur die Idee einer „sozialen Edition“ (Siemens 2011) mit der Debatte um „documentary editing“ (Robinson 2013; Gabler 2010; Pierazzo 2011) oder mit einem „multiplen Textbegriff“ als Grundlage der Edition (Sahle 2013) – zum anderen aber auch an den ganz praktischen Rahmenbedingungen der allermeist als drittmittelgeförderten, zeitlich begrenzten Projekte.

Es gilt daher, kritisch rückzublicken und zu reflektieren, in welcher Weise die Art der digitalen Erschließung die Erkenntnismöglichkeiten des Nutzers steuert, und ob sich die mit solchen Webpublikationen häufig verbundenen Ideen von „Offenheit“, „Erweiterbarkeit“, „Vernetzung“ und „Nutzerbeteiligung“ in der täglichen Arbeit der Projekte überhaupt realisieren lassen – oder ob diese Ideen nicht sogar teilweise auf zu sehr simplifizierten Voraussetzungen beruhen?

Innerhalb der Musikwissenschaft decken die digitalen Projekte des Detmold/Paderborner Virtuellen Forschungsverbunds Edirom und des Zen-

trums Musik – Edition – Medien sowie die damit assoziierten Vorhaben einen weiten Bereich der digitalen Aktivitäten im Fach ab. Es handelt sich einerseits um Projekte im Bereich der Musikedition (u.a. Weber-Gesamtausgabe, Freischütz Digital , Bargheer-Fiedellieder , Beethovens Werkstatt), andererseits um Methoden- und Softwareentwicklung (Edirom , ZenMEM , VideApp , WeGA-WebApp), sowie um neuartige Erschließungskonzepte (Detmolder Hoftheater) bzw. Beiträge zur Entwicklung von Codierungsstandards (TEI, MEI). Vor allem in Rückbindung an TEI und MEI sind dabei Publikationen entstanden, die weit über eine schlichte Digitalisierung hinausgehen: Metadaten, die wissenschaftlichen Ansprüchen gerecht werden, extensive Verknüpfungen und inhaltliche Auszeichnungen der Dokumente sind selbstverständlich. Ferner sind unterschiedliche Annotationspraktiken erprobt.

Das Bemühen der bisherigen Arbeiten war vor allem darauf gerichtet, die wissenschaftlichen Standards der „klassischen“ Edition und Informationsbereitstellung zu halten und „digitale“ Möglichkeiten wie Verknüpfungen zu externen Quellen und die Vereinheitlichung von Angaben durch Normdaten zu integrieren.

Aber genügt eine solche „Aufbereitung“ der gesammelten Daten und wie kommt dabei der „Nutzer“ ins Spiel bzw. wie kann seinen Erwartungen entsprochen werden? Mit Blick auf solche Fragen stellte Jeffrey T. Schnapp 2013 fest:

„Herein resides the challenge that I am referring to as storied collections and that I associate with the need to give rise to a humanistic culture of critical engagement with data and data architectures themselves as well as with the tools that analyze and translate them into argumentative or narrative forms.“ (Schnapp 2013)

D. h. es ist zu fragen:

- Wie bestimmen die verwendeten Schemata unsere Erschließung?
- Was verstehen wir unter und wie ermöglichen wir Partizipation der Nutzer?
- Wie verhält sich Partizipation zu unserem wissenschaftlichem Anspruch?
- Welche technischen Probleme stehen Partizipation (noch?) entgegen?
- Wie erreichen wir einen kritischen Umgang mit den bereitgestellten Daten?
- Wie erreichen wir durch bloße Informationsbereitstellung kritisches Wissen?
- Kann der Nutzer ohne Vorwissen solche Portale / Editionen effektiv nutzen?
- Inwiefern nehmen wir überhaupt auf verschiedene Erkenntnisinteressen Rücksicht?

Wenn dies auch Fragen sind, die z. T. alle Geisteswissenschaften betreffen, so sollen sie doch in dem vorgeschlagenen Panel aus Sicht der speziellen Anforderung der Musikwissenschaft betrachtet werden. Hierzu werden drei verschiedene musikwissenschaftliche digitale Projekte (Weber-Gesamtausgabe, Hoftheater-Projekt und Beethovens Werkstatt) ihren bisherigen Umgang mit den Standards und den digitalen Möglichkeiten kritisch erläutern. Ergänzt werden diese Überlegungen durch die kritische Reflexion der technischen Bedingungen von Partizipation und Konsistenz der Daten (Zentrum Musik – Edition – Medien).

Weber-Gesamtausgabe

Die digitale Edition der Schriften, Tagebücher und Schriften Carl Maria von Webers wurde 2011 der Öffentlichkeit vorgestellt und seitdem – sowohl in der TEI-Auszeichnung als auch in der HTML-Darstellung – kontinuierlich weiterentwickelt und angepasst. Erst im letzten Jahr z.B. wurden dabei „Themenkommentare“ ergänzt als Versuch, die inzwischen über 27.000 verfügbaren Dokumente stärker narrativ einzubetten bzw. zu verknüpfen. Grundsätzlich bleibt aber das Dilemma, dass ein starker Fokus auf der Standardisierung und Normalisierung der Auszeichnung liegt – das ermöglicht zwar auf einer globalen Ebene das Vernetzen mit anderen Repositorien z.B. durch GND-Beacon oder correspSearch und demonstriert somit die Möglichkeiten und Anschlussfähigkeiten digitaler Editionen, aus dem Blick geraten dabei aber oft die Besonderheiten (und Unsicherheiten) lokaler Phänomene.

Hoftheater-Projekt

Das sog. Hoftheater-Projekt („Entwicklung eines MEI- und TEI-basierten Modells kontextueller Tiefenerschließung von Musikalienbeständen am Beispiel des Detmolder Hoftheaters im 19. Jahrhundert (1825–1875)“) stellt einerseits in traditioneller Weise Informationen zu sehr heterogenen Beständen bereit und verknüpft diese andererseits untereinander in einer Form, die erst durch digitale Mittel möglich ist. D. h. neben der Präsentation von Digitalisaten, Metadaten, Incipits und Textübertragungen, werden die einzelnen Objekte durch die Auszeichnung nicht nur mit Elementen, sondern mit key-Attributen (und wenn möglich mit Normdaten) für Personen, Werke und Rollen miteinander in Verbindung gesetzt.

Es entsteht so ein Informationsnetz [1], das unterschiedliche Forschungsinteressen zulässt. Neben den „traditionellen“ Informationen zu den Quellen können Angaben zur Organisation des Theaterbetriebs, zur finanziellen Situation einzelner Personen, zur Theatersituation in den verschiedenen Spielorten etc. abgefragt werden. Die Daten können aber auch Basis soziologischer/historischer Studien werden, indem z. B. die Gehälter am Theater mit denen anderer Berufsgruppen in Beziehung gesetzt werden.

Es ergeben sich u. a. folgende Fragen für einen kritischen Umgang mit diesen Daten:

- die bisher verwendeten Auszeichnungselemente sind fachspezifisch gewählt
- Zweifel in der Übertragung werden ausgezeichnet [aber nicht angezeigt]. Bei den Auszeichnungen wird hingegen auf diese Angabe verzichtet bzw. bleiben Lücken, da das grundsätzliche Problem, wie inhaltliche Argumente „dargestellt“ werden können, noch nicht gelöst ist
- die XML-Dateien stehen innerhalb der Anwendung zur Verfügung, können aber nicht frei heruntergeladen werden

Beethovens Werkstatt

Die Zielsetzung der in „Beethovens Werkstatt“ entwickelten VideApp ist es, die Erkenntnisse und Beobachtungen des Projekts zur Textgenese ausgewählter Werke Beethovens möglichst direkt sichtbar zu machen, also eine Vermittlungsform zu finden, die jenseits verbaler Erläuterungen einen möglichst unmittelbaren und nachvollziehbaren Zugang zu den Inhalten bietet. Dabei zeigt sich, dass die Menge der zu treffenden Aussagen, verbunden mit der Neuartigkeit dieser Vermittlungsformen, leicht zu Orientierungsschwierigkeiten des Benutzers führt: Nicht immer erschließen sich gut gemeinte Funktionen so schnell wie erhofft, und besonders spannende Beobachtungen gehen in der Fülle an Details unter. Eine vor Jahren im Kontext des Edirom-Projekts entstandene, aber nie umgesetzte Idee aufgreifend versucht das Projekt daher inzwischen, dem Benutzer besonders relevante Aspekte der Editionen über geführte „Touren“ nahezubringen, ohne dessen eigenständige vertiefende Auseinandersetzung mit den Materialien einzuschränken.

ZenMEM

Im Verbundprojekt „Zentrum Musik – Edition – Medien“ (ZenMEM) beschäftigen sich Wissenschaftler/-innen der Universität Paderborn, der Hochschule für Musik Detmold und der Hochschule Ostwestfalen-Lippe mit den Veränderungen und den neuen Möglichkeiten beim Übergang von analogen zu digitalen Musik- und Medieneditionen.

Unbestritten sind an dieser Stelle die vielen Vorteile und Mehrwerte einer digitalen Edition gegenüber der klassischen, analogen Edition in Buchform. Die Digitalisierung von Musikeditionen schafft aber gleichzeitig auch ganz neue Problemstellungen und Herausforderungen. Im Projekt durchgeführte problemzentrierte Leitfadeninterviews mit Editoren zeigten bspw. deutlich das Spannungsfeld zwischen Offenheit und Abgeschlossenheit insbesondere bei der (Nach-)Nutzung:

- Wie erreicht man eine Form der Abgeschlossenheit bei der Publikation digitaler Musikeditionen, welche die ‚Wertigkeit‘ einer gedruckten Edition besitzt?
- Wie bringt man die gewünschte offene (Nach-)Nutzung einer digitalen Musikedition durch ein breites Publikum (neben Editoren auch Dirigenten, interessierte Laien, Studierende) in Einklang mit dem Wunsch nach Abgeschlossenheit?
- Wie gestaltet sich die Wertschöpfung im Bereich (offener) digitaler Musikeditionen und wie ist das Verhältnis von Editoren, Verlagen, Forschungs- und Gedächtnisinstitutionen zueinander?
- Wie stellt man eine dauerhafte Verfügbarkeit und Referenzierbarkeit (insbesondere im Hinblick auf eine offene, sich weiterentwickelnde) digitale Musikedition sicher?
- Wie geht man mit Autorschaft in einer gemeinschaftlich erarbeiteten Edition um? Bereits auf Annotationsebene?
- Wie geht man mit den Rechten an (externem) Quellenmaterial um?
- Wie kann man Nachhaltigkeit gewährleisten?

Die genannten Problemfelder ergeben sich zum Teil zwar schon direkt oder indirekt aus dem Übergang von analogen zu digitalen Musikeditionen und haben bereits Auswirkungen auf den Prozess des Edierens selbst, doch eine breitgefächerte (Nach)Nutzung muss frühzeitig mitbetrachtet werden, da zusätzliche Akteure mit unterschiedlichen Interessen in Einklang gebracht werden müssen.

Zusätzlich implizieren viele der Probleme auch sehr technische Herausforderungen, wie bspw. eine Revisionssicherheit (insbesondere: Berechtigungen, Schutz vor Veränderung und Verfälschung, Sicherung vor Verlust, Dokumentation

von Prozessen und Nachvollziehbarkeit) sowie Versionierung und Referenzierung von Arbeits- und Publikationsständen. Hier gilt es nun den nächsten Paradigmenwechsel von einzelnen digitalen Musikeditionen hin zu Editionsinfrastrukturen forschungsbegleitend zu vollziehen, um die genannten Herausforderungen überhaupt adäquat adressieren zu können.

Fußnoten

1. Siehe das Modell auf der Website: http://hof-theater-detmold.de/?page_id=1095 .

Bibliographie

Gabler, Hans Walter (2010): „Theorizing the Digital Scholarly Edition“, in: Literature Compass 7/2 (2010): 43–56

Pierazzo, Elena (2011): „A Rationale of Digital Documentary editions“. Literary and Linguistic Computing 26/4 (2011): 463–77

Robinson, Peter (2013): „Towards a Theory of Digital Editions“, in: Variants 10 (2013): 105–131

Sahle, Patrick (2013): „Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels“, Norderstedt 2013

Schnapp, Jeffrey T. (2003): „Knowledge Design. Incubating new knowledge forms / genres / spaces in the laboratory of the digital humanities.“ Keynote delivered at the Herrenhausen Conference „Digital Humanities Revisited – Challenges and Opportunities in the Digital Age“ (Dez. 2013)

Siemens, Ray / Timney, Meagan / Leitch, Cara / Koolen, Corina / Garnett, Alex (2011): „Toward Modelling the *Social* Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media“, in: Literary and linguistic computing 27/4 (2012): 445–461

Wiering, Frans / Crawford, Tim / Lewis, David(2006): „Digital Critical Editions of Music. A Multidimensional Model“, Methods Network Expert Seminar „Modern Methods for Musicology“, online unter http://www.methodsnetwork.ac.uk/redist/pdf/wiering.pdf

Vorträge

Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane

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

Betz, Katrin
katrin.betz@uni-wuerzburg.de
Universität Würzburg, Deutschland

Schlör, Daniel
schloer@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Die Definition von Gattungen sowohl im Sinne allgemeiner Gattungskonzepte als auch konkreter einzelner Gattungen ist ein altes und nach wie vor zentrales Problem der Literaturwissenschaft und wird immer noch debattiert (Kayser 1956: 330-387, Zymner 2003). Allgemein können Gattungsbegriffe als Sammelbegriffe verstanden werden, deren Aufgabe es ist, zu beschreiben, in welcher Hinsicht Texte zu Textgruppen zusammengefasst werden können. Oft ist dabei auf das Konzept von Gattungen als logischen Klassen zurückgegriffen worden, auch in vielen Untersuchungen zu literarischen Gattungen im Bereich der Digital Humanities, obwohl in der literaturwissenschaftlichen Gattungstheorie bereits seit den 60er-Jahren andere Vorschläge für das Verständnis von Gattungskategorien gemacht worden sind.

Im Sinne der “Kritik der digitalen Vernunft” ist das Ziel dieses Beitrags, die Problematik der Gattungsbeschreibung auf der theoretischen Basis alternativer Gattungstheorien und mit Hilfe informatischer Mittel aus einer neuen Perspektive zu betrachten. Dazu wird exemplarisch die Anwendbarkeit des Prototypenmodells als Gattungskonzept für digitale gattungsstilistische Studien überprüft. Als Testkorpus dient eine Sammlung von Texten aus der hispanoamerikanischen Romanliteratur des 19. Jahrhunderts.

Forschungsstand und Ziele

Viele Gattungstheorien beruhen auf der Grundannahme, dass sich konkrete Texte anhand von hinreichenden und notwendigen Attributen eindeutig disjunkten Klassen zuordnen lassen und oft auch, dass sich für Gattungen eine Taxonomie entwickeln lässt (vgl. Zymner 2003: 102-104). Allerdings weisen nicht alle Texte einer bestimmten Gattung gemeinsame Merkmale auf, noch sind die Beziehungen zwischen den einzelnen Gattungen und Texten statisch. Zudem gibt es die Vorstellung von Werken, die initiale, prägende Wirkung haben (z. B. *Waverley* von Walter Scott für den historischen Roman) oder besonders gute Vertreter einer Gattung sind (z. B. Goethes *Wilhelm Meisters Lehrjahre*). Es kann dann andere Vertreter geben, bei denen die für die Gattung als prototypisch angesehenen Merkmale nicht so stark ausgeprägt sind, die aber trotzdem der entsprechenden Textgruppe zuzuordnen sind. Kritik am Klassenkonzept äußert schon Vivas (1968), später u. a. Hempfer (2010). Neuere Gattungstheorien haben daher u. a. Wittgensteins Konzept der Familienähnlichkeit (Weitz 1956, Fowler 1982, Fishelov 1991) und die in der Kognitionspsychologie entwickelte Prototypentheorie (Rosch 1973) als alternative Modelle der Kategorisierung aufgegriffen.

In den Digital Humanities gibt es bereits einige Untersuchungen zu literarischen Gattungen, bei denen jedoch üblicherweise die Annahme zugrunde liegt, dass Gattungen als Klassen im logischen Sinn zu verstehen sind (Calvo Tello et al. 2017, Hettinger el al. 2016a, Hettinger et al. 2016b, Schöch et al. 2016, Schöch 2015, Schöch 2013).[1]

Im vorliegenden Beitrag wird exemplarisch dargestellt, inwiefern es möglich ist, das Prototypenmodell zu verwenden, um Einsichten in die Anordnung maschinell gruppierter Texte zu gewinnen, die über den klassischen Ansatz der Klassifikation nicht möglich wären. Der Beitrag leistet damit zum einen, einen konkreten Vorschlag für die Formalisierung des Prototypen-Ansatzes für gattungsstilistische Untersuchungen zu machen. Zum anderen zielt er darauf, durch die Berücksichtigung der internen Strukturierung von Gattungskategorien eine bessere Anbindung der computergestützten Verfahren an literatur-

geschichtliche Forschungsergebnisse zu ermöglichen.

Fallbeispiel: Untergattungen des hispanoamerikanischen Romans im 19. Jahrhundert

Korpus. Die Analyse wurde an 80 hispanoamerikanischen Romanen von 51 AutorInnen getestet, welche fünf verschiedenen Untergattungen zugeordnet sind: dem historischen Roman, dem Liebesroman, dem kostumbristischen Roman, dem Gaucho-Roman und dem Antisklaverei-Roman. Der historische Roman, der Liebesroman und der kostumbristische Roman sind im Hispanoamerika des 19. Jahrhunderts verbreitete Untergattungen (Janik 2008: 60-77) mit europäischen Vorläufern und Vorbildern. Der Gaucho-Roman und der Antisklaverei-Roman sind Romantypen, die in Hispanoamerika entstanden sind und nur in bestimmten Regionen vorkommen. Für alle fünf Untergattungen wird angenommen, dass sie vor allem auf einer inhaltlichen Ebene definiert sind (Álamo Felices 2011, Lichtblau 1959: 121-135, Rivas 1990).

Für die Zuordnung der Romane zu den verschiedenen Untergattungen wurde einschlägige Sekundärliteratur ausgewertet. Abb. 1 und 2 zeigen die Verteilung der Romane über die Zeit nach Untergattungen sowie nach Ländern. Bestandteil des Korpus sind neben hispanoamerikanischen Romanen auch fünf Texte aus Spanien sowie je ein Text aus England und Frankreich (in spanischer Übersetzung), welche aufgrund ihres Prototypen-Status einbezogen wurden. Das Korpus umfasst insgesamt 2,3 Mio. Token.[2]

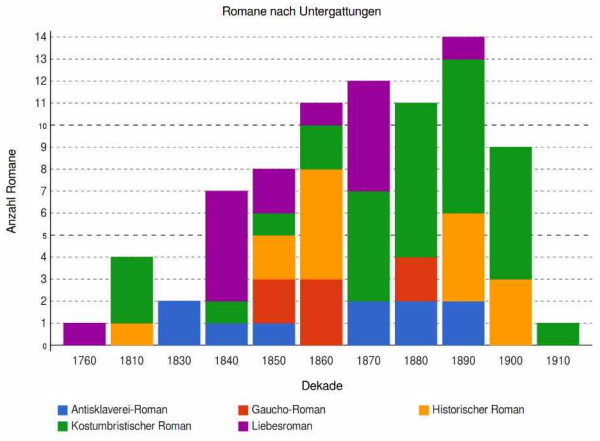


Abbildung 1: Verteilung der Romane über die Jahrzehnte und nach Untergattungen

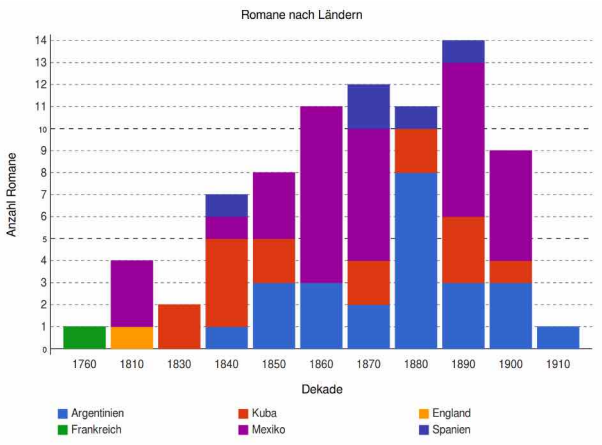


Abbildung 2: Verteilung der Romane über die Jahrzehnte und nach Ländern

Definition von Prototypen. Für jede der fünf Untergattungen wurde mindestens ein zur Untergattung gehörender Roman als Prototyp festgelegt.[3] Es wird zunächst also nicht die Idee vom Prototypen als dem durchschnittlichen Vertreter einer Kategorie verfolgt, sondern der Ansatz, dass ein oder mehrere bestimmte Exemplare einen besonderen, prototypischen Status haben. Ob ein Text eine Vorbild- oder Höhepunkt-Funktion innerhalb einer bestimmten Gattung hat, ist letztlich eine Frage der Perspektive und der Korpuszusammenstellung. Folgende Prototypen sind für die fünf Untergattungen gesetzt worden, wobei es für den kostumbristischen Roman eine ganze Reihe in Frage kommender Prototypen gibt:

Tabelle 1: Untergattungen und die für sie gesetzten Prototypen

Untergattung	Prototyp(en)	Art
Historischer Roman	• Waverley o hace sesenta años (Walter Scott, 1814)[4]	Vorbild
Liebesroman	• Julia o la nueva Eloísa (Jean-Jacques Rousseau, 1761)[5]	Vorbild
Kostumbristischer Roman	• El Periquillo Sarniento (José Joaquín Fernández de Lizardi, 1816) • Don Catrín de la Fachenda (Lizardi, 1818) • Noches tristes y día alegre (Lizardi, 1818) • La gaviota (Fernán Caballero, 1849) • El sombrero de tres picos (Pedro Antonio de Alarcón, 1874) • Pepita Jiménez (Juan Valera, 1874) • Sotileza (José María de Pereda, 1884) • Peñas arriba (Pereda, 1895)	Vorbilder
Gaucho-Roman	• Juan Moreira (Eduardo Gutiérrez, 1880)	Höhepunkt
Antisklaverei-Roman	• Sab (Gertrudis Gómez de Avellaneda, 1841)	Höhepunkt

Beim historischen Roman und beim Liebesroman handelt es sich bei den Prototyp-Texten um Übersetzungen der ursprünglich auf englisch und französisch verfassten Romane, die auch in Hispanoamerika als Vorbilder für diese Romantypen eingeschätzt werden. Kontroverser wird diskutiert, welche Texte Einfluss auf die kostumbristischen Romane ausgeübt haben. Auf der einen Seite wird der mexikanische Autor Lizardi als Pionier genannt, auf der anderen Seite werden Romane spanischer Autoren (Caballero, Alarcón, Valera, Pereda) angeführt (Calderón 2005). Eine Prototypenanalyse könnte hier Argumente für eine der beiden Thesen liefern. Im Falle des Gaucho-Romans und des Antisklaverei-Romans sind die gesetzten Prototypen als Höhepunkte der jeweiligen Gattung zu verstehen, da die weiteren diesen Untergattungen zugeordneten Romane im Korpus entweder als Vorstufen oder Nachfolger der besonders repräsentativen Gattungsvertreter beschrieben worden sind (Lichtblau 1959: 121-135, Rivas 1990).

Methoden. Um die für die Untergattungen relevanten Strukturen auf unterschiedlichen textlichen Ebenen zu analysieren und zu modellieren, wurden Topic Modelling und eine Analyse der häufigsten Wörter (MFW) angewandt.

Das Topic Modeling (vgl. Blei 2012) wurde mit MALLET[6] mit für das Korpus geeigneten Vorverarbeitungsschritten und Parametern durchgeführt (Lemmatisierung, Beschränkung auf Substantive, Segmentlänge von 1000 Wörtern, 30 Topics, 5000 Iterationen, Optimierung alle 10 Iterationen) und die Ergebnisse anschließend für die einzelnen Romane wieder aggregiert. Abb. 3 zeigt ein Beispiel-Topic aus dem entstandenen Modell:

Abbildung 3: Beispiel-Topic: carta-alma-corazón (Brief-Seele-Herz)

Im Ergebnis erhält man für jedes Topic in jedem Roman einen bestimmten Wahrscheinlichkeitswert; ein Roman ist durch die Reihe seiner einzelnen Topic-Wahrscheinlichkeiten repräsentiert. Anschließend sind die Ähnlichkeiten zwi-

schen den Topic-Verteilungen der Romane mit der Kosinus-Ähnlichkeit berechnet worden.[7] Darauf aufbauend konnten die Ähnlichkeiten aller Romane untereinander ermittelt werden, so auch zwischen einzelnen Vertretern einer Roman-Untergattung und den jeweiligen Prototypen.

Die zweite Dokument-Repräsentation wurde auf Basis der 10.000 häufigsten Wörter (MFW) in den Roman-Volltexten erstellt (nicht lemmatisiert, gewichtet mit TF-IDF, maximale Dokument-Frequenz von 90 %). Anhand dieser beiden Repräsentationen kann überprüft werden, welche Merkmale zentral für die Abbildung von Gattungsunterschieden zwischen den Texten sind.

Ergebnisse und Diskussion. Die auf der Grundlage der Topics und MFW ermittelten Ähnlichkeiten zwischen den Texten wurden hinsichtlich der Beziehungen der Romane zu ihren jeweiligen Prototypen ausgewertet. Abb. 4 zeigt beispielsweise die Abstände der einzelnen Gaucho-Romane zum Prototypen "Juan Moreira" von Eduardo Gutiérrez, auf der linken Seite die Topic-Ähnlichkeiten und auf der rechten Seite die MFW-Ähnlichkeiten. Insgesamt fällt auf, dass die Vorläufer und Nachfolger dem Prototypen hinsichtlich der Topics ähnlicher sind als in Bezug auf die MFW, der Gaucho-Roman also mehr durch gemeinsame Themen als (im weitesten Sinne) stilistische Eigenschaften zusammengehalten wird. In beiden Fällen sind die Abstände zum Prototypen jedoch relativ groß. Wir führen das darauf zurück, dass die Beschreibung der Gattung durch Lichtblau vor allem eine seiner Entwicklung ist: in den Vorläufern ("La familia de Sconner", "Amalia") kommt der Gaucho als Typ nur beiläufig bzw. nur in kurzen Passagen überhaupt vor. Bei den Topic-Ähnlichkeiten ist der späteste Gaucho-Roman im Korpus, "Arturo Sierra" von Julio Llanos, dem Prototypen am nächsten. Laut Lichtblau ist dies ein sentimentaler Sensations-Roman, in dem der Gaucho nur noch eine oberflächliche Rolle spielt (Lichtblau 1959: 134f). Dies zeigt, dass Oberflächenphänomene mit unserem Verfahren gut erfasst werden, die Texteigenschaften aber nicht in allen Fällen hinreichend sind, um die Gattungszugehörigkeit zu klären.

Abbildung 4: Ähnlichkeiten der dem Gaucho-Roman zuzuordnenden Texte zum Prototypen ("Juan Moreira", 1880); links Topics, rechts MFW

Der Boxplot in Abb. 5 zeigt über mehrere Untergattungen hinweg (hier: Liebesroman, Historischer Roman, Antisklaverei-Roman und Gaucho-Roman), wie ähnlich die Texte ihrem Prototypen im Durchschnitt sind und wie weit die Ähnlichkeiten der Romane zum Prototypen streuen (PT). Zusätzlich werden die durchschnittliche Ähnlichkeit und deren Varianz mit der durchschnittlichen Topic-Verteilung innerhalb der Untergattung verglichen und dargestellt (AVG). Damit soll einerseits die Ähnlichkeit der Romane zum Prototypen im Sinne eines besonderen, repräsentativen Romans derselben Untergattung und andererseits ihre Ähnlichkeit zu einem durchschnittlichen (hier fiktiven, berechneten) Text gezeigt werden. Die Liebesromane sind im Vergleich der verschiedenen Untergattungen ihrem Prototypen am unähnlichsten, möglicherweise, weil der Prototyp eine Übersetzung eines französischen Briefromans aus dem 18. Jahrhundert ist. Die hispanoamerikanischen Liebesromane des 19. Jahrhunderts sind in ihrer Topic-Struktur untereinander dagegen sehr ähnlich. Auch die historischen Romane und die Gaucho-Romane sind sich untereinander ähnlicher als ihren gesetzten Prototypen. Die in der Literaturgeschichte traditionell angenommenen Prototypen von literarischen Gattungen sind also in vielen Fällen eher Ausnahme-Romane denn typische Vertreter ihrer Art.

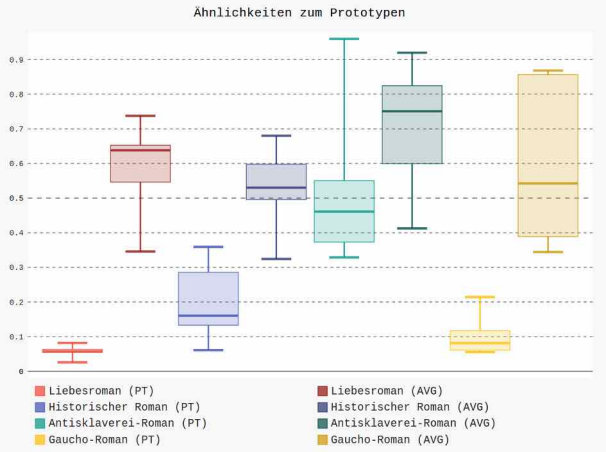


Abbildung 5: Topic-Ähnlichkeiten der Romane verschiedener Untergattungen zu ihren Prototypen (PT) sowie zur durchschnittlichen Topic-Verteilung innerhalb der Untergattung (AVG)

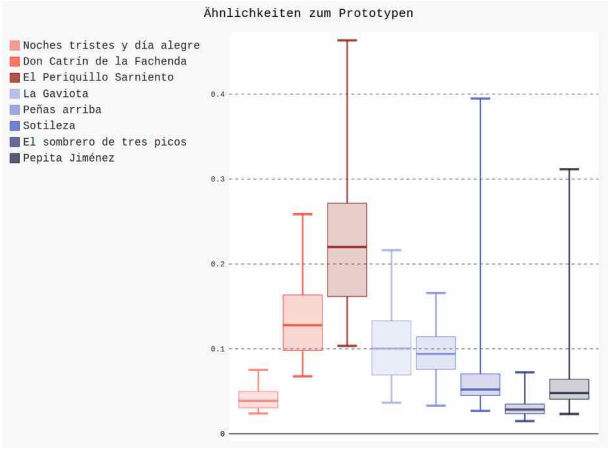


Abbildung 6: MFW-Ähnlichkeiten der kostumbristischen Romane zu möglichen Prototypen, mexikanische (rot) vs. spanische Romane (blau)

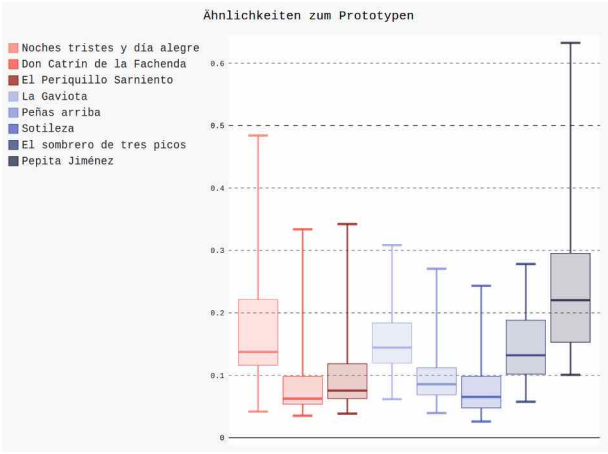


Abbildung 7: Topic-Ähnlichkeiten der kostumbristischen Romane zu möglichen Prototypen, mexikanische (rot) vs. spanische Romane (blau)

Um der umstrittenen Frage nachzugehen, welcher Prototyp aus der Liste der benannten Romane für die hispanoamerikanischen kostumbristischen Romane anzusetzen ist, sind in Abb. 6 die MFW-Distanzen zu den verschiedenen als Prototypen in Frage kommenden Romanen visualisiert. Die Ähnlichkeit ist zu dem Roman “El Periquillo Sarniento” des Mexikaners Lizardi am größten, gefolgt von dem ebenfalls von ihm verfassten Werk “Don Catrin de la Fachenda”. Das dritte Werk von Lizardi “Noches tristes y día alegre” ist den kostumbristischen Romanen zwar durchschnittlich unähnlicher, allerdings handelt es sich dabei um einen Roman in Dialogform. Die MFW als Merkmale stützen also eher die These der mexikanischen Vorbilder. Zieht man die Topics als Merkmale heran, rücken die spanischen Romane dem Prototypen etwas näher (vgl. Abb. 7). Die verschiedenen Positionen können auf unterschiedliche Texteigenschaften zurückgeführt werden (eher stilistisch vs. thematisch), so dass anzunehmen ist, dass die Vertreter der beiden Thesen jeweils unterschiedliche Schwerpunkte bei den Gattungsdefinitionen angesetzt haben. Es kann also an dieser Stelle die Kontroverse um die “wahren Prototypen” bestätigt und datengetrieben differenzierter betrachtet werden.

Für das Gesamtkorpus zeigen Abb. 8 (Topics) und 9 (MFW) die Ähnlichkeitsverhältnisse aller 80 Romane, wobei hohe Ähnlichkeit in rot dargestellt ist. Bei den Topics wird sichtbar, dass die Liebesromane (rechts oben) untereinander sehr ähnlich sind. Auch die Gaucho- (Mitte rechts) und Antisklaverei-Romane (unten links) sind als Blöcke in der Heatmap gut zu erkennen. Hinsichtlich ihrer Topic-Verteilungen weniger homogen sind die historischen Romane und die kostumbristischen Romane. Ähnlichkeiten in den Topic-Verteilungen sind außerdem nicht nur innerhalb einzelner Untergattungen vorhanden, sondern auch zwischen den Liebesromanen und anderen Untergattungen, insbesondere den Gaucho-Romanen und den Antisklaverei-Romanen. Die in den Liebesromanen zentralen Topics sind also auch in anderen Gattungen relevant, was auf Mischtypen hindeu-

tet und beispielsweise bei einer Klassifikation auf der Basis von Topics berücksichtigt werden sollte.

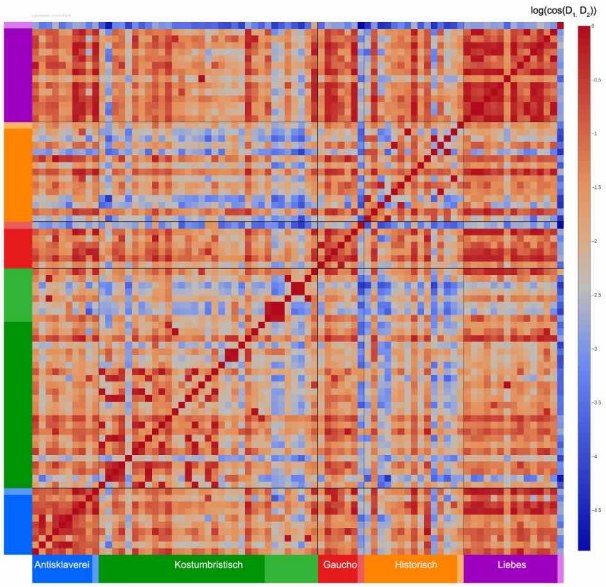


Abbildung 8: Heatmap mit Topic-Ähnlichkeiten aller Romane untereinander, Prototypen sind heller gefärbt

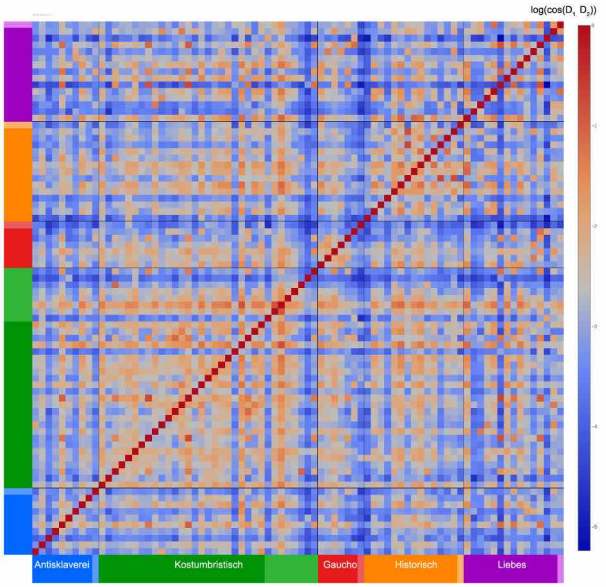


Abbildung 9: Heatmap mit Ähnlichkeiten der 10.000 MFW aller Romane untereinander

Die Heatmap auf der Grundlage der 10.000 MFW zeigt eine größere Homogenität für die kostumbristischen und historischen Romane, während die Antisklaverei-, Gaucho- und Liebesromane hier weniger Nähe zueinander aufweisen. Auch die historischen und kostumbristischen Romane untereinander weisen größere Ähnlichkeiten hinsichtlich ihrer MFW auf. Möglicherweise sind bei beiden Untergattungen spezifisches Vokabular oder stilistische Aspekte für die Gattungszugehörigkeit wichtiger. Besonders für die kostumbristischen Romane, in denen bei der Beschreibung regionaler und lokaler Gegebenheiten auch die Nachahmung von Dialekten und Umgangssprache eine Rolle spielt, ist diese Erklärung schlüssig. In einem auf Substantive beschränkten und lemmatisierten Korpus, wie es in das Topic-Modeling eingegangen ist, können solche Besonderheiten nicht aufgefangen werden.

Fazit und Ausblick

Die exemplarische Anwendung des Prototypenmodells auf die hispanoamerikanischen Romane verschiedener Untergattungen hat gezeigt, dass Ansätze zur Modellierung literarischer Gattungen, die über das Prinzip logischer Klassen hinausgehen, informatisch umgesetzt werden können. Fragen wie diejenige nach der Prototypizität einzelner Texte, Gattungsmischungen und nach differenzierten Nähe- und Distanzverhältnissen lassen sich erst auf dieser Grundlage angehen. Literaturgeschichtliche Aussagen zu Gattungszugehörigkeiten und Gattungsentwicklungen mit prototypensemantischem Bezug (wie hier die Entwicklung des Gaucho-Romans oder die Frage nach den die hispanoamerikanische Tradition prägenden kostumbristischen Romanen) können so hinsichtlich relevanter Textmerkmale genauer untersucht werden.

Künftig sollen weitere theoretische Gattungsmodelle, insbesondere das Prinzip der Familienähnlichkeit, auf ihre Anwendbarkeit für gattungsstilitische Untersuchungen hin getestet werden. Außerdem soll geprüft werden, welche Ergebnisse andere Textrepräsentationen als Topic-Modelle und MFW liefern, z.B. Word Embeddings. Zu diskutieren bleibt, wie offene Kategorisierungsmodelle evaluiert werden können.

Fußnoten

1. Kürzlich stellte jedoch Underwood bei einem Vortrag im Rahmen des DARIAH-Expertenworkshops “Distant Reading in Literary Texts” einen Ansatz für den Umgang mit wechselnden Perspektiven auf das Konzept von Gattung über die Zeit vor. Van Dalen-Oskam betonte die Bedeutung von LeserInnen-Meinungen bei Gattungszuordnungen, im Zusammenspiel mit formalen Merkmalen, vgl. Hagen et al. 2017.
2. Die Metadaten zum verwendeten Korpus und weitere Analysedaten sind unter https://githu-

b.com/cligs/projects2018/tree/master/prototypen-dhd [letzter Zugriff 13. Januar 2018] verfügbar. Die Auswahl der Texte geht auf ein größeres Korpus hispanoamerikanischer Romane und eine digitale Bibliographie zurück, welche im Projekt CLiGS erstellt wurden: https://github.com/cligs/textbox [letzter Zugriff 13. Januar 2018] und http://bibacme.cligs.digital-humanities.de [letzter Zugriff 13. Januar 2018]. Es wurden sämtliche verfügbaren Romane einbezogen, die den untersuchten Untergattungen zugeordnet werden konnten.
3. Wie die Gattungszuordnung ist auch die Entscheidung für die Prototypen auf der Grundlage der Auswertung von Sekundärliteratur getroffen worden (welche Texte werden häufig als prototypisch genannt?) und hat den Status einer Arbeitshypothese.
4. " *Waverley, or, 'Tis Sixty Years Since*". Übersetzung aus dem Englischen von Francisco Gutiérrez-Brito und Isidoro López Lapuya, o.J.
5. " *Julie ou la Nouvelle Héloïse*". Übersetzung aus dem Französischen von José Mor de Fuentes aus dem Jahr 1836.
6. http://mallet.cs.umass.edu/topics.php [letzter Zugriff 13. Januar 2018].
7. Im Ergebnis liegen die Werte zwischen 0 (keine Ähnlichkeit) und 1 (maximale Ähnlichkeit).

Bibliographie

Álamo Felices, Francisco (2011): *Los subgéneros novelescos. Teoría y modalidades narrativas.* Almería: Universidad Almería.

Blei, David M. (2012): "Probabilistic Topic Models", in: *Communications of the ACM* 55 (4): 77-84. DOI: 10.1145/2133806.2133826

Calderón, Mario (2005): "La novela costumbrista mexicana", in: Clark de Lara, Belem / Speckman Guerra, Elisa (eds.): *La república de las letras. Vol. 1: Ambientes, asociaciones y grupos. Movimientos, temas y géneros literarios.* México: UNAM 315-324.

Calvo Tello, José / Schlör, Daniel / Henny, Ulrike / Schöch, Christof (2017): "Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels" in: *DH2017: Annual Conference of the Alliance of Digital Humanities Organizations.* Montreal: McGill University & Université de Montréal 181-184 https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf [letzter Zugriff 13. Januar 2018].

Fishelov, David (1991): "Genre theory and family resemblance", in: *Poetics* 20: 123-138.

Fowler, Alastair (1982): *Kinds of Literature. An Introduction to the Theory of Genres and Modes.* Oxford: Clarendon Press.

Hagen, Thora / Huber, Michael / Tepavac, Daniel (2017): "Workshop on Distant Reading in Literary Texts", in: *DHdBlog* http://dhd-blog.org/?p=8905 [letzter Zugriff 13. Januar 2018].

Hempfer, Klaus W. (2010): "Zum begrifflichen Status der Gattungsbegriffe: Von 'Klassen' zu 'Familienähnlichkeiten' und 'Prototypen'", in: *Zeitschrift für französische Sprache und Literatur* 120 (1): 14-32.

Hettinger, Lena / Jannidis, Fotis / Reger, Isabella / Hotho, Andreas (2016a): "Classification of Literary Subgenres", in: *DHD 2016.* Leipzig: Universität Leipzig 154-58 http://dhd2016.de/boa.pdf [letzter Zugriff 13. Januar 2018].

Hettinger, Lena / Jannidis, Fotis / Reger, Isabella / Hotho, Andreas (2016b): "Significance Testing for the Classification of Literary Subgenres", in: *DH2016: Annual Conference of the Alliance of Digital Humanities Organizations.* Kraków: Jagiellonian University & Pedagogical University 218-220 http://dh2016.adho.org/abstracts/173 [letzter Zugriff 13. Januar 2018].

Janik, Dieter (2008): *Hispanoamerikanische Literaturen. Von der Unabhängigkeit bis zu den Avantgarden (1810-1930).* Tübingen: Narr.

Kayser, Wolfgang (1956 [4]): *Das sprachliche Kunstwerk. Eine Einführung in die Literaturwissenschaft.* Bern: Francke Verlag. Erste Auflage 1948.

Lichtblau, Myron I. (1959): *The Argentine Novel in the Nineteenth Century.* New York: Hispanic Institute in the United States.

Rivas, Mercedes (1990): *Literatura y esclavitud en la novela cubana del siglo XIX.* Sevilla: Escuela de Estudios Hispano-Americanos.

Rosch, Eleanor (1973): "On the internal structure of perceptual and semantic categories", in: E. T. Moore (ed.): *Cognitive development and the acquisition of language.* New York: Academic Press: 111-144.

Schöch, Christof, / Henny, Ulrike / Calvo Tello, José / Schlör, Daniel / Popp, Stefanie (2016): "Topic, Genre, Text. Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880-1930)", in: *DHD 2016.* Leipzig: Universität Leipzig, 235-239 http://dhd2016.de/boa.pdf [letzter Zugriff 13. Januar 2018].

Schöch, Christof (2015): "Topic Modeling French Crime Fiction", in: *DH2015: Annual Conference of the Alliance of Digital Humanities Organizations.* Sydney: The University of Western Sydney http://dh2015.org/abstracts/xml/SCHOCH_Christof_Topic_Modeling_French_Crime_Ficti/SCH_CH_Christof_Topic_Modelin-

g_French_Crime_Fiction.html [letzter Zugriff 13. Januar 2018].

Schöch, Christof (2013): "Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater", in: *DH2013: Annual Conference of the Alliance of Digital Humanities Organizations.* Lincoln: UNL http://dh2013.unl.edu/abstracts/ab-270.html [letzter Zugriff 13. Januar 2018].

Vivas, Eliseo (1968): "Literary Classes: Some Problems", in: *Genre* 1: 97-105.

Weitz, Morris (1956): "The Role of Theory in Aesthetics", in: *The Journal of Aesthetics and Art Criticism* 15 (1): 27-35.

Zymner, Rüdiger (2003): *Gattungstheorie. Probleme und Positionen der Literaturwissenschaft.* Paderborn: mentis Verlag.

Ambiguität und Annotation: Herausforderungen von Automatisierung und Digitalität

Zirker, Angelika

angelika.zirker@uni-tuebingen.de
Eberhard Karls Universität Tübingen,
Deutschland; Humboldt Universität zu Berlin

Der Vortrag beruht auf Überlegungen zur Theorie und Praxis der erklärenden Annotation literarischer Texte. Im Vordergrund steht also weniger die digitale Aufbereitung (im Sinne von Markup) als die Anreicherung von Texten durch Annotationen, wie sie in TEASys (Tübingen Explanatory Annotations System; s. Bauer / Zirker 2015 und 2017) entwickelt wurden. TEASys bietet erklärende Annotationen auf drei verschiedenen Komplexitätsebenen und ist strukturiert nach Kategorien, darunter wie sprachliche Erklärungen, Kontextinformationen, Intertextualität, intratextuelle Verweise, formale Aspekte, textphilologische Anmerkungen und Interpretationen. Das Aufkommen digitaler Annotationen eröffnet neue Möglichkeiten für die Gestaltung der erklärenden Annotationen von Texten, die es noch zu entdecken und aufzubereiten gilt. Ein wesentlicher Faktor liegt dabei im Informationsmanagement, insbesondere im Unterschied zu Annotationen im gedruckten Buch: das digitale Medium erlaubt eine schier unbegrenzte Menge an Informationen sowohl hinsichtlich der dargebotenen Inhalte wie auch durch Hyperlinks. TEASys wurde für die Annotation literarischer Texte entwickelt, soll langfristig aber auch für die Informationsanreicherung und Erläuterung nicht-literarischer Texte und anderer Disziplinen herangezogen werden (vgl. Bauer / Zirker 2015). Es wurde bei der DHd 2016 und 2017 bereits vorgestellt. Seither hat sich TEASys vor allem technisch weiterentwickelt: Die online bereitgestellten Annotationen werden in einer Datenbank gespeichert, die bei der Neuanlage von Annotationen in einem Text auch Vorschläge automatisiert anbietet. Der Vortrag für die DHd 2018 ergibt sich aus dieser technischen Weiterentwicklung des Projekts hinsichtlich eines theoretischen Problems, nämlich der Frage, wie in erklärenden Annotationen mit Ambiguität, d.h. sprachlicher aber auch textueller Mehrdeutigkeit, umzugehen ist, und widmet sich vor allem den Herausforderungen, die sich aus der Automatisierung von Annotationen im Zusammenhang mit Ambiguität ergeben. Ambiguität wird hier verstanden als Doppel- oder Mehrdeutigkeit, d.h. als distinkte Bedeutungen sprachlicher Einheiten (vgl. GRK 1808 2017).

Obwohl es sich bei Praxis der erklärenden Annotation um eine der ältesten Kulturtechniken bei der Aufbereitung von (literarischen) Texten handelt, wurden die theoretischen Probleme und Herausforderungen der erklärenden Annotation bisher nicht systematisch behandelt (vgl. Assmann 1995; Eggert 2009; van Peursen 2010; Drucker 2012; Parry 2012; Zirker / Bauer 2017). Dazu gehört insbesondere die Frage nach dem Verhältnis von Textteilen und Textganzem, Text und Kontext, Erklärung und Interpretation. Ambiguität ist für all diese Aspekte hoch relevant: die Annotation etwa eines Ausdrucks oder eines Textauszugs, z.B. in einem Roman, kann dazu beitragen, die Verbindung zwischen lokaler und globaler Textbedeutung zu zeigen. Ein Fallbeispiel dafür ist etwa die Geistergeschichte „To Be Taken with a Grain of Salt" von Charles Dickens: diese Erzählung ist in die Sammlung *Doctor Marigold's Prescriptions* eingebettet, woraus die Ambiguität des Titels resultiert, d.h. er kann wörtlich wie auch metaphorisch gelesen werden (Zirker 2014). Dies muss von einer Annotation entsprechend erläutert werden. Die Datenbank kann beispielsweise auf weitere (Kon)Texte der Verwendung und damit auf ein Spektrum möglicher Bedeutungen hinweisen, die wiederum auf den Text (zurück)bezogen werden können. [1]

Die Ambiguität – oder Mehrdeutigkeit von Wörtern, Ausdrücken, Sätzen, ganzen Texten – stellt somit eine besondere Herausforderung bei der digitalen erläuternden Annotation dar, vor allem wenn Annotationen automatisiert werden (s.

dazu auch Gius / Jacke 2017). Ein Negativbeispiel für die automatisierte Annotation literarischer Texte, das im Vortrag vorgestellt werden wird, findet sich bei Amazon x-ray (vgl. Bauer / Zirker 2017): dort werden häufig falsche Annotationen angeboten oder es wird bei der Weiterleitung auf die Disambiguierungsseiten von Wikipedia Wissen vorausgesetzt, das dem Textverstehen bereits zugrunde liegt. In TEASys tritt ein anderes Problem hervor: bei der Anlage neuer Annotationen werden dem Annotator aus der Datenbank Vorschläge zu dem Item aus der zugrunde liegenden Datenbank unterbreitet. Im Falle von Ambiguität tritt hier nun die Schwierigkeit auf, dass Textverstehen vorausgesetzt wird, um die ‚richtige' Annotation im jeweiligen Kontext zu wählen. Nimmt man etwa die Phrase „Let me not", die Shakespeares 116. Sonett einleitet, so kann sie sowohl von Sprecher an sich selbst gerichtet sein (analog zu einem Soliloquium) oder aber an einen Adressaten (im Sinne eines Imperativs). Die hier vorliegende Ambiguität hinsichtlich der Kommunikationssituation ist jedoch nicht automatisch auf andere (Kon)Texte übertragbar: „Let me not" wird auch von Hamlet in einem Soliloquium verwendet (Akt 1, Sz. 2) bzw. von Brutus in *Julius Caesar* in einem Dialog mit Cassius (Akt 1, Sz. 2). Eine Automatisierung wie auch Rekontextualisierung der Annotation zu „Let me not" ist deshalb schwierig gerade *aufgrund* ihres Potentials, mehrdeutig zu sein. Dies bedeutet aber auch, dass im Fall der Eindeutigkeit des Ausdrucks die Metadaten der jeweiligen Annotationen auf die Bedingungen für eine solche Disambiguierung verweisen sollten, damit dieser Hinweis beim Erstellen weiterer Annotationen erhältlich und nützlich bleibt.

Doch was passiert, wenn Annotationen Ambiguität berücksichtigen? Hierzu gibt es drei Szenarien: (1) Die erklärende Annotation disambiguiert die Textstelle / den Text. (2) Die erklärende Annotation weist den Nutzer auf die Ambiguität hin, insbesondere in einem literarischen Text (Bode 1988), und bietet distinkte Denotationen an. (3) Die erklärende Annotation führt (in strategischer Weise; s. dazu Bauer / Zirker, in Vorb.) die Wahrnehmung einer Ambiguität ein, die tatsächlich vorliegt, oder eben auch nicht. Im Fall von (1) kann dies in einer Vereindeutigung des Textes resultieren, die besondere Qualitäten literarischer Werke außer Acht lässt; möglicherweise ermöglicht dies aber auch eine klare Interpretation. Die Disambiguierung mag sogar erforderlich sein, etwa im Zuge von Sprachwandel (im Englischen denke man hier z.B. an die heute weniger geläufigen Bedeutungen von „gay" und „nice", die im 18. Jahrhundert völlig andere Denotationen besaßen). Im Fall von (2) kann die Annotation einer Ambiguität die ästhetischen Merkmale eines Textes in besonderer Weise hervortreten lassen, die sonst in den Hintergrund gerückt werden, bspw. die Ambiguität zwischen interner und externer Kommunikationsebene (man denke hier an dramatische Ironie). Im Fall von (3) kann der Annotator eine Erklärung eines potentiell ambigen Items anbieten, die das globale Textverstehen beeinflusst. Dies trifft etwa auf biographische Lesarten der Sonette Shakespeares zu, wo lokale Ambiguitäten strategisch in die Texte hineingelesen werden. Im 145. Sonett ist z.B. von „hate away" die Rede, was in einer Annotation als Verweis auf Shakespeares Frau, Ann Hathaway, interpretiert wird (Booth 1977: 501).

Die theoretischen Probleme, die an diese Überlegungen und kurzen Fallbeispiele anknüpfen, wurden bisher nicht systematisch reflektiert. Dies gilt auch hinsichtlich der Frage, inwieweit erklärende Annotationen der Komplexität und Ambiguität literarischer Texte gerecht werden können. Und während das digitale Medium und der damit (scheinbar) unbegrenzte Raum für Annotationen neue Möglichkeiten für die Lösung dieser Fragen und Probleme bietet, resultieren daraus aber auch wiederum neue Herausforderungen, nämlich die Reflexion über den Mehrwert und den Verlust, der sich aufgrund der geschilderten hermeneutischen Schwierigkeiten ergibt (s. auch Gius / Jacke 2015 und 2017). Wendet man sich nun der digitalen erklärenden Annotation im Verhältnis zu Ambiguität zu, so resultiert daraus – wie in den Fällen (1) bis (3) dargelegt – ein *trade-off* zwischen Klärung und Verdunkelung sowie die Notwendigkeit eines Mittelwegs zwischen der Überforderung bzw. Überfrachtung des Nutzers und der Gefahr, zu wenig Informationen anzubieten (s. dazu Berry 2012).

Der Vortrag widmet sich diesen Fragen und Herausforderungen anhand einiger ausgewählter Beispiele und versucht, folgende Aspekte zu präsentieren bzw. anzureißen: 1. die theoretischen Grundlagen der erklärenden Annotation und ihrer hermeneutischen Grundlagen im Hinblick auf die Ambiguität näher zu beleuchten; 2. das Verhältnis von digitalem Medium und literarischer Annotation weiter zu konzeptualisieren, und zwar vor dem Hintergrund der erschwerten Bedingungen durch Ambiguität; 3. die Automatisierung von Annotationen ambiger Items im digitalen Medium in theoretischer Hinsicht voranzutreiben.

Fußnoten

1. Die Phrase „grain of salt" wird z.B. von Henry James in The American metaphorisch verwendet (Kap. 9), während sie von Sir Alfred Tennyson in

der letzten Zeile seines Gedichts „Will“ zwar literal gebraucht, durch den Kontext ihrer Verwendung die Metaphorik durch die Rätselhaftigkeit des Ausdrucks „The city sparkles like a grain of salt“ aber aufgerufen wird.

Bibliographie

Assmann, Jan (1995). “Text und Kommentar: Einführung,” in: Assmann, Jan / Gladigow, Burkhard (eds.): *Text und Kommentar: Archäologie der literarischen Kommunikation IV*. München: Fink 9-33.

Bauer, Matthias / Zirker, Angelika (2015): “Whipping Boys Explained: Literary Annotation and Digital Humanities,” in: Siemens, Ray / Price, Kenneth M. (eds.): *Literary Studies in the Digital Age: An Evolving Anthology*. Ed. Ray Siemens and Kenneth M. Price. https://dlsanthology.commons.mla.org/whipping-boys-explained-literary-annotation-and-digital-humanities/ [letzter Zugriff: 11. September 2017].

Bauer, Matthias / Zirker, Angelika (2017): “Explanatory Annotation of Literary Texts and the Reader: Seven Types of Problems,” in: Zirker, Angelika / Bauer, Matthias (eds.): *International Journal of Humanities and Arts Computing*. 11.2 (2017): 212-32.

Bauer, Matthias / Zirker, Angelika (in Vorb.): *Strategies of Ambiguity*. Routledge [zur Publikation angenommen; erscheint 2018].

Berry, David M. (2012): “Introduction: Understanding Digital Humanities,” in: Berry, David M. (ed.): *Understanding Digital Humanities*. London: Palgrave Macmillan.

Bode, Christoph (1988): *Ästhetik der Ambiguität: Zur Funktion und Bedeutung von Mehrdeutigkeit in der Literatur der Moderne*. Tübingen: Niemeyer.

Booth, William (ed.) (1977). *Shakespeare’s Sonnets*. New Haven: Yale University Press.

Drucker, Johanna (2012). “Humanistic Theory and Digital Scholarship,” in: Gold, Matthew K. (ed.): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press 85-95.

Eggert, Paul (2009). “The Book, the E-text and the ‘Work-site,’” in: Deegan, Marilyn / Sutherland, Kathryn (eds.): *Text Editing, Print and the Digital World*. Aldershot: Ashgate 63-82.

Gius, Evelyn / Jacke, Janina (2015): “Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse’, *Zeitschrift für digitale Geisteswissenschaften*, 1 (2015), http://www.zfdg.de/sb001_006 [letzter Zugriff: 11. September 2017].

Gius, Evelyn / Jacke, Janina (2017): “The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis,” in: Zirker, Angelika / Bauer, Matthias (eds.): *International Journal of Humanities and Arts Computing*. 11.2 (2017): 233-54. [im Druck]

GRK 1808 (2017): “Forschungsprogramm des Graduiertenkollegs ‚Ambiguität: Produktion und Rezeption,” https://www.uni-tuebingen.de/forschung/forschungsschwerpunkte/graduiertenkollegs/grk-1808-ambiguitaet-produktion-und-rezeption/forschung/forschungsprogramm.html

Parry, Dave (2012). “The Digital Humanities or a Digital Humanism,” in: Gold, Matthew K. (ed.): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press 429-37.

van Peursen, Wido (2010): “Text Comparison and Digital Creativity: An Introduction,” in: van Peursen, Wido / Thoutenhoofdm Ernst D. / van der Weel, Adrian (eds.): *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*. Leiden: Brill 1-27.

Zirker, Angelika (2014): “‘To Be Taken with a Grain of Salt’: Charles Dickens and the Ambiguous Ghost Story,” in: Lennartz, Norbert / Koch, Dieter (eds.): *Texts, Contexts and Intertextuality: Dickens as a Reader*. Göttingen: Vandenhoeck & Ruprecht 163-80.

Zirker, Angelika / Bauer, Matthias (2017): “Guest Editors’ Introduction: Explanatory Annotation in the Context of the Digital Humanities,” in: Zirker, Angelika / Bauer, Matthias (eds.): *International Journal of Humanities and Arts Computing*. 11.2 (2017): 145-52.

Analysing Direct Speech in German Novels

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

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Würzburg, Deutschland

Zehe, Albin
zehe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Introduction

Detecting direct speech in fiction allows gaining insight into an important element of its narrative structure. In literary studies, there are assumptions on the factors influencing the distribution of direct speech, like genre, period and aesthetic complexity.

This paper aims to provide a detailed analysis of the use of direct speech across different time periods and domains. To create a reliable database for these analyses, we need to measure the usage of direct speech in a large and representative corpus. This task is more challenging than it may sound: While, nowadays, direct speech is often marked very explicitly by the use of quotes, this has not always been consistently the case. Many historical novels are not available in a well-edited form, meaning that there may be inconsistent use of quotation, or no quotation at all (Brunner, 2013). In this case, a more robust method for detecting direct speech is necessary.

Our first contribution is therefore a deep learning-based method to detect direct speech using large amounts of rule-based, but slightly flawed, labelled data extracted from raw text. This has multiple advantages over the use of manually annotated training data: First, manually annotating large amounts of text is very time-intensive and therefore costly. Furthermore, annotations for one type of texts may not be transferable to other types, leading to the necessity of new annotated data for new corpora. Being able to learn from the already existing weakly labelled data is therefore desirable, as this data can automatically be extracted for a new corpus.

Our second contribution is the application of this approach on curated texts to gain insight in trends of direct speech distribution. On one hand we try to look for development of direct speech over time, analysing a large dataset of novels from the nineteenth century, on the other hand we focus on differences in genre comparing contemporary high and low brow literature.

Related Work and Task Description

There have been several previous approaches to direct speech detection applying machine learning methods.

For example, Brunner (2013) tests rule-based and machine learning driven classification, as well as combinations of both, on German novels. She recommends using a pure machine learning approach (Random Forest), reaching an F1 score of 0.87.

Scheible et al. (2016) employ a simple greedy algorithm and a semi-Markov model, showing that the latter outperforms the previous state-of-the-art by achieving a precision of 0.88.

Although the results seem quite satisfying, these systems require a relatively large amount of labelled data for training. As stated above, this is problematic because of the need for expensive annotation and lack of transferability to other domains. Thus, our goal in this paper differs from that in previous work. We do not aim to set a new state-of-the-art in direct speech detection, but instead:

a) present a method that can leverage large amounts of weakly labelled data extracted from raw text, and

b) use this model for the analysis of different distributions of direct speech across genres or time-periods.

To the best of our knowledge, the second task has never been done on a large collection of texts.

Corpus and Resources

The following experiments are based on three German corpora. The first one is a large corpus containing 4600+ public domain novels including texts from the TextGrid digital library[1] and Project Gutenberg[2]. We will refer to this as the Corpus *Public Domain*, PD. The second one contains 800+ texts of current popular genres like romance, crime or science-fiction (Corpus *Low Brow*, LB). Finally, we use a corpus with 200 novels nominated for the *German Book Prize* or the *Georg Büchner Prize* (Corpus *High Brow*, HB).

In order to train and evaluate our classifiers, we need to obtain labels specifying which parts of the texts contain direct speeches. To this end, we chose two strategies:

For training our classifiers, we decided to extract weak labels using a simple rule based on quotation, implying everything written between quotation marks is direct speech. To yield high accuracy for this approach, it is necessary to use a well-edited collection of texts. Our PD corpus contains such a subset, which we refer to as our *Kerncorpus*. This *Kerncorpus* consists of 250 high and middle brow texts (those from the TextGrid digital library), has been manually edited and is assumed to have a mostly consistent use of quotation.

Using our quotation rule on the *Kerncorpus* resulted in a dataset where about 36% of tokens were marked as direct speech. In order to assess the quality of these weak labels, we gave 500 of the sentences to domain experts for manual correction. We found that there was an error-rate of about 3% in those sentences, mostly caused by nested direct speech or inscriptions being enclosed by quotation marks.

For further evaluation, we chose to annotate a smaller subset of the corpus *LB* by hand. We selected 50 snippets from texts of low brow literature. This dataset, referred to as *ALB*, is relatively skewed towards text outside direct speech, with only about 18% of tokens in a direct speech.

Experiments

The following experiments use both labelled subsets described above, the large *Kerncorpus*[3] and the smaller *ALB*. For all experiments, quotation marks are removed from the texts. This is done to avoid training models that rely only on the formal style of qualifying direct speech, but also consider implicit signs like the use of first person verbs or speech words.

We conducted experiments on two different levels, starting with a sentence classification task, which is then refined to detect direct speech on word-level.

Sentence-Level Classification

In our first classification task, documents are split into sentences and vectorised by storing each sentence in a bag-of-words representation. To create a baseline for measuring the advantage using deep learning for direct speech recognition, we compared the performance of traditional machine learning algorithms on our labelled datasets. Training and testing some of the most common machine learning classifiers to detect sentences containing at least one word of direct speech leads to an accuracy of **0.85** using Logistic Regression; for more results see Table 1.

Using the same setting and replacing machine learning with a combination of recurrent and convolutional neural networks (see Chollet 2017 and Goodfellow 2017) ended up with an accuracy of **0.84**.

Table 1: Results of traditional machine learning algorithms for direct speech detection.

Algorithm	Multinomial Naive Bayes	Random forest	SVM linear kernel	Logistic Regression	K-nearest neighbours	Passive Aggressive	Perceptron
Accuracy	0.84	0.78	0.80	**0.85**	0.67	0.78	0.76

Since we noticed that three of our classifiers all ended up with about the same score, we decided to give the task to two human annotators to establish an upper bound. We selected 250 sentences for manual annotation and again removed all quotation marks. Both annotators ended up with an accuracy comparable to that of the best machine learning methods, 84% and 82.8% respectively. From this result we concluded that it is not expedient to further optimise the sentence classification task, as we had already reached human-level accuracy.

Word-Level Classification

Because of the results from the previous section, we decided to modify our task to a word-level prediction, which enables us to include more context by ignoring sentence boundaries and at the same time make more fine-grained predictions. In this second classification task, each word is to be classified separately as inside or outside a direct speech. As baseline for this task, we trained a Linear Chain Conditional Random Field (CRF) that was only given the word itself and its part-of-speech tag. This CRF stagnated at a comparably low accuracy of 0 **.71** using cross-validation on the *Kerncorpus*.

Since our goal was to provide the classifier with more context, we chose to use an architecture based on recurrent neural networks, which are able to deal with relatively large contexts. Our assumption here is that, for a good classification, we need context from both before and after the target word itself, as markers for direct speech can be found at the beginning or the end of the direct speech. We thus designed a two-branch network, visualised in Figure 1. This network receives as input a text-segment, specifically the target word in its context. The words of the input are then passed through an embedding layer and split into two parts, where the first part contains the context up to the target word and the second part contains the context following the target word. The target word itself is contained in both parts. Each part is passed through three separate LSTM-layers. In the future-branch, the context is passed through the layers in reverse, so that the target word is the last word to be read in both branches. The LSTM-layers in the past-branch are stateful and can therefore theoretically retain the entire context of the novel up to the target word. The outputs of the final LSTM-layer of both branches are concatenated. The final prediction is made based on this concatenation by a fully connected layer.

In our best setup, we used 60 words before and after the target word as context.

Training on one half of the *Kerncorpus* and evaluating on the other half, this setup yielded an accuracy of **0.83**. Training on the full *Kerncorpus* and evaluating on the manually annotated *ALB* reached an even better accuracy of **0.90**.

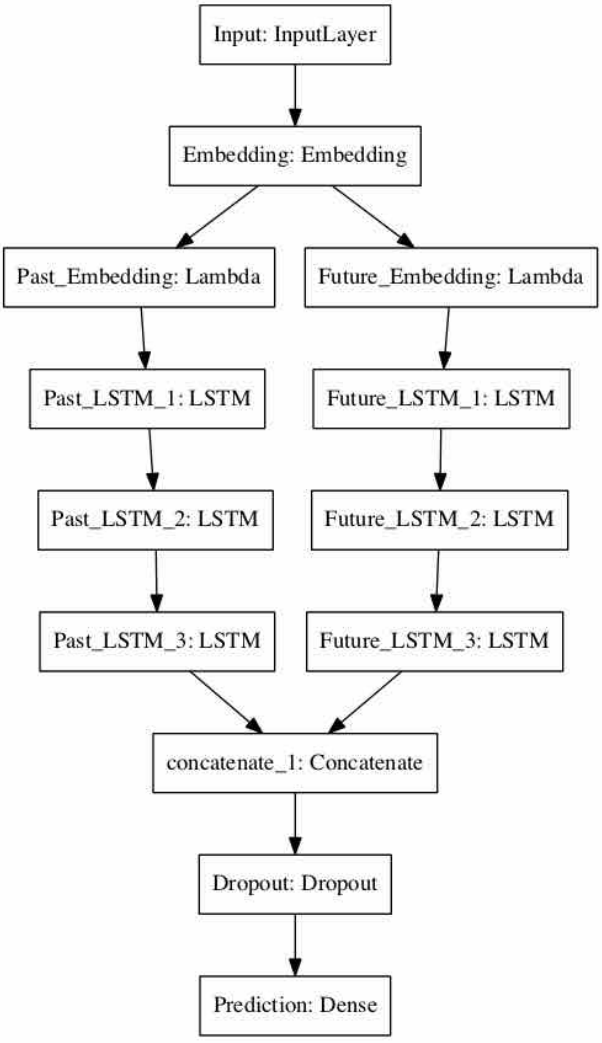


Figure 1: Architecture of recurrent network to detect direct speech.

Distribution of direct speech

In the following experiments, we used the model based on the architecture described above. We trained this model on the *Kerncorpus* and used it to detect direct speech in the complete corpora *PD*, *LB* and *HB*. Here, we describe our findings on these corpora.

Direct speech in 19th Century Fiction

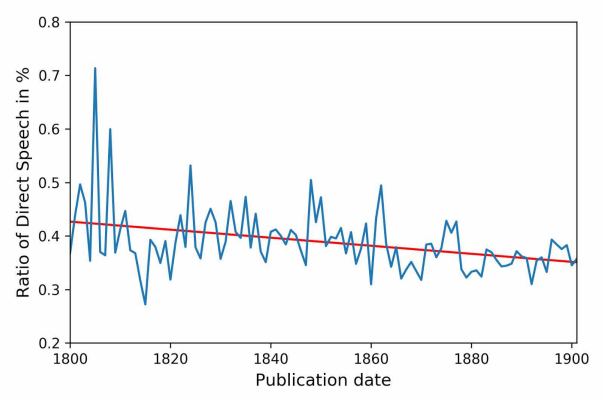


Figure 2: Ratio of direct speech in German novels from 1800 – 1900.

Figure 2 shows the ratio of direct speech in German novels from 1800 till 1900 based on the texts from Corpus *PD*. The regression line indicates a decline of direct speech over time; at the same time, we can observe a decrease of variance. The strong variations between certain years, especially in the early 19th century, are caused by low numbers of provided texts (see Fig. 3). For instance, the peak in 1805 can be explained by the first publication of Denis Diderots "Herrn Rameaus Neffe", a philosophical dialogue-based novel.

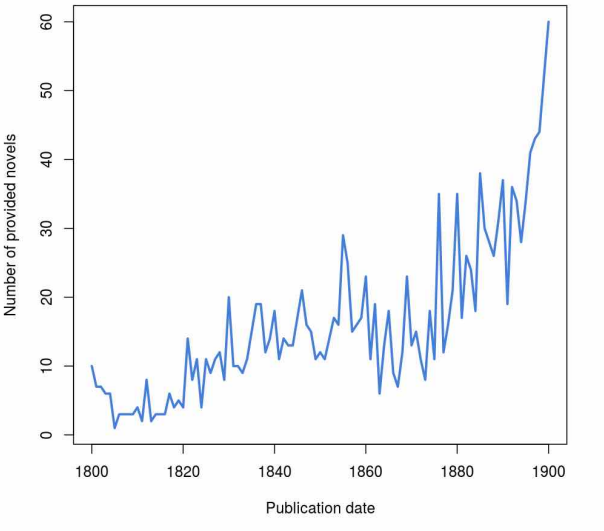


Figure 3: Number of provided novels per year.

Distribution of direct speech in low and high brow literature

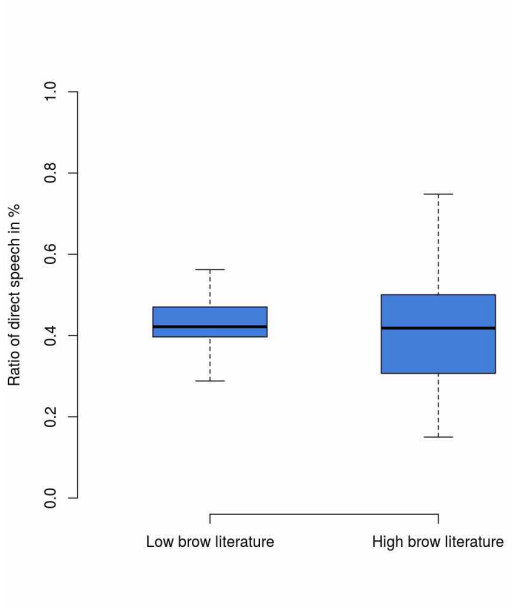


Figure 4: Ratio of direct speech in German low and high brow novels after 1945.

There is an assumption in literary studies that a huge amount of direct speech is an indicator of low brow fiction. Figure 4 shows the ratio of direct speech between Corpus *LB* and *HB*. While the mean usage of direct speech is nearly equal in both groups, the high brow literature is far more variable.

This finding is contrary to the assumption mentioned above. We propose that, while there is no clear difference in the average use of direct speech between high and low brow literature, authors in high brow literature are far more flexible in choosing how much direct speech they use in their novels. Low brow literature, on the other hand, is expected to have a rather constant amount of dialogue.

Conclusion and Future Work

In this paper, we introduced a neural network architecture that is able to learn the classification of direct speech by training on weakly labelled data. This network works purely on the raw text of a novel by taking into account a relatively large context. We also demonstrate that training on weakly labelled data leads to satisfying results.

While an accuracy of 0.9 is remarkable, there is still need for optimisation. Recent developments in the performance of neural networks by adding an attention mechanism (see Rush 2015) could improve the results.

We used our neural network to analyse the distribution of direct speech over time and genres. Besides algorithmic refinements, there is a lot of potential in adding more text to our corpus and refining metadata to allow more sophisticated research questions like differences between or development of direct speech in certain genres.

Fußnoten

1. https://textgrid.de/digitale-bibliothek
2. https://gutenberg.spiegel.de
3. We cannot use the remaining texts for either training or evaluation, as we do not have any reliable source of labels for these texts.

Bibliography

Brunner, Annelen (2013): “Automatic recognition of speech, thought, and writing representation in German narrative texts”, in *Literary and Linguistic Computing. Vol. 28 (2013).*

Chollet, Francois (2017): “Deep Learning with Python”. Manning Publications. New York. (Preprint: https://www.manning.com/books/deep-learning-with-python)

Goodfellow, Ian / Bengio Yoshua / Courville, Aaron (2016): “Deep Learning”. MIT Press. (URL: *http://www.deeplearningbook.org*)

Rush, Alexander M. / Chopra, Sumit / Weston, Jason (2015): “A Neural Attention Model for Abstractive Sentence Summarization”. *arXiv preprint arXiv:1509.00685.*

Scheible, C., Klinger, R. & Padó, S. (2016): “Model Architectures for Quotation Detection”, in *Proceedings of ACL (p./pp. 1736–1745).*

An den Grenzen der Interoperabilität: Eine kritische Reflexion über digitale Forschungsdaten und -anwendungen in der Online-Edition des Projekts "Die Schule von Salamanca"

Wagner, Andreas
wagner@rg.mpg.de
Max-Planck-Institut für europäische Rechtsgeschichte, Frankfurt am Main

Glück, David
glueck@rg.mpg.de
Akademie der Wissenschaften und der Literatur, Mainz; Johann Wolfgang Goethe-Universität Frankfurt am Main

Einleitung

Der Beitrag diskutiert kritisch, welche nicht allein arbeitsökonomischen, sondern vor allem intellektuellen, methodologischen und wissenschaftstheoretischen "Kosten" mit der Digitalisierung und dem dadurch etablierten Fokus auf Fragen der Interoperabilität entstehen. Die deutlichen wissenschaftlichen Vorteile von Interoperabilität müssen nämlich einem Umbau (und z.T. Rückbau) in wissenschaftlichen Modellierungen und im Verständnis wissenschaftlicher Erkenntnisvermehrung gegenüber gestellt werden. Dabei wird im Sinne des doppelten Genitivs "Kritik der digitalen Vernunft" sowohl das naheliegende Phänomen diskutiert, dass die digitale Transformation Hoffnungen weckt und Lösungen wissenschaftlich-methodologischer Schwierigkeiten nahelegt, die sich (z.T. erst bei der Implementierung) als wissenschaftlich nicht akzeptabel erweisen, als auch der umgekehrte Fall, dass die digitale Transformation ein kritisches Umdenken in der Ausrichtung der eigenen wissenschaftlichen Arbeit und eine Umorientierung wissenschaftlicher Ambitionen erzwingt.

Diese Diskussion wird im Durchgang durch einige beispielhafte Entwicklungen im seit vier Jahren laufenden Projekt "Die Schule von Salamanca" geführt, bevor ein Versuch der Verallgemeinerung unternommen wird. Dass Interoperabilität signifikante wissenschaftliche Kosten mit sich bringt, heißt im Übrigen nicht, dass diese nicht womöglich durch die Vorteile aufgewogen würden. Es ist allerdings in der Orientierung der Projektarbeit wichtig, sich an den „Grenzen der Interoperabilität“ abzuarbeiten und sich regelmäßig beide Seiten dieser Bilanz vorzulegen.

Interoperabilität

Mit der digitalen Transformation sind die Möglichkeiten des überregionalen und interdisziplinären Austauschs und der Weiterverwendung von Forschungsdaten in einer ganz neuen Weise möglich geworden. So hat sich der Begriff der Interoperabilität als zentrales Paradigma in den Digital Humanities etabliert, um den Anspruch zu beschreiben, diese Möglichkeiten methodisch auszubauen und ein Evaluationskriterium für Forschungsleistungen und -ergebnisse anzubieten. Interoperabilität ist auf mehreren Ebenen zu verstehen (vgl. Gradmann 2009) und wird auf den konkreten technischen und syntaktischen Ebenen vor allem durch Standards für Datenformate und Schnittstellen, sowie auf der pragmatischen Ebene u.a. durch den Bezug auf Normdaten realisiert. Mag Interoperabilität auf der semantischen Ebene schließlich vor einiger Zeit noch als ein utopisches Ziel angemutet haben (vgl. Baumann 2011), so haben sich etwa durch die Selbstreflexion digitaler geisteswissenschaftlicher Forschungs-Arbeit (Hughes et al. 2016) oder in der Beschreibung von kulturellen und geistigen Phänomenen (Le Boeuf et al. 2017; Stead et al. 2015) deutliche Fortschritte in der Modellierbarkeit geisteswissenschaftlicher Phänomene und in der Verfügbarkeit und Relationierbarkeit von semantischen Forschungsdaten und -ergebnissen ergeben.

Wie Forschungsprojekte den derart gewachsenen Interoperabilitätserwartungen gerecht werden können, kann jedoch kaum allgemein, sondern nur im Rahmen ihrer jeweils spezifischen Arbeit ermittelt werden (zu einer projektunabhängig formulierten Handreichung vgl. aber Beer et al. 2014). Dabei stellt sich die Frage nach dem Nutzen, den Grenzen und den Kosten interoperabler Techniken im Projekt "Die Schule von Salamanca" insofern auf besondere Weise, als dieses einerseits über einen langen Zeitraum und mit Blick auf eine langzeitverfügbare Erschließung von relativ großen (Text-)Datenmengen operiert,

andererseits aber auch die Erforschung innovativer Textaufbereitungsverfahren und Webanwendungsfunktionen zur Ermöglichung neuartiger Forschungserkenntnisse für die beteiligten Fachwissenschaften anstrebt.

Das Projekt

Im durch die Akademie der Wissenschaften und der Literatur | Mainz geförderten und insgesamt auf 18 Jahre angelegten Projekt "Die Schule von Salamanca. Eine digitale Quellensammlung und ein Wörterbuch ihrer juristisch-politischen Sprache" (Duve et al. 2013) werden voraussichtlich insgesamt etwa 120 Texte der gleichnamigen Schule iberischer Theologen und Juristen des 16. und 17. Jahrhunderts nach und nach digitalisiert und als Volltexte erfasst. Die in TEI-XML ausgezeichneten und aufwändig normalisierten Texte werden nicht nur strukturell, sondern auch im Hinblick auf als Linked Open Data (LOD) referenzierbare Entitäten – etwa Personennamen – erschlossen; dabei werden die Text-Digitalisate, die Volltexte und die LOD-Datensammlungen online bereit gestellt. Hinzu kommt ein digitales (und schließlich auch gedrucktes) Wörterbuch, in dem sowohl biographische Informationen zu den in der Edition vertretenen Autoren als auch zentrale Begriffe der Rechts- und politischen Ideengeschichte und deren Entwicklung im Diskussionszusammenhang der "Schule von Salamanca" erfasst werden. In diesen beiden Säulen des Projekts ist die Erschließung und Repräsentation der Struktur der internen Verweise (Autoren, die sich wechselseitig zitieren, Wörterbuchartikel, die auf Textstellen verweisen) eines der zentralen wissenschaftlichen Ergebnisse.

Die sowohl für die Benutzung als auch für die Bereitstellung der Daten als zentrales Portal dienende Webseite des Projekts (https://salamanca.school/) ist technisch in Form einer komplex modularisierten Webanwendung implementiert, die fachwissenschaftlichen BenutzerInnen eine Vielzahl von Funktionen bieten soll, z.B. eine geräteübergreifende und performante Darstellung der Editionstexte, eine intelligente Suchfunktion, die den frühmodern-lateinischen und -spanischen Volltexten angepasst ist, und eine feinkörnige Referenzierung der Texteinheiten. Durch technische Mechanismen (content negotiation, API, RESTful Microservice-Architektur), Exportfunktionen (abschnitts-, text- oder corpusweise, plaintext-, TEI- oder andere Formate) und die Orientierung an Formatstandards wird Anforderungen der Interoperabilität explizit Rechnung getragen. Der Quelltext der Webanwendung wird bis Januar 2018 ebenfalls veröffentlicht und fortan in Open Source weiterentwickelt werden.

Interoperabilitäts-"Konflikte": Einige Beispiele

Einhergehend mit der Open Source-Veröffentlichung der Webanwendung soll über einige repräsentative Aspekte der Projektarbeit reflektiert werden, in denen sich an vermeintlich technischen Herausforderungen Konflikte um die wissenschaftlichen Implikationen von Interoperabilität entzünden:

a) Textrepräsentation:

Während in diesem Kernbereich der Projektarbeit gegenwärtig noch mit einem eigens spezifizierten und den bisherigen Forschungsanforderungen entsprechenden "idiosynkratischen" TEI-Format gearbeitet wird und die Datenmodellierung somit eher als "research-driven" (vgl. Flanders/Jannidis 2016: 233) bezeichnet werden kann, wird eine langfristige Verfügbarmachung der Texte in einem interoperableren, "curation-driven" Format wie etwa *TEI Simple* (Text Encoding Initiative Consortium 2016) erwägt. Dies ist nicht zuletzt auch durch den Blick auf die Wissenschaftsförderung motiviert (vgl. etwa DFG 2015, Anhang), die die Interoperabilität von Textauszeichnungen dem erst noch nachzuweisenden Erkenntnisgewinn 'reicher' Annotationen gegenüberstellt. Nun beinhaltet das XML-Ökosystem der Edition allerdings auch Informationen wie etwa Metadaten zur Zeichenkodierung (die nicht zuletzt Interoperabilitäts-Funktionen erfüllen), die sich nach derzeitigem Stand nicht ohne weiteres in *TEI Simple* abbilden lassen. So ergibt sich eine nur durch einigen Aufwand aufzulösende Spannung zwischen detaillierter wissenschaftlicher Gegenstandsbeschreibung und Maßnahmen zur Förderung der Nachnutzbarkeit der eigenen Ressourcen (z.B. des Angebots mehrerer alternativer Datenformate), und es stellt sich die Frage, wann und durch wen jener Aufwand erbracht werden soll. Selbst eine gegenüber den "Experten"-Annotationen tolerantere Zielvorgabe, die sich etwa am Konzept des 'Interchange' (als einer durch menschliche Interpretation vermittelten Nachnutzung, vgl. Holmes 2017, Baumann 2011) orientiert und im Vergleich zum Modell einer bruchlosen Weiterverwertbarkeit der Daten durch automatische Prozesse gemäßigtere Ansprüche erhebt, sieht sich ähnlichen Fragen der

Standardisierung von Schnittstellen und Datenformaten ausgesetzt.

b) Modulare Infrastruktur:

Im Zuge der Einrichtung einer Linked Open Data-Infrastruktur haben wir *content-negotiation*-Mechanismen, Weiterleitungen und die Adressierung unterschiedlicher Funktionen über verschiedene Server und Server-Adressen eingeführt. Diese Entwicklung legt eine Fortsetzung nahe, die den Umbau der Web-Anwendung insgesamt in ein Ensemble von Microservices bedeuten würde (vgl. Wolff 2016), in dem Daten und Dienste aufs Engste verschränkt sind. Während dies die Interoperabilität, d.h. konversions- und barrierearme Nachnutzbarkeit der Daten enorm verbessert (etwa dadurch, dass auf verschiedenen Ebenen in verschiedenen Formaten und verschiedenen Granularitäten Daten und Dienste verfügbar sind, vgl. Turska et al. 2016) hat es jedoch auch den Nachteil, dass der Daten- und Anwendungszusammenhang – der eben auch *als Zusammenhang* eine Forschungsleistung darstellt – nur in einer sehr viel aufwändigeren Weise repliziert und ggf. archiviert werden kann: Während erste Infrastrukturen die Archivierung und langfristige Zugänglichkeit von Javascript-Anwendungen erlauben sollen (vgl. Bingert/Buddenbohm 2016; kritischer: Brunelle 2016), so ist dies für eine solche Infrastruktur mit mehreren kooperierenden und kommunizierenden Servern nur sehr viel schwerer vorstellbar.

c) Adressierung, Versionierung und Persistenz:

Im Zusammenhang mit der Adressierung von einzelnen Textpassagen haben wir ein System eingeführt, das sowohl semantische Aussagen über Text-Entitäten als auch die Realisierung der Verweisstrukturen auf implementierungs- und plattformunabhängige, intellektuell intuitive Weise erlaubt (im Anschluss an das Canonical Text Services-Schema, Blackwell/Smith 2014; vgl. Wagner 2016). Diese Struktur der komplexen Verknüpfung von Ressourcen und Entitäten untereinander verträgt sich aktuell nicht mit etablierten Methoden, Dokumente persistent zu identifizieren und die Überarbeitungshistorie der Ressourcen in einer Versionierung transparent und nachvollziehbar zu machen. Diese Methoden spezifizieren nämlich zumeist den Umgang mit einem ganzen Dokument und vernachlässigen den Bedarf, Entitäten innerhalb des Dokuments (persistent) zu referenzieren oder die Einbettung jenes Dokuments in ein Corpus zu verwalten. Wenn beispielsweise ein Dokument verändert wird, muss es unter anderem normalerweise eine neue persistente ID erhalten - damit entsteht ein Aktualisierungs- oder mindestens Kontrollbedarf bei den Querverweisen innerhalb des Dokuments sowie in allen weiteren Dokumenten des Corpus (und in unserem Falle in allen Artikeln des Wörterbuchs), welche Verweise auf das aktuelle Dokument enthalten. Mechanismen wie das Memento framework (van de Sompel/Nelson 2015) bieten eine transparente und flexible Versionierung, die für Web-Dokumente wie für Semantische Ressourcen gleichermaßen funktioniert, sind jedoch in ihrer Integration mit PID-Systemen ebenfalls noch nicht erprobt. Versteht man diese Motive in einem Interoperabilitäts-Zusammenhang, lesen sie sich wie ein Konflikt zwischen etablierten Lösungen auf verschiedenen Interoperabilitäts-Ebenen und es fragt sich, wie kanonische Verweissysteme (funktional-pragmatische Ebene nach Gradmann 2009) und Persistent Identifiers (technische Ebene) miteinander harmonieren können.

Diskussion

Die hier am Beispiel digitaler Projektarbeit aufgezeigten Probleme etwa einer gesteigerten Spannung zwischen „Interoperabilität und Expressivität“ (Baumann 2011) weisen auf grundlegendere wissenschafts- und erkenntnistheoretische Fragestellungen hin, die nicht zuletzt das Selbstverständnis der Digital Humanities und die mit ihnen verbundenen Hoffnungen berühren: Ist Expertenwissen am Ende vielleicht gar nicht "interoperabel"? Erweisen sich die Möglichkeiten, durch Verknüpfung digitaler Forschungsdaten und -anwendungen verschiedenster Kontexte neuartige Einsichten zu generieren, als stark begrenzt? Eine wissenschaftspolitische Pointe des Beitrags besteht so darin, auf die entscheidende Rolle aufmerksam zu machen, die die Erwartung darüber spielt, an welchem Ort – zwischen den Disziplinen oder “tief” in den fachwissenschaftlichen Spezialdiskursen – wissenschaftlicher Fortschritt erzielt wird.

Bibliographie

Baumann, Syd (2011): “Interchange vs. Interoperability”, in: *Proceedings of Balisage: The Markup Conference 2011* 7, https:// doi .org/ 10.4242/BalisageVol7.Bauman01 [letzter Zugriff: 13.01.2018].

Beer, Nikolaos / Herold, Kristin / Kolbmann, Wibke / Kollatz, Thomas / Romanello, Matteo / Rose, Sebastian / Walkowski, Niels-Oliver (2014): "Interdisciplinary Interoperability", DARIAH-DE Working Papers Nr. 3, http://nbn-resolving.de/urn.nbn.de.gbv:7-dariah-2014-1-0 [letzter Zugriff: 18.01.2018].

Bingert, Sven / Buddenbohm, Stefan (2016): "Die HDC-Anwendungskonservierung - ein Dienst zur Archivierung und Bereitstellung komplexer Forschungsergebnisse", in: *GWDG-Nachrichten* 11/2016: 7-9 https://www.gwdg.de/documents/20182/27257/GN_11-2016_www.pdf [letzter Zugriff: 23.09.2017].

Blackwell, Christopher / Smith, Neel (2014): "The Canonical Text Services protocol, version 5.0.rc.2", https://cite-architecture.github.io/cts_spec/ [letzter Zugriff: 25.09.2017].

Brunelle, Justin F. / Kelly, Mat / Weigle, Michele C. / Nelson, Michael L. (2016): "The impact of JavaScript on archivability", in: *International Journal on Digital Libraries* 17/2: 95–117, https://doi.org/10.1007/s00799-015-0140-8 [letzter Zugriff: 25.09.2017].

DFG (2015): "Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft", http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf [letzter Zugriff: 12.1.2018].

Duve, Thomas / Lutz-Bachmann, Matthias / Birr, Christiane / Niederberger, Andreas (2013): "Die Schule von Salamanca: eine digitale Quellensammlung und ein Wörterbuch ihrer juristisch-politischen Sprache. Zu Grundanliegen und Struktur eines Forschungsvorhabens". Mainz: Akademie der Wissenschaften und der Literatur. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:hebis:30:3-324011 [letzter Zugriff: 12.1.2018].

Flanders, Julia / Jannidis, Fotis (2016): "Data Modeling", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A New Companion to Digital Humanities*. Chichester: Wiley Blackwell 229-237.

Gradman, Stefan (2009): "Interoperability. A key concept for large scale, persistent digital libraries". *Digital Preservation Europe* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.363.6311&rep=rep1&type=pdf [letzter Zugriff: 25.09.2017].

Holmes, Martin (2017): "Whatever happened to Interchange?", in: *Digital Scholarship in the Humanities* 32 (Issue suppl_1): i63–i68 https://doi.org/10.1093/llc/fqw048 [letzter Zugriff: 25.09.2017].

Hughes, Lorna / Constantopoulos, Panos / Dallas, Costis (2016): "Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines", in: Susan Schreibman / Ray Siemens / John Unsworth (eds.), *A New Companion to Digital Humanities*, Wiley & Sons, 150–170, https://doi.org/10.1111/b.9781118680643.2016.00013.x [letzter Zugriff: 13.1.2018].

Le Boeuf, Patrick / Doerr, Martin / Ore, Christian Emil / Stead, Stephen et al. (2017): "Definition of the CIDOC Conceptual Reference Model", Version 6.2.2, http://www.cidoc-crm.org/sites/default/files/2017-09-30%23CIDOC%20CRM_v6.2.2_esIP.pdf [letzter Zugriff: 13.1.2018].

Schmidt, Desmond (2014): "Towards an Interoperable Digital Scholarly Edition", in: *Journal of the Text Encoding Initiative* 7, https://doi.org/10.4000/jtei.979 [letzter Zugriff: 13.1.2018].

Stead, Stephen / Doerr, Martin et al. (2015): "CRMinf: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation", http://www.cidoc-crm.org/crminf/sites/default/files/CRMinf-0.7%28forSite%29.pdf [letzter Zugriff: 13.1.2018].

Text Encoding Initiative Consortium (2016): "TEI Simple" https://github.com/TEIC/TEI-Simple [letzter Zugriff: 25.09.2017].

Turska, Magdalena / Cummings, James / Rahtz, Sebastian (2016): "Challenging the Myth of Presentation in Digital Editions", in: *Journal of the Text Encoding Initiative* 9 http://jtei.revues.org/1453 [letzter Zugriff: 25.09.2017].

van de Sompel, Herbert / Nelson, Michael L. (2015): "Reminiscing about 15 Years of Interoperability Efforts", in: *D-Lib Magazine* 21 (11/12) http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html [letzter Zugriff: 25.09.2017].

Wagner, Andreas (2016): "What's in a URI? Part I: The School of Salamanca, the Semantic Web and Scholarly Referencing" https://blog.salamanca.school/2016/11/15/whats-in-a-uri-part-1/ [letzter Zugriff: 22.09.2017].

Wolff, Eberhard (2016): "Microservices. Grundlagen flexibler Softwarearchitekturen". Heidelberg: dpunkt.verlag.

A Reporting Tool for Relational Visualization and Analysis of Character Mentions in Literature

Barth, Florian
florian.barth@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Kim, Evgeny
evgeny.kim@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Murr, Sandra
sandra.murr@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Klinger, Roman
roman.klinger@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Introduction and Motivation

The emergence of computational methods of text processing has created new paradigms of research in literary studies in recent years (Jockers & Underwood, 2016), for instance *distant reading* to find patterns and regularities (Moretti, 2005). Network analysis and extraction of information about relations between characters from literary texts is an example for distant reading methods. Such information can not only be helpful for better understanding of character interactions but can also facilitate the comparison of thereof in different texts.

Existing tools of text analysis and network visualization such as Voyant [1] or Gephi [2] are either missing modules for character network analysis or require preliminary steps on data preprocessing from the user and therefore are not easy-to-use for some humanities scholars who lack programming skills. Interactive tools in addition often lack features to ensure reproducibility of results.

We present our ongoing effort on closing this gap by developing a literary analysis reporting tool *rCAT* [3], whose primary purpose is to provide an easy-to-use, stable, and reusable solution for automatic extraction of relational information from text and to characterize these relationships automatically to provide the user with deeper qualitative insight. We opt for implementation as a web-based reporting tool instead of an interactive tool for two reasons: (1) automatically generated reports in PDF format can serve as a stable foundation for discussion and can be reused in publications and visualizations easily, and (2) the results are clearly connected to the chosen input parameters such that reproducibility of results is ensured.

As a use-case study, we apply *rCAT* to Johann Wolfgang von Goethe's epistolary novel *Die Leiden des jungen Werthers*. On the basis of this epistolary novel, we show that not only the network can be generated, but also the characteristic triangular relationship of the protagonists is easily identified. The goal is to automatically determine this triad in the original text and in the adaptations that have been published since the publication of *Werther* in 1774.

Previous Work

Previous research on social networks in literary fiction generally fall into one of the two categories: (1) works that explore methods for extracting and formalizing character networks (*cf.*, Elson et al. (2010), Agarwal et al. (2012, 2013), Park et al. (2012)), and (2) works that primarily focus on qualitative implications of network analysis (*cf.*, Rydberg-Cox (2011), Moretti (2011), Nalisnick & Baird (2013), Jayannavar et al. (2015)). It is common to address both tasks at the same time, as in Beveridge & Shan (2016), who introduce a number of formal measures for analyzing the centrality of the characters in *Game of Thrones* books, which results in both expected and surprising findings.

Building on graph theory extensively elaborated in the past fifty years (e.g., Bondy and Murty, 1976 or West, 2001), our work is similar to Beveridge & Shan (2016), in particular, in terms of the weighted degree measure, and to Park et al. (2012), in terms of distance measure for detecting closely related characters in a text.

Methods

In the following, we explain the different components in *rCAT*, which are available for text ana-

lysis. After that, we discuss the results based on a use-case study.

Character lists and character identification

To detect character mentions in the text we use a fundamental named-entity recognition approach based on dictionaries. This approach is suitable for scholars who analyze texts they already know. Consequently, we opt for a transparent and simple character recognition procedure: The user provides a list of character names to be included in the analysis specifying a canonical name form and all variations thereof she would like to take into account (*e.g.*, “Lotte” is the canonical name and “Lotten”, “Lottens”, “Lottgen”, “Lottchen”, “Charlotten S.”. are its variants).

Relation detection and context words

We define the closeness of relationship between two characters using a *distance measure dist X (p,q)*, where *p* and *q* are the strings corresponding to these characters and *X* is the number of tokens between them (Park et al., 2012). In addition, we introduce the *context measure cont Y (p,q)*, where *p* and *q* are the strings corresponding to these characters and *Y* is the number of tokens before the character *p* and after the character *q*. While the former measure allows for detecting those characters that are closely related to each other, the latter one enables a contextual analysis of their relationship.

Network analysis

We visualize the network of characters with an undirected graph *G=(V,E)*, where *V* are the vertices, each vertex corresponding to one character, and each edge $E=(V\,i, ,V\,j\,)$ corresponding to relations between pairs of characters. We output the following measures for each character node: *degree*, *edge weight*, *weighted degree* and *density*. The degree is the number of edges occurring with a given vertex. The edge weight, $w\,i,j \geq 0$, is defined as the number of interactions between the vertices V *i* and V *j*. The weighted degree is the sum of weights of the edges occurring with a vertex *i*. Density is the ratio of occurring edges between two vertices and all possible vertex pairs.

Word clouds

Word clouds are an approach to visualize the vocabulary of a text. The size of one word corresponds to its frequency. We use two different kinds of word clouds: For each character in the character list, we show word clouds based on the context of a window size *n.* For each pair of characters occurring in the network, we present a word cloud based on the words between them as well as on the words found in the context. Both types of word clouds can be filtered to the specific word fields (words from specific domains) which is helpful in gaining a focused insight into the characters relations.

Word Field developments

We plot the timeline of multiple predefined world fields (specified by word lists) in the text. This feature is helpful in representing how certain fields (*e.g.*, concepts, emotions) develop throughout the narrative (Kim et al., 2017).

Implementation

The tool was developed using Python v.3.6 and the Flask [4] web development framework. The tool outputs a single PDF report. The resulting document contains information from the analysis modules described in the previous section. Network graphs included in the report are generated with *graphviz*. Additionally, the tool can generate a CSV file that can be used as input to Gephi.

Use-case Demonstration

For a use-case analysis, we apply *rCAT* to *Die Leiden des jungen Werther* by Johann Wolfgang Goethe with the following parameters: X=8, Y=5, stop words removed (previous work focused on this analysis without rCAT, *cf.* Murr, 2017).

In Goethe's epistolary novel, the protagonist Werther describes his unhappy love for Lotte, who is engaged to Albert. The characteristic triangular relationship in the novel arises from this constellation (protagonist - beloved woman - antagonist). With *rCAT* we expect to identify and characterize this relationship. Figures 1 and 2 show a sample network analysis output (tables are shown only partly).

The protagonist Werther shows a degree of 21, which is the number of characters with whom he interacts. The closest relationship measured by edge weight (Figure 2) is observed between Werther and Lotte (81 interactions). The antagonist Albert has a low degree of 3. However, his weigh-

ted degree is 36 (third highest after Werther and Lotte), which confirms his important role in the triangular relationship.

Degrees

Character (Node)	degree	weighted degree
Werther	21	184
Lotte	12	101
Albert	3	36
Wilhelm	3	34
Vetter	2	3
Magd	1	1
Schreiber	2	2
Hans	1	1
Marianne	1	1
Grafen von M . .	1	1
Graf v. C.	4	17

Illustration 1: Degrees and weighted degrees for most important characters of Goethe's Werther

Weights for Edges

Character Pair (Edge)	Weight
Werther -- Lotte	81
Werther -- Albert	26
Werther -- Wilhelm	32
Werther -- Vetter	2
Werther -- Magd	1
Werther -- Schreiber	1
Werther -- Hans	1
Werther -- Marianne	1
Werther -- Graf v. C.	12

Illustration 2: Edge weights

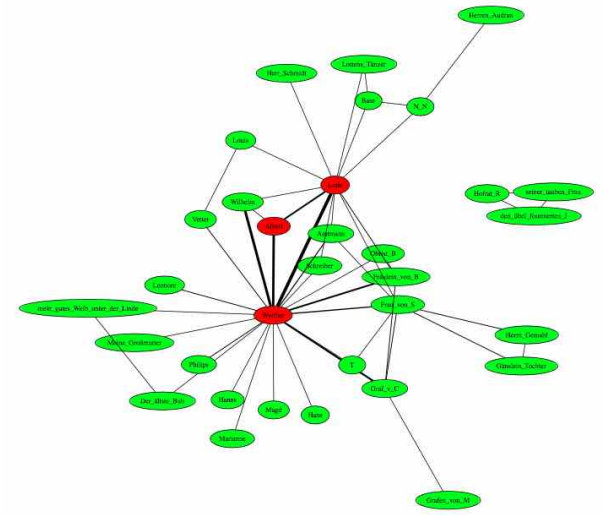


Illustration 3: Complete network of Goethe's Werther

Highlighted in red is the typical triangular relationship in Goethe's novel, which corresponds to the three highest weighted degrees. In further steps, we will use *rCAT* to analyze the adaptations of Goethe's novel with a focus on this triad.

To better characterize the edges, the tool outputs top- *n* word clouds sorted by edge weight (*n* is specified by the user) for character pairs and by degree for single characters. Figure 4 and 5 show examples of the word clouds for character pairs filtered to the words from the emotion domain.

Illustration 4: Word clouds for Werther-Lotte

Illustration 5: Werther-Albert

The word clouds enable first conclusions about the relationships of the characters. Werther and Lotte's word cloud characterizes their ambivalent relationship. The key words "Leidenschaft" and "Freude" reflect Werther's love, whereas the mentions of "sterben" and "Verblendung" are characteristic of the unrequited love, which leads Werther into his "disease unto death". As Werther and Albert's word cloud reveals, their relationship is dominated by the "Unruhe" that Werther feels through his adversary.

Additionally, the tool plots the development of the narrative (not bound to specific characters) based on the word fields, an example of which is shown on Figure 6. In this case we used words from the emotion domain (with emotion dictionaries by Klinger et al. (2016)).

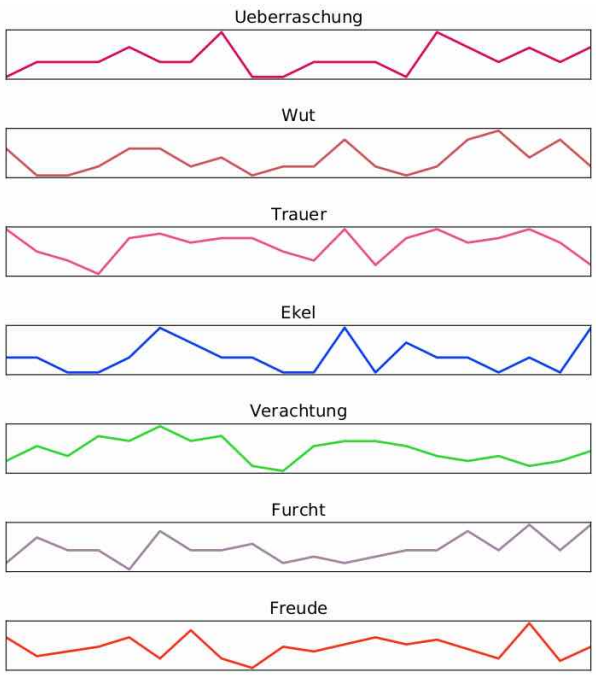


Illustration 6: Word field development for Goethe's Werther

The word field development can highlight the prevalence of individual emotion domains across the text. The accumulation of the negative emotion words (Wut,Trauer, Furcht) towards the end suggests, for example, that Goethe's novel has no "happy ending". The striking rash on "Freude", however, captures the last happy hours Werther spends with Lotte in the second part of the narration before he kills himself.

Future Work

The next version of the tool will include a character-oriented word field development calculated and plotted for the main characters of the stories. In addition, future releases will include more analysis features and bulk file processing.

Fußnoten

1. https://voyant-tools.org/
2. https://gephi.org/
3. *www.ims.uni-stuttgart.de/data/rcat*
4. http://flask.pocoo.org/

Bibliographie

Agarwal, A. / Corvalan, A. / Jensen, J. / Rambow, O. (2012): "Social Network Analysis of Alice in Wonderland", in: CLfL@ NAACL-HLT 88-96.

Agarwal, A. / Kotalwar, A. / Rambow, O. (2013): "Automatic Extraction of Social Networks from Literary Text. A Case Study on Alice in Wonderland", in: IJCNLP 1202-1208.

Beveridge, A. / Shan, J., (2016): "Network of thrones", in: Math Horizons, 23(4): 18-22.

Bondy, J.A. / Murty, U.S.R. (1976): Graph theory with applications (Vol. 290). London: Macmillan.

Burrows, J.F. (1987): "Word-patterns and story-shapes: The statistical analysis of narrative style", in: Literary & Linguistic Computing, 2(2): 61-70.

Elson, D.K. / Dames, N. / McKeown, K.R. (2010): "Extracting social networks from literary fiction", in: Proceedings of the 48th annual meeting of the association for computational linguistics 138-147. Association for Computational Linguistics.

Heuser, R., F. Moretti / E. Steiner (2016): The Emotions of London. Technical report. Stanford University. Pamphlets of the Stanford Literary Lab.

Jayannavar, P. / Agarwal, A. / Ju, M. and Rambow, O. (2015): "Validating Literary Theories Using Automatic Social Network Extraction", in CLfL@ NAACL-HLT 32-41.

Jockers, M.L. / Underwood, T. (2016): "Text-Mining the Humanities", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): A New Companion to Digital Humanities 291-306.

Kim, E. / Padó, S. / Klinger, R. (2017): "Investigating the Relationship between Literary Genres and Emotional Plot Development", in: Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 17-26.

Klinger, R. / Sulliya S.S. / Reiter N. (2016): "Automatic Emotion Detection for Quantitative Literary Studies -- A Case Study on Kafka's 'Das Schloss' and 'Amerika'", in: Digital Humanities (DH), Conference Abstracts, Kraków, Poland, 2016.

Michel, J.B. / Shen, Y.K. / Aiden, A.P. / Veres, A. / Gray, M.K. / Pickett, J.P. / Hoiberg, D. / Clancy, D. / Norvig, P. / Orwant, J. / Pinker, S. (2011): "Quantitative analysis of culture using millions of digitized books", in: science, 331(6014) 176-182.

Moretti, F. (2005): Graphs, maps, trees: abstract models for a literary history. Verso.

Moretti, F. (2011). Network theory, plot analysis. Stanford Literary Lab Pamphlet Series 2. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf

Murr, S. / Barth, F. (2017): Digital Analysis of the Literary Reception of J.W. v. Goethe's 'Die Leiden des jungen Werthers', in: Digital Humanities (DH), Conference Abstracts, Montreal, Canada 2017.

Nalisnick, E.T. / Baird, H.S. (2013): "Extracting sentiment networks from Shakespeare's plays", in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on IEEE 758-762.

Park, G.M. / Kim, S.H. / Cho, H.G. (2013): "Structural analysis on social network constructed from characters in literature texts", in: Journal of Computers, 8(9): 2442-2447.

Rydberg-Cox, J., (2011): "Social networks and the language of greek tragedy", in: Journal of the Chicago Colloquium on Digital Humanities and Computer Science (Vol. 1, No. 3).

West, D.B. (2001): Introduction to graph theory (Vol. 2). Upper Saddle River: Prentice Hall.

Auf der Suche nach der verlorenen Materialität. Kodikologie und Restaurierungswissenschaft im Zeitalter der (Massen-) Digitalisierung.

Busch, Hannah
buschh@uni-trier.de
Center for Digital Humanities, Universität Trier, Deutschland

Bös, Eva
ef_fa@web.de
Buchbinderei Mohr, Trier, Deutschland

Einleitung

Die Materialitäten von Handschriften und ihrer digitalen Abbilder bieten den Ausgangspunkt unseres Vortrags, der zwei Disziplinen in einen Dialog bringt, die sich mit der Materialität handgeschriebener Artefakte beschäftigen und an der Digitalisierung beteiligt sind: die digitale Kodikologie und die Restaurierungwissenschaft. Die handwerkliche Arbeit der Buchrestauratoren ist für Digitalisierungsprojekte unerlässlich und gehört zur Vor- und Nachbereitung: In der Regel geht jeder Kodex zunächst durch ihre Hände, da unter Umständen der Zustand des Objekts bewertet, Festigungsmaßnahmen oder andere Eingriffe vorgenommen werden müssen, um die Digitalisierung physisch zu ermöglichen. [1] Restauratoren sind aber mehr als bloße Dienstleister, sondern verfügen über einen speziellen Blick auf die Materialität von Originalen, der zusätzliche Informationen für digital basierte kodikologische Untersuchungen bereitstellen kann. Umgekehrt verfügen die Digital Humanities über Methoden, die für Anliegen der Restaurierung nützlich sein und gemeinsam weiterentwickelt werden könnten. Nur durch interdisziplinäre Ansätze, so die These, kann das Potenzial des Digitalisats erkannt und ausgeschöpft werden. Gleichzeitig führt die Synergie zur Kompetenzerweiterung der beteiligten Disziplinen.

Seit mehr als zwanzig Jahren werden sukzessive Handschriftenbestände digitalisiert. Zunächst beschränkte sich dieses Unternehmen auf besonders hervorzuhebende Bestände und Einzelstücke von erhöhtem öffentlichem Interesse. Es kann davon ausgegangen werden, dass diese Digitalisierungsmaßnahme in erster Linie der allgemeinen Bereitstellung und dem Schutz des Originals galt. Das digitale Faksimile ersetzt das Original im 'alltäglichen' Gebrauch und schont damit das materielle Objekt, zudem dient es zumindest als visuelle Sicherung im Falle eines Verlustes des Originals. Im Gegensatz zur Ausstellung im musealen Raum, in dem in der Regel nur ein auf eine Doppelseite begrenzter Einblick in das Kulturgut gegeben werden kann, bietet das Digitalisat die Möglichkeit, die Handschrift zu durchblättern. Durch Bereitstellung im World Wide Web wird zudem die örtliche und zeitliche Begrenzung aufgehoben und ein Zugang für Forscher weltweit ermöglicht. Der Einblick in die Handschrift kann ortsunabhängig nahezu unbegrenzt vielen Personen zur selben Zeit ermöglicht werden und auch beliebig oft, ohne Verschleißerscheinungen am Original und am Digitalisat zu hinterlassen. Seitdem die Anschaffung hochwertiger Scanner oder Konstruktionen, die mit digitaler Fotografie arbeiten [2], nicht nur für die großen Digitalisierungszentren erschwinglich geworden ist, werden auch immer mehr Gesamtbestände mittelalterlicher Bibliotheken digitalisiert und der Forschung sowie interessierten Personen zur Verfügung gestellt. Digitalisierung ermöglicht also, vergessene und/oder besonders fragile handschriftliche Kulturgüter wieder in den Fokus der öffentlichen Aufmerksamkeit zu bringen und bestandsübergreifende Forschung zu erleichtern.

Angesichts der gewaltigen Masse finanzieller und personeller Ressourcen, die Digitalisierungsprojekte innerhalb von Institutionen binden, erscheint das Nachdenken darüber, welche Bearbeitungsschritte und technischen Features in bestehende Workflows im Kontext der Digitalisierung eingebunden werden könnten, sowohl in wissenschaftsorientierter als auch in wirtschaftlicher Hinsicht lohnenswert. [3] Die Tatsache, dass im Rahmen von Digitalisierungen eine immense Zahl an Objekten einzeln in die Hand genommen und Seite für Seite durchblättert wird, birgt ein bislang noch nicht ausgeschöpftes Potential für interdisziplinäre Anschlussmöglichkeiten.

Gegen eine Theoretisierung der Materialität [4] – Materialität von analogem Original und digitalem Faksimile

Dem digitalisierten Objekt wird häufig seine Materialität abgesprochen; darüber zu diskutieren, dass das digitale Faksimile einer mittelalterlichen Handschrift nicht die gleiche synästhetische Erfahrung bieten kann wie das Original, ist obsolet. Es ist wahr, dass das digitale Faksimile zunächst nicht dazu geeignet scheint physische Eigenschaften, wie Lagenstrukturen, Linierung, den Erhaltungszustand oder Eigenschaften der Bindung zu untersuchen (vgl. Pierazzo 2016:94). Dass dennoch gerade eHumanities-Projekte, die sich im Feld der digitalen Kodikologie angesiedelt haben, materielle Eigenschaften von Handschriften ins Zentrum zu rücken verstehen, zeigen etwa das groß angelegte Projekt der Wasserzeichendatenbank , die Untersuchungen des Archimedes Palimpsest , VisColl zur Visualisierung von Lagenstrukturen oder eCodicology .
Das Potenzial des Digitalisats liegt im Material des Originals. Restaurierungswissenschaft kann Materialaspekte bestimmen; die Digital Humanities computergestützte Techniken bereitstellen, um diese Informationen zu erfassen, zu dokumentieren und so aufzubereiten, dass die Daten z.B. miteinander verglichen, auf unterschiedliche Weise systematisiert, visualisiert und eine Nachnutzung ermöglicht werden kann.
Der Beitrag möchte sich deshalb weniger der allgemeinen Defintion von Materialität widmen, sondern bisherige Digitalisierungspraktiken hinterfragen. Dabei steht immer die Frage im Hintergrund, welche Synergieeffekte sich durch den Austausch der Disziplinen ergeben, die in die Digitalisierungsabläufe involviert sind. [5]

Gemeinsame Aufgaben, gemeinsame Ziele, gemeinsame Probleme

Wirft man einen Blick auf die Unterfangen der Digitalisierung und der Restaurierung fällt auf, dass sie die Kernaufgabe, Kulturgut zu erhalten und zugleich der Forschung und dem allgemeinen Nutzen zugänglich zu machen, gemeinsam haben. Sie teilen aber auch die mit diesen gemeinsamen Aufgaben und Zielen einhergehenden Probleme der Finanzierung der Maßnahmen, das ungeklärte und teils unabsehbare Problem der Langzeiterhaltung und den Kampf um die nötige Aufmerksamkeit auf institutioneller und politischer Ebene. [6] Sammlungen und Einzelexemplare mit einer hohen Nutzungsintensität stehen in einem erhöhten Interesse der Zugänglichmachung und werden mit einer höheren Priorität der Bestandserhaltung zugeführt, während andere Handschriften ihr einsames und weitgehend ungeachtetes Dasein in den Magazinen von Bibliotheken und Magazinen fristen. Durch die Motivation, vollständige Handschriftenbestände zu digitalisieren – beispielsweise um dislozierte Bestände digital wieder zusammenzuführen – rücken jedoch inzwischen auch bisher wenig beachtete Bestände in den Fokus und erhalten den ‚Luxus' restauratorischer Begutachtung und Behandlung sowie der prioritären Bestandserhaltung mittels Digitalisierung . [7]Geht man einen Schritt weiter und richtet seinen Fokus auf die daraus resultierenden neuen Möglichkeiten für die Forschungsdisziplinen Kodikologie und Restaurierungswissenschaft, lassen sich weitere Gemeinsamkeiten erkennen: beide bedienen sich derselben Techniken – der Bildverarbeitung und Mustererkennung – um neue Erkenntnisse und Arbeitsabläufe zu kreieren (vgl. Weber / Hähner 2014: 139ff und Busch / Chandna 2017), beide stehen vor den selben Herausforderungen, wie dem Mengenbetrieb sowie der daraus resultierenden Verwaltung großer Datenmengen.

Aspekte des Einzelobjekts/-digitalisats

Hinsichtlich der authentischen Abbildung originaler Materialität im Digitalisat stellt sich immer auch die Frage nach dem ‚digitalen Erhaltungszustand'. Informationsverluste bei der Digitalisierung betreffen v.a. Farbigkeiten und Oberflächenstrukturen, letztere sind in herkömmlichen

Auflichtaufnahmen kaum wahrnehmbar, sodass Tiefendimensionen eingeebnet sind.
Die Restaurierung arbeitet im Rahmen von Dokumentationen mit Streiflicht und Durchlicht, um Unebenheiten, Knicke, Risse, Trockenstempel, Linierungen etc. sichtbar zu machen. Könnten solche Aufnahmetechniken in Digitalisierungsprozesse einbezogen werden, um kodikologische Merkmale im Digitalisat zu bewahren? Und wie können die Resultate von restauratorischen Dokumentationen erfasst und beispielsweise in die Metadaten digitalisierter Sammlungen einfließen?
Verschiedene Belichtungsarten (UV, IR etc.) eröffnen auch Zugänge zu nicht nur ‚einer Materialität', z.B. können durch die unterschiedliche Reichweite der Strahlungen frühere materielle Zustände – z.B. gelöschte oder überdeckte Schriftzeichen – rekonstruiert, sichtbar und darstellbar gemacht werden, vgl. Archimedes Palimpsest oder die Weimarer lösch- und brandgeschädigte Notenhandschriften (Weber / Hähner 2014: 140ff.).

Aspekte der Objekt-/Datenmengen

Die authentische Wiedergabe kodikologischer Merkmale ist nicht zuletzt für Methoden der digitalen quantitativen Kodikologie relevant (vgl. Chandna et al. 2015). Untersuchungen zur Miseen-page etwa (bspw. Verhältnis von beschriebenem und unbeschriebenem Raum einer Handschriftenseite) würden zu realistischeren Ergebnissen führen, wenn mit einbezogen werden könnte, ob eine Handschrift einer Trockenreinigung, Bleiche o.ä. unterzogen worden ist, da dies Einfluss auf den Ist-Zustand des Textträgers hat. Durch die durchgeführten Maßnahmen weichen Farbwerte von Handschriften gleicher Provenienz, gleichen Alters und gleicher Materialität voneinander ab; der Kontrast von Schrift zu Hintergrund verändert sich. Durch die Festlegung eines Koeffizienten zur Korrektur dieser Abweichung könnte dem entgegengewirkt und eine größere Vergleichbarkeit gewährleistet werden. Die Dokumentation von Restaurierungsmaßnahmen könnte ein solches Verfahren ermöglichen.
Andersherum könnten Techniken der Datenverarbeitung aus den DH genutzt werden, um restaurierungsrelevante Daten wie Schadensbilder, Zustandsbeschreibungen oder objektbezogene Informationen zum Einband etc. zu erfassen, die etwa für die Vorbereitung von großangelegten Restaurierungsmaßnahmen genutzt werden könnten.
In Prozessen der Bestandserhaltung und Restaurierung von Schriftgut – wie sie zum Beispiel vor der Digitalisierung oder nach Notfällen wie Bränden und Archiveinstürzen durchgeführt werden – stehen die materiellen Aspekte der Objekte bei der Schadensdokumentation im Mittelpunkt und werden in schriftlichen Schemata festgehalten; die Inhalte der Dokumentationen werden aber nur teilweise und eher selten in weiteren Projekten weitergenutzt oder virtuell zur Verfügung gestellt. In der HAAB Weimar wird an einer digitalen Form des dort analog verwendeten Dokumentationsschemas gearbeitet; im Bibliothekskatalog sind für die beim Brand geschädigten Exemplare Felder zur Materialbeschreibung und der Art des Schadens vorgesehen. [8] Umgekehrt werden bei der Digitalisierung Materialdaten nicht erfasst, die sowohl für die Forschung als auch für die Bestandserhaltung (Langzeitarchivierung und analoger Objekterhalt) von Bedeutung sein könnten.
Mit dem hier skizzierten Beitrag soll nicht nur eine Kritik an der bisherigen Digitalisierungspraxis ohne ausreichende Einbeziehung der Restaurierungswissenschaften geübt werden, sondern vielmehr das mögliche Potenzial der interdisziplinären Zusammenarbeit bei zukünftigen Initiativen an konkreten Anwendungsbeispielen aufzeigen. Darüber hinaus erhoffen wir uns, das Thema interdisziplinärer Synergieeffekte hinsichtlich der Erfassung von Materialität in Digitalisaten ins Gespräch zu bringen und Anregungen zu sammeln, inwiefern sich der vorgestellte Ansatz in Zukunft in der Praxis umsetzen lässt.

Fußnoten

1. Bsp. der Steppkapitale der neugebundenen Mattheiser Handschriften in Trier, die vor der Digitalisierung aufgelöst und hinterher wieder erneuert werden mussten, um den nötigen Aufschlagwinkel zu erreichen.
2. Den größten Bekanntheitsgrad haben hier vermutlich der Grazer Büchertisch und der Wolfenbütteler Buchspiegel, aber auch Aufsichtsscanner liefern digitale Aufnahmen die hohen Qualitätsansprüchen gerecht werden.
3. Nicht zuletzt entsteht so auch ein Instrument für die Institutionen, ihre Relevanz für die Öffentlichkeit nachzuweisen (z.B. durch die Zahl der Zugriffe auf die digitalisierten Bestände). In Zeiten, in denen eine Vielzahl kultureller Einrichtungen um verhältnismäßig wenige öffentliche finanzielle Mittel buhlen muss und Evaluierung einen Teil des Arbeitsalltags ausmacht, ist es unerlässlich geworden, ein öffentliches Inter-

esse der eigenen Tätigkeit nachweisen zu können, um mit einer Zuwendung bedacht zu werden (s. bspw. Knoche 2017).
4. Welchen Nutzen birgt das Digitalisat, abgesehen von der Bereitstellung und Zugänglichmachung von Texten inklusive eines Eindrucks des originalen Textträgers?
5. Welche Erkenntnisse kann uns die Materialität liefern? Wir können wir sie digital Erfassen und Präsentieren? Welchen Verlust der Materialität im Digitalen haben wir tatsächlich zu verzeichnen?
6. Siehe hierzu Initiativen wie den Berliner Appell aus dem Jahr 2013 und dem Weimarer Appell von 2014.
7. An dieser Stelle sei auf den im Dezember 2015 veröffentlichten Masterplan zur „Digitalisierung mittelalterlicher Handschriften in deutschen Bibliotheken“ verwiesen.
8. Die Verlust- und Schadensdokumentation der HAAB Weimar ist unter https://lhwei.gbv.de/DB=2.2/ einsehbar [letzter Zugriff 24. September 2017].

Bibliographie

Berliner Appell zum Erhalt des digitalen Kulturerbes, September 2013. http://www.berliner-appell.org [letzter Zugriff am 25. September 2017].

Busch, Hannah / Chandna, Swati (2017): “eCodicology: The Computer and the Mediaeval Library” in: Busch, Hannah / Fischer, Franz / Sahle, Patrick (Hrsg.): *Kodikologie und Paläographie im digitalen Zeitalter 4.* Norderstedt: BoD. urn:nbn:de:hbz:38-77742

Chandna, Swati / Tonne, Danah / Jejkal, Thomas / Stotzka, Rainer / Krause, Celia / Vanscheidt, Philipp / Busch, Hannah / Prabhune, Ajinkya (2015): “Software Workflow for the Automatic Tagging of Mediaeval Manuscript Images (SWATI)” in: Ringger, Eric K. / Bart Lamiroy (eds.): *SPIE Proceedings 9402. Document Recognition and Retrieval XXII*. San Francisco, February 11-12 2015.

Knoche, Michael (2017): „Rettung von Bücherschätzen. In guter Ordnung, aber schlechter Verfassung“ in: F.A.Z. vom 18.07.2017. http://www.faz.net/-gqz-8zv5x [letzter Zugriff 25. September 2017].

Pierazzo, Elena (2015): *Digital Scholarly Editing. Theories, Models and Methods.* London: Routledge.

Schreiber, Carolin / Fabian, Claudia (Red.) (2015): *Digitalisierung mittelalterlicher Handschriften in deutschen Bibliotheken.* Masterplan. München. http://www.handschriftenzentren.de/wp-content/uploads/2016/06/Priorisierungsfragen-Masterplan_pub.pdf

Weber, Jürgen / Hähner, Ulrike (Hrsg.) (2014): *Restaurieren nach dem Brand. Die Rettung der Bücher der Herzogin Anna Amalia Bibliothek*. Petersberg: Michael Imhof Verlag.

Weimarer Appell zur Erhaltung schriftlichen Kulturguts, September 2014.

Bildanalyse durch Distant Viewing - zur Identifizierung von klassizistischem Mobiliar in Interieurdarstellungen.

Donig, Simon
simon.donig@uni-passau.de
Universität Passau, Deutschland

Christoforaki, Maria
Maria.Christoforaki@Uni-Passau.De
Universität Passau, Deutschland

Bermeitinger, Bernhard
Bernhard.Bermeitinger@uni-passau.de
Universität Passau, Deutschland

Handschuh, Siegfried
Siegfried.Handschuh@uni-passau.de
Universität Passau, Deutschland

In den vergangenen Jahren haben digitale Forschungsinstrumente in Kunst-, Architektur- und Designgeschichte sowie den Material-Culture Studies an Bedeutung gewonnen (Berry 2017; Klinke 2016; Auslander 2005). Wie viele disruptive Technologien verändern neue Techniken im Bereich der Computer Vision (Bell & Ommer 2015; dies. 2016) die Arbeitsweise unsere Disziplinen.

Durch unsere Arbeit am Neoclassica-Framework (Donig et al. 2017a) möchten wir Forschenden solche neue Instrumente und Methoden für die Analyse und Klassifizierung materialer Kultur, konstruktiver Merkmale und ästhetischer Formen des Klassizismus an die Hand geben. In unserer Forschung konzentrieren wir uns dabei zunächst auf Raumkunst (insbesondere Mobiliar

und Innenausstattung) sowie Architektur und deren jeweilige visuelle Darstellungen.

Die Klassifizierung von einzelnen Artefakten und ihre Identifizierung in Raumdarstellungen bildet deshalb einen nicht unwichtigen Meilenstein für das Projekt als Ganzes.

In dem hier vorliegenden Beitrag beschreiben und reflektieren wir einen Zugang zur Klassifizierung von Objekten in Einzeldarstellungen bzw. zu ihrer Identifikation in Raumdarstellungen, aufbauend auf unseren Experimenten (Bermeitinger et al. 2017) und Analysen (Donig et al. 2017b) zur automatisierten Klassifizierung von materialer Kultur des Klassizismus mit *Tiefem Lernen* (Deep Learning) – hier konkret Faltenden Neuronalen Netzen (Convolutional Neural Networks, CNN) (Krizhevsky et al. 2012).

Interieuranalyse durch Distant Viewing – theoretische Überlegungen

Unter *Distant Viewing* verstehen wir in Analogie zur Notion des *Distant Reading* (Moretti 2013) eine wissenschaftliche Methode zur technischen Überbrückung sowohl zeitlicher wie räumlicher Distanz als auch Verbreiterung der Menge betrachtbarer Bilder. In diesem Sinn ermöglicht der ferne Blick als wissenschaftliche Methode es uns, ein Verständnis für die tatsächlich oder auch nur kontemporär imaginierte Beschaffenheit von vergangenen Räumen zu entwickeln.

Da Bildquellen als selbstreferentielle Systeme mit einem spezifischen Eigensinn ausgestattet sind, bedürfen sie einer besonderen methoden- und quellenkritischen Durchdringung, denn selbst Bildwerke, die ihrem Anspruch nach einen dokumentierenden Charakter haben, bleiben ästhetischen Zwängen des Mediums unterworfen (verwiesen sei etwa auf die kritische vergleichende Analyse der Raumwirkung von Aquarellen aus dem sogenannten Wittelsbacher Album durch (Langenholt 2002: 47–49)).

Durch die Ausweitung kultureller Produktion und Konsumption – sowohl von Texten wie von Bildern – an der Epochenschwelle 1800 stehen für Forschungen in diesem Bereich reichhaltige Quellenbestände zur Verfügung. Diese reichen von eher typisierenden Raumdarstellungen (wie etwa in Karikaturen oder häufig auch zeitgenössischen Darstellungen des Alltags unterbürgerlicher Schichten) über ihrem Bewusstsein nach eher historisch-dokumentierende Ansätze (beispielsweise Alben mit Raumansichten, wie sie in den Oberschichten etwa als Hochzeitsgeschenke für “ausheiratende” weibliche Familienmitglieder beliebt waren) bis hin zu visionären Raum- und Werkstücksentwürfen, die einem konsumierbaren Erzeugnis in ihrer Antikenrezeption zugleich auch eine gesellschaftliche Vision einschrieben (Pawlitzki, Bruer, Kunze 2009); (Kepetzis 2006); (Auslander 1996).

Auch wenn in der Forschungspraxis beide Ansätze stets aufeinander bezogen bleiben müssen, kann man idealtypisch zwischen einem Ansatz unterscheiden, der das Bildwerk vor allem im Hinblick auf seine Produktionsbedingungen und Produzenten befragt, und einem Ansatz, der das Bild als Quelle des Dargestellten analysiert.

Betrachtet man Bildwerke etwa als Zeugnisse vergangener Alltagskultur und ihrer Praktiken, wird es durch den hier besprochenen Zugang beispielsweise möglich, Veränderungen in der Ausstattung und -gestaltung einander ähnlicher Räume über längere Zeit zu verfolgen. Wir können so zugleich Cluster und Typen von Räumen aufgrund der Darstellung ihrer Beschaffenheit bilden, die es wiederum erlauben, tradierte Funktionszuschreibungen von Räumen kritisch zu hinterfragen.

Stellt man dagegen das Dargestellte als Repräsentationsform von Geschmack in den Mittelpunkt, wird es in ähnlicher Weise möglich, Artefakte mit anderen digitalisierten Korpora abzugleichen. Unabhängig davon, ob das dargestellte Objekt jemals zur Ausführung gelangt ist, kann unser Zugang so beispielsweise für die Rezeptionsforschung den Transfer spezifischer Darstellungsweisen durch Medien wie etwa die *Collection de Meubles et Objets de Goût* (La Mésangère 1761–1831) oder Musterbücher neu zugänglich machen.

Instrumentenentwicklung

Objekterkennung in Einzeldarstellungen

Datengewinnung, Aufbereitung und Säuberung

Ein erster wichtiger Meilenstein im Rahmen unseres Forschungsprogramms war die automatisierte Erkennung von Objekten in Einzeldarstellungen. Um zeitnah ein großes Trainingskorpus zu generieren, haben wir zunächst Abbildungen von materieller Kultur des Klassizismus – hauptsächlich Möbel und Kleinkunst (darunter etwa Bronzen, Silberarbeiten etc.) – aus den Beständen des Metropolitan Museum of Art gescrapt, welches diese Bildwerke Anfang 2017 als Public Domain zugänglich gemacht hat (The Metropolitan

Museum of Art 2017). Dadurch ist es möglich, die Trainings- und Testdaten an interessierte Dritte weiterzugeben, was die Reproduzierbarkeit des Experiments sicherstellt.

Da ein Faltendes Neuronales Netz Bilder als Ganzes klassifiziert, haben wir zunächst sichergestellt, dass jedes Bild lediglich ein Objekt in der Totalen zeigt. Dazu wurden alle Darstellungen von Möbelmerkmalen wie Nahaufnahmen ausgeschlossen sowie alle Bildwerke, die Interieurs, Ensembles und Möbel à la suite zeigen, derart aufgespalten, dass auf jeder Abbildung nur noch ein Möbelstück zu sehen ist. Diese Darstellungen haben wir gegebenenfalls so bearbeitet, dass noch sichtbare Teile benachbarter Objekte mit einer homogenen Farbe abgedeckt wurden. Im Endergebnis entstand so ein Korpus von 1.246 Bildern, der 379 Artefakte umfasst. Diese wurden gemäß der Neoclassica-Ontologie (Donig et al. 2016) annotiert.

Ergebnis

Zunächst haben wir mit der Standardimplementierung des VGG19-Layouts, das an einem Subset von 1.000 Klassen von ImageNet (Deng et al. 2009) vortrainiert wurde, dieses Korpus prozessiert. Dieses Verfahren resultierte in einer durchschnittlichen Genauigkeit von 0,82 und einem durchschnittlichen F1-Wert von 0,62. Wir wiederholten das Experiment 21 mal mit Aufteilungen des Korpus in verschiedene Trainings- und Testsets in einem 80/20 Verhältnis pro Klasse, wobei wir für jeden Durchlauf sehr ähnliche Ergebnisse erhalten haben (für weitere Details vgl. Bermeitinger et al 2017).

Überprüfung des Befunds an zeitgenössischen Möbelzeichnungen

Da das Korpus in seiner überwältigenden Mehrheit aus Fotografien des 20. Jahrhunderts besteht, beschlossen wir, durch ein Kontrollkorpus, das auf einer Kompilation von Thomas Sheratons Möbelzeichnungen beruht (Sheraton/Munro 1910), nachzuprüfen, ob zeitgenössische Grafiken (bzw. deren Reproduktionen) vergleichbar gute Resultate liefern können. Wiederum wurden alle Blätter, die mehrere Objekte zugleich zeigen, so aufgesplittet, dass daraus Einzeldarstellungen wurden.

Dieses relativ begrenzte Korpus von 64 Abbildungen wurde mit einer durchschnittlichen Genauigkeit von 0,63 beziehungsweise 0,78 in den Top-2 und 0,84 in den Top-3 Klassen erkannt. Wir schließen daraus, dass zeitgenössische Grafiken nicht alleine mit einer ähnlich großen Genauigkeit klassifiziert werden können wie moderne Fotografien, sondern dass damit vor allem auch der von uns gewählte Ansatz prinzipiell geeignet ist, um historische Interieurdarstellungen auszuwerten.

Artefakte in Interieurszenen

Ein Artefakt in einer Interieurszene automatisiert zu klassifizieren, stellt eine besondere Herausforderung dar. Idealtypisch muss dazu erstens eine interessante Region im Bild identifiziert und zweitens muss diese anschließend korrekt klassifiziert werden.

Ein derzeit verbreitetes Werkzeug für derartige Identifikations- bzw. Klassifizierungsaufgaben sind *Regionale Faltende Neuronale Netze* (Regional Convolutional Neural Network, RCNN) (Girshick et al. 2014). Für diese Aufgabe haben wir uns für das vergleichsweise aufwandsarm zu nutzende Objekterkennungsmodul (Huang et al. 2016) des Tensorflow-Frameworks (Abadi et al. 2016) entschieden, das dem State-of-the-Art im Bereich optischer Merkmalserkennung und -klassifizierung entspricht. Dabei verwenden wir die in diesem Modul verfügbare Implementierung *Faster-RCNN with ResNet101*, die für uns den besten Kompromiss zwischen Trainingsgeschwindigkeit und Performanz darstellt.

Für das Training des Netzes verwendeten wir wiederum das Korpus des Metropolitan Museum of Art, das zu diesem Zweck neu mit Polygonen annotiert wurde. Für die Annotation wurde der PyLabelMe Editor (PyLabelMe, 2011) eingesetzt. Dieses Annotationsinstrument kann ein Bild durch die Verknüpfung von Polygonflächen im Bild mit Labels annotieren, die dann als json-Dateien gespeichert werden.

Für das Experiment stellten wir eine Untermenge aus dem augmentierten MET-Korpus zusammen, die alle Klassen mit mindestens zwölf Bildern umfasst und damit 29 Klassen repräsentiert, welche insgesamt 618 Bildern entsprechen.

Wiederum folgten wir einem 80/20 Verhältnis bei der automatisierten Aufteilung der Klassen in Trainings- und Testset, was in dem ausgezeichneten arithmetischen Mittel der Präzision (aMP) von 0,94 resultierte.

Um dieses Ergebnis an unabhängigen Daten zu verifizieren, haben wir fünf Kontrollkorpora zusammengestellt. Diese umfassen erstens Darstellungen von Einzelobjekten, die relativ separiert vom Hintergrund sind, unterschieden in Fotografien, Druckgrafiken sowie (teils kolorierte) Zeichnungen. Zweitens Interieurs, die wir in Fotografien kontemporärer “Period Rooms” aus Museen bzw. von Ausstellungsflächen sowie schließlich

historische Raumansichten (Aquarell-, Gouache- und Ölmalerei, teils Lithografien) unterschieden haben.

Die Fotografien entstammen den Beständen des Victoria & Albert Museum und der Wallace Collection. Für historische Druckgrafiken legten wir neben dem Sheraton-Set ein weiteres Set basierend auf einer Neuauflage von Thomas Hepplewhites “The cabinet maker and upholsterer's guide” (Hepplewhite 1790) an.

Für die kolorierten Zeichnungen griffen wir auf eine digitale Reproduktion des sogenannten Bellange Albums zurück, das der Werkstatt von Pierre-Antoine Bellange (1757–1827) und seinem Sohn Louis-Alexandre (1797–1861) sowie mehreren anderen Künstlern zugeschrieben wird. (Für eine detaillierte Analyse des Albums (MET, asc. nr. 51.624.2) vgl. Cordier 2012).

Während die zeitgenössischen Fotos von Period Rooms aus den Beständen des MET stammen, haben wir die historischen Raumansichten vorwiegend der Thaw Collection des Cooper Hewitt, Smithsonian Design Museum entnommen.

Lediglich die Einzeldarstellungen von Objekten liegen dabei in annotierter Form vor, was uns erlaubt, numerische Qualitätskriterien für den Erfolg unserer Experimente zu benennen. Für die nicht annotierten Interieurs erfolgte eine visuelle Kontrolle der Befunde.

Ergebnisse und Analyse: Einzelobjekte

Die unten stehende Tabelle zeigt die Resultate des Klassifikationsexperiments für die Einzelobjekte.

Institution	Quellentyp	arith. Mittel der Präzision (aMP)	Zahl der genutzten Abbildungen
Victoria & Albert Museum	Primär Fotografien	0,60	371
Wallace Collection	Primär Fotografien	0,69	61
Sheraton	Druckgrafik	0,48	89
Hepplewhite	Druckgrafik	0,50	75
Bellange-Album	kolorierte Zeichnungen	0,64	24

Betrachtet man das arithmetische Mittel der Präzision, zeigen sich durchgängig gute Werte, die im wesentlichen denen des vorausgegangenen Experiments mit dem CNN ähneln und die bei den zweifarbigen Druckgrafiken am niedrigsten ausfallen. Dieser Effekt ist nicht alleine durch das Medium bzw. die Technik bedingt, sondern hängt auch mit der Annotationspraxis zusammen. Nicht direkt aus dem numerischen Qualitätskriterium ersichtlich ist etwa, dass zahlreiche inkorrekte Zuschreibungen innerhalb desselben konzeptionellen Hierarchie-Baums bleiben. Neben tatsächlich falschen Zuordnungen finden sich so auch nachvollziehbare “Fehler” wie die richtige Zuordnung von Artefakten mit falscher Ausgangsklassifikation (Abb. 01), sinngemäß richtige Zuordnungen wie die Klassifizierung eines im Restaurierungszustand ohne Polsterung abgebildeten Fauteuils durch das übergeordnete Konzept des Armlehnstuhls (Abb. 03), oder Verwechslungen zwischen benachbarten, visuell ähnlichen Klassen (ein Bücherschrank mit einem Schreibkabinett à abattant mit einem Bücherschrank mit Zylinderkompartement (Abb. 02)).

Ergebnisse und Analyse: Interieurdarstellungen

Unsere Ausgangshypothese war, dass das an den Einzelobjekten trainierte RCNN in der Lage sein müsste, auch Objekte, die sich in Interieurs befinden, korrekt zu identifizieren und anschließend zu klassifizieren.

Während die Hypothese generell als bestätigt betrachtet werden kann, können wir aus unseren Experimenten eine Reihe von Beobachtungen ableiten, die uns erlauben, den unterschiedlichen Grad von Erfolg und bleibende Herausforderungen zu benennen, die wir in unserer zukünftigen Forschung angehen möchten.

Generell lässt sich festhalten, dass sich die Aufgabe der Identifizierung für den von uns trainierten Klassifikator als deutlich komplexer erwiesen hat, als die der Klassifizierung, wie die guten bis sehr guten Werte für das aMP zeigen. Für eine erfolgreiche Klassifizierung scheint primär die Zahl und visuelle Ähnlichkeit der Objekte innerhalb einer Klasse entscheidend zu sein, wie aus den hervorragenden Werten der zahlenstärksten Klassen wie etwa Stühlen ersichtlich ist.

Die zentrale Herausforderung unserer Experimente lag vor allem in der Identifikation des Artefakts im Interieur. Wir gehen davon aus, dass die Qualität dieser Erkennung primär von drei Faktoren beeinflusst wird:

Erstens dem Grad der Separation von Objekt und Bildhintergrund. Fotos, die etwa in Ausstellungsflächen entstanden sind und eine starke Separation zwischen Vordergrund und Hintergrund aufweisen, erzielen im Schnitt sehr viel bessere Resultate als andere Raumdarstellungen, die durch perspektivische Verzerrungen, die Überlappung von Artefakten und oft kontrastarme

bzw. übertrieben kontrastreiche Lichtgestaltung geprägt sind (Abb. 05, 06, 07).

Zweitens der Zahl der Objekte pro Klasse und der Bandbreite von Betrachtungswinkeln, mit denen trainiert wurde. Mehrfachidentifikationen desselben Objekts zeigen deutlich, dass Klassen mit hohen Objektzahlen (z.B. Stuhl) bei der Erkennung einen höheren Vertrauensgrad aufweisen als Klassen mit niedrigen Objektzahlen (z.B. Klismosstuhl) (Abb. 07). Dadurch kann es vorkommen, dass in manchen Ansichten nur wenige Artefakte mit hohem Vertrauensgrad erkannt werden, während eine große Zahl anderer im Rauschen zahlreicher "false positives" untergeht, die einen ähnlichen, niedrigen Vertrauensgrad aufweisen (Abb. 09, 04b).

Drittens der Materialität, Modalität und der Technik der Ausführung. Während der Klassifikator mit Fotografien durchweg gute Ergebnisse erzielt, schwankt die Qualität der Objekterkennung in historischem Material sehr. Unterschiede im Erfolg der Identifizierung von Objekten in verschiedenartigen Bildern führen wir primär auf die geringe Anzahl von überhaupt verfügbaren Darstellungen für das Training zurück. Sobald historische Darstellungen dem Gros des Trainingsmaterials stärker ähneln, steigt die Zahl der identifizierten Objekte, etwa bei naturalistischen Ölgemälden (Abb. 04a) oder Einzelblättern mit Designs (Abb. 08). Besonders deutlich wird dieser Umstand etwa durch den unterschiedlichen Vertrauensgrad, den dasselbe Motiv zwischen einer Ausführung als Gravur und dem ihr zugrunde liegenden Ölgemälde erreicht (Abb. 04a / 04b).

Zusammenfassung und Ausblick

Unser Beitrag zeigt eine Reihe von Experimenten zur Identifizierung und Klassifizierung von separierten Einzelobjekten wie auch von Objekten in Interieurdarstellungen mittels Deep Learning. Die Ergebnisse dieser Experimente reichen von exzeptionell (im Fall separierter Einzelobjekte) bis hin zu weniger befriedigend (für historische Darstellungen komplex gestalteter Räume, die in spezifischen Materialien, Modi und Techniken zur Ausführung kommen).

Die Befunde sind wichtige Schritte auf dem Weg zu einer produktiven Instrumentenkritik, die sowohl interne Limitierungen – etwa der black-box Charakter des Instruments, der die Gründe für eine spezifische Klassifizierung nur begrenzt nachvollziehbar macht – als auch die Bedeutung der Kuratierung und Augmentierung der Trainings- und Kontrollkorpora unterstreicht. Methodologisch herausfordernd ist besonders das Fehlen von Trainingsmaterial, das zum einen durch die Natur des Korpus zustande kommt (dies betrifft sowohl das Vorkommen von Artefakten in der historischen Lebenswelt als auch die museale Kuratierungspraxis), aber auch durch die zögerliche Praxis vieler Kulturinstitutionen, Bildmaterial unter permissiven Lizenzen freizugeben.

Trotz dieser Limitierungen haben wir mit diesem Beitrag gezeigt, dass es prinzipiell möglich ist, digitale Werkzeuge aus dem Bereich der Bilddatenforschung mit Big Image Data in die Geistes- und Kulturwissenschaften zu übertragen. Unseres Wissens stellt unser Experiment den ersten Versuch da, ein Korpus aus einer spezifischen Epoche, das auf geteilten ästhetischen Formen beruht und das miteinander korrespondierende Artefakte in historischen Darstellungen von komplexen Räumen einschließt, mittels Deep Learning zu klassifizieren.

Abbildungen

Abb. 1: Sheraton/Munro (1910), S.1

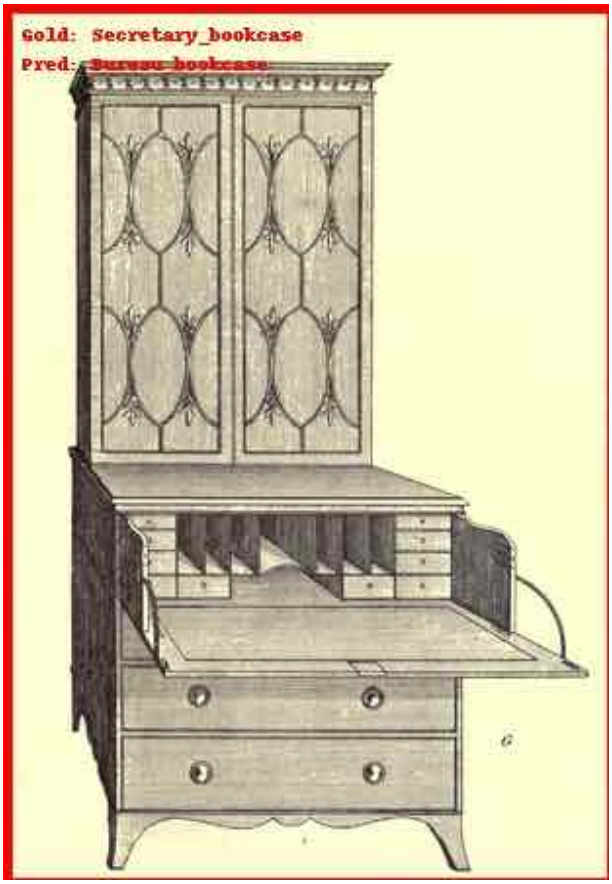


Abb. 2: Hepplewhite ((1)1790/(3)1897), plate 44

Abb. 3: The Wallace Collection, Asc. nr. F226

Abb. 4a: baron François Gérard (French, Rome 1770–1837 Paris): Charles Maurice de Talleyrand Périgord (1754–1838), Prince de Bénévent. Oil on canvas, Paris 1808. The Wrightsman Collection, Metropolitan Museum of Art, Accession Number: 2012.348,

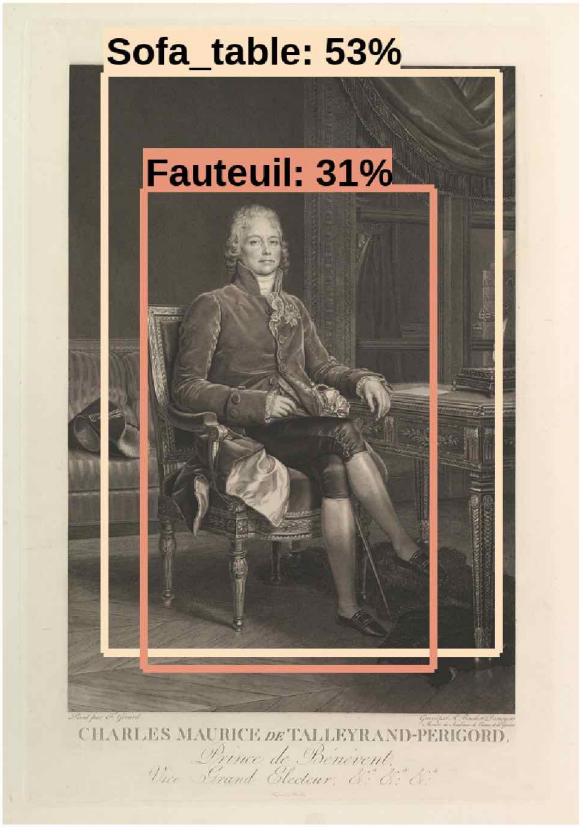


Abb. 4b:Auguste Gaspard Louis Boucher Desnoyers (French, Paris 1779–1857 Paris) after baron François Gérard (French, Rome 1770–1837 Paris) (after 1808): Portrait of Charles Maurice de Talleyrand-Périgord. Engraving with etching; third state of three. Metropolitan Museum of Art Accession Number:24.63.1051

Abb. 5: MET Objekte mit den Inventarnummern: 1986.449 (rot links), 63.143 (gelb), 1994.189 (mit rotem Kissen), 1984.126 (grün)

Abb. 6: Parlor from the William C. Williams House by Theophilus Nash (1810–11), Accession Number: 68.137 , https://www.metmuseum.org/art/collection/search/3411

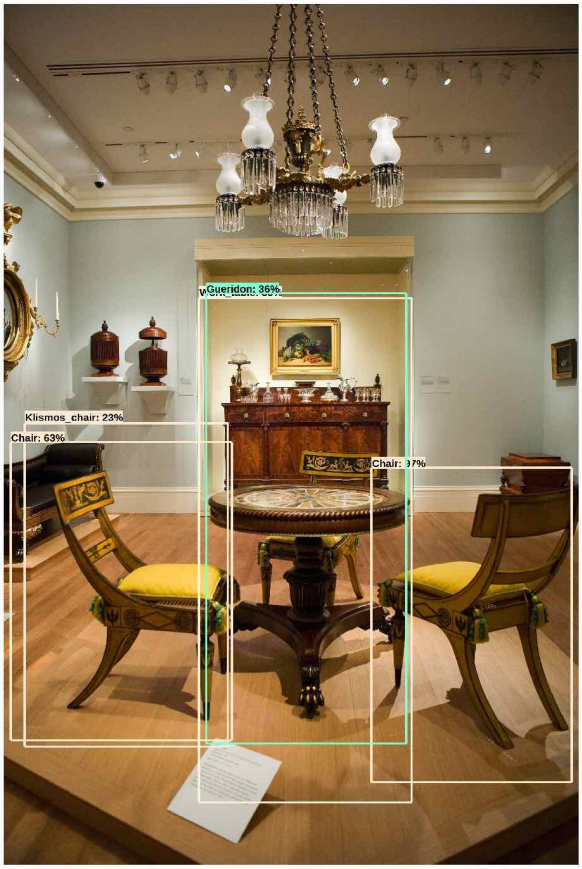


Abb. 7: MET Objekte mit den Inventarnummern: 68.96 (Gueridon), 65.167.5, 65.167.6, 65.167.8 (Klismosstühle)

Abb. 8: Drawing, Designs for Three Chairs, 1790; Attributed to Jean Démosthène Dugourc (French, 1749–1825); Germany; brush and watercolor, pen and black ink, graphite on white laid paper; Cooper Hewitt/Smithsonian: Thaw Collection. Accession Number 1921-6-136 , Object ID 18218673 , Short URL http://cprhw.tt/o/2BnM4/

Abb. 9: L. Lely: Drawing, A Salon in the Palazzo Satriano, Naples, 1829; brush and watercolor over graphite on white wove paper; Cooper Hewitt/Smithsonian: Thaw Collection; 2007-27-9; Accession Number 2007-27-9 , Object ID 18708105 , Short URL http://cprhw.tt/o/2DTgx/

Webseiten

Victoria and Albert Museum (https://www.vam.ac.uk/)

Wallace Collection (http://www.wallacecollection.org/)

Cooper Hewitt, Smithsonian Design Museum (https://www.cooperhewitt.org/)

The Neoclassica Project (*http://www.neoclassica.network/*)

Bibliographie

Abadi, Martin / Barham, Paul / Chen, Jianmin / Chen, Zhifeng / Davis, Andy / Dean, Jeffrey / Devin, Matthieu / Ghemawat, Sanjay / Irving, Geoffrey / Isard, Michael / Kudlur, Manjunath / Levenberg, Josh / Monga, Rajat / Moore, Sherry / Murray, Derek G. / Steiner, Benoit / Tucker, Paul / Vasudevan, Vijay / Warden, Pete / Wicke, Martin / Yu, Yuan / Zheng, Xiaoqiang (2016): TensorFlow: A system for large-scale machine learning in *OSDI16*: 265–283

Auslander, Leora (1996): "Taste and Power: Furnishing Modern France". London: UCP.

Auslander, Leora (2005): "Beyond Words" in *The American Historical Review* 110, Nr. 4: 1015–45. doi:10.1086/ahr.110.4.1015.

Bell, Peter / Ommer, Björn (2015): "Training Argus. Ansätze zum automatischen Sehen in der Kunstgeschichte" in *Kunstchronik* 68, Nr. 8: 414–420.

Ommer, Björn / Bell, Peter (2016): Digital Connoisseur? How Computer Vision Supports Art History, in: Stefan Albl / Alina Aggujaro (Hgg.): *Il metodo del conoscitore - approcci, limiti, prospettive Connoisseurship nel XXI secolo*, Roma: Artemide: 187–200.

Bermeitinger, Bernhard / Donig, Simon / Christoforaki, Maria / Freitas, André / Handschuh, Siegfried (2017): "Object Classification in Images of Neoclassical Artifacts Using Deep Learning", DH2017, Montréal, Canada https://dh2017.adho.org/abstracts/590/590.pdf [letzter Zugriff 12. Januar 2018].

Berry, David M. / Fagerjord, Anders (2017): "Digital Humanities: Knowledge and Critique in a Digital Age". Cambridge, UK ; Malden, MA, USA: John Wiley & Sons.

Cordier, Sylvain (2012): "The Bellangé Album and New Discoveries in French Nineteenth-Century Decorative Arts" in: *Metropolitan Museum Journal Volume 47*: 119–147, 10.1086/670144.

Deng, J. / Dong, W. / Socher, R. / Li, L.-J. / Li, K. / Fei-Fei, L. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". in: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*: 248–255.

Donig, Simon / Christoforaki, Maria / Handschuh, Siegfried (2016): "Neoclassica – A Multilingual Domain Ontology. Representing Material Culture from the Era of Classicism in the Se-

mantic Web" in: Bozic, B. / Mendel-Gleason, G. / Debruyne, C. / O'Sullivan, D. (eds.): *Computational History and Data-Driven Humanities*. CHDDH 2016. IFIP Advances in Information and Communication Technology, vol 482. Cham, Springer: 41–53. doi: 10.1007/978-3-319-46224-0_5

Donig, Simon / Christoforaki, Maria / Bermeitinger, Bernhard / Handschuh, Siegfried (2017a): "Neoclassica – an Open Framework for Research in Neoclassicism". Montréal, Canada. https://dh2017.adho.org/abstracts/384/384.pdf [letzter Zugriff 12. Januar 2018]

Donig, Simon / Christoforaki, Maria / Bermeitinger, Bernhard / Handschuh, Siegfried (2017b): "Visual artefacts through the Black Box: Analysing Deep Learning classifcation of Neoclassical furniture images", München, 2017 https://f.hypotheses.org/wp-content/blogs.dir/1856/files/2017/03/16_Donig_Visual-Aretfacts-Black-Box.pdf [letzter Zugriff 12. Januar 2018]

Girshick, Ross / Donahue, Jeff / Darrell, Trevor / Malik, Jitendra (2015): "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*: 580–587.

Hepplewhite, A[lice]. & Co. (1790): "The Cabinet Maker and Upholsterer's Guide; or, Repository of Designs for Every Article of Household Furniture" Third edition. London: Reprinted by Batsford, B.T. 1897. ark:/13960/t2d796t9g.

Huang, Jonathan / Rathod, Vivek / Sun, Chen / Zhu, Menglong / Korattikara, Anoop / Fathi, Alireza / Fischer, Ian / Wojna, Zbigniew / Song, Yang / Guadarrama, Sergio / Murphy, Kevin (2016): "Speed/accuracy trade-offs for modern convolutional object detectors". ArXiv:1611.10012 [Cs] http://arxiv.org/abs/1611.10012 [letzer Zugriff 14. Juli 2017].

Kepetzis, Ekaterini (2006): "Antike als Vision und Rekonstruktion. Das klassische Altertum als Projektion einer idealen Gegenwelt". In: Kohle, Hubertus / Dogerloh, Annette / Arnold-Becker, Alice (eds.): *Geschichte der bildenden Kunst in Deutschland. (7) Vom Biedermeier zum Impressionismus*. Darmstadt: WBG: 325–346.

Klinke, Harald (2016): "Big Image Data within the Big Picture of Art History" in: *International Journal for Digital Art History*, Nr. 2. doi:10.11588/dah.2016.2.33527.

Krizhevsky, Alex / Sutskever, Ilya / Hinton, Geoffrey E. (2012): "ImageNet Classification with Deep Convolutional Neural Networks" in: Pereira, F. / Burges, C. J. C. / Bottou, F. / Weinberger, K. Q. (eds.): *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems* 2012: 1097–1105.

La Mésangère, Pierre de (1761-1831) (Ed.): "Collection de meubles et objets de goût comprenant: fauteuils d'appartement et de bureau, chaises garnies, canapés, divans, tabourets, lits, draperies de croisées, tables, commodes, secrétaires, bibliothèques, toilettes d'homme". Paris: Bureau du Journal des Dames.

Langenholt, Thomas (2002): Das Wittelsbacher Album: das Interieur als kunsthistorisches Dokument am Beispiel der Münchner Residenz im ersten Drittel des 19. Jahrhunderts, Norderstedt: BoD – Books on Demand.

Lin, Tsung Yi / Maire, Michael / Belongie, Serge / Hays, James / Perona, Pietro / Ramanan, Deva / Dollár, Piotr / Zitnick, C. Lawrence (2014): "Microsoft COCO: Common objects in context" in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS. pp. 740–755 doi:10.1007/978-3-319-10602-1_48.

Moretti, Franco (2013): "Distant Reading". New York / London: Verso Books.

Pawlitzki, Brigitte / Bruer, Stephanie-Gerrit / Kunze, Max (2009) (eds.): "Antik wird Mode: Antike im bürgerlichen Alltag des 18. und 19. Jahrhunderts". Wiesbaden: Harrassowitz.

PyLabelMe (2011) https://github.com/mpitid/pylabelme [letzer Zugriff 12. Januar 2017]

Sheraton, Thomas (1910): "The furniture designs". (J. Munro Bell (ed.)). London: Gibbings. https://archive.org/details/furnituredesigns00sheruoft [letzter Zugriff 01. September 2017].

The Metropolitan Museum of Art (2017): "The Met Makes Its Images of Public-Domain Artworks Freely Available through New Open Access Policy" http://www.metmuseum.org/press/news/2017/open-access letzer Zugriff 12. Januar 2018].

Burrows Zeta: Varianten und Evaluation

Schöch, Christof

schoech@uni-trier.de
Universität Trier, Deutschland

Zehe, Albin

zehe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Calvo Tello, José
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Der vorliegende Beitrag enthält methodische Überlegungen und Experimente zu "Zeta", einem von John Burrows (2007) vorgeschlagenen Maß für die Distinktivität oder "keyness" von textuellen Merkmalen (Wortformen, Lemmata, etc.). Mit solchen Maßen werden Merkmale ermittelt, die für eine bestimmte Gruppe von Texten gegenüber einer Vergleichsgruppe charakteristisch sind.

Das Exposé gibt einen Überblick zu solchen Maßen, bevor die Funktionsweise von Zeta erläutert wird. Aufbauend auf einer Neu-Implementierung in Python ("pyzeta", https://github.com/cligs/pyzeta) und Vorarbeiten (Schöch im Druck) liegt der spezifische Forschungsbeitrag dann in den folgenden Schritten: erstens werden mehrere Varianten von Zeta vorgeschlagen und implementiert; zweitens werden Verfahren zum Vergleich und der Evaluation der Ergebnisse erprobt. Ziel ist es, Zeta in seiner Funktionsweise und in seiner Beziehung zu vergleichbaren Maßen besser zu verstehen und vorhandene Nachteile des Maßes durch gezielte Modifikationen zu beheben.

Überblick und Stand der Forschung

Die vergleichende, kontrastierende Analyse zweier Gruppen von Texten ist ein in den Sprach- und Literaturwissenschaften weit verbreitetes Verfahren. Entsprechend wurden zahlreiche Maße der Distinktivität oder "keyness" von Merkmalen entwickelt und für vielfältige Fragestellungen eingesetzt. Die grundlegende Annahme solcher Maße ist, dass ein Merkmal nicht schon durch seine reine Häufigkeit in einer Textgruppe für diese charakteristisch ist, sondern dass dies auch davon abhängt, wie häufig das Merkmal in einer Vergleichsgruppe ist. Diejenigen Merkmale bekommen einen besonders hohen Wert zugewiesen, die in der einen Gruppe sehr häufig sind und zugleich in der Vergleichsgruppe sehr selten sind (Scott 1997, 236). Man kann vier Arten von Verfahren unterscheiden:

1. Verfahren, welche erwartete und beobachtete Werte vergleichen (wie "log-likelihood-ratio"; siehe Rayson und Garside 2000);
2. Verfahren, die eine Gewichtung der Häufigkeiten vornehmen (wie "tf-idf", "term frequency / inverse document frequency"; siehe Robertson 2004);
3. Statistische Hypothesentests, die Verteilungseigenschaften vergleichen (wie "Welch's t-Test"; siehe Bortz und Schuster 2010);
4. Dispersionsmaße, die nicht die Häufigkeit, sondern den Grad der konsistenten Verwendung von Merkmalen in Beziehung setzen (wie "deviation of proportions"; Gries 2008).

Die praktische Bedeutung von Distinktivitätsmaßen ist daran erkennbar, dass Korpusanalyse-Software meist eine entsprechende Funktion anbietet, so "keyness" in WordCruncher (Scott 1997) oder "spécificity" in TXM (Heiden et al. 2012). Kilgariff 2004 und Lijfijt et al. 2014 sind wichtige Arbeiten zur Evaluation von Distinktivitätsmaßen.

Was ist Zeta?

Das von John Burrows (2007) vorgeschlagene "Zeta" beruht auf einem Dispersionsmaß. Vor der Berechnung werden die Texte in kleinere Segmente gesplittet, wobei die Segmentlänge ein wichtiger Parameter ist. Dann wird für jedes Merkmal der Anteile der Segmente erhoben, in denen das Merkmal mindestens einmal vorkommt (die "document proportion"). Von diesem Anteil in der untersuchten Gruppe wird der entsprechende Anteil in der Vergleichsgruppe subtrahiert, woraus sich ein Zeta-Wert zwischen -1 und 1 ergibt.

Ein Effekt dieser Berechnungsweise ist, dass Zeta Inhaltswörter als distinktive Wörter favorisiert, Funktionswörter sowie Eigennamen hingegen penalisiert. Daraus ergibt sich eine hohe Interpretierbarkeit der Ergebnisse, die Zeta im Vergleich zu anderen Maßen für die (digitalen) Literaturwissenschaften besonders attraktiv macht. Ein Nachteil ist, dass Merkmale durch die Subtraktion niemals einen Zeta-Wert bekommen können, der höher ist als ihre "document proportion" in der untersuchten Textgruppe, selbst wenn sie gegenüber der Vergleichsgruppe deutlich überrepräsentiert sind (Abbildung 1, Wörter in den roten Rahmen; Schöch im Druck).

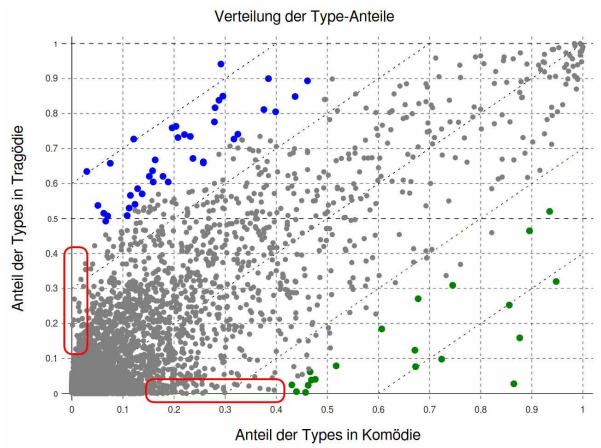


Abbildung 1: Scatterplot der Wörter in zwei Textgruppen (französische Komödien und Tragödien): "document proportions" der Wörter in zwei Textgruppen (x- und y-Achse) und resultierende Zeta-Werte (Distanz von der Diagonale).

Eine bekannte Implementierung von Zeta existiert im stylo-Paket für Rin der Funktion "oppose()" (Eder et al. 2016). Abbildung 2 zeigt für ein Beispiel die Ergebnisdarstellung in der hier verwendeten "pyzeta"-Implementierung.

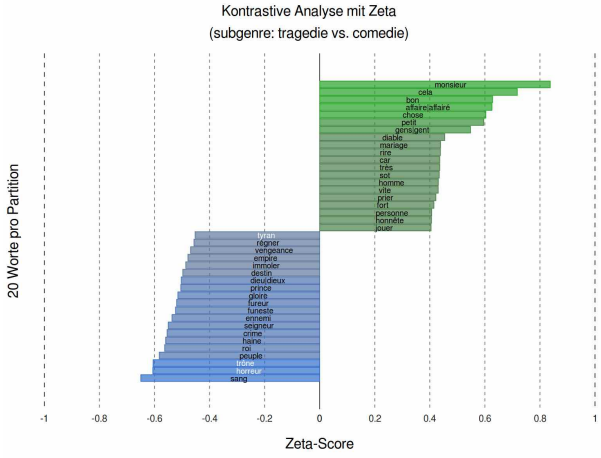


Abbildung 2: Positive und negative Keywords für französische Komödien (rechts) im Vergleich mit Tragödien (links). Zeta-Werte auf der horizontalen Achse.

Anwendungsbeispiele von Zeta gibt es in der Shakespeare-Forschung (Craig und Kinney 2009), der modernen englischsprachigen Literatur (Hoover 2010; Weidman und O'Sullivan 2017) und der Romanistik (Schöch im Druck). In der zuletzt genannten Arbeit zum französischen Theater der Klassik und Aufklärung konnte nicht nur die erwartbare, klare Differenzierung von Komödien und Tragödien gezeigt werden. Vielmehr wurde auch die spezifische Verortung der Tragikomödien deutlich, die nicht als Mischform zwischen Komödien und Tragödien zu verstehen sind, sondern eine besondere Affinität zur Tragödie aufweisen (Abbildung 3).

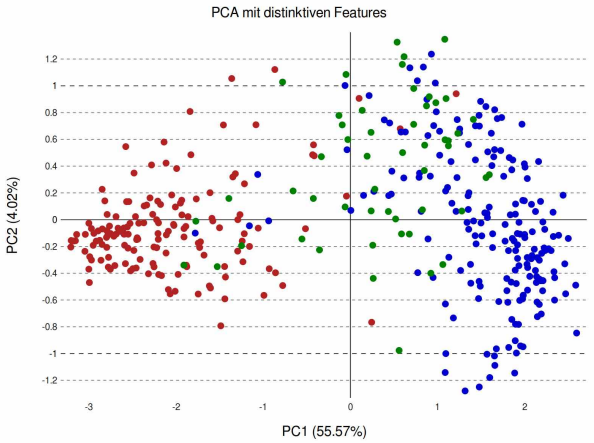


Abbildung 3: Hauptkomponentenanalyse auf Grundlage der 50 Wörter, die für Komödien und Tragödien die höchsten Zeta-Werte erhalten. Komödien in rot, Tragödien in blau, Tragikomödien in grün. Quelle: Schöch im Druck.

Varianten von Zeta

Ausgehend von der ursprünglichen Formulierung von Zeta durch Burrows als Subtraktion der "document proportions" lassen sich mehrere Faktoren identifizieren, die zur Formulierung von Varianten von Zeta geeignet erscheinen:

1. Statt "document proportions" werden relative Häufigkeiten verwendet;
2. Statt der Subtraktion erfolgt eine Division;
3. Statt nicht-transformierter Werte wird eine log2-Transformation der Werte vorgenommen.

Die Kombination dieser Faktoren ergibt 8 Varianten von Zeta (Tabelle 1).

	document proportions	relative Häufigkeiten		
	keine Transformation	log2-Transformation	keine Transformation	log2-Transformation
Subtraktion	sd0	sd2	sr0	sr2
Division	dd0	dd2	dr0	dr2

Tabelle 1: Übersicht über die getesteten Varianten von Zeta. Die Variante mit Label „sd0“ entspricht Burrows‘ Zeta.

Einige der Varianten sind mathematisch gut motivierbar und versprechen, den oben genannten Nachteil der begrenzten Werte für bestimmte Wörter auszugleichen und damit Zeta zu verbessern, es wurden aber alle implementiert und auf zwei Datensätzen evaluiert.

Datensätze

Es wurden zwei unterschiedliche Korpora verwendet. Erstens ein Korpus aus der textbox-Sammlung (Schöch et al. 2017), das Romane enthält, die zwischen 1880 und 1940 veröffentlicht wurden: jeweils 24 Texte aus Spanien und aus Lateinamerika (ca. 2,8 Millionen Tokens). Zweitens, ein Teil der Sammlung *Théâtre classique* (Fièvre 2007-2017) mit französischen Dramen: 134 Tragödien und 158 Komödien aus Klassik und Aufklärung (ca. 4,9 Millionen Tokens).

Evaluation

Die 8 Varianten führen zu unterschiedlichen Wortlisten, geordnet nach absteigenden Zeta-Werten. Vergleicht man den Beginn der Wortlisten für zwei Varianten, fällt auf, dass es wie erwartet zu Verschiebungen im Rang der distinktivsten Wörter kommt.

Ähnlichkeit der Varianten

Um den Grad der Abweichung der Ergebnisse für alle Varianten zueinander auf der Grundlage längerer Wortlisten zu erheben, ist ein quantifizierendes Verfahren unerlässlich. Ein Ansatz ist, ein Clustering der Maße auf Basis der Zeta-Werte ihrer Wörter vorzunehmen (Abbildung 4).

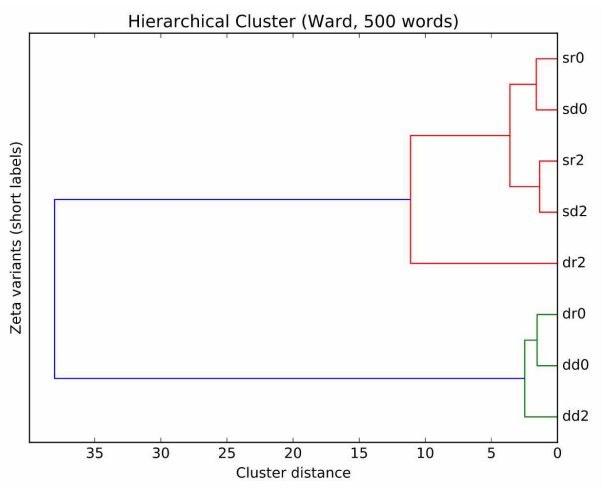


Abbildung 4: Dendrogramm auf Grundlage einer Cluster Analyse der Zeta-Werte für die 8 Zeta-Varianten (Théâtre-classique-Datensatz; 500 distinktive Wörter; Ward-Verfahren).

Abbildung 4 zeigt, dass der wichtigste Faktor für die Unterschiedlichkeit der Varianten ist, ob subtrahiert oder dividiert wird (zwei Haupt-Cluster). Die beiden anderen Variablen spielen eine viel kleinere Rolle. (Die Ergebnisse weiterer Analysen, u.a. auf Basis der RBO-Ähnlichkeit (“ranked biased order”, Webber et al. 2010), werden aus Platzgründen hier nicht diskutiert.)

Evaluation mit Klassifikationstask

Unabhängig von den Beziehungen der Varianten zueinander stellt sich die Frage, welche der Varianten von Zeta besonders gut distinktive Wörter identifiziert. Dabei kann zur Evaluation nicht auf einen Goldstandard zurückgegriffen werden: eine händische Annotation der Wörter nach dem Grad ihrer Distinktivität ist nicht möglich, weil niemand das zugrunde liegende Korpus überblicken kann. Die Qualität eines Distinktivitätsmaßes kann aber evaluiert werden, indem es als Merkmalsselektor für einen Klassifikationstask verwendet wird.

Wenn die durch Zeta am höchsten bewerteten Wörter als Features für einen Klassifikator verwendet werden, sollte dieser Klassifikator eine höhere Genauigkeit erreichen als bei einfacher Verwendung der häufigsten Wörter. Tatsächlich lässt sich dieser Effekt auf dem Korpus der spanisch-sprachigen Romane nachweisen (Tabelle 2). Zur Ermittlung einer Baseline wurde für die Klassifikation in spanische und lateinamerikanische Romane ein linearer SVM-Classifier auf den häufigsten 80, nach TF-IDF gewichteten Wörtern (ohne Stoppwörter) trainiert. Dieser Classi-

fier erreichte lediglich eine Klassifikationsgüte (F1-Score) von 0.49, ist also nicht vom Zufall zu unterscheiden.

Trainiert man stattdessen auf den 40 distinktivsten Wörtern nach Zeta (oder einer der Varianten), lassen sich Genauigkeiten von deutlich über 90% erzielen. Diese Genauigkeit kann nicht als tatsächliches Klassifikationsergebnis gesehen werden, da die distinktivsten Merkmale auf dem gesamten Korpus extrahiert wurden, ohne Aufteilung in Trainings- und Testdaten. Dennoch zeigt das Ergebnis, dass die von Zeta selektierten Merkmale tatsächlich sehr nützlich für eine Klassifikation sind. Zudem zeigen sich deutliche Unterschiede in der Performanz je nach verwendeter Variante: während mit "sd0" (=Burrows Zeta) 81% Genauigkeit erreicht wird, erhöht sich dieser Wert bei der Variante mit log2-Transformation, "sd2", auf 98%.

baseline	sd0	sd2	sr0	sr2	dd0	dd2	dr0	dr2
0.49	0.81	0.98	0.48	0.83	0.79	0.85	0.75	0.79

Tabelle 2: Klassifikationsergebnisse bei Verwendung einer linearen SVM, trainiert auf den 40 am höchsten gerankten Wörtern verschiedener Maße im Vergleich zur Baseline. Alle Werte sind der Durchschnitt einer dreifachen Kreuzvalidierung.

Fazit

Wichtigste Ergebnisse dieses Beitrags sind ein differenziertes Verständnis davon, wie Zeta im Kontext anderer Distinktivitätsmaße einzuordnen ist und wie bestimmte mathematischen Parameter sich auf die Ergebnislisten auswirken: als ein auf dem Grad der Dispersion der Merkmale beruhendes Maß, dessen entscheidende Eigenschaft die Subtraktion der Werte ist. Ein weiteres wesentliches Ergebnis sind die beiden vorgeschlagenen Strategien zum Vergleich und der Evaluation von Distinktivitätsmaßen, wenn eine direkte Evaluation auf Goldstandard-Daten nicht möglich ist.

Nächste Schritte: Wir möchten als weitere Evaluationsstrategie künstliche Texte generieren, in denen wir kontrolliert einzelne Wörter mit unterschiedlich stark abweichender Verteilung einfügen. So können verschieden Zeta-Varianten direkt dahingehend evaluiert werden, wie gut sie diese Wörter korrekt identifizieren. Zudem möchten wir neben der "document proportion" von Zeta ein weiteres Dispersionsmaß, die von Gries (2008) vorgeschlagene "deviation of proportions" als Grundlage für eine weitere Zeta-Variante verwenden. Schließlich möchten wir untersuchen, ob die hohe Interpretierbarkeit des Original-Zeta bei den Varianten mit noch höherer Klassifikationsgüte erhalten bleibt.

Eine separate Untersuchung ist in Vorbereitung zu zwei eng zusammenhängenden Fragen: wie sich unterschiedliche Segmentlängen einerseits auf die Ergebnisse auswirken, und wie sich die Ergebnisse verändern, wenn unterschiedlich lange Texte nicht mit allen Segmenten in die Berechnung eingehen, sondern aus jedem Einzeltext zufällig eine identische Anzahl von Segmenten gesampelt wird.

Übergeordnetes Ziel all dieser Arbeiten zu Zeta ist es letztlich weniger, ein perfektes Distinktivitätsmaß zu identifizieren, als ein justierbares Maß vorzuschlagen, bei dem in Abhängigkeit von Daten und Forschungsfragen dynamisch Parameter verändert und die resultierenden Verschiebungen in den Ergebnissen visualisiert werden können.

Bibliographie

Bortz, Jürgen, and Christof Schuster (2010). *Statistik für Human- und Sozialwissenschaftler*. 7. Auflage. Berlin: Springer.

Burrows, John (2007). "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22, no. 1: 27–47. doi:10.1093/llc/fqi067.

Craig, Hugh, and Arthur F. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. 1st ed. Cambridge University Press.

Eder, Maciej, Mike Kestemont, and Jan Rybicki(2016). "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 16, no. 1: 1–15.

Fièvre, Paul, ed. (2007-2013). "Théâtre classique." Paris: Université Paris-IV Sorbonne. http://www.theatre-classique.fr.

Gries, Stefan Th. (2008). "Dispersions and Adjusted Frequencies in Corpora." *International Journal of Corpus Linguistics* 13, no. 4: 403–37. doi:10.1075/ijcl.13.4.02gri.

Heiden, Serge (2010). "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–98. Sendai: Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764/en.

Hoover, David L. (2010). "Teasing out Authorship and Style with T-Tests and Zeta." In *Digital Humanities Conference*. London:

ADHO. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.

Kilgarriff, Adam (2001). “Comparing Corpora.” *International Journal of Corpus Linguistics* 6, no. 1: 97–133. doi:10.1075/ijcl.6.1.05kil.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2014). “Significance Testing of Word Frequencies in Corpora.” *Digital Scholarship in the Humanities* 31, no. 2: 374–97. doi:10.1093/llc/fqu064.

Rayson, Paul, and R. Garside (2000). “Comparing Corpora Using Frequency Profiling.” In *Proceedings of the Workshop on Comparing Corpora*, 1–6. Hong Kong: ACM.

Robertson, Stephen (2004). “Understanding Inverse Document Frequency: On Theoretical Arguments for IDF.” *Journal of Documentation* 60, no. 5 : 503–20.

Schöch, Christof (im Druck). “Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie.” In *Quantitative Verfahren in der Literaturwissenschaft. Von einer Scientia Quantitatis zu den Digital Humanities*, edited by Andrea Albrecht, Sandra Richter, Marcel Lepper, Marcus Willand, and Toni Bernhart. Berlin: de Gruyter. https://cligs.hypotheses.org/files/2017/09/Schoech_2017-preprint_Zeta-fuer-die-kontrastive-Analyse.pdf.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, and Stefanie Popp (angenommen). “The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI.” *Journal of the Text Encoding Initiative* http://cligs.hypotheses.org/files/2017/09/Schoech-et-al_2017_Textbox.pdf.

Scott, Mike (1997). “PC Analysis of Key Words and Key Key Words.” *System* 25, no. 2: 233–45.

Webber, William, Alistair Moffat, and Justin Zobel (2010). “A Similarity Measure for Indefinite Rankings.” *ACM Trans. Inf. Syst.* 28, no. 4: 20:1–20:38. doi:10.1145/1852102.1852106.

Computationale Beschreibung visuellen Materials am Beispiel des Graphic Narrative Corpus

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Dubray, David
ddubray@uni-potsdam.de
Universität Potsdam, Deutschland

Krügel, André
kruegel@uni-potsdam.de
Universität Potsdam, Deutschland

Einleitung

Die digitale Revolution in den Geisteswissenschaften hat diesen eine Reihe neuer Methoden eröffnet. Durch die heute verfügbaren großen Datenmengen und intelligenten Algorithmen haben sich zwar einige bisher offengeliebene geisteswissenschaftliche Kernfragen beantworten lassen, jedoch wird mancherorts eine Methodenfokussiertheit und Theoriemangel kritisiert (Gumbrecht, 2014). Ob die Digitalen Geisteswissenschaften zu einer darüber hinausgehenden tiefergreifenden Veränderung der geisteswissenschaftlichen Erkenntnis führen werden, bleibt abzuwarten; derzeit scheint es, als werde das Potenzial der neuen methodischen Zugänge erst noch ausgelotet. In anderen Wissenschaften hat aber die Verfügbarkeit computationaler Modelle und der damit einhergehende Zwang, implizite Annahmen zu explizieren und Theorien formal testbar zu machen, signifikant zur Theoriebildung und -prüfung beigetragen (cf. Myung & Pitt, 2002; Lewandowsky & Farrell, 2011). Deshalb besteht die begründete Hoffnung, dass computationale Modellierung auch die Geisteswissenschaften bereichern wird.

Ein Großteil der Forschung in den digitalen Geisteswissenschaften beschäftigt sich mit Text. Es gibt hier eine fruchtbare interdisziplinäre Zusammenarbeit von Literaturwissenschaften und Computerlinguistik; im Umfeld des "Distant Reading” sind umfangreiche Werkzeuge entstanden, mit

denen sich etwa stilometrische Analysen oder Topic Modeling computergestützt vornehmen lassen (Blei, 2012; Juola, 2006). Auch netzwerkanalytische Methoden aus der theoretischen Physik und computationalen Soziologie haben hier interessante neue Perspektiven eröffnet (Schich et al., 2014). Dagegen ist die digitale Analyse visuellen Materials noch relativ wenig entwickelt oder standardisiert, obwohl dieses für Disziplinen wie z.B. Kunstgeschichte oder Archäologie von zentralem Interesse ist. In den letzten Jahren wurden durch Entwicklungen im Bereich der Convolutional Neural Networks (CNN) die Möglichkeiten automatisierter Bildanalyse revolutioniert. Während in klassischen Ansätzen der maschinellen Bildverarbeitung ein hohes Ausmaß an Expertenwissen notwendig war, um Merkmale zu definieren, mit denen sich das Material sinnvoll beschreiben ließ ("engineered features"), lernen CNNs die Merkmale durch Fehlerrückführung (Backpropagation) selbst.

Convolutional Neural Networks sind eine besondere Klasse künstlicher neuronaler Netze, die sich durch eine 2D-Anordnung der Neuronen, innerhalb einer Schicht geteilte Gewichte und lokale Konnektivität auszeichnen. Sie eignen sich insbesondere für die Analyse von Bildmaterial. Die Netze sind typischerweise auf einer großen Anzahl von Fotos in Objektklassifikationsaufgaben trainiert worden, dabei bilden sich auf verschiedenen Ebenen der CNNs Repräsentationen aus, die denen im menschlichen visuellen System ähnlich sind. Neuronen auf niedrigen Ebenen des Netzwerks haben oft eine Filterantwort, die relativ einfache Merkmale kodiert, vergleichbar z.B. mit Kantendetektoren im frühen visuellen Kortex, während Neuronen auf höheren Ebenen recht komplexe Merkmale kodieren können, z.B. Texturen oder Teile von Gesichtern. Da diese Merkmale relativ generisch sind, ist zu erwarten, dass Transfer auf neuartiges Material gelingt. Es sind heute einige derart vortrainierte Netzwerke verfügbar, die sich mit relativ wenig Aufwand an neues Material anpassen lassen. Der Vergleich der Gewichte für verschiedene Materialtypen erlaubt dann auch Rückschlüsse über deren Unterschiede.

Fragestellung

Generalisieren die auf Fotos vortrainierten Netzwerke auch auf zeichnerisches Material? Wir berichten von Experimenten, in denen wir das Material des Graphic Narrative Corpus (GNC, Dunst et al., 2017) mit CNNs beschreiben. Der Graphic Narrative Corpus repräsentiert das erste digitale Korpus von englischsprachigen Graphic Novels mit derzeit 130 Titeln. Die ersten Kapitel dieser Werke werden von menschlichen Kodierern annotiert, dabei werden u.a. die Identität und der Ort zentraler Charaktere, Orte von Panels, Sprechblasen und Textboxen (Captions) und Onomatopeia sowie der Text selbst notiert. Außerdem werden Blickbewegungen von Lesern erhoben (Eye-Tracking), um Aufschluss über die Aufmerksamkeitsverteilung auf Seite der Rezipienten zu erhalten.

Die Beschreibung des GNC mit CNNs hat verschiedene Ziele. Erstens erhoffen wir uns Aufschluss über stilistische Unterschiede zwischen Werken und Genres, z.B. mittels Berechnung von Distanzmaßen basierend auf den Modellparametern. Allgemeiner könnte so der Weg zu einer visuellen Stilometrie aufgezeigt werden, die auch für inhaltliche Bereiche außerhalb der Graphic Novels relevant ist, etwa im Sinne einer computationalen Kunstgeschichte (Saleh & Elgammal, 2015; Manovich, 2015). Zweitens ermöglicht die Beschreibung mit Hilfe der Merkmale tiefer CNNs durch sogenannte Region Proposal Networks (Girshick et al., 2013) die Detektion von Objektklassen. Beispielsweise könnten sich Sprechblasen oder handelnde Charaktere lokalisieren lassen. Wenn Klassen von Objekten automatisiert lokalisiert werden können, erleichtert dies die Arbeit der Annotatoren sehr. Die Ergebnisse können also zurück in das Annotationswerkzeug fließen, um eine Teilautomatisierung zu ermöglichen. Drittens ist aus kognitionspsychologischer Perspektive interessant, welche Merkmale die Aufmerksamkeit auf sich ziehen. Die Korrelation der Netzwerkbeschreibung mit den Blickbewegungsdaten ermöglicht eine Modellierung der Aufmerksamkeitssteuerung auf einem deutlich höheren Auflösungsgrad als die subjektive Beschreibung.

Methode

Für die Modellierung des Materials nutzen wir die Architektur VGG der Visual Geometry Group in Oxford (Simonyan & Zisserman, 2014), insbesondere VGG-16 und VGG-19. Diese Wahl ist motiviert durch die Einfachheit der Architektur, die die Interpretation der Gewichte erleichtert. Das zugrundeliegende Netzwerk lässt sich aber prinzipiell austauschen; andere Architekturen wie ResNet (He et al., 2015) oder Inception (Szegedy et al., 2015) sind denkbar und sollten ähnlich gute Ergebnisse liefern. Für die Vorhersage der Aufmerksamkeitsverteilung der Leser nutzen wir die Architektur Deep Gaze II (Kümmer et al., 2016). Deep Gaze II ist ein neuronales Netz, das auf VGG-19 aufsetzt und die Antwort einiger dessen Schichten nutzt, um "empirische Salienz" vorher-

zusagen. Empirische Salienz ist operationalisiert durch Messung von Mauspositionen beim Aufdecken eines verschwommenen Bildes bzw. Messung von Blickbewegungsdaten beim Betrachten von Fotos natürlicher Szenen. Die Fotos sind andere, als die für das Training von VGG-19 benutzten. Man beachte, dass sowohl VGG-19 als auch Deep Gaze II auf Fotos trainiert wurden, also nie Graphic Novels gesehen haben. Da sie jedoch Merkmale und Gewichte herausgebildet haben, die für die Interpretation (von Bildern) der menschlichen Umwelt nützlich sind, kann man vermuten, dass sie sich auch für die Analyse von Zeichnungen eignen. Zwar sind Zeichnungen Abstraktionen, haben aber als solche einen Bezug zur visuellen (Photo-)Realität.

Ergebnisse

Die Ergebnisse zeigen, dass sich mit Hilfe von Neuronen auf höheren Ebenen der tiefen CNNs recht gut bestimmte Klassen von Objekten lokalisieren lassen. Beispielsweise eignen sich einige Kombinationen von Merkmalen zuverlässig als Sprechblasendetektoren (Abb. 1). Dies ist insofern bemerkenswert, als die Detektion von Sprechblasen sich für klassische Ansätzen der maschinellen Bildbverarbeitung als schwieriges Problem dargestellt hat (Rigaud et al., 2013).

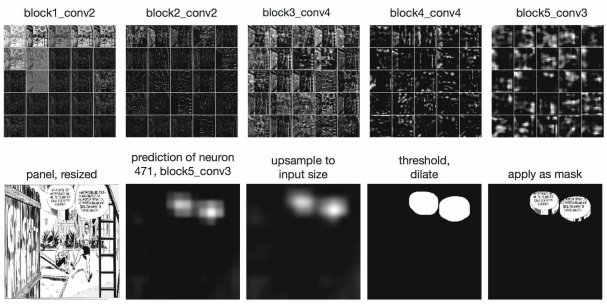


Abbildung 1. Sprechblasendetektion mithilfe von Convolutional Features.

Auch für die Erkennung gezeichneter Gesichter eignen sich CNNs, allerdings ist hier ein Training auf Ansichten in verschiedenen Perspektiven (Frontal, Profil) notwendig. Und schließlich lässt sich die empirische Fixationsverteilung mit Deep Gaze II insgesamt sehr überzeugend reproduzieren (Abb. 2). Die CNN-Features kodieren also aufmerksamkeitsrelevante Merkmale.

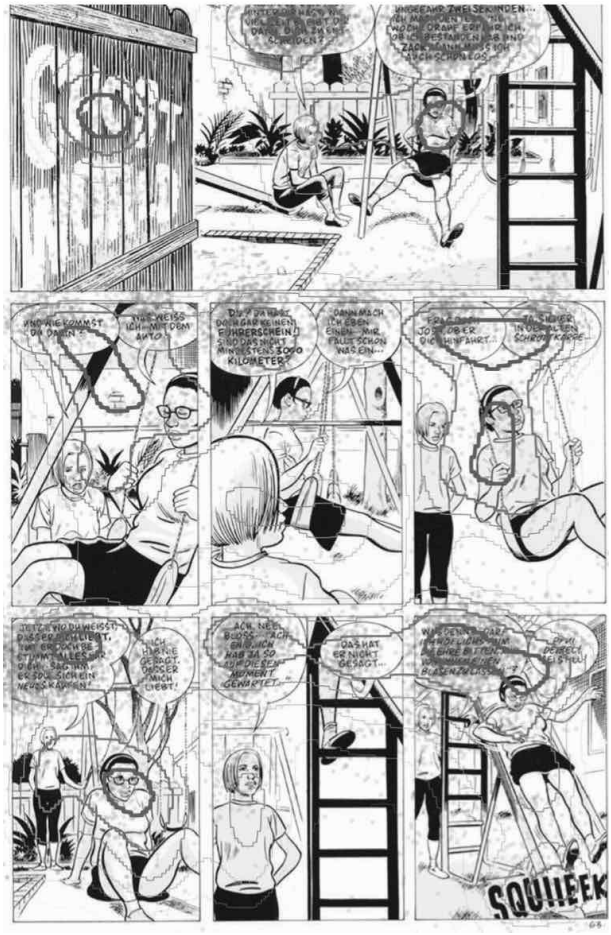


Abbildung 2. Empirische Fixationsverteilung von 100 Lesern (Punkte) und DeepGaze II-Vorhersagen (Konturlinien).

Insgesamt eigenen sich auf Fotos trainierte CNNs schon ohne spezifisches weiteres Training recht gut zur Beschreibung gezeichneten Materials in Graphic Novels.

Diskussion

Die "objektive" Beschreibung eröffnet vielfältige Anwendungen. Einerseits kann, wie oben skizziert, die Annotation visuellen Materials durch Nutzung von vorgeschlagener Regionen deutlich erleichtert werden, etwa vergleichbar mit dem bei verbalen Material durch Verwendung von Optical Character Recognition (OCR) ermöglichten Übergang von kompletter Transkription zum Korrekturlesen. Hier soll angemerkt werden, dass vielversprechende CNN-basierte Ansätze zur Textlokalisation (Sudholt & Fink, 2016) und OCR existieren (Lee & Osindero, 2016). Andererseits sind durch das Vorliegen visueller Merkmale (Features) vielfältige stilometrische Anwendungen denkbar. Zum Beispiel lassen sich auf-

grund der Merkmale Ähnlichkeiten verschiedener Zeichner und Künstler berechnen und durch Gruppierung (Clustering) im Merkmalsraum auch Stile definieren. Auch die weitergehende Exploration der Repräsentation auf verschiedenen Schichten des Netzwerks scheint eine vielversprechende Aufgabe weiterer Forschung. Beispielsweise könnte der Vergleich der Antworten auf fotographische versus zeichnerisch abstrahierte Abbilder von Exemplaren einer Kategorie Hinweise auf das Wesen der Abstraktion geben, oder es lassen sich visuelle Merkmale identifzieren, die in besonderem Ausmaß die Aufmerksamkeitszuwendung im Leseprozess und bei der Rezeption von Zeichnungen leiten.

Wir haben beispielhaft aufgezeigt, wie sich Werkzeuge der mathematisch-computationalen Modellierung eignen, um grafisches Material zu analysieren und zu beschreiben. Die Hoffnung ist, dass eine visuelle Stilometrie die Digitalen Geisteswissenschaften im Bereich visuellen Materials in ähnlicher Art und Weise bereichert wie computerlinguistische Ansätze im Bereich der Textanalyse. Digitale Analysen liefern mächtige neue Werkzeuge, die mittel- bis längerfristig auch eine neue Theoriebildung fördern könnten.

Bibliographie

Blei, D. (2012). Probabilistic Topic Models. Communication of the ACM, 55, 77-84.

Dunst, A., Hartel, R. & Laubrock, J. (im Druck). The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities. *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017).*

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*:1311.2524.

Gumbrecht, H. U. (2014) **.** Das Denken muss nun auch den Daten folgen. *Frankfurter Allgemeine Zeitung, 12.03.2014, S. 14*, http://www.faz.net/aktuell/feuilleton/geisteswissenschaften/neue-serie-das-digitale-denken-das-denken-muss-nun-auch-den-daten-folgen-12840532.html

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Juola, P. (2006) **.** Authorship Attribution. *Foundations and Trends in Information Retrieval, 1*, 233-334.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016) **.** Deepgaze II: Reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563.

Lee, C. & Osindero, S. (2016). Recursive recurrent nets with attention modeling for OCR in the wild. *CoRR, abs/1603.03101* (CVPR 206).

Lewandowsky, S. & Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks: SAGE.

Manovich, L. (2015). Data Science and Digital Art History. *International Journal for Digital Art History, 1*, 13-35. http://dx.doi.org/10.11588/dah.2015.1.21631.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In J. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology (Third Edition), Volume IV (Methodology)* (pp. 429–459). New York: John Wiley & Sons.

Rigaud, C., Burie, J. C., Ogier, J. M., Karatzas, D., & Weijer, J. V. D. (2013). An active contour model for speech balloon detection in comics. *Proceedings of the 12th International Conference on Document Analysis and Recognition, 1240–1244*.

Saleh, B. & Elgammal, A. M. (2015) **.** Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855.

Schich, M., Song, C., Ahn, Y. Y., Mirsky, A., Martino, M., Barabási, A. L., & Helbing, D. (2014). A network framework of cultural history. *Science, 345(6196)*, 558-562.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Sudholt, S. & Fink, G. A. (2016). Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 277–282.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.

Contextualizing Bandera: Ein Distant Watching Ansatz

Bermeitinger, Bernhard
Bernhard.Bermeitinger@uni-passau.de
Lehrstuhl für Informatik mit Schwerpunkt Digital Libraries and Web Information Systems, Universität Passau, Deutschland

Howanitz, Gernot
Gernot.Howanitz@uni-passau.De
Lehrstuhl für Slavische Literaturen und Kulturen, Universität Passau, Deutschland

Radisch, Erik
Erik.Radisch@uni-passau.de
Lehrstuhl für Digital Humanities, Universität Passau, Deutschland

Einleitung

Zahlreiche geisteswissenschaftliche Projekte im Kontext der Digital Humanities sind textlastig. Ein Grund dafür ist das breite Spektrum an etablierten Verfahren, das für solche Fragestellungen zur Verfügung steht. 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 Medien, 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 des zu untersuchenden Phänomens eingehen, erscheinen deshalb vielversprechend.

Neben Texten spielen Bilder in zahlreichen kulturellen Zusammenhängen eine tragende Rolle, so auch im Internet: Bilder und Videos werden kopiert, bearbeitet, geteilt und damit zu sogenannten Memen (Shifman 2013). Dabei entsteht eine große Zahl an Bildern bzw. Videos, die sich aufgrund ihrer schieren Masse einem traditionellen *Close Reading* entzieht. Gleichzeitig liegen die Bilder und Videos digitalisiert vor und sind zum Teil durch die verwendete Web-Plattform (z. B. YouTube) mit Metadaten annotiert. Aus diesen Gründen sind Bilder-Meme und virale Videos prädestiniert für den Einsatz quantitativer Verfahren.

Die hier vorgestellte neue Methode setzt *Distant Watching* um und versucht, automatisiert den Bildinhalt zu erfassen. Damit wird im Vergleich zu bisherigen Ansätzen, die entweder nur sehr generische Informationen wie Schnittkurven (Howanitz 2015) oder Farben (Burghardt/Wolff 2016) herauslesen bzw. ganz auf manueller Annotation beruhen (Dunst/Hartel 2016), eine wesentliche Verbesserung erreicht. Ein State-of-the-Art *Regional Convolutional Neural Network* (RCNN) wird auf konkrete vorselektierte Symbole in Videos trainiert, um diese in einem großen Videokorpus automatisiert erkennen zu können. Damit wird erstmals der Bildinhalt von Videos automatisiert erfass- und quantitativ messbar. Je nach (Co-)Präsenz oder Absenz von Symbolen können Rückschlüsse auf den Inhalt des Videos gezogen werden.

Diese Ausweitung des methodischen Repertoires auf Bilder bzw. Videos ermöglicht den Kulturwissenschaften quantitative Perspektiven auf Malerei, Photographie und Film. Auch Objekte oder performative Handlungen können über Bild- bzw. Videodokumentationen einer quantitativen kulturwissenschaftlichen Analyse zugeführt werden. Diese quantitative methodische Innovation muss allerdings durch eine qualitative ergänzt werden. Die vorliegende Studie setzt sich zum Ziel, diese Innovationen anzustoßen.

MultiPath Network

Unsere Studie beruht auf einer Weiterentwicklung eines *Convolutional Neural Networks* (CNN). Ein konventionelles CNN, beispielsweise *VGG19* (Russakovsky et al. 2015), ist in der Lage, ein Eingabebild anhand eines vorher durchgeführten Trainings in genau eine vordefinierte Klasse einzuordnen. Für ideale Eingabebilder, etwa jene aus *MNIST*- (Lecun et al. 1998) oder *CIFAR100-Korpus* (Krizhevsky 2009), ist das ein praktisch gelöstes Problem. Enthält das Bild allerdings Instanzen mehrerer Klassen, produziert ein konventionelles CNN keine verwertbaren Ergebnisse. CNNs wurden deshalb zu *Regional Convolutional Neural Networks* (RCNN) weiterentwickelt, wie beispielsweise zu *MultiPath Network* (Zagoruyko et al. 2016).

RCNNs operieren zweistufig: Zuerst werden automatisch Regionen in einem Bild vorgeschlagen und intern noch verfeinert. Anschließend werden diese vorgeschlagenen Ausschnitte klassifiziert. *MultiPath* ist standardmäßig durch die generischen Bilder des *COCO*-Korpus (Lin et al. 2014) vortrainiert und erkennt alltägliche Dinge (Katzen, Flugzeuge, Speisen, usw.). Wie kleine explorative Experimente gezeigt haben, ist es notwendig, auf dem allgemeinen Training aufbauend eigene Trainingsläufe für jene visuellen Symbole zu entwickeln und durchzuführen, die uns für die konkrete kulturwissenschaftliche Fragestellung interessieren. Dies bedeutet einen hohen Aufwand an Rechenzeit und verlangt spezialisierte Hardware; dafür lassen sich RCNNs dann aber auf jegliche Arten von Symbolen oder andere visuell unterscheidbare Merkmale trainieren und können zu deren Identifizierung eingesetzt werden.

Stepan Bandera

Zentrum der Untersuchung dieses Papers ist die Rezeption des ukrainischen Nationalisten Stepan Bandera, die in sich die Ambivalenz ukrainischer Erinnerungskultur vereint und die im gegenwärtigen Ukrainekonflikt immer wieder polarisiert: Für das prorussische Lager ist er ein Faschist und Massenmörder, seine Anhänger werden als *Banderovcy* mit Faschisten gleichgestellt. Für die ukrainisch-nationalistische Seite ist Bandera ein idealisierter Held, der kompromisslos für die nationale Unabhängigkeit kämpfte. Neue Medien werden intensiv genutzt, um die von der jeweiligen Seite präferierte Sicht auf Bandera durchzusetzen. Eine erste Untersuchung zeigte, dass sich diese Instrumentalisierung durch alle größeren digitalen Medien zieht und bereits vor 2014 immanent war (Fredheim et al. 2014). Unser Paper baut auf dieser Vorarbeit auf; wir vergleichen das Youtube-Korpus vor dem Kriegsausbruch in der Ukraine mit einem heutigen Korpus, um aufzuzeigen, ob und wenn ja, wie der Ukrainekonflikt die bereits vorhandene unterschiedliche Instrumentalisierung verändert hat. Als Korpus dienen uns die jeweils 200 ersten *YouTube*-Suchresultate für die Begriffe "Stepan Bandera" und "Степан Бандера". Dass die *YouTube*-Suche keine objektive Übersicht über den Datenbestand liefert, sondern die Ergebnisliste je nach Land, Browser und anderen Details des Suchenden anpasst, sei angemerkt, kann an dieser Stelle allerdings nicht weiter diskutiert werden.

Neben der propagandistischen Instrumentalisierung ist ebenso auf die Ebene der "post-memory" (Hirsch 2012) zu verweisen. Marianne Hirsch beschreibt mit diesem Konzept eine Auseinandersetzung mit einer traumatischen Vergangenheit, die man selber nicht erlebt hat. Dabei spielen visuelle Medien eine entscheidende Rolle, weil sie, so Hirsch, emotional aufladbarer sind als Texte. Wie diese emotionale Komponente im Rahmen des *Distant Watchings* mitbedacht werden kann, ist sowohl aus qualitativer als auch als quantitativer Sicht zu klären.

Methode

Automatische Lokalisierung und Klassifizierung erfordern eine genaue Definition der Objekte, die gefunden werden sollen. Wir haben eine Reihe von 12 typischen Symbolen festgelegt, die im Kontext der Auseinandersetzung über Bandera häufig verwendet werden, darunter Symbole des russischen oder ukrainischen Nationalismus bzw. des Faschismus.

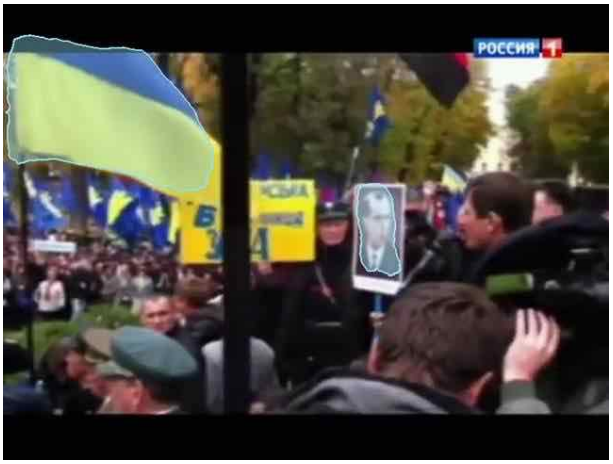
Symbole des ukrainischen Nationalismus:	Symbole des *Faschistische Symbole:*	Symbole des polnischen Nationalismus	Symbole des *russischen (sowjetischen) Nationalismus:*
----------	----------	----------	----------
ukrainisches Wappen (182)	Hitler-Bilder (95)	polnisches Wappen (38)	Hammer & Sichel (111)
Bandera-Bilder (110)	Hakenkreuz (190)	Falanga (95)	Georgsband (147)
ukrainische Flagge (168)	SS-Rune (107)		
Flagge der UPA (48)			
Swoboda-Symbol (129)			

Die Auswahl der Symbole erfolgte aus einem Close Watching einer Reihe von Beispiel-Videos zu Bandera heraus. Es handelt sich hierbei um wiederkehrende Symbole, die klar dafür genutzt wurden, um eine wertende Aussage im Bildprogramm zu platzieren. Die Symbole werden manuell anhand von Bildern aus den Beispielvideos sowie an Bildern aus dem Internet annotiert. Bisherige Tests zeigen, dass die Symbole auf mindestens 80 Bildern annotiert werden müssen, um robuste Ergebnisse erzielen zu können. Die automatische Klassifikation erlaubt es, das gesamte Korpus nach den trainierten Symbolen durchsuchen zu lassen. Außerdem gibt es Auskunft, wie viel Platz es in dem Video eingenommen hat und wie lange es sichtbar war. Mithilfe dieser Daten kann das Korpus statistisch analysiert und beispielsweise festgestellt werden, in welchem symbolischen Kontext Bandera gezeigt wird.

Das experimentelle Korpus umfasst 813 Bildern mit insgesamt 1483 Annotationen. Das ergibt im Mittel 123 annotierte Objekte pro Kategorie. Eine Annotation besteht aus Punktkoordinaten, die den Umriss des Objekts angeben und den zugehörigen Name der Klasse. Einem Bild sind zwischen 1 und 13 Annotationen zugeordnet. Im Durchschnitt sind es 1,7; der Median beträgt 1. Um Overfitting zu vermeiden, wird das Korpus, wie üblich, zufällig in Trainings- und Evaluationsdaten in einem Verhältnis 80/20 aufgeteilt.

Die Evalutationsmetrik der ersten Stufe wird mit dem numerischen Maß *Intersection over Union* (IoU) aus dem Intervall von 0 bis 1 angegeben. Je mehr dieser Wert gegen 1 geht, desto mehr stimmt die vorgeschlagene Region mit der vordefinierten Region überein.

Experimente mit den beiden Unterstufen der ersten Stufe (Lokalisierung von Objekten und deren Verfeinerung) zeigen einen durchschnittlichen IoU von 0,68 (Median 0,76). Für Symbole, die in deutlich über 80 mal in Bildern annotiert werden konnten ist der IoU mit 0,74 (Median 0,76) nochmals höher.

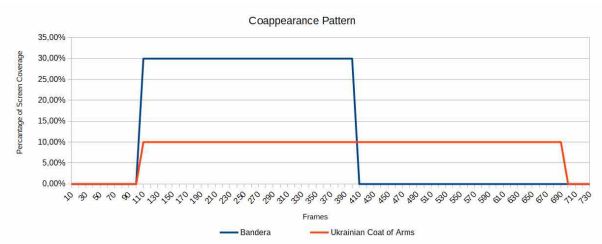


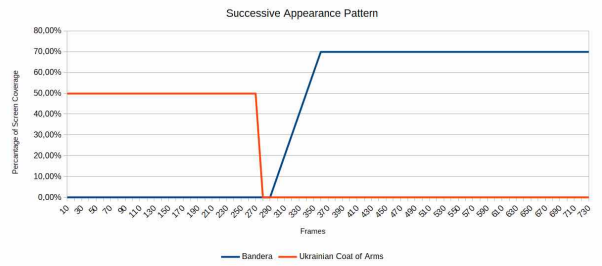
Auf den hier gezeigten Beispielbildern sind automatisch erkannte Regionen trainierter Symbole farblich hervorgehoben. Auf dem ersten Bild wurde die ukrainische Flagge sowie das Konterfei Banderas erkannt, auf dem zweiten Bild Hammer und Sichel, auf dem dritten Bild das Gesicht Adolf Hitlers. Aber nicht nur Vorhandensein und Nichtvorhandensein der Symbole ist feststellbar, auch Position und Größe können extrahiert werden.

Unsere Programme sollen es ermöglichen, Videos direkt aus dem Stream heraus in MultiPath zur Verarbeitung laden zu lassen, weil eine Speicherung der Videos auf der Festplatte augenblicklich nicht möglich ist, da dies gegen die AGB von YouTube verstoßen würde. Unsere Abwandlung von MultiPath Network erzeugt aus den Informationen des Videostreams eine Beschreibungsdatei im JSON-Format mit Informationen, welche Symbole in welchen Frames erkannt wurde und wie viel Fläche es eingenommen hatte.

Dies ermöglicht eine statistische Auswertung welche Symbole zusammen mit anderen gezeigt werden. und welche nicht. Für jedes Frame wird berechnet, wie viel Platz ein Symbol im Frame einnimmt. Diese Werte können dann für das gesamte Video ausgewertet werden, wie exemplarisch unten in zwei Bildern gezeigt wird. Solche statistischen Auswertungen ermöglichen die automatische Kontextualisierung der Bandera-Videos. Je nach Symbolen, die gleichzeitig, oder im Umfeld mit, Bandera gezeigt werden, lassen sich Aussagen treffen, ob das Video pro-russisch oder pro-ukrainisch einzuordnen ist.

Auch lässt sich auf diese Weise untersuchen, ob mit der Zeit bestimmte Symbole (zum Beispiel faschistische) in den Videos zu- oder abnahmen.





Zusammenfassung, Ausblick, Kritische Reflexion

Die Methode steht und fällt mit der Zusammenstellung der zu trainierenden Symbole. Werden wichtige Symbole beim Training außen vor gelassen, hat dies große Auswirkungen auf die interpretatorische Aussagefähigkeit. Ähnlich wie bei Texten ergeben sich auch bei visuellen Medien erst durch die Kombination von Close und Distant Watching Synergie-Effekte (Hayles 2010).

Unser Experiment konnte zeigen, dass der Ansatz, YouTube-Videos mit *MultiPath* "aus der Ferne" zu betrachten, funktioniert. Derzeit wird das Training von *MultiPath* optimiert, um bestmögliche Resultate zu generieren. Nächster Schritt ist dann die komplette Auswertung des Korpus und eine Überprüfung der Resultate durch ein Close Watching ausgewählter Videos. Die Ergebnisse dieser Verfeinerung werden mit in den Vortrag einfließen.

Bibliographie

Burghardt, M. / Wolff, C. (2016). “Digital Humanities in Bewegung. Ansätze für die computergestützte Filmanalyse” in *DHd 2016: Book of Abstracts*, 108-112.

Dunst, A. / Hartel, R. (2016). “Die Corpusanalyse multimodaler Erzählungen am Beispiel graphischer Romane” in *DHd 2016: Book of Abstracts*, 120-122.

Fredheim, R. / Howanitz, G. / Makhortykh, M. (2014). “Scraping the Monumental: Stepan Bandera through the Lens of Quantitative Memory Studies” in *Digital Icons* 12 (2014), 25-53. http://www.digitalicons.org/wp-content/uploads/issue12/files/2014/11/DI12_2_Fredheim.pdf [letzer Zugriff 11. September 2017].

Hayles, N. K. (2010): “How We Read: Close, hyper, Machine” in: *ADE Bulletin, 150*: 62-79.

Hirsch, M. (2012). “The Generation of Postmemory: Writing and Visual Culture After the Holocaust”, New York: Columbia University Press.

Howanitz, G. (2015). “Jožin z Bažin – Ein Mem, aus der Distanz betrachtet” in: Simonek, Stefan / Doschek, Jolanta (eds.): *Slawische Popkultur*. Wien: PAN, 63-80.

Shifman, L. (2013): “Memes in Digital Culture”. Cambridge (MA): MIT Press.

Krizhevsky, A. (2009). “Learning Multiple Layers of Features from Tiny Images” https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf [letzter Zugriff 12. Januar 2018].

Lecun, Y. / Bottou, L. / Bengio, Y. / Haffner, P. (1998): “Gradient-based learning applied to document recognition” in: *Proceedings of the IEEE*, 86(11): 2278–2324 doi: https://doi.org/10.1109/5.726791 .

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: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS. pp. 740–755 doi:10.1007/978-3-319-10602-1_48. http://arxiv.org/abs/1405.0312 [letzter Zugriff 12. Januar 2018].

Russakovsky, O. / Deng, J. / Su, H. / Krause, J. / Satheesh, S. / Ma, S. / Huang, Z. (2015): “ImageNet Large Scale Visual Recognition Challenge” in: *International Journal of Computer Vision*, 115(3): 211–252 doi:10.1007/s11263-015-0816-y.

Zagoruyko, S. / Lerer, A. / Lin, T.-Y. / Pinheiro, P. O. / Gross, S. / Chintala, S. / Dollár, P. (2016): “A MultiPath Network for Object Detection”. *BMVC* http://arxiv.org/abs/1604.02135 [letzter Zugriff 12. Januar 2018].

Critical Digital Cultural Studies: Digitale Kulturwissenschaft und die Kritik des Mem-Begriffs

Ernst, Thomas
t.ernst@uva.nl
University of Amsterdam, Niederlande

Die Digitalisierung hat weltweit nicht nur positive Effekte hervorgebracht, sondern auch vielfach kritisierte Probleme verursacht – man denke nur an die monopolistische Marktmacht intransparenter Firmen der digitalen Information (Google, Facebook), die neuen Formen der Überwachung (Snowden vs. NSA) oder die Popularisierung politischer Öffentlichkeiten und deren Beitrag zu unerwarteten Entwicklungen (Präsident Trump, Brexit, AfD). Während zwar einerseits die digitalen Medien und Kommunikationsverhältnisse den Alltag der meisten europäischen Bürgerinnen und Bürger bestimmen, hat sich andererseits noch keine digitale Kultur etabliert, in

der die Bürger auf breiter Ebene kritisch und bewusst mit der Macht von Algorithmen oder den Inhalten Allgemeiner Geschäftsbedingungen umgingen.

In einer solchen Periode eines umfassenden gesellschaftlichen Medienwandels ist es die Aufgabe einer Digitalen Kulturwissenschaft, die sich neben allgemeinen kulturellen Gegenständen auch mit digitalen Öffentlichkeiten, Medien und Methoden kulturtheoretisch und -analytisch beschäftigt, sowohl innerhalb der Geisteswissenschaften als auch in die Gesellschaft hinein den Stand der digitalen Kultur kritisch zu reflektieren. Dazu gehört die Frage des Ausgleichs zwischen dem Schutz und der Öffnung persönlicher und institutioneller Daten, die Frage nach klarer Zuweisung von Identitäten oder der Möglichkeit der anonymisierten Mediennutzung, nach Formen der kollaborativen digitalen Arbeit und ihrer Regulierung oder nach den Online-Marktverhältnissen und der (Nicht-)Offenlegung von Algorithmen, die öffentliche Massenmedien regulieren – sowie deren Auswirkungen auf unterschiedliche Kulturen. Mit solchen Themen beschäftigen sich intensiv die Medien- und Medienkulturwissenschaften (vgl. Engemann 2003, Lovink 2008, Reichert 2013, Schäfer 2011), aber auch Künstler und Publizisten (Lanier 2014, Morozov 2013) sowie Politiker und politisch Bewegte (Albrecht 2014, Wagner 2017).

Für den Bereich der Kulturwissenschaft ist eine grundsätzlichere theoretische Begriffsarbeit notwendig, bevor eine Digitale Kulturwissenschaft, die sich explizit als (digital)kulturkritisch versteht, grundiert werden kann. Um solche – in einem internationalen, interdisziplinären und komparatistischen Kontext zu verortende und von mir hiermit eingeführte – *Critical Digital Cultural Studies* zu begründen, muss zunächst ein angemessener Begriff der Kritik bestimmt werden. Dieser lässt sich in einem dreifachen Verfahren konturieren: Zunächst kann ein kulturwissenschaftlicher Begriff der Kritik in einer kritischen Lektüre der Auseinandersetzungen von Theodor W. Adorno, Max Horkheimer und Michel Foucault mit den Begriffen der Kritik und der Aufklärung ermöglichen, einen kulturtheoretisch fundierten Begriff einer Kritik der digitalen Vernunft zu entwickeln.

Max Horkheimer hat 1937 mit seiner Unterscheidung von traditioneller und kritischer Theorie „die mathematische Naturwissenschaft, die als ewiger Logos erscheint," einer geisteswissenschaftlich-begrifflich ausgerichteten „kritische[n] Theorie der bestehenden Gesellschaft" (Horkheimer 1995: 215) gegenübergestellt. Als Folge des Nationalsozialismus und des Holocaust denken Horkheimer und Theodor W. Adorno in ihrer *Dialektik der Aufklärung* über die Konsequenzen aus diesem Zivilisationsbruch nach: Fortan müsse das Denken von der „Selbstzerstörung der Aufklärung" ausgehen, zugleich sei allerdings – so die dialektische Figur – „die Freiheit in der Gesellschaft vom aufklärenden Denken unabtrennbar" (Horkheimer/Adorno 2000: 3). Eine starke Skepsis gegen die Massenmedien und die technische Entwicklung prägen diesen Ansatz, der die Selbstzerstörung der Aufklärung durch mehr (individuelle) Aufklärung überwinden will.

Es mag überraschen, dass Michel Foucault in seiner Arbeit über einen diskursanalytisch fundierten Begriff der Kritik explizit an diese Vernunftkritik der Kritischen Theorie anschließt, wobei die Technikkritik weniger in seinem Fokus steht. Foucault stellt den Begriff der Kritik jenem der Regierung, der „Regierbarmachung der Gesellschaft" (Foucault 1992: 17), entgegen: Kritik als „die Kunst nicht dermaßen regiert zu werden." (Ebd.: 12) Innerhalb dieses Verständnisses der Kritik stehen die Begriffe des Wissens und der Macht zentral: Wissen bezeichnet bei Foucault „alle Erkenntnisverfahren und -wirkungen [...], die in einem bestimmten Moment und in einem bestimmten Gebiet akzeptabel sind"; Macht wiederum jene Mechanismen, „die in der Lage scheinen, Verhalten oder Diskurse zu induzieren" (ebd.: 32). Foucault plädiert somit für die kritische Analyse jener gesellschaftlichen Vernetzungen, die der Wissensproduktion und zugleich der Legitimation dieses Wissens – und somit der Machtausübung – dienen (vgl. ebd.: 37). Mit Foucault ließe sich ein kulturtheoretisch und diskursanalytisch fundierter Begriff der Kritik formulieren, der gerade nicht – wie noch bei Adorno und Horkheimer – technik- und massenfeindlich ist, sondern vielmehr auch bei der Analyse digitaler Massenmedien genutzt werden und zugleich kritisch gegen die Wissensproduktion der traditionellen und der Digitalen Kulturwissenschaft selbst gewendet werden kann (vgl. auch Foucault 1995, Foucault 2003).

Ein solchermaßen konturierter kritischer Begriff des Wissens kann zusätzlich bereichert werden durch bisherige Arbeiten aus verwandten Schulen und Teildisziplinen. Dazu zählen unter anderem die *Critical Code Studies*, die ihre Analysen stärker auf Code und Software selbst konzentrieren, die *Critical Digital Studies* (vgl. Kroker/Kroker 2013) und die *Critical Cultural Studies*. Dies würde zugleich ermöglichen, Gegenstände, Felder und Methoden der *Critical Digital Cultural Studies* zu differenzieren, wobei hier insbesondere eine klare Bestimmung der Objekte der Kritik, eine klare Differenzierung der Kriterien der Kritik sowie eine selbstreflexive Kritik der Kritik (vgl. auch Ullmaier 2017: 71) unabdingbar erscheinen. Eine solche kritische Form der Geis-

teswissenschaft, die auf digitales Wissen und digitale Methoden fokussiert und zugleich auf kulturtheoretisches Wissen zurückgreifen kann, wird nicht nur eine kritische Funktion in der Gesellschaft einnehmen können, sondern auch innerhalb der Kulturwissenschaften bestehende Konzepte, Theorien und Methoden problematisieren können – und zugleich auch in die Digital Humanities selbst wirken.

Die Produktivität einer solchen Engführung von kulturwissenschaftlichen Begriffen der Aufklärung und der kritischen Geisteswissenschaft mit Theoremen, Methoden und Themen der Digital Humanities kann schließlich an einem repräsentativen Beispiel vorgeführt werden. Der Biologe Richard Dawkins führte 1976 den ‚Mem'-Begriff ein, um analog zur Genetik auch die soziokulturelle Evolution begrifflich fassen zu können (vgl. Dawkins 1976). Dieses Konzept hat sich inzwischen auch in den Medien- und Kommunikationswissenschaften sowie teilweise auch in den Kulturwissenschaften durchgesetzt, insbesondere um virale Internetphänomene konzeptionell beschreiben zu können (vgl. u.a. Shifman 2014). Die breite Kritik an der Memtheorie, die unterkomplex sei und viele kultur- und sozialwissenschaftliche Erkenntnisse missachte, kann hier in einer direkten Konfrontation mit der machtkritischen Diskursanalyse Michel Foucaults geleistet werden, die ein wesentlich komplexeres Modell sozialer Entwicklung und Kommunikation zur Verfügung stellt, das über den Diskursbegriff die Produktion, Legitimation und Distribution von Wissen unter den jeweiligen Machtverhältnissen zu beschreiben versucht (Foucault 1995, Foucault 2003). Damit trägt der Vortrag zu einer selbstreflexiven Diskussion einer Erkenntniskategorie der Digital Humanities bei, indem er zwar nicht selbst digitale Methoden oder Daten einsetzt, diese jedoch kulturtheoretisch reflektiert – und somit für eine kulturtheoretische Auseinandersetzung mit ihren Erkenntniskategorien auch innerhalb der Digital Humanities plädiert, die perspektivisch ihre gesellschaftliche Relevanz befördern kann.

Bibliographie

Albrecht, Jan Philipp (2014): *Finger weg von unseren Daten! Wie wir entmündigt und ausgenommen werden*. München: Droemer Knaur.

Bal, Mieke (2006): *Kulturanalyse*. Frankfurt am Main: Suhrkamp.

Dawkins, Richard (1976): *The Selfish Gene*. Oxford: Oxford University Press.

Engemann, Christoph (2003): *Electronic Government – vom User zum Bürger. Zur kritischen Theorie des Internet*. Bielefeld: transcript.

Foucault, Michel (1992): *Was ist Kritik?* Aus dem Französischen von Walter Seitter. Berlin: Merve.

Foucault, Michel (1995): *Archäologie des Wissens*. Übersetzt von Ulrich Köppen. 7. Aufl., Frankfurt am Main: Suhrkamp.

Foucault, Michel (2003): *Die Ordnung des Diskurses*. Aus dem Französischen von Walter Seitter. Mit einem Essay von Ralf Konersmann. 9. Aufl., Frankfurt am Main: S. Fischer.

Horkheimer, Max (1995): „Traditionelle und kritische Theorie (1937)", in: Ders.: *Traditionelle und kritische Theorie. Fünf Aufsätze*. Frankfurt am Main: S. Fischer 205-259.

Horkheimer, Max / Adorno, Theodor W. (2000): *Dialektik der Aufklärung. Philosophische Fragmente*. 12. Aufl., Frankfurt am Main: S. Fischer.

Lanier, Jaron (2014): *Wem gehört die Zukunft? Du bist nicht die Zukunft der Internet-Konzerne. Du bist ihr Produkt*. Hamburg: Hoffmann und Campe.

Lovink, Geert (2008): *Zero Comments. Elemente einer kritischen Internetkultur*. Bielefeld: transcript.

Morozov, Evgeny (2013): *Smarte neue Welt. Digitale Technik und die Freiheit des Menschen*. München: Blessing.

Reichert, Ramón (2013): *Die Macht der Vielen. Über den neuen Kult der digitalen Vernetzung*. Bielefeld: transcript.

Schäfer, Mirko Tobias (2011): *Bastard Culture! How User Participation Transforms Cultural Production*. Amsterdam: Amsterdam University Press. http://oapen.org/download?type=document&docid=371358 [letzter Zugriff 14. Januar 2018].

Schäfer, Mirko Tobias / van Es, Karin (eds., 2017): *The Datafied Society. Studying Culture Through Data*. Amsterdam: Amsterdam University Press.

Shifman, Limor (2014): *Meme. Kunst, Kultur und Politik im digitalen Zeitalter*. Berlin: Suhrkamp.

Ullmaier, Johannes (2017): „Kategorien der Kritik. Detaillierte Inhaltsübersicht zu einer ungeschriebenen Studie", in: *Testcard. Beiträge zur Popgeschichte* 25: 70-88.

Wagner, Thomas (2017): *Das Netz in unsere Hand! Vom digitalen Kapitalismus zur Datendemokratie*. Köln: PapyRossa.

Cäsar Flaischlens „Graphische Litteratur-Tafel“ – digitale Erschließung einer großformatigen Karte zur Deutschen Literatur

Börner, Ingo
ingo.boerner@univie.ac.at
Universität Wien, Österreich

Fischer, Frank
ffischer@hse.ru
National Research University Higher School of Economics, Moskau, Russland

Hechtl, Angelika
angelika.hechtl@wu.ac.at
Wirtschaftsuniversität Wien, Österreich

Jäschke, Robert
r.jaschke@sheffield.ac.uk
University of Sheffield, UK

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam, Deutschland

Vorhaben

Cäsar Flaischlens „Graphische Litteratur-Tafel“ von 1890 stellt den Versuch dar, die Entwicklung der Deutschen Literatur mit ihren Einflüssen aus anderen Nationalliteraturen graphisch in der Form eines Flusses darzustellen. Gegenstand des Vortrags ist die digitale Edition und Bereitstellung des Vorwortes und der Karte sowie der entwickelte Workflow: Für die Edition wurde das graphische Karteninventar kodiert. Mithilfe computergestützter Bildanalyse können nicht-textuelle Informationen der Visualisierung erfasst und einer quantitativen Analyse zugeführt werden.

Visualisierung von Literaturgeschichte

In den letzten Jahren lässt sich ein Trend innerhalb der Literaturwissenschaft – u.a. der Literaturgeschichtsschreibung – ausmachen, Fragestellungen auf der Grundlage von großen Datenkorpora zu beantworten. Charakteristisch für diese Art von Literaturwissenschaft ist ein Methodenimport aus Natur- und Sozialwissenschaften, der sich nicht zuletzt in den Darstellungsformen deutlich zeigt.

Auch wenn sich diese Zugänge gegenwärtig großer Beliebtheit erfreuen, sind Darstellungsweisen wie jene in Morettis einflussreichem Buch „Kurven, Karten, Stammbäume: Abstrakte Modelle für die Literaturgeschichte“ (Moretti 2007) keineswegs ein Phänomen der Gegenwart, denn Literaturgeschichtsschreibung bedient sich bereits seit der Antike Bildmedien und anderer – nicht rein textueller – Präsentationsformen. Darstellungen von AutorInnen, wie etwa jene auf Raffaels berühmtem Parnassfresko in den Vatikanischen Museen lassen sich aus heutiger Perspektive wie Diagramme lesen. Information zu Relevanz sowie Verbindungen einzelner AutorInnen sind hier im Bildmedium kodiert. (vgl. Hölter/Schmitz-Emans 2013, Hölter 2005)

Das Parnassfresko ist nur eine Art, wie sich Kanonbildung, Rezeption und die Zugehörigkeit zu einer AutorInnengruppe darstellen lassen. Häufig gewählte Darstellungsformen sind (Stamm-)Baum (Lima 2014) und Fluss. Die Literaturwissenschaft greift damit Formen auf, die erst mit Fortschritten in der Buchproduktion durch die Entwicklung der Lithographie möglich geworden sind und zunächst in der Geschichtsschreibung Anwendung gefunden haben (vgl. Rosenberg/Grafton 2010).

Cäsar Flaischlens „Graphische Litteratur-Tafel“

Wie produktiv sich diese tradierten Denkbilder auf die Konzeptualisierung von (Literatur-)Karten auswirken können, zeigt die „Graphische Litteratur-Tafel” (1890) des deutschen Autors Cäsar Flaischlen (1864–1920). Flaischlens großformatige Karte (58x86,5 cm) visualisiert den – wie es im Untertitel heißt – „Einfluss fremder Literaturen” auf die deutsche Literatur und bedient dafür die Denkfigur geschichtlicher Prozesse als Fluss, bestehend aus einer Summe von Einflüssen. Er knüpft damit an eine Darstellungstradition von Weltgeschichte an, wie sie durch die Graphik

„Strom der Zeiten“ (1804) des österreichischen Historiographen Friedrich Strass maßgeblich geprägt wurde (vgl. Rosenberg und Grafton, 2010).

In Cäsar Flaischlens Œuvre nimmt die aufwendig gestaltete Litteratur-Tafel eine Sonderstellung ein: Die für ihre Zeit ungewöhnliche literaturwissenschaftliche Arbeit erschien beinahe zeitgleich mit seiner Promotion 1899 und sollte Flaischlens einzige Publikation zur Literaturgeschichte bleiben. Flaischlen verließ die akademische Welt und war als Mitherausgeber und Redakteur von Literatur- und Kunstzeitschriften tätig. Heute ist der Autor hauptsächlich für seine Mundartgedichte und Erzählungen bekannt.

Auf der Karte wird die deutschsprachige Literatur von ihren Anfängen bis in Flaischlens Gegenwart dargestellt. Was in der Grafik zunächst als zwei sich schlängelnde Bäche der „Volks- und Kunstpoesie“ um 750 beginnt, entwickelt sich im Laufe der Jahrhunderte zu einem breiten Strom, in welchen über den gesamten (Zeit-)Verlauf Zuflüsse aus anderen (hauptsächlich) europäischen Nationalliteraturen einmünden. In der Legende der Karte führt Flaischlen folgende Einflüsse auf: Altes- und Neues Testament, Englisch, Französisch, Nordisch, Orientalisch, Klassisches Altertum, Spätlateinisch, Niederländisch, Italienisch, Spanisch, Schwedisch/Dänisch/Norwegisch, Russisch.

Als Vertreter positivistischer Denkrichtung unternimmt Cäsar Flaischlen also den Versuch, die Darstellung der Zeit als Fluss mit den exakten Wissenschaften zu verbinden. Die Tatsache, dass er keine Quellen für die Zusammenstellung seiner Tafel nennt, spricht dafür, dass er den gängigen Kanon abbildet. Im 8-spaltigen Vorwort zur Tafel spielt Flaischlen zwar den Zusammenhang von quantitativem Befund und Visualisierung herunter – so gibt er etwa zu bedenken, dass die Breite des Flusses „nicht mathematisch berechnet” sei, die Platzierung der Autoren folgt jedoch einem gewissen Prinzip: Die Autoren habe er am „Höhepunkt“ ihres Schaffens eingezeichnet.

Der Informationsgehalt der „Litteraturtafel“ ist sehr hoch: Auswahl, Platzierung und Größe der beeinflussenden und beeinflussten Autoren ermöglichen die Rekonstruktion von Flaischlens Datengrundlage und lassen Rückschlüsse auf die hierfür verwendeten Quellen zu.

So sind etwa in der Tafel Namen von Autoren, kanonischen Texten, literarischen Gruppierungen und literarischen Schulen enthalten. Ferner nutzt Cäsar Flaischlen typographische Gestaltungsmöglichkeiten wie Schriftart, Schriftgröße, Farbwahl und Unterstreichung), Symbole (Kreise in verschiedenen Größen, römische und lateinische Ziffern) und die farbliche Schraffur der (Zu-)Flüsse. Am unteren Rand der Karte ist zwar eine Legende angebracht, diese weist jedoch lediglich die Bedeutung der Schraffur aus (z.B. blau für Einflüsse aus der Englischen Literatur, rot für Einflüsse aus der Französischen Literatur, etc.). Weitere Angaben, insbesondere Erläuterungen zu verwendeten Schriftarten und -größen fehlen jedoch.

Technische Umsetzung

Das vorgestellte Projekt „Cäsar Flaischlens Graphische Litteratur-Tafel digital“ unternimmt den Versuch eines ‚reverse engineering‘ und erschließt dieses Dokument früher Visualisierung literaturgeschichtlicher Daten mit Methoden der Digital Humanities.

Dazu wurden die Koordinaten von angeführten Personen, Texten, literarischen Strömungen und Schulen unter Verwendung des GIMP ImageMap-Editors ermittelt und entsprechend den in den TEI P5 Guidelines (TEI Consortium 2017) definierten Transkriptionskonventionen („Advanced Uses of surface and zone“) erfasst, um die spatiale Dimension (Kodierung von Information über räumliche Anordnung) ebenfalls zugänglich machen zu können. Die Personen- und Werkreferenzen wurden mit den entsprechenden Normdaten (GND, VIAF, wikidata) verknüpft. Das kartographische Inventar der Karte wurde unter Rückgriff auf CSS innerhalb von @style erfasst. Dies ermöglicht nun, neben den textuellen Informationen zusätzlich die typographische Gestaltung des Textes auszuwerten.

Die TEI-Daten werden über ein github-repository bereitgestellt und als Webseite aufbereitet. Das Interface fügt die drei separaten Abschnitte von Flaischlens Karte zusammen. In einer Print-Ausgabe wäre der zusammengefügte Fluss beinahe drei Meter lang und somit nur schwer lesbar. Die Web-Version erlaubt über das Scroll-Interface jedoch einen komfortablen Zugang. Die kodierten Informationen sind über Register erschlossen. Ein Prototyp der Edition ist unter http://litteratur-tafel.weltliteratur.net zugänglich.

Computergestützte Analyse des kartographischen Inventars (Zwischenergebnisse)

Durch eine Analyse des kartographischen Inventars lässt sich das Zeichensystem rekonstruieren, das zudem gängigen kartographischen Konventionen folgt. Beeinflusste und beeinflussende Autoren werden durch Unterstreichung unter-

schieden. Die Farbe der Unterstreichung entspricht der farblichen Kennzeichnung der Zuflüsse und ordnet somit die Autoren einer der beeinflussenden Nationalliteraturen zu. Durch die Verwendung unterschiedlicher Schriftgrößen wird die „Relevanz“ eines Autors bzw. Textes für die deutsche Literatur zum Ausdruck gebracht. Besonders wichtige deutsche Autoren sind in Konturschrift ausgeführt, vergleichbar der Beschriftung von Städten in Abhängigkeit von ihrer Einwohnerzahl (vgl. Kohlstock 2004: 100). EIne Analyse der Typografie erlaubt es somit, Flaischlens „Verständnis“ der deutschen Literatur zu rekonstruieren.

Um die farblich kodierten Informationen der Karte ebenfalls berücksichtigen zu können, wurde ein Zugang über Verfahren aus dem Bereich computergestützter Bildanalyse gewählt. Mittels der Open Source Computer Vision Library (openCV) wurden die Pixel nach Farbbereichen klassifiziert und quantitativ ausgewertet, um die „Einflüsse“ zu messen und ihre Veränderungen im Laufe der Zeit visualisieren zu können. Abbildung 1 zeigt jene Pixel, der Kategorie „rot“ und somit dem französischen Einfluss zugeordnet wurden.

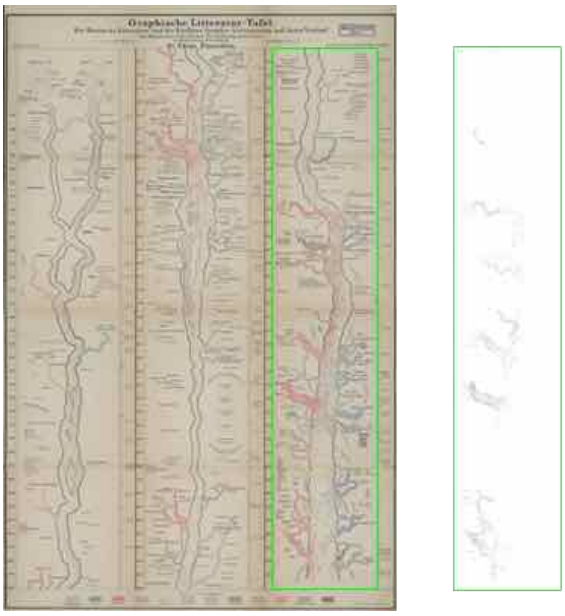


Abbildung 1: Original und klassifizierte Pixel

Für Abbildung 2 wurde wurden die Pixel nach Jahren gruppiert und als Liniendiagramm visualisiert. Deutlich erkennbar sind Spitzen um 1620, 1665, 1715 und 1800, die sich verschiedenen Epochen der französischen Literaturgeschichte zuordnen lassen: Dem Französischen Klassizismus und der Aufklärung. Der Peak in den 1860er Jahren ist mit dem französischen Naturalismus verbunden.

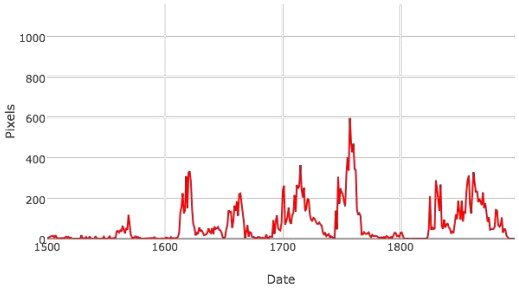


Abbildung 2: Französischer Einfluss auf Basis der Pixelklassifikation

Ausblick

Das Projekt erschließt nicht nur einen inspirierenden Vorläufer gegenwärtiger Versuche von Visualisierung literaturgeschichtlicher Daten mithilfe von ‘Graphen, Karten und Stammbäumen’, sondern erprobt auch auf einer methodologischen Ebene Möglichkeiten und Workflows zur Erschließung und Kodierung älterer graphischer Darstellungen von (Literatur-)geschichte. Darüber hinaus liefert es Impulse für die Arbeit mit der TEI für das Encoding von Bildmaterial, indem die Eignung von primär zur Kodierung von Manuskripten eingesetzten Tags für Grafiken und Diagramme überprüft werden.

Bibliographie

Flaischlen, Cäsar (1890): *Graphische Litteratur-Tafel. Die deutsche Litteratur und der Einfluß fremder Litteraturen auf ihren Verlauf vom Beginn der schriftlichen Ueberlieferung an bis heute in graphischer Darstellung*. Berlin: Behr's Verlag.

Hölter, Achim (2005): „Überlegungen zu Raffaels Parnass-Fresko als Kanonbild“ in: Heimböckel, Dieter / Werlein, Uwe (eds.): *Der Bildhunger der Literatur. Festschrift für Gunter E. Grimm*. Würzburg 51-68.

Hölter, Achim / Schmitz-Emans, Monika (eds.) (2013): *Literaturgeschichte und Bildmedien.* Heidelberg: Synchron.

Kohlstock, Manuel (2004): *Kartographie. Eine Einführung.* Stuttgart: UTB.

Lima, Manuel (2014): *The Book of Trees. Visualizing Branches of Knowledge*. New York: Princeton Architectural Press.

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

openCV: *Open Source Computer Vision Library*. http://opencv.org [letzter Zugriff 24. September 2017].

Rosenberg, Daniel / Grafton, Anthony (2010): *Cartographies of Time*. New York: Princeton Architectural Press.

Strass, Friedrich (1803): *Der Strom Der Zeiten.* http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~281767~90054624 [letzter Zugriff 24. September 2017].

TEI Consortium (2017). *TEI P5 – Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0.* http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [letzter Zugriff 25. September 2017].

Das neue "Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft" und seine Auswirkungen für Digital Humanities

Kamocki, Pawel
pawel.kamocki@gmail.com
WWU Münster, Germany; ELDA, France; IDS Mannheim

Ketzan, Erik
eketzan@gmail.com
Birkbeck, University of Lonfon

Wildgans, Julia
j.wildgans@googlemail.com
IDS Mannheim; Universität Mannheim

Witt, Andreas
witt@ids-mannheim.de
IDS Mannheim; Universität zu Köln

Forschungsdaten im Bereich der Digital Humanities sind bekanntlich häufig urheberrechtlich bzw. durch das sui-generis-Recht für Datenbanken geschützt. Dementsprechend ist eine Erhebung und Verwendung der Daten nur rechtlich zulässig, wenn der Rechteinhaber seine Zustimmung erteilt hat oder eine gesetzlich vorgesehene Schrankenregelung eingreift. Die Einholung der notwendigen Lizenzen ist allerdings häufig sehr aufwendig und nicht zuletzt kostspielig; um den mit der Nutzung der Daten verbundenen Aufwand zu verringern, führte der Gesetzgeber nun Schrankenregelungen für die Wissenschaft ein.

Auf EU-Ebene eröffnete die Urheberrechtsrichtlinie (RL 2001/29/EG) den Mitgliedstaaten die Möglichkeit, Ausnahmeregelungen zu schaffen, um die Vervielfältigung von Werken und ihre öffentliche Zugänglichmachung für nicht-kommerzielle Zwecke zu ermöglichen. Einzige Voraussetzung dafür war die Angabe der jeweiligen Quelle. Eine ähnliche Ausnahmeregelung für Forschungszwecke ist in Art. 9 b der Richtlinie 96/9/EG über den rechtlichen Schutz von Datenbanken vorgesehen; allerdings erlaubt diese lediglich die Entnahme (und nicht die Weiterverwendung) von Daten aus einer der Öffentlichkeit - in welcher Weise auch immer - zur Verfügung gestellten Datenbank.

Damit diese Regelungen in den Mitgliedstaaten verbindliche Geltung erlangen können, müssen die Richtlinien von den nationalen Gesetzgebern in nationales Recht umgesetzt werden. Dabei haben sie allerdings einen weiten Spielraum: Die Richtlinie ist lediglich hinsichtlich ihres Ziels verbindlich. Die Mitgliedstaaten können also selbst entscheiden, ob und inwieweit sie die genannten Ausnahmeregelungen in ihren nationalen Rechtsordnungen aufnehmen (denn diese können, müssen aber nicht eingeführt werden). Um die Interessen von Wissenschaftlern auf der einen Seite und Rechteinhabern (d.h. insbesondere den Verlagen) auf der anderen Seite auszugleichen, entscheiden sich die nationalen Gesetzgeber häufig für die Einführung enger Schrankenregelungen. So ist beispielsweise in Deutschland gem. § 52a UrhG lediglich die Nutzung von veröffentlichten “kleinen Teilen” eines Werkes (also - richterrechtlich festgelegt - bis zu 25 % eines Werkes bis max. 100 Seiten) bzw. Werken “geringen Umfangs” (also Werke mit weniger als 25 Seiten, einzelne Bilder und Musikstücke) für nichtkommerzielle Zwecke zur Forschung erlaubt. Damit verbunden ist allerdings zwingend ein Vergütungsanspruch des jeweiligen Rechteinhabers, der nur durch eine Verwertungsgesellschaft geltend gemacht werden kann; die dazu notwendigen Verhandlungen zwischen den Universitäten und der VG Wort dauerten viele Jahre und mussten schließlich durch einen Richter geklärt werden. Erst 2006 konnte ein Rahmenvertrag unterzeichnet werden, der den Preis vergleichsweise tief festsetzte: 0,008 EUR pro Seite pro Nutzer. In der Praxis ergab sich aber bald das Problem, dass

die Ausnahmeregelung (und der damit verbundene Vergütungsanspruch) durch eine vertragliche Regelung umgangen wurden - Denn wurde der Inhalt aufgrund eines Vertrags (z.B. einer Lizenz) zugänglich gemacht, so konnte eine Regelung des Vertrags dem Nutzer einfach verbieten, das Werk in der gem. § 52a UrhG gestatteten Weise zu nutzen. Dies hatte zur Folge, dass den Wissenschaftlern alle Vorteile der Schrankenregelung wieder verloren gingen.

Im Jahr 2017 entschied sich der deutsche Gesetzgeber zu handeln: Das Bundesministerium der Justiz und für Verbraucherschutz erarbeitete einen Entwurf für das sog. Urheberrechts-Wissensgesellschafts-Gesetz, das letztlich - nach einem bemerkenswert kurzen Gesetzgebungsprozess - vom Bundestag verabschiedet wurde. Ab März 2018 wird der alte § 52a UrhG (und weitere Normen, die urheberrechtliche Nutzung von Werken in der Forschung, Archiven und Bibliotheken zum Gegenstand hatten) durch die neuen §§ 60a-60h UrhG ergänzt. Dies gilt zunächst für einen Zeitraum von 5 Jahren. Danach muss der Gesetzgeber entscheiden, ob er die Gültigkeit explizit verlängert oder durch andere Regelungen ersetzt (was nicht völlig unwahrscheinlich ist, wenn eine neue EU-Richtlinie für den digitalen Binnenmarkt erlassen wird).

Von besonderem Interesse für die Digital Humanities sind dabei § 60c und § 60d UrhG.

§ 60c UrhG erlaubt nun ausdrücklich die Vervielfältigung, die Verbreitung und die öffentliche Zugänglichmachung von bis zu 15 % eines Werkes zum Zwecke der nicht-kommerziellen wissenschaftlichen Forschung (also nicht wie die bisherige Rechtsprechung 25 %). Die Regelung beinhaltet also keine Seitenanzahl mehr, wodurch sie dem digitalen Zeitalter angepasst wird. Für die eigene wissenschaftliche Forschung (also ohne Veröffentlichung) dürfen sogar bis zu 75 Prozent eines Werkes vervielfältigt werden.

Unabhängig von diesen Regelungen dürfen Abbildungen, einzelne Beiträge aus Zeitungen oder Zeitschriften, sonstige Werke geringen Umfangs und vergriffene Werke vollständig genutzt werden.

§ 60d UrhG erlaubt das Data Mining für nichtkommerzielle Forschungszwecke (die Beschränkung “nicht kommerziell” stammt dabei aus der zugrundeliegenden Richtlinie und darf daher vom nationalen Gesetzgeber nicht übergangen werden - sonst liegt ein Verstoß gegen EU-Recht vor!): Ursprungsmaterial darf also auch automatisiert und systematisch vervielfältigt werden, um daraus insbesondere durch Normalisierung, Strukturierung und Kategorisierung ein auszuwertendes Korpus zu erstellen und dieses einem bestimmt abgegrenzten Kreis von Personen (vermutlich den Mitgliedern des eigenen Forschungsteam) für die gemeinsame wissenschaftliche Forschung zur Verfügung zu stellen. Nach Abschluss des Forschungsprojekts ist das gesamte Korpus zu löschen oder einem Archiv oder eine Bibliothek zur dauerhaften Aufbewahrung zu übermitteln. Diese Ausnahmeregelung betrifft nicht nur urheberrechtlich geschützte Werke, sondern erfasst auch Werke, die durch das sui-generis-Recht für Datenbanken geschützt sind: Obwohl die Richtlinie 96/9/EG keine Schrankenregelung für die Weiterverwendung zu Forschungszwecken vorsieht, fand der nationale Gesetzgeber einen geschickten Weg, diese Einschränkung zu umgehen.

Anzumerken ist, dass die neuen Schrankenregelungen nicht mehr durch vertragliche Regelungen umgangen werden können, d.h. auf Vereinbarungen, die erlaubte Nutzungen nach den §§ 60a bis 60f UrhG zum Nachteil der Nutzungsberechtigten beschränken oder untersagen, kann sich der Rechteinhaber nicht berufen (vgl. § 60g UrhG). Allerdings sind die Nutzungen zu vergüten; dieser Anspruch kann erneut nur durch die Verwertungsgesellschaften geltend gemacht werden. Auch die angemessene Höhe dieser Vergütung wird wahrscheinlich Gegenstand langer Verhandlungen werden, die Preisfestsetzung wird jedenfalls eine abschreckende Wirkung haben. Beispielsweise setzte eine Vereinbarung zwischen den Bibliotheken und den Verwertungsgesellschaften im Jahr 2006 die Vergütung für die Digitalisierung und öffentliche Zugänglichmachung von Büchern fest auf 120 % des Nettopreises des Buches.

Die neuen Schrankenregelungen sind jedenfalls ein entscheidender Schritt in die richtige Richtung. Es ist wichtig, die DH-Gemeinschaft über diese aktuellen Entwicklungen zu informieren. Um allerdings das volle Potenzial des Data Mining in der EU zur Entfaltung zu bringen, ist der EU-Gesetzgeber gefragt. Tatsächlich gab es Ende 2016 einen Vorschlag für eine neue Richtlinie für den digitalen Binnenmarkt, die eine verbindliche Schrankenregelung für das Data Mining öffentlicher Forschungseinrichtungen (wie z.B: Universitäten) vorsieht, auch zu kommerziellen Zwecken. Derselbe Richtlinienvorschlag enthält auch einige vernünftige Einschränkungen für den Zugang zu Material - insbesondere verpflichtet er die Lizenznehmer zur regelmäßigen Information des Lizenzgebers über die Nutzung ihrer Werke. Allerdings muss auch er zunächst vom EU-Gesetzgeber vollständig ausgearbeitet und verabschiedet und anschließend von den Mitgliedsstaaten umgesetzt werden. Zum jetzigen Zeitpunkt ist eine Vorhersage darüber, wie die finale Version der Richtlinie aussehen wird, absolut nicht möglich. Allerdings

sollte die DH-Community unbedingt zeitnah diesbezüglich mit Informationen versorgt werden.

Bibliographie

Kamocki, Pawel (2016): "Allow Mining!" The Argument for "Orthogonal Uses" of Intellectual Works in the Debate on Text and Data Mining. In: Revue Internationale du Droit d'Auteur 247. S. 4-85.

Lehmberg, Timm/Rehm, Georg/Witt, Andreas/Zimmermann, Felix (2008): Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation. In: Library Trends 57/1. Urbana-Champaign: University of Illinois, 2008. S. 52-71.

Lehmberg, Timm/Chiarcos, Christian/Rehm, Georg/Witt, Andreas (2007): Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In: Rehm, Georg/Witt, Andreas/Lemnitzer, Lothar (Hrsg.):Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Proceedings of the Biennial GLDV Conference 2007. Tübingen: Narr, 2007. S. 93-102. IDS-Publikationsserver

Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG) vom 1. September 2017. BT-Drs. 18/13014.

Data models for Digital Editions: Complex XML versus Graph structures

Bruder, Daniel
dmb77@cam.ac.uk
University of Cambridge, Vereinigtes Königreich

Teufel, Simone
sht25@cl.cam.ac.uk
University of Cambridge, Vereinigtes Königreich

In terms of longevity and collation of textual data in the humanities, digital data, notwithstanding its potential, still falls short the qualities of the traditionally printed book.

To streamline the diverse and idiosyncratic Digital Editions of the time and to establish a cross- and re-usable, durable digital archive of textual cultural artifacts, in 1988 the Text Encoding Initiative (TEI) was established with the goal to present a commonly shared standard for the transcription of literary, scientific and other forms of text.

As data model, the extensible markup language XML was chosen to assure longevity and exchangeability of the data. However, it turns out that XML, and with it, the data model of the hierarchically ordered tree are questionable choices for the recording of complex texts – as they are commonly found in the humanities – by potentially rendering the data ambiguous on semantic level.

The abstract idea behind the commonly shared tag set for the description of textual data is reflected in the *TEI abstract model* (TEI Consortium 2016b) which uses XML as a serialisation format – but to which it is not bound:

> The rules and recommendations made in these Guidelines are expressed in terms of what is currently the most widely-used markup language for digital resources of all kinds: the Extensible Markup Language (XML) [...]. However, the TEI encoding scheme itself does not depend on this language [...], and may in future years be re-expressed in other ways as the field of markup develops and matures.

In the following, fundamental limitations of the tree data model are highlighted in spotlight fashion and contrasted with a graph based model for the sustainable recording and long-term archiving of complex textual data.

Limitations of the tree model

Paradoxically, Digital Editions as well as digital archives, tools, platforms and data repositories are not as interoperable in practice as one would theoretically expect from standardised sources. To be able to cross- or re-use data or tools between projects, in practice, serious refactoring and rededication is necessary – e.g. existing web platforms cannot readily be re-used by another project, notwithstanding the fact that the data repositories are fully validating, validating TEI-P5 sources. How is this possible?

As will be shown, this paradoxical situation of factually unattainable interoperability of editions and tools are a direct consequence of the choice of data model.

The decision towards XML and the tree data model is based on the OHCO assumption of text as an Ordered Hierarchy of Content Objects (DeRose et al. (1990); revised in Renear, Mylonas, and Durand (1993)). Contrasting the original goals (TEI Consortium 2016c) of interoperable long-term archivable data repositories with the status quo, this decision towards XML as the serialisation format needs to be critically questioned – particularly since the TEI Guidelines themselves very early on make clear that the assumption of data model be-

hind XML is an improper simplification (TEI Consortium 2016a):

> Surprisingly perhaps, this grossly simplified view of what text is […] turns out to be very effective for a large number of purposes. It is not, however, adequate for the full complexity of real textual structures, for which more complex mechanisms need to be employed.

Already two most basic constellations can lead to a necessary departure from the tree paradigm which could be described as 'Complex XML'.

These situations are commonly resolved by using workarounds (TEI Consortium 2016d). Although *syntactically* permissible on the level of XML markup, these workarounds establish structures beyond the data model of the tree and can lead to misrepresentation of the data on *semantic*, modelling level, seriously harming effective re-use and long-term archiving.

- Data as well as tools inevitably become idiosyncratic, i.e. they irrevocably need to be handled on individual, project-specific basis; projects increasingly develop 'private dialects' and couple philologists and data scientists for actually accessing the data; data and tools are inaccessible to cross- and re-use between projects; finally, the possibility of a common digital archive is lost beyond recall.
- Complex textual structures demand additional annotation to help and guide downstream tooling to not misrepresent the data. The transcription – in spite of valid, conforming data w.r.t. to the XML Schema – cannot automatically, i.e. without human intervention, be unambiguously resolved into its textual variants.
- The necessary supplementary annotation to one-unambiguously describe and model the source sets in motion a vicious circle of exponentially growing complexity in the data. Project-specific, idiosyncratic tools become necessary and must match this complexity. Moreover, such repositories typically suffer from overtagging (Hanrahan 2015), or, in the worst case need to be abandoned entirely (Schmidt et al. 2006).
- Any further annotation or commentary only ever increases the complexity: any further annotation must match the existing complexity of the amended tree structure to accordingly be integrated; data and tools suffer from a 'Heisenberg-Effect' in that any further, more precise description of the source makes the data only ever more imprecise.

Complex XML

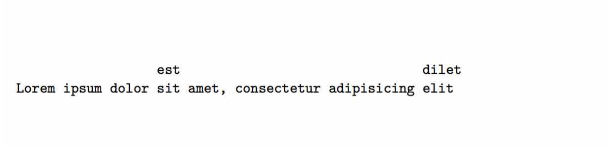
In contrast to a simple edition, i.e. one of linear text without any further annotation, the need for 'Complex XML', on most fundamental, level arises through:

1. the edition of a non-linear text
2. the edition of a linear text, open for annotation

In essence, anything that is beyond linear text free of annotation cannot adequately be represented by a mono-hierarchical tree model and will need "more complex mechanisms" (TEI Consortium 2016a).

Complex XML through non-linear text

Non-linear text results from editorial operations such as insertions, deletions, substitutions. For instance, recording the genealogical writing process of two undecided variants within the same sentence, yields four different, non-linear potential readings.



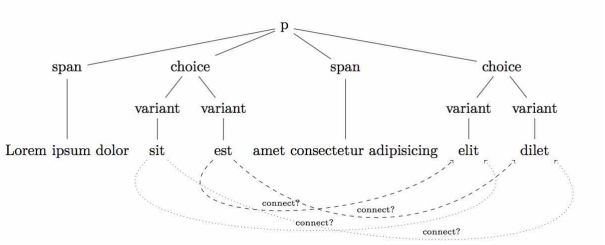
These four different readings derived from mechanical re-combination potentially are not intended and to be reduced to specific readings only.

```
> Lorem ipsum dolor sit amet, consectetur adipisicing elit
> Lorem ipsum dolor est amet, consectetur adipisicing elit
> Lorem ipsum dolor sit amet, consectetur adipisicing dilet
> Lorem ipsum dolor est amet, consectetur adipisicing dilet
```

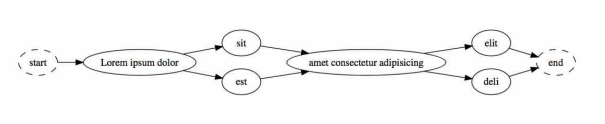
Constraining these combinatorial permutations cannot be done in general ways within the mono-hierarchical tree data model. The tree model exposes a general limitation – even without the prevalence of overlapping structures.

```
<p>
  Lorem ipsum dolor
  <choice>
    <variant id="1-a" connect-with="...">sit</variant>
    <variant id="1-b" connect-with="...">est</variant>
  </choice>
  amet, consectetur adipisicing
  <choice>
    <variant id="2-a">elit</variant>
    <variant id="2-b">dilet</variant>
  </choice>
</p>
```

While interconnecting nodes across the tree's boundaries by (ab-)using attributes is syntactically possible it nevertheless makes the data idiosyncratic on semantic level, i.e. project-specific rules are introduced and must individually be followed when working with the data.



These interconnections to constrain the combinatorics to specific readings cannot formally be made part of the tree structure itself. To build a tree, any node in the tree must have exactly one parent. A different data model and data structure is necessary to model more than one parent for one node, namely the data model of the graph.



Complex XML through meta-data

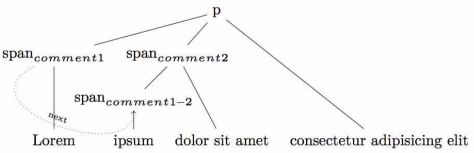
Complex XML can also result from linear text, open for annotation. The following schematic example shows a linear text with overlapping annotation:

```
Lorem ipsum dolor sit amet, consectetur adipisicing elit
`---------' comment1
      `------------------' comment2
```

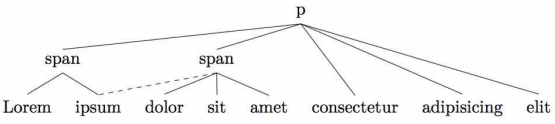
Corresponding serialisation using XML and the segmentation method (TEI Consortium 2016d):

```
<p>
<span id="comment1" next="comment1-2">Lorem</span>
  <span id="comment2">
    <span id="comment1-2">ipsum</span>
    dolor sit amet
  </span>
, consectetur adipisicing elit
</p>
```

The necessary interconnection and recombination of fragmented nodes cannot be modelled within the tree structure in general ways:



Another representation shows how one node in the tree is made the child of two parents:



The relationship between graphs and trees

Trees and graphs are closely related: An ordered tree is a special form of graph with the properties of *a)* it is a directed graph without cycles, *b)* has one designated root node and *c)* any node has exactly one parent node.

As was shown in the previous basic examples, there is strictly no possibility to interconnect nodes of the tree across branches of the tree. By trying to associate two parents to one node, the tree paradigm is effectively abandoned, and results in a permanent need for case-specific handling to resolve potential ambiguities in the data.

Conclusion

Digital Editions wanting to model more than just simple structures can – notwithstanding the syntactical possibilities of XML – not be represented in interoperable ways within the paradigm of the tree data model, making longevity and uniformly re-usable digital archives impossible.

Alternative, graph-theoretic attempts to solve this problem have been suggested and could implement the *TEI abstract model* through an adequate data structure (Huitfeldt 1994; Barnard et al. 1995; Sperberg-McQueen and Huitfeldt 2000; Huitfeldt and Sperberg-McQueen 2001; Durusau and O'Donnell 2002; Tennison and Piez 2002; Dipper 2005; Dekhtyar and Iacob 2005; Banski and Przepiórkowski 2009; Di Iorio, Peroni, and Vitali 2010; Di Iorio, Peroni, and Vitali 2011; Schmidt and Colomb 2009; Schmidt 2014; Götze and Dipper 2006; Peroni, Vitali, and Di Iorio 2009; Witt 2007; Kuczera 2016).

Yet, the question of an adequate serialisation and exchange format to any such data structure remains open. To be able to give guarantees of long term storage and archiving, any such serialisation format must be able to one-unambiguously represent the source as well as data structure. Ideally, any such serialisation format should be both machine readable as well as human intelligible and independent of existing computer hardware and software.

Previous graph-based approaches for the recording of complex textual data either did not catch on or have been abandoned for reasons of complexity in implementation or usage.

Because of the choice of data model, current repositories are idiosyncratic and tools and data must be handled on individual basis. In order to be able to build general digital archives fully interoperable data repositories are necessary. Interoperability is closely connected to the choice of data model. The TEI abstract model should be implemented as a graph structure, however, the graph structure is in need of a suitable exchange and serialisation format.

The commonly shared property between former graph-based approaches is the use of embedded markup. It is conjectured that future research on suitable serialisation formats for graph-based approaches should re-evaluate standoff based markup for the durable recording of Digital Editions.

Bibliography

Banski, Piotr, and Adam Przepiórkowski. 2009. “Stand-Off TEI Annotation: The Case of the National Corpus of P *olish.” In Proceedings of the Third Linguistic Annotati*on Workshop, 64–67. Association for Computational Linguistics.

Barnard, David T, Lou Burnard, Jean-Pierre Gaspart, Lynne A Price, CM Sperberg-McQueen, and Giovanni Battista Varile. 1995. “Hierarchical Encoding of Text: Technical Problems and SGML Solutions.” In *Text Encoding Initiative*, 211–31. Springer.

Dekhtyar, Alex, and Ionut E Iacob. 2005. “A Framework for Management of Concurrent XML Markup.” *Data & Knowledge Engineering* 52 (2). Elsevier:185–208.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. “What is text, really.” *Journal of Computing in Higher Education* 1 (2). Springer Nature:3–26. https://doi.org/10.1007/bf02941632 .

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali. 2010. “Handling markup overlaps using OWL.” In *Knowledge Engineering and Management by the Masses*, 391–400. Springer.

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali. 2011. “A Semantic Web Approach to Everyday Overlapping Markup.” *Journal of the American Society for Information Science and Technology* 62 (9). Wiley Online Library:1696–1716.

Dipper, Stefanie. 2005. “XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation.” In *Berliner XML Tage*, 39–50.

Durusau, Patrick, and M Brook O'Donnell. 2002. “Concurrent Markup for XML Documents.” In *Proc. XML Europe*.

Götze, Michael, and Stefanie Dipper. 2006. “ANNIS: Complex Multilevel Annotations in a Linguistic Database.” In *Proceedings of the 5* th *Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, 61–64. Association for Computational Linguistics.

Hanrahan, Elise. 2015. “Over-Tagging with XML in Digital Scholarly Editions.” In *DHd2015 Conference – von Daten Zu Erkenntnissen. Book of Abstracts.*, edited by Various, 162–65. Graz, Austria.

Huitfeldt, Claus. 1994. “Multi-Dimensional Texts in a One-Dimensional Medium.” *Computers and the Humanities* 28 (4-5). Springer:235–41.

Huitfeldt, Claus, and CM Sperberg-McQueen. 2001. “TexMECS: An Experimental Markup Meta-Language for Complex Documents.” *URL Http://Www. Hit. Uib. No/Claus/Mlcd/Papers/Texmecs. Html*.

Kuczera, Andreas. 2016. “Digital Editions Beyond Xml – Graph-Based Digital Editions.” In *Proceedings of the 3* rd *Histoinformatics Workshop on Computational History (Histoinformatics 2016)*, edited by Johannes Preiser-Kappeller Marten Düring Adam Jatowt.

Peroni, Silvio, Fabio Vitali, and Angelo Di Iorio. 2009. “Towards markup support for full GODDAGs and beyond: the EARMARK approach.” https://doi.org/10.4242/BalisageVol3.Peroni01 .

Renear, Allen H, Elli Mylonas, and David Durand. 1993. “Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies.” Oxford University Press.

Schmidt, Desmond. 2014. “Towards an Interoperable Digital Scholarly Edition.” *Journal of the Text Encoding Initiative*, no. 7. Text Encoding Initiative Consortium.

Schmidt, Desmond, and Robert Colomb. 2009. “A Data Structure for Representing Multi-Version Texts Online.” *International Journal of Human-Computer Studies* 67 (6). Elsevier:497–514.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. “Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources.” In *6* th *E-Meld Workshop, Ypsilanti*.

Sperberg-McQueen, C Michael, and Claus Huitfeldt. 2000. “GODDAG: A Data Structure for Overlapping Hierarchies.” In *Digital Documents: Systems and Principles*, 139–60. Springer.

TEI Consortium. 2016a. “A Gentle Introduction to XML” In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG152 .

———. 2016b. “About These Guidelines.” In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html .

———. 2016c. “Design Principles.” In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html#ABTEI2 .

———. 2016d. “Non-Hierarchical Structures.” In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html .

Tennison, Jeni, and Wendell Piez. 2002. “The Layered Markup and Annotation Language (LMNL).” In *Extreme Markup Languages*.

Witt, Andreas. 2007. “Guideline: Multiple Hierarchies.” In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Der Sammlung gerecht werden: Kritisch-generative Methoden zur Konzeption experimenteller Visualisierungen

Dörk, Marian
doerk@fh-potsdam.de
Fachhochschule Potsdam, Deutschland

Glinka, Katrin
k.glinka@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz, Deutschland

Einleitung

Das Versprechen hinter Digitalisierungsprojekten in sammelnden Institutionen ist oft die Erweiterung des Zugangs zum kulturellen Erbe, sei es für die Forschung oder im Sinne der Vermittlung. Bei kritischer Betrachtung fast aller Benutzerschnittstellen für Sammlungen scheint es aber an Ansätzen zu fehlen, reichhaltige Informationsräume einladend bereitzustellen. Diese Diskrepanz zwischen Digitalisierung von Sammlungen und ihrer digitalen Verfügbarmachung lässt sich dadurch begründen, dass Kultureinrichtungen selten über die nötige Kapazität verfügen, eigenständig Benutzerschnittstellen zu konzipieren und umzusetzen. Die Zielstellung des dreijährigen Forschungsprojektes *Visualisierung kultureller Sammlungen* (VIKUS) an der Fachhochschule Potsdam [1] war die Erforschung graphischer Benutzerschnittstellen zur explorativen Sichtung von Kulturobjekten. Im Projekt wurden in Kooperation mit Kultur- und Technologiepartnern Erkenntnisse zur visuellen Exploration digitalisierter Sammlungen gewonnen (Glinka et al. 2017a), die Entwicklung nachhaltiger Technologielösungen behandelt (Glinka et al. 2017b) und in experi-

mentellen Settings kritisch-generative Methoden erprobt. Zu letzterem zählt die Übertragung der Forschungsfragen in den Kontext eines interdisziplinären Lehrformats, welches wir in unserem Beitrag diskutieren. Wir gehen insbesondere auf den produktiven Zusammenhang zwischen Kritik und Konzeption von Informationsvisualisierungen ein und zeigen auf, welche Potenziale ein bewusster Bruch mit disziplinären Konventionen und Sehgewohnheiten mit sich bringt.

Hintergrund und Herangehensweise

Mit der Digitalisierung von Sammlungen erhält eine breite Öffentlichkeit Zugriff auf zahlreiche Kulturobjekte, welche zuvor hauptsächlich für Wissenschaftler*innen zugänglich waren. Dieser Zugriff basiert häufig auf digitalisierten Museumskatalogen oder Archivsystemen, welche ursprünglich als Bestandsnachweis und der Ortung physischer Originale und nicht als eigenständige „Repräsentation“ der Sammlungsobjekte dienten. Das Forschungs- und Lehrprojekt VIKUS stellte sich der Frage, wie die digitale Repräsentation als eine für sich stehende Perspektive auf Sammlungen zu begreifen sein könnte und welche Interfacekonzepte diese unterstützen würden. Hier eröffnet sich die Gelegenheit, die Stärken des Digitalen bei der Bereitstellung kultureller Sammlungen zu berücksichtigen. So kann zum Beispiel die vergleichsweise statische Anordnung in Ausstellungen in digitalen Benutzeroberfllächen mittels dynamischer Arrangements durchbrochen werden. Obwohl die Auswahl und Anordnung von Objekten bei der Gestaltung von Ausstellungsräumen große Aufmerksamkeit erfährt, werden diese Überlegungen im digitalen Kontext häufig noch vernachlässigt. Vor kultureller Intention stehen zumeist technische Konventionen, die dem vielschichtigen Gehalt der Sammlung kaum entsprechen.

Der Ansatz dieses Projekts liegt in der Verknüpfung technologischer Möglichkeiten mit kulturwissenschaftlichen Überlegungen, um kritisch-generative Methoden zur Visualisierung zu entwickeln, welche alternative Perspektiven auf Kultursammlungen eröffnen. Dabei sind solche Visualisierungen ebenso als Kulturartefakte zu betrachten, die es in ihrer Funktion nicht nur zu konstruieren, sondern ebenso zu kritisieren gilt, wobei der »performative Charakter der Interpretation« (Drucker 2013) insbesondere bei *Humanities Interfaces* relevant wird. Entlang der für das Projekt zentralen disziplinären Perspektiven—Interface Design, Information Retrieval und Digital Humanities—haben sich drei Fragestellungen herausgebildet:

1.) Wie können Visualisierungen der kulturellen Signifikanz einer Sammlung gerecht werden? Bislang orientieren sich digitale Zugänge zu Kultursammlungen eher an technischen Gegebenheiten von Datenbanklösungen als an der kulturellen Bedeutung der Objekte. So werden diese herkömmlicherweise in einer tabellarischen Auflistung in Anlehnung an einen Leuchttisch gezeigt. Statt solche Standardlösungen auf alle Arten von Sammlungen anzuwenden, untersuchen wir, inwiefern die spezifischen Eigenschaften einer Sammlung reflektiert werden können.

2.) Wie kann offenes Stöbern in komplexen Informationsräumen angeregt werden? Herkömmliche Sammlungsinterfaces sind auf gezielte Suche mit expliziter Anfrageformulierung ausgerichtet, was eine von Neugier getriebene Exploration erschwert. Als Gegenentwurf zu solch „geizigen” Zugängen sind „freigebige” Oberflächen vonnöten (Whitelaw 2015), die durch visuelle Arrangements der Sammlungsobjekte entlang ihrer Facetten die Benutzer*innen zum Stöbern einladen. Diesem Anspruch folgend entwickelt das VIKUS-Projekt Szenarien für die offene Exploration digitaler Sammlungen weiter.

3.) Wie können digitale Methoden die Analyse visueller Sammlungen unterstützen? Als Ergänzung zur Forschung an textbasierten Quellen in den Digital Humanities widmen wir uns reichhaltigen Sammlungen, in denen bildliche Aspekte einen wichtigen Stellenwert einnehmen und erproben die Übertragung des Konzepts „Distant Reading“ (Moretti 2005) auf andere geisteswissenschaftliche Disziplinen. Über das interessierte Stöbern hinaus kann Visualisierung auch für Expert*innen die Sichtung von Kulturobjekten und deren Analyse entlang verschiedener Fragestellungen unterstützen (Yamaoka et al. 2011).

Interdisziplinäre Projektkurse

Um diesen Fragestellungen nachzugehen, wurde im Laufe des Forschungsprojekts die Herangehensweise der iterativen Gestaltung verfolgt. Dazu zählen der Einsatz von Co-Creation, stufenweise Ideenentwicklung im interdisziplinären Austausch und das Durchführen von Nutzerstudien (Glinka et al. 2017a). Insbesondere durch die Einbindung von Studierenden und Projektpartnern im Rahmen von Workshops (Chen et al. 2014) und Projektkursen hat sich das Spektrum an Erkenntnissen signifikant erweitert. Im Folgenden gehen wir auf das Kursformat ein und umreißen die Ergebnisse.

Methodik

Die Zielstellung des interdisziplinären Projektkurses, welcher seit 2014 bereits vier mal stattfand, folgt der umrissenen Herangehensweise des Forschungsprojektes. In interdisziplinären Teams erforschen fortgeschrittene Studierende aus Design, Kulturarbeit, Europäischer Medienwissenschaft und Informationswissenschaften innovative Darstellungsformen zur explorativen Sichtung von Sammlungen. Kursteilnehmer*innen verfügen über Grundlagen und Erfahrungen in mindestens einem der Kernthemen des Kurses—Informationsvisualisierung und Kultursammlungen—und haben Interesse an dem jeweilig anderen. Der Kurs verfolgt die Ideen des fächerübergreifenden »Forschenden Lernens«, indem aktuelle Forschungsfragen das Kursthema motivieren und Interdisziplinarität neue Erkenntnisse verspricht.

Die Studierenden nähern sich dem Kursthema aus ihrem jeweiligen disziplinären Hintergrund, entwickeln selbständig Fragestellungen und bringen ihre fachliche Perspektive ein. Die Kursprojekte werden in Teams von 2-4 Studierenden bearbeitet, welche jeweils aus verschiedenen Studiengängen kommen müssen. Es findet anfangs eine kritische Auseinandersetzung mit existierenden Sammlungszugängen statt. Nach drei Vorlesungsterminen, in denen die zentralen Konzepte und Forschungsfragen des Kurses vorgestellt werden, stellen alle Studierende Webseiten ihrer Lieblingsmuseen vor und untersuchen diese in Hinblick auf funktionale und ästhetische Gestaltungselemente. Durch die Analyse von existierenden Sammlungsinterfaces wird ein kritisches Bewusstsein geschult, welches den Studierenden bei ihrer eigenen Projektentwicklung zu Gute kommt. Um der Forschungsfrage nach adäquaten, ansprechenden und aktuellen Repräsentation digitaler Sammlungen an konkreten Beispielen nachzugehen, stehen dem Kurs als Datengrundlage eine Auswahl an digitalisierten Beständen zur Verfügung, welche die Kooperationspartner zur Verfügung stellen. Zu den Datensätzen zählten bislang u.a. Sammlungen der *Stiftung Preußische Schlösser und Gärten (SPSG)*, der *Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW)*, des *Syrian Heritage Archives*, des *Deutschen Forums für Kunstgeschichte Paris (DFK)* und des *Münzkabinetts der SMB* (siehe https://uclab.fh-potsdam.de/vikus/). Über die freie Verfügbarmachung der Daten hinaus beteiligen sich Mitarbeiter*innen der jeweiligen Partnerinstitutionen an Workshops, Austauschtreffen und Präsentationen und unterstützen die studentischen Projektteams mit regelmäßigem Feedback, zusätzlichem Material und wissenschaftlicher Beratung. Jedes Forschungsprojekt umfasst die klassischen Forschungsphasen: von der Entwicklung einer Forschungsfrage zur Literaturrecherche, Entwicklung einer Methodik bis zur Diskussion und Präsentation der Ergebnisse.

Im größeren Rahmen der Forschungsthematik „Visualisierung kultureller Sammlungen“ erfolgt die Konkretisierung der Themen, Fragestellungen und Methoden durch die Teilnehmer*innen in Absprache mit dem Dozenten. Anstelle von frontaler Wissensvermittlung wird versucht, einen konstruktiven und kritischen Möglichkeitsraum für Forschung und Studium aufzuspannen. Dafür werden auch Professor*innen aus den Fachbereichen als Gastkritiker*innen in den Kurs eingeladen, um Feedback bei Zwischen- und Endpräsentationen zu geben. Wichtiges Element des selbstorganisierten Lehrformats sind von Studierenden vorbereitete Methodeninputs, welche bezugnehmend auf die konkreten Anforderungen der Projekte relevante Techniken und Theorien vorstellen. Die vorgestellten Methoden in den vergangenen Kursen reichten von praktischen Werkzeugen zur Entwicklung von Interfaceprototypen bis zu theoretischen Konzepten zu Fotografie, Kuratierung und Interfacekritik.

Ergebnisse

Die Kursergebnisse reichen von Visualisierungskonzepten über Studien zur Wireframe-Analyse von Sammlungswebseiten (Kreiseler et al. 2017) bis hin zu funktionstüchtigen Webprototypen.

Eines der hervorzuhebenden Kursergebnisse ist ein Konzept zu einem Datenbestand der BBAW, welches mit seiner reduzierten Gestaltung und seiner Fokussierung auf Zeitgenossenschaft bewusst mit klassischen Archiv-Interfaces bricht (siehe Abb. 1). Anstatt die einzelnen Lebensdaten der Personen als Liste darzustellen, wird ein zeitlich geordneter Überblick geboten, der die Beziehungen zwischen Personen der »Berliner Klassik« in den Vordergrund rückt.

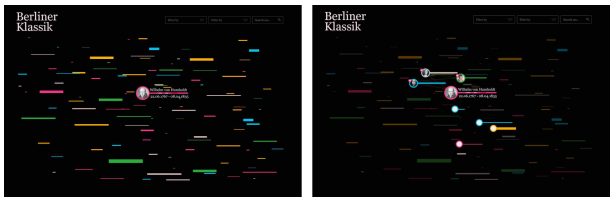


Abb. 1: Biographische Textdaten der Berliner Klassik,BBAW (Sebastian Schuth, Tatjana Tsĕrnõhh, Andreas Waleczek, Alexander Zöller).

Ein Projekt zur ostasiatischen Porzellansammlung der SPSG hatte zum Ziel, sowohl die Exploration der Bestände aus wissenschaftlichem Interesse zu unterstützen als auch kontextualisierende Narrative über die Entstehung der Objekte und der Sammlung anzubieten (siehe Abb. 2). Statt ausschließlich Informationen zu den Objekten in Form von Bild- und Metadaten zugrunde zu legen, haben die Studierenden gemeinsam mit den Kuratorinnen der Sammlung illustrierende Inhalte textuell und visuell erarbeitet, wodurch sie eine Vermittlungsebene in die wissenschaftlich erschlossenen Sammlungsdaten integrierten.

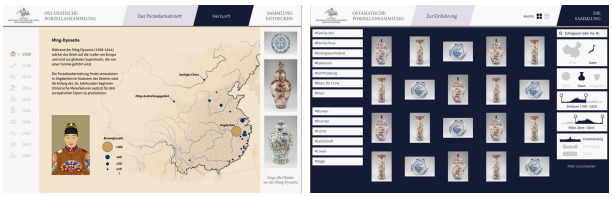


Abb. 2: Narrative und explorative Visualisierung einer Porzellansammlung der SPSG (Jana Klausberger, Mark-Jan Bludau, Swann Nowak, Constantin Eichstaedt).

Eine interaktive Visualisierung des Münzkabinetts zeigt, wie das Loslösen von Darstellungskonventionen dazu beitragen kann, auf spielerische Art Erkenntnisse aus einem für Laien eher schwer zugänglichen Bestand zu gewinnen (siehe Abb. 3). Sowohl im physischen Ausstellungskontext als auch im Datenbankinterface werden Münzen zumeist in rigiden Tableaus dargestellt. Die Projektgruppe näherte sich dem numismatischen Bestand über ihren alltäglichen Blick auf das Material, nämlich über physische Haufen unterschiedlicher Münzen, in denen das Einzelstück zunächst untergeht. Die daraus resultierende Visualisierungsumgebung erlaubt es, aus haptisch anmutenden Arrangements über verschiedene Darstellungsmodi die Sammlung als Ganzes nach verschiedenen Merkmalen zu filtern und zu sortieren. Es können Zusammenhänge zwischen Prägungsort, Material, zeitlichen Verläufen und anderen Facetten in verschiedenen organischen Layouts untersucht werden.

Abb. 3: Visualisierung der numismatischen Sammlung des Berliner Münzkabinetts [2] (Flavio Gortana, Daniela Guhlmann, Franziska von Tenspolde).

Diskussion

Disziplinäre Konventionen und die „digitale Vernunft“ in Sammlungsinstitutionen legen bei Digitalisierungs- und Erschließungsprojekten häufig den Fokus auf die strukturierte und möglichst vollständige wissenschaftliche Erfassung von Metadaten und die Entwicklung von Datenbankstrukturen, welche meist für interne Prozesse optimiert wird. Obwohl dieses Vorgehen im Sinne einer infrastrukturellen Entwicklung von digitalen Forschungsumgebungen zu Recht weite Verbreitung findet, werden auf diesem Wege kritische und gestalterische Ansätze nicht begünstigt. In den ersten Beschäftigungen mit den in den Kurs eingebrachten Datensätzen wird angeregt, dass sich die Studierenden einerseits „sensibel“ mit den Spezifika der Sammlungen auseinandersetzen und sich in einen engen Austausch mit den Wissenschaftler*innen der datengebenden Institutionen begeben, gleichzeitig aber im Sinne eines kritisch-generativen Annäherns auch neue und möglicherweise „ungewöhnliche“ Sichten auf die Sammlungen in Betracht ziehen. In den einführenden Sitzungen finden daher ebenso Impulsvorträge zur musealen Praxis des Ausstellens, zur Infragestellung disziplinärer Deutungshoheit und zur Diskrepanz zwischen Materialität von Sammlungen und deren Distanz schaffenden Präsentationsformen in Vitrinen, Schaukästen oder gesicherten Displays statt. Nach teilweiser Skepsis vonseiten der Institutionen am Anfang der Zusammenarbeit mit den Projektgruppen führt diese kritisch-generative Annäherung jedoch immer dazu, dass auch die Expert*innen neue Blicke auf „ihre“ Sammlungen gewinnen können. Einige der Projekte werden, auch auf Wunsch der Sammlungsinstitutionen, über das Ende des Kurses hinweg weiterentwickelt und zeugen somit davon, dass alternative Entwürfe zur gängigen Darstellungspraxis und das Hinterfragen des Status Quo für alle Seiten ein Zugewinn bieten kann: für die Expert*innen in den Sammlungsinstitutionen, den mit den Sammlungen arbeitenden Wissenschafter*innen, der breiteren Öffentlichkeit, welche über interaktive und anregende Visualisierungen Zugang zu den Sammlungen erhält und für die Studierenden, die selbständig neuartige Zugänge zu spannenden Beständen entwickeln und erforschen.

Fußnoten

1. https://uclab.fh-potsdam.de/project/vikus/
2. https://uclab.fh-potsdam.de/coins

Bibliographie

Chen, Ko-Le / Dörk, Marian / Dade-Robertson, Martyn (2014): „Exploring the promises and potentials of visual archive interfaces“. In: Proceedings of the 2014 iConference. iSchools.

Drucker, Johanna (2013): „Performative Materiality and Theoretical Approaches to Interface“. In: DHQ: Digital Humanities Quarterly, 7(1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html [letzter Zugriff 25. September 2017].

Glinka, Katrin / Pietsch, Christopher / Dörk, Marian (2017a): „Past visions and reconciling views: Visualizing time, texture and themes in cultural collections“. In: DHQ: Digital Humanities Quarterly, 11(2). http://www.digitalhumanities.org/dhq/vol/11/2/000290/000290.html [letzter Zugriff 25. September 2017].

Glinka, Katrin / Pietsch, Christopher / Dörk, Marian (2017b). „Von sammlungsspezifischen Visualisierungen zu nachnutzbaren Werkzeugen“. In: Konferenzband zur DHd 2017 Bern - Digitale Nachhaltigkeit.

Moretti, Franco (2005). Graphs, Maps, Trees: Abstract models for a literary history. Verso.

Kreiseler, Sarah / Brüggemann, Viktoria / Dörk, Marian (2017). „Tracing exploratory modes in digital collections of museum web sites using reverse information architecture“. In: First Monday, 22(4). http://firstmonday.org/ojs/index.php/fm/article/view/6984 [letzter Zugriff 25. September 2017].

Yamaoka, So / Manovich, Lev / Douglass, Jeremy / Kuester, Falko (2011). „Cultural analytics in large-scale visualization environments“. In: Computer, 44(12):39–48.

Whitelaw, Mitchell (2015). „Generous interfaces for digital cultural collections“. digital humanities quarterly, 9(1). http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html [letzter Zugriff 25. September 2017].

Die Angst vor dem „Elektronengehirn“: Topoi der Kybernetik-Kritik in der bundesdeutschen Nachkriegsphilosophie

Heßbrüggen-Walter, Stefan
shessbru@hse.ru
National Research University Higher School of Economics, Russland

Mein Beitrag greift einen Diskurs auf, der meines Wissens selbst in der Wissenschaftsgeschichte der Kybernetik noch keine Beachtung gefunden hat und erst recht für die ‚Vorgeschichte‘ der digital humanities im deutschsprachigen Raum noch nicht ausgewertet worden ist: die philosophische Kritik der Kybernetik in der Bundesrepublik der 60er Jahre. Ich beschränke mich dabei exemplarisch auf die Analyse dreier Aufsätze, die in der Zeitschrift für philosophische Forschung zwischen den Jahren 1965 und 1970 veröffentlicht worden sind. Die Reihe eröffnete der Gründer und Herausgeber der Zeitschrift Georgi Schischkoff (Schischkoff 1965). Der zweite Text wurde vom in Karlsruhe lehrenden österreichischen Philosophen Simon Moser verfasst (Moser 1967). Sein Tübinger Kollege Walter Gölz äußerte sich abschließend (Gölz 1970). Mein Beitrag soll zur Genealogie heutiger Kritik digitaler Geisteswissenschaft beitragen.

In der Kybernetik wurden Regelungstheorie, Informationstheorie und Theorie der Nachrichtenverarbeitung zusammengeführt (Steinbuch 1963: 317). Sie sollte als „zukünftige Universalwissenschaft“ (Steinbuch 1963: 340) den „Weg zu einer neuen Einheit der Wissenschaften“ (Steinbuch 1963: 319) bahnen, indem sie Modelle und Erklärungsansätze der Technik auf die Erklärung von Lebewesen, insbesondere des Menschen, und Gesellschaften überträgt (Kline 2015: 11-12). Jedenfalls zielte sie auf den Abbau von Barrieren zwischen Disziplinen und Verbesserung des Austauschs zwischen ihnen beitragen (KLine 2015: 63). Der institutionelle Erfolg der Kybernetik trug jedoch zur Verwischung ihres Profils bei. Zugleich wuchs der Zweifel, ob ihre Versprechungen überhaupt einlösbar erschienen (Kline 2015: 179-181, Aumann 2015: 25).

Ähnlich wie die digitalen Geisteswissenschaften nutzte die Kybernetik also formalwissenschaftliche Werkzeuge zur Behandlung von Fragen, die auf den ersten Blick einer solchen Behandlung nicht zugänglich sind (Steinbuch 1963, 317). Hierzu zählten Analysen der Wahrnehmung ästhetischer Information (Steinbuch, 1963, 280) genauso wie der Plan einer formalen, aber qualitativen Informationstheorie (Steinbuch 1963, 315). Zu verweisen ist auch auf die Affinität zwischen Kybernetik und der Entwicklung strukturalistischer Linguistik, etwa bei Jakobson (Kline 2015, 41). „Das Eindringen der Kybernetik in die Geisteswissenschaft ist ein Markstein in der Geschichte der Wissenschaften." (Steinbuch 1963, 339) In den USA wurden Thesen der Kybernetik auch in der analytischen Philosophie und Wissenschaftstheorie rezipiert, nicht immer positiv (Kline 2015: 98-99).

Die Kybernetik-Kritik bundesdeutscher Philosophen fällt jedoch um einiges grundsätzlicher aus. Zunächst ist auf methodischer Ebene auf die Zuschreibung von ‚Kompetenz-Asymmetrie' hinzuweisen. Der Philosoph darf sich zur Kybernetik äußern, ohne über einschlägiges Fachwissen zu verfügen, denn dieses ist „hauptsächlich von technologischer Natur, und ihre Ausgangspunkte [sc. der Kybernetik] sowie speziell die anthropologischen Überlegungen und Analogien lassen sich im Rahmen allgemeiner philosophisch-methodologischer Interpretationen behandeln." (Schischkoff 1965: 251) Umgekehrt ist dem Kybernetiker die Beteiligung am philosophischen Diskurs zu verwehren, „weil ja die dürftigen philosophischen Voraussetzungen der Kybernetiker für eine Interpretation auf breiter Basis naturgemäß nicht ausreichen" (Schischkoff 1965: 251). Ergänzend hinzu tritt die Argumentationsfigur der ‚Schuld durch Assoziierung'. Durch Reduktion kybernetischer Thesen auf bekannte und im Diskurskontext als widerlegt geltende Positionen wird deren Haltlosigkeit offengelegt. So gilt der in der Kybernetik angeblich vorausgesetzte Physikalismus als „ein Beispiel höchster Steigerung des positivistischen Radikalismus" (Schischkoff 1963: 252). 254-256). Der Materialismus ist falsch: die Psychologie setze Bewusstsein notwendig voraus und könne auf Introspektion nicht verzichten (Moser 1967, 65). Der Begriff des Wissens sei ebenfalls auf Bewusstsein verwiesen und deswegen kybernetischer Analyse prinzipiell unzugänglich (Gölz 1970, 256).

Auf der Sachebene kreisen die Argumente der Kybernetik-Kritik, sofern sie hier einschlägig sind, in der Hauptsache um drei Themen: den Begriff der Information, das Sprachverstehen und die Rolle der Formalwissenschaften Logik und Mathematik. Shannons Informationstheorie verzichtet, darin auch innerhalb der Kybernetik nicht unumstritten, auf jede Einbeziehung der Bedeutung sprachlicher Ausdrücke (Kline 2015: 15). Demgegenüber beharrt die Kybernetik-Kritik auf der Subjektgebundenheit des Begriffs: „Information über etwas gibt es nur von einem bewußten Wesen an ein anderes." (Moser 1965: 66) Sie ist „objektivierter Geist", der immer nur „in bezug auf den Menschen Sinn und Bedeutung hat" (Gölz 1970, 257). Entsprechend erfordern „geistig fundierte Texte" – im Gegensatz zu bloßen inhaltlichen Mitteilungen – die Erfassung durch einen lebendigen Adressaten (Schischkoff 1965, 259). Eigentliches Lehren bedürfe der „lebendigen Sprache des Lehrers",die „einen eigenen bildenden Wert hat" (Schischkoff 1965, 267). Die Überschätzung der formalen Wissenschaften führe schließlich zu einem übersteigerten Rationalismus: die Kybernetik übersehe, dass sich die „innere geistig unerschöpfliche Sphäre des eigentlichen menschlichen Seins" nicht in „rational erfaßbare[n] Strukturen" abbilden lasse (Schischkoff 1965: 262). Die Mathematisierung der Kybernetik erzeuge „die Gefahr einer formallogischen und formal mathematischen Verdünnung des System- und Modellbegriffes gegenüber den materialkonkreten Bedürfnissen der Physik, Physiologie und Technik" (Moser 1967: 67). Oder nach Gölz: „Die Fähigkeit der Maschinen, solche – im weitesten Sinne! – mechanischen Denkprozesse zu bewältigen, bestätigt aber nicht den ‚Geist' der Maschinen, sondern den mechanischen Charakter gewisser formaler Denkprozesse." (Gölz 1970: 259)

Ein solcher ‚seelenloser Materialismus' ist nicht nur abstrakt und theoretisch, sondern auch praktisch und politisch gefährlich. Der kybernetische Materialismus stehe im Bunde mit dem „östlichen",also historischen oder dialektischen, Materialismus (Schischkoff 1965: 256). Besondere Aufmerksamkeit erhielt hier die wissenschaftspolitische Dimension. Die Entwicklung der Kybernetik wäre ohne politische Förderung und Patronage nicht möglich gewesen (Kline 2015: 99). Häufig diente der Begriff ‚Kybernetik' als Schlagwort, „um Entscheidungsträgern modernes und zukunftszugewandtes Denken zu demonstrieren." (Aumann 2015: 32). Dies blieb auch ihren philosophischen Kritikern nicht verborgen. Ein großes Risiko bildet hierbei nach Schischkoff die technische Anwendbarkeit der Kybernetik: „Datenverarbeitungsanlagen, die auch als ‚Elektronengehirne' bezeichnet werden, lernende Automaten und Rechenmaschinen sind dafür weitbekannte Beispiele." (Schischkoff 1965: 250) Philosophen hingegen dürften wohl kaum über Patente verfügen (Schischkoff 1965: 250). Es folge vermutlich die „rasche Errichtung von Lehrstellen für Kybernetik, für die sich die finanziel-

len Mittel viel leichter finden, als etwa für die Errichtung neuer geisteswissenschaftlicher und philosophischer Lehrstühle.“ (Schischkoff 1965: 250) Dies gefährde die „bisherige Vordergrundstellung der klassischen Disziplinen des Geistes zumindest hinsichtlich deren praktischer Förderung“ (Schischkoff 1965: 250). Kybernetik schicke sich an, „an Stelle der Geisteswissenschaften und Philosophie treten zu können“.Dies werde zum „Aussterben der geistigen Elite“ führen (Schischkoff 1965: 266). Einzig die Philosophie erscheint einstweilen vor dem Zugriff eines solchen Imperialismus gefeit: „Die Forschungsarbeit tiefschürfenden philosophischen Denkens kann ihrem Wesen nach zum Glück nicht zu einem Spezialistentum der Modelltechnik führen, so daß also ein direkter Verrat des eigenen Faches undenkbar erscheint.“ (Schischkoff 1965: 275).

Die hier referierten methodischen Einwände erscheinen aus heutiger Perspektive sophistisch: eine durchgreifende Kritik digitaler Vernunft ist ohne vertiefte Kenntnis des kritisierten Sachgebiets kaum denkbar. So wäre auch der Schuldzuschreibung durch Assoziierung vorzubeugen: zu behandeln sind die konkreten Erzeugnisse digitaler Forschung, nicht deren vorgeblicher Zusammenhang mit angeblich haltlosen Lehrgebäuden. Bedenkenswert erscheint hingegen weiterhin die Frage, in welchem Ausmaß digitale Forschung in den Geisteswissenschaften die lebensweltliche Verankerung verwendeter Begriffe in Frage stellen darf, wie dies die kybernetische Informationstheorie vorschlug. Umgekehrt muss sich manche Kritik der digitalen Geisteswissenschaften vielleicht die Rückfrage gefallen lassen, inwiefern sie ähnlich wie ihre Vorläufer einem überkommenen Elite-Verständnis anhängt, das die Gabe zu geisteswissenschaftlicher Forschung in der ‚geistig unerschöpflichen Sphäre‘ besonders begabter Individuen verortet.

Auf disziplinpolitischer Ebene erlaubt die hier vorgeschlagene Rekonstruktion ebenfalls einige Schlussfolgerungen. So wie sich der Aufstieg der Kybernetik konkretem politischem Willen verdankte, war auch ihr Abstieg nicht zuletzt der immer größer werdenden Diskrepanz zwischen Anspruch und Wirklichkeit und dem damit einhergehenden Entzug politischen Wohlgefallens geschuldet. Wollen die digitalen Geisteswissenschaften diesem Schicksal entgehen, sollte die Sorge aber nicht allein ihrer materiellen und politischen Basis gelten, sondern auch ihrem theoretischen Überbau. Eine vorurteilsfreie Bestimmung der Grenzen digitaler Vernunft ist hierfür sicherlich ein erster Schritt, eine von gedanklicher Offenheit geprägte Auseinandersetzung über das Verhältnis digitaler und außerdigitaler Forschungspraxen vielleicht der zweite. An beidem hat es in der Auseinandersetzung über die Rolle der Kybernetik sicherlich gemangelt. Es wäre an uns, zumindest diese Fehler nicht zu wiederholen.

Bibliographie

Aumann, Philipp (2015): „Neues Denken in Wissenschaft und Gesellschaft: Die Kybernetik in der Mitte des 20. Jahrhunderts“ in: Jeschke, Sabina / Dröge, Alicia / Schmitt, Robert (eds.): *Exploring Cybernetics: Kybernetik im interdisziplinären Diskurs*, Wiesbaden: Springer Fachmedien 21-40

Gölz, Walter (1970): „Philosophisches Problembewußtsein und kybernetische Theorie“,in: *Zeitschrift für philosophische Forschung* 24: 253-264

Kline, Ronald R (2015): *The Cybernetics Moment. Or Why we Call Our Age the Information Age*. Baltimore: Johns Hopkins University Press

Moser, Simon (1967): „Zur philosophischen Diskussion der Kybernetik in der Gegenwart“,in: *Zeitschrift für philosophische Forschung* 21: 64-77

Schischkoff, Georgi (1965): „Philosophie und Kybernetik. Zur Kritik am kybernetischen Positivismus“,in: *Zeitschrift für philosophische Forschung* 19: 248-278

Steinbuch, Karl (1963): *Automat und Mensch: Kybernetische Tatsachen und Hypothesen*. Berlin / Göttingen / Heidelberg: Springer Verlag OHG

Die guten ins Töpfchen: Zur Anwendbarkeit von Burrows’ Delta bei kurzen mittelhochdeutschen Texten nebst eines Attributionstests zu Konrads ‚Halber Birne‘

Dimpel, Friedrich Michael
mail@dimpel.de
FAU Erlangen-Nürnberg, Deutschland,
Germanistik, und TU Darmstadt,
Computerphilologie und Mediävistik

Einleitung

Die Anwendbarkeit von Burrows' Delta (Burrows 2002) als Autorschaftstest für das Deutsche ist in Validierungstestreihen wiederholt eindrucksvoll demonstriert worden (Büttner et alia 2017, Eder 2013a/b, Evert et alia 2015. Evert et alia 2016); auch im Mittelhochdeutschen ist Delta anwendbar (Dimpel 2016/2018). Die Stabilität des Verfahrens wurde in Noise-Tests belegt: Wenn man etwa 12% aller Wörter durch Fremdmaterial austauscht, sinkt die Erkennungsquote kaum (Dimpel 2017a/2018).

Bei nicht-normalisierten mittelhochdeutschen Texten steigt die Erkennungsquote in einem Validierungstest von 80% auf 91%, wenn man die bei Evert et alia (2016) entwickelte Methode der Z-Wert-Begrenzung mit einem von mir zusammengestellten Normalisierungswörterbuch kombiniert (Dimpel 2017a). Kontraintuitiv ist, dass nur die Kombination dieser Optimierungsverfahren zu einer Verbesserung um 11% führt, während in diesem Setting nur der Einsatz der Z-Wert-Begrenzung zu einer minimalen Verschlechterung führt; der Einsatz nur des Normalisierungswörterbuchs führt nur zu einer Verbesserung um 5,6%. Dieser Befund wird unter dem Stichwort „Delta-Rätsel" in einem Dariah-de-Working-Paper (Dimpel 2017b) ausführlich analysiert. Bei der Rätsel-Analyse wurde – ein Serendipitätseffekt – eine Möglichkeit entdeckt, wie man bei einem konkreten Vergleich von drei Texten die Wortformen identifizieren kann, die eine korrekte Autorschaftserkennung begünstigen oder behindern – dazu im Weiteren.

Gute und schlechte Wortformen

Beim Delta-Test berechnet man aus den Wortfrequenzen für ein Korpus jeweils die zugehörigen Z-Werte. Beim Vergleich von zwei Wortformen aus zwei Texten wird die Differenz der jeweiligen Z-Werte gebildet und der Betrag dieser Differenz genommen. Delta ist schließlich der Mittelwert der absoluten Z-Wert-Differenzen für alle Wortformen.

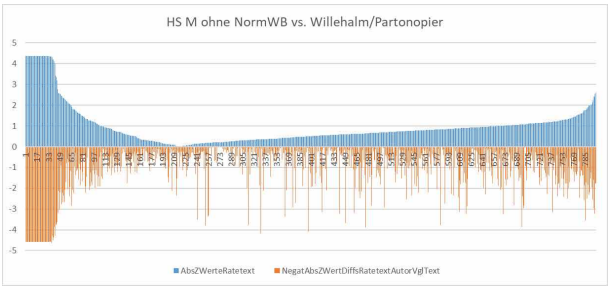


Abb. 1: Ratetext-Z-Werte (blau) sowie Z-Wert-Differenzen Ratetext–Autor-Vergleichstext (orange)

Abb. 1. zeigt oben die Z-Werte der Handschrift M von Wolframs ‚Parzival' in einem Test, in dem sich im Vergleichskorpus neben Wolframs ‚Willehalm' noch weitere 19 Distraktortexte von anderen Autoren befinden (ausführlich zum Testverfahren Dimpel 2017b). Der ‚Parzival' soll dem Autor-Vergleichstext (Wolframs ‚Willehalm') zugeordnet werden und nicht etwa Konrads ‚Partonopier'. Im oberen linken Viertel sind positive Z-Werte blau aufgetragen und nach der Höhe der Z-Werte sortiert. Ab der Stelle, an der die blauen Balken auf 0 zurückgehen, folgt rechts der Betrag der negativen Z-Werte (blau). Unten stehen (orange) die absoluten Z-Wert-Differenzen zwischen dem Ratetext und dem Autor-Vergleichstext (Differenzen der Z-Werte von Wolframs ‚Parzival' und Wolframs ‚Willehalm').

Man könnte A) den Verdacht haben, dass Wortformen bei hohen blauen Balken „gut" sind, um einen Text von Distraktortexten zu unterscheiden, da hohe Z-Werte auf erhebliche Abweichung von den übrigen Korpusfrequenzen hindeuten. Man könnte auch B) den Verdacht haben, dass Wortformen bei hohen orangen Balken „schlecht" für die Autorerkennung sind: Unterschiede zwischen dem Ratetext und Autor-Vergleichstext (also Unterschiede von zwei Texten des gleichen Autors) sollten eher niedrig sein, damit die Erkennung funktioniert. Allerdings sind bei hohen blauen Balken relativ oft auch hohe orange Balken vorhanden – auch in anderen Tests (Dimpel 2017b). Dieses Diagramm erlaubt also keine Aussage darüber, welche Wortformen gut für die Autorerkennung sind; hohe Z-Werte allein erlauben noch keine Aussage darüber, ob ein Wort hier gut geeignet ist, um einen Autor zu charakterisieren.

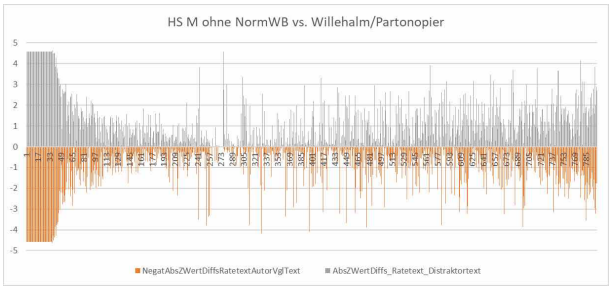


Abb. 2: Z-Wert-Differenzen Ratetext–Autor-Vergleichstext (orange: Wolframs ‚Parzival'– Wolframs ‚Willehalm') und Z-Wert-Differenzen Ratetext–Distraktortext (grau: Wolframs ‚Parzival' – Konrads ‚Partonopier')

Neu ist in Abb. 2 nur die obere Hälfte: Sie enthält Z-Wert-Differenzen des Ratetexts zum Distraktortext (‚Partonopier'). Diese grauen Unterschiede sollten bei funktionierender Autorerkennung eher groß sein; gleichzeitig sollten die orangen Unterschiede der Texte vom gleichen Autor niedriger sein als die grauen. Dort, wo die grauen Balken genauso hoch sind wie die orangen, hilft das Wort nicht bei der Autorerkennung – dies ist bei sehr hohen positiven Z-Werten der Fall. Sind die orangen Balken höher als die grauen, stört die Wortform die Autorerkennung: Die Differenzen zwischen Texten verschiedener Autoren müssen größer sein als die Differenzen zwischen Texten gleicher Autoren, wenn die Autorschaftserkennung funktioniert.

Die Differenz zwischen orange und grau sei ‚Level-2-Differenz' genannt: „Differenz aus der Z-Wert-Differenz zwischen Ratetext und Distraktortext einerseits und der Z-Wert-Differenz zwischen Ratetext und Autor-Vergleichstext andererseits". Bei positiven Level-2-Differenzen ist eine Wortform vorteilhaft für die Autorerkennung – mit Blick auf den einen untersuchten Distraktortext. Bei negativen Level-2-Differenzen ist die Wortform schlecht für die Autorerkennung. Über diese Differenz kann man „gute" und „schlechte" Wortformen einzeln identifizieren.

Use-Case-Szenario ‚Halbe Birne'

Konrads Autorschaft wurde der ‚Halben Birne' trotz Selbstnennung im Epilog (*von Wirzburc maister Kuonrat*) abgesprochen (Lachmann 1820, Laudan 1906, de Boor 1973, de Boor / Janota 1997; ‚Konrad' mit Fragezeichen bei Grubmüller 1996) – aufgrund des „obszönen" Inhalts und sprachlicher Merkmale; anders Feistner 2000.

Die stilometrische Analyse ist in mehrfacher Hinsicht eine Herausforderung: Eine gattungsübergreifende Attribution ist mangels anderer Vergleichstexte nötig (nach Schöch 2014 wäre eine Gattungsmischung möglichst zu meiden). In Konrads Oevre herrscht eine Vielfalt an Themen, Frivoles wie in der ‚Halben Birne' ist eher selten – auch im einzigen anderen Märentext Konrads: im ‚Herzmäre' bleibt die Liebe unerfüllt, es kommt zum doppelten Minnetod. Zudem ist die ‚Halbe Birne' recht kurz: sehr gute Quoten erreicht Delta ab 5.000 Wortformen in einer Bag-of-Words (vgl. Abb. 3 sowie Eder 2013a und Eder 2013b).

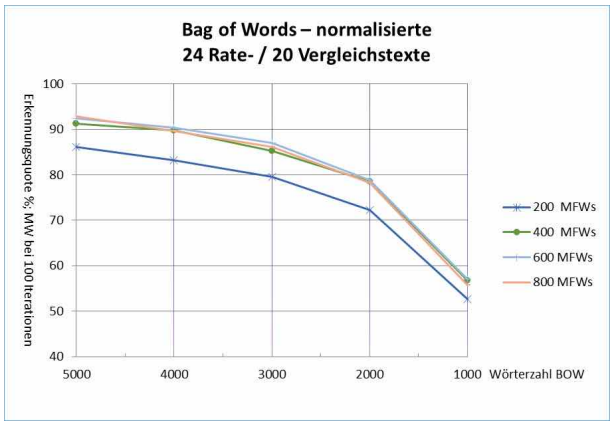


Abb. 3: zum Setting vgl. Dimpel 2018.

Die ‚Halbe Birne' enthält jedoch nur 2.469 Wortformen. Wenn man nun die ‚Birne' gegen ein Konrad-Korpus testet, kann man entweder die Wörter mit hoher Level-2-Differenz, die einer Erkennung von Konrad entgegenstehen, aus der Liste der untersuchten Most-Frequent-Words (MFWs) streichen. Oder man kann eine Positivliste mit „guten" Wörtern bevorzugt verwenden – Wörter mit hoher positiver Level-2-Differenz.

Vorab wird das Verfahren validiert: In einer Ermittlungsgruppe (vier Konrad-Texte) werden „gute" und „schlechte" Wörter identifiziert.[1] In einer Kontrollgruppe (vier andere Konrad-Texte) zeigt sich, dass die Erkennungsquote durch dieses Verfahren bei Bag-of-Words mit 2.000 Wortformen steigt – beim bevorzugten Verwenden „guter" Wörtern stärker als beim Aussortieren der „schlechten". Danach werden alle acht Konrad-Texte erneut zur Bildung der Listen der „guten" und „schlechten" Wortformen herangezogen. Als geeignete Parameter haben sich gezeigt:

„Gute Wörter": Level-2-Differenzen >+2,31 in 6 von 7 Ermittlungsgruppen-Ratetexten, 304 items

„Schlechte Wörter": Level-2-Differenzen <-1,2 in 2 von 7 Ermittlungsgruppen-Ratetexten, 174 items

Attributionstest 1: Halbe Birne und Herzmäre im Vergleichskorpus vs. Konrad-Ratekorpus

Erkennungsquote in %. Parameter: Bag-of-Words mit 2.000 Wortformen und mit 100 Iterationen, 400 MFWs, mit Normalisierungswörterbuch, Pronomina-Entfernung

	Halbe Birne	Herzmäre
Ohne Wortliste	28,6%	4,5%
Negativliste: Schlechte Wörter	35,8%	10,3%
Positivliste: Gute Wörter	83,8%	42,5%

Im Attributionstest 1 wird die ‚Halbe Birne' als Autor-Vergleichstext verwendet, als Ratetexte werden die acht Konrad-Texte sowie das ‚Herzmäre' verwendet; im ‚Herzmäre'-Test bleibt es bei acht Konrad-Ratetexten; das ‚Herzmäre' ist Autor-Vergleichstext. Hier erreicht das ‚Herzmäre' nur 4,5%, ein schlechter Wert, obwohl hier die Autorschaft nicht infrage gestellt wurde. Dagegen liegt die Erkennungsquote bei der ‚Halben Birne' auch ohne zusätzliche Wortlisten bereits über dem Zufallswert: Wenn ein Konrad-Text aus dem Ratekorpus nun nicht einem der 20 Texte von anderen Autoren zuordnet wird, sondern der ‚Halben Birne', dann stehen die Chancen dafür 1 zu 21. Wenn es also auf den Zufall zurückzuführen wäre, dass ein Text dem richtigen Autor zugeordnet wird, dann müsste die Erkennungsquote bei 5% liegen – so beim ‚Herzmäre'. 83,8% bei der ‚Halben Birne' sind ein ordentlicher Wert, wenn man bedenkt, dass nur kurze Bag-of-Words mit 2.000 Wortformen getestet werden können und dass gattungsübergreifend getestet wird.

Beim Attributionstest 1 befand sich die ‚Halbe Birne' im Vergleichskorpus. Im Ratekorpus waren inklusive ‚Herzmäre' 9 Konrad-Texte. Nun werden umgekehrt ‚Halbe Birne' bzw. ‚Herzmäre' als Ratetexte verwendet. Ins Vergleichskorpus gebe ich zu den 20 Distraktortexten in separaten Tests jeweils einen Konrad-Text als Autor-Vergleichstext ins Vergleichskorpus.

Attributionstest 2:

Autor-Vergleichstext	Ratetext Halbe Birne	Ratetext Herzmäre
Alexius	99,0	89,0
Engelhard	69,0	97,0
Turnier von Nantes	99,0	3,0
Pantaleon	94,0	88,0
Partonopier	86,0	89,0
Schwanritter	97,0	87,0
Silvester	95,0	100,0
Trojan. Krieg	94,0	71,0
Herzmäre	2,0	

Im Attributionstest 2 übersteigen die meisten Werte 86%. Es gibt lediglich zwei deutliche Ausreißer, an denen jeweils das ‚Herzmäre' beteiligt ist. Dieses Minneleid-und-Minnetod-Märe fügt sich nicht zur politischen Propagandadichtung ‚Turnier von Nantes'. Auch zur ‚Halben Birne' passt das ‚Herzmäre' nicht: Dort geht es um eine Dame, die einen Ritter abweist, weil er beim Birnenverzehr keine Tischmanieren an den Tag legt. Die Dame schläft mit einem vermeintlich taubstummen Hofnarren, der sich jedoch später als der abgewiesene Birnen-Ritter entpuppt. Interessante Fehlattributionen (etwa ‚Birne' zu ‚Häslein' statt zum ‚Herzmäre') werden im Vortrag vorgestellt.

Ein kleiner Schritt für die Attribution der ‚Halben Birne' an Konrad

Als Katharina Zeppezauer-Wachauer (Salzburg) mir einige Mären aus der Mittelhochdeutschen Begriffsdatenbank überlassen hat (vielen Dank dafür!), hat sie notiert: „Vielleicht können Sie ja wirklich, wie Edith Feistner gefordert hat, ‚Konrad seine Birne wiedergeben'!" Auch wenn die Zahlen in beiden Attributionstests trotz der geringen Textlänge und trotz der Gattungsproblematik überraschend eindeutig sind, möchte ich bei einer vorsichtigen Interpretation bleiben. Zwar ist die Wahrscheinlichkeit sehr gering, dass die gefundene Nähe der ‚Halben Birne' zum Konrad-Korpus auf dem Zufall beruht. Allerdings wären ‚Kontrollpeilungen' (Eibl 2013) wünschenswert: Eine Attribution sollte nicht auf einem einzelnen Test mit einer Methode erfolgen, wünschenswert wären Bestätigungen mit anderen Methoden. Immerhin aber geht es hier nicht um eine blinde Attribution, sondern lediglich um Widerspruch gegen eine Athetese der Forschung. Eine Attribution stünde in Einklang mit Konrads Selbstnennung in fünf von sieben überlieferten Textzeugen.

Zudem würde ich den Test gerne mit einem größeren Mären-Korpus wiederholen, in dem idealerweise längere Texte wären und mehr Texte, die näher an Konrads Schaffenszeit liegen. Dass die Birne nicht zu Kaufringer clustert, könnte auch dem zeitlichen Abstand geschuldet sein, der durch gemeinsame groteske oder frivole Inhaltselemente nicht überlagert wird.

Wichtig ist mir auch das Verfahren: Bislang ist eine Feature-Eliminierung oder Feature-Selektion häufig auf dem Weg des maschinellen Lernens erfolgt (Büttner et alia 2016) – mit dem Nachteil, dass der Weg der Kategorisierung teilweise im Dunk-

len bleibt. Ermittelt man „gute“ oder „schlechte“ Wörter via Level-2-Differenzen, so ist transparent, wie man zu den Parametern kommt und wie auf dieser Basis die weiteren Berechnungen erfolgen.

Fußnoten

1. Im Vergleichskorpus verwende ich hier und für die folgenden Attributionstests 7 Romane und 13 Mären: Barlaam, Daniel, Lanzelet, Meleranz, Parzival, Tristan, Wigalois; Frauentreue, Haeslein, Heidin_B, JvFreiberg_Raedlein, Kaufringer_Moerderin, Kaufringer_Rache, Kaufringer_listige_Frauen, Pyramus, Rosenpluet_Pfarrer, Schlegel, Schueler_Paris, StudentenAbenteuer_A, Zwickauer_Moenches_Not.

Bibliographie

Büttner, Andreas / Dimpel, Friedrich Michael / Evert, Stefan / Jannidis, Fotis / Pielström, Steffen / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (forthcoming 2017): „„Delta“ in der stilometrischen Autorschaftsattribution“, in: ZfdG.

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

De Boor, Helmut (1967): „Die Chronologie der Werke Konrads von Würzburg, insbesondere die Stellung des Turniers von Nantes“, in: *PBB* 89: 210–269.

De Boor, Helmut / Janota, Johannes (1997): *Geschichte der deutschen Literatur von den Anfängen bis zur Gegenwart. Band III /,1. Die deutsche Literatur im späten Mittelalter: Epik, Lyrik, Didaktik, geistliche und historische Dichtung: 1250–1350*, 5., neubearb. Aufl. von Johannes Janota. München.

Dimpel, Friedrich Michael (2016): *„Burrows’ Delta im Mittelalter: Wilde Graphien und metrische Analysedaten“*, in: Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts zur DHd-Tagung 2016 in Leipzig, http://dhd2016.de/: 65–70.

Dimpel, Friedrich Michael (2017a): *„Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch“*, in Stolz, Michael (Hrsg.): *Konferenzabstracts DHd 2017 Bern. Digitale Nachhaltigkeit*. Bern: 100-103. http://www.dhd2017.ch/programm.

Dimpel, Friedrich Michael (forthcoming 2017b): „Ein Delta-Rätsel: Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an?“, in: *Dariah-de-Working Papers* n.n.

Dimpel, Friedrich Michael (forthcoming 2018): *„Stabile Autorschaft trotz handschriftlicher Varianz? Die Erfolgsquote von Burrows' Delta bei nichtnormalisierten mittelhochdeutschen Texten optimieren“* (in Begutachtung, n.n.).

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

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

Eibl, Karl (2013): „Ist Literaturwissenschaft als Erfahrungswissenschaft möglich? Mit einigen Anmerkungen zur Wissenschaftsphilosophie des Wiener Kreises“, in: Philip Ajouri [u. a.] (Hrsg.): *Empirie in der Literaturwissenschaft, Münster* (Poetogenesis. Studien zur empirischen Anthropologie der Literatur 8): 19–45.

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): *„Towards a better understanding of Burrows’s Delta in literary authorship attribution“*, in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, CO: Association for Computational Linguistics: 79–88. 10.5281/zenodo.18177. http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [Abruf 20.8.2015].

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Reger, Isabella/ Schöch, Christof / Vitt, Thorsten (2016): *„Burrows Delta verstehen“*, in: *Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts zur DHd-Tagung 2016 in Leipzig*, http://dhd2016.de/: 61–65.

Edith Feistner (2000): „Kulinarische Begegnungen. Konrad von Würzburg und ‚Die halbe Birne‘“ in: Dorothea Klein et al. (Hrsg.): *Vom Mittelalter zur Neuzeit. FS Horst Brunner.* Wiesbaden: 291–304. Grubmüller, Klaus (1996): *Novellistik des Mittelalters. Märendichtung*. Frankfurt/Main 1996 (Bibliothek des Mittelalters 23).

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

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): *„Improving Burrows’ Delta – An Empirical Evaluation of Text Distance Measu-*

res“, in: *Digital Humanities Conference 2015*, Sydney. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Lachmann, Karl (1820): *Auswahl aus den Hochdeutschen Dichtern des dreizehnten Jahrhunderts*, Berlin.

Laudan, Hans (1908): „‚Die Halbe Birne‘ nicht von Konrad von Würzburg“, in: *ZfdA* 50: 158–166.

Schöch, Christof (2014): *„Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“* in: Christof Schöch und Lars Schneider (Hrsg.): *Literaturwissenschaft im digitalen Medienwandel*, Berlin (Philologie im Netz, Beiheft 7): 130–157.

Die Ontologie historischer deutschprachiger Berufs- und Amtsbezeichnungen. Interoperationalität und Berufsklassifizierung durch semantisches Topic Modeling

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

Moeller, Katrin
katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Forschungsstand:

Berufsbezeichnungen sind eine der häufigsten Angaben von individualspezifischen Quellen. Besonders in den Sozial- und Politikwissenschaften, den Geisteswissenschaften und einigen naturwissenschaftlichen Disziplinen (Sozialtopografie, Medizin, Arbeitsmedizin, Epidemiologie etc.) bieten Berufsbezeichnungen einen wichtigen Bezugspunkt sozialstruktureller Analysen. Dazu kommen verschiedene Formen von Berufsklassifikationen zum Einsatz. Während für die Berufswelten des 20./21. Jahrhunderts verschiedene normierte Klassifikationsmodelle als Standard sowohl auf nationaler wie internationaler Ebene existieren, kann die interdisziplinäre Forschung für deutschsprachige, historische Berufe nicht auf ein solches Normsystem zurückgreifen. Insgesamt gibt es bisher keinen gültigen Standard. Zu den renommiertesten und qualitativ hochwertigsten historischen Klassifikationsmodellen gehören bisher die Historical International Standard Classification of Occupations (HISCO) (van Leeuwen et al. 2002), das PST-System der Cambridge Group um Wrigley und Davies (Wrigley 2010). Sie zeichnen sich durch eine theoretisch nachvollziehbare Konzeption der Tätigkeit aus, beinhalten bisher aber kaum deutschsprachige Berufsnamen. Berufsklassifikationssysteme wie zu Altona 1803 (Brandenburg et al.1991) und das Berufsklassifikationsmodell zur Analyse von Bürgerlichkeit von Schüren und Hettling (Schüren 1989, Hettling 1999) liefern Ansätze für deutschsprachige Berufe. Daneben existieren etliche andere, im Zuge der Städteforschung entstandene Systeme, die jedoch eher als Teilaspekt einer Arbeit entstanden sind (bspw. François 1982; Kill 2001; Rödel 1985; Sachse 1987).

Aufgrund des hohen personellen Aufwands werden Systematiken häufig induktiv entwickelt und die Einordnung der Berufe (Abklären des genauen historischen Tätigkeitsprofils; Zuordnung zur richtigen Beschreibungsform der Berufsklassifizierung) gar nicht durchgeführt bzw. nicht dokumentiert. Daher entwickeln Forschungsprojekte jeweils neue Klassifizierungsmodelle, was eine Vielzahl nicht vergleichbarer Ergebnisse und für das einzelne Projekt ein enorm zeitaufwendiges Verfahren produziert. Dies schmälert den wissenschaftlichen Mehrwert, da die Ergebnisse von sozialstrukturellen Aussagen und Analysen letztlich nicht reproduzierbar sind. Zudem unterbleibt bei solchen Vorhaben sowohl eine konzeptionelle als auch eine direkte Anbindung an moderne Klassifikationsmodelle, da diese keine unmittelbare „Schnittstelle“ zu historischer Beruflichkeit bieten.

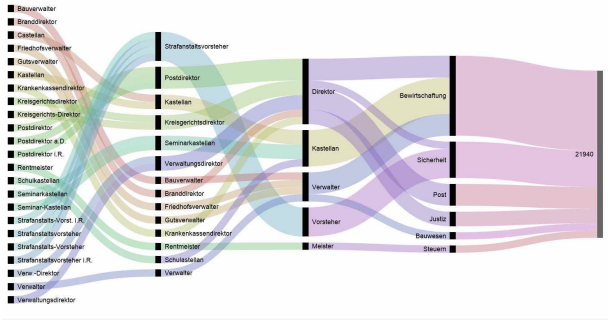
Mit der Ontologie deutschsprachiger Berufs- und Amtsbezeichnungen möchten wir diese Lücke schließen und eine Berufssystematik für historische, deutschsprachige Berufe entwickeln, die sowohl an die (modernen) Klassifikationssysteme anknüpfen (Klassifikation der Berufe 2010: Wiemer et al. 2011) als auch die deutschen Berufsbezeichnungen für internationale, historische Klassifikationssysteme (HISCO, PST) anschlussfähig

macht und die sowohl als maschinenlesbare Ontologie als auch mittels eines Webservice für einen individuellen Zugriff genutzt werden kann.

Ziel des Projekts

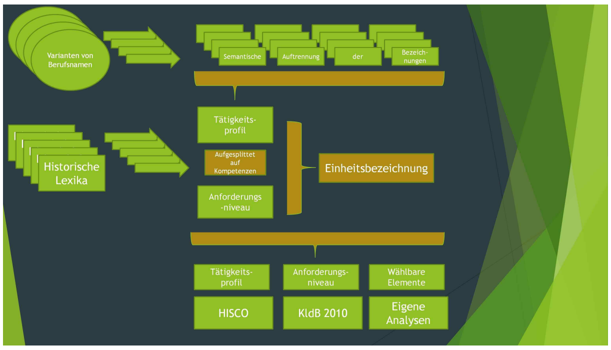
Ziel des Projektes ist es einen nationalen Standard zur Normierung und multiperspektivischen Klassifizierung von deutschsprachigen, historischen Berufsbezeichnungen zu entwickeln und als webbasierte, offene Ressource für die Nachnutzung in Forschungsprojekten und für Infrastrukturen des Forschungsdatenmanagements zur Verfügung stellen. Der hohe Mehrwert für die wissenschaftliche Analyse und die enorme Arbeitserleichterung sollte erheblich dazu beitragen, auf positive Weise tatsächlich einen anreizbildenden Standard zu etablieren, zumal dieser dann auch international anschlussfähig sein wird. Da sich die meisten Klassifizierungsmodelle auf Berufssysteme des 19./20. Jahrhundert konzentrieren, möchten wir diese Systematiken um Beruflichkeit der Frühen Neuzeit und des Spätmittelalters erweitern. Im Einzelnen umfasst dieser Service:

- Einen Standard zur Normierung von deutschsprachigen Berufsbezeichnungen. Dazu stehen ca. 200.000 Varianten von historischen Berufsschreibungen zur Verfügung, die sich in ca. 20.000 normierte Einheitsbezeichnungen zusammenfassen lassen. Dieser enorme Korpus ist Ergebnis jahrelanger eigener Datenproduktionen und Kooperationen mit qualitativ hochrangigen Projekten. Im Gegensatz zu vielen Handbüchern, Lexika und anderen gedruckten Kompendien (Ebener 2015; Gerholz 2005; Haemmerle 1933; Molle 1975; Palla 2014; Puchner / Stadler 1935; Reith 1991; Ulm-Sanford 1975) werden die Originalschreibungen der Varianten erhalten. Dies ermöglicht einen hohen Mehrwert für interdisziplinäre Zugriffe z.B. linguistische und etymologische Analysen, aber auch für den technischen Vorgang der Normierung und Zuweisung der Kodierung in einer Vielzahl von Erschließungsprojekten. Damit bietet das Werkzeug eine außerordentliche gute Grundlage, zur Erschließung (Lemmatisierung und Kodierung) von textuellen Quellen jeder Art.
- Um die Klassifizierung zu erleichtern, wie auch einen Informationsverlust durch die Klassifizierung zu verhindern, werden alle Berufsbezeichnungen zunächst in kategorial geordnete semantische Einheiten aufgetrennt. Dies erleichtert erheblich die Nutzbarkeit des Angebots für unterschiedlichste Anforderungen. Diese semantische Trennung und Normierung der Namenseinheiten ermöglicht später eine individuell flexible inhaltliche Auswahl verschiedener logischer Einheiten (Topic Modeling) nach eigenen Prinzipien. In unserem Vortrag möchten wir die Logik dieses Modells erläutern und dem Fachpublikum zur Diskussion stellen, um Anregungen für eine interdisziplinär optimal zu nutzende Methode zu erhalten. Diese Vorgehensweise unterscheidet das Projekt von allen anderen rein auf ID-Identifikatoren zugeschnittenen Verfahren und macht es wesentlich flexibler und breiter für verschiedenste Bedürfnisse geistes- und sozialwissenschaftlicher Forschung anwendbar. Bei herkömmlichen Verfahren kann lediglich über eine fest zugewiesene ID-Nummer auf den verschiedenen Ebenen des Kodierungssystems klassifiziert und sortiert werden. In dem hier vorgeschlagenen Modell erfolgt zusätzlich auch auf der semantischen Ebene die Möglichkeit zum intuitiven Systematisieren, Anordnen und Auswählen. Dies gilt auch für alle weiteren Informationen (Geschlecht, Herrschaft, Rechtsbeziehungen, Ruhestand, Karriere etc.), die mit typisch frühneuzeitlichen, quellenbasierten Berufsbezeichnungen einhergehen und in moderne Berufssystematiken meist keinen Eingang finden. Moderne Klassifikationssysteme für frühneuzeitliche Berufsnamen sind daher tendenziell mit einem sehr hohen Informationsdefizit belastet und können unter Umständen in ahistorischen Analysen enden.



- Zur Klassifizierung von Berufen werden aus zeittypischen Lexika des 17./18. Jahrhunderts Tätigkeitsbeschreibungen abstrahiert, welche die Anforderungsprofile und Kompetenzen einzelner Berufe definieren. Sie dienen dem Abgleich mit modernen Tätigkeitsprofilen und erlauben erstmals überhaupt eine tätigkeitsgenaue Zuordnung historischer Berufe. Im Vortrag möchten wir zeigen, welche Möglichkeiten zur Systematisierung und Kategorisierung Lexika der frühen Neuzeit bie-

ten und welche inhaltlichen Anknüpfungspunkte sie für moderne Klassifikationssysteme liefern. Dieser Aspekt bietet zudem hervorragende Möglichkeiten zur fachwissenschaftlichen Analyse, werden auf diese Weise doch Definitionskriterien frühneuzeitlicher Beruflichkeit überhaupt ermittelt und mittels einer logischen Auszeichnung auffindbar und auswertbar (xml/TEI). Wir möchten zeigen, welche Probleme dabei auftreten und welche Lösungsansätze wir hierfür entwickelt haben.



Weitere Ziel des Projektes (aber nicht unmittelbar unseres Vortrages, auch wenn wir auf die einzelnen Punkte sicherlich hinweisen) sind zudem folgende Punkte:

- Die Tätigkeitsprofile ermöglichen die nachvollziehbare, transparente Einordnung von Berufsbezeichnungen zum bereits existierenden englischsprachigen Berufsklassifizierungssystems HISCO. Momentan gibt es in HISCO lediglich 1.306 deutschsprachige Berufsbezeichnungen. Die Daten wurden nicht von einem Muttersprachler kodiert, weshalb viele Zuordnungen korrigiert werden müssten. Die Kodierung von HISCO ermöglicht das Mapping weiterer historischer Klassifikationsmodelle wie HISCLASS oder PST. Da HISCO moderne Beruflichkeit misst, ist eine Erweiterung und Präzisierung für die Berufe der Frühen Neuzeit unerlässlich (van Leeuwen et al. 2002; HISCO Tree Of Occupational Groups [http://historyofwork.iisg.nl/major.php]).
- Zudem werden die Einheitsbezeichnungen der KldB 2010 zugeordnet. Sie bieten nicht nur einen Zugriff auf (weitere) moderne Berufsklassifikationsmodelle, sondern durch den theoretischen Perspektivwechsel auch eine Systematisierung nach Anforderungsprofilen (Wiemer 2011).
- Die offenen Daten ermöglichen es Nutzern, transkribierte, quellenbasierte Originalberufsbezeichnungen zu normalisieren und automatisiert in ein ausgewähltes Kodierungssystem zu überführen. Ein Usecase wäre bspw. ein Forschungsprojekt, welches eine Netzwerkanalyse durch Heiratsverbindungen vornimmt. Der Beruf bietet hier häufig den einzigen Hinweis auf die sozialstrukturelle Dimension von Gruppen. Über das Werkzeug können die Daten in einem gewünschten Klassifizierungsmodell ausgegeben werden, indem die Originalschreibungen der Berufe in das Tool geladen und automatisiert kodiert werden.
- Varianten die keine automatisierte Kodierung beinhalten, können anschließend recherchiert und kodiert und in das Kompendium übertragen werden. Das Werkzeug ist damit beliebig erweiterbar. Das Projekt entwickelt ein Geschäftsmodell, wie solche Daten als Teil eines Serviceangebots nachkodiert und integriert werden können. Als Teil der langfristigen Infrastruktureinrichtung DARIAH soll die Ontologie der Berufe durch die Niedersächsische Staats- und Universitätsbibliothek auf Dauer angeboten werden.

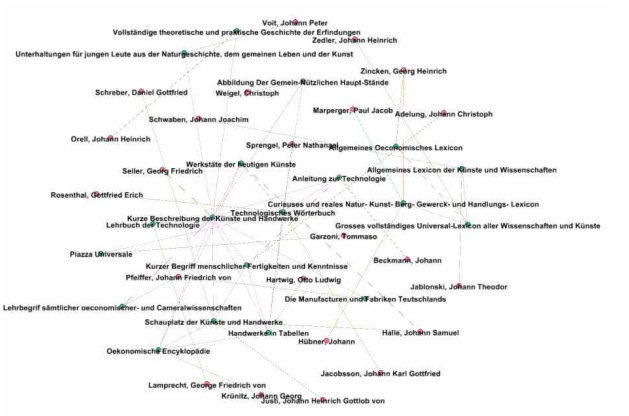
Zwischenergebnisse

- Im Bereich der Berufssegmentierung wurde ein Prototyp eines Datenschemas entwickelt, um die Berufe in adäquate Informationseinheiten aufzutrennen.
- Der Informationsgehalt der Lexika bezogen auf berufskundliche Informationen wurde tiefergehend untersucht und erste Aussagen können darüber getroffen werden.

	N	BerFac	Syn	Obj	Roh	Wer	BerOrd	Prü	Kom	Pre	Ort	Ausb	VerTit	SpeTit	Recht	Pri	Zer	Ver	HuKTit	Qua	HuK	Eig	For	Eide	NebTät	Mittel
Zincke	3	100%	67%	100%	0%	100%	100%	100%	100%	0%	67%	100%	100%	50%	N/A	0%	100%	0%	33%	50%	33%	100%	100%	100%	0%	65%
Seiler	3	100%	67%	100%	100%	100%	100%	100%	33%	67%	67%	100%	67%	100%	N/A	0%	0%	N/A	50%	50%	50%	0%	33%	0%	100%	63%
Krünitz	9	100%	100%	100%	100%	43%	75%	100%	86%	33%	57%	100%	60%	63%	100%	100%	0%	0%	17%	33%	25%	25%	50%	0%	0%	57%
Zedler	10	90%	90%	60%	33%	38%	60%	25%	75%	57%	50%	0%	0%	50%	100%	50%	50%	33%	43%	44%	40%	44%	40%	0%	0%	45%
Adelung	3	100%	67%	50%	100%	100%	100%	67%	33%	33%	33%	100%	33%	0%	N/A	0%	0%	N/A	50%	50%	50%	0%	0%	0%	0%	44%
Weigel	8	63%	50%	75%	50%	71%	60%	50%	29%	57%	14%	25%	17%	43%	0%	50%	50%	100%	50%	0%	25%	0%	0%	50%	50%	41%
Jablonksi	6	83%	83%	67%	50%	50%	40%	50%	40%	75%	33%	0%	0%	60%	N/A	100%	50%	0%	20%	0%	33%	20%	0%	0%	0%	37%
Höpfner	3	100%	67%	0%	0%	100%	100%	50%	33%	50%	33%	50%	50%	0%	N/A	0%	0%	100%	0%	0%	0%	33%	0%	0%	0%	33%
Garzoni	10	80%	80%	40%	67%	63%	20%	0%	50%	71%	13%	0%	33%	25%	0%	0%	25%	0%	14%	44%	0%	33%	20%	50%	50%	32%
Marperger	7	71%	86%	50%	100%	20%	25%	33%	67%	40%	33%	0%	0%	33%	N/A	0%	33%	0%	20%	17%	0%	17%	0%	0%	0%	28%
Beier	5	80%	40%	33%	100%	33%	33%	33%	40%	25%	40%	0%	0%	25%	0%	50%	33%	0%	0%	0%	25%	0%	0%	0%	0%	25%
Jacobsson	2	100%	50%	100%	N/A	0%	0%	100%	0%	0%	50%	0%	100%	0%	0%	0%	0%	N/A	0%	0%	0%	0%	0%	N/A	0%	24%
Mittel	69	89%	70%	65%	64%	60%	59%	59%	49%	42%	41%	40%	38%	37%	33%	29%	28%	26%	25%	24%	23%	23%	20%	18%	17%	41%

33%-Quantile		
oberes	mittleres	unteres
≥ 44%	bis 33%	bis 24%

Des Weiteren ist ein erstes Netzwerk von Autorenschaften und Informationsflüssen nachweisbar.



- Die Taxonomie der Klassifikationssysteme von HISCO und der KldB 2010 wurden ausführlich untersucht und erste Ansätze einer eigenen Systematik können vorgestellt werden.
- Mehrere Workflowalternativen von der Quelle bis zum Ergebnis der Inhaltsanalyse wurden erprobt, bewertet und können mit ihren Vor- und Nachteilen präsentiert werden. An dieser Stelle platziert sich am deutlichsten die „Kritik an der digitalen Vernunft“, vor allem in Bezug auf die
  - Arbeit mit OCR-Erkennung (am Beispiel der OCR-Programmsammlung „ocropy“)
  - dem Pre- und Postprocessing von Quellendigitalisaten (im Sinne von Ordnerstrukturen und Bildaufarbeitung)
  - XML vs. QDA-gestützten hermeneutischen Verfahren der Inhaltsanalyse (zur Implementierung und Weiterverarbeitung der Ergebnisse und Daten) und
  - der Nutzung von Objekt- oder relationalen Datenbanken zur Datenverwaltung und Verarbeitung.

Hierbei wird über die Vermeidung von epistemischen Fallstricken durch informationstechnische Automatisierung reflektiert und die Effizienzsteigerung digitaler Werkzeuge kritisch betrachtet, aber auch über die neuen Möglichkeiten zur Bewältigung von Big Data und dem dazugehörigen Erkenntnisgewinn referiert.

Bibliographie

Brandenburg, Hajo / Gehrmann, Rolf / Krüger, Kersten / Künne, Andreas / Rüffer, Jörn (1991): Berufe in Altona. Berufssystematik für eine präindustrielle Stadtgesellschaft anhand der Volkszählung. Kiel: Arbeitskreis für Wirtschafts- und Sozialgeschichte Schleswig-Holsteins.

Brückner, Carola / Möhle, Sylvia / Pröve, Ralf / Roschmann, Joachim (1988): Vom Fremden zum Bürger: Zuwanderer in Göttingen 1700-1755. In: Hermann Wellenreuther (Hg.): Göttingen 1690-1755. Studien zur Sozialgeschichte einer Stadt. Göttingen (Göttinger Universitätsschriften. Serie A, Schriften, Bd. 9), S. 88–174.

Bundesanstalt für Arbeit (1988): Klassifizierung der Berufe. Systematisches und alphabetisches Verzeichnis der Berufsbenennungen. Nürnberg. Online verfügbar unter http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Klassifikation-der-Berufe/KldB1975-1992/Generische-Publikationen/KldB1988-Systematischer-Teil.pdf, zuletzt geprüft am 07.06.2016.

Bundesanstalt für Arbeit (2011): Klassifikation der Berufe 2010. Systematischer und alphabetischer Teil mit Erläuterungen. 2 Bände. Nürnberg (1). Online verfügbar unter https://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Klassifikation-der-Berufe/KldB2010/Printausgabe-KldB-2010/Generische-Publikationen/KldB2010-Printversion-Band1.pdf, zuletzt geprüft am 07.06.2016.

Ebeling, Dietrich (1987): Bürgertum und Pöbel. Wirtschaft und Gesellschaft Kölns im 18. Jahrhundert. Köln (Städteforschung. Reihe A, Darstellungen, Bd. 26).

Ebner, Jakob (2015) Wörterbuch historischer Berufsbezeichnungen. Berlin u.a.

Fischer, Volker (2000): Stadt und Bürgertum in Kurhessen. Kommunalreform und Wandel der städtischen Gesellschaft 1814-1848. Kassel (Hessische Forschungen zur geschichtlichen Landes- und Volkskunde, 35).

François, Etienne (1982): Koblenz im 18. Jahrhundert. Zur Sozial- und Bevölkerungsstruktur einer deutschen Residenzstadt. Göttingen (Veröffentlichungen des Max-Planck-Instituts für Geschichte, Bd. 72).

Gerber, Roland (2001): Gott ist Burger zu Bern. Eine spätmittelalterliche Stadtgesellschaft zwischen Herrschaftsbildung und sozialem Ausgleich. Weimar (Forschungen zur mittelalterlichen Geschichte, 39).

Gerholz, Heinrich (2005): Gerholz-Kartei, Eine Sammlung alter Berufsbezeichnungen, Lübeck.

Haemmerle, Albert (1966): Alphabetisches Verzeichnis der Berufs- und Standesbezeichnungen vom ausgehenden Mittelalter bis zur neueren Zeit, (Reprografischer Nachdruck der Ausgabe München 1933), Hildesheim.

Hahn, Hans-Werner (1991): Altständisches Bürgertum zwischen Beharrung und Wandel. Wetz-

lar, 1689-1870. München (Stadt und Bürgertum, Bd. 2).

Hettling, Manfred (1999): Politische Bürgerlichkeit. Der Bürger zwischen Individualität und Vergesellschaftung in Deutschland und der Schweiz von 1860 bis 1918. Göttingen.

ILO (1958): International Standard Classification of Occupations. Geneva. Online verfügbar unter http://www.ilo.org/public/libdoc/ilo/1958/58B09_81_engl.pdf, zuletzt geprüft am 31.05.2016.

ILO (1969): International Standard Classification of Occupations. Revised Edition 1968. Geneva.

ILO (2004): ISCO-88. Main Objectives. Online verfügbar unter http://www.ilo.org/public/english/bureau/stat/isco/isco88/publ1.htm, zuletzt geprüft am 06.07.2016.

ILO (2008): Resolution Concerning Updating ISCO 1 Resolution Concerning Updating the International Standard Classification of Occupations. Online verfügbar unter http://www.ilo.org/public/english/bureau/stat/isco/docs/resol08.pdf, zuletzt geprüft am 31.05.2016.

Jägers, Regine (2001): Duisburg im 18. Jahrhundert. Sozialstruktur und Bevölkerungsbewegung einer niederrheinischen Kleinstadt im Ancien Regime (1713-1814). Köln (Rheinisches Archiv, 143).

Kill, Susanne (2001): Das Bürgertum in Münster 1770-1870. bürgerliche Selbstbestimmung im Spannungsfeld von Kirche und Staat. München.

Kroll, Stefan (1997): Stadtgesellschaft und Krieg. Sozialstruktur, Bevölkerung und Wirtschaft in Stralsund und Stade 1700 bis 1715. Göttingen (Göttinger Beiträge zur Wirtschafts- und Sozialgeschichte, Bd. 18).

Krüger, Kersten (1986): Sozialstruktur der Stadt Oldenburg 1630 und 1678. Analysen in historischer Finanzsoziologie anhand staatlicher Steuerregister. Oldenburg.

Laufer, Wolfgang (1973): Die Sozialstruktur der Stadt Trier in der frühen Neuzeit. Bonn (Rheinisches Archiv, 86).

Lundgreen, Margret Kraul; Ditt, Karl (1988): Bildungschancen und soziale Mobilität in der städtischen Gesellschaft des 19. Jahrhunderts. Göttingen.

Manke, Matthias (2000): Rostock zwischen Revolution und Biedermeier. Alltag und Sozialstruktur. Rostock (Rostocker Studien zur Regionalgeschichte, Bd. 1).

Molle, Fritz (1975): Wörterbuch der Berufs- und Berufstätigkeitsbezeichnungen, Wolfenbüttel .

Müller, Christina (1992): Karlsruhe im 18. Jahrhundert. Zur Genese und zur sozialen Schichtung einer residenzstädtischen Bevölkerung. Karlsruhe (Forschungen und Quellen zur Stadtgeschichte, 1).

Palla, Rudi (2014): Verschwundene Arbeit. Das Buch der untergegangenen Berufe. Wien.

Paulus, Wiebke / Matthes, Britta (2013): Klassifikation der Berufe. Struktur, Codierung und Umsteigeschlüssel. Online verfügbar unter http://doku.iab.de/fdz/reporte/2013/MR_08-13.pdf, zuletzt geprüft am 11.06.2016.

Puchner, Karl / Stadler, Josef Klemens (1935): Lateinische Berufsbezeichnungen in Pfarrmatrikeln und sonstigen orts- und familiengeschichtlichen Quellen, Hirschenhausen (Obby).

Raschke, Helga (2001): Bevölkerung und Handwerk einer thüringischen Residenzstadt. Gotha zwischen 1640 und 1740. 1. Aufl. Jena (Palmbaum Texte. Kulturgeschichte, Bd. 9).

Reith, Reinhold (1991): Lexikon des alten Handwerks. Vom späten Mittelalter bis ins 20. Jahrhundert. München.

Rödel, Walter Gerd (1985): Mainz und seine Bevölkerung im 17. und 18. Jahrhundert. Demographische Entwicklung, Lebensverhältnisse und soziale Strukturen in einer geistlichen Residenzstadt. Stuttgart (Geschichtliche Landeskunde, Bd. 28).

Sachse, Wieland (1987): Göttingen im 18. und 19. Jahrhundert. Zur Bevölkerungs- und Sozialstruktur einer deutschen Universitätsstadt. Göttingen (Studien zur Geschichte der Stadt Göttingen, Bd. 15).

Schüren, Reinhard (1989): Soziale Mobilität. Muster, Veränderungen und Bedingungen im 19. und 20. Jahrhundert. St. Katharinen.

Schüren, Reinhard (1989): Soziale Mobilität: Muster, Veränderungen und Bedingungen im 19. und 20. Jahrhundert. St. Katharinen.

Statistisches Bundesamt (1992): Klassifikation der Berufe 1992 (KldB 92). Gliederungsstruktur bis zur 4 Steller-Ebene. Stuttgart: Metzler-Poeschel. Online verfügbar unter https://www.destatis.de/DE/Methoden/Klassifikationen/Berufe/klassifikationkldb92_4st.pdf?__blob=publicationFile, zuletzt geprüft am 06.06.2016.

Straubel, Rolf (1995): Frankfurt (Oder) und Potsdam am Ende des Alten Reiches. Studien zur städtischen Wirtschafts- und Sozialstruktur. 1. Aufl. Potsdam (Quellen und Studien zur Geschichte und Kultur Brandenburg--Preussens und des Alten Reiches, Bd. 2).

Ulm-Sanford, Gerlinde (1975): Wörterbuch von Berufsbezeichnungen aus dem siebzehnten Jahrhundert, gesammelt aus den Wiener Totenprotokollen der Jahre 1648 - 1668 und einigen weiteren Quellen, Bern 1975.

van Leeuwen, Marco H.D. / Maas, Ineke / Miles, Andrew (2002): Historical international standard classification of occupations. Leuven.

Weichel, Thomas (1993a): Die Berufsstruktur der Städte - erste Ergebnisse und Vergleiche. In: Lothar Gall (Hg.): Stadt und Bürgertum im Übergang von der traditionalen zur modernen Gesellschaft. München (Historische Zeitschrift. Beihefte, n.F., Bd. 16), S. 51–74.

Weichel, Thomas (1993b): Die Bürger in ihrer beruflichen und sozialen Stellung. In: Lothar Gall (Hg.): Stadt und Bürgertum im Übergang von der traditionalen zur modernen Gesellschaft. München (Historische Zeitschrift. Beihefte, n.F., Bd. 16), S. 93–103.

Weichel, Thomas (1997): Die Bürger von Wiesbaden. von der Landstadt zur "Weltkurstadt", 1780 1914. München.

Digitale Differenz. Luhmanns Zettelkasten als physisch-historisches Objekt und als vernetzter Navigationsraum

Goedel, Martina
mgoedel@uni-koeln.de
Cologne Center for eHumanities

Zimmer, Sebastian
sebastian.zimmer@uni-koeln.de
Cologne Center for eHumanities

Schmidt, Johannes
johannes.schmidt@uni-bielefeld.de
Niklas Luhmann-Archiv, Fakultät für Soziologie, Universität Bielefeld

Einleitung

Inwieweit verändert der Einsatz digitaler Verfahren die sozialwissenschaftliche Forschung? Diese Frage lässt sich am Beispiel der Digitalisierung des Zettelkastens Niklas Luhmanns stellen, die im Rahmen des Forschungsprojektes “Niklas Luhmann - Theorie als Passion. Wissenschaftliche Erschließung und Edition des Nachlasses”[1] erfolgt. Das Langzeitvorhaben (2015-2030), in dem die Fakultät für Soziologie der Universität Bielefeld mit dem Cologne Center for eHumanities kooperiert, wird im Akademienprogramm durch die Nordrhein-Westfälische Akademie der Wissenschaft und der Künste gefördert.

Luhmann (1927-1998) zählt zu den bedeutendsten Soziologen des 20. Jahrhunderts. Im Laufe seiner 35-jährigen Forschungstätigkeit entwickelte er eine universale Sozial- und Gesellschaftstheorie, die er in annähernd fünfzig Monographien und 500 Aufsätzen publiziert hat. Als Basis für diese erstaunliche Produktivität diente Luhmann ein Zettelkasten, den er über vierzig Jahre lang systematisch gefüllt und gepflegt hat. Im Zuge der seit 2015 laufenden Nachlasserschließung wurden die ca. 90.0000 Zettel digitalisiert, in einem zweiten Schritt werden sie nun transkribiert und fachwissenschaftlich editiert sowie in eine Internetpräsentation überführt. Ziel dieses Prozesses ist zunächst eine digitale Reproduktion des Zettelkastens, die aber zugleich die Möglichkeiten der modernen digitalen Technik nutzt, um die nicht linear strukturierte Sammlung lesbar und ihre Genese nachvollziehbar zu machen. Indem die Digitalisierung über die reine Reproduktion hinausgeht, macht sie den Kasten selbst zu einem Forschungsobjekt.

Der (analoge) Zettelkasten Niklas Luhmanns

Die Zettelsammlung ist durch vier Merkmale gekennzeichnet, deren Kombination das besondere theoretische Kreativitätspotential der Sammlung begründet:

(a) **Nichthierarchische Ordnungsstruktur**: Luhmann verzichtet weitgehend auf eine vorher festgelegte systematische Ordnung der Sammlung; diese ist primär ein historisches Produkt seiner Lektüre- und Forschungsinteressen. Aufgrund seines spezifischen Einstellprinzips führt die jeweilige thematische Erstfestlegung nämlich nicht zu einer monothematischen Reihung von Zetteln: Die für den Zettelkasten konstitutive Idee ist, dass **ein Zettel thematisch** nur in irgendeiner Weise an den **vor ihm stehenden anschließen muss**, ohne sich an einer übergeordneten (systematischen) Themenstruktur zu orientieren. Damit korrespondiert eine spezifische Art der Notizgenerierung, bei der Luhmann Nebengedanken weiterverfolgt, indem er diese zusätzlichen Notizen, auf einen an dieser Stelle einzuschiebenden Zettel notiert, so dass ein Wachstum des Zettelkastens ‚nach innen‘ erfolgt.

(b) **Nummerierungssystem**: Mit der skizzierten Ablagetechnik in einem konstitutiven Zusammenhang steht das besondere Nummerierungssystem: Jeder Zettel erhält eine von seinem

jeweiligen Einstellplatz abhängige Nummerierung, die zugleich auf das Problem reagiert, wie Luhmann neue Zettel in den Altbestand einfügen kann, ohne die schon bestehende Nummerierung in Frage zu stellen. Zunächst erfolgt eine einfache Durchnummerierung der Zettel entsprechend des Zeitpunkts des jeweiligen Eintrags (1, 2, 3 usw.). Ein später erstellter Zettel, der einen einzelnen Aspekt (also gerade nicht zwingend monothematisch) von Zettel 1 weiterverfolgt, wird mit 1a nummeriert und zwischen den Zettel 1 und 2 eingeschoben. Daran kann wiederum monothematisch 1b anschließen oder aber eine thematisch abzweigende weitere Verzettelung folgen, beispielsweise 1a1. Das führt dazu, dass zwischen zwei ursprünglich direkt hintereinander eingestellten Zetteln im Extremfall bis zu 1000 später erstellte Zettel mit einem entsprechend komplexen Nummernsystem eingestellt werden.

(c) **Verweisungssystem**: Das skizzierte Ablagesystem macht es außerdem nötig, dass die thematisch oder konzeptionell miteinander zusammenhängenden, aber eben verstreut in der Sammlung stehenden Zettel aufeinander verweisen, indem auf den Zetteln jeweils die entsprechenden Zettelnummern notiert werden. Neben Einzelverweisen findet man häufig am Beginn eines thematischen Abschnitts auf einem einleitenden Zettel auch eine ganze Sammlung von Verweisen, die thematisch verwandte Bereiche des Zettelkastens systematisch erschließen. Aufgrund einer stichprobenartigen Auszählung kann man davon ausgehen, dass sich in der Sammlung insgesamt ca. 50.000 Verweise befinden. Luhmann selbst nennt diese netzwerkartige Verweisungsstruktur ein „spinnenförmiges System“.

(d) **Schlagwortverzeichnis**: Um Einstiegspunkte in dieses Verweisungsnetz zu erhalten, hat Luhmann ein ca. 4000 Einträge umfassendes Schlagwortverzeichnis erstellt. Dieses Schlagwortregister war das zentrale Werkzeug für seine Nutzung des Kastens, da nur so Notizen zu einem bestimmten Thema zuverlässig wiedergefunden werden konnten. Das Schlagwortregister des Kastens beansprucht dabei aber keinen Anspruch auf Vollständigkeit hinsichtlich der Erfassung der einschlägigen Stellen in der Sammlung. Vielmehr notiert Luhmann in der Regel nur maximal drei Systemstellen, an denen der jeweilige Begriff zu finden ist, da er annimmt, dass man dann über das interne Verweisungsnetz schnell die anderen relevanten Stellen findet.

Technische Umsetzung

Datenmodell und Arbeitsumgebung

Der Anlage des Zettelkastens folgend verstehen wir die Zettel als semantisch freie Einheiten und erzeugen für jede Zettelseite eine XML-TEI-Datei, in der die Transkription des Zettelinhalts erfolgt. Der Kopf (<teiHeader>) bietet Raum für Metainformationen, in einem umfangreichen Anhang zur Transkription (<back>) betten wir den Zettel in das Zettelnetzwerk ein. Wir unterscheiden mehrere Navigationswege durch den Kasten: Erstens eine physische Navigation in der von Zettelvorderseite zu folgender Vorderseite gesprungen wird. Zweitens geht es um übergeordnete Gliederungsverläufe. In einer dritten Navigation modellieren wir logische Abfolgen. In Form von Sprungzielen wird hier der nächste, in einem Gedankenstrang folgende Zettel angegeben, dabei werden von Luhmann eingeschobene Zettel zunächst ausgeblendet. Diese Abfolgen werden in eigenen Feldern klassifiziert und explizit gemacht.

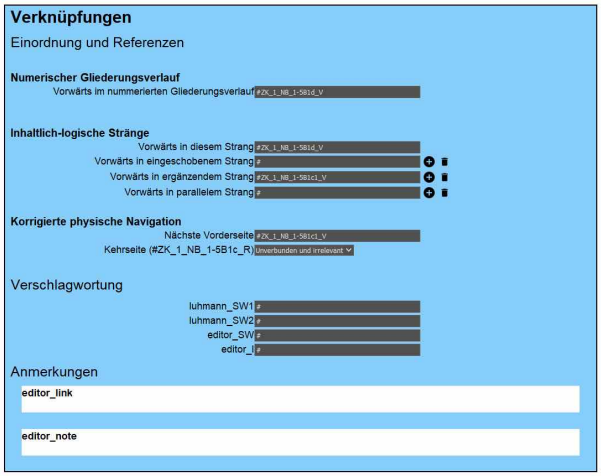


Abbildung 1: Ausschnitt aus der oXygen-Arbeitsoberfläche zur Transkription und Bearbeitung von Zettelseiten (Zettel 1-5B1c).

Um den Fachwissenschaftlern die Arbeit mit den XML-Texten zu erleichtern, arbeiten wir mit dem Author-Mode des oXygen XML-Editors.[2] Die von uns für die Zetteltranskription und Einbettung in das Netzwerk entwickelte Arbeitsoberfläche setzt im Verknüpfungsanhang auf fest eingerichtete Standardfelder, in die die Editoren IDs vorausliegender Zettel eintragen. Auf Basis dieser Einträge - die im Arbeitsprozess einer stetigen Weiterbearbeitung unterliegen - ergänzen wir über Skripte die Verbindung zu den davor liegenden Zetteln, um für die Portalnavigation beide Richtungen anbieten zu können.

Für die Transkription selbst stellen wir im Framework ein Mixed-Content Feld zur Verfügung, in dem weitere Textphänomene und von

Luhmanns selbst explizit formulierte Zettelverweise über Buttons ausgezeichnet werden können.

Forschungsportal

Das Forschungsportal soll dem Benutzer eine intuitive Möglichkeit bieten, durch den Zettelkasten zu navigieren und ihn dabei nicht nur so benutzen zu können, wie Luhmann es vorgesehen hat, sondern die Nutzbarkeit durch die digitalen Möglichkeiten zu erweitern. Dazu wird für jeden Zettel dessen Ort in der im Rahmen des Projekts erstellten Inhaltsübersicht sowie eine interaktive Visualisierung seiner inhaltlich-logischen Einordnung angezeigt. Über indikative Buttons ist erkennbar, welche Navigationsmöglichkeiten sich für den momentan angezeigten Zettel für den Benutzer bieten (siehe Abbildung 2). Die Buttons sind gruppiert nach dem Navigationstypus.

Innerhalb der Transkription sind Schlagwörter, bibliographische Angaben, Personennamen und vor allem Verweise zu anderen Zetteln verlinkt. Außerdem ist eine facettierte Volltext-Suche (z.B. nach Schlagworten oder bibliographischen Angaben) über vom Benutzer definierbare Bereiche des Zettelkastens möglich.

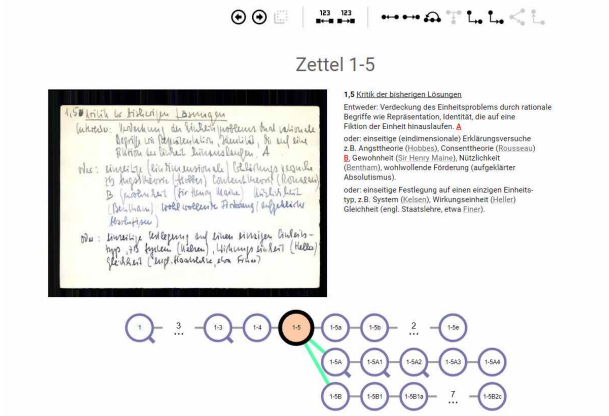


Abbildung 2: Ansicht eines Zettels im Forschungsportal

Neben optimierten Darstellungen für einzelne Zettelseiten gibt es Visualisierungen, die größere Bereiche des Kastens abdecken. Dazu zählt eine Visualisierung der im Rahmen der Edition erstellten inhaltlich-logischen Einordnungs- und Navigationsstruktur des Zettelkastens (siehe Abbildung 3), sowie ein Arc Diagram[3] zur Darstellung von Zettelverweisen innerhalb verschiedener Bereiche (siehe Abbildung 4), um die Dichte der Vernetzung innerhalb und zwischen den verschiedenen Abteilungen des Zettelkastens zu verdeutlichen.

Die Datenaufbereitung der TEI-XML-Dateien geschieht jeweils über im Rahmen des Projekts entwickelte Node.js-Skripte[4], bevor die Visualisierungen mit der JavaScript-Bibliothek D3.js[5] erstellt werden.

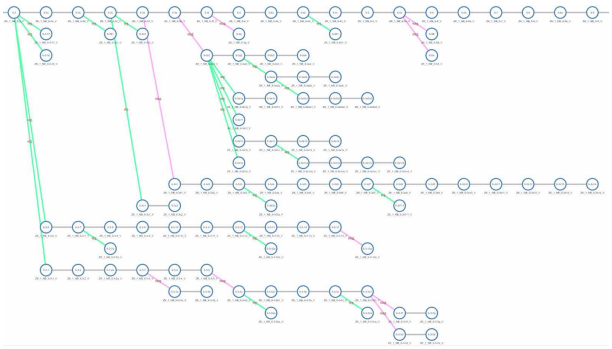


Abbildung 3: Ausschnitt einer Visualisierung der im Rahmen der Edition erstellten inhaltlich-logischen Einordnungs- und Navigationsstruktur des digitalen Zettelkastens.

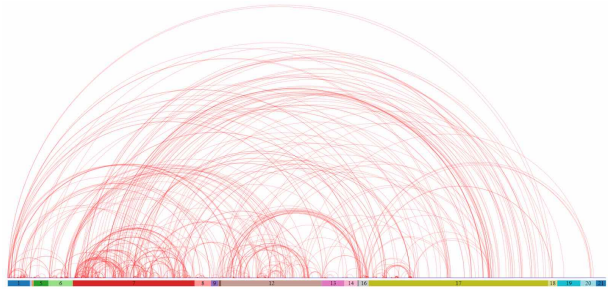


Abbildung 4: Visualisierung der internen Zettelverweise des ersten Auszugs von Zettelkasten I (ca. 3300 Zettel). Man erkennt die deutlichen Unterschiede der Vernetzungsdichte innerhalb und zwischen den 21 Abteilungen.

Fazit: Der Zettelkasten als Subjekt und Objekt der Forschung

Die Digitalisierung des Zettelkastens erfolgt als ein Editionsprojekt im Rahmen einer Nachlasserschließung, insofern ist das primäre Ziel eine digitalisierte **Reproduktion** mit der Intention, die Nutzbarkeit im Luhmannschen Sinne zu rekonstruieren und zu erleichtern. Die dafür notwendige digitale Modellierung hatte eine vertiefte Reflexion über das Design des Zettelkastens zur Folge, insbesondere aufgrund der damit einhergehenden Notwendigkeit einer eindeutigen Typendifferenzierung von Zettelanschlüssen. Durch

diese Differenzierung, die ihren Niederschlag in einer entsprechend komplexen Navigationsstruktur des digitalen Kastens gefunden hat , die eine erleichterte Nutzbarkeit des Kastens ermöglicht, kam es zu einer deutlichen Konkretisierung der zweiten fachwissenschaftlichen Intention der Publikation: der Möglichkeit einer werkgenetischen Lesbarkeit des Kastens über eine Rekonstruktion der Einstellhistorie. Gerade weil die Zettel selbst undatiert sind, ist eine solche historisierende Les**art** nur über die (fachwissenschaftliche) Identifizierung von ursprünglichen Zettelanschlüssen und späteren Einschüben möglich; die Les**barkeit** der dadurch implizierten Zettelfolgen, die sich von der physischen Stellordnung der Zettel in Teilen unabhängig macht, ist dann aber nur aufgrund der entsprechenden technischen Umsetzung möglich.

Die technische ‚Aufrüstung' des digitalen Kastens führt allerdings auch dazu, dass diese Version kein reines Abbild des analogen Kastens mehr darstellt, sondern den Zettelkasten nun in einer Weise verfügbar macht, wie Luhmann selbst ihn nie genutzt hat. Diese **Differenz zum ursprünglichen Zettelkasten** wird verstärkt durch die weiteren o.g. Recherchemöglichkeiten im Rahmen der digitalisierten Version sowie durch die Visualisierungsmöglichkeiten der Verweisungsstruktur zwischen den Zetteln, die zudem die Netzwerkförmigkeit der Sammlung transparent macht. Diese Form der Edition des Zettelkastens rekonstruiert den Kasten selbst also nicht mehr als das Forschungs**subjekt**, das er für Luhmann als ‚Denkmaschine' und Theorieapparat war, sondern macht den Kasten selbst schon bei der Editionsarbeit zu einem Forschungs**objekt** und bietet die für die Beantwortung von jetzt noch unbekannten Forschungsfragen notwendigen Instrumentarien an.

Fußnoten

1. *www.niklas-luhmann-archiv.de*
2. Vgl. oXygen Visual (WYSIWYG) XML Editors: https://www.oxygenxml.com/xml_author/WYSIWYG_Editors.html
3. Arc diagram: https://en.wikipedia.org/wiki/Arc_diagram
4. Node.js: https://nodejs.org/en/
5. D3.js: https://d3js.org/

Bibliographie

Krajewski, Markus (2002): *ZettelWirtschaft. Die Geburt der Kartei aus dem Geiste der Bibliothek* . Berlin: Kadmos.

Schmidt, Johannes F.K. (2016): "Niklas Luhmann´s Card Index: Thinking Tool, Communication Partner, Publication Machine", in: Alberto Cevolini (ed.): Forgetting Machines. Knowledge Management Evolution in Early Modern Europe. Leiden: Brill, 289-311.

Watts, Duncan (2004): "The »new« science of networks", in: Annual Review of Sociology 30, 243-270.

Digitale Methoden sind weder digital noch innovativ

Raunig, Michael
michael.raunig@uni-graz.at
Universität Graz, Österreich

Höfler, Elke
elke.hoefler@uni-graz.at
Universität Graz, Österreich

Ausgangslage

Das „Digitale" hat den allgemeinen Sprachgebrauch unabhängig von unterschiedlichen gesellschaftlichen Systemen erreicht: Man spricht etwa von der Digitalisierung der Wirtschaft, konstatiert einen allgemeinen Trend zur digitalen Transformation oder gar einen „digital turn", im Bildungsbereich etablieren sich Netzwerke und Initiativen wie das Hochschulforum Digitalisierung, in dessen Rahmen man sich über Anforderungen, Potenziale und Hindernisse im „digitalen Zeitalter" austauscht (https://hochschulforumdigitalisierung.de/). In Österreich wurde 2017 mit der Digital Roadmap Austria „[d]ie digitale Strategie der österreichischen Bundesregierung" für den Bildungssektor vorgelegt (https://www.digitalroadmap.gv.at/). Auch in den unterschiedlichen Wissenschaftsdisziplinen ist eine verstärkte Auseinandersetzung mit „Digitalität" und ihren Folgen zu beobachten, nicht zuletzt und neuerdings auch verstärkt in den Geisteswissenschaften (z. B. http://digitalitaet-geisteswissenschaften.de/).

„Digitale Methoden"

Vielfach wird der Anschein erweckt, dass mit den Prozessen der Digitalisierung des Wissenschaftsbetriebs im Lehren und Lernen auch ein

Aufkommen genuin digitaler Methoden und Praktiken verbunden sei, die sich einerseits grundlegend von den tradierten Methoden unterscheiden und andererseits neue Formen der Auseinandersetzung mit den unterschiedlichen Disziplinen ermöglichen. Dieser Eindruck wird durch einen unbefangenen Sprachgebrauch in Bezug auf das Attribut „digital“ („digitale Arbeit“, „digitale Strategien“, „digitales Lernen“, „digitale Forschung“ bis hin zu neuen, „digitalen“ Wissenschaftsdisziplinen) und ungelenk-plakative Metaphorik (etwa der bereits erwähnte „digital turn“) verstärkt.

Den vielfach mit „neuen Medien“ gleichgesetzten „digitalen Medien“ wird generell ein innovatives, wenn nicht gar revolutionäres Potential zuerkannt. Ähnliches wird von den „digitalen Methoden“ erwartet. Ist aber im Bereich der Methoden eine Gegenüberstellung, wie sie bei den traditionellen Medien im Gegensatz zu den neuen Medien bzw. bei den analogen im Gegensatz zu den digitalen Medien infolge des Computereinsatzes gerechtfertigt wird, zielführend? Sind bei den „digitalen Methoden“ ähnliche Umbrüche zu erwarten, und macht es Sinn, diese von den bisherigen Methoden abzugrenzen? Dagegen spricht einerseits, dass früher nicht von „analogen Methoden“ gesprochen wurde und mit einer solchen Etikettierung wenig Erkenntnisgewinn verbunden wäre. Andererseits ist es - historisch und empirisch gesehen - keineswegs der Gebrauch, sondern die technische (d. h. elektronisch-digitale) Realisierung von Medien, die diese als „neue“ qualifiziert; insofern ist es naheliegend, die Digitalität weniger in den Methoden als vielmehr in den (didaktisch und didaktisiert eingesetzten) Werkzeugen zu verorten.

Dass Methoden per se nicht digital sind, ist einigermaßen trivial, da digital (im engeren, technischen Sinn und im Gegensatz zum Begriff „analog“) lediglich eine Form der Datenrepräsentation (Dale & Lewis 2016: 57ff) benennt. Selbst die sogenannten „digitalen Medien“, deren sich digitale Methoden bedienen, sind nur in einem abgeleiteten Sinn digital, indem ihre technische Infrastruktur auf der Verarbeitung digitaler bzw. binär codierter Daten beruht. Die wahrgenommene Umgebung, die die „digitale Medien“ konstituieren („Multimedia“), entspricht vielmehr den Mustern analoger Datenrepräsentation, da die wahrgenommenen Inhalte aufgrund ihrer hohen „Auflösung“ kontinuierlich und „naturnah“ erscheinen – seien es auch „virtuelle“ oder dynamische/interaktive Gegenstände. Man nimmt nach wie vor Text, Sprache, Ton, (Bewegt-)Bild sowie (simulierte) Objekte und Situationen über die Sinne wahr; das Digitale an den „digitalen Medien“ ist lediglich für den Computer relevant. Insofern kann die Rede von digitalen Methoden - wenn man nicht den Methodenbegriff auf maschinelle Prozesse ausdehnen will - nur auf die maßgebliche Verwendung von technischen Hilfsmitteln (PCs, mobile Geräte, Internet, World Wide Web und aufbauende Technologien) in der Wissenschaft abzielen. Kerres (2016) deutet im Anschluss an seine Problematisierung des Begriffs der „digitalen Bildung“ - der nicht ernst gemeint sein bzw. buchstäblich genommen werden könne - an, dass die gegenwärtigen, mit der Digitalisierung verbundenen Transformationsprozesse nicht als Umstellung von traditionellen/analogen auf digitale Modi (im Sinne einer Dichotomie) zu verstehen sind, sondern vielmehr als „Durchdringung“ des Digitalen, die letztlich Digital-Attribuierungen selbstverständlich und gleichzeitig überflüssig machen würden.

Thesen

These der Unabhängigkeit der Methode. Methodologisch betrachtet sind die eingesetzten Werkzeuge und Plattformen unerheblich (vgl. Kerres et al.: 2002). Die Konzeption einer wissenschaftlichen Praxis, die Methode des Forschens oder Lehrens ist auf einer grundlegenderen Ebene verortet als jener der Hervorbringung und Repräsentation ihrer Ergebnisse und Erzeugnisse. Methodische Aspekte haben vielmehr mit den Gegenständen, der Genese und der Konfrontation wissenschaftlicher Disziplinen (etwa im Fall der „Digital Humanities“) zu tun. Es wäre im Zusammenhang der Digitalität dringend geboten, die Unterscheidung von Medium (als unhintergehbare Wahrnehmungsbedingung) und Werkzeug (als einfaches Hilfsmittel) zu präzisieren und deren unterschiedliche pragmatische Implikationen und Auswirkungen erneut zu überdenken. Neuartige Medien (man denke beispielsweise an die Einführung der Schrift, die geradezu das Paradigma mediengeschichtlicher Zäsuren darstellt) üben einen größeren Einfluss auf die Methodologie aus als bloß der Einsatz neuer Werkzeuge.

These der Einheitlichkeit der Methode. Digital genannte Methoden sind weitestgehend traditionell. Lehre beispielsweise war schon immer mediengestützt, neu ist nur die jeweilige Repräsentation. Ob ein Memory zum Wiederholen von Inhalten traditionell-analog in Papierform oder über digitale Werkzeuge (wie beispielsweise LearningApps.org, http://learningapps.org/) eingesetzt wird, ist nicht relevant; die Methode ist dieselbe (siehe hierzu beispielsweise die „Thesen zum Lernen im digitalen Wandel“, die Philip Stade in einem differenzierten Blogpost formuliert; Stade 2017). Es gibt folglich keine „digitalen“ Methoden, da der Einsatz digitaler Werkzeuge

zwar neue Dimensionen in der Anwendung eröffnet, aber kein hinreichendes Unterscheidungskriterium zu den traditionellen Methoden liefert.

These der Unbeeinflussbarkeit der Methode. Methodologische Aspekte sind von der digitalen Vermittlung weitgehend unberührt. Es gibt zwar eine Reihe von Begleiterscheinungen und Effekten des „digitalen Arbeitens“ (s. u.), es gibt jedoch keine methodologischen Implikationen bei der Verwendung digitaler Werkzeuge. Die Veränderungen, die sich in einem wissenschaftspraktischen Setting (in den vielfältigen Ausprägungen von „Forschung“, aber auch in den unterschiedlichsten Formen und Modellen von Lehre) durch den Einsatz digitaler Werkzeuge ergeben, sind bloß akzidentieller Natur; im methodischen Handeln sind keine wesentlichen Änderungen zu erwarten.

These der digitalen Repräsentation. Digital (repräsentiert) sind nur die verarbeiteten Daten, die mediale Umgebung der „digitalen Arbeit“ ist komplett „analog“. Die Ausgabegeräte von Computern sprechen dieselben Sinne an, die auch analoge Medien ansprechen. Es gibt (außer bei Maschinen und Bots) kein digitales Denken, Forschen, Lehren und Lernen, wie auch Wampfler (2017) herausstreicht.

Mehrwert und Ausblick

Im Gegensatz zu ihrem methodologischen Beitrag sind Mehrwert – wenngleich auch der Begriff „Mehrwert“ zu problematisieren ist, wie Brandhofer (2017) verdeutlicht – und Potenziale des Einsatzes digitaler Technologien zum Lernen und Lehren jedoch unbestritten. Diese umfassen

1. die erleichterte Herstellung, unkomplizierte Verbreitung und Verfügbarkeit von Forschungs- und Lehrinhalten (nicht nur, aber besonders im Sinn von „Open Educational Resources“), damit verbunden auch die Erschließung neuer Zielgruppen,
2. die Automatisierung bzw. maschinelle Unterstützung bestimmter Arbeitsschritte und Prozesse, wobei die durch Computerunterstützung erzielte Komplexität und Kapazität bestimmter Faktoren die menschlichen („analogen“) Grenzen überschreiten kann,
3. eine Tendenz zu kollaborativen (ortsunabhängigen und synchronen) Arbeitsweisen, Offenheit, Austausch und Partizipation („Science 2.0“, „Open Science“),
4. die Ent-Professionalisierung bzw. Demokratisierung sowohl der Produktion als auch der Nutzung von wissenschaftlichen Daten und Werkzeugen (was umgekehrt jedoch auch eine Reihe von nicht traditionell-wissenschaftlichen Fertigkeiten und Tätigkeiten bedingt) sowie
5. eine Pluralität von Werkzeugen und Möglichkeiten zur Ausübung wissenschaftlich-methodischer Praktiken.

Die (gewiss nicht vollständig) angeführten Effekte der Nutzung digitaler Medien in der wissenschaftlichen und unterrichtlichen Arbeit sind auch dasjenige, was ihren innovativen Charakter ausmacht und eine „Digitalisierung“ von Wissenschaft und Hochschule begrüßenswert macht. Innovative Methoden ergeben sich daraus jedoch keine.

Im Anschluss an die Vorstellung der Thesen versucht der Vortrag, diese mit aktuellen Beispielen innovativer „digitaler“ Methoden und Arbeitsgebiete unter Fokussierung der Medienpädagogik und -didaktik zu illustrieren bzw. zu untermauern. Die Perspektive der Autorin / des Autors in dieser Auseinandersetzung ist – aufbauend auf eine klassische geisteswissenschaftliche Sozialisierung – mediendidaktisch bzw. bildungstechnologisch motiviert. Empirische Gegenbeispiele (insbesondere aus den Digital Humanities) und konzeptuelle Gegenthesen sind höchst willkommen; Ziel ist es letztlich, die Berechtigung des Begriffs der „digitalen Methoden“ zu problematisieren und diesen in der Medientheorie und reflektierten Auseinandersetzung mit „digitalen Medien“ entsprechend zu verorten.

Bibliographie

Brandhofer, Gerhard (2017): „Zur Problematik des Begriffes Mehrwert“. http://www.brandhofer.cc/mehrwert/ [letzter Zugriff 19. September 2017].

Dale, Nell / Lewis, John (2016): *Computer science illuminated*. Sixth edition. Burlington, MA: Jones & Bartlett Learning.

Kerres, Michael (2016): „E-Learning oder Digitalisierung in der Bildung: Neues Label oder neues Paradigma?“, in: *Grundlagen der Weiterbildung - Praxishilfen* (7.30.10.80): 159–171.

Kerres, Michael / De Witt, Claudia / Stratmann, Jörg (2002): „E-Learning. Didaktische Konzepte für erfolgreiches Lernen“, in: Schwuchow, Karlheiz / Guttmann, Joachim (eds.): Jahrbuch Personalentwicklung & Weiterbildung 2003. Neuwied: Luchterhand.

Stade, Philip (2017): „Thesen zum Lernen im digitalen Wandel – Lernen im digitalen Wandel“. https://herrstade.wordpress.com/2017/03/26/thesen-zum-lernen-im-

digitalen-wandel/ [letzter Zugriff 19. September 2017].

Wampfler, Philippe (2017): „Der Kahoot-Sog und die Gefahr der Quizifizierung der digitalen Bildung – Schule und Social Media“. https://schulesocialmedia.com/2017/05/19/der-kahoot-sog-und-die-gefahr-der-quizifizierung-der-digitalen-bildung/ [letzter Zugriff 19. September 2017].

Digitale Modellierung von Figurenkomplexität am Beispiel des Parzival von Wolfram von Eschenbach

Braun, Manuel
manuel.braun@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart

Klinger, Roman
roman.klinger@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Padó, Sebastian
pado@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Viehhauser, Gabriel
viehhauser@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart

Einleitung

Figuren gehören zu den wichtigsten Bestandteilen literarischer Erzählungen. Narratologische Analysen haben sich bislang insbesondere mit zwei Aspekten von Figuren beschäftigt: ihrer strukturellen Bedeutung und ihrer Charakterisierung. Während der erste Aspekt in den Digital Humanities bereits modelliert worden ist (etwa im Rahmen von Netzwerkanalysen, vgl. Jannidis et. al. 2016, Piper et. al. 2017), steht die datengetriebene Untersuchung des zweiten noch ganz am Anfang, obwohl die hermeneutisch arbeitende Literaturwissenschaft wiederholt auf seine Bedeutung hingewiesen hat (etwa Jannidis 2004, 2009). Dieses Desiderat lässt sich darauf zurückführen, dass Figuren auf die unterschiedlichste Art gekennzeichnet werden – etwa durch ihr Handeln und Reden, aber auch durch Beschreibungen und Bewertungen des Erzählers – und dass neben expliziten Charakterisierungen schwerer zu greifende implizite stehen. Diese Informationen lassen sich nur schwer erheben und in einem umgreifenden Modell verrechnen.

Um dennoch einen Einstieg in die digitale Erfassung der Figurencharakteristik zu ermöglichen, nehmen wir eine vergleichsweise einfach zu modellierende Facette der Figurendarstellung in den Blick, und zwar das Konzept der Figurenkomplexität, das auch in der traditionellen Theoriebildung eine prominente Rolle spielt. Nach der ebenso verbreiteten wie vielkritisierten Kategorisierung von Forster (1927) lassen sich Figuren grundsätzlich in *flat* und *round characters* einteilen. Trotz der zahlreichen Differenzierungsversuche späterer Forscher prägt dieses Modell noch heute Annahmen über die literaturgeschichtliche Entwicklung von Figurenentwürfen. So gelten die Protagonisten der mittelhochdeutschen Literatur klischeehaft als *flat characters*, die in erster Linie auf ihre Funktion für das Handlungsgefüge hin konstruiert werden und nicht die Tiefe moderner Individuen erreichen (vgl. zusammenfassend Schulz 2012). Im Widerspruch hierzu hat die mediävistische Literaturwissenschaft aber auch herausgestellt, dass die Konstruktion komplexer Figuren zu den grundlegenden Gestaltungsprinzipien mittelhochdeutscher Texte gehören kann.

Als Paradefall für Figurenkomplexität kann der um 1200/1210 abgefasste Artusroman ‚Parzival‘ angesehen *werden*, dessen Autor Wolfram von Eschenbach im Prolog das Konzept eines ‚gemischten‘, nicht eindeutig als gut oder böse bewerteten Menschentypus entwirft, von dem er im Folgenden handeln will. Dementsprechend lassen sich im ‚Parzival‘ immer wieder Figuren finden, die nicht dem Ideal (nach ihm sind adelige Protagonisten nicht nur politisch mächtig, sondern zum Beispiel auch ethisch gut und körperlich schön) entsprechen, sondern in ihrer Darstellung widersprüchlich erscheinen und daher als komplex angesehen werden können.

Im Folgenden schlagen wir eine Methodik vor, mit der diese Einschätzungen der hermeneutischen Forschung digital modelliert und überprüft werden kann.

Methode

Wir untersuchen, ob man bereits mit extrem einfachen Methoden Vorhersagen zur Figurenkomplexität treffen kann, und verwenden dazu *distributionelle Analysen* (Harris 1954, Miller und Charles 1991). Distributionelle Analysen repräsentieren die Bedeutung eines Zielwortes durch Eigenschaften der Kontexte seines Vorkommens in Textkorpora – im einfachsten Fall durch die Häufigkeiten aller Kontextwörter innerhalb eines Fensters um die Vorkommen des Zielwortes (‚Kontextvektor').

Spezifisch betrachten wir zwei Arten frequenzbasierter Kontextvektoren: (a) *lexikalische Kontextvektoren*, berechnet über lemmatisierte Kontexte; (b) *entitätsbasierte Kontextvektoren*, berechnet über Eigennamen in den Kontexten. Dabei misst (a) die *lexikalische* Vielfalt der Kontexte, in denen die jeweilige Figur vorkommt, und (b) entsprechend die *soziale* Vielfalt der Kontexte. Beide Arten von Repräsentationen verwenden als Kontexte die sogenannten *Dreißiger*-Abschnitte (die Unterteilung des ‚Parzival' in Abschnitte von 30 Versen nach Karl Lachmann) und greifen nicht auf sprachliche Strukturen innerhalb der Dreißiger wie Wortarten, Syntax, Satzgrenzen oder Anaphern zurück.[1]

In der Computerlinguistik werden Kontextvektoren typischerweise miteinander verglichen, um die semantische Ähnlichkeit der Zielwörter zu modellieren und finden breite Anwendung (Pantel und Turney 2010). In unserer Studie betrachten wir stattdessen den *Informationsgehalt* der Kontextvektoren als Prädiktor für die Figurenkomplexität und messen ihn per *Entropie*. Entropie eruiert den Informationsgehalt $H(p)$ einer Wahrscheinlichkeitsverteilung p, definiert als $H(p) = -\Sigma \times p(x) \log p(x)$ (Shannon 1948). Entropie kann verstanden werden als die Anzahl an Bits, die an Information in der Verteilung enthalten sind. Damit enthalten diejenigen Kontextvektoren eine höhere Entropie, die in gleichmäßiger verteilten Kontexten vorkommen, als die, die häufig in ähnlichen Kontexten auftreten.

Dies entspricht der (aus theoretischer Perspektive natürlich stark vereinfachenden) Annahme, wonach Figuren als komplexer wahrgenommen werden, wenn sie in reicheren und verschiedeneren Kontexten vorkommen. Bei lexikalischen Kontexten (s.o.) sagen wir also voraus, dass höhere lexikalische Variabilität auf Komplexität hinweist – wobei Entropie freilich nicht zwischen lexikalisch reichen und semantisch widersprüchlichen Kontexten unterscheiden kann. Bei entitätsbasierten Kontexten übernimmt diese Rolle das gemeinsame Auftreten einer Figur mit mehreren anderen Figuren.

Methodisch ist anzumerken, dass Frequenz einen Störfaktor bei der Interpretation von Entropie darstellt: Häufigere Zielwörter kommen – ceteris paribus – häufiger in verschiedenen Kontexten vor und erhalten damit eine höhere Entropie. Da dieser Zusammenhang aber nicht linear ist, ist eine einfache Normalisierung nicht möglich. In einem ersten Schritt sind also nur Zielwörter mit ähnlicher Frequenz hinsichtlich ihrer Entropie gut vergleichbar.

Experiment

Als Textgrundlage verwenden wir den ‚Parzival' nach der 5. Auflage der Ausgabe Lachmanns (1891) in der digitalisierten, mit Lemmata und der Auszeichnung von Eigennamen versehenen Fassung von Yeandle (2014).

Für unser Hauptexperiment beschränken wir uns auf die sieben im Text am häufigsten namentlich genannten Frauenfiguren Cundrîe, Herzeloyde, Jeschûte, Cunnewâre, Arnîve, Bêne und Itonjê. Hinzu kommt Sigûne als ein in der Forschung oft genanntes Beispiel für eine weniger komplexe Figur im ‚Parzival'. Sieben der acht Figuren liegen in einem engen Frequenzband zwischen 30 und 40 Nennungen, während Sigûne nur 14 Mal vorkommt. Wir präsentieren die Ergebnisse per Streudiagramm, mit Entropie und Frequenz als den beiden Achsen.

Abbildung 1 und 2 zeigen die Ergebnisse für die wichtigsten weiblichen Hauptfiguren. Diese lassen sich zwanglos mit den Einschätzungen der hermeneutischen Literaturwissenschaft in Einklang bringen. Die höchsten Werte weisen Cundrîe und Herzeloyde auf, auf deren Komplexität die ‚Parzival'-Forschung immer wieder hingewiesen hat: Bei Cundrîe handelt es sich um die Gralsbotin, die zwar kultiviert und nach höfischen Sitten gekleidet auftritt, zugleich aber monströs hässlich ist. Sie klagt Parzival zwar berechtigterweise an, fällt dabei aber aus dem höfischen Rahmen. Herzeloyde zieht ihren Sohn Parzival aus (übermäßiger?) Trauer um ihren im Kampf gefallenen Ehemann fern von der höfischen Welt in einer Waldeinöde auf, um zu verhindern, dass auch er einst Ritter wird – ein Verhalten, das die Forschung unterschiedlich bewertet. Mittelwerte erreichen Cunnewâre und Jeschute, die in der Forschung durchaus kontrovers diskutiert werden. Bêne, Arnîve und Itonjê lassen sich hingegen eher als Nebenfiguren bezeichnen, für die keine hohe Komplexität zu erwarten war.

Die in der Trauer um ihren Geliebten Schionatulander verharrende Sigûne, die gerade in ih-

rer Geradlinigkeit einen Gegenentwurf zu den auf Ruhm und Ehre versessenen Artusrittern darstellt, zeigt erwartungsgemäß keine hohen Entropiewerte. Allerdings erlaubt dieser Befund keine starke Interpretation, da für Sigûne aufgrund der niedrigeren Frequenz von vorneherein niedrigere Entropiewerte zu erwartet sind. Um dennoch eine Vorhersage zur Komplexität Sigûnes zu erhalten, vergleichen wir in den Abbildungen 3 und 4 ein breiteres Spektrum an Figuren im Frequenzbereich zwischen 10 und 20 Nennungen bezüglich ihrer Entropie. Dies entspricht einer erweiterten Version unserer Hypothese: Figuren, die *verglichen mit anderen Figuren des gleichen Frequenzbereiches* besonders hohe bzw. niedrige Entropien aufweisen, sind *relativ zu diesen* mehr bzw. weniger komplex.

Interessanterweise liefern die beiden Kontextdefinitionen für Sigûne abweichende Vorhersagen: Die unterdurchschnittliche soziale Kontextvielfalt entspricht der Isoliertheit der Figur, die sich aus Trauer von der Welt zurückzieht und wenn, dann fast nur noch mit Parzival interagiert. Die vergleichsweise hohe lexikalische Vielfalt könnte demgegenüber darauf zurückzuführen sein, dass Sigûne neben ihrer Trauer auch die weitere Funktion hat, Parzival über seine Herkunft aufzuklären: Während Sigûne als Figur eher statisch erscheint, sind ihre Erzählungen über die Gralswelt informationshaltig. Für eingehendere Untersuchungen wären daher weitere Faktoren der Figurendarstellung wie Figurenbeschreibung und Figurenrede mit einzubeziehen sowie das Verhältnis von lexikalischer und sozialer Kontextvielfalt näher zu bestimmen.

Diskussion

Unsere Ergebnisse legen nahe, dass sich die Analyse der lexikalischen und ‚sozialen' Vielfalt von Figurenkontexten durchaus als Maß für eine erste Annäherung an das Konzept der Figurenkomplexität eignet. Damit konnte mit erstaunlich einfachen Mitteln ein Zugang zum komplexen narratologischen Themenfeld der Figurencharakterisierung gewonnen werden. Auf der technischen Ebene bleiben zwei Fragen offen, zum einen die nach den sprachlichen Mechanismen, die zur Korrelation von Entropie und Figurenkomplexität führen, zum anderen die nach dem tatsächlichen Zusammenhang von Frequenz und Entropie: Gibt es also auch selten auftretende *round characters* oder häufig vorkommende flache Figuren?

Auf der konzeptuellen Ebene bieten unsere Ergebnisse den Einstieg in eine differenziertere Modellierung von Figurenkomplexität, deren Erarbeitung auf die traditionelle narratologische Theoriebildung zurückwirken kann: Die digitale Modellierung macht es notwendig, die Faktoren, aus denen sich das Konzept der Figurenkomplexität zusammensetzt (etwa Figurenhandeln, -rede, -beschreibung und -bewertung), genauer und expliziter zu bestimmen sowie über ihre Gewichtung nachzudenken. Außerdem wäre zu klären, was das Konzept der Komplexität genau meint und wie es sich zu verwandten Konzepten wie denen der Hybridität, der Dynamik oder der Aktantenhaftigkeit von Figuren verhält. Im Sinne einer kritisch reflektierten digitalen Methodik werden somit vielschichtige literaturwissenschaftliche Phänomene wie die Charakterisierung von Figuren durch den formalisierenden Zugang nicht nivelliert, vielmehr führt die Verbindung von quantitativen und qualitativen Methoden zur wechselseitigen Erhellung.

Abbildung 1: Entropie nach Lemmata im Dreißiger, weibliche Hauptfiguren

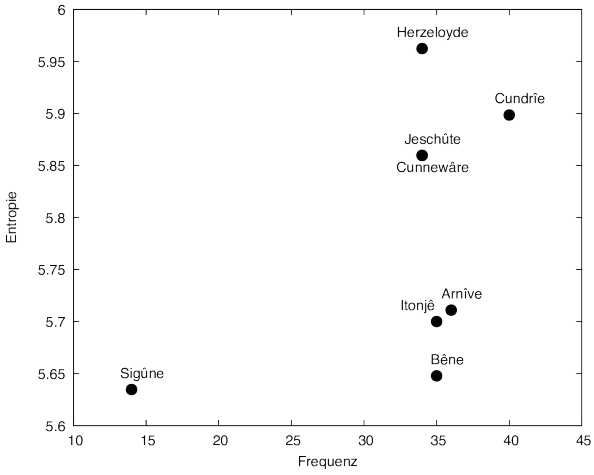


Abbildung 2: Entropie nach Figuren im Dreißiger, weibliche Hauptfiguren

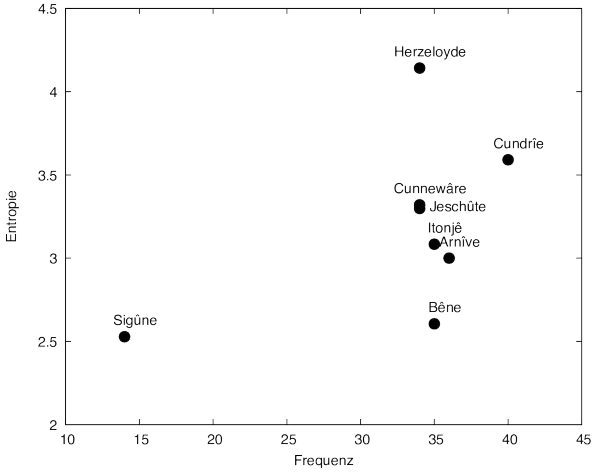


Abbildung 3: Lexikalische Entropie im Dreißiger, Nebenfiguren

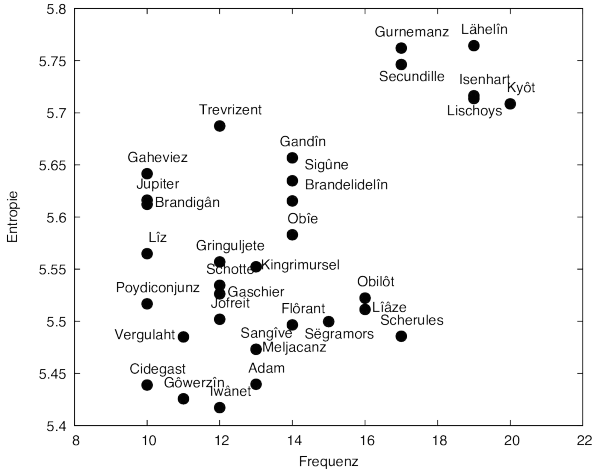
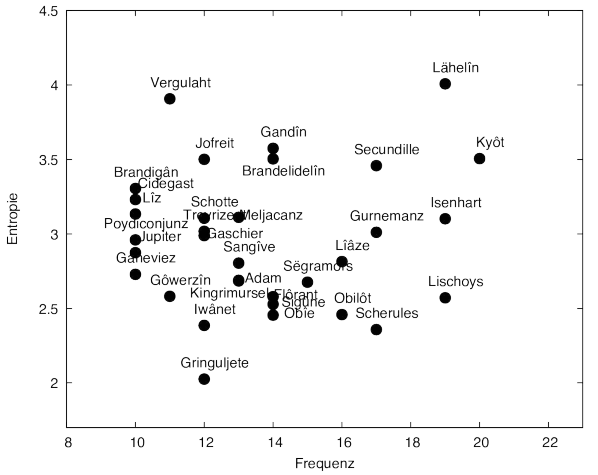


Abbildung 4: Soziale Entropie im Dreißiger, Nebenfiguren



Fußnoten

1. Experimente mit kleineren Kontexten lieferten schlechtere Ergebnisse.

Bibliographie

Forster, Edward M. (1927): *Aspects of the Novel.* New York: Harcourt.

Harris, Zellig S. (1954): "Distributional Structure", in: *Word* 10(2-3): 146–162.

Jannidis, Fotis (2004): *Figur und Person.* Beitrag zu einer historischen Narratologie. Berlin: de Gruyter.

Jannidis, Fotis (2009): „Character", in: Hühn, Peter / Pier, John / Schmid, Wolf / Schönert, Jörg (eds.) Handbook of Narratology. Berlin: de Gruyter 14-29

Jannidis, Fotis / Reger, Isabella / Krug, Manfred / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): „Comparison of Methods for the Identification of Main Characters in German Novels", in: *Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków* 578-582.

Manning, Christopher D. / Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Miller, George A. / Charles, Walter G. (1991): "Contextual Correlates of Semantic Similarity", in: *Language and Cognitive Processes* 6(1): 1–28.

Piper, Andrew / Algee-Hewitt, Mark / Sinha, Koustuv / Ruths, Derek / Vala, Hardik (2017): "Studying Literary Characters and Character Networks", in: *Digital Humanities 2017: Conference Abstracts. Montreal, Canada, August 8-11, 2017.*

Schulz, Armin (2012): Erzähltheorie in mediävistischer Perspektive. Berlin: De Gruyter.

Shannon, Claude E. (1948): "A Mathematical Theory of Communication", in: *Bell System Technical Journal* 27(3): 379–423.

Turney, Peter D. / Pantel, Patrick (2010): "From Frequency to Meaning: Vector Space Models of Semantics", in: *Journal of Artificial Intelligence Research* 37(1): 141–188.

Yeandle, David (o. J.): *Stellenbibliographie zum 'Parzival' Wolframs von Eschenbach für die Jahre 1753–2004*. http://wolfram.lexcoll.net/homeidx.htm [letzter Zugriff 9. Januar 2017].

Digitale Vernunft zwischen Text und Diagramm Digital Mapmaking als Hilfsmittel zur Erklärung historischer Ereignisse

Frank, Ingo
frank@ios-regensburg.de
Leibniz-Institut für Ost- und Südosteuropaforschung (IOS)

Einleitung

Mit meinem Vortrag möchte ich zu einer Kritik der digitalen Methoden beitragen. Gegenwärtige Forschung im Bereich der Digital Humanities konzentriert sich vorwiegend auf den Einsatz

quantitativer digitaler Methoden. Ich betrachte diesen Schwerpunkt auf 'Big Data', 'Distant Reading' und die Beschränkung auf statistische Verfahren zur Erklärung geisteswissenschaftlicher Phänomene kritisch. In einer methodologischen Kritik der digitalen Geisteswissenschaften werde ich dazu aus wissenschaftstheoretischer Perspektive anhand Beispielen historischer Forschung die Grenzen der quantitativen Ansätze aufzeigen, um dann Möglichkeiten vorzustellen, wie qualitative Methoden im Sinne eines **Augmenting Human(ist) Intellect**-Ansatzes (Engelbart 1962) digital unterstützt und erweitert werden sollten.

Vorgehensweise

Als erstes werde ich die Methoden der Digital Humanities mit Hilfe der NeMO (NeDiMAH Methods Ontology) wissenschaftstheoretisch verorten (Abb. 1).

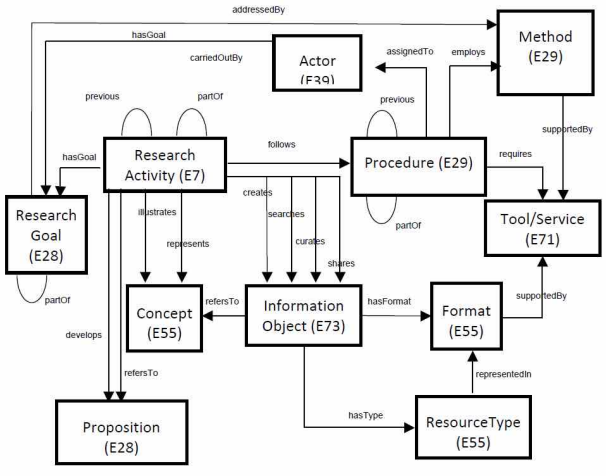


Abb. 1: NeMO (NeDiMAH Methods Ontology) zur Modellierung von Forschungsprozessen, digitalen Methoden, Werkzeugen, Ressourcen (Dokumente, Forschungsdaten, etc.) usw. (aus Benardou et al. 2010)

NeMO dient als Rahmen zum Aufbau einer Taxonomie digitaler Methoden und Werkzeuge. Ein erster Schritt dazu ist die Unterscheidung zwischen quantitativen und qualitativen Methoden. Da diese Unterscheidung auf methodischer Ebene und nicht etwa auf epistemologischer oder theoretischer Ebene erfolgt, erscheint die gängige scharfe Trennung in quantitative und qualitative Forschung nicht gerechtfertigt. Auf Ebene der Epistemologie muß man bei einer bestimmten erkenntnistheoretischen Position bleiben und kann dabei qualitative und quantitative Methoden kombinieren, ohne daß das methodologisch problematisch wäre (Crotty 1998). Bei digitaler Forschung kommt es auf den Ansatz der formalen Modellierung von Forschungsgegenständen an. Formale Modellierung bildet den Kern der Digital Humanities und ist Voraussetzung um überhaupt digital transformierte Foschungsmethoden einsetzen zu können – egal ob quantitativ oder qualitativ. Man könnte daher anstatt von **digitalen** Methoden besser von **formalen** bzw. **formalisierbaren** Methoden sprechen. Im Vortrag werde ich mich auf 'Digital Mapmaking' als Methode konzentrierten, die den Einsatz von Diagrammen zur qualitativen Forschung umfaßt. Um den Mehrwert von diagrammatischer Darstellung zu zeigen, werde ich auf Beispiele aus der Sozialgeschichte zurückgreifen. Skocpols "States and Revolutions: A Comparative Analysis of France, Russia, and China" (Skocpol 1979) eignet sich sehr gut, um zu zeigen, daß Text nicht ausreicht, um die komplexen kausalen Narrative, die zu sozialen Revolutionen führen, gut nachvollziehbar zu repräsentieren. In der methodologischen Forschung dazu wird sogar argumentiert (Mahoney 1999; George / Bennett 2005), daß die Darstellung der kausalen Prozesse in diagrammatischer Form notwendig ist, um die kausalen Zusammenhänge in der Präsentation der Forschungsergebnisse explizit zu machen und darüber hinaus den Forschungsprozess selbst durch den Zwang zur Formalisierung zu unterstützen. Am Beispiel Skocpols vergleichender Studie werde ich demonstrieren, wie die Methoden QCA (Qualitative Comparative Analysis) und Process Tracing in den Digital Humanities durch den Einsatz von geeigneten diagrammatischen ‚Denkwerkzeugen' wie Hypertext-Karten, Fuzzy Cognitive Maps und Dynamischer Netzwerkanalyse unterstützt und erweitert werden können.

Beispiele

Um überhaupt sinnvoll von ‚digitalen' Methoden reden zu können, muß der Aspekt der formalen Modellierung als zentral erachtet werden. Formalisierung ist die Voraussetzung für die „Mechanisierung angeblich geistiger Tätigkeiten" (Brauer et al. 1989). Piotrowski (2016) definiert Digital Humanities wie folgt: "The digital humanities study the means and methods of constructing formal models in the humanities." Ein Modell ist dabei als Repräsentation eines geisteswissenschaftlichen Untersuchungsgegenstands zu verstehen und ‚formal' bedeutet **logisch kohärent**, **nicht mehrdeutig** und **explizit** (Piotrowski 2016).

Mir geht es im folgenden um digitale Methoden des 'Mapmaking':

> [Mapmaking] is a combinatorial method, for it involves mapping information gathered from other sources: generally secondary data or surveys, though texts are another possibility. It must thus reflect the strengths and weaknesses of its sources of data. As with hermeneutics, mapmaking involves a set of techniques beyond those involved in the methods it draws upon. Its obvious strength lies in terms of spatiality. There are, quite simply, some spatial movements that cannot be properly visualized nor comprehended without recourse to maps […] (Szostak 2004)

Dazu kann mit NeMO Quellen- und Archivmaterial, das zur Erstellung von Karten herangezogen wird, formal erfaßt werden, was im Kontext von Digital Libraries und Forschungsdatenmanagement relevant ist.

Einführen werde ich **Digital Mapmaking** mit einem für die Geschichtswissenschaft naheliegenden Diagrammtyp (Champagne 2016): Zeitleisten. Synchronoptische Visualisierungen (Abb. 2) dienen der Kontextualisierung großer Datenmengen aus verschiedenen Quellen und ermöglichen die Generierung von Hypothesen durch Abduktion (Frank 2017a). Eine Erklärung – z. B. in Form der Beschreibung eines sozialen Mechanismus, der ein historisches Ereignis hervorbringt – kann jedoch nur durch weiterführende Forschung gefunden werden.

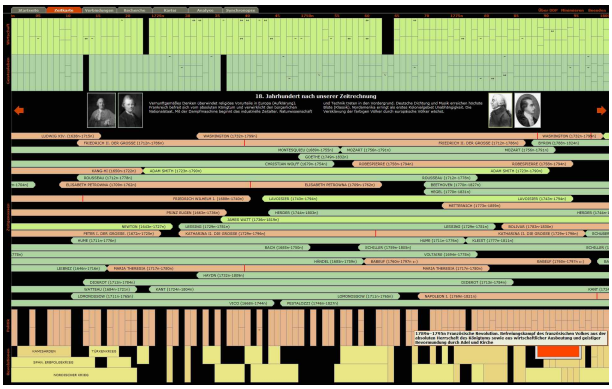


Abb. 2: Parallele Zeitleisten in der Zeitkarte aus der Digitalen Edition (Behrendt et al. 2010) von Peters' Synchronoptischer Weltgeschichte (Peters / Peters 1952) als ‚Hypothesen-Generator'

Im weiteren Verlauf werde ich zeigen, wie Visualisierungsansätze nicht nur **explorative** Analyse unterstützen können, sondern wie sie darüber hinaus auch die Möglichkeiten zur **Erklärung** historischer Ereignisse erweitern können. Sehr gut veranschaulicht werden kann ein solcher Ansatz mit den diagrammatischen Darstellungen der kausalen Narrative in Skocpols Theorie sozialer Revolutionen (Skocpol 1979).

Goertz und Mahoney (2005) argumentieren, daß "a failure to appropriately conceptualize levels and relationships between levels" zu vielen Fehlinterpretationen von Skocpols Theorie geführt haben. George und Bennett (2005) empfehlen tatsächlich "diagrams to present clearly the argument of causal narratives, to make the causal claim more explicit". Um die laut Goertz und Mahoney (2005) oft in den Interpretationen vorkommende Verwechslung von Kausalität und Konstitution zu vermeiden, eignet sich zur Darstellung von **Two-Level Theories** der Einsatz von Diagrammen (Abb. 3):

> For example, the examination of an ontological relationship between levels allows the analyst to explore the specific defining properties of the basic-level concepts that actually affect the outcome of interest. In this case of an ontological relationship, the specific properties identified in the secondary level are "mechanisms" that explain why the basic-level variables have the effects they do. (Goertz / Mahoney 2005)

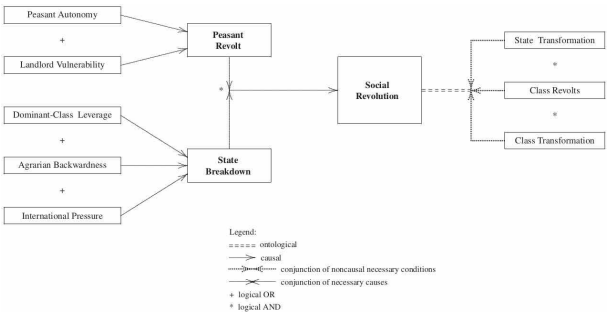


Abb. 3: Diagramm für Two-Level Theories zu Ursachen und Konstitution sozialer Revolutionen (aus Goertz / Mahoney 2005)

Im Rückentext des Geschichtstheorie-Lehrbuchs von Kolmer (2008) steht dazu treffend: „Wer sich nicht von der Beredsamkeit der Historiker blenden lassen will, muss das Gerüst entdecken können, das ihre Erzählungen trägt." Meine Idee ist nun, die kausalen Narrative nicht nur als Diagramme zu visualisieren, sondern die historischen Narrative der komplexen kausalen Zusammenhänge als Hypertext zu modellieren. Der Mehrwert dieses **Hypertext Mapping**-Ansatzes wird durch einen weiteren Verweis auf Szostak (2004) deutlich: "Mapmaking can give unique insight into spatial and temporal aspects of causal systems (among other things)."

Offensichtlich bietet die diagrammatische Visualisierung (Abb. 4) von Mahoney (1999) deutliche Vorteile bei der Darstellung der komple-

xen kausalen Zusammenhänge und erleichtert dadurch das genaue Nachvollziehen und Verstehen der historischen Vorgänge, die zum Zusammenbruch des französischen Staates führten, erheblich.

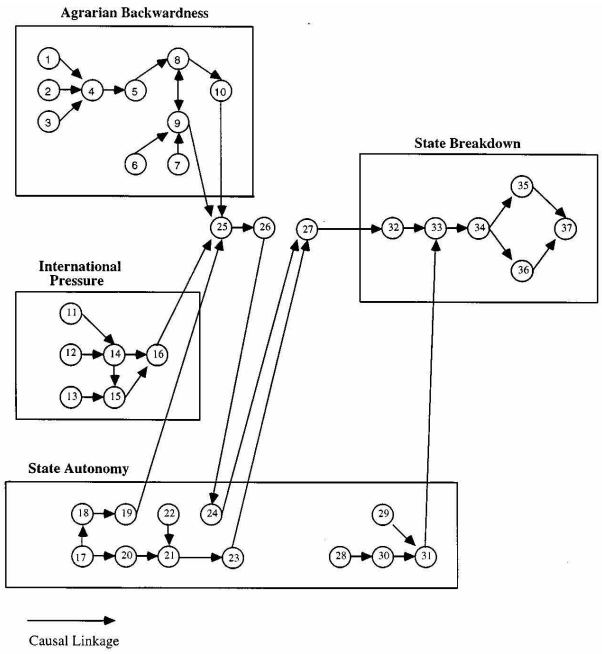


Abb. 4: Diagramm des kausalen Narrativs über den Prozess des Zusammenbruchs eines Staates (aus Mahoney 1999)

Allerdings berücksichtigt das Diagramm nach Mahoney (1999) keine komplexen Rückkopplungsschleifen zwischen Ereignissen und außerdem fehlt die Darstellung der Gewichtung einzelner Ereignisse und die Möglichkeit, die Details der kausalen Prozesse zu erfassen. Diese vernachlässigten Aspekte könnten durch Causal Loop Diagrams (siehe Beispiel in Abb. 5) oder Fuzzy Cognitive Maps (Carvalho 2013) oder auch Dynamische Netzwerkanalyse (Lemercier 2015b) modelliert werden.

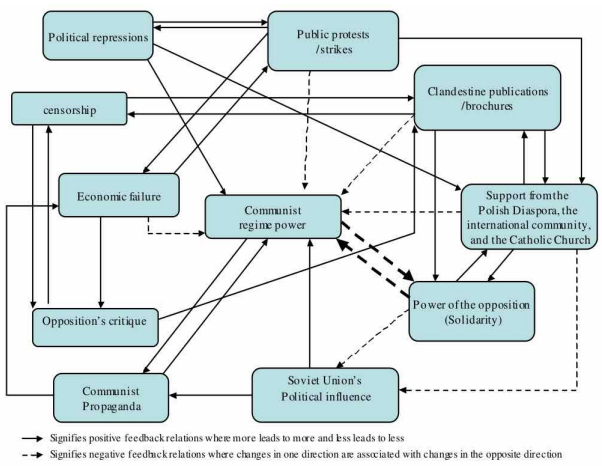


Abb. 5: Causal Loop Diagram des Konflikts zwischen der polnischen Regierung und Solidarność (aus Coleman et al. 2006)

Die ersten fünf von Mahoney (1999) aus Skocpol (1979) extrahierten kausalen Faktoren, sollen zur Veranschaulichung einer zusätzlichen Möglichkeit zur expliziten Strukturierung mit Hypertext Maps ausreichen:

1. Property relations prevent introduction of new agricultural techniques (S. 55)
2. Tax system discourages agricultural innovation (S. 55)
3. Sustained growth discourages agricultural innovation (S. 55)
4. Backwardness of French agriculture (esp. vis-à-vis England) (S. 56)
5. Weak domestic market for industrial goods (S. 55–56)

Die Kausalkette von 4 nach 5 wird von Skocpol (1979) z. B. mit einer Reihe von historischen Studien belegt. Diese Verweise können in einem Hypertext-Narrativ mit typisierten Links (Peroni / Shotton 2012) explizit gemacht werden.

Wie ein kritischer Rückblick in die Geschichte der Hypertext-Forschung zeigt, sind wesentliche Anforderungen an Hypertext-Systeme bisher immer noch nicht zufriedenstellend erfüllt. Ich werde zeigen, wie einige der sieben offenen Punkte/Issues von Halasz (2001) mit aktueller Semantic Web- und Informationsvisualisierungstechnologie angegangen werden können. Suche unter Berücksichtigung der Hypertextstruktur (Issue 1) kann auf Basis von Ontologie-basiertem Hypertext im RDF-Datenmodell realisiert werden. Berechnungen aufgrund der Hypertextstruktur (Issue 4) könnten durch den Einsatz von Dynamischer Netzwerkanalyse und Fuzzy Cognitive Maps (Carvalho 2012) umgesetzt werden.

Ergebnisse

Die ‚digitale Vernunft' erfordert es schließlich, daß geisteswissenschaftliches **Erklären** bzw. **Verstehen** (von Wright 1971) in einem angemessen methodologischen (Schützeichel 2015) und ontologischen Rahmen (Little 2010) abläuft. Mechanistische Erklärung bietet ein explanatorischen Rahmenwerk, um z. B. die verschiedenen ‚Ismen' der Politikwissenschaft zu vereinen (Bennett 2013).

Die Relevanz von Digital Mapmaking als Hilfsmittel zur Erklärung historischer Ereignisse zeigt sich insbesondere beim Einsatz von Hypertext Maps zur Strukturierung von historischen Narrativen. Historische Narrative können dabei durchaus als mechanistische Erklärungen historischer Ereignisse aufgefaßt werden (Glennan 2010). Analog zum Periodensystem chemischer Elemente kann ein diagrammatischer Hypertext auf Lücken im theoretisch fundierten Gerüst eines kausalen Narrativs hinweisen, um diese durch Methoden wie Process Tracing (Bennett 2010) zu füllen. Zusätzlich zu mechanistischer Erklärung unterstützt Hypertext multiperspektivische Erklärung historischer Ereignisse (Shaw 2013; Jensen 2013; Krameritsch 2009).

Die Ergebnisse bringt schließlich ein erneuter Blick in die Hypertext-Geschichte auf den Punkt: "Linearity was never an option for historical writing; hypertextuality can make complex structure concrete, clear and responsive to both the author and the reader." (Eastgate Systems 2005)

Zusammenfassung und Ausblick

Etablierte qualitative Methoden wie QCA, Process Tracing und Netzwerkanalyse werden in den Digital Humanities vernachlässigt oder, wie Netzwerkanalyse (Haug 2008), nicht kritisch genug eingesetzt – im Sinne von zu wenig formaler Modellierung (Lemercier 2015a). Der Einsatz dieser Methoden im Bereich der Digital Humanities setzt die formale Modellierung der geisteswissenschaftlichen Untersuchungsgegenstände voraus.

Die Rolle von formaler Ontologie für die Digital Humanities beim Aufbau von Regionalontologien für die Geisteswissenschaften (Gnoli 2008) und der Repräsentation von Wissen aus verschiedenen diziplinären Perspektiven und Kontexten wird an dieser Stelle deutlich (Frank 2017b). Bisher gibt es nur wenig Ansätze, die die Komplexität der geisteswissenschaftlichen Realität dabei angemessen berücksichtigen (Grossner 2010; Fokkens et al. 2016; Garbacz 2015).

Im Vortrag werden abschließend erste eigene formal-ontologische Modellierungsbeispiele und diagrammatische Visualisierungen kausaler Narrative historischer Ereignisse gemäß dem vorgestellten Ansatz präsentiert und offene Probleme und Fragen zur Diskussion gestellt.

Bibliographie

Behrendt, Hans R. / Burch, Thomas / Weinmann, Martin (2010): *Der Digitale Peters: Arno Peters' Synchronoptische Weltgeschichte*. Frankfurt am Main: Zweitausendeins.

Benardou, Agiatis / Constantopoulos, Panos / Dallas, Costis / Gavrilis, Dimitris (2010): A Conceptual Model for Scholarly Research Activity, in: *iConferencePapers 2010*.

Bennett, Andrew (2010): Process Tracing and Causal Inference, in: Brady, Henry E. (Hrsg.) / Collier, David (Hrsg.): *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Rowman & Littlefield Publishers.

Bennett, Andrew (2013): The mother of all isms: Causal mechanisms and structured pluralism in International Relations theory, in: *European Journal of International Relations* 19, 3: 459–481.

Brauer, Wilfried / Haacke, Wolfhart / Münch, Siegfried (1989): *Studien- und Forschungsführer Informatik*. Springer.

Carvalho, João P. (2012): *Rule Based Fuzzy Cognitive Maps in Humanities, Social Sciences and Economics*. 289–300, in: Seising, Rudolf (Hrsg.) / González, Veronica S. (Hrsg.): *Soft Computing in Humanities and Social Sciences*, Springer.

Carvalho, João P. (2013): On the Semantics and the Use of Fuzzy Cognitive Maps and Dynamic Cognitive Maps in Social Sciences, in: *Fuzzy Sets and Systems* 214: 6–19.

Champagne, Marc (2016): Diagrams of the Past: How Timelines Can Aid the Growth of Historical Knowledge, in: *Cognitive Semiotics* 9, 1: 11–44.

Coleman, Peter T. / Vallacher, Robin R. / Nowak, Andrzej / Bui-Wrzosińska, Lan (2006): Protracted Conflicts as Dynamical Systems, in: Kupfer Schneider, Andrea (Hrsg.) / Honeyman, Christopher (Hrsg.): *The Negotiator's Fieldbook: The Desk Reference for the Experienced Negotiator*. American Bar Association, (Section of Dispute Resolution) 61–74.

Crotty, Michael (1998): *The Foundations of Social Research: Meaning and Perspective in the Research Process*. SAGE Publications.

Eastgate Systems (2005): Hypertext and the Linearity of History. – URL http://www.eastgate.com/HypertextNow/archives/History.html

Engelbart, Douglas C. (1962): Augmenting Human Intellect: A Conceptual Framework. – Forschungsbericht.

Fokkens, Antske S. / ter Braake, Serge / Maks, Isa / Ceolin, Davide (2016): *On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change*, in: Darányi, S. (Hrsg.) / Hollink, L. (Hrsg.) / Meroño Peñuela, A. (Hrsg.) / Kontopoulos, E. (Hrsg.): Proceedings of Drift-a-LOD.

Frank, Ingo (2017a): Digital Humanities und Semiotik oder wie man die Unterstützung und Erweiterung geisteswissenschaftlichen Denkens durch Computerprogramme semiotisch erklären kann, in: Deutsche Gesellschaft für Semiotik (DGS) e. V. (Hrsg.): *15. Internationaler Kongress der Deutschen Gesellschaft für Semiotik (DGS)* 70. – Panel 4: Chancen und Grenzen Digitaler Geisteswissenschaften, Sektion Digital Humanities.

Frank, Ingo (2017b): Interdisciplinary Knowledge Organization as Intersection between Information Science and Digital Humanities, in: Burghardt, Manuel (Hrsg.) / Kattenbeck, Markus (Hrsg.): *ISI 2017 Satellite Workshop on the Relationship of Information Science and the Digital Humanities* 25–33.

Garbacz, Pawel (2015): Challenges for Ontological Engineering in the Humanities – A Case Study of Philosophy, in: Garoufallou, Emmanouel (Hrsg.) / Hartley, Richard J. (Hrsg.) / Gaitanou, Panorea (Hrsg.): *Metadata and Semantics Research*. Cham: Springer International Publishing 27–38.

George, Alexander L. / Bennett, Andrew (2005): Case Studies and Theory Development in the Social Sciences. MIT Press, (BCSIA Studies in International Security).

Glennan, Stuart (2010): Ephemeral Mechanisms and Historical Explanation, in: *Erkenntnis* 72, 2: 251–266.

Gnoli, Claudio (2008): Categories and Facets in Integrative Levels, in: *Axiomathes* 18, 2: 177–192.

Goertz, Gary / Mahoney, James (2005): Two-Level Theories and Fuzzy-Set Analysis, in: *Sociological Methods & Research* 33, 4: 497–538.

Grossner, Karl (2010): *Representing Historical Knowledge in Geographic Information Systems*, University of California, Santa Barbara, Dissertation.

Halasz, Frank G. (2001): Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems, in: *ACM Journal of Computer Documentation* 25, 3: 71–87.

Haug, Sonja (2008): Migration Networks and Migration Decision-Making, in: *Journal of Ethnic and Migration Studies* 34, 4: 585–605.

Jensen, Anthony K. (2013): *Nietzsche's Philosophy of History*. Cambridge University Press.

Kolmer, Lothar (2008): UTB Profile. Bd. 3002: *Geschichtstheorien*. Fink.

Krameritsch, Jakob (2009): Die fünf Typen des historischen Erzählens – im Zeitalter digitaler Medien, in: *Zeithistorische Forschungen* 3: 413–432. – URL http://www.zeithistorische-forschungen.de/3-2009/id=4566

Lemercier, Claire (2015a): Formal network methods in history: why and how? In: *Social Networks, Political Institutions, and Rural Societies*. Brepols 281–310. – URL https://halshs.archives-ouvertes.fr/halshs-00521527

Lemercier, Claire (2015b): Taking time seriously: How do we deal with change in historical networks?, in: Düring, Marten (Hrsg.) / Gamper, Markus (Hrsg.) / Reschke, Linda (Hrsg.): *Knoten und Kanten III*. Transcript Verlag 183–211. – URL https://hal.archives-ouvertes.fr/hal-01445932

Little, Daniel (2010): *Historical Concepts and Social Ontology*. 41–72, in: *New Contributions to the Philosophy of History*. Dordrecht: Springer Netherlands.

Mahoney, James (1999): Nominal, Ordinal, and Narrative Appraisal in Macrocausal Analysis, in: *American Journal of Sociology* 104, 4: 1154–1196.

Peroni, Silvio / Shotton, David (2012): FaBiO and CiTO: ontologies for describing bibliographic resources and citations, in: *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 17: 33–43. – URL http://speroni.web.cs.unibo.it/publications/peroni-2012-fabio-cito-ontologies.pdf

Peters, Arno / Peters, Anneliese (1952): *Synchronoptische Weltgeschichte*. Frankfurt am Main: Universum-Verlag.

Piotrowski, Michael (2016): *Digital Humanities, Computational Linguistics, and Natural Language Processing*. Lectures on Language Technology and History. – URL http://stp.lingfil.uu.se/~nivre/docs/michael_piotrowski_2016.pdf

Schützeichel, Rainer (2015): *Pfade, Mechanismen, Ereignisse. Zur gegenwärtigen Forschungslage in der Soziologie sozialer Prozesse*. 87–147, in: Schützeichel, Rainer (Hrsg.) / Jordan, Stefan (Hrsg.): Prozesse: Formen, Dynamiken, Erklärungen. Wiesbaden: Springer VS.

Shaw, Ryan (2013): A Semantic Tool for Historical Events, in: *Proceedings of the 1st Workshop on Events: Definition, Detection, Coreference, and Representation*, Association for Computational Linguistics 38–46.

Skocpol, Theda (1979): *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press.

Szostak, Rick (2004): *Information Science and Knowledge Management*. Bd. 7: *Classifying Science: Phenomena, Data, Theory, Method, Practice*. Springer Netherlands.

von Wright, Georg H. (1971): *Explanation and Understanding*. Cornell University Press.

Digital HUMANities - Eine benutzerzentrierte Perspektive

Mayr, Eva
eva.mayr@donau-uni.ac.at
Donau Universität Krems, Österreich

Schreder, Günther
guenther.schreder@donau-uni.ac.at
Donau Universität Krems, Österreich

Windhager, Florian
florian.windhager@donau-uni.ac.at
Donau Universität Krems, Österreich

Wenn von Digital Humanities (DH) die Rede ist, liegt der Fokus oft auf dem Digitalen, auf den neuen Möglichkeiten, welche die technischen Entwicklungen der letzten Jahre eröffnen. Jedoch sollte bei Digital Humanities nicht primär digital im Vordergrund stehen, sondern die Human- und Geisteswissenschaften, die sich digitaler Methoden zur Unterstützung ihrer wissenschaftlichen Forschung bedienen (vgl. Siemens, 2016). Im Zentrum dieser Forschung stehen trotz digitaler Optionen dennoch die etablierten Fragestellungen der Humanwissenschaften. Technische Entwicklungen können neue Wege der Erkenntnis eröffnen oder bestehende Methoden vereinfachen und erleichtern, doch ohne fundierten ExpertInnen der Geisteswissenschaften in diesem Prozess eine zentrale Rolle zuzugestehen, können digitale Forschungs-Systeme die Bedürfnisse und Zielsetzungen ihrer BenutzerInnen nicht (gut genug) unterstützen und werden auch nicht nachhaltig von diesen aufgenommen.

Welche Möglichkeiten gibt es, den Einfluss von prä-, non-, oder postdigitalen Konzeptionen der Geisteswissenschaft in den DH zu stärken? Wie können Projekte in den DH weniger technologie- und stärker menschen- oder inhaltsgetrieben geplant und durchgeführt werden? Eine Antwort darauf können benutzerzentrierte Gestaltungsprozesse (“user centered design”) geben, in denen den BenutzerInnen der Systeme eine zentrale Rolle zukommt - von der Planung bis zur Evaluation der technologischen Entwicklungen.

Im Folgenden werden die Grundprinzipien und Methoden des benutzerzentrierten Designs erörtert und ein aktuelles Projekt - als Anwendungsfall eines benutzerzentrierten Designprozesses - vorgestellt.

Benutzerzentrierte Gestaltungsprozesse

Im Gegensatz zu technikzentrierter Entwicklung, steht in benutzerzentrierten Gestaltungsprozessen die BenutzerIn, ihre Bedürfnisse und Aufgaben, ihr Fühlen und Denken im Vordergrund. Ausgangspunkt sind daher auch nicht die technischen Möglichkeiten, sondern eine Analyse der Zielgruppe: Welche Eigenschaften und welche Arbeitsweisen zeichnet sie aus? Für welche Probleme bedarf es einer technischen Lösung? Erst danach werden geeignete Technologien entwickelt und laufend unter wiederholter Einbeziehung der Zielgruppe getestet.

Maguire (2011) definierte vier Schlüsselprinzipien für benutzerzentriertes Design: (1) das aktive Miteinbeziehen der BenutzerInnen, sowie ein klares Verständnis der Zielgruppe und ihrer Bedürfnisse, (2) eine geeignete Aufteilung der Funktionen und Prozesse zwischen Benutzer und System, (3) iterative Entwicklung und Testung der Technologie, und (4) Zusammenarbeit in einem inter- bzw. transdisziplinären Team.

Was bedeutet das umgelegt auf DH? Es gibt Stimmen, die generell daran zweifeln, dass Benutzer ihre Bedürfnisse verbalisieren können bzw. dieses Wissen von Nutzen für technologische Entwicklungen ist. Kemman und Klappe (2014) befragten daher Geisteswissenschaftler nach ihren Anforderungen für eine DH-Anwendung. Sie stellten fest, dass diese ihre Bedürfnisse gut verbalisieren konnten, sie fanden aber auch etliche (aus ihrer Sicht) irrelevante Bedürfnisse und vermissten eine Vorstellungskraft dafür, welche Möglichkeiten über den Standard-Forschungsprozess hinaus mithilfe neuer Technologien erschlossen werden könnten. Unserer Meinung nach bedarf es daher anderer Methoden als einer reinen Befragung nach den Bedürfnissen der Benutzer: Möglichkeiten für eine solche erweiterte Bedarfsanalyse sind Beobachtungen der Forschungsprozesse, aber auch Literaturstudien. Um die technischen Möglichkeiten und die geisteswissenschaftlichen Bedürfnisse in die Definition der Anforderungen mit einzubeziehen, bedarf es innovativer Ansätze, wie etwa Design-Sprints (Venturini, Munk & Meunier, 2017), in denen alle Beteiligten kollaborativ eine gemeinsame Perspektive entwickeln (Vertreter der Zielgruppe, Geisteswis-

senschaftlerInnen, DH-ExpertInnen und ComputerwissenschaftlerInnen). Da in der Zusammenarbeit zwischen ComputerwissenschaftlerInnen und GeisteswissenschaftlerInnen zwei sehr heterogene Welten aufeinander prallen (unterschiedliche Wissenschaftskultur, Terminologien, Epistemiken, sowie unterschiedliche Wege des Erkenntnisgewinns), ist es offensichtlich entscheidend, in einen aktiven und strukturierten Dialog miteinander zu treten, Herausforderungen arbeitsteilig zu lösen, den Wissensaustausch zu fördern, aber auch idealerweise diesen Prozess durch Personen zu mediieren, die in beiden Kulturen sozialisiert und fachsprachlich versiert sind.

Benutzerzentrierte Gestaltung definiert dabei nur ein Vorgehensmodell, in dem die Bedürfnisse der BenutzerInnen am Anfang stehen und in dem diese wiederholt einbezogen werden, es bestimmt jedoch nicht die zur Anwendung kommenden Forschungsmethoden, sondern bedient sich der jeweils passenden Methoden, zum Beispiel aus der Usability-Forschung: Benutzertests mit Prototypen, Beobachtungen, Befragungen, Fokusgruppen, cognitive walkthroughs, lautes Denken oder heuristische Evaluationen (siehe z.B. Barnum, 2008; Sarodnik & Brau, 2006). Im Folgenden soll die Spezifikation dieses Prozesses anhand einer Fallstudie zur Visualisierung kultureller Sammlungen exemplifiziert werden.

Fallstudie: Visualisierung kultureller Sammlungen

In den letzten 10 Jahren wurden kulturelle Sammlungen im großen Stil digitalisiert und aggregiert (z.B. Europeana, DPLA) mit dem Ziel die Zugänglichkeit zum kulturellen Erbe zu verbessern. Diese digitalen Sammlungen stellen jedoch für Benutzer ohne fachliche Expertise große Hürden dar (Walsh & Hall, 2015): Die Oberflächen sind zumeist von einer Suchfunktion dominiert, die eine Kenntnis der Datenbankstruktur voraussetzt, und das Suchergebnis wird als unstrukturierte Liste präsentiert, was das Gewinnen eines Überblicks über die Sammlung erschwert. Neuere Ansätze fordern daher “großzügigere” Benutzeroberflächen (Whitelaw, 2015), die ein “Flanieren” durch die Informationen ermöglichen (Doerk et al., 2011). Das Projekt *polycube* (Windhager et al., 2016) entwickelt diesen Anforderungen folgend Informationsvisualisierung von kulturellen Sammlungen zur Verbesserung der Zugänglichkeit und des Verständnisses dieser Sammlungen für die allgemeine Bevölkerung.

Für eine bessere Einbettung dieses Forschungsvorhabens wurde der aktuelle Stand der Technik erhoben (Windhager et al., 2017) und 48 Publikationen zur Visualisierung kultureller Sammlungen unter anderem mit Hinblick auf die Einbeziehung der Zielgruppe in den Gestaltungsprozess bewertet: In 6 Publikationen wurden rein technische Aspekte besprochen, Zielgruppen wurden dabei nicht genannt. In 17 Publikationen wurden zwar die Zielgruppen erwähnt, aber es fanden sich keine Informationen, ob und wie sie in die Entwicklung einbezogen wurden. In 5 Publikationen wurden (geplante) Usertests erwähnt – allerdings ohne weitere Details. Nur in 20 Publikationen wurde von Studien berichtet, die in ihrem Umfang stark variieren - von Einzelfallanalysen bis hin zu elaborierten Testungen mit großen Benutzergruppen (vgl. Abbildung 1).

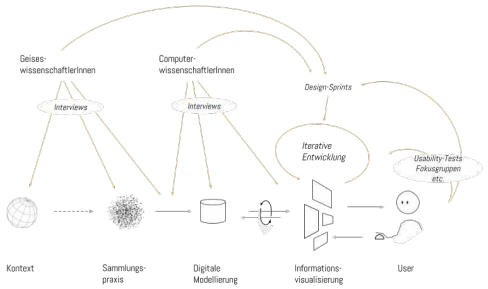


Abbildung 1: Überblick über verschiedene Methoden zur Einbeziehung der BenutzerInnen

Im Projekt *polycube* (https://www.donau-uni.ac.at/de/polycube) stand am Anfang eine Data-Users-Tasks-Analyse (Miksch & Aigner, 2014). Es wurden ExpertInneninterviews mit HistorikerInnen und DH-ForscherInnen durchgeführt zu den *Daten*: Welche Zusammenhänge gibt es zwischen den Objekten der Sammlung? Welche Daten sind digital vorhanden, welche Informationen fehlen in der digitalen Datenbank? Eine Literaturstudie widmete sich den *BenutzerInnen* und ihren *Aufgaben* (Mayr et al., 2016a): Da es sich um alltägliche NutzerInnen handelt, gestaltete sich die Definition von relevanten Aufgaben als besonders herausfordernd. Stattdessen wurden relevante Informationsbedürfnisse und Verhaltensmuster definiert.

Auf diesen Erkenntnissen aufbauend wurde ein erster Prototyp entwickelt, der derzeit in einer qualitativen Studie getestet wird. Im Mittelpunkt steht dabei die Frage, wie die entwickelten Infor-

mationsvisualisierungen die BenutzerInnen beim Aufbau eines mentalen Modells über die kulturelle Sammlung unterstützen (Mayr et al., 2016b). Iterativ werden die gewonnenen Erkenntnisse in die Entwicklung des Prototypen einfließen und in zwei weiteren Experimenten gegen alternative Informationsvisualisierungen getestet werden.

Das gewählte benutzerzentrierte Vorgehen soll dazu beitragen, dass die entwickelten Informationsvisualisierungen ein besseres Verständnis der kulturellen Sammlungen vermitteln, damit intuitiv interagiert werden kann und diese zu einer weiteren Auseinandersetzung mit den Informationen anregen.

Diskussion

In der Abgrenzung zu den nicht-digitalen Geisteswissenschaften versuchen die DH "die Prozesse der Gewinnung und Vermittelung neuen Wissens unter den Bedingungen einer digitalen Arbeits- und Medienwelt weiter zu entwickeln" (DHd). Dabei agieren ihre AkteurInnen nicht selten mit einem Fokus auf Technik- und Infrastrukturentwicklung statt mit einem Fokus auf geisteswissenschaftliche Prozesse und Methoden des Erkenntnisgewinns. Um diesen Prozessen einen stärkeren Stellenwert in DH-Projekten einzuräumen, haben wir in diesem Beitrag benutzerzentrierte Gestaltungsprozesse als eine Herangehensweise diskutiert, in der geisteswissenschaftliche ExpertInnen und andere Zielgruppen in die Entwicklung neuer Technologien intensiv miteinbezogen werden.

Warum ist ein benutzerzentrierter Gestaltungsprozess gerade in der DH von Vorteil?

1. Unterschiedliche Wege des Erkenntnisgewinns und Forschungsmethoden in den Geisteswissenschaften und den Computerwissenschaften erschweren ein Verständnis der Probleme und Anliegen der jeweils anderen Disziplinen. Durch die intensive Zusammenarbeit und Koordination in einem benutzerzentrierten Designprozess können die Beteiligten in einen intensiven Wissensaustausch eintreten und ein besseres Verständnis füreinander aufbauen.
2. Die Definition von zu lösenden Problemen aus Sicht der GeisteswissenschaftlerInnen am Beginn eines Projektes erlaubt die Entwicklung von innovativen Technologien im Dienste der Geisteswissenschaft, anstatt geisteswissenschaftliche Daten als Anwendungsfeld für neue technische Entwicklungen zu instrumentalisieren.
3. Die Iteration von Entwicklungen und Testungen führen zu einer regelmäßigen Evaluation der Technologien in verschiedenen Entwicklungsstadien und ermöglichen die Korrektur von Fehlentwicklungen bereits früh im Projektverlauf. Im Gegensatz dazu bleiben summative Evaluationen am Ende von Projekten oft ohne Einfluss auf die entwickelte Technologie bzw. erlauben nur mehr geringfügige Adaptierungen. Eine Iteration von Entwicklungs- und Evaluationsphasen erhöht die Anzahl der explorierten Design-Optionen und erlaubt die Auswahl und Weiterentwicklung der am besten geeigneten Varianten.

Benutzerzentriertes Design ist aber nicht immer das Mittel der Wahl. Um etwa radikal innovative Produkte zu erschaffen, besitzen zukünftige BenutzerInnen oft nicht die Vorstellungskraft, welche Möglichkeiten sich durch die neuen technischen Entwicklungen ergeben (Norman, 2010; vgl. die Beobachtungen von Kemman & Klappe, 2014). Hier empfiehlt es sich, die Benutzer erst in der Optimierung der Produkte mit einzubeziehen. Auch zur Lösung von institutionellen, finanziellen, oder politischen Problemen sollten statt benutzerzentrierter eher transdisziplinäre Methoden gewählt werden.

Unsere Erfahrung in DH Projekten zeigt, dass der methodengestützte Dialog zwischen Computer- und Geisteswissenschaften essentiell ist und sein Potenzial für gute und nachhaltige technische Entwicklungen in benutzerzentrierten Gestaltungsprozessen besonders gut entfalten kann. Die präsentierte Analyse von Publikationen zur Visualisierung kultureller Sammlungen (Windhager et al., 2017) zeigt, dass eine Einbeziehung der BenutzerInnen in den DH keine Selbstverständlichkeit ist und dass eine benutzerzentrierte Entwicklung derzeit nicht zum Stand der Technik gehört. Das Potenzial dieses Vorgehensmodells ist jedoch sehr groß, wenn es darum geht die Bedürfnisse der NutzerInnen zu erfüllen und die Technologie soweit daran anzupassen, dass deren Akzeptanz und nachhaltige Nutzung sichergestellt werden kann.

Danksagung

Die beschriebene Arbeit wurde durch den Wissenschaftsfonds FWF P.No. P28363 gefördert.

Bibliographie

Barnum, Carol M. (2008): *Usability testing and research*. Allyn & Bacon Series in Technical Communication, Longman.

Dörk, Marian / Carpendale, Sheelagh / Williamson, Carey (2011): "The information flaneur: A fresh look at information seeking", in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1215-1224.

Kemman, Max / Kleppe, Martijn (2014, October): "Too many varied user requirements for digital humanities projects", in: *3rd CLARIN ERIC Annual Conference* 24-25.

Maguire, Martin (2001): "Methods to support human-centred design", in: *International Journal of Human-Computer Studies, 55*, 587-634.

Mayr, Eva / Federico, Paolo / Miksch, Silvia / Schreder, Günther / Smuc, Michael / Windhager, Florian (2016a): "Visualization of cultural heritage data for casual users", in: *Proc. of the 1st IEEE VIS Workshop on Visualization for the Digital Humanities*. Baltimore, MD.

Mayr, Eva / Schreder, Günther / Smuc, Michael / Windhager, Florian (2016b): "Looking at the representations in our mind: Measuring mental models of information visualizations", in: *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization*. ACM, 96-103.

Miksch, Silvia / Aigner, Wolfgang (2014): "A matter of time: Applying a data-users-tasks design triangle to visual analytics of time-oriented data", in: *Computers & Graphics 38*: 286-290.

Norman, Donald A. (2010). "Technology first, needs last: The research-product gulf", in: *Interactions 17*: 38-42.

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

Siemens, Roy (2016): "Communities of practice, the methodological commons, and digital self-determination in the humanities", in: C. Crompton, R. J. Lane, & R. Siemens (Eds.), *Doing Digital Humanities: Practice, training, research*. Milton Park: Routledge xxi-xxxiii.

Venturini, Tommaso / Munk, Anders / Meunier, Axel (2017): "Data-sprints: A public approach to digital research", in: C. Lury / P. Clough / M. Michael / R. Fensham / S. Lammes / A. Last / E. Uprichard (Eds.), *Routledge Handbook on Interdisciplinary Research Methods*. Forthcoming.

Walsh, David / Hall, Mark M. (2015): "Just looking around: Supporting casual users initial encounters with Digital Cultural Heritage", in: *Supporting Complex Search Tasks*. Vienna: CEUR-WS 1338.

Whitelaw, Mitchell (2015): "Generous interfaces for digital cultural collections", in: *Digital Humanities Quarterly 9*(1).

Windhager, Florian / Mayr, Eva / Schreder, Günther / Smuc, Michael / Federico, Paolo / Miksch, Silvia (2016): „Reframing cultural heritage collections in a visualization framework of space-time cubes", in: *Proceedings of the 3rd HistoInformatics Workshop*. CEUR- WS 1632.

Windhager, Florian / Federico, Paolo / Schreder, Günther / Glinka, Katrin / Doerk, Marian / Miksch, Silvia / Mayr, Eva (2017): „Visualization of cultural heritage collection data: State of the art and future challenges", *Manuscript under review*.

Dokumentenarbeit mit hierarchisch strukturierten Texten: Eine historisch vergleichende Analyse von Verfassungen

Knoth, Alexander
alexander.knoth@uni-potsdam.de
Universität Potsdam, Deutschland

Stede, Manfred
stede@uni-potsdam.de
Universität Potsdam, Deutschland

Hägert, Erik
haegert@uni-potsdam.de
Universität Potsdam, Deutschland

1. Einleitung: Verfassungsvergleich als Spiegel staatlichen Wandels?

Staatlich verfasste Gesellschaften sind komplex und differenziert (Mayntz 1997; Schimank 1999). Will man etwas über die „Identität" von Staaten und deren Wandel erfahren, dann eignen sich Verfassungen, da diese spezfische Dokumentensorte soziologisch als kodifizietre Selbstbeschreibung von Gesellschaften verstanden werden kann (Boli-Bennett und Meyer 1978; Go 2003; Heintz und Schnabel 2006; Boli-Bennett 1979).

Moderne Staaten produzieren aber nicht nur Unmengen an amtlichen Dokumenten, sondern sind darüber hinaus durch ihre konstitutionelle sowie rechtsstaatlich-bürokratische Verfasstheit grundsätzlich textlich strukturiert (Weber 1972). Für die historsiche Dokumentenanalyse spielen die Erstellung des Korpus und die Auswahl der Untersuchungsmethoden wichtige Rollen, um sowohl die Textlichkeit, als auch den Kontext angemessen zu berücksichtigen.

In diesem Vorhaben wird mit Verfassungsdokumenten europäischer Staaten gearbeitet, anhand derer staatlicher Wandel von der ersten Verfassungsgebung bis heute sichtbar gemacht wird. Verfassungen beinhalten u.a. Vorstellungen darüber, wie die Gesellschaft beschaffen ist. Konkret wird der historisch-wissenssoziologischen Frage nach der sozialen Konstruktion des (Staats-)Bürgers nachgegangen. Denn historisch betrachtet reflektieren Verfassungen den sukzessiven Umbau von ständisch stratifizierten hin zu souveränen Bürgergesellschaften und damit reflektieren sie ebenso den Wandel des gesellschaftlichen Personals, das es in Form von Personenkategorien und Zugehörigkeitsdimensionen aus dem Material herauszuarbeiten gilt. Was aber genau unter „Verfassung“ verstanden wurde, wie sich die Staaten und ihr ‚Personal’ über dieses Dokument selbst beschreiben und welches Wissen in selbiges eingeht, variiert erheblich (Gosewinkel, Masing und Würschinger 2006; Vorländer 2007). Daher bedarf es eines geeignetne methodisch-analytischen Instrumentariums, um die aufgeworfene Fragen zu beantworten.

2. Dokumentenanalyse im Schnittfeld von historischer Soziologie und Computerlinguistik

Um Verfassungen strukturell und inhaltlich untersuchen zu können, werden Ansätze der historischen Wissenssoziologie (Thelen 1999, 2002; Jepperson 1991) und der Computerlinguistik (z.B. Hausser 2014; Lobin 2010) miteinander verschränkt. Die Entwicklung dieses methodischen Werkzeugs zur „Dokumentenarbeit“ umfasst Verfahrensschritte der Datenerhebung, -aufbereitung und -auswertung, wobei hier vor allem auf methodologische Herausforderungen, d.h. die Korpuserstellung und die semi-automatische Analyse von Dokumentenstrukturen eingegangen wird.

Rechtstexte im Allgemeinen und Verfassungen im Besonderen, weisen eine Dokumentenlogik auf, die stark durch eine formale hierarchische Struktur gekennzeichnet ist. Bei dieser Dokumentenart ist daher davon auszugehen, dass der Struktur eine besonders sinnstiftende Bedeutung zukommt, die es vor allem bei vergleichenden Untersuchungen (synchron wie auch diachron) zu berücksichtigen gilt. Insofern sollte ein computerlinguistisches Verfahren die Strukturinformationen bspw. in welche (Sinn-)Abschnitte sich ein Dokument gliedert für den Vergleich nutzen. Diese Strukturauswertungen können dann wiederum mit statistischen Häufigkeits- und Ähnlichkeitsberechnungen von Worten innerhalb von Fließtexten – wie das u.a. die gängigen Vektorraummodelle (Manning, Raghavan und Schütze 2008; Salton, Wong und Yang 1975) oder insbesondere die derzeit populären „word embedding“ Modell (z.B. Mikolov et al. 2013) machen – kombiniert werden.

Bei der hierarchischen Struktur von Dokumenten anzusetzen bietet einen klaren Ausgangspunkt für die systematische Analyse von großen Textmengen und stiftet zugleich Orientierung im Feld der inhaltsanalytischen Methoden (Kuckartz 2012; Mayring 2015). Diese unterscheiden sich vor allem in Bezug auf ihre Anlage, d.h. entweder Häufigkeiten zählende oder hermeneutisch interpretierende Ausrichtung, und firmieren in den Sozialwissenschaften oftmals unter dem Label „Dokumentenanalyse“. Zwar verbindet alle diese Ansätze, dass sie sich durch eine ständige Korrespondenz von Forschungsfrage und Arbeit am Material auszeichnen und in der Regel mehrere Iterationen durchlaufen, bevor valide Ergebnisse vorliegen. Dennoch bringt vornehmlich die manuelle Bearbeitung von umfangreichen Textmengen, etwa in Form von Kodier- und Kategorisierschritten der *Grounded Theory* (Strauss und Corbin 1996), Probleme der methodisch kontrollierten Auswertung und damit der Reliabilität der Ergebnisse mit sich.

Außerdem lassen sich über den strukturellen Zugang Fragen erschließen, die über den „reinen“ Inhalt hinausgehen, und die die Verwendung wie auch die Art und Weise in den Vordergrund rücken, in der Verfassungen im Zeitverlauf politisch unter Druck geraten, sich also aufgrund wechselnder politischer Machtverhältnisse wandeln. Damit werden die Relation von Dokument und (Entstehungs-)Kontext und besonders die Verfasser von Dokumenten und deren Konstruktion sozialer Wirklichkeit durch die schriftliche Fixierung gesellschaftlichen Wissens (Prior 2011) fokussiert.

Von der Dokumentenstruktur auszugehen heißt, zunächst methodologisch zu fragen, inwieweit sich Dokumente formal wie auch inhaltlich ähneln. Das setzt wiederum voraus, dass

sich Dokumente überhaupt vergleichen lassen und so einer Analyse etwaiger struktureller Ähnlichkeiten und spezifischer Unterschiede allererst zugänglich gemacht werden. Hier bieten computergestützte Verfahren einen produktiven Ausgangspunkt, um einerseits bestehende Methoden zu reflektieren und andererseits ein dann auch verallgemeinerbares Werkzeug zur Dokumentenanalyse von Rechtstexten zu entwickeln. Zur Beantwortung der formulierten Frage wird eine innovative, von uns eigens entwickelte Software vorgestellt, mit der sich Verfassungen in ihrer historischen Entwicklung vergleichen lassen. Hierdurch werden Impulse zur Generierung neuer methodischer Ansätze gegeben werden.

3. Arbeitsschritte: Vom Download zur Analysesoftware

Um mit der Auswertung der Dokumente beginnen zu können, muss das Korpus erstellt werden. Ein normales PDF beinhaltet in der Regel kaum explizite Strukturinformationen, lediglich einzelne Worte ließen sich automatisch erfassen, nicht aber die zugrunde liegenden Strukturen abbilden. Hierfür bedürfte es bspw. der Kennzeichnung von Überschriften, Absätzen oder inhaltlich unterscheidbaren Abschnitten. Das Dokument muss also mit weiteren Informationen in seiner Struktur beschrieben werden.

Das konkrete methodische Vorgehen, das hier als „Dokumentenarbeit“ zur Korpuserstellung bezeichnet wird, gliedert sich in drei Schritte: das Zusammenstellen der Ausgangsdaten als HTML-Dateien, die Transformation der Ausgangsdaten in das XML-Format sowie die eigentliche Auszeichnung der Ausgangsdaten mit Metadaten zur Modellierung der Struktur.

Entgegen der Annahme, dass derart politisch relevante Dokumente wie Verfassungen als elektronische Ausgangsdaten vorliegen sollten, müssen diese zunächst hergestellt und in ein bearbeitbares Datenformat transformiert werden. Zwar finden sich aktuelle Verfassungen als elektronisch veröffentlichte Ressourcen bspw. in den Rechtsdatenbanken und -portalen der jeweiligen Staaten, im deutschen Fall bspw. „Juris“.Es existiert jedoch kein lückenloser, chronologischer Verlauf, aus dem sich alle Änderungen computergestützt entnehmen ließen.

Auf der Seite www.verfassungen.org lassen sich die meisten Verfassungen online (auf Deutsch) abrufen und downloaden. Zudem beinhalten die dortigen Dokumente farblich abgesetzte Änderungen in Textform und nicht etwa als Kommentar oder gesonderte Liste sowie jeweils Totalrevisionen als separate Dokumente. Diese Dokumente werden dann mit offiziell veröffentlichen Verfassungen abgeglichen, um gegebenenfalls inhaltliche Fehler aufzuspüren und zu beheben. Anschließend werden die Daten manuell bereinigt und standardisiert, d.h. nicht benötigte Beschreibungen der Autor*innen der Webseiten oder andere irrelevante Informationen werden entfernt. Diese Dokumente bilden sodann die Grundlage für das Korpus. Die einzelnen Verfassungen beinhalten eine Fülle an textlichen Ergänzungen, Streichungen und anderweitigen textlichen Veränderungen, die einerseits schwer zu identifizieren sind und andererseits nicht chronologisch sortiert, sondern der Texthierarchie folgend vorliegen. Aus diesen Gründen wird zuerst eine Ausgangsverfassung des Jahres 2011, also dem Ende des Untersuchungszeitraums, erstellt, um ausgehend davon jede weitere Änderung als eigenständige, nicht offizielle, „Phantom-Verfassung“ zu rekonstruieren. Durch diesen iterativen, historischen Rekonstruktionsprozess wird schließlich die Datengrundlage geschaffen.

Die Verfassungen liegen zunächst als HTML-Dokumente, mit einer sehr flachen Dokumentenstruktur vor. Die zu entwickelnde Analysesoftware benötigt jedoch ein Datenformat, das Metadaten mit Dokumentendaten assoziieren kann. Dem aktuellen Entwicklungsstand entsprechend verwenden wir ein XML-Format (Bubenhofer und Scharloth, 2015).

Deshalb wird im nächsten Schritt der HTML-Code mittels eines XSL-Skripts (Extensible Stylesheet Language) in das technische XML Format (vgl. XML Schema 2001; XQuery 2002; XSLT 1999) überführt und formatiert. Bei XSL bzw. XSLT handelt es sich um eine Programmiersprache zur Transformation (und Formatierung) von XML Derivaten – in unserem Fall die HTML-Dateien – in XML Dokumente. Das XML Format eignet sich in erster Linie dafür, die informationsarmen Ausgangsdaten mit Metadaten (z.B. Attribute, Codes oder Variablen) zur systematischen Beschreibung der Strukturen und der Inhalte anzureichern. Bspw. ließe sich das Ausgangsdatum „Herbert“ mit dem Attribut „Vorname“ verknüpfen und so systematisch alle Vornamen erschließen.

Die Umwandlung von HTML zu XML ist die Grundlage des Mappings der einzelnen Versionen auf einander. Hierfür wurde kein existierender Standard verwendet, sondern das Format so entwickelt, dass es die Struktur der Texte möglichst treu abbildet. Dafür sollen möglichst wenige Elemente verwendet und unnötig tiefe Einbettungen vermieden werden.

Jeder Version wird ein Vorspann vorangestellt (<front>) der den Titel (<docTitle>) und das Datum der jeweiligen Version (<docEdition>) enthält.

Diesem Vorfeld folgt der eigentliche Text der Verfassung (<body>). Dieser ist zunächst in Hauptteile gegliedert (<div n="" type="v-teil”>). Diese können wiederum aus Sektionen (<div n="" type="sektion”>) bestehen, welche die Artikel ( ) der Verfassung enthalten. Die Artikel setzen sich aus Sätzen ( ) und gegebenenfalls auch aus listenartigen Aufzählungen (<div n="" type="aufzaehlung”>) zusammen. Letztere bestehen aus einer Reihe von Listenelementen ( ).

Schwesterknoten werden mit eins anfangend durchnummeriert (n). Ein Hauptteil mit n="0" ist eine Präambel. Diese enthalten keine Artikel, sondern eine Reihe von Sätzen ( ) und gegebenenfalls Aufzählungen. Hauptteile, Sektionen und Artikel weisen jeweils ein Element für ihre Überschriften auf. Darüberhinaus enthalten nur Paragraphenelemente ( ) Text. Eine Validierung gegen eine DTD findet nicht statt.

Es ergibt sich folgendes Format:

```
<text>
<front>
<titlePage>
<docTitle></docTitle>
<docEdition></docEdition>
</titlePage>
</front>
<body>
<div n="0" type="v-teil">
<head></head>
<p n="1" type="praeamb-satz"></p>
<div n="1" type="aufzaehlung">
<p n="1" type="aufzaehlung-item"></p>
</div>
<div>
<div n="1" type="v-teil">
<head></head>
<div n="1" type="artikel">
<head></head>
<p n="1" type="satz"></p>
<div n="1" type="aufzaehlung">
<p n="1" type="aufzaehlung-item"></p>
</div>
</div>
<div n="1" type="sektion">
<div n="1" type="artikel">
<head></head>
<p n="1" type="satz"></p>
</div>
</div>
</div>
<body>
<text>
```

An dieser Stelle der Beschreibung der Ausgangsdaten im XML Format setzen wiederum qualitative Dokumentenarbeitsschritte ein, die sich an der analytischen Strategie des Kodierens und Kategorisierens anlehnen. In diesen Vorgang fließen einerseits Kontextinformationen ein, andererseits werden während des Arbeitsschritts wichtige empirische Beobachtungen gemacht, die in Form von Kodiermemos dokumentiert werden. So können die gewonnenen Informationen zu einem späteren Zeitpunkt für die Tiefenstrukturanalyse des Materials oder die Ausdifferenzierung der Analysesoftware genutzt werden.

Mapping als Strukturvergleich

Das Mapping der Strukturelemente der im Vergleich stehenden Versionen aufeinander wird automatisiert vollzogen, indem jedes Element einer strukturellen Ebene (Hauptteil, Sektion, Artikel) mit jeder anderen entsprechenden Ebene der Vergleichsversion abgeglichen wird. Dafür verwenden wir das gängige Cosinus-Maß, das Text-Ähnlichkeit durch Modellierung im hochdimensionalen Vektorraum misst. Auf diese Weise wird eine Matrix von Ähnlichkeitswerten aufgebaut.

Währen des Aufbaus der Matrix wird die Anzahl der Berechnungen reduziert, indem, sobald eine Ähnlichkeit vom Wert 1 zu einem Element der Vergleichsversion gefunden wird, also eine perfekte Übereinstimmung vorliegt, das Elementpaar als unveränderte Übereinstimmung abgespeichert und die Elemente aus dem weiteren Vergleich ausgeschlossen werden.

Liegt keine genaue Übereinstimmung vor, wird getestet, ob die beiden zu vergleichenden Texte unterschiedlicher Länge sind. Ist dies der Fall, wird ferner geprüft, ob der kürzere der beiden Texte einen Teilstring des längeren bildet. In solchen Fällen werden die Texte einander als Änderungen (Erweiterungen oder Kürzungen) zugeordnet, abgespeichert und ebenfalls aus der weiteren Berechnung ausgeschlossen.

Schließich bleibt eine Matrix der Ähnlichkeitswerte ausschließlich der Elemente übrig, für die keine Entsprechung gefunden werden konnte.

Die Paare, die die höchsten Cosinusähnlichkeiten (< 1) aufweisen, werden als Änderungen abgespeichert. Die hiernach übrig bleibenden Elemente, denen kein Element der Vergleichsversion zugeordnet werden konnte, sind entweder Tilgungen (keine Entsprechung in der zeitlich späteren Version) oder Hinzufügungen (keine Entsprechung in der früheren Version). Abbildung 1 illustriert den Prozess in tabellarischer Form.

1.

	Art.1	Art2	Art.3	Art.4	Art.5
Art.1	1.0	–	–	–	–
Art.2	–	–	–	–	–
Art.3	–	–	–	–	–
Art.4	–	–	–	–	–

$Art.1 \Rightarrow Art.1$

2.

	Art2	Art.3	Art.4	Art.5
Art.2	0.1	–	–	–
Art.3	–	–	–	–
Art.4	–	–	–	–

$Art.2 \Rightarrow Art.2$

3.

	Art.3	Art.4	Art.5
Art.3	0.8	0.4	0.6
Art.4	0.3	0.7	0.9

4.

$Art.3 \Rightarrow Art.3$
$Art.4 \Rightarrow Art.5$
Neu: Art.4

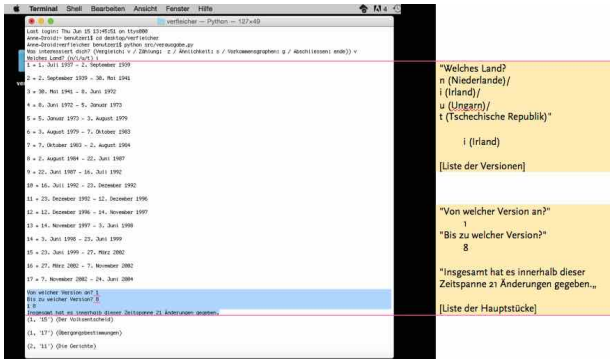
Entwicklungsergebnisse

Die Software zur Verfassungsanalyse ist in der Programmiersprache *Python* geschrieben. Im Zuge der ersten Entwicklungsphase lassen sich rein formal die hierarchische Struktur und die jeweiligen Abschnittslängen der Dokumente vergleichen und auch quantifizieren.

In der vorliegenden Fassung des Werkzeugs können vier verschiedene Operationen ausgeführt werden:

1. Vergleichen

Für den Vergleich wird zunächst ein Land und ein Zeitraum ausgewählt für den die Änderungen ausgegeben werden sollen. Um die Suche weiter einzuschränken, wird zunächst gezeigt, wie viele Änderungen in welchen Hauptstücken in dem angegebenen Zeitraum stattgefunden haben. Nach der Auswahl eines Hauptteils werden die gefundenen Änderungen (in Sektionen und Artikeln), Tilgungen und Hinzufügungen ausgegeben (Abbildung 2).

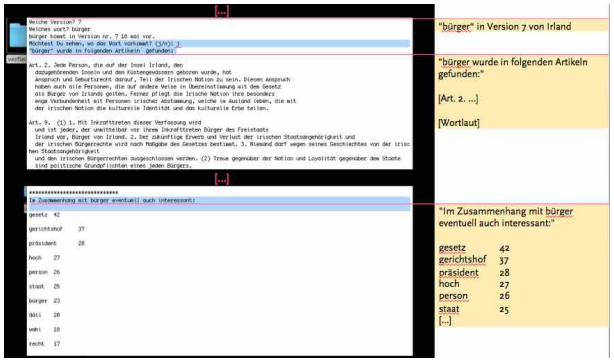


2. Cosinusähnlichkeiten

Mit dieser Funktion (Abbildung 3) lassen sich die Cosinusähnlichkeiten ganzer Versionstexte untereinander berechnen und ausgeben.

3. Wortzählungen und Wortprofile

Der Nutzer kann sich unter Angabe der Version, die von Interesse ist, die Aufkommenshäufigkeiten von Wörtern ausgeben lassen. Zudem lassen sich die Textstellen, die das gezählte Wort enthalten, zusammen mit einer Liste der Wörter ausgeben, die häufiger als ein definierter Schwellenwert (bspw. fünf Mal) in derselben textuellen Umgebung vorkommen (Abbildung 4).



Das Beispiel zeigt die politische Kernkategorie, den Bürger, in Version 7 der irländischen Verfassung. Die textuelle Umgebung ist durch Begriffe wie Gesetz (42 mal), Gerichtshof (37 mal) und Präsident (28 mal), die Hinweise dazu liefern in welchen Sinnzusammenhängen der Bürger theamtisiert wird, gekennzeichnet. Die Begriffe Person (26 mal) und Staat (25 mal) weisen darauf hin, dass es sich beim Bürger offenbar tatsächlich um eine Kategorisierung als Person handelt, die wiederum in irgendeiner Beziehung zum Staat steht. Dieser Zusammenhang, die Beziehung von Bürger und Staat kann nun mithilfe von (historischen) Kontextrecherchen und Literatur basierten Konzepten und Theorien genauer untersucht werden.

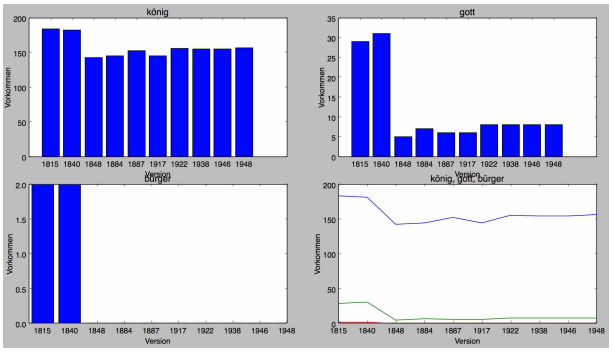
Zur Erstellung der Wortprofile, d.h. für die Anreicherung der Daten mit bspw. Lemmata, POS-Taggs und Dependenzrelationen wurde die Pipeline des DARIAH-DKPro-Wrapper [1] des NLP-Toolkits DKPro Core (vgl. Eckart de Castilho und Gurevych (2014)) benutzt. Das ermöglicht die Ausgabe der syntaktischen Relationen, die das gezählte Wort mit anderen Wörtern eingeht, zusammen mit ihren Häufigkeiten (Abbildung 5).

```
Welche Version? 1
Welches wort? mitglied
mitglied kommt in Version nr. 1 78 mal vor.
SUBJEKT:
'mitglied' kommt 2 mal als Subjekt von 'stimmen' vor.

'mitglied' kommt 2 mal als Subjekt von 'empfangen' vor.

'mitglied' kommt 1 mal als Subjekt von 'stehen' vor.

'mitglied' kommt 1 mal als Subjekt von 'reichen' vor.

'mitglied' kommt 1 mal als Subjekt von 'nehmen' vor.

'mitglied' kommt 1 mal als Subjekt von 'legen' vor.

'mitglied' kommt 1 mal als Subjekt von 'erhalten' vor.

'mitglied' kommt 1 mal als Subjekt von 'bekleiden' vor.

GENITIVATTRIBUT:
'mitglied' kommt 2 mal als Genitivattribut von 'drittel' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'wahl' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'vollmacht' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'teil' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'pension' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'mehrheit' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'hälfte' vor.

'mitglied' kommt 1 mal als Genitivattribut von 'gesamtzahl' vor.

AKKUSATIVOBJEKT:
'mitglied' kommt 1 mal als Akkusativobjekt von 'hinzufügen' vor.

'mitglied' kommt 1 mal als Akkusativobjekt von 'ernennen' vor.
```

Das verwendete Beispiel, die Personenkategorie Mitglied, kommt in der gewählten Verfassungsversion 78 mal vor. Die Informationen, wovon Mitglied das Subjekt ist oder inwiefern Mitglied als Genitivattribut oder Akkusativobjekt von bestimmten Termini vorkommt können u.a. dabei helfen, die Eigenschaften dieser Kategorie oder auch die Prozesse in die diese kategorie eingebunden sein kann, genauer zu bestimmen. So kann ein Mitglied hinzugefügt oder auch ernannt werden. In weiteren Betrachtungen kann dann herausgearbeitet werden, wozu ein Mitglied ernannt oder hinzugefügt werden kann. Das Genitivattribut der Kategorie König gibt bspw. Auskunft darüber, von welcher Bezugsgruppe diese Person überhaupt der König ist. Diese automatisiert verfügbaren Informationen tragen dazu bei, die kategoriale Wissensbestände der Verfassungsstaaten historisch-vergleichend zu untersuchen.

4. Graphische Darstellungen

Nutzende können sich in der aktuellen Version die Auftretensverteilungen von Wörtern in einem anzugebenden Zeitraum als Graph ausgeben lassen. Dabei werden pro eingegebenes Wort ein Balkendiagramm sowie ein Diagramm generiert, das die Kurven aller angegebenen Wörter in einem Graph zugleich darstellt (Abbildung 6).



In dem verwendeten Beispiel werden die Verkommenshäufkeiten von drei Personenkategorien – Gott, König und Bürger – in den Verfassungen der Niederlande für den Zeitraum 1815 bis 1948 dargestellt. Dabei fällt auf, dass der Bürger im Vergleich zum König als Repräsentation des Souveräns eine deutlich untergeordnete Rolle spielt und ab 1848 bis 1948 im Prinzip nicht vorkommt. Die Verfassungen spiegeln das politische System der Monarchie und nicht die Staatsbürgergesellschaft wieder. Ab 1848 nimmt auch das Vorkommen der Kategorie Gott signifikant ab. Während 1840 Gott noch 30 Mal vorkommt, verringert sich die Häufigkeit in den darauffolgenden 100 Jahren auf durchscnittlich sieben. Anahnd dieser Ergebnisse lassen sich ganz verschiedene Interpretationen und tiefergehende Analysen anschließen, um bspw. Erkenntniss über die religiöse Semantiken staatlicher Selbstbeschreibung (Gottesbezog, Gründungsmythen, Vorstellungen der Nation usw.) oder den Wandel des politischen Gemeinwesens und politischer Zugehörigkeit (wer gehört eigentlich dazu?) zu erlangen.

Diese Funktion wird demnächst um weitere bereichert werden, um die Potentiale von Visualisierungen als darstellende Klammer des Strukturvergleichs, der Wortsuche und der syntaktischen Wortrelationen wie auch als analytisches Werkzeug (Lupton 2014) selbst zu sondieren.

4. Zusammenfassung und Ausblick

Derzeit kann die Software alle Änderungen – unabhängig von formalen Totalrevisionen innerhalb der historischen Verfassungsentwicklungen aufzeigen. So lässt sich bspw. feststellen, welche Teile besonders häufig geändert werden oder welche Teile bis heute unangetastet geblieben sind.

Die Vorteile dieser Software gegenüber frei im Internet zugänglichen Versionierungstools (github, gitlab o.ä.) liegen auf der Hand: Zwar bieten solche Programme relativ einfach die Möglichkeit Textänderungen nachzuverfolgen, das gezielte Nachvollziehen von der Änderungshistorie spezifischer Textabschnitte ist ungleich schwieriger. Darüber hinaus bieten die beschrieben Funktionalitäten viel weitgehendere Auswertungsszenarien, als das reine Mapping, das durch ein Versionierungstool angeboten wird.

Beispielsweise kann mit dem Programm untersucht werden, welche neuen (normativen) Vorgaben Einzug in die Verfassung finden. Diese Änderungen in den Zeitreihen lassen sich dann ihrerseits im Zuge der weitergehenden Untersuchung historisch kontextualisieren und bspw. mit Blick auf staatlichen Wandel oder die Institutionalisierung bzw. Legitimierung neuer Werte, Normen, staatlicher Handlungsverpflichtungen und kultureller Leitideen (Meyer et al. 2005) interpretieren. Das Programm stellt das technische Werkzeug dafür dar, beide Ebenen, Mikro- und Makroebene, gleichermaßen zu betrachten, indem die Änderungshistorie einzelner Abschnitte ins Verhältnis zur Distanzsicht auf die gesamttextlichen Änderungen vieler Verfassungen gesetzt werden können.

Künftig sollen auch Fallvergleiche zwischen verschiedenen europäischen Staaten möglich sein, bspw. indem für frei wählbare Textbereiche die Cosinusähnlichkeit berechnet wird. Dadurch wird u.a. die Herausforderung des Vergleichs verschiedener historischer Kontexte tangiert. Wie kann ein informationstheoretisches Modell aussehen, das verschiedene Vektorräume, die definierte temporäre Sequenzen umfassen, zusammenbringt und miteinander vergleicht? Wie können Okkurrenzen kategorisiert werden, wenn diese bspw. häufiger auftreten?

Die dargestellte Form der Dokumentenarbeit macht Verfassungen nicht nur einer breiten Öffentlichkeit und vielfältigen wissenschaftlichen Erhebungen zugänglich. Vielmehr reflektiert sie Methoden der Dokumentenanalyse, indem sie der spezifischen Dokumentengattung „Verfassung“ besondere Aufmerksamkeit schenkt. Hermeneutisch-interpretative Verfahren versuchen Kontextwissen zumindest zu Beginn der Analyse weitestgehend auszublenden, wohingegen beim skizzierten Vorgehen eben dieses Wissen über die Dokumentenart, deren struktureller Aufbau sowie etwaige kulturell-historische Besonderheiten in die Auszeichung des Textes mit Metadaten für die computerbasierte Bearbeitung einfließt.

Insgesamt leistet die methodische Verschränkung von historischer Wissenssoziologie und Computerlinguistik als Dokumentenarbeit und Entwicklung einer Analysesoftware einen Beitrag zur Untersuchung der Ko-Fabrikation von Sprache und Verfassungsrecht in Europa, indem über einzelene Begriffe und Begriffskombinationen spezifische Wissensbestände und Semantiken in den Blick genommen werden können. Dieses Vorgehen kann dazu beitragen, neue Textkorpora zu erschließen und weitere gesellschaftliche Wissensbestände (bspw. Bibelversionen, Dramen usw.) zu erkunden. Diese Analysen ließen sich mit anderen methodisch ähnlichen, aber gegenständlich anders ausgerichteten Untersuchungen koppeln. Bspw. könnten Verfassungen in Beziehung zu Presseartikeln und den sich darin ablesbaren Diskursen gesetzt und miteinander verglichen werden.

Fußnoten

1. Im Zuge dieses Entwicklungsschrittes wurden u.a. folgende Tagger und Parser verwendet: Open NLP Segmenter, Mate Tools POS-Tagger, Mate Tools Lemmatizer, Open NLP Chunker, Mate Tools Morphological Analyzer, Hyphenation Annotator, CoreNLP Named Entity Recognizer, Mate Tools Dependency Parser.

Bibliographie

Boli-Bennett, John (1979): *The Ideology of Expanding State Authority in National Constitutions, 1870-1970, in: Meyer, John W. / Michael Thomas Hannan (eds.): National development and the world system: educational, economic and political change. Chicago: University of Chicago Press 222-237.*

Boli-Bennett, John / John W. Meyer (1978): The ideology of childhood and the state: Rules distinguishing children in national constitutions, 1870-1970, in: *American Sociological Review 43:* 797–812 .

Bubenhofer, Noah / Joachim Scharloth (2015): Themenheft „Maschinelle Textanalyse“, *in: Zeitschrift für germanistische Linguistik, 43.1.*

Eckart de Castilho, Richard / Gurevych, Iryna (2014): A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, 1-11, Dublin, Ireland.

Go, Julian (2003): A Globalizing Constitutionalism? Views from the Postcolony, 1945-2000, in: *International Sociology* 18.1: 71-95.

Gosewinkel, Dieter / Johannes Masing / Andreas Würschinger (2006): *Die Verfassungen in Europa 1789-1949.* München: Beck .

Hausser, Roland (2014): *Foundations of Computational Linguistics:* Human-Computer Communication in Natural Language . Berlin: Springer.

Heintz, Bettina / Annette Schnabel (2006): Verfassungen als Spiegel globaler Normen? Eine quantitative Analyse der Gleichberechtigungsartikel in nationalen Verfassungen, in: *Koelner Zeitschrift für Soziologie und Sozialpsychologie,* 58.4, 685-716.

Jepperson, Ronald L (1991): *Institutions, Institutional Effects, and Institutionalism, in: DiMaggio, Paul / Walter W. Powell (eds.): The New Institutionalism in Organizational Analysis. Chicago: University of Chicago Press 143-163.*

Kuckartz, Udo (2012): *Qualitative Inhaltsanalyse.* Methoden, Praxis, Computerunterstützung. Weinheim / Basel: Beltz.

Lobin, Henning (2010): *Computerlinguistik und Texttechnologie.* Paderborn / München: Fink.

Lupton, Deborah (2014): *Digital sociology.* New York: Routledge.

Manning, Christopher D. / Prabhakar Raghavan / Hinrich Schütze (2008): *Introduction to information retrieval.* New York: Cambridge University Press.

Mayntz, Renate (1997): *Soziale Dynamik und politische Steuerung:* theoretische und methodologische Überlegungen. Frankfurt/Main: Campus.

Mayring, Philipp (2015): *Qualitative Inhaltsanalyse.* Grundlagen, Techniken. Weinheim / Basel: Beltz.

Meyer, John W. (2005): *Die Weltgesellschaft und der Nationalstaat, in: ders., Weltkultur: wie die westlichen Prinzipien die Welt durchdringen. Frankfurt/Main: Suhrkamp 85-132.*

Mikolov, Thomas / Wen-tau Yih / Geoffrey Zweig (2013): Linguistic Regularities in Continuous Space Word Representations, in: *Proceedings of the HLT-NAACL conference* 746-752.

Prior, Lindsay (2011): *Using documents in social research.* Los Angeles: Sage.

Salton, Gerard / Andrew Wong / Shungshu Yang (1975): A vector-space model for information retrieval, in: *Journal of the American Society for Information Science* 18: 613-620.

Schimank, Uwe (1999): *Funktionale Differenzierung und Systemintegration der modernen Gesellschaft:* Soziale Integration. Opladen: Westdeutscher Verlag.

Strauss, Anselm L. / Juliet M. Corbin (1996): *Grounded theory:* Grundlagen qualitativer Sozialforschung. Weinheim: Beltz.

Thelen, Kathleen (1999): Historical Institutionalism in Comparative Politics, in: *Annual Review of Political Science* 2.1: 369-404.

Thelen, Kathleen (2002): *The explanatory power of historical institutionalism, in: Mayntz, Renate (eds.): Akteure-Mechanismen–Modelle. Zur Theoriefähigkeit makro-sozialer Analysen. Frankfurt / New York: Campus) 91-107.*

Vorländer, Hans (2007): Europas multiple Konstitutionalismen, in: *Zeitschrift für Staats- und Europawissenschaften* 5.2: 160-180.

Weber, Max (1972): *Wirtschaft und Gesellschaft:* Grundriss der verstehenden Soziologie. Tübingen: Mohr.

„XML Schema". World Wide Web Consortium (W3C) http://www.w3c.org/XML/Schema , [letzter Zugriff] 10.09.2017).

„XQuery 1.0: An XML Query Language". World Wide Web Consortium (W3C) http://www.w3.org/TR/xquery [letzter Zugriff 10.09.17].

„XSL Transformation Version 1.0". World Wide Web Consortium (W3C) http://www.w3c.org/TR/xslt [letzter Zugriff 10.09.17].

Eine nachhaltige Präsentationsschicht für digitale Editionen

Fechner, Martin

fechner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Einleitung

Die Anforderungen an die Erstellung von Editionen sind im letzten Jahrzehnt gestiegen. [1] Es ist unstreitig, dass digitale Editionen Annäherungen an den edierten Text ermöglichen, die weit über den statischen Druck hinausgehen (Sahle 2016). Gleichzeitig gibt es aber auch noch immer Anforderungen, etwa die Zitierbarkeit, die im Druck gelöst sind, zu denen es aber in der Webpublikation noch keine einheitliche Entsprechung gibt. Insbe-

sondere stellt sich die Frage nach der Nachhaltigkeit und der Langzeitarchivierung. [2]

Ein Lösungsansatz besteht darin, Techniken der Data Curation zu etablieren (Pempe 2012: 141f.). Damit ist aber ein Aufwand verbunden, der linear mit der Zahl der Webpublikationen wächst und sogar polynomial ansteigt, wenn auch die Schnittstellen zu verknüpften Ressourcen gepflegt werden müssen. Eine besondere Herausforderung stellt sich, wenn man sinnvollerweise annimmt, dass die Darstellung und die Funktionalitäten Teil der Edition selbst sind (Ralle 2016, Pierazzo 2015: 127-146, Porter 2016 und Turska et al. 2016). Wenn also nur die Forschungsdaten nachhaltig archiviert werden, geht die ursprüngliche Präsentation letztendlich verloren. Es wird zwar momentan auf die Einführung von Standards gesetzt, [3] einen wirklichen Mehrwert für den Gebrauch von digitalen Editionen entfalten diese aber erst, wenn sie durch das Angebot von technischen Schnittstellen unterstützt werden. [4] Bisher fehlt es jedoch an einem klar definierten Interface, welches die Forschungsdaten in eine nachhaltige funktionale Präsentationsform mit allen Aspekten einer digitalen Edition übersetzt. [5]

Entwurf einer Schnittstelle

Hier wird nun ein System für die Nachhaltigkeit von Webpublikationen entworfen. Es verbindet die Daten mit der Präsentationsschicht und kann mithilfe einer projektspezifischen, archivierbaren Konfiguration über eine Schnittstelle gesteuert werden. Durch den Einsatz einer solchen Schnittstelle können Webpublikationen inklusive der entsprechenden Funktionalitäten aus den Forschungsdaten reproduziert werden. [6] Digitale Editionen unterscheiden sich technisch zwar zurzeit noch stark voneinander (Robinson 2016), der hier gemachte Vorschlag und das sich anschließende Software-Beispiel zeigen daher, wo es schon jetzt Möglichkeiten zur Standardisierung gibt und wie diese aussehen könnten.

Vorbild für das System ist der erfolgreiche IIIF-Standard, der zu einem ähnlichen Zweck für Bilder eingeführt wurde (Cramer 2011). Der IIIF-Standard sieht eine Aufteilung von Server- und Clientstruktur vor, und als Austauschformat fungiert eine Manifest-Datei. Übertragen auf Digitale Editionen heißt dies, dass eine archivierbare Manifestdatei notwendige Definitionen festhält, mit denen es einem Viewer möglich ist, die vollständige Präsentation und Funktionalität der jeweiligen digitalen Edition herzustellen.[7]

Manifestdatei

Der Vorschlag für eine Manifestdatei, aus der die Funktionalitäten hergestellt werden können, lautet wie folgt (vgl. auch die nachstehende Tabelle):

Ganz konkret sollten die notwendigen Metadaten der digitalen Edition definiert werden. Dabei kann darüber diskutiert werden, welche Informationen für eine digitale Edition notwendig und welche für wünschenswert gehalten werden.

Mit der Definition einer Gliederung der Materialien, also von Editionstexten, Kommentaren und Begleitmaterial, können sinnvolle hierarchische Navigationselemente in der Darstellung umgesetzt werden.

In der Präsentationsoberfläche muss es möglich sein, zu den verschiedenen Datentypen zu navigieren und entsprechende Überblickslisten anzeigen zu lassen. Dafür werden die Typen von Datenobjekten (etwa Textsorten oder Register) definiert. Für die Unterstützung von facettierten Filtern, müssen diese entsprechend festgelegt werden.

Die Darstellung der Dokumente in der Einzelansicht kann unter Einbindung von Schnittstellen für die Transformation geschehen. [8] Auch können für die Navigation zu Abschnitten im Dokument, die entsprechenden Teile definiert werden. Zu jedem Objekttyp sollten auch die vorhandenen Beziehungen zu Teilen, als auch zu anderen Objekttypen festgehalten werden. Damit können verschiedene Forschungsdaten in einer dynamischen Ansicht zusammengeführt werden.

Für eine nachhaltige Einbindung externer Ressourcen sollte definiert werden, auf welche Ressourcen die Edition Bezug nimmt und welche Schnittstellen dazu genutzt werden. Die Integration externer Ressourcen kann problematisch sein, wenn deren Verfügbarkeit noch nicht gesichert ist. [9]

Zu definierende Elemente	Zu definierende Eigenschaften	Steht in Beziehung zu
Metadaten	Titelei, Impressum, Lizenzangaben	
Gliederung der Edition	Hierarchie, Verweise auf Objekte / Objekttypen	Objekttypen
Gesamtschau	Typ, Sortierungen	
Filterdefinitionen	Name, Typ, ID Kontext: Zugehöriger Objekttyp Wurzel des Filterobjekts (XPath) Pfad zum Filterwert (XPath oder XQuery) Labelingfunktion (XPath oder XQuery)	
Dokumente und Objekte	Name, Typ Schnittstelle: Zugang zu den Objekten Objekt-Wurzel Pfad zur Objekt-ID (XPath oder XQuery) Pfad zum Objekttitel Thumbnail: Funktionsdefinition für Kurzanzeige (HTML)	Gesamtschau, Filter, Einzelpräsentation, Binnenstruktur, Beziehungen, Schnittstellen, Zitation
Einzelpräsentation	Schnittstelle: Formate und Transformationskripte (ODD)	
Binnenstruktur	Name, ID Wurzel der referenzierbaren Objektteile (XPath) Pfad zur jeweiligen ID (XPath) Titel und Thumbnail für die Darstellung in einer Liste	
Beziehungen	Name, ID Subjekt, Prädikat, Objekt-Beziehungen	Objekttypen, Teile von Objekttypen
Schnittstellen	Label Pfad zu den Elementen (XPath) Schnittstellen (URLs)	

Tabelle 1: Definitionen für die Schnittstelle

Prototyp

Im Forschungprojekt „ediarum“ wird momentan ein Prototyp entwickelt, der das vorgestellte Konzept umsetzt und die Anforderungen der Darstellung erfüllen soll (Dumont/Fechner 2014 und http://www.bbaw.de/telota/software/ediarum). Dieser enthält eine Programmbibliothek, die die Funktionalitäten zur Anzeige bereitstellt. Kern der spezifischen Darstellung einer digitalen Edition wird durch eine Manifestdatei nach obigem Konzept gebildet. Mit diesem Prototypen ist es bereits möglich mithilfe der Manifestdatei und wenigen Anpassungen, die vor allem das Layout betreffen, eine Webseite für eine Digitale Edition zu erstellen. Mit dem Einsatz des Prototyps für mehrere Editionen werden die einzelnen Funktionalitäten und Konfigurationsmöglichkeiten ausgetestet und verbessert. Schließlich soll er als Viewer zur Verfügung stehen, der zur Präsentation lediglich die Manifestdatei und Zugang zu den Daten über eine entsprechende Serverinfrastruktur benötigt. Durch das Zusammenspiel der Kernkomponente, die alle Funktionalitäten bereitstellt, und der projektspezifischen Komponente, die im Layout angepasst werden kann, wird eine hohe Flexibilität erreicht. Somit kann für jedes Projekt ein individueller Auftritt erzeugt werden.

```
<config xmlns="http://www.bbaw.de/telota/software/ediarum/web/appconf">
    <project>
        <name>example</name>
    </project>
    <object xml:id="handschriften">
        <name>Handschriften</name>
        <collection>/Handschriften</collection>
        <item>
            <namespace id="tei">http://www.tei-c.org/ns/1.0</namespace>
            <root>tei:TEI</root>
            <id>@xml:id</id>
            <label type="xpath">.//tei:titleStmt/tei:title/string()</label>
        </item>
        <filters>
            <filter xml:id="country">
                <name>Land</name>
                <type>single</type>
                <xpath>tei:body//tei:msIdentifier/tei:country</xpath>
                <label type="xquery">function($string) {normalize-space($string)}</label-function>
            </filter>
            <filter> ... </filter>
        </filters>
        <inner-navigation>
            <navigation xml:id="ms">
                <name>Verlinkte Handschriften im Text</name>
                <xpath>//tei:ref[@type='otherms']</xpath>
                <id>@cRef</id>
                <order-by>label</order-by>
                <label type="xquery">function ($node as node()) { $node/text() }</label-function>
            </navigation>
            <navigation xml:id="section">
                <name>Abschnitte</name>
                <xpath>//tei:body/tei:msDesc/*</xpath>
                <id>name()</id>
                <order-by>position</order-by>
                <label type="xquery">function ($node as node()) { ... return $string }</label-function>
            </navigation>
        </inner-navigation>
        <backlinks>
            <backlink xml:id="ms-backlink">
                <name>Verweisende Handschriften</name>
                <object ref="handschriften"/>
                <condition type="xquery">function($node as node(), $this as node()) as xs:boolean {
                        let $sequence := for $ref in $node//tei:ref[@type='otherms'] return $ref/@cRef
                        return index-of($sequence, $this/@xml:id) > 0 }
                </condition>
                <parameters/>
            </backlink>
        </backlinks>
    </object>
    <object> ... </object>
</config>
```

Abbildung 1: Manifestdatei des Prototyps

Fazit

Hier wird die Entwicklung einer neuen Standardschnittstelle vorgeschlagen, um die Nachhaltigkeit digitaler Editionen zu verbessern. Denn es braucht für die Langzeitarchivierung digitaler Editionen auch eine Standardisierung ihrer Funktionalitäten. Editionen sind zwar sehr unterschiedlich, doch mit dem hier beschriebenen Interface und dem zugehörigen Austauschformat wird beispielhaft ein praktikabler Ansatz vorgestellt, um diese Lücke zu schließen. Eine Weiterentwicklung des hier gemachten Entwurfs und die Integration weiterer Standards unter Einbindung unterschiedlicher Editionsprojekte kann in Zukunft die Unabhängigkeit digitaler Editionen von einzelnen technischen Systemen erhöhen und unterstützt damit die Langzeitarchivierung.

Fußnoten

1. Ziele von Editionen finden sich bei (Förderkriterien 2015, Eggert 2016, Ralle 2016 und Sahle 2016).
2. Für die Forschungsdaten selbst ist die Langzeitarchivierung grundsätzlich gelöst. Digitalen Editionen wird jedoch nur geringe Zuverlässigkeit zugeschrieben (Pierazzo 2015: 169).
3. So wird das Format der Text Encoding Initiative (TEI) verbreitet eingesetzt, das nur ein erster Schritt zur Standardisierung ist (Holmes 2017). Weitere Standards und Identifikatoren sind etwa die GND für Personen, GeoNames-IDs für Orte oder der Canocical Text Service (CTS) für Zitationen, das DITA- oder DocBook-Format für (technische) Dokumentationen. Als Langzeitarchivierungsformat für Metadaten gibt es etwa LMER (Steinke 2005).
4. Gute Ansätze einer Präsentationsoberfläche für Einzeldokumente bietet der "teiPublisher", der auf dem Datenformat ODD aufbaut (Meier 2017, Meier/Turska 2016 und Turska et al. 2016).
5. Die Benutzbarkeit digitaler Editionen leidet unter mangelnden Interfaces (Robinson 2016). Pierazzo sieht es als Nachteil, dass viele digitale Editionen unterschiedliche User Interfaces besitzen, hält aber eine zukünftige Angleichung für wahrscheinlich (Pierazzo 2015: 162).
6. Es gibt dabei sehr unterschiedliche Anforderungen, die an die Präsentation digitaler Editionen gestellt werden (Shillingsburg 2016, Ralle 2016: 154f. und Sahle 2014).
7. Für die Langzeitarchivierung könnte eine Kompatibilität der Manifestdatei etwa mit LMER hergestellt werden (Steinke 2005).
8. TEI-Dokumente können etwa mit ODD zur Einzelpräsentation transformiert werden (Meier 2017).
9. Externe Daten können auch in eine "Standalone" Version überführt werden (Holmes 2017).

Bibliographie

Cramer, Tom (2011): The International Image Interoperability Framework (IIIF): Laying the Foundation for Common Services, Integrated Resources and a Marketplace of Tools for Scholars Worldwide. Blogpost. URL: https://www.cni.org/topics/information-access-retrieval/international-image-interoperability-framework

Dumont, Stefan / Fechner, Martin (2014): “Bridging the Gap: Greater Usability for TEI encoding”, in: Journal of the Text Encoding Initiative 8. URL: http://jtei.revues.org/1242

Eggert, Paul (2016): “The reader-oriented scholarly edition”, in: Digital Scholarship in the Humanities 31, 4: 797–810. DOI: 10.1093/llc/fqw043

Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft. In: Informationen für Geistes- und Sozialwissenschaftler/innen (11) 2015.

Holmes, Martin (2017): “Whatever happened to interchange?” In: Digital Scholarship in the Humanities 32, suppl_1: i63–i68. DOI: 10.1093/llc/fqw048

Meier, Wolfgang (2017): teiPublisher. The instant publishing toolbox. Version v2.2.0, Stand 8. September 2017. URL: http://teipublisher.com/index.html

Meier, Wolfgang / Turska, Magdalena (2016): “TEI Processing Model Toolbox: Power To The Editor”, in: Digital Humanities 2016: Conference Abstracts: 936. URL: http://dh2016.adho.org/abstracts/401

Pempe, Wolfgang (2012): „Geisteswissenschaften“, in: Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme. Hg. v. Heike Neuroth et al., Version 1.0, Stand 2012: 137-159. URN: urn:nbn:de:0008-2012031401

Pierazzo, Elena (2015): Digital Scholarly Editing: Theories, Models and Methods. Abingdon, England.

Porter, Dot (2016): "What is an edition anyway?" My Keynote for the Digital Scholarly Editions as Interfaces conference, University of Graz. Blogpost, in: Dot Porter Digital. URL: http://www.deporterdigital.org/?p=309

Ralle, Inga Hanna (2016): „Maschinenlesbar – menschenlesbar. Über die grundlegende Ausrichtung der Edition“, in: Editio 30, 1: 144-156. DOI: 10.1515/editio-2016-0009

Robinson, Peter M. W. (2016): “Project-based digital humanities and social, digital, and scholarly editions”, in: *Digital Scholarship in the Humanities* 31, 4: 875–889. DOI: 10.1093/llc/fqw020

Sahle, Patrik (2014): Kriterienkatalog für die Besprechung digitaler Editionen. Version 1.1, Stand Juni 2014. URL: https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/

Sahle, Patrik (2016): “What is a Scholarly Digital Edition?” In: Digital Scholarly Editing: Theories and Practices: 19-40. DOI: 10.11647/obp.0095.02

Shillingsburg, Peter (2016): “Reliable social scholarly editing”, in: *Digital Scholarship in the Humanities* 31, 4: 890–897. DOI: 10.1093/llc/fqw044

Steinke, Tobias (Redaktion) (2005): LMER Langzeitarchivierungsmetadaten für elektronische Ressourcen.Version 1.2, Stand 7. April 2005. URN: urn:nbn:de:1111-2005041102

Turska, Magdalena / Cummings, James / Rahtz, Sebastian (2016): Challenging the Myth of Presentation in Digital Editions, in: Journal of the Text Encoding Initiative 9. DOI: 10.4000/jtei.1453

Endstation Digital?! Herausforderung Metadaten und Nachhaltigkeit in musikwissenschaftlichen Datenbanken

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

Rettinghaus, Klaus
rettinghaus@bach-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Projekt Bach-Repertorium

I.

Ausgangspunkt war eine reine Metadaten-Sammlung zu Werken und Quellen Johann Sebastian Bachs (1999-2008: „Göttinger Bach-Katalog“), die zum Abschluss der ‚Neuen Bach-Ausgabe‘ am Johann Sebastian Bach-Institut Göttingen erfolgte. Dieses Metadatensammlung war 2008 die Basis für den Projektstart von „Bach digital“. Seitdem erfolgte mittels kontinuierlicher Förderung durch die DFG ein mehrstufig angelegter Ausbau:

Das erste Digitalisierungsprojekt umfasste die sogenannten Originalquellen zu Johann Sebastian Bachs Musik, also Autographen und originales Aufführungsmaterial Bachs, die sich zu etwa 90 % im Besitz der oben genannten Bibliotheken befinden. 2010 ging diese erste Stufe als www.bach-digital.de online.

Daran schloss sich von 2013 bis 2016 die Digitalisierung von sogenannten Sekundärquellen Bachscher Musik aus der Generation der Bach-Söhne und -Schüler an, ein Bestand, der besonders viel Tastenmusik J. S. Bachs umfasst, die vielfach nicht autograph überliefert ist und damit Forschungen zu individuellen Fassungen ermöglicht sowie Bachs Arbeitsweise in der Klavier- und Orgelmusik zwischen Kunstwerk und Unterrichtspraxis transparent zu machen hilft.

Mittlerweile wurde die dritte Stufe gezündet: die konsequente Ausweitung der Datenbank in Metadaten und Digitalisaten auf die Musik der Bach-Söhne im Projekt „Quellenkorpus Bach-Söhne – Erschließung und Digitalisierung der Primärüberlieferung zu Werken Wilhelm Friedemann, Carl Philipp Emanuel, Johann Christoph Friedrich und Johann Christian Bach sowie deren Einbindung in das zu erweiternde Portal Bach digital“.

Daneben werden Schritt für Schritt auch Werkverzeichnisse der Publikationsreihe „Bach-Repertorium“ sowie das derzeit neu erarbeitete „Bach-Werke-Verzeichnis III“ integriert. Mittelfristig werden dazu auch musikalische Incipits implementiert bzw. suchbar gemacht.

Diese stufenweise Bearbeitung eines Kernbestandes der Musik des 18. Jahrhunderts strukturiert und systematisiert das gesamte musikalische Quellenmaterial und bildet den soliden Ausgangspunkt für eine quellenbasierte Forschung zur Musik der Bach-Familie: nicht nur als wichtiges Hilfsmittel der Bach-Forschung, sondern auch z. B. als Vergleichsobjekt für Studien zu anderen Repertoires, als unterstützendes Material für Forschungen zu Mitteldeutschland, als Kernbestandteil zur Provenienzforschung wichtiger Sammlungen wie Poelchau, Breitkopf etc. etc. Die Nutzungsmöglichkeiten sind in den vergangenen zehn Jahren stark angewachsen; und damit auf die Verantwortung, ein möglichst den diversen Anforderungen gerecht werdendes Material bestmöglich aufzubereiten.

Das MyCoRe-basierte Projekt wurde dabei von Anfang an durch eine Dokumentation begleitet, die es problemlos nachnutzbar macht: https://ww-

w.bach-digital.de/content/documentation.xml?XSL.lastPage.SESSION=/content/documentation.xml.

Abseits der Bach-Forschung bzw. Musikwissenschaft werden Daten und Digitalisate von „Bach digital“ auch von einer breitgefächerten Bach-Community gesucht: das sind die weltweit großen Nutzerkreise musikinteressierter Laien sowie auch Musiker, die z. T. direkt nach originalen Quellen-Digitalisaten musizieren. Das Spektrum der Nutzer weitet sich nach unserer Erfahrung mit statistischen Daten zur Datenbank: je divergenter das ins Netz gestellte Material, desto vielfältiger die Nutzung. Inwiefern eine ursprünglich für die Bach-Forschungscommunity entwickelte Datenbank dieser Entwicklung noch stärker Rechnung tragen soll, wäre zu diskutieren.

II.

Die Menge der Datensätze an sich (es sind Stand Januar 2018 immerhin 8230 Musik-Quellen zu 3870 Werken der Bach-Familie) sowie der Umfang der Metadaten innerhalb eines Datensatzes ist nur mit hohem personellem Aufwand auf dem neuesten Stand der Forschung zu halten. Digitalisate und Metadaten werden so gut es geht laufend überprüft, auch mithilfe von Nutzer-Feedbacks – besonders jenen für die Bach-Forschung so wichtigen Power-Usern aus aller Welt. Dies ist eine ständige Anforderung, die die Daten selbst stellen, sobald sie öffentlich sichtbar sind. An der Aktualisierung der Datensätze aufgrund von Neuerkenntnissen der Bach-Forschung sollen deshalb nun auch mehr Mitarbeiter in der Forschungsabteilung des Bach-Archivs beteiligt werden als es Projektmitarbeiter für „Bach digital“ gibt. Ziel ist es, der Veraltung von Forschungsdaten entgegenzuwirken. Das Bach-Archiv sieht sich hier in der Verantwortung, die einmal publizierten Forschungsdaten mit den „Bach digital“-Nutzern möglichst zu teilen. Hierzu gehört auch die Mehrsprachigkeit, die derzeit nur mit Hilfe von strukturierten Daten umgesetzt werden kann. Fließtexte zu übersetzen ist mangels dafür vorhandener Projektmittel nur sehr begrenzt möglich. Alle anderen Daten sind aber mittlerweile auch in Englisch, Japanisch, Französisch (und Anfang 2018 auch Italienisch und Spanisch) recherchierbar. Hierbei sind wiederum die Nutzer der Datenbank selbst behilflich. Geplant ist als nächstes eine Nutzerbefragung, die über die Interessen und Wünsche sowie Kritik oder weitere Formen der Common Science-Beteiligung Auskunft geben soll. Inwieweit dieses Ergebnis zu einer Umstrukturierung von Daten oder der Präsentation von Modulen führen wird oder muss, ist derzeit noch offen.

Bei dieser prinzipiell optimistischen Sicht auf „Bach digital“ sollen weitere kritische Punkte nicht außer Acht gelassen werden, die aus dieser langjährigen Erfahrung mit den Metadaten resultieren:

Die Datenbank-Struktur suggeriert Eindeutigkeit, suggeriert, dass die Daten dem - in letzter Zeit in den Polit-Medien - so beliebten Faktencheck standhalten. Die Herkunft der Daten, gerade auch bei Neuerkenntnissen, wird dabei oft nicht präzise offengelegt. Die Datenbankstruktur suggeriert indes meist, dass es hier um Fakten geht. Unsicherheiten können nur sehr begrenzt formuliert werden, gerade im Fall von strukturierten Daten. Gerade auch die gegenüber den Printmedien so einfach zu handhabende Datenänderung ist also ein Problem für die Transparenz von Forschungsdaten.

Ein einfacher Daten-Austausch per Schnittstelle, sicher allgemein gewünscht und praktiziert, ist nur insofern dauerhaft praktikabel, so lange ermöglicht wird, dieses Procedere mehrfach zu wiederholen, gerade auch bei Richtigstellungen von Forschungsdaten. Ansonsten finden sich mehrere Versionen von Quellen- oder Werkdaten im Netz, die sicherlich unerwünscht sind, selbst wenn man mit Versionierungsangaben arbeitet.

III.

Immer mehr Datenbanken zu Musikern und musikalischen Quellen tummeln sich im Netz. Doch selbst wenn es inhaltliche Überschneidungen gibt, arbeiten sie zumeist aneinander vorbei. Dabei ist das größte Problem nicht einmal die Vergeudung von Ressourcen, sondern die prinzipielle Unmöglichkeit eines Datenaustauschs bzw. einer einfachen Nachnutzbarkeit der Metadaten – selbst wenn sie in den sogenannten „Quasi-Standards“ TEI oder MEI vorliegen. Bei öffentlich geförderten Projekten ist heutzutage Grundvoraussetzung, dass die „langfristige Sicherung von“ und der „grundsätzlich offene Zugang zu“ Forschungsdaten gewährleistet sein muss, es aber bislang unklar ist, was genau dies heißt. Ist der Zugang schon „offen“ wenn man die Informationen im Internet finden kann, oder erst dann, wenn sie über eine Schnittstelle bereitgestellt werden? Solange Forschungsprojekte nur „digitale Inseln“ errichten, bringt das „Digitale“ keinen wirklichen Mehrwehrt.

Zwar existieren bereits verschiedene Formate, die speziell für den Datenaustausch gedacht sind, wie z. B. MARC21 und METS/MODS, doch sind diese nur sehr eingeschränkt für (musikwissen-

schaftliche) Forschungsprojekte und Datenbanken einsetzbar. Auch spezielle Ontologien stehen bereit, die vom W3C zu den „Good Ontologies“ gezählt werden, also Ontologien, die vollständig dokumentiert, dereferenzierbar, von unabhängigen Datenlieferanten verwendet und möglicherweise von bestehenden Tools unterstützt werden („ontologies that are fully documented, dereferenceable, used by independent data providers and possibly supported by existing tools“). Beispiele dafür sind „ *Dublin Core*“ und „ *The Music Ontology* “ . Doch auch hier bleibt das Problem, dass diese Formate zu flexibel, zu schwammig gestaltet sind, um einen sinnvollen, nachvollziehbaren Datenaustausch zu gewährleisten, oder aber spezielle Forschungs-Erkenntnisse nicht hinreichend darin abgebildet werden können – ganz abgesehen davon, dass derlei Lösungen überhaupt erst einmal implementiert werden müssen. Das vielgepriesene RDF, das versucht, einige dieser Probleme zu lösen (oder zu umschiffen), kann dabei kein Selbstläufer sein.

Auch können Projekte a-priori nicht immer vorhersehen, welche Daten genau anfallen werden, bzw. welche von anderen Forschern oder Projekten nachgenutzt werden könnten. Es ist also nicht unbedingt zielführend, Daten in allen möglichen Formaten anbieten zu wollen, selbst wenn die Ressourcen es gestatten verschiedene Daten-Export-Möglichkeiten bereitzustellen (und zu pflegen).

Können RESTful APIs die Lösung aller Probleme sein? Diese ermöglichen es zwar, sehr spezielle Kombinationen aus Metadaten zusammen zu stellen. Dennoch bleibt das Problem der intern verwendeten Formate bestehen; beschreibt ein Feld „date“ ein Aufführungsdatum oder das Datum der Werkgenese?

Um digitale Gräber zu verhindern, sind spezielle, klar definiert und strukturierte Datenformate vonnöten, die für klar definierte Anwendungsfälle einen echten Austausch ermöglichen und somit auch erstmals dezentrale Suchmaschinen ermöglichen. Solche Suchmaschinen können abseits von Google überhaupt erst wirkliche Interdisziplinarität herstellen, denn mit wachsender Zahl an digitalen Projekten – so begrüßenswert dies auch sein mag – steigt die Gefahr, dass man „den Wald vor lauter Bäumen nicht sieht“, also Ergebnisse anderer (vielleicht fachfremder) Projekte nicht wahrnimmt, und dadurch möglicherweise den eigenen Erkenntnisprozess behindert.

Der aktuelle Umgang mit gesammelten Metadaten soll am Beispiel von „Bach digital“ gezeigt sowie mögliche Auswege skizziert und diskutiert werden. Vorgestellt werden dabei standardisierte Formate, die bereits heute den Informationsaustausch und -fluss ermöglichen und aufzeigen, was dadurch zukünftig möglich sein könnte, aber auch, wo die größten Lücken und dringendsten Desiderate bislang bestehen blieben.

'Exakt Historisch' im Digitalen? Versuch einer Anleihe

Schilz, Andrea
andrea.schilz@uni-passau.de
Universität Passau, Deutschland

Thematik und Ziel

„Der Schlaf der Vernunft gebiert Monster“, wusste Francisco de Goya. Ein *Schlaf* der Quellenkritik auch. Deshalb ist eine dem Digitalen angepasste, auf Daten erweiterte quellenkrititische Methodik üblich in den Digital Humanities und verwandten Fächern. Akzeptanz für eine kontextuell orientierte Quellenkritik im Digitalen ist auch im erweiterten Diskurs detektierbar, wenn auf einer abstrakteren Ebene für kulturkritische Perspektiven mit verstärkt ganzheitlichen Sichtweisen plädiert wird (Liu 2012, Presner 2015). In diesem Zusammenhang steht das Ziel des Vortrags, der sich in zwei Blöcke gliedert: Einer Analyse von Quellenspezifika im Digitalen folgt, vergleichend und übertragend, die Skizze eines digital-quellenkritischen Leitfadens, der Kriterien der *exakt historischen Methode* auf Born Digital spiegelt. Damit nimmt das transdisziplinäre Experiment methodisch Anleihe an der volkskundlichen *Münchner Schule*, die wiederum auf “Klassiker” der Quellenkritik zurückgreift, etwa Johann Gustav Droysen. Diese dezidiert *historische* Sichtweise wird eingenommen, da auch digitale Quellen historisch bedingt sind und ihre Deutung - im Sinne einer Ganzheitlichkeit - dem Rechnung tragen sollte. Es erscheint sinnvoll, eine Systematik anzuwenden, die hilft, Kontexte entsprechend zu identifizieren und transparent in hermeneutische Prozesse miteinzubeziehen.

Stand

Die historische Dimension des Digitalen (das Internet: Brügger 2017; Born Digital, Webseiten: z. B. Nanni 2017; Twitter: Sternfeld 2014) ist ebenso Gegenstand in den Digital Humanities wie die kulturwissenschaftliche (Klawitter et al 2012). Quellenkritikim Digitalenpräzisieren

Handbücher(Crompton et al 2016; Griffin, Hayler 2016), Angebote wie „compas. Strukturiertes Forschen im Web“ (infoclio.ch, Baumann/Hügi 2017) führen niederschwellig ein in„Quellenkritik bei Quellen aus dem Internet“. Zwei Beiträge seien hier herausgestellt: Eva Pfanzelter (2010)vergleicht explizit historische Quellenkritik („innere/äußere Kritik“, Pfanzelter 2010: 43) mit Quellenkritik im Digitalen und beleuchtet den daraus notwendig resultierenden „kritischen Umgang mit digitalen Ressourcen“. Peter Haber Peter Haber (2011)rekurriert in „Digital Past“ explizit auf Droysens Methodik.

Spezifika

Vorangestellt sei ein Diktum von Alan Liu, *„(…) the virtual is indeed fully material“* (Liu 2014: 276). Dem wird zugestimmt *und* aufgezeigt, dass das Ungreifbare Folgen hat für die Einordnung von Inhalt und Kontext: Digitale Quellen weisen spezifische Eigenschaften auf, die sich auf Autorschaft, Stoffliches und Zeitliches beziehen. Diese Kriterien werden anhand digitaler Quellenarten herausgearbeitet und in Bezug auf kritische Methodik betrachtet, um dann die methodische Übertragung zu zeichnen.

Daten und Autorschaft

Konventionell bezieht sich Autorschaft im Digitalen auf sekundäre oder primäre Quellen, bei denen Publizierende konservativ verzeichnet sind, sowie auf Schwarmprodukte mit fließenden Autorschaften, Schichtungen und Intentionen, deren Identifikation synoptische Auswertungsprozesse verlangt.
Da Autorschaft und Intentionalität als kritisches Moment eng hängen zusammenhängen, ist die Frage der Autorschaft im gegebenen Kontext zu erweitern auf Daten: Datenkritik. Das „Ethos der Statistik“, das sich auf Erfassungsparameter und Algorithmen genauso wie auf Fragestellungen und Operationalisierungen bezieht, ist im quellenkritischen Sinn zu erweitern auf hermeneutische Interpretationen. Am Beispiel von Malte Rehbeins (2017) Kritik des Projekts *Charting Culture* wird der Wert dieser kritischen Verschränkung deutlich.

Digitalisat und Stofflichkeit

Digitalisate bedürfen als vom Analogen ins Digitale transformierte Quellen besonderer Kritik, sowohl bezüglich des Objekts als auch der Metadaten. Im Analogen bildet die Dualität von Medium und Text Information aus, bei Daten als Träger von Information fallen Medium und Botschaft im McLuhanschen Sinne zusammen. Deshalb wohnt digitalen Repräsentationen immer ein Informationsverlust inne, dem Erfassung und Modellierung lediglich entgegenwirken. So teilt ein digitales Faksimile mehr über die *kritische* Physis der analogen Quelle mit (z. B. Alterung) als es die Homogenität eines OCR-prozessierten Textes vermag (immanente Schriftinformationen).

Born Digital und Zeitlichkeit

Born Digital hat keine Rückbindung an Greifbares und ist selbst potentiell ungreifbar. Ihrem Wesen nach sind diese Quellen fluid: Zum einen unterliegen sie ständigen Alternierungsprozessen, die der Rezipient bestenfalls passiv zur Kenntnis nehmen kann. Folgen für das Erfassen und Tradieren, die Domäne der Webarchivierung, sind Selektion, motiviert durch permanente „Vervielfältigung“ der sich im Turnus oder unregelmäßig verändernden Quellen, daraus resultierende Lücken sowie Probleme bei Datenspeicherung bzw. -vorhaltung. Zum anderen oszilliert Born Digital zwischen Ewigem Leben (vgl. Recht auf Vergessen) und spontanem Verschwinden. Diese Eigenschaften haben in Summe Konsequenzen für Korpusvalidität, Datierungen bzw. Ordnungen (Zeugenschaften), die Kritik von Inhalten sowie für die potentielle geschichtliche Dimension der Quellenart als solcher. Aufgrund der besonderen Bedeutung als „Quelle der Zukunft“ und ihrer komplexen Beschaffenheit stellt Born Digital eine besondere Herausforderung dar.

Methode

Das präzise Sondieren dynamischer kulturhistorischer Phänomene ist sowohl den Digital Humanities als auch der Volkskunde eigen, in der der hier diskutierte methodische Bezugspunkt Mitte der 1950er Jahre gesetzt wurde. Hans Moser und Karl-Sigismund Kramer initiierten die als *Münchner Schule* bezeichnete Perspektive. Sie trug zu einer Neuaufstellung nach der NS-Zeit bei, in der etliche Fach-Akteure die Blut-und-Boden Ideologie mitgestaltet hatten. Anstelle der Suche nach Absolutem im (germanischen) Vergangenem („Ursprungsforschung“) trat die *exakt historische Methode* als „exakte Geschichtsschreibung der Volkskultur“ mit definierten Quellen, Räumen und Zeiten.
Damals teils polarisierend, forderte Hermann Bausinger (*Tübinger Schule*) zeitnah eine Orien-

tierung am Aktuellen, der „technischen Welt“ – in Kombination führten u. a. diese beiden Ansätze zu einer Art vektorialen Denkens in der Volkskunde: Heutige Phänomene methodisch *historisch* zu lesen.

Prozess

Die Historische Quellenkritik staffelt sich zuerst in „äußere“und „innere Kritik“. Der„äußeren Kritik“(vgl. zum Begriff: Pfanzelter 2010: 43) zuzuordnen sind Aspekte der Multimodalität - das Zusammenspiel von Text, Bild, Audiovisuellem, Interaktion - was imFolgenden nicht dezidiert vertieftwird; derBlick geht vielmehr *exakt historisch* von Außen nach Innen. Karl-Sigismund Kramer formuliert 1968 modellhaft Kriterien der Quellenkritik, die Übertragung folgt dieser Systematik.

Der Quellenkritik vorgelagert ist eine Material-Kritik zur Unterscheidung „objektiven oder subjektiven Zeugniswerts“ bzw. von „Mischlagen“, was an der individuellen Quelle zu beurteilen ist. Übertragen auf Born Digital, erscheinen komplexere Formen wie Blogs und Foren, die ausgeprägt durch „Mischlagen“ charakterisiert sind, probat: „Objektiv“ bezieht sich auf inhaltlich definierte Themenkomplexe, „subjektiv“ auf eine erste Grobordnung nach Tendenzen.

Es folgen die drei Stufen der Quellenkritik (nach Droysen; vgl.: Haber):

- „1. Kritik der Echtheit“; dies verlangt den kritischen Abgleich von Traditionen bezüglich falscher Sachverhalte. Z. B.: Das Erzeugen einer Authentizitäts-Anmutung, die Optimierungsprozessen geschuldet ist und von einem spezialisierten Microtask-Markt mitgetragen wird.
- „2. Kritik des Früheren und Späteren“; das Prüfen zeitlicher Schichtung hat bei der Dynamik der gegebenen Quellen zufolge, dasslineare Vergleiche nur anhand systematisch eingehegter Archivierung möglich sind – diese Stufe der Kritik verweist auf die Notwendigkeit einer solchen zur Herstellung der Arbeitsbasis.
- „3. Kritik des Richtigen, d. h. die Frage nach dem Grad der Verzeichnung [eines objektiven Verhalts, d. Verf.], die (...) besonders durch subjektive und tendenziöse Verfärbung eingetreten sein kann“; aufbauend auf der bereits erfolgten Material-Kritik werden Themenkomplexe weiter aufgesplittet und granularer „objektiv“ kategorisiert. Dieses Extrapolieren von Konnotationen benötigt eine intermediale, dezidiert historische Lesart.

Auf dieser Basis kann die Interpretation in vier Stufen vorgenommen werden:

- „1. Pragmatische Interpretation, d. h. die Herstellung des sachlichen Zusammenhanges innerhalb des Forschungsgegenstandes (ob Einzelerscheinung oder Gesamtaspekt), wie er sich aus dem kritisch geordneten Material ergibt.“ Hier wird Verlinkung im Kontext der Korpusvalidität angesprochen – wie ist „Gesamtheit“ im Terrain von Born Digital bewertbar?
- „2. Interpretation der Bedingungen, (…) Umwelteinflüsse im engeren Umkreis der lokalen, wirtschaftlichen, rechtlichen, sozialen, technischen und allgemein geistigen Bedingungen, die auf das Werden der Erscheinung eingewirkt haben und ihre Funktionen bestimmen“. Das Beispiel Fake News verweist auf Fragen, die hier Relevanz haben.
- „3. Psychologische Interpretation, d. h. Versuch der schärferen Erkenntnis der seelischen Konstitution der Umwelt, in der die Erscheinung beheimatet ist, und des Willens und der Gefühle der aktiv oder passiv beteiligten Personen und Gruppen.“ Bezugsrahmen für Subjektives, das psychosozial eingeordnet und kontextuell decodiert wird, ist hier Twitter.
- „4. Interpretation nach den bewegenden sittlichen und politischen Mächten, (...) überindividuelle und [auf] den engeren Umkreis übergreifenden Impulse, die auf das Volksleben einwirken, es bewegen und gestalten“: Hier erfolgt die kulturkritische Einbettung in größere gesamtgesellschaftliche bzw. gobale Kontexte und theoretische wie empirische Metaperspektiven.

Fazit

Das transdisziplinäre Experiment versteht sich als „synkretistischer“ Versuch, eine tradierte Denkschule auf den Raum des Digitalen zu projizieren. Die Übertragung gibt Impulse für methodische Vertiefungen (konzeptionelle Verdichtung, Use Cases) und für Adaptionen in der Lehre (geisteswissenschaftliche Grundlagen und Analytik, Kritikfähigkeit).

Bibliographie

Baumann, Jan / Hügi, Jasmin (2017): „compas. Strukturiertes Forschen im Web. Ein Pro-

jekt von infoclio.ch.“; ebd.:„2. 5. 2. Quellenkritik bei Quellen aus dem Internet“. 25.04.2017 http://www.compas.infoclio.ch/de/kompas/2-5-2-quellenkritik-bei-quellen-aus-dem-internet/164 [letzter Zugriff 11. September2017].

Brückner, Wolfgang (1985): „Hans Mosers Bedeutung für die Volkskunde“, in: Moser, Hans (1985): Volksbräuche im geschichtlichen Wandel. Ergebnisse aus fünfzig Jahren volkskundlicher Quellenforschung. Berlin-München: Deutscher Kunstverlag X-XI

Crompton, Constance / Lane, Richard J. / Siemens, Ray(eds.) (2016) : „Doing Digital Humanities. Practice, Training, Research“. London: Routledge

Nanni, Federico (2017): „Reconstructing a website’s lost past. Methodological issues concerning the history of Unibo.it.“, in: *Digital Humanities Quarterly* 2017 Volume 11 Number 2

Gerndt, Helge (ed.) (1987): Volkskunde und Nationalsozialismus. Referate und Diskussionen einer Tagung der Deutschen Gesellschaft für Volkskunde, München 23. bis 25. Oktober 1986. München: Münchner Vereinigung für Volkskunde, 1989²

Griffin, Gabriele / Hayler, Matt (2016): Research methods for reading digital data in the digital humanities. Edinburgh: Edinburgh University Press

Haber, Peter (2011): Digital Past. Geschichtswissenschaft im digitalen Zeitalter. München: Oldenbourg Verlag 104-112

Kaschuba, Wolfgang (1999/2003): Einführung in die Europäische Ethnologie. München: Beck 2006³ 83-85

Klawitter, Jana / Lobin, Henning / Schmidt Torben(2012): „Kulturwissenschaftliche Forschung – Einflüsse von Digitalisierung und Internet“, in: Diess. (eds.): Kulturwissenschaften digital. Neue Forschungsfragen und Methoden. Frankfurt am Main: Campus Verlag 9-29

Köstlin, Konrad: „Historische Methode und regionale Kultur“ in: Ders.(ed.) (1987): Historische Methode und regionale Kultur. Karl-S. Kramer zum 70. Geburtstag. Regensburger Schriften zur Volkskunde, B. 4. Berlin-Vilseck: Tesdorpf Verlag 7-23

Kramer, Karl-Sigismund(1968): „Zur Erforschung der historischen Volkskultur“, in: *Rheinisches Jahrbuch für Volkskunde*, 1968, 19. Jahrgang. Bonn: Ferdinand Dümmler Verlag 7-41

Liu, Alan (2012): „Where Is Cultural Criticism in the Digital Humanities?“, in: Gold, Matthew K. (ed.): Debates in the Digital Humanities. Minneapolis-London: University of Minnesota Press 490-509. http://dhdebates.gc.cuny.edu/debates/text/20 [letzter Zugriff 11. September 2017].

Liu, Alan (2014): „The Big Bang of Online Reading“, in: Arthur, Paul Longley / Bode, Katherine (eds.): Advancing Digital Humanities. Research, Methods, Theories. Basingstoke: Palgrave Macmillan 275-290

Mittler, Elmar (2012): „Wissenschaftliche Forschung und Publikation im Netz. Neue Herausforderungen für Forscher, Bibliotheken und Verlage“, in: Füssel, Stephan (ed.): Medienkonvergenz – Transdisziplinär. Media Convergence, Band 1. Berlin-Boston: Walter de Gruyter Verlag 31-80

Pfanzelter, Eva (2010): „Von der Quellenkritik zum kritischen Umgang mit digitalen Ressourcen“, in: Gasteiner, Martin / Haber, Peter (eds.): Digitale Arbeitstechniken für die Geistes- und Kulturwissenschaften. Wien: UTB 39-49

Presner, Todd (2015): „Critical Theory and the Magle of Digital Humanities“, in: Svensson, Patrik (ed.):Between humanities and the digital. Cambridge, Mass: MIT Press 55-67

Rehbein, >Malte (2015): Forum: „Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht“, in: H-Soz-Kult, 27.11.2015 http://www.hsozkult.de/debate/id/diskussionen-2905 [letzter Zugriff 11. September 2017].

Rehbein, Malte (2017): „Geschichtsforschung im digitalen Raum. Über die Notwendigkeit der Digital Humanities als historische Grund- und Transferwissenschaftwissenschaft“, in: Herbers, Klaus / Trenkle, Viktoria (eds.): Papstgeschichte des hohen Mittelalters: Digitale und hilfswissenschaftliche Zugangsweisen zu einer Kulturgeschichte Europas (im Druck).

Schaller, Martin (2015): „Arbeiten mit digitalisierten Quellen. Herausforderungen und Chancen“, in: Schmale, Wolfgang (ed.): Digital humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität. Stuttgart: Steiner 15-30

Schich, Maximilian / Song, Chaoming / Ahn, Yong Yeol /Mirsky, Alexander / Martino, Mauro / Albert Barabási, Albert László / Helbing, Dirk (2014): „A network framework of cultural history“, in: *Science* 345 (6196). DOI: 10.1126/science.1240064. 558–562

Sternfeld, Joshua (2014): „Historical Understanding in the Quantum Age“, in: *Journal of Digital Humanites* Vol. 3, No. 2 Summer 2014

Weber, Matthew S. (2017): „The tumultuous history of news on the web“, in: Brügger, Niels / Schroeder, Ralph (eds.): The Web as History. Using Web Archives to Understand the Past and the Present. London: UCL Press 83-100

Fachspezifische Herausforderungen in der Realisierung des webbasierten digitalen Archivs THESPIS.DIGITAL

Löcker-Herschkowitz, Johannes A.
j.a.loecker-herschkowitz@univie.ac.at
Universität Wien, Österreich

Wagner, Christian
christian.wagner@univie.ac.at
Universität Wien, Österreich

Ausgehend von Dramen und Libretti des italienischen Dramatikers Giacinto Andrea Cicognini (1606–1649) beschäftigt sich das FWF-Forschungsprojekt „Making of a Repertoire for German Theatre (1650–1730): The Reception of Cicognini“ (Projektleiter Univ.-Prof. Dr. Stefan Hulfeld) mit Theaterrepertoires des deutschsprachigen Wandertheaters im 17. und beginnenden 18. Jahrhundert. Dabei werden vorhandene Daten überprüft und neue (Meta-)Daten zu Dramen und deren Aufführungen generiert. Wandertheaterforschung ist aufgrund einer problematischen Quellenlage und -überlieferung ein höchst komplexes Vorhaben. Informationen zu Aktivitäten der Wandertruppen können über administrative Unterlagen, Rechnungsbücher von Städten und Fürstenhöfen, Theaterzettel, Manuskripte und zeitgenössische Schriften gefunden werden. Seit mehr als einem Jahrhundert sammeln und veröffentlichen Wissenschaftler_innen Daten zu deutschsprachigen Wandertruppen. Diese finden sich in Studien zu einzelnen Wandertruppen, in Arbeiten zu Dramensammlungen/Codices, in Untersuchungen zu einzelnen Spielorten sowie in Darstellungen spezifischer Zeiträume. Problematisch ist hier die Anordnung und Auswertung der genannten Quellen: Unzählige Artikel und Bücher müssen herangezogen werden, um Fehler und Missinterpretationen von zuverlässigen Fakten zu unterscheiden. Für die Überprüfung der Daten werden Bibliotheken und Archive konsultiert, da die entsprechenden Unterlagen nicht über Websites verfügbar sind. Die Entwicklung eines webbasierten Archivs, in der die validierten Daten gesichert werden, ist aufgrund der oben beschriebenen Herausforderungen naheliegend. Nähere Informationen zum Forschungsprojekt finden Sie unter https://thespis.digital/Forschungsprojekt/ .

Beginnend mit einer Darstellung der Ausgangslage des Forschungsprojekts inklusive der im Forschungsfeld existierenden Datenbanken sowie bekannten Vorgangsweisen, welche als Vorbilder auch für unser digitales Archiv impulsgebend waren, werden wir in unserem Vortrag über die Gründe und Argumente für die Entscheidung des Einsatzes von Semantic MediaWiki (SMW) als webbasiertes Werkzeug zur kollaborativen Arbeit referieren. Dabei werden wir Anforderungen an ein webbasiertes Archiv, Möglichkeiten zur Kollaboration, Verwendung von offenen Standards und Formaten ebenso vorstellen wie die für das Projekt so entscheidende Vorgehensweise hinsichtlich der Entwicklung eines Datenmodells in enger Zusammenarbeit mit dem Forschungsteam.

Für die Erfassung der erhobenen und validierten Daten kommt ein Semantic MediaWiki (SMW) zum Einsatz – THESPIS.DIGITAL. Verzeichnet sind Daten und Metadaten zu Repertoirestücken, Aufführungen, Dokumenten, Personen und Datensätze, welche keinem Repertoirestück zuordenbar sind, allerdings wichtige Informationen für die Wandertruppenforschung darstellen. Das zu Grunde liegende Datenmodell wurde in enger Zusammenarbeit mit dem Forschungsteam erarbeitet; die Art der agilen Entwicklung wird in unserem Vortrag erörtert. Zentrale Bedingung für die Erfassung von Dokumenten bzw. Aufführungen ist der Bezug zur Quelle. Einzelne Quellen (Digitalisate von handschriftlichen Dramen, Theaterzettel etc.) sind im digitalen Archiv als Index- und Objektlinks verzeichnet. Digitalisate, die nicht über die Archive direkt im WWW zugänglich sind, werden von uns im Repositorium Phaidra (Digital Asset Management System mit Langzeitarchivierungsfunktion der Universität Wien, https://phaidra.univie.ac.at) abgelegt. Personen und Orte werden als Links zur deutschsprachigen Wikipedia verzeichnet. Das starke Interesse der Forschungsgruppe, Forschungsergebnisse einer interessierten Öffentlichkeit zugänglich zu machen, führt dazu, dass Personenartikel der Schauspieler_innen und Theaterleiter_innen direkt in der Wikipedia angelegt bzw. überarbeitet werden. Durch die so entstehende Verknüpfung mittels Verzeichnissen von Normdaten, über Wikipedia und THESPIS.DIGITAL zu Digitalisaten in Archiven wird eine Kette von Linked Open Data (LOD) hergestellt. Mit dem Blog https://thespis.hypotheses.org treten wir in Kommunikation nach außen und ergänzen so den Austausch mit Wissenschaftler_innen und der Öffentlichkeit. Ein weiterer wichtiger Aspekt beim Wissensaustausch mit Fachkolleg_innen im Feld

der Wandertruppenforschung stellt der Einsatz des benutzerfreundlichen Werkzeugs SMW dar. Diese Form der Kollaboration wird verstärkt nach Projektende zu tragen kommen: der Datenbestand wird durch Wissenschaftler_innen aus dem Bereich der Wandertheaterforschung kontinuierlich ausgebaut. THESPIS.DIGITAL wird als zentrales Tool zur Erfassung, Auswertung und Visualisierung von (Aufführungs-)Daten im Zusammenhang mit Wandertheater im 17. und 18. Jahrhundert etabliert.

Bei SMW handelt es sich um eine Open-Source-Erweiterung für MediaWiki. Es stellt ein leistungsfähiges und flexibles Wissensmanagement-System zur Verfügung, in welchem das Speichern und Abfragen von Daten innerhalb von Wiki-Seiten möglich ist. Der entscheidende Vorteil von SMW liegt darin, dass Daten auf mehreren Ebenen mit semantischen Informationen angereichert und erstellte Daten über Semantic Web Standards veröffentlicht werden. Durch die Maschinenlesbarkeit der Daten besteht eine enorme Anschlussfähigkeit für weiterführende Anwendungen und Auswertungen. Die Verwendung offener Standards (Darstellung, Export, Softwarebasis, Dokumentation, Lizenz) fördern eine zukunftssichere Verwendung über das Projektende hinaus. SMW verfügt über eine Benutzer- bzw. Rechteverwaltung zur Unterstützung der Kollaboration durch die Vergabe differenzierter Benutzerrechte. Durch eine umfassende Versionskontrolle kann jeder Bearbeitungsschritt transparent nachvollzogen werden – sowohl bei der Entwicklung als auch bei den Änderungen der inhaltlichen Eingaben.

Wir werden aber nicht nur auf Stärken und technische Funktionsweisen des Werkzeugs eingehen. Vor allem wollen wir die Vorgangsweise im Projekt als eine Möglichkeit präsentieren, wie mit dem kritischen Anspruch der Theaterwissenschaft die technologische Entwicklung des Werkzeugs dominiert wird. Dies betrifft insbesondere die Einbeziehung des Forschungsteams in jede Phase des Entstehungsprozesses. Die Inputs hinsichtlich Anpassungen/Änderungen/Erweiterungen des Datenmodells wurden von einer wissenschaftlichen Perspektive geleitet und unabhängig von technologischen Bedingtheiten zur Diskussion gestellt. Dabei spielt die technologische Basis zwar eine grundlegende Rolle, viel stärker treten allerdings fachwissenschaftliche Diskussionen und Übersetzungsstrategien bei Entscheidungen in den Vordergrund. Mit THESPIS.DIGITAL wurde ein Instrument geschaffen, dass sich in jeder Phase der Entwicklung an den Vorstellungen des Forschungsteams orientiert hat und theoretische Implikationen umsetzt, anstatt bestimmte Möglichkeiten oder Praktiken vorzugeben. Entstanden sind Funktionsweisen und Abläufe die dem Forschungsgegenstand entsprechen. Die Umsetzung erfolgt anhand der Möglichkeiten von SMW, offener Standards und Technologien.

Im Zuge der agilen Entwicklung des Datenmodells und der verwendeten Werkzeuge konnten gemeinsam mit dem Projektteam auch die zu verwendenden Standards erarbeitet und Entscheidung getroffen werden, welche eine sehr offene Verwendung von Daten (sowohl die Projektdaten als auch Daten bspw. von Wikipedia bzw. Wikidata) ermöglicht. Unsere Daten stehen in der Creative Commons Lizenz CC BY-NC-SA 4.0 zur Verfügung und können mittels offener Formate wie CSV, JSON, SPARQL, RDF und API ausgetauscht werden. Im Projektteam wurde es kritisch gesehen, dass die meisten „Datenbanken“ keinen offenen Zugang zu den Daten und somit keine automatische (maschinengestützte) Weiterverwendung der Daten ermöglicht. Aus diesem Grund war die Entscheidung, auf offene Standards und Daten zu setzen, nur folgerichtig. In einer Demonstrationsphase geben wir Einblick in die Funktionsweise von THESPIS.DIGITAL und zeigen, wie die Vorteile des Werkzeugs zum Tragen kommen.

Funktionale und deklarative Programmierung-basierte Methode für nachhaltige, reproduzierbare und verifizierbare Datenkuration.

Barabucci, Gioele
gioele.barabucci@uni-koeln.de
Universität zu Köln, Deutschland

Einleitung

Durch die wachsende Menge an digitalen Daten im Bereich Digital Humanities bedarf es zunehmend der Arbeit von Datenkuratoren. Aufgrund der ständigen Steigerung der Zahl und des Umfangs der zu verarbeitenden Quellen ist jedoch

der übliche Modus Operandi der Datenkuratoren (manuelle Konversionen und konsekutive Anpassungen) untragbar geworden.

Dieser Vortrag stellt eine neue Methode für die Kuration digitaler Daten vor, die auf den Prinzipien der funktionalen Programmierung, der unix-Tools und der XML-Technologien basiert. Diese Methode wurde vom Cologne Center for eHumanities der Universität zu Köln seit 2014 im Rahmen des Lazarus-Projekts (CCeH 2014) und danach in verschiedenen anderen DH-Projekten angewendet.

Die Besonderheit dieser Methode besteht in der Aufgliederung des Kurationsprozesses in eine Pipeline von Miniprogrammen, von denen jedes einzelne einen präzisen Schritt des Kurationsprozesses darstellt. Wird ein Fehler in den resultierenden Dateien bemerkt, kann ein neuer Schritt geschrieben werden und in die Kette eingebaut werden, oder die existierenden Schritte korrigiert werden. Anschließend wird die Kurationspipeline neu laufen gelassen.

Der Hauptgrundsatz lautet: Keine Datei wird „manuell“ modifiziert, jede Operation muss von einem Miniprogramm ausgeführt werden.

Konsequenz dieses Grundsatzes und Hauptvorteil der vorgestellten Methode ist, dass die ganze Arbeit des Kurators — mitsamt seiner Entscheidungen und seiner bevorzugten Arbeitsweisen — in dieser Pipeline von Miniprogrammen dokumentiert ist und in sie eingebettet ist. Des weiteren sind der Kurationsprozess und seine Ergebnisse einfach zu reproduzieren und zu verifizieren. Das führt zu einer besseren Nachvollziehbarkeit der Kurationsarbeit, auch wenn die Kuration von einem Team durchgeführt wird.

Kuration von Digitalen Daten

Eine der Aufgaben der Datenkuration und der Datenkuratoren ist: »[to] intervene in the research process in order to translate or migrate data into new formats, to enhance it through additional layers of context or markup, to create connections between data sets, and to otherwise ensure that data is maintained in as highly-functional a form as possible.« (Flanders & Muñoz 2017).

Praktisch können wir die Arbeit der Datenkuratoren auf folgende Art und Weise grob zusammenfassen:

- Die Daten werden von den Forschern zu den Kuratoren übertragen.
- Die Kuratoren studieren die Daten, sowohl ihren Inhalt als auch ihr Format.
- Die Kuratoren reichern die Daten mit den nötigen Metadaten an und sie sorgen dafür, dass eventuelle Inkohärenzen zwischen den originalen Formaten und den Zielformaten ausgeglichen werden.
- Die so verarbeiteten Daten werden archiviert, publiziert oder in anderen Projekten verwendet.

Datenkuration am CCeH: Das Cologne-Sanskrit-Lexicon-Projekt

Ein praktisches Beispiel von Datenkuration sind die Wörterbücher des Cologne Sanskrit Lexikons. Diese wurden von verschiedenen Wissenschaftlern der Universität zu Köln seit den 90er Jahren (pre XML und pre Unicode) erarbeitet, innerhalb des Lazarus-Projektes vom CCeH kuratiert (d.h. in TEI/Unicode umgewandelt) und 2015 zugänglich gemacht.

Die anfängliche Arbeitshypothese, welche im Verlauf jedoch fallen gelassen wurde, sah einen eher klassischen Workflow vor: Die originalen Dateien in XML umwandeln, danach eine XSLT-Transformation nutzen, um diese in TEI umzuwandeln und schließlich die durch die Transformation entstandenen kleinen Fehler per Hand verbessern.

Wir haben es vorgezogen, diesen Weg aus zwei Gründen nicht einzuschlagen.

Erstens: Von Beginn an sind verschiedene Versionen der zu kuratierenden Dateien aufgetaucht. Hätten wir in der Zwischenzeit neue Versionen der Dateien entdeckt, wäre die bis dahin geleistete Arbeit vergebens gewesen.

Zweitens: Die Arbeitsgruppe bestand aus drei Personen aus unterschiedlichen Fächern und mit unterschiedlichen Herangehensweisen an das Thema Kuration. Einen einheitlichen Stil bei zu behalten wäre nicht einfach — wahrscheinlich unmöglich — gewesen.

Die Arbeitsgruppe hat sich dann für eine andere Arbeitsweise entschieden: Eine „wiederholbare Pipeline“-Methode, die auf der funktionalen Programmierung basiert, statt manueller Konversionen und Anpassungen.

In dieser Methode ist jeder Arbeitsschritt formell durch ein sehr kleines Programm beschrieben, das in einer funktionalen und deklarativen Programmiersprache implementiert ist und im Durchschnitt aus weniger als 15 Instruktionen besteht. In dieser Art sind sehr unterschiedliche Schritte implementiert, z.B. die Normalisierung der Lemmata, die Behebung von technischen Fehlern, die Integration von externen Quellen.

Diese Programme, die die Schritte der Kurationsarbeit darstellen, sind im Sinne einer Pipeline organisiert, d.h. der Output des einen ist der Input eines anderen. Die Kuratoren erklären, welches die kommenden Schritte sind und welche Abhängigkeiten zwischen den einzelnen Schritten bestehen (z.B. dass die Umwandlungsschritte dem Abrufschritt folgen sollen). Der folgende Abschnitt enthält verschiedene konkrete Beispiele von Kurationspipelines und Schritten/Miniprogrammen.

Alle diese Schritte sind reine *idempotente* Funktionen, d.h. dass ihr Ergebnis nur von den Input-Daten anhängig ist. Konkret bedeutet das, dass man die Kurationspipelinemehrmals durchlaufen kann, und immer das gleiche Ergebnis erhält. Dies steht im Gegensatz zu den klassischen Skript-basierten Methoden.

Lazarus-Kurationsworkflow: XML-Pipelines, Makefiles und Schematron

In der Essenz bedeutet das konkret, dass der Kurationworkflow, der im Lazaruz-Projekt und in anderen folgenden CCeH-Projekten benutzt wurde, aus drei großen Komponenten besteht:

1. Die Makefiles. Ein Makefile ist eine Datei, die den gesamten Kurationsprozess mittels des unix-Tools make steuert. Der Makefile erklärt, wie man eine Datei X (genannt **Target**) durch die Dateien A, B und C (genannt **Abhängigkeiten von X**) herstellen kann. Im Fall des Cologne-Sanskrit-Lexicon-Projekts erklären die Makefiles wie man die target TEI-Deteien herstellen und testen kann, d.h. wo man die originalen Dateien mit den Sanskrit-Wörterbüchern finden kann, wie man sie herunterladen kann, wie man den Konversionprozess durch die Konversionspipelines durchführen kann usw.
2. Die Konversionspipelines. Jede Pipeline ist für die Konversion bestimmter Dateien verantwortlich und besteht aus verschieden Schritten. Jeder Schritt ist implementiert durch eine XSLT-Transformation (die funktionalen Miniprogramme, wie oben erwähnt). Die Pipelines selbst sind XProc-basierte XML-Pipelines (Walsh 2007).
3. DieTests. Verschiedeneautomatisierte Schematron-basierte Tests kontrollieren, dass die hergestelltenDateien valide sind, dass keine altenschon behobenen Fehler erneut eingepflegt werden, sowie dass keine Informationen verloren gehen.

Um die Methode und die Beziehungen zwischen den verschieden Komponenten besser zu verstehen, stellen wir hier ein konkretes Beispiel vor: Das Monier Sanskrit-English Wörterbuch, Teil des Cologne-Sanskrit-Lexicon-Projekts.

Die ursprünglichenDateien mit den Digitalisatenund den Transkriptionen des Monier-Wörterbuches, Nachlass der Arbeit von Thomas Malten et al., sind auf einem Server der Universität zu Köln gespeichert und archiviert. Um dieses Wörterbuch in dasneue Cologne Sanskrit-Lexicon zu integrieren, müssen die Kuratoren folgenden Operationen durchführen:

1. Die Originaldateien vom Server abrufen;
2. Kleine Markup-Fehler beheben;
3. Die Daten in TEI/XML umwandeln;
4. Verweise zu externen Datenbanken/Quellen integrieren;
5. Prüfen, ob Fehler unterlaufen sind, bzw. dass kein Lemma verloren wurde;
6. Die kuratiertenDaten zur Verfügung stellen, sodass sie auf den Produktionsserver hochgeladen werden können.

Diese Operationen, die denKurationsprozess grob zusammenfassen, sind imMakefile monier.mkbeschrieben. Die Operationensind im Makefile in der folgenden Form ausgedrückt: „1) Datei X wird ausgehend von Datei Y hergestellt. 2) Wenn X fehlt oder älter als Y ist, wird X hergestellt, indem man Y anProgramm K mit bestimmten Parametern übergibt“. Abbildung 1 zeigt einige Regeln, die im Vergleich zum Original vereinfacht dargestellt sind.

```
STEPS = $(wildcard $(XPROC_ROOT)/*.xsl)
PIPELINE = (XPROC_ROOT)/monier/conversion.xpl
EXTRAS += monier/abbreviations.tei
EXTRAS += monier/authorities.tei
EXTRAS += monier/authorities-links.tei
EXTRAS += monier/greek.tei

monier.tei: $(STEPS)
monier.tei: $(PIPELINE) | $(XPROC_EXECUTABLE)
monier.tei: $(EXTRAS)
monier.tei: monier.xml
        $(XPROC) -isource=$< result-url=$(abspath $@) $(PIPELINE)

monier.split.tei: $(STEPS)
monier.split.tei: $(PIPELINE) | $(XPROC_EXECUTABLE)
monier.split.tei: $(EXTRAS)
monier.split.tei: CHUNK = 1 # Overridable at runtime
monier.split.tei: monier.split-$(CHUNK).xml
        $(XPROC) -isource=$< result-url=$(abspath $@) $(PIPELINE)
```

Abbildung 1: Makefile für die Kuration des Monier Wörterbuches. In den ersten Zeilen werden verschiedene Parameter eingerichtet. Dann werden die Abhängigkeiten der target-Datei monier.tei beschrieben. Schließlich wird das Kommando eingerichtet, das man benötigt, um die target-Datei herzustellen.

Die Makefiles geben in *imperativer*Weise vor, welche Daten die *funktionale*Pipeline durchlaufen sollen und wo das finale Ergebnis gespeichert werden soll. Dies macht unter anderem das Testen der Kurationspipeline anhand eines Auszugesdes Wörterbuches möglich, ohne die Pipeline an sich zu verändern; es werden lediglich Änderungen einiger Parameter in den Makefiles vorgenommen.

Operationen 2, 3 und 4 (die Behebung von Fehlern, die Umwandlung der originalen Daten in TEI/XML und deren Verlinkung mit externen Datenbanken) sind das Herzstück des Kurationsprozesses und wird durch verschiedeneXProc-Pipelines durchgeführt. Diese Pipelinesbestehen insgesamt aus 35 verschiedenen XSLT-Transformationen, von denen jede einzelne einen spezifischen Schritt des Kurationsprozesses darstellt. Beispiele für diese Schritte sind:fehlplatziertes Markup verschieben, falsch kodierte Devanagari-Buchstaben richtig stellen, bibliographische Referenzen hinzufügen. Abbildungen2 und 3 zeigen Ausschnitte von zwei XSLT-Transformationen.

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                version="2.0">
  <xsl:import href="../common/identity.xsl"/>
  <xsl:import href="../common/chars.xsl"/>

  <!--
    This transformation reworks the placement of `;` and
    whitespace in cases where there is more than one `<ls>`.

    The result is more similar to the facsimile.
  -->

  <!--
    Turns <ls>Ka1t2h. ; </ls><ls>Anukr.</ls>
    into  <ls>Ka1t2h.</ls>; <ls>Anukr.</ls>
  -->

  <xsl:template match="ls[ends-with(.,' ; ')]">
  <xsl:variable name="first-part"
                select="concat('', substring-before(., ' ; '))"/>

  <ls>
    <xsl:value-of select="$first-part"/>
  </ls>
  <xsl:text>;</xsl:text>
  <xsl:value-of select="$char-space"/>
</xsl:template>
```

Abbildung 2: Ein Miniprogramm. Dieser Schritt korrigiert nur einen bestimmten Fehler.

```
<!-- ls elments preceded by abE are ignored:
     they have already been used inside abE -->
<xsl:template match="ls[preceding-sibling::node()[1][self::abE]]"/>

<!-- Literary quotes and sources -->
<xsl:template match="ls">
  <cit type="literary_source">
    <bibl xml:lang="{$csdl-language}">
      <ref>
        <xsl:apply-templates/>
      </ref>
    </bibl>
  </cit>
</xsl:template>
```

Abbildung 3: Auszug aus der Haupt-TEI-Transformation. Da die vorherigen Schritte die kleinen Fehler schon behoben haben, kann diese Transformation kurz und bündig sein.

Diese Miniprogramme lesen oderschreiben keine Datei. Sie fungieren als rein funktionale Filter, die die Daten empfangen, einige Anteile modifizieren, und die modifizierten Daten zurückgeben. Wie bereits erwähnt ist die Aufgabe des Makefiles zu entscheiden, welche Daten in die Pipeline eingespeist werden sollen und wo die Ergebnisse gespeichert werden sollen. Die Pipeline kümmert sich nicht um diese Details. Dies verringert den Aufwand für die Entwicklung der Miniprogramme erheblich.

Die Miniprogramme, aus denen die Pipeline besteht, spiegeln die Entscheidungen des Teams wider, das diese Daten kuratiert hat (Peter Dängeli, Martina Gödel und Gioele Barabucci, mit der wissenschaftlichen Kollaborationvon Felix Rau). Weil die Entwicklung des Codes der Pipeline und der Miniprogramme mittels eines Repositorium auf GitHub stattgefunden hat,ist es möglich, den Entwicklungsprozessder Kurationsarbeit nachzuvollziehen. Insbesondere ist es möglich, zu sehen, wie einzelneSchritte,die sich alsirrtümlich herausgestellt haben, durch bessere Schritte ersetzt werden, ohne das die übrigen Teile der Pipeline verändert werden müssen.

Schließlich Operation 5 (testen, dass kein Fehler untergelaufen ist)ist durch Schematron-basierte Testsimplementiert. Während dieser Operation wird getestet, 1) dass der Umwandlungsprozess keine Daten unabsichtlich entfernthat und, 2) dass alte Fehler, die schon korrigiert worden sind, wiedereingeführt wurden. Die Wiedereinführung von alten Fehlernist ein Ereignis, das in über Jahre andauerndenProjekten leider häufig vorkommt. Diese Art von Test ist von den Best Practices der Continous Integration in der Softwareentwicklung inspiriert.

Vorzüge der vorgestelltenMethode

Der Gebrauch dieser Methode hat viele Vorteile, sowohl im Hinblick auf die methodologische Stringenz als auch die Technik an sich:

- Jede einzelne Handlung der Kuratoren ist formalisiert und dokumentiert (durch XSLT-Code und Code-Kommentare). Die geringe Größe der Miniprogramme macht deren Code praktisch selbst dokumentierend.
- Jeder Wissenschaftler kann unabhängig nachvollziehen, wie die Ergebnisse entstanden sind.
- Jeder Schritt kann einzeln getestet werden.

- Man kann zurückverfolgen, welcher Schritt ein bestimmtes Konstrukt in den Ergebnissen generiert hat.
- Die Wiederverwendung von Schritten ist möglich und leicht nachzuvollziehen.
- Eine methodologische Kohärenz kann über Jahre hinweg beibehalten werden, auch wenn neue Kuratoren diese Daten verwalten werden.
- Dank der Speicherung verschiedener Schritte in Versioning-Systemen wie Git, kann man sehen, wie der Kurationsprozess entwickelt worden ist.

Die Hauptregel dieser Methode, dass keine Datei „manuell“ modifiziert wird und alles durch Miniprogramme ausgeführt wird, garantiert, dass jede Operation an den gegebenen Daten klar definiert und ausdrücklich formuliert ist.

Die Rolle der Kuratoren

Diese Methode ändert nicht die Rolle oder die Verantwortung der Kuratoren, aber sie verändert grundlegend ihre tägliche Arbeit. Die Arbeit der Kuratoren besteht nicht mehr in dem Modifizieren von Dateien in einem Editor, sondern in dem Schreiben von Arbeitsschritten und in der korrekten Verwaltung von den Abhängigkeiten zwischen Arbeitsschritten.

Die Kurationsarbeit ist dann in zwei Teile aufgeteilt. Der erste Teil besteht im Schreiben und in der schrittweisen Präzisierung von Kurationsprogrammen, welches die Hauptaufgabe der Kuratoren darstellt. Hierin zeigt sich die Fähigkeit, die Erfahrung der Kuratoren sowie die von ihnen präferierten anwendbaren Richtlinien. Der zweite Teil ist die Generierung von kuratierten Daten, welche in sterilerArt und Wiese von einemKoordinationsprogramm vollzogen wird. Es führt die verschiedenen Schritte in der von Kuratoren bestimmten Reihenfolge innerhalbweniger-Minuten durch.

Eine letzte wichtige Auswirkung dieser Methode ist, dass was registriert/gespeichert wird, nicht nur die Endergebnisse sind, sondern der ganze Kurationsprozess: Vom Abrufen der originalen Daten bis zu der Speicherung der kuratierten Daten. Es besteht die Möglichkeit diesen Prozess der Öffentlichkeit zugänglich zu machen, nicht nur um eine bessere Transparenz zu schaffen, sondern auch um die Zusammenarbeit mit externen Kuratoren zu erleichtern.

Zusammenfassend ergibt sich das Besondere dieserMethode aus der Tatsache, dass nicht nur das Ergebnis der Kuration, sondern auch der Kurationsprozess dokumentiert wird.

Verwandte Arbeiten

Für die Kuration digitaler Daten wurde oft ein einfachererWorkflow vorgeschlagen: die Daten ändern und danach die Ergebnisse in einem Git-Repositorium speichern. (Reeve 2016, Crowley et at. 2017). Die Idee der Befürworter dieses Workflows ist, dass die Speichrung des Datenstatuses nach jedem Arbeitsschritt den Kurationsprozess adäquat widerspiegeln. Das greift jedoch zu kurz. Git speichert nur was geändert wurde, nicht mit welcher Absicht eine Änderung durchgeführt wurde. Natürlich könnten diese Absichten und die entsprechenden Begründungen in einer Commit-Nachricht beschrieben werden, aber oft sind sie es nicht und in jedem Fall können Commit-Nachrichten nicht so präzise wie ein Stück Code sein. Des weiteren löst die Änderungen mithilfe von Git nachzuvollziehen nicht die Probleme, welche entstehen, wenn die originalen Daten verändert werden: In diesem Fall muss die ganze Arbeit von vorne begonnen werden.

Workflows wie der in diesem Beitrag beschriebene, in welchen die Hauptaufgabe der Kuratoren ist, Pipelines zu schreiben, finden sich häufig in der Informatik (Doltra & Löh 2008; Schoen & Perry 2014) und in der Physik (Peng 2009).

Diese sind auch im Bereich Digital Scholarly Editing vorgeschlagen worden, z.B. von van Zundert (2016) oder Barabucci und Fischer (2017).

Bibliographie

Barabucci, Gioele / Fischer, Franz (2017): „The formalization of textual criticism: bridging the gap between automated collation and edited critical texts“, in: Advances in Digital Scholarly Editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp. Sidestone Press.

CCeH (2014). „sanskrit-dict-to-tei: TEI-fy existing Sanskrit dictionaries.“, https://github.com/cceh/sanskrit-dict-to-tei (Das Repository wird im Laufe des Jahres 2018 veröffentlicht werden).

Crowley, Ronan / Reeve, Jonathan / Schäuble, Johannes (2017). open-editions/corpus-joyce-ulysses-tei: Zenodo release (Version v0.1.1). Zenodo. 10.5281/zenodo.583139

Dolstra, Eelco / Löh, Andres (2008). „NixOS: A purely functional Linux distribution“, in: ACM Sigplan Notices, 43(9): 367-378.

Duvall, Paul M. / Matyas, Steve / Glover, Andrew (2007). „Continuous integration: improving software quality and reducing risk“. Pearson Education.

Flanders, Julia / Muñoz, Trevor (2017). „An Introduction to Humanities Data Curation“, http://guide.dhcuration.org/contents/intro/ [letzter Zugriff 2018-01-10].

Peng, Roger D (2009). „Reproducible research and biostatistics“, in: Biostatistics, 10(3): 405-408.

Reeve, Jonathan (2016). „Git-Lit: an Application of Distributed Version Control Technology toward the Creation of 50,000 Digital Scholarly Editions“, in: Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 657-658.

Schoen, Seth / Perry, Mike (2014). „Why and how of reproducible builds: Distrusting our own infrastructure for safer software releases“, https://air.mozilla.org/why-and-how-of-reproducible-builds-distrusting-our-own-infrastructure-for-safer-software-releases/ [letzter Zugriff 2018-01-10].

Walsh, Norman (2007). „XProc: An XML pipeline language“ in: XML Prague 2007.

van Zundert, Joris J. (2016). „Close Reading and Slow Programming — Computer Code as Digital Scholarly Edition.“, in: ESTS 2016.

Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Federbusch, Maria
maria.federbusch@sbb.spk-berlin.de
Staatsbibliothek zu Berlin Preußischer Kulturbesitz, Deutschland

Herrmann, Elisa
herrmann@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Neudecker, Clemens
clemens.neudecker@sbb.spk-berlin.de
Staatsbibliothek zu Berlin Preußischer Kulturbesitz, Deutschland

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

Einleitung

Die Verwendung von Referenzdaten für das Training und die Auswertung statistischer Annotations- und Analyseverfahren ist ein Kernmerkmal empirischer Forschung, zu der auch die Digital Humanities zählen möchten.[1] Die wichtigste Grundlage für den erfolgreichen Einsatz statistischer Verfahren liegt in der Verwendung geeigneter, den Algorithmen zugrunde liegender Modelle. Für deren Erstellung ist neben einem passenden Lernverfahren das Vorhandensein von Trainingsdaten eine wesentliche Voraussetzung. Werden Forschungsdaten, die mit quantitativen Methoden entstanden sind, mit Referenzdaten verifiziert und interpretiert, ist ein kritischer Blick auf Auswahl, Erstellung und Umgang mit selbigen ein häufig vernachlässigter Bereich.

Vor diesem Hintergrund richtet der vorliegende Beitrag einen kritischen Blick auf die Rolle von und den Umgang mit Ground-Truth-Daten im Bereich der Digital Humanities. Drei Beobachtungen sollen zur Diskussion gestellt werden:

1. Es fehlt den Digital Humanities an einheitlichen und etablierten Ground-Truth-Datensets für die Evaluierung von Forschungsergebnissen auf Basis quantitativer Methoden.
2. Es fehlt den Digital Humanities an Richtlinien zur Erstellung und Verfahren zur Verifizierung von Ground Truth.
3. Es fehlt den Digital Humanities an akzeptierten und operationalisierbaren Metriken zur Qualitätsbestimmung von Ground Truth und abgeleiteten Datenanalysen.

Anhand der automatischen Texterfassung, die als Analogie für ein allgemeines Modell eines empirischen Forschungsprozesses[2] gesetzt wird, sollen im Folgenden die Probleme diskutiert und Möglichkeiten gezeigt werden, wie diesen Defiziten begegnet werden kann.

Abbildung : Gegenüberstellung eines Modells eines Forschungsprozesses aus der empirischen

Sozialforschung[3] und dessen Anwendung auf den Prozess der automatischen Texterfassung

Unter Ground Truth wird in diesem Kontext die Dokumentation ausgewählter Merkmale (Zeichen, Zeilen, Absätze, Spalten, Abbildungen, usw.) des Textes in Form einer digitalen Transkription verstanden. Dabei ist je nach Anwendung zwischen allgemeineren Referenz- und spezifischeren Trainingsdaten zu unterscheiden.

Die Volltext-Digitalisierung von Archiv- und Bibliotheksbeständen, größeren Dokumentsammlungen oder Korpora wird heute von unterschiedlichen Seiten verfolgt: So werden beispielsweise seit 2005 massenhaft Bibliotheksbestände von Google im Rahmen von öffentlich-privaten Partnerschaften sowohl als Bild als auch als Text digitalisiert. Daneben unterstützen Stiftungen, Fördereinrichtungen wie die DFG sowie die Haushaltsmittel der Institutionen die Digitalisierungen im Rahmen spezifischer Projekte. Im Ergebnis dieser Digitalisierungsbemühungen stehen Volltextsammlungen höchst unterschiedlicher Qualität, Vollständigkeit, Interoperabilität und Nachnutzbarkeit.

Mit der OCR-D-Initiative wird erstmals versucht, die technischen und organisatorischen Grundlagen dafür zu schaffen, einen breiten heterogenen Bestand von Drucken aus dem 16. – 18. Jahrhundert vollständig und einheitlich in elektronischen Volltext umzuwandeln und frei zur Verfügung zu stellen. Unter Einbeziehung einzelner Modulprojekte wird vom DFG-geförderten Koordinierungsprojekt[4] die Transformation der Drucke in strukturierten Volltext konzeptuell und prototypisch vorbereitet. Dazu werden im Rahmen von OCR-D Anwendungen, Adaptionen und Weiterentwicklungen von Verfahren der Optical Character Recognition (OCR) für historische Drucke geprüft bzw. implementiert und in einer finalen, prototypischen Produktionsumgebung kombiniert. Eine zentrale Aufgabe von OCR-D besteht dabei in der Bereitstellung eines umfassenden Ground-Truth-Korpus, das sowohl Referenz- als auch Trainingsdaten sowie Richtlinien zur Transkription von Texten für deren Verwendung als Ground Truth umfasst. Damit können sowohl Texte bezüglich ihrer Zeichengenauigkeit transparent geprüft als auch spezielle statistische Modelle für die Text- und Strukturerkennung trainiert werden.

Diskussion

1. Gegenstand der Digital Humanities ist das digitale Objekt, beispielsweise digitaler Text.[5] Vergleicht man den Bereich der automatischen Texterfassung mit einem in der empirischen Sozialforschung etablierten Modell des Forschungsprozesses, wird deutlich, dass die Referenzdaten in beiden Prozessen, im Besonderen in der Phase der Evaluation nach Abschluss der Datenanalyse, eine bedeutende Rolle spielen. In dieser Phase wird auf entsprechende Referenzdaten oder -systematiken zurückgegriffen, die in der Phase der Theoriebildung identifiziert und angesammelt wurden. Anzumerken ist, dass der Forschungsprozess nicht isoliert zu betrachten ist, da er schon zu Beginn bei Formulierung und Auswahl des Forschungsproblems auf vorhandene Forschungsdaten zurückgreift. Der traditionelle Forschungsprozess hat durch ein System von Referenzendaten (u. a. Wörterbücher, Editionen, Nachschlage- und Quellenwerke) ein System der *referenzbasierten* Evaluation geschaffen. Dieses System wird gestützt durch Konventionen der Zitierung, Dokumentation, Verzeichnung und Aufbewahrung in Institutionen sowie spezifischen Publikationsformen. Der Forschungsprozess in den Digital Humanities kann auf diesen Hintergrund nur teilweise bzw. gar nicht zurückgreifen, da bisher zu wenige digitale Daten vorliegen.

Auf der anderen Seite haben die Digital Humanities bezüglich der Verfügbarmachung von Quellcode und Forschungsdaten einen großen Vorsprung gegenüber anderen datenbasiert arbeitenden Disziplinen (etwa der kognitiven Psychologie). *Reproducible Science* wird sowohl gefördert als auch propagiert. Problematisch sind aber die fehlende Vereinheitlichung und Transparenz bei der Datenerhebung sowie der Einsatz mangelhaft erfasster Daten bzw. deren ad-hoc Surrogaten bedingt durch die mangelnde Verfügbarkeit an (insbesondere annotierten) Forschungsdaten.

Die Arbeitsweise mit Referenzdaten, wie sie in den Naturwissenschaften und Teilen der Geisteswissenschaft Anwendung findet, möchte neben der Evaluation, Verifikation gerade die Vergleichbarkeit der Forschungsergebnisse stützen. Das setzt voraus, dass diese Daten in ihrer Qualität diesen Ansprüchen genügen müssen und durch entsprechende Normungen Interpretationsspielräume definiert sind.

Trotz des wissenschaftlichen Anspruches der Digital Humanities auf Objektivität und dem Bemühen Referenzdaten zu schaffen, werden unterschiedliche Interpretationen, die auf Grundlage von mangelhaften Referenz- und Trainingsdaten entstehen, möglich. Solch ein „hinzunehmendes Übel“ wird erkannt und mit „Pragmatismus“, mit der „Flexibilität“ oder „Austauschbarkeit von Konzepten im Konkreten der Texte“ gerechtfer-

tigt[6], eine Vergleichbarkeit der so entstandenen Texte bzw. Forschungsergebnisse damit aber wesentlich behindert. Um eine Vergleichbarkeit im Sinne von *Reproducible Science* zu erreichen, ist somit der Gegenstand der Digital Humanities um die Fragen der Objekterstellung zu erweitern.

2. Mit dem Begriff OCR wird üblicherweise der Gesamtprozess der automatischen Texterfassung bestehend aus den Teilaufgaben Bildvorverarbeitung, Struktur- und Texterkennung sowie gegebenenfalls (automatisierter) Nachkorrektur bezeichnet. Damit eine OCR vorgenommen werden kann, sind entsprechende Erkennungsmodelle notwendig. Diese werden durch sogenanntes Training unter Nutzung von Ground Truth induziert. Generische Modelle werden vom Anbieter der OCR-Software zur Verfügung gestellt, domänenspezifische Modelle müssen trainiert werden. Das Training dient immer der Verbesserung der Endergebnisse des Erkennungsprozesses. Somit sollten bei einem universellen Anspruch Ground-Truth-Daten sowohl spezifische als auch allgemeine Normungen enthalten, damit sowohl generische als auch spezifische Modelle trainiert werden können. Ziel ist es, die unterschiedlichen Bedürfnisse und Schwerpunkte in den Digital Humanities zu bedienen.

3. Um diesem sehr breiten Anspruch gerecht zu werden, reicht eine Akklamation, dass diese oder jene Sammlung von Daten als Ground Truth bezeichnet und genutzt werden kann, nicht aus. Es bedarf hingegen eines Ground-Truth-Konzepts. Dieses Konzept dokumentiert sowohl dessen inneren Aufbau als auch dessen Erstellung. Damit können im Bereich der Texterfassung beispielsweise folgende Punkte im Umgang mit Ground Truth erreicht werden:

1. (weitere) Reduktion bzw. Standardisierung der Freiheitsgrade bei der Transkription (z. B. langes s, Ligaturen, Zeilenumbruch)
2. weitergehende Operationalisierung der Überprüfbarkeit der Validität der Transkription
3. Ergänzung von (weiteren) Anweisungen zum Umgang mit koordinatenbasierten Phänomenen

Illustriert werden kann ein solches Konzept am Beispiel der *OCR-D Ground-Truth-Guidelines*. Um die Freiheitsgrade von Interpretationen (Punkt A) zu normieren werden drei Level von Erfassungsgraden angeboten. Die einzelnen Level sollen nachvollziehbare Interpretationsentscheidungen sowohl festlegen als auch dokumentieren und damit die Möglichkeit der maschinellen Überprüfbarkeit eröffnen.

Tabelle : Beispiel der Anwendung der Level bei der Ligatur ct.

Zeichen	Level 1	Level 2	Level 3
	ct Die Ligatur wird in zwei einzelne Zeichen aufgespalten.	ct Die Ligatur wird aufgespalten und mit einer zusätzlichen Annotation, dass es sich um eine Ligatur handelt, im PAGE-Format versehen. textStyle{offset:0;length:2;ligatur:true;}	 Die Ligatur wird als **ein** Zeichen interpretiert und mit dem entsprechenden Unicode-Zeichen wiedergegeben.

Damit werden Anweisung und Richtlinien weitgehend operationalisiert und eine computergestützte Validierung umsetzbar (Punkt B). Das entsprechende Ground-Truth-Korpus liegt im XML-basierten PAGE-Format[7] vor. Das PAGE-Format hat sich im Rahmen des EU-Projekts IMPACT[8] sowie durch seine Verbreitung im Rahmen von Wettbewerben bei wissenschaftlichen Konferenzen (z.B. ICDAR, ICFHR, DAS) als de-facto Standard für XML-basierter Ground Truth etabliert. Mit Hilfe von Schematron[9]-Regeln kann nun, wie das Beispiel zeigt, geprüft werden, nach welchem Level die „ct“-Ligatur kodiert ist:

```
<pattern id="ct_ligatur">
<let name="x"
value="(//page:Unicode[text()[contains(.,'ct')]])"/>
<rule context="//page:Unicode[text()[contains(.,'ct')]]
<report test="$x" role="WARNING">[W0001]
The document contains splitted ligature ct.
OCR-D Level 1 </report>
</rule>
</pattern>
```

Die Wahl des PAGE-Formates ermöglicht im Unterschied zum TEI-basierten Basisformat des DTA (DTABf)[10] eine unkomplizierte Lösung für die Repräsentation von koordinatenbasierten Phänomenen (Punkt C).

```
<Word id="word_1479724691818_218" language="German" custom="re
 structure {type:signature-mark;} textStyle {offset:0; length:
    <Coords points="1111,2184 1037,2184 1037,2123 1111,2123"/>
    <TextEquiv>
      <Unicode>a</Unicode>
    </TextEquiv>
    <TextStyle fontFamily="Antiqua" fontSize="26.0"/>
</Word>
```

Fazit

Ground Truth stellt im Texterkennungsprozess und im dazu betrachteten Forschungsprozess in den Digital Humanities eine entscheidende Rolle dar. Ergebnis dieser beiden Prozesse sind immer Publikationen, die als Forschungsdaten in neuen Forschungszusammenhängen nachgenutzt werden. Gelingt es den Digital Humanities nicht, ein Referenzsystem mit Konventionen zur Prüfung, Dokumentation, Verzeichnung und Aufbewahrung ihrer Daten in Institutionen sowie spezifischen Publikationsformen aufzubauen, ist die Vergleichbarkeit der Forschungsergebnisse nicht immer gegeben. Dabei bietet der Aufbau so genannter Forschungsdatenrepositorien erste positive Entwicklungen in Richtung dokumentierter Forschungsprozesse.[11]

Der Gegenstand der Digital Humanities ist nicht nur auf das digitale Objekt zu beschränken, sondern auch auf dessen Erstellung zu erweitern. Die Erstellung ist ein Prozess, der von den Digital Humanities zu dokumentieren ist, da nur so eine Referenzierbarkeit und Reproduzierbarkeit der Forschungsergebnisse möglich wird. Der kritische Umgang mit diesen Daten stellt eine Aufgabe der Wissenschaft dar.

Daher täten die Digital Humanities gut daran, in einen aktiven Dialog mit digitalisierenden Einrichtungen und Förderern einzutreten um gemeinsame Standards und Richtlinien zur Dokumentation zu etablieren, die transparent Auskunft über die Provenienz eines digitalen Objekts geben sowie klar über Möglichkeiten sowie Einschränkungen zu dessen Nachnutzbarkeit informieren. Entscheidungen, Kriterien, Daten und Ergebnisse sollten, im Unterschied zur bisherigen, in diesem Beitrag kritisierten Praxis, zukünftig möglichst transparent, operationalisierbar sowie validierbar digital zur Verfügung gestellt werden.

Fußnoten

1. Vgl. dazu These 1.1: „Die Digital Humanities bereichern die traditionellen Geisteswissenschaften konzeptionell und methodisch - ihre Werkzeuge und Verfahren ergänzen das „Wie“ unserer Praxis um *eine empirisch ausgerichtete Epistemologie.*“ [Thesenpapier des Fachverbands „Digital Humanities im deutschsprachigen Raum“ 2014]
2. siehe [Schnell, Hill, Esser 2011: 4]
3. siehe Fußnote 2
4. OCR-D: Koordinierungsprojekt zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR), http://www.ocr-d.de/.
5. „Es werden weitere korpusbasierte Analysen angestrebt. Diese erfordern aber eine Voraussetzung: Im Zuge der ‚digitale[n] Wende‘ [Schöch 2014: 130] ist es weiterhin wünschenswert und erforderlich, dass immer mehr literarische Texte digital zur Verfügung stehen oder diese durch leichte und praktikable Verfahren der Texterkennung (OCR, optical character recognition) digitalisierbar gemacht werden können.“ [Mihm 2016: 200]
6. „Dass die von uns einbezogenen Online-Repositorien hinsichtlich der editionsphilologischen Textqualität variieren ist ein hinzunehmendes Übel, dem wir zum einen pragmatisch (Wahl der bestmöglichen verfügbaren Ausgabe; Ziel, die Fehlermarge unter 2% zu halten), zum anderen unter Hinweis auf die flexible Struktur des Korpus (Austausch durch eine qualitativ hochwertigere Version ist möglich) begegnen. Durch die nahtlose Dokumentation des Korpus wird zudem die nötige Transparenz gewährleistet um auch Nachnutzern flexible Kontrolle der Daten zu ermöglichen.“ [Herrmann-Wolf, Lauer 2016: 159]
7. Page Analysis and Ground Truth Elements, siehe http://www.primaresearch.org/publications/ICPR2010_Pletschacher_PAGE und https://github.com/PRImA-Research-Lab/PAGE-XML.
8. Vgl. https://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/.
9. ISO/IEC-Standard 19757-3:2006
10. Deutsches Textarchiv, DTA-Basisformat, http://www.deutschestextarchiv.de/doku/basisformat/.
11. Vgl. z.B. https://rdmorganiser.github.io/

Bibliographie

Herrmann-Wolf, J. Berenike / Lauer, Gerhard (2016): „Aufbau und Annotation des Kafka/Referenzkorpus“ in: Burr Elisabeth (ed): *DHd 2016 : Modellierung, Vernetzung, Visualisierung : die Digital Humanities als fächerübergreifendes Forschungsparadigma* : Konferenzabstracts : Universität Leipzig 7. bis 12. März 2016. [Duisburg]: nisaba 158-160.

Mihm, Melanie (2016): „Weibliches Erzählen im Expressionismus? Eine Stilometrie von Mela Hartwigs Prosa“ in: Burr Elisabeth (ed): *DHd 2016 : Modellierung, Vernetzung, Visualisierung :*

die Digital Humanities als fächerübergreifendes Forschungsparadigma : Konferenzabstracts : Universität Leipzig 7. bis 12. März 2016. [Duisburg]: nisaba 198-200.

Schöch, Christof (2014): „Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“, in: Schöch, Christof / Schneider, Lars (eds.): *Literaturwissenschaft im digitalen Medienwandel* (=Philologie im Netz Beiheft 7) 130-157 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf [letzter Zugriff 25.September 2017].

Schnell, Rainer / Hill, Paul B. / Esser, Elke (2011): Methoden der empirischen Sozialforschung. 9., aktualisierte Aufl. München: Oldenbourg

Thesenpapier des Fachverbands „Digital Humanities im deutschsprachigen Raum“ (DHd) (2014): „Digital Humanities 2020“, vorgestellt im März 2014 auf der ersten Jahrestagung des Verbandes in Passau. http://dig-hum.de/digital-humanities-2020.

Hinterlistig – schelmisch – treulos – Sentiment Analyse in Texten des 19. Jahrhunderts: Eine exemplarische Analyse für Länder und Ethnien

Wodausch, David
wodausch@uni-hildesheim.de
Universität Hildesheim, Deutschland

Fiedler, Maik
fiedler@gei.de
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung (GEI)

Heuwing, Ben
heuwing@uni-hildesheim.de
Universität Hildesheim, Deutschland

Mandl, Thomas
mandl@uni-hildesheim.de
Universität Hildesheim, Deutschland

Abstract

Sentiment Analyse ist eine Standard-Methode des Text Mining und prüft, was in Texten meinungsbehaftet behandelt wird. Damit ist es auch für Digital Humanities sehr interessant. Viele Ansätze basieren auf einfachen Listen, sogenannten Sentiment Lexika, welche sich aber nicht vollständig auf andere Domänen und auch kaum auf andere Zeiträume übertragen lassen. Dieser Beitrag zeigt beispielhaft, wie eine solche Liste methodisch für Schulbücher des 19. und frühen 20. Jhds. erstellt werden kann. Am Beispiel der Behandlung des Judentums in den historischen Schulbüchern wird gezeigt, dass automatische Ansätze des Text Mining alleine nicht ausreichen, um die Komplexität der Bewertung zu erfassen. Erst durch die Verknüpfung intellektueller und digitaler Ansätze entstehen interessante Ergebnisse. Zudem wird gezeigt, dass ein aktuelles Sentiment Lexikon für diese historischen Epoche ungeeignet ist.

Einleitung

Text Mining ist in den Digital Humanities eine häufig eingesetzte Technologie mit vielen Facetten. Die Sentiment Analyse versucht zu analysieren, welche Meinungen ein Text zu einer Entität ausdrückt. Dazu werden die im Kontext von Erwähnungen der Entität vorkommenden Stellen ausgewertet (Liu 2012). Häufig ist eine Meinung nicht leicht maschinell zu erkennen, so dass leichtgewichtige linguistische Ansätze (Struß 2016) oder auch tiefgehende Analysen notwendig sind (Somasundaran et al. 2008). Für viele Anwendungen werden aber einfache listenbasierte Ansätze genutzt, welche das Vorkommen von positiven und negativen Wörtern einer Sprache im Umfeld einer Entität oder einen Konzept, das mehrere Wörter berücksichtigt, auswerten.

In der hier geschilderten Untersuchung sollte geprüft werden, ob eine Liste für historische Texte effizient erstellt werden könnte und inwieweit damit fachwissenschaftliche Fragestellungen analysiert werden könnten. Im Rahmen des Projektes „Welt der Kinder“ (De Luca 2014) stand eine Kollektion von über 3500 Schulbüchern des 19. und frühen 20. Jhd. ebenso zur Verfügung wie Werkzeuge zur Analyse (wdk.gei.de). Die Basis für die Systementwicklung bildete eine Analyse der Informationsbedürfnisse der von Historikern (Heuwing et al. 2016). Unter anderem wurden Topic Models und darauf basierende Filter entwickelt (Schober & Gurevych 2015).

Schulbücher eignen sich besonders als Gegenstand der Forschung zum 19. Jhd. Diese Bücher waren oft die einzigen Medien, welche Kinder ausgesetzt werden. Sie prägten somit ihr Weltbild nachhaltig (Fuchs 2014). Die Dokumente sind durch OCR in Volltexte umgewandelt. Dabei ergibt sich insbesondere für Fraktur eine Fehlerrate. Diese ist jedoch gering und eine quantitative Auswertung ist gleichwohl möglich.

Methode

Ziel des Vorgehens war es, ein erstes und vorläufiges Sentiment Wörterbuch aus Schulbüchern des 19. Jhds. zu erstellen und zu prüfen, inwieweit dies sich von einer aktuellen Liste unterscheidet. Dazu wurde folgendes methodische Vorgehen gewählt, das zwar nicht ausreicht, ein vollständiges Sentiment Lexikon für die Kollektion zu erstellen, aber doch zu interessanten Resultaten führt. Durch die Auswahl von Begriffen, die häufig mit Sentiment verbunden sind, sollten Textstellen identifiziert werden, die dann einer manuellen Auswertung unterzogen wurden. Bei dieser Bewertung wurden die meinungstragenden Phrasen als positiv oder negativ bewertet.

Konkret wurde zunächst aus dem Gesamtkorpus drei Untermengen ausgewählt, wobei Anfragen gewählt wurden, die nach ersten Hypothesen aus Perspektive der Geschichtswissenschaft zu Aussagen mit Meinungen führen würden.

- In einer Untermenge, welche zu Völkern und Staaten generiert wurde, erfolgte eine Recherche nach den Suchbegriffen Deutschland und Frankreich.
- Der Name Napoleon Bonaparte wurde in Jungen und Mädchenbüchern getrennt recherchiert, um so mögliche unterschiedliche Darstellungen zu erkennen.
- Mit dem Wortfeld Jude wurde in Büchern aus zwei Zeiträumen gesucht. Diese umfassten 1850-1859 und 1910-1919. Dabei stand die Hypothese im Raum, dass die verwendete Sprache des Zeitraums 1850 bis 1859 einen weniger radikalisierten Antisemitismus aufweist als die Analyse der Textquellen von 1910 bis 1919. Diese Einschätzung ist auf den ab 1879 bis nach den Ersten Weltkrieg sich radikalisierenden Antisemitismus zurückzuführen.

Das Werkzeug AntConc erlaubt es, Konkordanzen zu extrahieren. Die Suchbegriffe wurden als zentrale Wörter dargestellt und AntConc liefert dann den Kontext der Phrase. Die Sortierung der Phrasen erfolgt nach dem Maß Mutual Information.

So wurden zu jeder Suche und damit zu jedem der sechs Sub-Korpora jeweils ca. 100 Phrasen manuell ausgewertet und klassifiziert. Dabei wurde festgehalten, ob eine meinungstragende Phrase vorliegt. Zudem wurde über die Polarität entschieden, also ob die Meinung positiv oder negativ ist. Diese manuelle Annotation semantischer Werte in Textabschnitten erfolgte, um darauf aufbauend das Sentiment Lexikon zu erstellen.

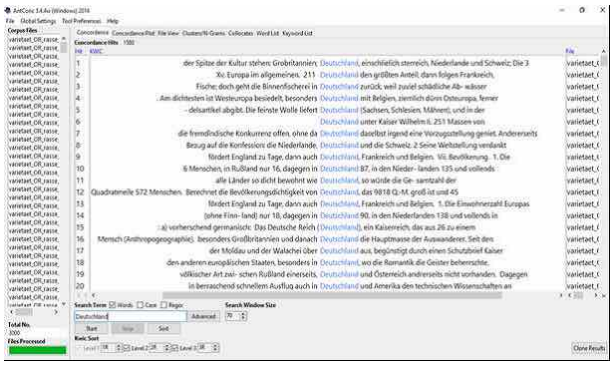


Abbildung 1: Kollokationen in AntConc

Ergebnisse

Für die erste Textmenge ergaben sich dabei 270 unterschiedliche Kollokationen, die eine positive oder negative interpretierte Meinung zu Deutschland oder Frankreich besitzen. Dabei wird Frankreich negativer dargestellt, 49% der meinungstragenden Phrasen sind negativ. Bei Deutschland sind es 39%. Diese beziehen sich beispielsweise auf die Situation vor 1871 und kennzeichnen diese als Zersplitterung.

Für die Textstellen zu Napoleon Bonaparte konnten 174 wertende Phrasen erkannt werden. Allerdings ergaben sich keine nennenswerten Unterschiede in der Darstellung der Person in Jungen und Mädchenschulbüchern.

Für die qualitative Analyse der beiden Stichproben aus den Subkorpora für das Judentum wurden für den Veröffentlichungszeitraum von 1850 bis 1859 insgesamt 720 und beim zweiten Subkorpus von 1910 bis 1919 insgesamt 825 Konkordanzen untersucht. Dabei wurden insgesamt 231 Kollokationen aus 155 wertenden Konkordanzen gefunden. Insgesamt waren die meisten Vorkommen von Begriffen aus dem Wortfeld Jude und Judentum also nicht meinungsbehaftet.

Die Auswertung ergab, dass durchschnittlich 76,5% aller wertenden Kollokationen in den untersuchten Konkordanzen der Subkorpora als negativ klassifiziert wurden (Subkorpus (1): 102 negativ (77%); Subkorpus (2): 75 negativ (76%)). Somit ist das Verhältnis zwischen negativ wer-

tenden und positiv wertenden Kollokationen beider Sub-Korpora annähernd gleich, wobei negative Äußerungen klar überwiegen. Allerdings ergibt sich aus dieser rein quantitativen Betrachtung keine Meinungsentwicklung gegenüber Juden während des Kaiserreichs, wie sie die Hypothese vermutet hatte.

In der Folge wurden die Kollokationen manuell in sechs, in sich homogene Gruppen geclustert, welche jeweils den lokalen Kontext und das Auftreten einer Kollokation beschreiben: Religion; Synonym für Jude; positive Eigenschaft Jude; negative Eigenschaft Jude; Juden im Gesellschaftsleben und Sonstiges. Entwickelt bzw. festgelegt wurden die sechs Cluster auf Grundlage der Kollokationen.

Die Analyse der Vorkommenshäufigkeit der sechs Cluster ergab eine zeitliche Veränderung. Im Zusammenhang mit Religion wurde das Judentum im früheren Subkorpus ab 1850 häufiger erwähnt als im Zeitraum von 1910 bis 1919. Im späteren Zeitraum wird das Judentum häufiger im Zusammenhang mit dem gesellschaftlichen Leben erwähnt. Diese spiegelt möglicherweise den Wandel des Judenhasses während der zweiten Hälfte des 19. Jahrhunderts auf Grundlage religiöser Motive zu einer biologisch-rassischen Begründung.

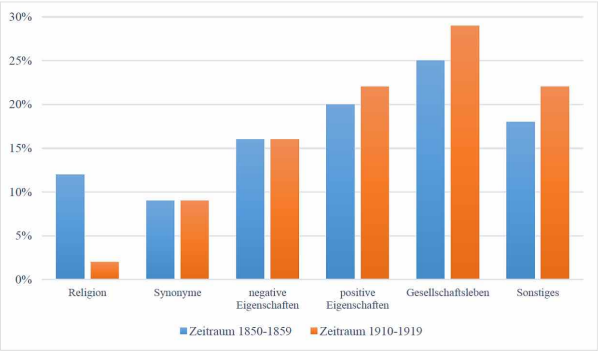


Abbildung 2: Cluster zum Wortfeld Jude/Judentum

Insgesamt wurden 225 einmalig auftretende meinungstragende Begriffe gefunden. Von diesen sind nur 44% in einer aktuellen Liste des Deutschen vorhanden (Sentiment Wortschatz der Universität Leipzig). Zwar ist die verwendete Liste deutlich länger und aus umfangreicheren Studien hervorgegangen, doch allein dieser Vergleich zeigt, dass das historische Vokabular sich deutlich unterscheidet. Nicht enthalten sind beispielsweise: verachtet, zerstreut, unkünstlerisch oder halsstarrig.

Auffällig ist, dass im historischen Korpus deutlich mehr meinungstragende Substantive enthalten sind als im aktuellen Korpus, in dem Adjektive überwiegen. Dies müsste weiter untersucht werden. Möglicherweise wurden Substantive mit spezifischen Zielen wie etwa einer oberflächlichen Objektivierung eingesetzt. Beispiele für Substantive sind: Zarenwahnsinn, Blitzesschnelle, Engherzigkeit oder Schlappe.

Die vorläufigen Ergebnisse zeigen, dass nur eine Verknüpfung des iterativen Einsatzes analoger manueller und digitaler automatischer Methoden verschiedene Perspektiven auf Texte einnehmen kann und zu sinnvollen Ergebnissen führen kann. Dies fordert auch Stephen Ramsay unter dem Schlagwort Algorithmic Criticism (Ramsay 2008).

Mit Criticism wird dabei nicht die Überprüfung hermeneutischer Hypothesen mit algorithmischen Analysen gemeint, sondern die Reflexion, Explikation und Differenzierung von Geisteswissenschaften und Algorithmen sowie die Einbindung multiperspektivischer Zugänge zu Untersuchungsgegenständen (vgl. Bender, 2016, S. 300f.). Möglichkeiten für weitere Forschung liegen in der Verknüpfung mit anderen Quellen für historisches Vokabular.

Bibliographie

Bender, Michael (2016): Forschungsumgebungen in den Digital Humanities: Nutzerbedarf, Wissenstransfer, Textualität. Walter de Gruyter GmbH & Co KG

De Luca, Ernesto William (2014): Welt der Kinder. Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung. URL: http://welt-der-kinder.gei.de

Fuchs, Eckhardt (2014): Das Schulbuch in der Forschung. Analysen und Empfehlungen für die Bildungspraxis. Georg-Eckert-Institut für internationale Schulbuch-forschung. Band 4. V & R unipress in Göttingen

Heuwing, Ben / Mandl, Thomas / Womser-Hacker, Christa (2016): “Combining contextual interviews and participative design to define requirements for text analysis of historical media”. In: Information Research 21 (4) http://www.informationr.net/ir/21-4/isic/isic1606.html

Liu, Bing (2012): Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers

Ramsay, Stephen (2008): “Algorithmic Criticism”. In: A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, http://www.digitalhumanities.org/companionDLS/

Schnober, Carsten / Iryna Gurevych (2015): “Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project.” In Proceedings of

the 2015 Workshop on Topic Models: Post-Processing and Applications, TM '15. New York, NY, USA: ACM, 2015, S. 11–20.

Somasundaran, Swapna / Josef Ruppenhofer / Janyce Wiebe (2008) "Discourse level opinion relations: An annotation study." In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics.

Struß, Julia Maria (2016): Multilinguales aspektbasiertes Opinion Mining: Entwicklung eines ressourcenarmen Extraktionsverfahrens und Untersuchung von Nutzerperspektiven. Dissertation Universität Hildesheim.

Hin zu einer Visuellen Stilometrie: Automatische Genre- und Autorunterscheidung in graphischen Narrativen

Dunst, Alexander
alexander.dunst@gmail.com
Universität Paderborn, Deutschland

Hartel, Rita
rst@upb.de
Universität Paderborn, Deutschland

1. Einleitung

Stilometrische Untersuchungen haben eine lange Tradition in der Literaturwissenschaft und erleben mit der Digitalisierung von Textkorpora und Analysemethoden in den letzten drei Jahrzehnten einen erneuten Aufschwung (Holmes und Calle-Martín & Miranda-García). Ähnliche Tendenzen sind in den vergangenen Jahren auch in der Kunstgeschichte zu verzeichnen (Kohle; Qu, Taeb & Hughes; Manovich). Im Gegensatz dazu sind stilometrische Untersuchen visueller Erzählungen noch nicht etabliert. In den Bereich visueller, oder multimodaler, Narrative fallen etwa Comics, Filme und Fernsehserien, aber auch zu einem gewissen Grad Computerspiele, und damit viele der populärsten Erzählformen des 20. und 21. Jahrhunderts. Das Fehlen einschlägiger Forschung gründet einerseits auf die technischen Herausforderungen digitaler Bildanalyse, andererseits auf den Fokus auf qualitative und ideologiekritische Zugänge in großen Teilen der Medien- und Kulturwissenschaft. Auf Basis eines Corpus von 138 graphischen Romanen (Comicbüchern in Romanlänge) stellt unser Vortrag eine Methode zur automatischen Genre- und Autorunterscheidung vor. Weiters erörtern wir die Herausforderungen stilometrischer Methoden für Erzählformen, die Text und Bild kombinieren, und diskutieren abschließend die potenzielle Übertragbarkeit des vorgestellten Zugangs auf andere Medien.

2. Datenbasis & Methode

Die Basis der Untersuchungen stellt das erste repräsentative Corpus graphischer Romane dar, dass sich derzeit im Aufbau befindet (Dunst, Hartel & Laubrock). Zum Zeitpunkt der Untersuchungen im Mai 2017 waren 160 Volltexte digitalisiert; davon wurden aufgrund mangelnder Scanqualität bei einigen Titeln 138 Bücher mit einer Gesamtmenge von rund 33.000 Seiten für die Corpusstudie herangezogen. Der Fokus auf Bildanalyse ergab sich dabei sowohl aus methodischen als auch aus praktischen Überlegungen: einerseits sind Methoden zur stilometrischen Textanalyse bereits etabliert und werden laufend weiterentwickelt. Viele davon lassen sich direkt auf Comicstext anwenden oder dafür adaptieren. Andererseits bereitet eine zuverlässige, automatisierte Textlokalisierung und -Erkennung in Comics weiterhin Probleme. Damit bleibt derzeit nur der Weg über eine semi-automatische und zeitintensive Textannotation. Für größere Korpora bietet sich diese Methode daher nicht an. Drei visuelle Maße stellten die Basis unserer Berechnungen dar. Für all diese Maße wurde das Bild zunächst mithilfe einer linearen Näherung der Luminanz in eine Graustufen-Darstellung des Bildes umgewandelt.

- Median der Helligkeit einer Seite: der mittlere Grauton aller Pixel eines Bildes
- Shannon-Entropie: Entsprechend der Shannon-Entropie aus der Informationstheorie: $H(X) := -\Sigma_{i=1}^{n} P(x_i) * log_2(P(x_i))$. Hierbei ist die betrachtete Nachricht X die Liste aller Helligkeitswerte/Grautöne eines jeden Pixels. Ein schwarz-weiß Bild hat eine Entropie zwischen 0 und 1 (da es nur ‚binäre' Informationen (schwarz oder weiß) enthält). Die höchste Entropie (den höchsten Informationsgehalt) hat ein Bild, in dem alle Graustufen möglichst gleich verteilt vorkommen.
- Anzahl der Flächen: Hierzu wird das Bild in fünf binäre Bilder aufgeteilt, entsprechend

fünf unterschiedlicher Helligkeitsstufen. Das Bild B_i zur Helligkeitsstufe H_i enthält an den Stellen ein Bit „1“, an denen das Originalbild ein Pixel der Helligketisstufe H_i enthält. Anschließend werden alle zusammenhängenden Flächen von 1-Bits gezählt und für alle fünf Helligkeitsstufen aufaddiert.

Im Anschluss an die Kalkulation dieser Grundmaße berechneten wir aus den Ergebnissen pro Seite die Mediane für jedes einzelne der 138 Werke. Um stilistische Abweichungen innerhalb eines Werkes zu messen, berechneten wir die Standardabweichung von den drei Maßen. Um die stilistische Entwicklung innerhalb eines Werkes zu berechnen, wurden in einem ersten Schritt manuell Inhalts- von Funktionsseiten getrennt und letztere von der Kalkulation ausgeschlossen. Hierzu betrachten wir die Kurvenverläufe der drei visuellen Maße über alle Seiten eines graphischen Romans. Für jedes visuelle Maß ermittelten wir: die Anzahl Extrema je 100 Seiten, die Standardabweichung aller Messwerte, die Standardabweichung innerhalb der lokalen Minima bzw. Maxima, die Regelmäßigkeit einer Kurve (also die Standardabweichung des Abstands in Seiten zweier aufeinanderfolgender Extrema) und eine Klassifikation des Kurven-Beginns und -Endes (die erste/letzte Seite hat einen größeren/kleineren Messwert als die nächste/vorherige Seite). Letztere Berechnungen für die stilistische Entwicklung ergaben sieben Werte je Maß, also insgesamt 21 Ergebnisse pro graphischem Roman. Diese reduzierten wir mit Hilfe von skalierter Hauptkomponentenanalyse (Principal Component Analysis [PCA]). Die errechneten Maße zogen wir im Anschluss für die Analyse folgender literatur-, bzw. kulturwissenschaftlicher Konzepte heran: Genre, Autorschaft, und Publikationsformat. Aus Zeitgründen stellen wir hier nur Teilergebnisse für die ersten beiden Kategorien vor.

3. Ergebnisse & Diskussion

3.a. Genre

Unser Korpus enthält sowohl fiktionale als auch nicht-fiktionale Texte, die trotz dieser Unterschiede oftmals als graphische Romane bezeichnet werden. Tatsächlich erscheint es sinnvoller hier von graphischen Narrativen zu sprechen. Dabei handelts es sich um durchgehende Erzählungen im Medium Comics, die sich an ein erwachsenes Publikum wenden und einen Umfang von mehr als 64 Seiten haben, womit sie sich vom traditionellen Seitenformat von Comicbüchern abheben. Einzelnen Büchern wurde auf Basis von Klappentexten, Verlagsinformaionen und Zusammenfassungen eines der folgenden Genres zugewiesen: Graphic Novel, Graphic Memoir, andere nicht-fiktionale Texte (Sachbücher). Texte, denen die Subgenres Superhelden, Science Fiction, Märchen, Fantasy, Mystery und Horror zugewiesen wurden, fassten wir in der Sammelkategorie Graphic Fantasy zusammen. Eine kleine Zahl an Texten, die unter keine dieser Klassifikationen fielen, fassten wir unter dem Begriff „miscellaneous“. Die Ergebnisse der Untersuchen wurden mit Hilfe von Welch Two Sample T-Tests mit $p<0,05$ auf ihre statistische Signifikanz untersucht.

Dabei zeigte sich, dass sich die Genres Graphic Novel, Graphic Memoir und Graphic Fantasy signifikant voneinander unterscheiden. Im speziellen zeichnen sich graphische Romane und Memoiren durch einen deutlich regelmäßigeren visuellen Stil aus. Titel mit der niedrigsten Anzahl an Flächen und niedriger Standardabweichung von der durchschnittlichen Helligkeit gehören meist zu diesen beiden Genres. Graphische Memoiren sind außerdem deutlich heller als die Erzählungen, die wir unter Graphic Fantasy zusammengefasst hatten. Illustration 1 zeigt, dass sich auch eine historische Entwicklung nachweisen lässt. So werden graphische Memoiren seit Mitte des letzten Jahrzehnts deutlich heller – eine Entwicklung, die möglicherweise durch die beispielgebende Wirkung von mittlerweile kanonischen Titeln wie Alison Bechdels *Fun Home* bedingt ist.

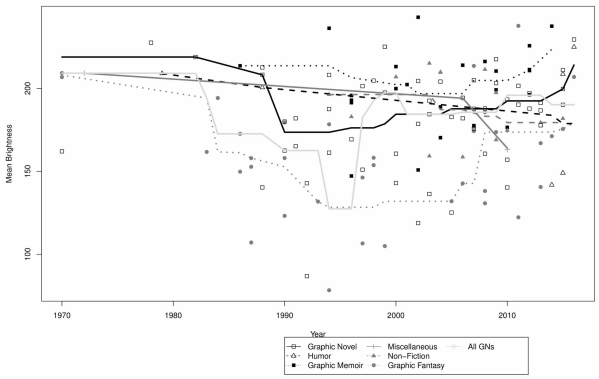


Illustration 1: Historische Entwicklung der durchschnittlichen Helligkeit per Genre

3.b. Autorschaft

Ähnlich wie im Fall der Genreunterscheidungen zeigten sich signifikante visuelle Unterschiede zwischen unterschiedlichen Autoren. Illustration 2 visualisiert die stilistische Handschrift von

Künstlern, die mir mehr als drei Titeln in unserem Corpus vertreten sind. Jene, die einen von Werk zu Werk vergleichsweise konstanten Stil bevorzugen, nehmen dabei eine kleinere Fläche in der Matrix ein. Darunter fallen etwa die bekannten Autoren Jason Lutes, Chester Brown und der japanische Manga-Autor Ozamu Tezuka. Hingegen nehmen Frank Miller, Dave McKean und David Mazzuchelli aufgrund ihrer visuellen Bandbreite eine größere Fläche im stilistischen Raum (Manovichs „style space“) graphischer Erzählungen ein. Die Einbeziehung von Tezuka zeigt als Testfall aber auch die momentanen Grenzen dieser Methode auf. Die verwendeten Maße erlauben keine Differenzierung zwischen dem angloamerikanischen graphischen Roman und dem japanischen Manga, zwei Erzählformen, die trotz gegenseitiger Beeinflussung deutliche stilistische Unterschiede aufweisen.

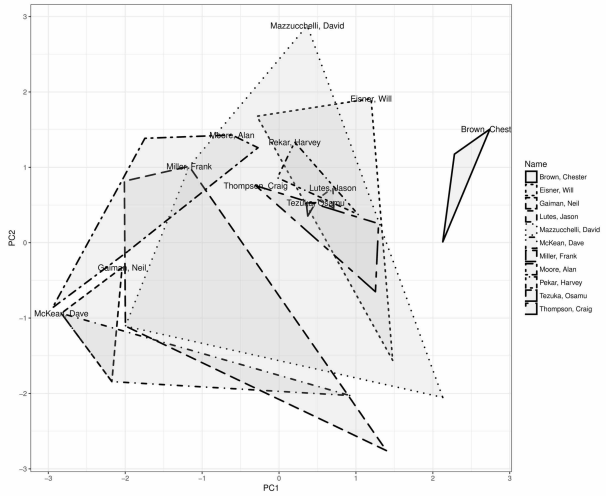


Illustration 2: PCA stilistischer Handschrift von Comicsautoren

4. Zusammenfassung & Ausblick

In diesem Vortrag haben wir automatische Bildanalysen vorgestellt, die stilometrische Unterscheidungen zwischen Genres und Autoren auf den Bereich visueller Erzählungen übertragen. Dazu ist anzumerken, dass sich dieser Zugang in einem experimentellen Stadium befindet. So werden wir demnächst zusätzliche visuelle Maße zu den hier verwendeten erproben. In einem ersten Schritt ist dabei an Farbigkeitsmaße gedacht. Die Datenbasis wird mit Fortschreiten der Retrodigitalisierung unseres Corpus auf rund 250 Werke erweitert. Bei den Genreunterscheidungen hat sich die relativ aufwändige Berechnung der internen Entwicklung eines Werkes als wenig aussagekräftig herausgestellt. Allerdings fließt diese bei den Autorenunterscheidungen in die PCA ein. Auch die Genrekategorien bedürfen einer weiteren Verfeinerung: bei Versuchen, die interne Kohärenz der Genres zu berechnen (Distanzmaße vom Zentroiden eines Clusters) hat sich die Sammelkategorie Graphic Fantasy als weniger kohärent als eine zufällige Vergleichsgruppe herausgestellt. Hier sollten bei neuerlichen Berechnungen Subgenres wie Superheldennarrative herangezogen werden. Als graphische Kunstwerke, die von Hand gezeichnet werden, unterscheiden sich Comics deutlich von Filmen, Fernsehserien, aber auch von Computerspielen. Dennoch birgt die hier präsentierte Methode unserer Meinung nach Potenzial für diese Medien, da Fragen zur Stilistik von Genre und Autorschaft auch dort eine Rolle spielen. So könnte etwa erforscht werden, ob sich das Genre des Film Noir tatächlich stilistische Kohärenz aufweist, oder ob sich komplexe Qualitätsserien wie *The Wire* oder *The Sopranos* visuell von anderen TV-Erzählungen unterscheiden.

Fußnoten

Zwar hatte Manovich stilistische Bildanalysen auch auf Manga angewandt. Diese Studien blieben aber weitgehend explorativ und ließen keine Rückschlüsse auf Konzepte wie Autorschaft oder Genre zu (Manovich, Douglas & Zepel).
Diese bietet die Beta-Version unseres graphischen XML-Editors an, der unter folgender Adresse frei verfügbar ist: https://groups.uni-paderborn.de/graphic-literature/wp/?page_id=3592.

Bibliographie

Calle-Martin, J. & A. Miranda-García (2012). „Stylometry and Authorship Attribution: Introduction to the Special Issue“, *Stylometry and Authorship Attribution*, special issue of *English Studies* 93-3: 251-58.

Dunst, A., R. Hartel & J. Laubrock (2017). „The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities” *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)* [im Druck].

Holmes, D (1998). „The Evolution of Stylometry in Humaninities Scholarship“ *Literary and Linguistic Computing* 13-3: 111-17.

Kohle, H. (2013). *Digitale Bildwissenschaft*. Glückstadt: VWH.

Manovich, L. (2011) “Style Space: How to Compare Image Sets and Follow their

Evolution", http://manovich.net/index.php/projects/style-space.

Manovich, L., J. Douglas, T. Zepel (2011). „How to Compare One Million Images", http://manovich.net/index.php/projects/how-to-compare.

Qi, Hanchao, Armeen Taeb & Shannon M. Hughes , „Visual stylometry using background selection and wavelet-HMT-based Fisher information distances for attribution and dating of impressionist paintings" *Signal Processing* 93-3 (2013): 541-53.

Historische Zeitungen kollaborativ erschließen: Die älteste, noch erscheinende Tageszeitung der Welt "under annotation"

Resch, Claudia
claudia.resch@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Kampkaspar, Dario
dario.kampkaspar@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Schopper, Daniel
daniel.schopper@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Die digitale Erschließung historischer Zeitungen lag vor wenigen Jahrzehnten noch außerhalb des Vorstellbaren. Inzwischen ist die Zeitung als Forschungsgegenstand etabliert und damit für viele verschiedene Disziplinen ins Zentrum gerückt: NutzerInnen sehen Nachrichtenblätter nicht nur als Quelle, die punktuell und komplementär zu anderen Texten befragt wird, sondern haben auch ein besonderes Interesse an der diachronen Entwicklung von historischen Zeitungen und ihren Themen, von Textsorten und natürlich von Sprache ganz im Allgemeinen. Die Transformation historischer Zeitungen in ein digitales Format, die von Bibliotheken, Archiven und Forschungseinrichtungen in den letzten Jahren stark vorangetrieben worden ist, befördert Fragestellungen dieser Art und überlässt den geistes- und sozialwissenschaftlichen Disziplinen einen reichen Schatz an Daten.

Im Zuge des wachsenden Interesses an historischem Zeitungsmaterial in digitaler Form haben Bibliotheken eigene Portale mit Bild-Digitalisaten und Kalenderübersichten eröffnet (vgl. etwa die Staats- und Universitätsbibliothek Bremen [1], die Staatsbibliothek zu Berlin [2], die Universität Bonn [3] oder die Österreichische Nationalbibliothek [4]) oder stellen ihre Daten Europeana [5] zur Verfügung. Um historische Zeitungen besser durchsuchbar zu machen, wird vereinzelt bereits an der sorgfältigen Volltexterschließung besonderer historischer Zeitungen gearbeitet, vgl. etwa das Projekt „Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712-1848)" [6] (Schuster, Wille 2016: 7-29) oder das „Mannheimer Korpus für Historische Zeitungen und Zeitschriften in COSMAS II" [7]. Internationale Vernetzungsinitiativen [8] machen einerseits auf die Herausforderungen bei der Erschließung historischer Zeitungen aufmerksam (enorme Textmenge, fehlerhafte OCR-Ergebnisse, wenig Trainingsdaten) und lassen andererseits die Entwicklung gemeinsamer Strategien und Standards bei der Aufbereitung erkennen.

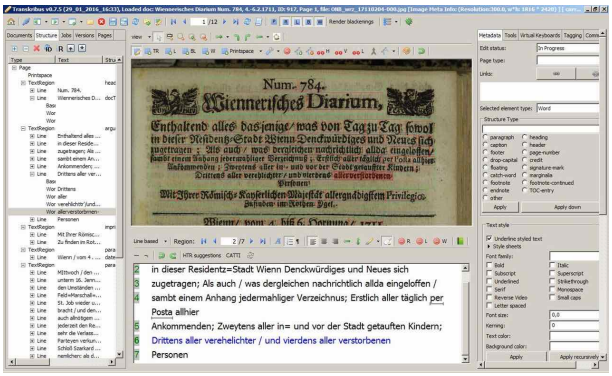
Die AutorInnen des vorliegenden Beitrags gehen davon aus, dass historische Zeitungen von Institutionen zwar „digital verfügbar" gemacht werden, aber bislang kaum an die Erkenntnisinteressen potentieller NutzerInnen angepasst sind - auch, aber nicht nur aufgrund der oben genannten Herausforderungen und des erheblichen Aufwandes, der mit der Erschließung einer historischen Zeitung verbunden ist. Bevor ein aufwändiges Projekt startet, ist daher zu überlegen, wie eine Erschließung geplant sein muss, sodass sie idealerweise für mehrere Disziplinen von Nutzen ist. Bei der Konzeption ist darauf zu achten, dass keine (für den überwiegenden Teil der antizipierten Disziplinen) relevanten Informationen vernachlässigt werden oder verloren gehen: Die Entscheidungen, die zu treffen sind, beginnen (1) bei der Auswahl der Ausgaben, setzen sich (2) bei den Transkriptionsrichtlinien fort und lassen sich (3) bis zu den Annotationskonzepten weiterführen.

Anhand eines laufenden Projektes, das sich der Volltextdigitalisierung ausgewählter Nummern des „Wien[n]erischen Diariums" [9] aus dem 18. Jahrhundert widmet, sollen all diese Aspekte kritisch hinterfragt und anhand von Beispielen, Erfahrungswerten und Zwischenergebnissen dar-

gestellt werden. Das methodische Konzept, auf dem dieses konkrete Vorhaben beruht, wurde ausgehend von der Überzeugung gestaltet, dass insbesondere bei einer vielseitig nutzbaren Ressource wie dem “Wiennerischen Diarium” die einzelnen Fachdisziplinen und NutzerInnen bereits möglichst früh in den interdisziplinären Erschließungsprozess einzubeziehen sind. Im Vortrag soll darüber berichtet werden, welche Maßnahmen bereits gesetzt wurden, um die Fachwissenschaften zu vernetzen, welche Tools im Projekt entwickelt wurden, um Personen außerhalb des Kernteams zu involvieren, und durch welche digitalen Angebote diese Kollaboration ermöglicht wird und gelingen kann.

Auswahl der Ausgaben: Um ein ausgewogenes Korpus von mehreren hundert Ausgaben verteilt über das 18. Jahrhundert zu erstellen, waren sowohl ExpertInnen mehrerer geisteswissenschaftlicher Disziplinen als auch LeserInnen der heutigen Wiener Zeitung dazu aufgefordert, jene Nummer(n) online zu nominieren, die sie als besonders relevant einstufen würden. Bei der Auswertung dieses Calls hat sich erneut bestätigt, wie breit das Themenspektrum und damit die individuellen Erkenntnisinteressen sind: Nominiert und zur Volltextdigitalisierung empfohlen wurden Ausgaben mit Geburten, Taufen und Sterbefällen bekannter Persönlichkeiten, Geburts- und Namenstage, Krönungen und Erbhuldigungen, kirchliche und weltliche Feste, Ankündigungen, Eröffnungen und Einweihungen sowie die Besuche prominenter Gäste in der Residenzstadt. Errungenschaften im weitesten Sinn – wie die Erklärung der Menschenrechte oder der Beginn der Luftfahrt – waren ebenso in der Auswahl wie das medienimmanente Thema der Herausgeberschaft. Die Ergänzungen, die das Projektteam letztlich vorgenommen hat, betrafen daher weniger die Vielfalt der angesprochenen Themen, sondern waren darauf ausgerichtet, zeitliche Lücken zu füllen. Um das Wien[n]erische Diarium als Korpus in einer kontinuierlich chronologischen Jahresabfolge plausibel dokumentieren zu können, wurden die Nominierungen dahingehend komplettiert, dass sich die Umbrüche und Wendungen in einer Periode sich verändernder politischer, sozialer, wissenschaftlicher und künstlerischer Bedingungen im 18. Jahrhundert, idealerweise auch korpusgestützt nachvollziehen lassen.

Texterstellung und -transkription: Ausgangspunkt für die Weiterverarbeitung sind jene Image-Digitalisate, welche die Österreichische Nationalbibliothek in ANNO (Austrian Newspapers Online) zur Verfügung stellt. [10] Die Erstellung des digitalen Textes für die ausgewählten Nummern erfolgt mit Transkribus [11], genauer mittels der Handwritten Text Recognition (HTR). Anhand einer kleinen Zahl an Ausgaben, die im Rahmen von projektbezogenen Vorstudien (vgl. Resch et al. 2016) bearbeitet wurden, konnte ein erstes Modell für das Diarium erstellt werden. Je nach Beschaffenheit der Digitalisate liegt die Genauigkeit hiermit zwischen 70% und 95%. Um die Qualität zu steigern, werden mehrere Tranchen von 40 bis 50 Ausgaben von zwei unterschiedlichen externen Dienstleistern nach den für dieses Projekt erstellten Transkriptionsrichtlinien erfasst, sodass das Modell weiter trainiert und die Erfassungsgenauigkeit erhöht werden kann.



Ein „Reporting-Tool“ [12], das direkt auf die Transkribus-Plattform zugreift, dokumentiert den Bearbeitungsstatus der ausgewählten Einzelnummern, gibt Auskunft über deren Umfang (Zählung von Regionen, Zeilen und Wörtern) und informiert die Fachgemeinschaft tagesaktuell über den Fortschritt des Projekts.

Erprobung von Annotationskonzepten: Parallel dazu arbeitet das Kernteam an einer digitalen Arbeits- und Annotationsoberfläche: Für die Präsentation der Texte wird nach derzeitigem Projektstand eine auf eXist basierende Umgebung genutzt, die auch in anderen Projekten des Instituts und in anderen Institutionen Anwendung findet. In die HTML-Präsentation integriert ist die Möglichkeit zur Annotation des Textes. Dabei kann entweder der Text korrigiert, eine Entität ausgezeichnet und/oder identifiziert oder eine Volltextanmerkung geschrieben werden. Die Möglichkeiten der Annotation, die Modellierung der annotierten Entitäten, ihre Verwaltung und Darstellung sollen im Rahmen eines „Annote-a-thons“ in Zusammenarbeit mit den NutzerInnen weiterentwickelt werden. Für das Projektteam ist es etwa wichtig zu erfragen, welche unterschiedlichen Sichtweisen es seitens der verschiedenen Disziplinen auf den Text gibt, welche Aspekte bei der Erschließung von besonderer Relevanz sind oder ob es so etwas wie einen “kleinsten, ge-

meinsamen Nenner" aller Annotationskonzepte geben kann. Die Einbeziehung von ExpertInnen im Annotationsprozess ist dem Projektteam ein besonderes Anliegen: Es begreift das Annotieren als höchst anspruchsvolle Forschungsleistung, die ein profundes historisch-kulturelles Wissen bei der Beurteilung erfordert und für eine Quelle wie dem „Wiennerischen Diarium" bei genauerer Betrachtung eigentlich nur gemeinsam erbracht werden kann. Ein kollaboratives Annotieren technisch vorzusehen und zu ermöglichen, ist hierbei die besondere Herausforderung. Die dafür entstehende benutzerfreundliche Präsentations- wie auch die Annotationsumgebung wird unter einer freien Lizenz zur Nachnutzung zur Verfügung gestellt werden.

Fußnoten

1. Vgl. *https://www.suub.uni-bremen.de/ueber-uns/projekte/alte-zeitungen/* [letzter Zugriff 14. Januar 2018]
2. Vgl. *http://zefys.staatsbibliothek-berlin.de/* [letzter Zugriff 14. Januar 2018]
3. Vgl. *http://digitale-sammlungen.ulb.uni-bonn.de/ulbbnz/date/list/229854* [letzter Zugriff 14. Januar 2018]
4. Vgl. *http://anno.onb.ac.at/* [letzter Zugriff 14. Januar 2018]
5. Vgl. *http://www.europeana-newspapers.eu/* [letzter Zugriff 14. Januar 2018]
6. Vgl. *https://kw.uni-paderborn.de/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistische-und-allgemeine-sprachwissenschaft/schuster/forschung/projekte/der-hamburgische-unpartheyische-correspondent-volltextdigitalisierung/* [letzter Zugriff 14. Januar 2018]
7. Vgl. *http://repos.ids-mannheim.de/mkhz-beschreibung.html* [letzter Zugriff 14.01.2018]
8. Vgl. etwa die Einrichtung einer "Special Interest Group Newspapers" bei der TEI-Konferenz 2016 in Wien, das CLARIN-Vernetzungstreffen "Working with Digital Collections of Newspapers" 2016 in Leuven oder das "Transatlantic Digitised Newspaper Symposion" 2017 in London.
9. Das „Wien[n]erische Diarium" ist am 8. August 1703 erstmals erschienen, als „Wiener Zeitung" bis heute erhältlich und somit die älteste, noch existierende Tageszeitung der Welt. Insbesondere im 18. Jahrhundert lässt sich die Entwicklung der Zeitung von den Anfängen des modernen Journalismus in einer spannenden Zeit und unter sich verändernden politischen, sozialen und künstlerischen Bedingungen gut nachverfolgen und mitvollziehen.
10. Vgl. http://anno.onb.ac.at/cgi-content/anno?aid=wrz [letzter Zugriff 14. Januar 2018]
11. Vgl. https://transkribus.eu/Transkribus/ [letzter Zugriff 14. Januar 2018]
12. Vgl. https://www.oeaw.ac.at/acdh/projects/wienerisches-diarium-digital/ [letzter Zugriff 14. Januar 2018]

Bibliographie

ANNO – AustriaN Newspapers Online. http://anno.onb.ac.at/ [letzter Zugriff 14. Januar 2018]

Harald Burger / Luginbühl, Martin (2015): Mediensprache , Berlin / Boston, 2015, 39-45.

Reisner, Andrea / Schiemer, Alfred (2016): „Das Wien(n)erische Diarium und die Entstehung der periodischen Presse" in: *Österreichische Mediengeschichte 1*: 87-112.

Digitalisierung der vollständigen deutschsprachigen Zeitungsbestände des 17. Jahrhunderts der Staats- und Universitätsbilbiothek Bremen: https://www.suub.uni-bremen.de/ueber-uns/projekte/alte-zeitungen/ [letzter Zugriff 14. Januar 2018]

Europeana Newspapers: http://www.europeana-newspapers.eu/ [letzter Zugriff 14. Januar 2018]

Mannheimer Korpus für Historische Zeitungen und Zeitschriften in COSMAS II: http://repos.ids-mannheim.de/mkhz-beschreibung.html [letzter Zugriff 14. Januar 2018]

Newspapers Sammlung der Universitäts- und Landesbibliothek Bonn: http://digitale-sammlungen.ulb.uni-bonn.de/ulbbnz/date/list/229854 [letzter Zugriff 14. Januar 2018]

Projekt „Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712-1848): https://kw.uni-paderborn.de/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistische-und-allgemeine-sprachwissenschaft/schuster/forschung/projekte/der-hamburgische-unpartheyische-correspondent-volltextdigitalisierung/ [letzter Zugriff 14. Januar 2018]

Resch, Claudia / Schopper, Daniel / Hannesschläger, Vanessa / Wohlfarter, Eva / Mader, Anna / Fischer, Nora (2016): Wienerisches Diarium Digital: Unlocking a historic newspaper for interdisciplinary studies with the TEI Guidelines. https://goo.gl/a9SmLk

Schuster, Britt-Marie / Wille, Manuel (2016): Von der Kanzlei- zur Bügersprache? Textsortengeschichtliche Betrachtungen zur „Staats- und ge-

lehrten Zeitung des Hamburgischen unpartheyischen Correspondenten“ im 18. Jahrhundert in: Jahrbuch für Kommunikationsgeschichte 17 Stuttgart: Franz Steiner Verlag 7-29.

ZEFYS Das Zeitungsinformationssystem der Staatsbibliothek zu Berlin: http://zefys.staatsbibliothek-berlin.de/ [letzter Zugriff 14. Januar 2018]

Horizontales Lesen: Das "Verdi-Requiem" und die deutsche Kritik

Roeder, Torsten
mail@torstenroeder.de
Universität Würzburg, Deutschland

1. Thema

Dieser Beitrag stellt anhand eines konkreten Fallbeispiels eine Methode zur inhaltlichen Analyse von heterogenen Textkorpora vor. Das Verfahren entstand im Rahmen einer Dissertation im Fach Musikwissenschaft und operiert einerseits mit X-Technologien, andererseits bezieht es geisteswissenschaftliche Ansätze mit ein. Der Bezug zum Tagungsthema besteht vor allem in dem Unterfangen, eine Brücke zwischen digitalem Material und hermeneutischen Analyseansätzen zu schlagen.

Das Vorhaben befasste sich mit der **Messa da Requiem** des italienischen Komponisten Giuseppe Verdi (1813–1901). Dieses geistliche Werk für Chor, großes Orchester und vier Solisten stellt eine Ausnahme in dem Schaffen des Opernkomponisten dar. Gewidmet ist es dem italienischen Schriftsteller Alessandro Manzoni, der ein Zeitgenosse und Freund Verdis war (vgl. Schweikert 2013) und 1873 verstarb. Das deshalb so genannte "Manzoni-Requiem" verbreitete sich schnell und mit großem Erfolg in ganz Europa, vor allem in Frankreich, Österreich und (kurze Zeit später) auch im Deutschen Reich.

Der Fokus der Analyse liegt auf der deutschsprachigen Musikkritik im Zeitraum der Erstaufführungen (1874–1878). Mit Hinweis auf den oft emotionalen und bildhaften Charakter der Musik wurde das Werk vor allem im Deutschen Reich abschätzig beurteilt: Ein Kritiker aus Köln meinte beispielsweise, das Werk sei "kein Requiem nach deutscher Art" (August Guckeisen, Kölnische Zeitung, 12.12.1875). Diese dichotomische Abgrenzung einer "deutschen Art" gegen eine "italienische Art" ist nicht nur als Reflex von Wagnerismus und reichsdeutschem Kulturkampf zu interpretieren, sondern auch im Kontext romantischer Kirchenmusikästhetik und des zeitgenössischen Realismusbegriffs zu verstehen (vgl. Kirsch 2013). Darüber hinaus regten sich aber gerade auch in Österreich solche Stimmen, die dem geforderten "deutschen Ernst" nur wenig abgewinnen konnten und für mehr Emotionalität in der geistlichen Musik plädierten. Auch gegenwärtig führt die Ambivalenz des Werkes, das zwischen spiritueller und szenischer Musik changiert, immer wieder zu Diskussionen hinsichtlich der "richtigen" Interpretation. Das Nachdenken über historische Auffassungen und Begriffsbildungen gibt uns die Möglichkeit, unsere aktuellen Positionen und deren Ursprünge zu hinterfragen. Dies bildet den Ausgangspunkt der Untersuchung.

2. Material

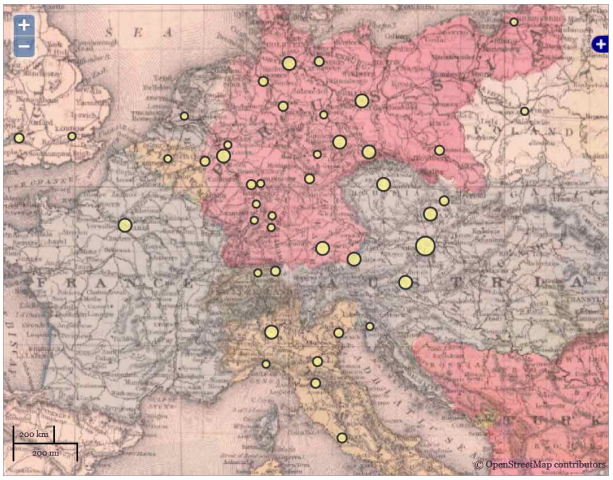
Für die Grundlage der Analyse wurden Artikel aus Tageszeitungen und Musikfachblättern aus den Jahren 1874–1878 zu einem Textkorpus zusammengestellt. Auswahlkriterien waren erstens eine Erwähnung der *Messa da Requiem* und zweitens die Sprache Deutsch. Durch diese Offenheit entstand ein Korpus aus Texten, die hinsichtlich des Umfangs, der Gattung und der inhaltlichen Qualität stark variieren: Die Textsorten decken ein Spektrum von lapidaren Konzertanzeigen bis hin zu seitenlangen Werkbesprechungen mit zahlreichen Notenbeispielen ab. Für die Forschungsarbeit erfüllen die Texte unterschiedliche Zwecke: Die knapperen Texte beschränken sich meist auf Hinweise zu Aufführungen, Mitwirkenden oder Veranstaltungsorten und dienten deshalb vorrangig der Rekonstruktion einer Aufführungsgeschichte. Dabei gaben präzise Datumsangaben oft einen Anhaltspunkt für die Recherche nach ausführlicheren Quellen in lokalen Tageszeitungen. Für die Untersuchung der Rezeptionsgeschichte sind die längeren Texte weitaus interessanter und ergiebiger. Dabei bestand eine Herausforderung darin, dass sehr häufig (in ca. 70% der Fälle) nichts über den Hintergrund des Autors bekannt ist, so dass der Entstehungskontext bei der Analyse nicht einbezogen werden konnte. Bisherige Analysen berücksichtigten deshalb vor allem namhafte Autoren, deren (wie beispielsweise Eduard Hanslick, Heinrich Adolf Köstlin und Emil Naumann; vgl. z. B. Kreuzer 2005). Digitale Auswertungsverfahren sollten nun dabei helfen, auch die Äußerungen unbekannter Autoren in einem breiteren Spektrum einordnen und bewerten zu können.

Die Texte wurden dazu nach Digitalisaten oder Fotokopien (20% der Fälle) transkribiert und die daraus resultierenden 950.000 Zeichen Reintext mit TEI-Markup versehen. Dabei wurde der Schwerpunkt, im Sinne von "smart data" (vgl. Schöch 2013), auf die Erschließung semantischer Einheiten gelegt. Dazu wurde ein halbautomatisches Verfahren eingesetzt, das mit Listen von bereits bekannten Eigennamen operierte. Das automatische Tagging wurde auf Richtigkeit überprüft und Sonderfälle (z.B. Pseudonyme oder Abkürzungen) manuell verzeichnet. Im Korpus konnten insgesamt 8.000 Entitäten identifiziert und mit Normdaten versehen werden. Die vier Register beinhalten 530 Personen, 142 Aufführungen, 96 Ortschaften und 135 Werke der Musik.

3. Auswertung

3.1. Metadaten

Die Analyse der Metadaten mithilfe geographischer und einfacher statistischer Visualisierungstools (z. B. OpenLayers/MapWarper und Google Chart API) gab Aufschluss über die geographische und chronologische Verteilung der Texte. Die Presseresonanz ist in den Monaten der Erstaufführungen an den jeweiligen Orten wie erwartet am stärksten. Dabei fiel auf, dass die Verteilung im Deutschen Reich großflächiger ist als in dem stark auf Wien zentrierten Österreich (vgl. Abbildung 1). Die Anzahl der Berichte an einem Ort ist in Österreich durchschnittlich höher als im Deutschen Reich, da dort durch das Digitalisierungsprojekt ANNO erheblich mehr Periodika digitalisiert wurden und relevante Texte durch Suche im OCR-Volltext viel leichter aufzufinden sind als in realen Zeitungsbänden. Im Deutschen Reich überwiegt deshalb auch der Anteil an Texten aus den bereits umfassend digitalisierten Musikfachzeitschriften (vor allem aus Leipzig) um so mehr.

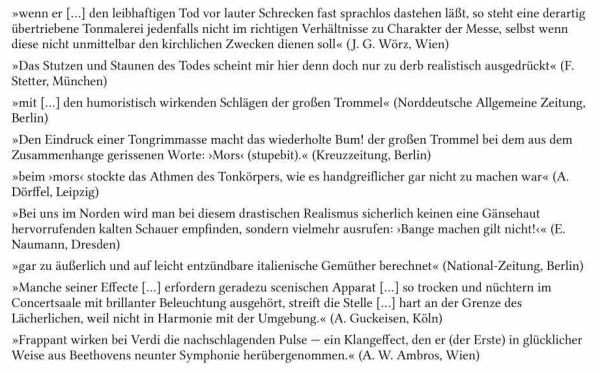


(Abbildung 1: Übersicht aller Spielorte, die im deutschsprachigen Raum erwähnt wurden)

3.2. Textanalyse

Die Textanalyse knüpfte zunächst an mehrere Vorarbeiten zur Dichotomieanalyse und Rezeptionsgeschichte an (vgl. Sponheuer 2001) und entwickelte diese anhand des digitalen Materials weiter. Es wurde das Experiment unternommen, mithilfe eines vorgefertigten Sachvokabulars die relevanten Diskurskategorien zu erschließen und dichotomische Muster zu identifizieren. Dies musste im Rahmen des damaligen Vorhabens verworfen werden, da es in vielen Einzelfällen zu einer deutlichen Verzerrung der Textbedeutung kam, die durch den relativ geringen Umfang des Textkorpus leider nicht ausgeglichen werden konnte.

Deshalb wurde der Fokus in einem zweiten Versuch auf die "Named Entities" gesetzt, also die bereits durch das Tagging identifizierten sachlichen Einheiten. Eine solche Einheit konnte beispielsweise ein Werkteil wie das "Kyrie" oder das "Agnus Dei" sein, zu dem dann aus allen Texten die unmittelbaren Kontexte ausgezogen und zu einem Spektrum zusammengestellt wurden. Die Kontexte zur stark umstrittenen Nummer "Mors stupebit" aus dem zweiten Teil des Werkes zeigen sowohl eine grundsätzliche Ablehnung und eine starke Grenzziehung zwischen "nordischen" und "italienischen" Stereotypen auf.



(Abbildung 2: Kontexte für "Mors stupebit")

Aus geisteswissenschaftlicher Perspektive genügt bereits diese Aufstellung, die ich hier "semantische Sichtachse" nennen möchte, um ein Spektrum von unterschiedlichen Positionen aufzuzeigen, in das auch die Haltungen der nicht näher bekannten Autoren mit einfließen. Vergleicht man die hier gezeigte Sichtachse zum "Mors stupebit" mit der Sichtachse des fast überall bewunderten "Agnus Dei", kommt ein Kontrast zwischen einer idealisierenden und einer realistischen Repräsentation des Todes deutlich zum Vorschein. Weitere aufschlussreiche Sichtachsen können beispielsweise zu Vergleichswerken (etwa die Requiem-Kompositionen von Mozart oder Cherubini), Komponistennamen (im Sinne einer stilistischen Referenz) und auch zu Schlagwörtern wie z. B. "Realismus" erzeugt werden.

Der Zweck dieses Verfahrens ist es also, mithilfe der Textdatenbank vergleichbare Kontexte innerhalb des Korpus zu finden und diese gegenüberzustellen. Die komparative Analyse der Einzelkontexte eröffnet die Möglichkeit, übergreifende Rezeptionsmuster zu identifizieren. Ich bezeichne das hier rein metaphorisch als "horizontales Lesen" (in vager Anlehnung an den Begriff des "distant reading", vgl. Moretti 2016), da es die Texte gleichwertig nebeneinander anordnet und weniger den einzelnen Text, sondern mehr gemeinsame Bezugspunkte im gesamten Korpus betrachtet. Auf der Grundlage dieser vergleichbaren Extrakte könnte übrigens der erste, zunächst verworfene Ansatz wieder ins Spiel kommen – dies ist jedoch Zukunftsmusik.

Ich füge noch einige Ergebnisse aus musikwissenschaftlicher Sicht an: Beispielsweise wurde das sehr beliebte "Agnus Dei", welches in über 60 Texten erwähnt wurde, auffallend oft als Abklatsch verschiedener Opernnummern gedeutet, um die schöpferische Qualität herabzusetzen. Hinsichtlich des Fugenstils im "Sanctus" verhielten sich die Berliner und Dresdner Kritiker auffallend empfindlich, während die Kritiker in Wien und auch in Köln in dieser Hinsicht liberaler waren. Eine ähnliche Aufteilung ist auch hinsichtlich des Einsatzes von musikalischen Effekten und Affekten festzustellen. Dabei wurden sowohl allgemeine Unterschiede zwischen dem Deutschen Reich und Österreich sowie dem katholischen und dem protestantischen Raum deutlich. Hinzu treten mehrere lokale Besonderheiten: So kam am Hamburger Stadt-Theater ein Bühnenbild zum Einsatz, welches das Innere einer katholischen Kirche zeigte, was die Hamburger Kritik durchaus faszinierte. In Salzburg schlossen sich mehrere Musikvereine zusammen, um das Werk für einen karitativen Zweck aufzuführen, weshalb die Kritik wohlwollender war als andernorts.

Die insgesamt sehr unterschiedlichen Kritiken können deshalb erst aus den jeweiligen Kontexten heraus angemessen bewertet werden. Durch einen flächendeckenden Vergleich ist es möglich, allgemeine Tendenzen der Rezeption zu identifizieren und zu verorten, auch wenn die individuellen Hintergründe der Autoren nicht immer bekannt sind. Die Texte von unbekannten Autoren sind auf diese Weise leichter zu kontextualisieren, zu bewerten und in das Gesamtbild einzuordnen.

4. Ausblick

Das Textmaterial ist für das ursprüngliche musikwissenschaftliche Vorhaben zunächst ausgeschöpft, bietet aber Potenzial für weitere Analysen mit anderen Methoden. Wärenbeispielsweise positive und negative Kritiken mithilfe stilometrischer Verfahren voneinander unterscheidbar? Geben Netzwerkvisualisierungen von Vergleichswerken in den Texten möglicherweise einen Hinweis auf Strukturen kanonischer Bezugssysteme? Deckt sich dies mit den bisherigen Beobachtungen? Wäre es für die Präsentation denkbar, eine in MEI kodierte Partitur takt- oder abschnittsweise mit den diversen Kommentaren der Kritiker zu versehen? Es ist zu hoffen, dass in Zukunft ähnliche Projekte entstehen, um das digital aufbereitete Datenmaterial für weitere Forschung nutzen können – etwa mit dem Fokus auf der Rezeption des Werkes in anderen Sprachgebieten oder Zeiträumen, oder auch auf anderen Requiem-Vertonungen. Zu diesem Zweck wird das Material digital publiziert, offen lizenziert sowie in Hackathons und THATcamps eingebracht. Die Publikationsplattform wird zum Konferenztermin bekanntgegeben.

Bibliographie

Kirsch, Winfried (2013): "Kirchenmusikreform, Cäcilianismus und Palestrina-Renaissance", in: Wolfgang Hochstein und Christoph Krummacher (eds.): Geschichte der Kirchenmusik, Bd. 3, Laaber: Laaber 56–71.

Kreuzer, Gundula (2005): "Oper im Kirchengewande"? Verdi's Requiem and the Anxieties of the Young German Empire, in: Journal of the American Musicological Society 58,2: 399–450.

Moretti, Franco (2016): Distant Reading, Konstanz: Konstanz University Press.

Schöch, Christof (2013): "Big? Smart? Clean? Messy? Data in the Humanities", in: Journal of Digital Humanities 2, No. 3: 2–13.

Schweikert, Uwe (2013): Messa da Requiem, in: Anselm Gerhard and Uwe Schweikert (eds.): Verdi-Handbuch, Kassel: Bärenreiter; Stuttgart/Weimar: Metzler (2., überarb. und erw. Auflage) 557–565.

Sponheuer, Bernd (2001): "Über das ›Deutsche‹ in der Musik. Versuch einer idealtypischen Rekonstruktion", in: Hermann Danuser / Herfried Münkler (eds.): Deutsche Meister – böse Geister? Nationale Selbstfindung in der Musik, Schliengen: Edition Argus 123–150.

Im Netz der Möglichkeiten - Wechselwirkungen in der Entwicklung von Theorie, Methode und Tools in den Digital Humanities am Beispiel der TEI

Schaßan, Torsten
schassan@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

In der Beschreibung der DHd-Konferenz 2018 „Kritik der digitalen Vernunft“ wird die These formuliert, dass die Digital Humanities„häufig als digital transformierte Bearbeitung von Fragestellungen aus den verschiedenen beteiligten Fächern beschrieben“ werden und „in weiten Teilen eine daten-, algorithmen- und werkzeuggetriebene Wissenschaft sei[en], die von ihren unmittelbaren Möglichkeiten und ihren Praktiken dominiert“ werden, welche den Prinzipien kritischer Wissenschaftlichkeit möglicherweise nicht genüge. Am Beispiel der Entwicklung der Guidelines sowie der Analyse einzelner konkreter Anwendungsfälle der Text Encoding Initiative (TEI; Synonym auch für die Richtlinien der Initiative, den „Guidelines for (Electronic) Text Encoding and Interchange“) soll untersucht werden, welche Wechselwirkungen es hierbei zwischen Theorie, Methode und Tool-Entwicklung gegeben hat.

Der Gegenstand

Der Beobachtungsgegenstand dieser Untersuchung sind die Guidelines der TEI selbst, die daraus resultierenden Schemata sowie deren Anwendung in ausgewählten Anwendungsfeldern der DH. Die TEI dokumentiert die Entwicklung der Guidelines seit der Einführung 1990 mit der Version P1 in unterschiedlicher Tiefe. Während für die älteren Versionen meist nur deren Endprodukt in Form der Document Type Definition (DTD) bzw. des Textes der Guidelines vorliegt, kann man die Entwicklung des de-facto-Standards seit der Version P4 und dann ab 2007 in listenförmiger Dokumentation der vorgenommenen Änderungen nachvollziehen. Die TEI dokumentiert außerdem die Diskussionen, die in ihren Gremien sowie in der Community über Mailingliste geführt werden, in je eigenen Archiven. Damit lässt sich ein nahezu lückenloses Bild der Entwicklungsschritte hin zu der geltenden Version nachvollziehen. Diese Materialien können vor dem Hintergrund der oben genannten Thesen untersucht und mit den traditionellen Wissenschaften in Kontext gesetzt werden, um Aufschlüsse über die Wechselwirkungen zu erhalten.

Ein weiterer Gegenstand besteht in der konkreten Anwendung der Guidelines bzw. Schemata der TEI in Projekten. Hier sollen beispielhaft Editionsprojekte sowie die Verwendung der TEI zur Speicherung von Metadaten betrachtet werden, um die Wechselwirkungen zu analysieren. Es soll damit der Frage nachgegangen werden, inwieweit die Definition einer Markup-Sprache, die Anwendung der Sprache bei der Erstellung konkreter Dokumente und die Validierung der Ergebnisse durch technische Rahmenbedingungen vorgegeben oder durch außerhalb der DH liegenden Theoriebildung beeinflusst werden oder die Theoriebildung und Methodenentwicklung selbst vorantreiben.

Theoriebildung

Bereits 1994 wies Sperberg-McQueen auf die theoretischen Implikationen der Textauszeichnung, analog dazu: des Gebrauchs einer Wissenschaftssprache, hin: „Like any notation, the TEI Guidelines inevitably make it easy to express certain kinds of ideas, and concomitantly harder to express other kinds of ideas, about the texts we encode in electronic form. Any notation carries with it the danger that it must favor certain habits of thought --- in the TEI's case, certain approaches to text --- at the expense of others. No one should use TEI markup without being aware of this danger --- any more than we should use the English language, or any other, without realizing that it favors the expression of certain kinds of ideas, and discourages the expression, and even the conception, of other ideas.“ (Sperberg-McQueen 1994) Die TEI-Community versucht allerdings dennoch, diesen Gefahren mit der offenen Diskussion über die semantischen Dimensionen des Markups, mit Best-Practice-Beispielen für die Anwendung entgegen zu wirken. Die Community beweist immer wieder die Fähigkeit, den etablierten Standard im Fall erweiterter Anforderungen an neue Aufgaben anzupassen. Paradigmenwechsel in der Editionsphilologie können dadurch adaptiert, in der Auszeichnungspraxis angewendet und somit wissenschaftlich nutzbar gemacht werden, ohne deshalb vorhandene Dokumente ungültig werden zu lassen oder die bisherige Anwendung des Standards zu konterkarieren. Beispiele hierfür sind die Inkorporation von Markup-Subsystemen für genetische Editionen, zur Beschreibung von Handschriften oder auch zur Dokumentation von Kommunikationsvorgängen, etwa in Briefen, mit der Einführung von <correspDesc>. Das System der TEI wurde dabei prinzipiell erweitert als grundlegend verändert. Die Abschaffung von Elementen oder Attributen, die etwa die Falsifikation nach Popper gleich zu setzen wären, ist daher sehr viel seltener als der Proof of Concept, mit welchem zunächst die Tauglichkeit von Markup zur Repräsentation bestimmter Phänomene getestet wird, bevor die Aufnahme in den Standard erfolgt und die Angebote durch die Community weiter genutzt werden können.

Gleichzeitig bestimmen die Anforderungen der inhaltlichen bzw. theoretischen Weiterentwicklung eines Faches die technische und theoretische Weiterentwicklung der TEI mit. Ein rezentes Beispiel hierfür ist die theoretische Ausrichtung des *Material Turns* in der Editionsphilologie und den historischen Wissenschaften. Die Hinwendung zum Objekt wird in der TEI mit der Einführung von <facsimile> und dem Modul für genetische Editionen gespiegelt. Auch Überlegungen zur Verallgemeinerung des Beschreibungsstandards von Dokumenten und Objekten mittels der Strukturen von <msDesc> wären hier zu nennen.

Methodenbildung

Ein Beispiel für die Anwendung neuer Methoden der DH ist die Anreicherung von Texten um bestimmte Kontexte. Wenn in einem Text benannte Entitäten wie Personen, Orte, Objekte oder Ereignisse ausgezeichnet werden, dann entspricht dies zunächst der Registerarbeit traditioneller Publikationsverfahren. Wenn diese Entitäten allerdings mit Normdaten angereichert und diese somit zu nachnutzbaren Bestandteilen im Sinne der Linked Open Data (LOD) werden, darf man wohl von der Anwendung einer neuen Methode sprechen. Die entstandenen Dokumente verändern ihren Charakter von traditionell erstellten und händisch ausgewerteten Objekten zu konzeptionell gestalteten und automatisiert nutzbaren. Die Dokumente werden in die Lage versetzt, auch außerhalb bestimmter, inhaltlich vorgegebener Auswertungskontexte mit anderen Methoden ausgewertet zu werden.

Die Konformität zu den TEI-Richtlinien spielt in der Anwendung der neuen Methode eine wesentliche Rolle. Im System der TEI ist die *Customisation* als Normalfall vorgesehen. Unter *Customisation* wird die kontextabhängige Auswahl von Modulen, Elementen und Attributen oder die Vorgabe von Wertemengen für Attribute verstanden. Diese Vorgaben können in der TEI in sogenannten „One Document Does it all“-Dateien (ODD) definiert und dokumentiert werden, um dann anschließend daraus eigene Datenstrukturdefinitionen in Form von Schema-Dokumenten zu generieren. Zunehmend werden hierbei neben den klassischen Schemasprachen zur Validierung von Datenstrukturen (DTD, RelaxNG, XML Schema) auch Regelwerke definiert, welche eine Validierung des Inhalts eines Dokumentes ermöglichen. (Schematron) ODD sind durch die Dokumentationsleistung und eine normierte Beschreibungssprache ein starkes Hilfsmittel, um die Verwissenschaftlichung von textuellem Markup und dem unterliegenden Textverständnis zu erreichen. Wie aber der Grad der TEI-Konformität zu messen und zu beschreiben sei, ist in der TEI bereits seit Längerem Gegenstand von Diskussionen.

Tool-Entwicklung

Wo nun mehr Texte in TEI-konformer Auszeichnung vorliegen, umso wichtiger wird die Existenz von Tools, um diese Daten zu nutzen. Wo Text- und Image-Digitalisierung Quellen einfacher zugänglich macht, werden diese Materialien potentiell als Gegenstand der Digital Humanities nützlich. Ähnlich wie bei der Theorie- oder Methodenbildung befruchten sich zur Verfügung stehende Materialien und daran anzulegende Fragestellungen gegenseitig. Die Kritik der digitalen Methoden wird hier ansetzen müssen, wo das Markup vom Ende her zu denken ist, weil bestimmte Funktionalitäten gewünscht sind: Dokumente sollen nach bestimmten Kriterien sortiert werden? Was bedeutet dies für die Aufbereitung der Daten? – Dokumente sollen im Rahmen des LOD genutzt werden? Welche Eigenschaften müssen sie hierfür haben?

DH, eine Wissenschaft wie jede andere?

Die in einer Wissenschaft zur Verfügung stehenden Quellen und deren Verarbeitungsmechanismen haben sich immer schon gegenseitig sowie die Theorie- und Methodenbildung beeinflusst. An den Beispielen der Untersuchung soll gezeigt werden, dass sich dies in den DH nicht anders ausnimmt.

Bibliographie

Lou Burnard / **Sebastian Rahtz** (2004): "RelaxNG with Son of ODD", in: *Proceedings of Extreme Markup Languages 2004*. http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html

Lou Burnard / **C. Michael Sperberg-McQueen** (1995): "The Design of the TEI Encoding Scheme", in: *Computers and the Humanities* 29 (1) p. 17–39. 10.1007/BF01830314

Guidelines for Electronic Text Encoding and Interchange, edited by TEI Consortium. 1990-, P1--P5. http://www.tei-c.org/Guidelines/ , http://www.tei-c.org/Vault/

Patrick Sahle (2013): *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* 3 Bde. Norderstedt: BoD. https://www.i-d-e.de/publikationen/schriften/s7-9-digitale-editionsformen/

C. Michael Sperberg-McQueen (1994): *Textual Criticism and the Text Encoding Initiative* . (First draft of a paper presented at MLA, San Diego, 1994) http://www.tei-c.org/Vault/XX/mla94.html

C. Michael Sperberg-McQueen / **Henry Thompson** (2000): *XML-Schema.* https://www.w3.org/XML/Schema

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

RelaxNG (2008): *ISO/IEC 19757-2:2008. Information technology -- Document Schema Definition Language (DSDL) -- Part 2: Regular-grammar-based validation -- RELAX NG.* http://standards.iso.org/ittf/PubliclyAvailableStandards/c052348_ISO_IEC_19757-2_2008(E).zip

Schematron (2016): *ISO/IEC 19757-3:2016. Information technology -- Document Schema Definition Languages (DSDL) -- Part 3: Rule-based validation – Schematron.* http://standards.iso.org/ittf/PubliclyAvailableStandards/c055982_ISO_IEC_19757-3_2016.zip

Interpretation und Unschärfe bei der semantischen Erschließung von historischen Quellen

Hamisch, Juliane
j.hamisch@gnm.de
Germanisches Nationalmuseum, Nürnberg

Große, Peggy
p.grosse@gnm.de
Germanisches Nationalmuseum, Nürnberg

Einleitung

Um im Rahmen eines Forschungsprojekts die anfallenden Daten in einer virtuellen Forschungsumgebung semantisch zu erschließen, wird unter Berücksichtigung der Fragestellung sowie der gewählten Bearbeitungsmethode für die formalisierte Erfassung und Darstellung eine Modellierung erarbeitet. Bei der Erschließung historischer

Quellen werden Bearbeiter häufig mit inhaltlichen Unschärfen konfrontiert, die in der Regel eine Interpretation der Quellen in Hinblick auf die eigene Fragestellung unter Einbeziehung schon vorhandener Forschungsergebnisse erfordern.

Es stellt sich folglich das Problem, wie die gewonnenen Forschungsdaten einerseits semantisch erschlossen und formalisiert werden können, um eine möglichst große Vergleichbarkeit der Daten zu gewährleisten, andererseits aber der Vorgang der Interpretation so transparent gehalten werden kann, dass er schließlich auch für Dritte nachvollziehbar bleibt. Das geschilderte Problem beginnt mit der Konzeption und Modellierung der Forschungsumgebung und reicht über die Erfassung durch Projektmitarbeiter bis zur Rezeption und gegebenenfalls Weiterverarbeitung der Daten durch Dritte.

Im Folgenden soll dieses Problem anhand zweier Projekte erläutert werden, die mit der virtuellen Forschungsumgebung ‚Wissenschaftliche KommunikationsInfrastruktur' (WissKI) [1] arbeiten. WissKI wurde speziell für die Forschung im Bereich des kulturellen Erbes entwickelt und arbeitet mit Linked Open Data sowie Semantic Web-Technologien. Als ‚Semantisches Backend' und somit Referenzontologie wird das CIDOC Conceptual Reference Model (CIDOC CRM, ISO21127) [2] des International Council of Museums genutzt (vgl. Doerr/Lampe/Krause 2011, Görz 2011, Hohmann 2011, Hohmann/Schiemann 2013, Hohmann/Fichtner 2015).

Erste Fallstudie: Digitale Edition der Posse-Tagebücher

Das vom Deutschen Zentrum für Kulturgutverluste geförderte Projekt "Kommentierte Online-Edition der fünf Reisetagebücher Hans Posses (1939-1942)" wertet die Reisetagebücher des Kunsthistorikers Hans Posse aus. In seiner Funktion als Sonderbeauftragter Adolf Hitlers war Posse für den Aufbau einer Sammlung für das ‚Führermuseum Linz' sowie die Vorbereitung und Umsetzung eines Verteilungsprogramms von NS-Raubkunst auf die Museen im Deutschen Reich zuständig. Da Posse seine Aufzeichnungen meistens vor Ort vornahm, handelt es sich um schwer lesbare Kurznotizen, Listen und Aufzählungen, nur selten um Fließtext.

Ausgehend von den in den Tagebüchern beschriebenen historischen Ereignissen wird es sich bei der Erarbeitung und Rezeption der Edition um das Produkt diverser Interpretationsvorgänge handeln (vgl. Sahle 2013a: 208). Die erste Interpretationsebene sind dabei die Aufzeichnungen von Posse selbst. Da die Tagebücher nie zur Veröffentlichung gedacht waren, sondern ihm lediglich als Gedächtnisstützen dienten, kann davon ausgegangen werden, dass es sich bei Posse im Rahmen der ihm zur Verfügung stehenden Informationen um einen zuverlässigen Erzähler handelt.

Die zweite Interpretationsebene erfolgt durch die Modellierung der Daten, die der Erfassung der vorgefundenen Inhalte und deren Auswertung eine Struktur verleiht. Die dritte Interpretationsebene ist die der inhaltlichen Bearbeitung des historischen Materials. So nutzte Posse in seinen Aufzeichnungen ungebräuchliche Kürzel, z.B. Rbdt. für Rembrandt oder Hbst. für den Kunsthändler Karl Haberstock. Er schrieb Namen nach Gehör und vergab für Kunstwerke, die er sich vor Ort ansah, selbst erfundene Titel. Zudem sind die Aufzeichnungen, bedingt durch ihre Funktion als Gedächtnisstütze, lückenhaft. Es fehlen beispielsweise vollständige Namensangaben oder Angaben zu Kunstwerken wie der Künstler, die Datierung etc. Durch das Hinzuziehen von Forschungsliteratur lassen sich viele der fehlenden Angaben rekonstruieren. Dennoch sind die Zuschreibungen durch die Bearbeiter mit kleineren oder größeren Unsicherheiten behaftet. Das konkrete Problem ist also folgendes: Einerseits sollen durch die Annotierung und Erfassung von Personen, Körperschaften, Orten und Werken möglichst viele Informationen formalisiert und damit vergleichbar und nachnutzbar werden. Andererseits sind nicht alle der vorgenommenen Zuschreibungen gleich wahrscheinlich. Auch die Unschärfen und Auslegung des Textes müssen also (semantisch) modelliert und für die späteren Nutzer nachvollziehbar sein, um ihnen eine angemessene Einordnung – auf der vierten Interpretationsebene - der vorgefundenen Informationen zu ermöglichen (vgl. Sahle 2013b: 184).

Konkret wird dieses Problem in der Modellierung mit zwei Lösungswegen angegangen. Vor allem bei den von Posse begutachteten Kunstwerken ist aufgrund der lückenhaften Informationen eine eindeutige Zuschreibung oft unmöglich. Auf semantischer Ebene dient deswegen ein Zuschreibungsereignis (attribute assignment) als Basis für die Erfassung eines Kunstwerks, das mit allen aus dem Text hervorgehenden Kontextinformationen verknüpft wird. Zusätzliche Informationen zu identifizierten Werken können bei Bedarf ergänzend hinzugefügt werden.

In der Transkription werden unsichere Zuschreibungen nicht direkt im Text vorgenommen, sondern mit zusätzlicher Erläuterung in den Fußnoten und dort annotiert. Um die Zuweisung auch im Rückschluss, ausgehend von den Entitäten wie Personen, Institutionen etc., nachvollziehbar zu machen, wird ein Zuschreibungsereignis zwi-

schengeschaltet, das entweder auf den Tagebuchtext oder die Fußnoten verweist.

Diese Maßnahmen führen allerdings dazu, dass die semantische Modellierung der Daten sehr komplex wird und eine sehr gute Kenntnis der verwendeten Ontologie, in diesem Fall des CIDOC CRM voraussetzt. Ohne ausführliche erläuternde Texte zur Vorgehensweise wird die Bewertung und Einordnung der Information trotz einer Berücksichtigung in der Darstellung kaum möglich sein.

Zweite Fallstudie: Repräsentationen des Friedens in der Vormoderne

Das von der Leibniz-Gemeinschaft seit Juli 2015 geförderte internationale Kooperationsprojekt „Repräsentationen des Friedens im vormodernen Europa“ erforscht Friedensbilder im Zeitraum vom 16. bis 18. Jahrhundert. Friedensvereinbarungen mussten über den reinen Vertragstext hinaus erklärt, begründet und vermittelt werden. Das übernahmen Friedensrepräsentationen, die ein multimediales Phänomen der Frühen Neuzeit waren. Folglich beschäftigt sich das Forschungsprojekt mit visuellen Darstellungen, sprachlichen Bilder sowie musikalischen Ausprägungsformen.

Um abstrakte Konzepte wie Frieden, Gerechtigkeit oder Wohlstand darzustellen, verwendeten Künstler, Dichter oder Komponisten einen Kanon von Motiven, die europaweit genutzt und verstanden wurde. Dieses ‚Vokabular‘ des Friedens soll beispielhaft erschlossen und über Gattungs- und Genregrenzen hinweg analysiert werden. Zudem wurden gemeinsame Fragestellungen zu transmedialen Rezeptionsvorgängen, Veränderungen der Motivik im Zusammenhang mit unterschiedlichen Friedensschlüssen entwickelt. Durch die Verwendung von WissKI ist ein transdisziplinäres und ortsunabhängiges Arbeiten an den Quellen möglich. Es erlaubt eine formular- und textbasierte Erfassung der Bestände, die durch Reproduktionen der graphischen und numismatischen Objekte sowie Verlinkungen auf Textquellen ergänzt wird. Die Erfassung der Themen, Motive und Friedensereignisse erfolgt im Objektformular, um eine möglichst hohe Auffindbarkeit zu garantieren.

Sowohl die Historizität der Quellen, als auch die Auslegung bzw. Deutung von visuellen und sprachlichen Symbolen durch die Projektmitarbeiter ist mit Unschärfen behaftet, die nur eingeschränkt in einer formularbasierten Erfassung in einer Datenbank für Dritte sichtbar gemacht werden können. Anhand von Bildquellen soll diese Problematik im Folgenden kurz erläutert werden.

Die Deutung von Bildern erfolgt in mehreren Interpretationsebenen: Zunächst muss sich der Betrachter versichern, was er sieht (Büttner 2014: 12). Doch ist diese erste vordergründig neutrale Auswertung einer Quelle nicht objektiv, sondern kann die individuelle Betrachtung von kulturellen Mustern und subjektiven Erfahrungen geprägt sein (Büttner 2014: 17). Die zweite Interpretationsebene ist die Identifizierung des Themas bzw. des Motivs und des Entstehungskontextes. Hier gilt es die vom Künstler oder Auftraggeber intendierte Funktion und Bedeutung visueller Symbole einzubeziehen bzw. dieser Absicht möglichst nahe zu kommen. Um sich der Intention anzunähern, ist der Entstehungskontext eines Werkes wichtig, der ausgehend von den Forschungsfragen vor allem das entsprechende Friedensereignis betrifft, für das ein Werk geschaffen wurde. Das kann mal explizit (durch Nennung im Text oder Bildbeischriften, Datierung) und mal weniger eindeutig sein. Anders als in einem beschreibenden Text kann der Weg, der zum Eintrag eines Themas oder eines Friedensereignisses im Formular führt, nicht ohne viel Aufwand abgebildet werden, sodass die eingetragene Information immer ein Endprodukt eines Interpretationsvorganges darstellt.

Auf der Ebene der semantischen Datenmodellierung wurde sowohl bei der Zuweisung von Motiven als auch bei Ereignissen durch die Einbindung eines Zuschreibungsereignisses (attribute assignment) der Vorgang der Interpretation ausgedrückt. Um eine Einheitlichkeit in der Benennung von Motiven und Themen zu erzielen wurde für die standardisierte Erfassung das weitverbreitete Klassifizierungssystem Iconclass hinterlegt. [3]

Für die Nachnutzung der Informationen durch Dritte muss erläutert werden, nach welchen Kriterien die Auslegung der Quellen erfolgte und wie dabei vorgegangen wurde. Für die weitere Benutzung zwar standardisierter, aber mit Unschärfen behafteter Daten, braucht es deshalb ein hohes Maß an Reflexion und Kenntnis der in der Erfassung involvierten Fachdisziplinen, um Informationen auswerten und einordnen zu können.

Fazit

Als Fazit lässt sich festhalten, dass mit der semantischen Modellierung auf Basis des CIDOC CRM eine Basis geschaffen wurde, auf struktureller Ebene Daten zu historischen Quellen vergleichbar zu erfassen.

Trotz der semantischen Datenmodellierung auf Basis einer standardisierten Ontologie und der dadurch gegebenen Vorstrukturierung der Inhalte unterliegt jedes System jedoch immer der In-

terpretation eines Forschungsgegenstandes. Um diesen Vorgang dem Nutzer zu vermitteln, bedarf es im Allgemeinen ausführlicher Erläuterungen in Form von Rahmentexten oder Fußnoten.

Auch wenn sich durch die semantische Modellierung auf Basis einer standardisierten Ontologie ein in sich logisches System auf Projektebene aufbauen lässt, so stellt sich doch die Frage, inwiefern verschiedene Forschungsumgebungen - trotz ähnlicher Forschungsgegenstände - noch vergleichbar im Sinne des Semantic Web-Gedanken bzw. des Linked Open Data sind. Eine inhaltliche Vergleichbarkeit ist nur bei grundlegenden Konzepten wie Personen, Orten oder Körperschaften gegeben. Für die Bewertung der Inhalte bleibt nach aktuellem Stand die genaue Kenntnis des Kontexts eines Forschungsprojekts unabdingbar.

Fußnoten

1. http://wiss-ki.eu/.
2. http://www.cidoc-crm.org/.
3. http://www.iconclass.nl/about-iconclass/what-is-iconclass.

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.

Görz, Günther (2011): „WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung" in: *Kunstgeschichte, Open Peer Reviewed Journal*, urn:nbn:de:bvb:355-kuge-167-7 [letzter Zugriff 10.01.2018].

Hohmann, Georg (2011): „Die Anwendung von Ontologien zur Wissensrepräsentation und -kommunikation im Bereich des Kulturellen Erbes" in: Schomburg, Silke u.a. (eds.): *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland.* Köln: Hochschulbibliothekszentrum NRW 33-39.

Hohmann, Georg / Schiemann, Bernhard (2013): „An Ontology-Based Communication System for Cultural Heritage. Approach and Progress of the WissKI Project in: Hans Bock u.a. (eds.): *Scientific Computing and Cultural Heritage*. Berlin: Springer 127-135.

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.

Sahle, Patrick (2013a): *Digitale Editionsformen - Teil 1: Das typografische Erbe: Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Norderstedt: BoD.

Sahle, Patrick (2013b): *Digitale Editionsformen – Teil 2: Befunde, Theorie und Methodik. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Norderstedt: BoD.

Ist kooperativ jetzt umsonst? Die Ausweisung von Datenautorenschaft als neue Form wissenschaftlicher Reputation zur Förderung offener Forschungsdatenkulturen

Moeller, Katrin
katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Vorbemerkung

Die Bedeutung von Forschungsdatenmanagement ist mittlerweile umfänglich in der Wissenschaftskultur angekommen. Die Vorteile der Open Data Sciences überzeugen schnell, auch wenn die jüngst vom Rat für Informationsinfrastrukturen (RfII) pointiert formulierten Herausforderungen (RfII 2016, RfII 2017) weiterhin bestehen. Sie betreffen vor allem den Umgestaltungsprozess hinsichtlich des Aufbaus einer Landschaft von dauerhaften, stabilen Infrastrukturangeboten der Langzeitarchivierung, die durch die Anpassung und Koordination von Fördermechanismen, Personalentwicklung, Qualitätssicherung und der

Entwicklung einer neuen "Forschungsdatenkultur" geprägt sind (AG Datenzentren 2017: 2-3). Den letzten Punkt möchte ich mit meinem Vortrag aufgreifen und aus dezidiert geisteswissenschaftlicher Perspektive einer datenproduzierenden und -bewahrenden Institution einen Beitrag zum Forschungsdatenmanagement formulieren, der meines Erachtens bisher kaum reflektiert wird: Es geht um die Entwicklung neuer Forschungsdatenkulturen, um Anreize und vor allem darum, welche wissenschaftliche Reputation sich mit der Nachnutzung von Forschungsdaten für den Urheber von Daten verbindet. Welche Anreize können Forschenden geboten werden, Daten tatsächlich zur möglichst flexiblen Nachnutzung frei zu geben? Diskutiert wird dazu nicht die Datenautorenschaft selbst, die über das Urheberrecht längst etabliert ist (Beer u.a. 2014: 4) und die heute bereits durch die Veröffentlichung von Forschungsdaten auf einem Forschungsdatenrepositorium über Lizenzen geregelt wird (CreativeCommons 2017, Beer u.a. 2014: bes. 24f.). Hier geht es vielmehr um eine Diskussion, welchen weiteren Weg die Datenautorenschaft nach der Veröffentlichung in einem anerkannten Repositorium nimmt und wie Datenautorenschaft innerhalb der Wissenschaft mehr Anerkennung, Sogkraft und Reputation entfalten. Grundlegend kann dies zu einer größeren Bereitschaft von Forschenden zum Data-Sharing beitragen.

Was beinhaltet ausgewiesene Datenautorenschaft?

Momentan werden alle Urheber eines wissenschaftlichen Forschungsergebnisses in Textpublikationen unabhängig vom inhaltlichen Beitrag als Autoren genannt. Es ist sinnvoll, hier ein Unterscheidungskriterium einzuführen. Dabei wird strikt nach den eigentlichen Textautoren und allen anderen Beiträgern eines wissenschaftlichen Forschungsergebnisses unterschieden. Textautoren sind ausschließlich diejenigen, die maßgeblich den Inhalt eines Textes verfassen, der das Forschungsergebnis und die Analyse repräsentiert. Sie sind für die Inhalte des Beitrags verantwortlich. Neben diesen Autoren für den Text werden dann alle weiteren Autoren als "Datenautor(en)" oder "Datengeber" geführt. Sie haben ebenfalls am Zustandekommen des Forschungsergebnis wesentlichen Anteil. Diese Beteiligung kann bspw. darin bestehen, prinzipiell urheberrechtlich geschützte Daten über Lizenzen offen für eine Analyse zur Verfügung zu stellen. Wichtig ist dabei, dass die Daten des Datengebers - anders als beim Zitat - in einem wesentlichen Umfang (mehr als ein Drittel) Verwendung finden. Es ist dahingehend nicht entscheidend, ob die Daten im Datenkonvolut des Datennutzers auch einen wesentlich Bestandteil bilden oder letztlich nur einen kleinen Baustein ausmachen.

Neben die Autoren, Herausgeber und Übersetzer würde also eine weitere qualitative Gruppe der in einer Textpublikation genannten Autoren treten, die nun allerdings explizit keinen eigenen Beitrag am Text, wohl aber einen unmittelbaren, wesentlichen Beitrag zur Quellenbasis eines Forschungsergebnisses leisten. Dieses Kriterium ist das wesentliche Unterscheidungsmerkmal zum Textautor. In den Metadaten einer Publikation würden diese Autoren analog zu den Herausgebern eines Sammelwerks mit einem geeigneten Kürzel benannt (DA) und ggf. auch separat ausgewiesen. Für den Datenautor zählt diese Nennung aber als weitere Publikation, wenn auch mit einem abgestuften Renommee.

Wozu brauchen wir extra ausgewiesene Datenautorenschaft?

a) Vergleich verschiedener Methoden zur Messung wissenschaftlicher Leistung

Es ist relativ einfach zu zeigen, warum viele WissenschaftlerInnen zögerlich bleiben, Daten zu teilen. Vor allem in der Geisteswissenschaft gibt es dafür eine Reihe von Gründen, von denen hier nur einige wenige knapp skizziert werden sollen:

- Datenproduktionen in den Geisteswissenschaften sind kosten-, zeit- und personalintensiv, für Datenbereinigung und Dokumentation müssen zusätzliche Aufwendungen gemacht werden
- Erschließungsprozesse von Medien und Quellen sind Teil des wissenschaftlichen Forschungsprozesses mit eigener Fachdisziplin (z. B. den Grund- und Hilfswissenschaften in der Geschichtswissenschaft)
- Forschungsdaten unterstehen damit - vielleicht auch im Unterschied zu sensorgestützten Datenproduktion in einigen naturwissenschaftlichen Sparten - dem Urheberrecht, da sie über eine eigenständige wissenschaftliche Leistung mit ausreichender Schöpfungshöhe verfügen. Dies gilt umso mehr, wenn Forschungsdaten nicht nur die Wiedergabe einer einzelnen Quelle in Form von Transkriptionen repräsentieren, sondern über den Erschließungsprozess mit einer Vielzahl von

Annotationen und editionskritischen Anmerkungen versehen werden oder komplexe Datenstrukturen einer ganzen Serie von Quellen kombinieren.

Zunächst möchte ich kurz einen Vergleich des Umgangs mit der Datenautorenschaft in den verschiedenen Forschungsdisziplinen anstellen. Grundlegende Unterschiede liegen in der Messbarkeit der Forschungsleistung über mehr oder weniger umstrittene Methoden der Auszählung von Zitationsraten, die in den STM-Fachdisziplinen (Naturwissenschaften, Technik, Medizin) eine etablierte Forschungspraxis repräsentieren. Abgesehen davon, dass der Journal Impact Factor auch in den STM-Fächern keineswegs unumstritten und grundsätzlich davon auszugehen ist, dass Daten in Repositorien wie auch in Open Access-Veröffentlichungen durch die bevorzugten Evaluationsmetriken nicht gleichwertig erfasst werden (Herb 2010, Kap. 1.3), ist dieses Verfahren in den Geisteswissenschaften bisher grundsätzlich nicht anwendbar (Wissenschaftsrat 2006: 48ff. Jehne 2009: 59). Die wissenschaftliche Reputation durch die Zitation von Daten greift in den Geisteswissenschaften daher nicht in der gleichen Weise, wie in den STM-Fächern. Auch in den Naturwissenschaften entstehen durch die Vervielfachung der Co-Autorenschaft erhebliche neue Unschärfen bei der Leistungsbewertung von wissenschaftlichen Ergebnissen, was wiederholt zu Reglungsbedarf der Hochschulen und der DFG führte (DFG 2013: 20). Eine Trennung von Text- und Datenautorenschaft auf der Ebene der Metadatenhaltung brächte hier wesentliche Vorteile für alle Wissenschaftsdisziplinen, da sie Beiträge an Forschungsergebnissen durch klarere Definitionskriterien wieder transparenter ausweisen würde.

b) Zitation versus Datenautorenschaft

Im zweiten Teil möchte ich durch einen Vergleich von Zitation und Datenautorenschaft die Vor- und Nachteile beider Prinzipien diskutieren und überlegen, wie man die Vorzüge beider Verwendungsweisen miteinander kombinieren könnte.

Das Grundprinzip des Zitats ist die Belegfunktion. Ein Zitat ist nach dem Urheberrecht dann zugelassen, wenn es eigene Ideen oder Gedanken unterstützt bzw. Ideen anderer aufgreift und in den eigenen Text integriert. Ist dem Recht genüge getan, verliert nach einer Faustregel der Text auch ohne das Zitat nicht an Sinn (Schwenke 2011). Das Urheberrecht regelt im § 52a, dass in der wissenschaftlichen Forschung "kleine Teile eines Werkes, Werke geringen Umfangs sowie einzelne Beiträge aus Zeitungen oder Zeitschriften" für die wissenschaftliche Forschung genutzt werden können (Bundesministerium 2013, § 52a). Auch wenn es kein festgesetztes Limit für den Umfang eines Zitats gibt, sollte es klar begrenzt sein. Als Faustregel führt ein Zitat nicht mehr als ein Drittel eines Textes auf (Schwenke 2011). Die Verwendung großer Teile von Daten oder ganzer Datenkonvolute zur Produktion eines Forschungsergebnisses verlässt den Rahmen einer Zitation deutlich.

Der wesentliche Vorteil des Zitierens ist neben seiner festen Kanonisierung in allen Wissenschaftsdisziplinen das Prinzip der Kontaktlosigkeit. Zudem wird klargestellt, dass nur der Autor inhaltliche Verantwortung für ein Forschungsergebnis trägt. Während Autoren sich untereinander abstimmen, Inhalte diskutieren und Rechte klären müssen, kann ein Wissenschaftler durch die Zitation Ergebnisse anderer unter den genannten Voraussetzungen in seine eigene Leistung einbinden und ausweisen. Die Autorenschaft ist daher organisatorisch aufwändiger und setzt die Erreichbarkeit des Urhebers voraus, wiewohl heute mit internetbasierten Referenzsystemen wie OrcID (ORCID 2017) Voraussetzungen dafür geschaffen werden.

Letztlich verzichtet der Autor durch die Lizenzierung von Forschungsdatensätzen in Forschungsrepositorien weitgehend auf alle Nutzungs- und Verwertungsrechte zugunsten einer möglichst breiten Nachnutzung von Daten. Neben anonymen Lizenzen ist der meistgewählte Typ an Lizenzen vermutlich die Auflage zur Namensnennung, die über das Zitat erfolgt. Dies ist möglich, indem Datenrepositorien die Freistellung der Daten von allen Verwertungsrechten nach § 15 des Urheberrechtsvertrages in ihre Nutzungsverträge übernehmen. Während Wissenschaftler und Wissenschaftlerinnen mit solchen Formen der Freistellung von Daten oft keine größeren Probleme haben, möchten sie aber weitgehend nicht auf die wissenschaftliche Verwertung ihrer Daten und die damit in Verbindung stehende wissenschaftliche Reputation verzichten. Sie stellt die eigentliche Währung der Wissenschaft dar.

Zudem möchte kein Datenautor für die Verwendung seiner Daten verantwortlich gemacht werden: Neben der formalen Trennung von Daten- und Textautor, kommt auch die inhaltliche zum Tragen. Das Forschungsergebnis verantwortet hier nur der Textautor.

Daher sollten Lizenzmodelle in Forschungsrepositorien idealerweise die Vorteile der Autorenschaft (wissenschaftlicher Mehrwert) und die Vorteile des Zitierens (Kontaktlosigkeit, keine rechtliche Regelung, Klarstellung der Autoren-

schaft) miteinander verbinden. Möglich wäre dies, indem Datenrepositorien die Datenautorenschaft von allen Rechten freistellen (wie dies in Lizenzen der Fall ist) aber entsprechend einer Verwendung von Daten in großen Teilen die Nennung als Datenautor oder Datengeber vorschreiben. Letztlich würde sich an der bisher diskutierten Praxis der Repositorien nichts Wesentliches ändern, mit Ausnahme der veränderten Notation im Sinne eines Autors des wissenschaftlichen Ergebnisses.

Datennutzer müssten lediglich zur Meldung von Veröffentlichungen beim Repositorium verpflichtet werden. Der Datengeber kann sich auf diese Weise bei Wunsch über eigene Veröffentlichung informieren.

Die Unterscheidung in Text- und Datenautorenschaft ist damit vor allem eine transparente Präzisierung der einzelnen Forschungsleistungen, die rechtlich über die Nutzungsverträge der Datenrepositorien Regelung erfährt. Dieses Verfahren findet bei jenen Daten Anwendung, deren Lizenzen eine freizügige, verschneidbare Nutzung ermöglichen. Auch bei anderen Daten ist natürlich eine Trennung von Datenautorenschaft und Textproduzent möglich, hängt jedoch durch die urheberrechtlichen Beschränkungen nach wie vor an einem persönlichen Austausch der beteiligten Personen.

Bibliographie

AG Datenzentren des DHd (2017): Stellungnahme der DHd AG Datenzentren und des DHd-Verbands zur Nationalen Forschungsdateninfrastruktur (NFDI), URL: http://dig-hum.de/sites/dig-hum.de/files/DHd_NFDI_Stellungnahme_2017-07-31.pdf [letzter Zugriff 22. September 2017].

Bundesministerium der Justiz und für Verbracherschutz (2013): Gesetz über Urheberrecht und verwandte Schutzrechte, in: *Juris*, URL: https://www.gesetze-im-internet.de/urhg/index.html#BJNR012730965BJNE023100377 [letzter Zugriff 22. September 2017].

Deutsche Forschungsgemeinschaft (2013): Sicherung guter wissenschaftlicher Praxis. Denkschrift, Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft", Weinheim 2013, URL: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf [letzter Zugriff 22. September 2017].

Martin Jehne (2009): Publikationsverhalten in der Geschichtswissenschaft, in: Alexander von Humboldt Stiftung (Hg.), *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen 12*, 2. erw. Auflage, S. 59-61, URL: https://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publikationsverhalten2_kompr.pdf [letzter Zugriff 22. September 2017].

Nikolaos Beer, Kristin Herold, Wibke Kolbmann u.a. (2014): Datenlizenzen für geisteswissenschaftliche Forschungsdaten. Rechtliche Bedingungen und Handlungsbedarf, in: *GOEDOC: Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen*, Göttingen, URL: http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2014-6.pdf [letzter Zugriff 22. September 2017].

Rat für Informationsinfrastrukturen (RfII) (2016): Leistung aus Vielfalt, Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, URL: http://www.rfii.de/de/category/dokumente/ .

Thomas Schwenke (2011): Texte richtig zitieren, statt plagiieren (Anleitung mit Checkliste), in: *ILAWit. Blog zum Social Media-, Marketing, Online- und Datenschutzrecht*, Berlin, URL: http://rechtsanwalt-schwenke.de/texte-richtig-zitieren-statt-plagiieren-anleitung-mit-checkliste/ [letzter Zugriff 22. September 2017].

Ulrich Herb 2010: Open Access, zitationsbasierte und nutzungsbasierte Impact Maße: Einige Befunde. Tagungsband der 12. Tagung der Deutschen ISKO, (International Society Knowledge Organization), URL: http://scidok.sulb.uni-saarland.de/volltexte/2010/3307 [letzter Zugriff 22. September 2017].

Wissenschaftsrat 2006: Empfehlungen zur Entwicklung und Förderung der Geisteswissenschaften in Deutschland, Köln, URL: http://www.wissenschaftsrat.de/download/archiv/geisteswissenschaften.pdf [letzter Zugriff 22. September 2017].

ORCID (2017): Distinguish yourself in three easy steps, URL: https://orcid.org / [letzter Zugriff 22. September 2017]..

"Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen

Schmidt, Thomas
thomas.schmidt@sprachlit.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Burghardt, Manuel
manuel.burghardt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Dennerlein, Katrin
katrin.dennerlein@uni-wuerzburg.de
Institut für Deutsche Philologie, Julius-Maximilians-Universität Würzburg

Sentiment Analyse und Dramenanalyse

Sentiment Analyse (SA) beschreibt eine Reihe von computergestützten Methoden zur Prädiktion der Polarität eines Texts, versucht also vereinfacht gesagt automatisiert herauszufinden, ob ein Text ein positives oder negatives Gefühl ausdrückt (Liu 2016). Darüber hinaus werden teilweise auch komplexere emotionale Kategorien (wie z.B. Zorn und Freude) betrachtet (Mohammad & Turney 2010). Zentrale Anwendungsfelder der SA sind bislang vor allem die Analyse von Online-Reviews (McGlohan, Glance & Reiter 2010) und Social Media-Daten (Kouloumpis, Wilson & Moore 2011).

Zur Analyse von literarischen Texten mittels SA-Techniken finden sich bislang nur wenige Studien, z.B. zu Märchen (Alm, Roth & Sproat 2005) und Romanen (Kakkonen & Kakkonen 2011; Elsner 2012; Jannidis et al. 2016). Auf größeren Textkorpora wurde getestet, inwiefern SA-Werte eines Textes und Emotionskurven von Texten zur Genreklassifikation verwendet werden können (Kim, Padó & Klinger 2017) und wie begriffsgeschichtliche Bedeutungsverschiebungen in literarischen Texten mithilfe von erweiterten SA-Methoden erforscht werden können (Buechel, Hellrich & Hahn 2017). In Dramentexten hat man bisher die Verteilung von emotionalen Kategorien (Mohammad 2011) oder die Entwicklung von Figurenbeziehungen (Nalisnick & Baird 2013) in Shakespeare-Dramen untersucht. Auch der vorliegende Beitrag beschäftigt sich mit dem Einsatz von SA im Bereich der Dramenanalyse. Es werden erstmals systematisch verschiedene Methoden der SA für Dramen getestet und evaluiert. Zudem wird exploriert, inwiefern bisher in der Literaturwissenschaft erforschte Aspekte von Dramen mithilfe der SA erfasst werden und inwiefern die SA auch für die Gewinnung neuer literaturwissenschaftlicher Erkenntnisse eingesetzt werden kann.

Das im Rahmen dieser Studie verwendete Lessing-Korpus umfasst ein mit Strukturinformationen annotiertes Dramenkorpus mit 11 Dramen, bestehend aus insgesamt 8224 Einzelrepliken. Sämtliche Dramen wurden über die Plattform *TextGrid* [1] bezogen, so dass alle im Rahmen dieses Beitrags entwickelten Tools auch auf andere *TextGrid*-Dramen anwendbar sind. Mit dem am besten evaluierten SA-Verfahren wurde eine webbasierte Anwendung zur Analyse und Visualisierung von Sentiment-Verteilungen und -Verläufen implementiert.

Evaluation unterschiedlicher SA-Verfahren

Lexikonsbasierte SA

Innerhalb der SA unterscheidet man zwei wesentliche Ansätze: (1) die Nutzung maschinellen Lernens und (2) die Verwendung lexikonbasierter Verfahren. Für das erstgenannte Vorgehen ist typischerweise ein mit Sentiment-Informationen annotiertes Trainingskorpus notwendig (D'Andrea et al. 2015), welches für die Dramenanalyse bislang nicht vorliegt. Aus diesem Grund werden in der vorliegenden Arbeit lexikonbasierte Verfahren eingesetzt. Ein Sentiment-Lexikon ist dabei eine Wortliste, in der für jedes Wort Sentiment-Informationen angegeben sind (Liu 2016: 10), also z.B. ob es positiv oder negativ konnotiert ist und in welchem Ausmaß (Polaritätsstärke). Ein derartiges Wort nennt man auch *sentiment bearing word* (SBW; Liu 2016: 189).

SA-Parameter

Folgende SA-Optionen wurden in unterschiedlichen Kombinationen systematisch evaluiert:

i) Lexika – Es wurden fünf zentrale Sentiment-Lexika für den deutschsprachigen Bereich herangezogen: *SentiWortschatz* (SentiWS; Remus, Quasthoff & Heyer 2010), die *Berlin Affective Word List – Reloaded* (Bawl-R; Vo et al. 2009), die deutsche Version des *NRC Emotion-Association Lexicon* (NRC, Mohammad & Turney 2010), ein Lexikon von Clematide & Klenner (2010; im folgenden CK genannt) und das *German Polarity Clues* (GPC; Waltinger 2010). SentiWS, Bawl-R und CK enthalten Polaritäten und Polaritätsstärken, das NRC und GPC nur Polaritätsangaben. Das NRC enthält des Weiteren Annotationen zu acht unterschiedlichen Emotionen (Zorn, Furcht, Erwartung, Freude, Vertrauen, Ekel, Traurigkeit, Überraschung).

ii) Historisch-linguistische Varianten – Über ein Tool des Deutschen Text-Archivs von Jurish (2011) wurde die Option der Lexikon-Erweiterung mit historischen linguistischen Varianten der Originalwörter untersucht.

iii) Stoppwortlisten – Analog zu Saif et al. (2014) wurde der Einfluss der Verwendung von insgesamt drei unterschiedlichen Stoppwortlisten auf die Qualität der SA untersucht. Grund hierfür ist, dass durch verschiedene Kombination der Verfahren Sentiment-tragende Stoppwörter entstehen. Neben herkömmlichen Stoppwörtern wurden dabei auch Listen mit hochfrequenten Wörtern des Korpus untersucht. Dadurch wird der Einfluss von Wörtern analysiert, die zwar als sentiment-tragend in SA-Lexika ausgezeichnet werden, aber aufgrund der häufigen Nutzung im Korpus ein ungleichmäßiges Sentiment-Gewicht erzeugen (z.B. Herr, Fräulein).

iv) Lemmatisierung – Eine weitere untersuchte Verarbeitungsform für die SA ist die Lemmatisierung. Als Lemmatisierer werden der *Pattern-Lemmatisierer* (De Smedt & Daelemans 2012) der Python-Bibliothek *textblob* und der Python-Wrapper des *treetagger*-Tools (Schmid 1995) evaluiert. Viele SA-Lexika enthalten lediglich Grundformen. Aufgrund der Probleme und Schwierigkeiten der Lemmatisierung im Deutschen (Eger, Gleim & Mehler 2016) soll vergleichend untersucht werden, welcher Lemmatisierer die besten Ergebnisse in Kombination mit Lexika erzielt. Ferner enthalten einige SA-Lexika manuell angegebene flektierte Wortformen. Es wird somit auch die automatische Lemmatisierung mit der manuellen Erweiterung verglichen.

SA-Metriken

Alle nachfolgenden Berechnungen wurden bezüglich aller kombinatorischen Möglichkeiten der soeben beschriebenen SA-Parameter durchgeführt. Dabei werden die jeweiligen SA-Metriken nach Term-Zähl-Methodik (Kennedy & Inkpen 2006) berechnet, d.h. ein Text wird hinsichtlich vorhandener SBWs untersucht, positive und negative Wörter ausgezählt und für einen Polaritätswert die positive von der negativen Zahl subtrahiert. SA-Metriken wurden auf folgenden Ebenen über die jeweils zugehörigen Texte kalkuliert: Drama, Akte, Szenen, Repliken sowie Sprecher und Sprecherbeziehungen pro Drama, Akt, Szene und Replik. Die Beziehungen zwischen den Figuren wurden nach einer Heuristik von Nalisnick & Baird (2013) berechnet.

Erstellung des Gold Standards

Zur systematischen Evaluation der Prädiktionsleistung der verschiedenen SA-Ansätze wurde ein Evaluationskorpus bestehend aus 200 Repliken erstellt. Bei der Auswahl der Repliken wurde darauf geachtet, dass die dramenspezifische Verteilung berücksichtigt wird, längere Dramen sind also mit mehr Repliken vertreten. Ferner wurden nur solche Repliken aufgenommen, die mindestens 19 Wörter umfassen. Diese Länge entspricht etwa -25% des Mittelwerts des Gesamtkorpus und vermeidet damit die Selektion von zu kurzen Repliken. Es wurde insgesamt auf eine gleichmäßige Längenverteilung geachtet.

Die Repliken wurden von insgesamt fünf Personen (4 weiblich, 1 männlich; alle jeweils mit Deutsch als Muttersprache) jeweils unabhängig voneinander bezüglich deren Polaritätswirkung bewertet. Die Polarität jeder Replik wurde jeweils sechswertig (sehr negativ, negativ, neutral, gemischt, positiv, sehr positiv) und binär (positiv, negativ) bewertet. Die Annotationen wurden bezüglich des Übereinstimmungsgrades analysiert. Dazu wurden das Übereinstimmungsmaß Fleiss' Kappa (Fleiss 1971) sowie der Durchschnittswert der prozentualen Übereinstimmung aller Annotatoren und Annotatorinnen berechnet (vgl. Tabelle 1).

Annotationsskala	*Fleiss' Kappa*	*Prozentuale Übereinstimmung*
Polarität-sechswertig	0,22	40%
Polarität-binär	0,47	77%

Tabelle 1. Annotator agreement.

Man erkennt eine geringe Übereinstimmung für die Bewertungsskala mit sechsstufiger Polarität

und eine moderate Übereinstimmung für die binäre Variante. Die Ergebnisse verhalten sich konform zu verwandten Studien bei der Interpretation literarischer Texte (Alm & Sproat 2005). Als finale Annotation für eine Replik wird die binäre Polarität gewählt, die die Mehrheit der Annotatoren und Annotatorinnen ausgewählt haben (Endresultat: 139 negativ, 61 positiv).

Evaluationsmaße

Als Evaluationsmaße wurden Genauigkeit (accuracy), Recall, Precision und F-Werte (Gonçalves et al. 2013) herangezogen. Abb. 1 zeigt einen Ausschnitt aus den je fünf besten Kombinationen pro Lexikon, geordnet nach Genauigkeit. [2]

Metric	DTAExtension	Lemmatization	Stopwords	accuracy	F-MeasureAvera
polaritySentiWS	dtaExtended	textblob	noStopwordList	0,67	0,6373626374
polaritySentiWS	dtaExtended	tokens	noStopwordList	0,665	0,5775402755
polaritySentiWS	dtaExtended	treetagger	noStopwordList	0,65	0,6042514699
polaritySentiWS	dtaExtended	treetagger	enhancedList	0,615	0,5558247527
polarityCd	dtaExtended	treetagger	enhancedList	0,595	0,5644107445
polarityCd	dtaExtended	treetagger	enhancedFiltered	0,585	0,5607419756
polarityCd	dtaExtended	textblob	enhancedList	0,565	0,5556577032
polarityGpc	noExtension	textblob	enhancedFiltered	0,56	0,5397008055
polarityGpc	dtaExtended	textblob	enhancedFiltered	0,56	0,5397008055
polarityCd	dtaExtended	treetagger	standardList	0,55	0,5499549955
polarityCdDichotom	dtaExtended	treetagger	enhancedList	0,535	0,5102040816
polarityCdDichotom	dtaExtended	treetagger	enhancedFiltered	0,53	0,5123975516
clearlyPolarityCombined	dtaExtended	textblob	enhancedList	0,51	0,5028409091
clearlyPolarityCombined	dtaExtended	treetagger	enhancedList	0,505	0,4854336131
polaritySentiWSDichotom	dtaExtended	tokens	noStopwordList	0,5	0,486863711

Abbildung 1: Ausschnitt aus der detaillierten Ergebnistabelle zur Evaluation der SA-Kombinationsmöglichkeiten.

Ergebnisse der Evaluation

Nachfolgend erfolgt eine überblicksartige Zusammenstellung einiger zentraler Ergebnisse aus der Evaluation:

- Eine explizite Lemmatisierung führt zu einer verbesserten Leistung. Beide Lemmatisierer erzielen dabei meist ähnliche Ergebnisse. Die Lexikonerweiterung durch historische Varianten macht die explizite Lemmatisierung jedoch weitestgehend unnötig, da hierbei auch eine grundlegende Lemmatisierung inkludiert ist.
- Es zeigt sich eine konsistente Verbesserung durch die Lexikonerweiterung mittels der Wort-Varianten aus dem Tool von Jurish (2011).
- Stoppwortlisten haben nur auf vereinzelte Lexika (GPC, CK) einen merklich positiven Einfluss.
- Lexika mit Polaritätsstärken sind meist besser als reine Term-Zähl-Verfahren desselben Lexikons.
- Das Lexikon, dass die höchsten Genauigkeiten für die SA erzielt, ist SentiWS
- Die beste Leistung (unter Analyse aller Metriken) erzielt das erweiterte SentiWS mit den Polaritätsstärken, lemmatisiert mittels Pattern-Lemmatisierer und ohne Stoppwortliste (Genauigkeit = 0,67; F-Wert = 0,64). Die Erkennungsrate ist besser als die random baseline von 0,576 aber schlechter als viele Erkennungsraten auf anderen Anwendungsgebieten der SA (Vinodhini & Chandrasekran 2012).

Aufgrund der Tatsache, dass hier ein verhältnismäßig simpler SA-Ansatz gewählt wurde und bereits menschliche Annotatoren und Annotatorinnen Schwierigkeiten mit der Polaritätsbestimmung haben, sind die Ergebnisse insgesamt durchaus positiv zu bewerten.

Online-Tool

Abschließend wurde auf Basis des besten SA-Ansatzes ein Web-Tool für die SA bei Dramen entwickelt. Dieses bietet interaktive Visualisierungen der Sentiment-Verteilungen und -Verläufe für alle berechneten Ebenen. Neben den SentiWS-Metriken wurden auch die Emotionskategorien des NRC integriert. Über das Tool kann man erste Fallstudien auf Dramen-, Akt-, Szenen-, Repliken-, Sprecher- und Sprecherbeziehungsebene durchführen. Die SA-Komponente ist online verfügbar. [3]

Trotz der historischen Differenz stimmen die Ergebnisse der automatischen SA tendenziell mit dem überein, was man in der Dramengeschichte über Bewertungen von Figuren und deren Verhalten weiß. Zusätzlich ist aber ein wichtiger heuristischer Mehrwert zu beobachten: eine Analyse allein auf der Basis von Sentiment-Zuschreibungen führt dazu, dass man das Augenmerk gezielt auf Fakten des Textes richtet, die bisher nicht berücksichtigt wurden.

Im Folgenden einige Beispiele für die Bestätigung bekannter Ergebnisse und für Entscheidungen von Analysefragen:

Fallstudie: Minna von Barnhelm

Die Analyse von Minna von Barnhelm zeigt, dass die negativen emotionalen Bewertungen insgesamt gegenüber den positiven deutlich überwiegen (vgl. Abb. 2). Dieser Befund bestätigt die bekannte Erkenntnis, dass Lessing das Schema des rührenden Lustspiels verwendet hat. Während die Komik im Stück eher das Ergebnis von Schlussprozessen ist, geht es auf der wörtlichen

Ebene überwiegend um ernste Vorwürfe und drohenden Identitäts- und Beziehungsverlust.

Abbildung 2: Polaritätsverteilung im Drama – Minna von Barnhelm

Es ist verschiedentlich behauptet worden (Saße 1993), Minna und nicht Tellheim sei die lächerliche Figur des Stücks. Die Sympathielenkung auf der wörtlichen Ebene des Textes, die in der unten stehenden Sentimentverteilung pro Akt abgebildet ist, kann dazu herangezogen werden, diese Frage negativ zu bescheiden (vgl. Abb. 3). Es ist eine auffällige Abweichung der Polarität im zweiten Akt erkennbar. In diesem Akt tritt Minna von Barnhelm zum ersten Mal auf, Tellheim jedoch nicht.

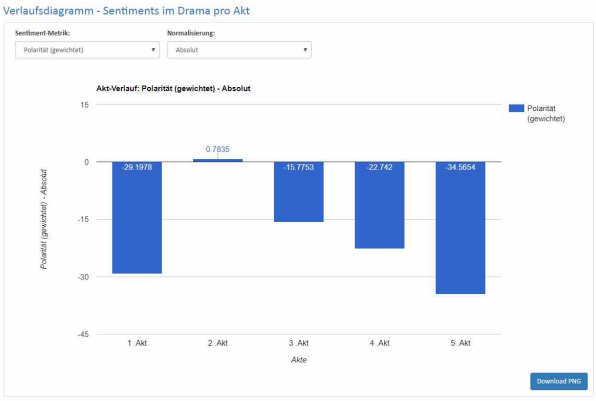


Abbildung 3: Polaritätsverlauf pro Akt – Minna von Barnhelm

Fallstudie: Emilia Galotti

Die letzte Visualisierung kann genutzt werden die Frage zu diskutieren, warum Emilia in Lessings Drama „Emilia Galotti“ sterben muss (vgl. Abb. 4). Auffällig ist hier die starke negative Bewertung Emilias im zweiten Akt. Entgegen bisheriger Interpretationen, in denen nur die Intrige des Prinzen und Marinelli dafür verantwortlich gemacht werden, dass Emilia um ihre Tugend fürchten und ihren Vater dazu bringen muss, sie umzubringen, wird dadurch die Abwertung allein durch die Avancen des Prinzen sichtbar, die später sowohl Emilias als auch für Odoardos Einschätzung der Ehrbarkeit Emilias in ihrem zukünftigen Leben bestimmen.

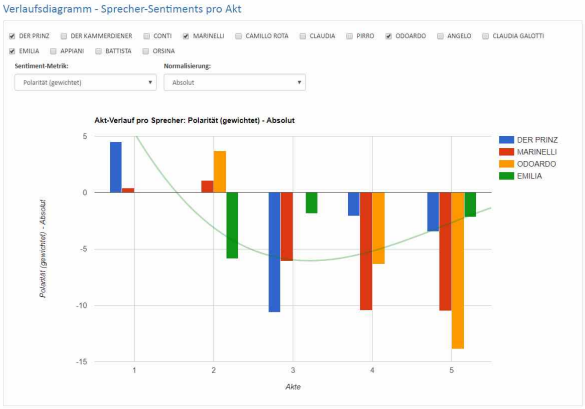


Abbildung 4: Polaritätsverlauf von Sprechern pro Akt – Emilia Galotti

Fazit

Insgesamt sind die ersten Analyse-Ergebnisse über das Web-Tool sehr vielversprechend. Dabei ist zu bedenken, dass über die Verwendung von SA-Lexika ein sehr einfacher SA-Ansatz gewählt wurde. Über ML- oder Hybrid-Ansätze können Besonderheiten der poetischen und veralteten Sprache möglicherweise besser beachtet werden. Ferner ist fraglich, ob eine Reduktion auf das sonst in der SA übliche binäre System positiv/negativ ausreichend ist für komplexe Interpretationen von Emotionen in Dramen.

Durch Optimierung des SA-Verfahrens, Ausbau der Funktionen im Front-End und Erweiterung des Tools mit zusätzlichen Dramen sollen künftig Möglichkeiten und Nutzen der SA in der Dramenanalyse weiter exploriert werden.

Fußnoten

1. https://textgridrep.org/repository.html; Hinweis: alle im Beitrag erwähnte URLs wurden zuletzt am 12.1.2018 überprüft
2. Die vollständige Tabelle ist online verfügbar unter https://drive.google.com/open?id=1cvyqiiL-J03XT1VNaWgSDoajeTE3wgeqxxr2PXp-VM4w

3. http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa_selection.html

Bibliographie

Alm, Cecilia Ovesdotter / Sproat, Richard (2005): "Emotional sequencing and development in fairy tales.", in: *International Conference on Affective Computing and Intelligent Interaction* 668-674.

Alm, Cecilia Ovesdotter / Roth, Dan / Sproat, Richard (2005): "Emotions from text: machine learning for text-based emotion prediction.", in: *Proceedings of the conference on human language technology and empirical methods in natural language processing* 579-586.

Buechel, Sven / Hellrich, Johannes / Hahn, Udo (2017): "The Course of Emotion in Three Centuries of German Text – A Methodological Framework.", in: *Digital Humanities 2017* 176-179.

Clematide, Simon / Klenner, Manfred (2010): "Evaluation and extension of a polarity lexicon for German.", in: *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* 7-13.

D'Andrea, Alessia et al. (2015): "Approaches, tools and applications for sentiment analysis implementation.", in *International Journal of Computer Applications* 125.3: 26-33.

De Smedt, Tom / Daelemans, Walter (2012): "Pattern for python.", in: *Journal of Machine Learning Research* 13: 2063-2067.

Eger, Steffen / Gleim, Rüdiger / Mehler, Alexander. (2016). "Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art.", in: *LREC* 1507–1513.

Elsner, Micha (2012): "Character-based kernels for novelistic plot structure.", in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* 634-644.

Fleiss, Joseph L. (1971): "Measuring nominal scale agreement among many raters.", in: *Psychological bulletin* 76.5: 378-382.

Gonçalves, Pollyanna, et al. (2013): "Comparing and combining sentiment analysis methods.", in: *Proceedings of the first ACM conference on Online social networks* 27-33.

Jannidis, Fotis, et al. (2016): "Analyzing Features for the Detection of Happy Endings in German Novels.", in: *arXiv preprint* arXiv:1611.09028

Jurish, Bryan (2011): *Finite-state canonicalization techniques for historical German*. Diss. Universitätsbibliothek der Universität Potsdam.

Kakkonen, Tuomo / Kakkonen, Gordana Galić (2011): "SentiProfiler: creating comparable visual profiles of sentimental content in texts.", in: *Language Technologies for Digital Humanities and Cultural Heritage* 62-67.

Kennedy, Alistair / Inkpen, Diana (2006): "Sentiment classification of movie reviews using contextual valence shifters.", in: *Computational intelligence* 22.2: 110-125.

Kim, Evgeny / Padó, Sebastian / Klinger, Roman (2017): "Investigating the relationship between Literary Genres and Emotional Plot Development.", in: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* 17–26.

Kouloumpis, Efthymios / Wilson, Theresa / Moore, Johanna D. (2011): "Twitter sentiment analysis: The good the bad and the omg!.", in: In *Proceedings of the Fifth International Conference on Weblogs and Social Media* 538-54.

Liu, Bing (2016): *Sentiment analysis: Mining opinions, sentiments, and emotions.* New York: Cambridge University Press.

McGlohon, Mary / Glance, Natalie S. / Reiter, Zach (2010) "Star Quality: Aggregating Reviews to Rank Products and Merchants.", in: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)* 114-121.

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.

Mohammad, Saif M. / Turney, Peter D. (2010): "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon.", in: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* 26-34.

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.

Remus, Robert / Quasthoff, Uwe / Gerhard, Heyer (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.", in: *LREC* 1168-1171.

Saif, Hassan, et al. (2014): "On stopwords, filtering and data sparsity for sentiment analysis of twitter.", in: *Proc. 9th Language Resources and Evaluation Conference (LREC)* 810-817.

Saße, Günter (1993): *Liebe und Ehe: oder, wie sich die Spontaneität des Herzens zu den Normen der Gesellschaft verhält. Lessings Minna von Barnhelm.* Tübingen: Niemeyer.

Schmid, Helmut (1995): "Improvements in part-of-speech tagging with an application to German.", in: *Proceedings of the acl sigdat-workshop*.

Vinodhini, G. / Chandrasekaran, R. M. (2012): "Sentiment analysis and opinion mining: a survey.", in: *International Journal of Advanced Research in Computer Science and Software Engineering* 2.6: 282-292.

Võ, Melissa LH, et al. (2009): "The Berlin affective word list reloaded (BAWL-R) ", in: *Behavior research methods* 41.2: 534-538.

Waltinger, Ulli (2010): "Sentiment Analysis Reloaded-A Comparative Study on Sentiment Polarity Identification Combining Machine Learning and Subjectivity Features.", in: *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*.

Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning

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

In den Geisteswissenschaften werden schon heute grosse Mengen an Text durchsucht, weiterverwertet und analysiert. Die Aufbereitung von Scans und Fotografien mit *optical character recognition* (OCR) bei gedruckten Texten wird erwartet. Gleichzeitig stehen selten Überlegungen im Zentrum, wie Text erkannt, welche Annahmen und vor allem welches Textverständnis vorausgesetzt wurde. Mit Hilfe des Sahleschen Textrads analysiert der Beitrag die automatisierte Erkennung von Texten mit neuronalen Netzen und fordert eine höhere Transparenz der verwendeten Methoden.

Sowohl Korpuslinguistik als auch Literatur- und Geschichtswissenschaften sind interessiert am Auffinden von Einzelbelegen, Mustern oder Entitäten in grossen Datenmengen. Tausende Seiten oder hunderte von Büchern lassen sich etwa mittels topic modeling klassifizieren (Schöch 2017). Der Prozess der Erkennung, der Weg zu den durchsuchbaren Texten, steht in den Überlegungen der Fächer jedoch meistens nicht im Zentrum. Obwohl Probleme der OCR-Erkennung angemerkt werden, ist die Textgüte nur bedingt ein Feld der Reflexion, die über Klagen hinausläuft (Ausnahme: Piotrowski 2012). Bedingt durch die Nutzung kommerzieller Produkte und entsprechend kaum offengelegter Prozesse, wird die Erkennleistung als gegeben angenommen und höchstens im *post-processing* die Qualität der Texte verbessert (bspw. PoCoTo: Vobl 2014).

Welcher Text?

Über das Monieren von Fehllesungen heraus, fehlt eine Reflexion, wie mit automatisch erkannten Texten umgegangen werden soll, gänzlich. Die Frage „welcher Text erkannt werden soll“, wird nicht thematisiert. Das hängt auch damit zusammen, dass Textverständnisse im digitalen Raum geprägt sind durch die Editionswissenschaften, einer Fachschaft, die aus einer anderen Richtung das digitale Feld bearbeitet. Die Qualität der Texterkennung, etwa der Transkription handschriftlicher Dokumente, steht nicht im Fokus, da bei menschlicher Erkennung durch Experten von einer Güte um 99,99% ausgegangen werden kann. Die Unsicherheiten, die unsicheren Lesungen, sind höchstens Teil von Paläographie orientierter fachspezifischer Debatten und nicht grundsätzlich ein qualitativer Messwert.

Die Editionswissenschaften waren es auch, die im Zuge der Digitalisierung Überlegungen zum Verständnis von Text hervorbrachten und sich – ganz im Sinne post-moderner Texttheorie – darauf einigten, dass es nicht den neutralen zu edierenden Text gibt. Ausgefaltet und visuell umgesetzt in Form eines „Textrads“ durch Patrick Sahle (Sahle 2013: 45-52). Erst das Verständnis von der Mehrschichtigkeit und Formbarkeit des Textbegriffes machte mehr oder minder konsequente Umsetzungen von digitalen Editionen überhaupt möglich und entspannte die Diskussion zwischen Philologie und Geschichtswissenschaft zu den je eigenen Vorstellungen von Text(-aufbereitung). Das Verständnis von Text unterscheidet sich dabei stark. Es kann vom Text als materiellem Ding ebenso ausgegangen werden, wie Text als Werk oder als Folge von Zeichen. Gerade für Fragen zur Umsetzung von Editionen, hilft das Textrad bei der Identifikation von Schwerpunkten, die Editionsentscheidungen unterstützen.

Im Rahmen von Projekten zur Texterkennung und Handschriftenerkennung durch grosse EU-Infrastrukturprojekte (IMPACT und READ) wurden derweil ebenso selten Überlegungen zum Text als Ressourcen angestellt und mehr auf Nachfragen beziehungsweise der Übernahme impliziter Vorstellungen abgestellt, die durch Informatiker oder Mathematiker bei der Entwicklung von Erkennungsalgorithmen eingebracht wurden. Dadurch muss auch deren Perspektive bei

einer Theoretisierung der Texterkennung mitberücksichtigt werden.

Die Frage nach dem Verständnis von Text bei automatisierten Erkennvorgängen, hilft weiter bei der Abwägung zum Verhältnis von Mensch und Maschine beim Zugänglichmachen von Texten und stellt auch althergebrachte editorische Praktiken in Frage.

Automatisch erkannter Text

Die Anwendung des Sahleschen Textrads auf automatisierte Texterkennung macht deutlich, dass gewisse Textformen auch mit besten Erkennmethoden nicht isoliert werden können: Ausgangspunkt ist immer eine Textversion (Druck- oder Manuskriptseite), die in Form eines Faksimilie/Digitalisat vorliegen muss. Zeichenhaftigkeit und auch intellektuelle Bezüge sind daraus ableitbar, Text als Werk etwa, wie es rekonstruiert oder abstrahiert wird, lässt sich dagegen nicht erkennen.

Das digitalisierte Objekt agiert bei der automatisierten Erkennung jeweils als Ausgangspunkt, das erneut konsultiert werden kann und bei Kontrolle und Überprüfung hilft. Das bedingt jedoch, dass auch die Art und Weise der Digitalisierung (Auflösung und Farbechtheit, aber auch Format und Aufnahmeverfahren), bei einer Kritik berücksichtigt werden müssen. Bereits der zugrundeliegende „Text“ ist also technisch geprägt.

Mit Fokus auf die Ausgabe des Erkennprozesses, werden erkannte Strings ins Zentrum gesetzt. Dabei muss nicht zwangsläufig nur ein String („die beste Lesung“) vorgelegt werden, sondern Varianten, also eine Reihe von Strings, die mehrere mögliche Lesungen enthalten, sind extrahierbar. Ergänzt um die durch die maschinell errechnete Wahrscheinlichkeit der Erkennung wird eine Matrix (sog. *confidence matrix*) an möglichen Lesungen und deren Wahrscheinlichkeit erstellt, die ebenfalls durchsucht werden kann. Insbesondere innerhalb von grossen Quellenmassen lassen sich so potente Suchen (Volltextsuchen ohne zugrunde liegendem Volltext sozusagen) realisieren, die ausgesprochen gute Ergebnisse erzielen. Mit dem Nachteil, dass je nach Suche auch *false-positive* Variantenlesungen vorgelegt werden. Die Methode wird daher nur bedingt für Auswertungen nutzbar, die auf Quantifizierung beruhen. Dank der Konfidenzen wird die Anzahl an Zeichenfehler, als zentralem Tool zur Messung von Textgüte relativiert, da ein alternativer Zugang zu den Strings besteht.

Innerhalb des Vorgangs zur Erkennung von Text kommt der eigentliche Erkennprozess jedoch erst an zweiter Stelle. Ebenso zentral und ebenfalls fehleranfällig, ist die Identifikation des Layout bzw. die Unterscheidung zwischen texttragenden und textfreien Zonen auf den zu erkennenden Digitalisaten, eine Aufgabe die für Menschen spielerisch einfach gelöst werden kann. Obwohl für reguläre Layouts von handschriftlichem Material in den vergangenen Monaten erhebliche Fortschritte erzielt wurden (Grüning 2017), bleiben Probleme im Umgang mit komplexen Layouts und insbesondere Tabellen, die in keine oder nur die ungenügende Identifikation von Zeilen mündet.

Der Punkt der Layouterkennung wird noch problematischer, da nur schwierig ausgewiesen werden kann, welche Teile als „texttragend“ identifiziert wurden. Bei Wettbewerben in den Computerwissenschaften werden unterschiedliche Messwerte angenommen. Etwa die Abweichung von einer manuell gezogenen Baseline [cBad] oder die Zuordnung zu Pixeln [DIVA-HisDB]. Allen Verfahren gemein ist der Bezug auf von Menschen hergestellte, relativ subjektive Grundlagen. Fakt ist, alles was nicht als Teil des Layouts identifiziert wird, kann im darauffolgenden Prozess nicht als Text erkannt werden.

(Vor-)Entscheidungen

Alle diese Überlegungen stellen nicht mehr als Grundlagen beziehungsweise Vorannahmen dar, die getroffen werden, bevor eine Erkennung überhaupt stattfinden kann. Im Gegensatz etwa zu händisch erstellten digitalen Editionen, ist eine Anpassung des Textbegriffs nicht im Erstellprozess möglich, sondern höchstens vor Beginn oder beim Abschluss der Bearbeitung.

Noch deutlicher wird die Abgeschlossenheit, nähert man sich der Texterkennung aus einer technischen Perspektive. Insbesondere für die Texterkennung von Handschriften und frühen Drucken lohnt sich der Einsatz von *machine learning*, konkret rekurrenten neuronalen Netzen (RNN). Das Training solcher Netze muss selbstredend vor der eigentlichen Erkennung erfolgen. Die trainierte Ausgabe entspricht dabei einem nachgeahmten, determinierten menschlichen Input. Je nach Grösse des Trainingssets und der Variabilität der Schriften wird dies mehr oder minder genau erreicht. Zentral im Prozess ist das erwartete Resultat, oder anders formuliert die Art und Weise, wie Text aufbereitet wird. Die Aufbereitung selbst, beispielsweise die stillschweigende Auflösung von Abkürzungen, Normalisierung von Schreibungen oder das Einfügen bzw. Zusammenführen von Zeilenumbrüchen, wird Konsequenzen auf die Ergebnisse haben. Auch der verwendete Zeichensatz (etwa die Codierung in Unicode) oder Vereinheitlichungen wird das Resultat beeinflussen.

Aus technischer Sicht gibt es innerhalb des Trainingsprozesses selbst nur einige wenige Parameter, die kontrolliert werden können. Entsprechend ist der gesamte Rest, Teil einer *blackbox*, die auch nicht näher analysiert werden kann, da das Funktionieren der einzelnen Neuronen in einem Netz nur schwer und häufig ohne Einsichten zum Funktionieren der gesamten Erkennung beobachtet werden können.

Ein Kontrollmechanismus findet sich einzig im standardisierten Testen der trainierten Modelle mit Hilfe von Testsets, also nach gleichem Muster hergestellte Seiten, die nicht fürs Training verwendet wurde und entsprechend Auskunft über die Leistungsfähigkeit eines Modells geben können. Zentrale Messwerte dabei sind *Character Error Rate* und *Word Error Rate*, die auf so genannter Ground-Truth basiert, also von Menschen hergestellte „korrekte" Lesungen der Texte.

Eine weitere Form der Einflussnahme besteht in der Verwendung von Wörterbüchern, die bei Unsicherheit herangezogen werden und plausiblere Lesungen (= im Wörterbuch) gegenüber anderen Strings bevorzugen. Für historische Schreibformen bestehen zwar Korpora, jedoch ist der Einsatz für Texte vor Ende des 19. Jahrhunderts (insbesondere für vormoderne Texte) umstritten, da keine Konventionen bestanden und die Gefahr der Hyperkorrektur aufgrund des verwendeten Wörterbuchs besteht.

Auch wenn es bislang nur beschränkte Erfahrungen mit dem Einsatz von *machine learning* bei der automatischen Erkennung von Text gibt, ist absehbar, dass die Verbesserungen zum Einsatz der Technologie führen werden. Mit dem Preis, dass Probleme des *machine learnings* mit eingeführt werden. Die *biases*, insbesondere die Perspektiven im Moment der Aufbereitung, von Trainingsdaten etwa, werden übernommen (Zundert 2016: 341). Im Kontext von automatisierter (Vor-)Aufbereitung von Bewerbungsdossiers oder nicht gender-gerechten Auswertungen von Informationsmassen wird das Probleme des Datenbias bei Methoden des *machine learnings* rasch sichtbar (Siehe dazu einen jüngeren Artikel aus dem britischen Guardian, Devlin 2017). Bei der Texterkennung mögen die Konsequenzen gesellschaftlich weniger gravierend ausfallen, problematisch und kritisch zu analysieren sind sie nichtsdestotrotz.

Ansprüche an automatische Texterkennung

Wie der kurze Abstecher in die Welt des maschinellen Lernens zeigte, ist eine Kontrolle der Erkennleistung nur ganz bedingt und basierend auf wenigen Faktoren möglich, entsprechend lässt sich zusammenfassend im Umgang mit automatisierten Texterkennungsalgorithmen eine Reihe von Forderungen ableiten, damit erkannte Texte kritisch eingeordnet werden können.

Messwerte basierend auf Testsets müssen ausgewiesen werden: Character Error Rate und Word Error Rate geben Aufschlüsse zur Qualität des erkannten Textes. Darüber hinaus ist eine Einschätzung sinnvoll, in welchem Bereich Falschlesungen häufig identifiziert wurden (Eigennamen, Zahlen etc.). Insgesamt sollte dadurch einsichtig werden, in welchen Bereichen Qualitätsprobleme zu erwarten sind.

Standards und Kernfragen an die publizierten Texte offen dokumentieren: In den Editionswissenschaften bereits praktiziert, wird der Umgang mit Frageperspektiven verbesserte Einsichten liefern, vor welchem Hintergrund Textkorpora erstellt wurden. Daran schliesst sich die Forderung nach **Offenlegung der zugrunde liegenden Ground-Truth zur Erstellung von Test- und Trainingssets** an: Damit wird nachvollziehbar, was zur Modellerstellung genutzt und auch, welchen Standards, Richtlinien und Gepflogenheiten dabei gefolgt wurde.

Durch den kritischen Umgang mit automatisch erkannten Texten eröffnet sich ein fundierter Umgang mit denselben, der mit gewissen Sicherheiten eine Weiternutzung von Text ermöglicht und die textzentrierten Teile der *digital humanities* in eine kritikfähige Zukunft führt.

Bibliographie

Devlin, Hannah (2017): AI programs exhibit racial and gender biases, research reveals. In: The Guardian vom 13.04. URL: https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals.

Grüning, Tobias, Labahn, Roger, Diem, Markus, Kleber, Florian, Fiel, Stefan (2017): READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents. [Preprint submitted to DAS2018] arXiv:1705.03311. URL: https://arxiv.org/abs/1705.03311.

Piotrowski, Michael (2012): Natural language processing for historical texts. Morgan & Claypool, San Rafael.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung., Schriften des IDE. BoD, Norderstedt.

Schöch, Christoph (2017): Topic Modeling Genre: An Exploration of French Classical and En-

lightenment Drama. Digital Humanities Quarterly 11.

Vobl, Thorsten, Gotscharek, Annette, Reffle, Uli, Ringlstetter, Christoph, Schulz, Klaus U. (2014): PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts, in: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, DATeCH '14. ACM, New York, NY, USA, pp. 57–61. doi:10.1145/2595188.2595197

Zundert, Joris J. van (2016): Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics, in: Schreibman, Susan, Siemens, Ray, Unsworth, John (Eds.), A New Companion to Digital Humanities. John Wiley & Sons, pp. 331–347.

Kritik der Digitalität (am Beispiel der digitalen Textwissenschaft)

Garcés, Juan

juan.garces@slub-dresden.de
Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), Deutschland

Bräuer, Johannes

johannes.braeuer@tu-dresden.de
Technische Universität Dresden, Institut für Philosophie

Welche Konsequenzen birgt die Digitalisierung der textbasierten Geisteswissenschaften? Zweifelsohne bringt die computerbetriebene Verarbeitung digital erschlossener Texte einen beachtlichen quantitativen Fortschritt mit sich: in dem Maße, in dem historische Texte digital erschlossen bzw. eine kritische Masse an digital erzeugten Texten zur Verfügung stehen, lassen sich aus den größeren, nun handhabbaren Datenmengen jetzt neues Wissen extrahieren und analysieren, das dieser Masse an Informationen gerecht werden kann. Neben den für das Druckzeitalter typischen akribischen Lektüren (*close reading*) gesellen sich jetzt beispielsweise „Lektüren aus der Distanz" (Moretti: *distant reading*; vgl. Jockers' *macroanalysis*), eine Kontrastierung, die sich in der Folge als allzu holzschnittartig erweisen wird. Nur schwerlich wird sich diese Unterscheidung aufrecht erhalten lassen, ohne das „close reading" als einen Anachronismus zu begreifen, der aber nur nachträglich in Kontrast zum distant reading funktioniert.

Der Ausdruck des *distant reading* ruft das Bild einer Neuorientierung in Bezug auf den wissenschaftlichen Gegenstand hervor. Die neugefundene Distanz zu den Gegenständen ermöglicht „eine spezifische Form der Erkenntnis" (Moretti), die im Falle Morettis einem Schritt der visualisierten Abstraktion (beispielsweise als Kurven, Karten und Stammbäume) und einer anschließend darauf basierenden Analyse entspringt. Grundlegend scheint hierbei der Zugang zum Wissensgegenstand über ein abstrahiertes digitales Objekt, das als Surrogat für den Gegenstand (beim Text, beispielsweise, das „Werk", jetzt zunehmend die „Werke") selbst steht. Die Neuorientierung zeichnet sich also modellartig nicht nur durch eine neue Perspektive (Distanz), sondern gleichzeitig auch durch ein neuartige Form von Gegenständlichkeit und demzufolge durch einen neuen mediatisierten Weltbezug aus, der von Alexander Galloway folgendermaßen beschrieben wird:

„[I]n order to be in a relation with the world informatically, one must erase the world, subjecting it to various forms of manipulation, preemption, modeling, and synthetic transformation. The computer takes our own superlative power over worlds as the condition of possibility for the creation of worlds. Our intense investment in worlds – our acute fact finding, our scanning and data mining, our spidering and extracting – is the precondition for how worlds are revealed. The promise is not one of revealing something as it is, but in simulating a thing so effectively that 'what it is' becomes less and less necessary to speak about, not because it is gone for good, but because we have perfected a language for it." (Interface Effect, 13)

Der wissenschaftliche Weltbezug durch ein medientechnologisch geprägtes Modell ist freilich kein Novum des digitalen Zeitalters. Dies hat man im Rahmen der digitalen Editorik überzeugend für das Druckzeitalter und im Allgemeinen für jedwede Texterschließung dargestellt:

„Der Textbegriff ist eine Funktion von Fragestellungen (Sichten auf den Text) und jeweils gegenwärtiger (also historischer) textmedialer Sozialisation. Bestimmte Textbegriffe werden immer durch bestimmte Texttechnologien gefördert oder behindert: Was der Text ist, ist eine ontologische Fragestellung, die von unterschiedlichen Technologien unterschiedlich beantwortet wird. Die Evolution der Techniken ist eine Evolution der Textbegriffe. Dabei neigen neue Techniken dazu, zunächst ihre Vorgängertechnologien zu imitieren (deren Textbegriff zu übernehmen), für bestimmte Probleme neue Lösungen anzubieten und damit für eine Verschiebung des Textbegriffes zu sorgen. Die Historizität der Technologien

erhellt die Relativität der Textbegriffe. “ (Sahle, Bd. 1, S. 391)

Das Potential der digitalen Erschließung wird aber nicht nur als die neueste Entwicklung einer medientechnologischen Evolution, sondern auch als qualitativen Sprung gesehen, der sich über den Medienwechsel hinaus „vielmehr als eine Befreiung von medialer Gebundenheit“ (Sahle, Bd. 2, S. 281) überhaupt beschreiben lässt.

Der so konstruierte digitale Textbegriff lässt sich in politischer und ökonomischer Hinsicht beschreiben, stellt sich aber in der Folge auch als ein epistemologisches bzw. wissenschaftstheoretisches Problem dar:

Auf politischer und ökonomischer Ebene übernimmt unsere Beschreibung wesentliche Elemente der Lyotardschen Analyse des Informationsbegriffs (vgl. aber auch die „kypernetische Hypothese“ Galloways) in Bezug auf seine institutionelle Realität bzw. seine Funktion in der gegenwärtigen Organisation des Wissens. Lyotard diagnostizierte einen Statuswechsel des Wissens, der zeitgleich mit der technologischen Transformation postindustrialisierter Gesellschaften einhergeht. In der posttechnisierten Gesellschaft „kann“ Wissen „die neuen Kanäle nur dann passieren und einsatzfähig gemacht werden, wenn die Erkenntnis in Informationsquantitäten übersetzt werden kann“ (Lyotard, S. 30). Wissen, welches diesen Übersetzungsprozeß nicht überlebt, also sich nicht „der Bedingung der Übersetzbarkeit etwaiger Ergebnisse in die Maschinensprache unterordnen“ kann, wird „vernachlässigt“: „Mit der Hegemonie der Informatik ist es eine bestimmte Logik, die sich durchsetzt, und daher auch ein Gefüge von Präskriptionen über die als ‚zum Wissen‘ gehörig akzeptierten Aussagen gegeben.“ (Lyotard, S. 31; Hervorhebung von uns)

Dabei spielt der Begriff der Information für unser Vorhaben insofern eine besondere Rolle, als dass er einen Anker für den Textbegriff in den digitalen Textwissenschaften darstellt. Zweifelsohne ist die strukturelle Rückführbarkeit des Textbegriffs auf den Informationsbegriff nicht ohne Folgen etwa für die Funktion der Wissensorganisation an Hochschulen wie auch an Bibliotheken und Erinnerungsinstitutionen. Ein besonders deutliches Beispiel für diese Transformation des Textbegriffs ist die Förderung wissenschaftlicher Projekte, welche (zurecht!) ein Erschließungsprojekt – und zunehmend auch großflächige Textanalysen – ohne informatisch adäquate Rahmenbedingungen (Erschließungsstandards, Datenmanagement, Langzeitarchivierung) geradezu undenkbar macht. Insofern funktioniert der Begriff der Information also als ein theoretischer Dispositiv (Foucault) und führt uns so zu genuin epistemologischen bzw. wissenschaftstheoretische Fragestellungen.

Epistemologisch führt die Bestimmung eines Textes in Rekurs auf den Begriff der Information dazu, dass jegliches, modellhaft vermitteltes, Wissen in der Form der Information erscheint. Dieses ist von beeindruckender Produktivität bzw. zeichnet sich aus durch eine „Handlichkeit“ (digitale Datenverarbeitung), evakuiert allerdings so etwas wie das Subjekt der Handlung. Insofern liegt darin ein strukturelles Element der Entfremdung, das sich zurückführen lässt auf die Einführung eines im weitesten Sinne mathematischen Paradigmas in den Geisteswissenschaften. Dieses Paradigma aber all zu stark gegen das „klassische“ geisteswissenschaftliche auszuspielen setzt sich einem Ideologieverdacht aus, dem wir uns schon zu Beginn erwert haben. Worum es uns, jetzt wissenschaftstheoretisch, geht, ist einige Effekte zu benennen, die aus der Anwendung des Paradigmas in den Geisteswissenschaften folgen.

Dabei ist es bemerkenswert, dass auch diejenigen Geisteswissenschaften, die sich mit der Kunst beschäftigen zunehmend auf die Wissenschaftstheorie positivistischer Prägung rekurrieren. Die Vermittlung des Wissens durch ein Modell führt aber dazu, dass die Sache, die damit erklärt werden soll, gerade nicht mehr erfahren werden muss. Wendet man diese Form der Vermittlung auf genuine Gegenstände der Geisteswissenschaft an, betrachtet man sie als Gegenstände, die zur Ordnung der Natur gehören, also im Grunde als unmittelbare. Das Gegenteil ist aber der Fall, sie sind höchstvermittelt, d.h. sie sind Singulare.

Bibliographie

Berry, David M. (2011): *The Philosophy of Software: Code and Mediation in the Digital Age*, Houndmills & New York.

Berry, David M. (2014): *Critical Theory and the Digital*, New York & London.

Dilthey, Wilhelm (1883): *Einleitung in die Geisteswissenschaften: Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschichte*, Leipzig.

Evens, Aden (2015): *Logic of the Digital*, London u.a..

Foucault, Michel (1978): *Dispositive der Macht: Über Sexualität, Wissen und Wahrheit*, Übers. Jutta Kranz et al., Berlin.

Galloway, Alexander R. (2012): *The Interface Effect*, Cambridge & Malden.

Galloway, Alexander R. (2014): “The Cybernetic Hypothesis”, in: *differences* 25: 107-131.

Lyotard, Jean-François (2012): *Das postmoderne Wissen: Ein Bericht*, 7., überarb. Aufl., Wien.

Moretti, Franco (2009): *Kurven, Karten, Bäume: Abstrakte Modelle für die Literaturgeschichte*, übers. Florian Kessler, Frankfurt am Main.

McCarty, Willard (2005): *Humanities Computing*, Houndmills & New York.

Ramsay, Stephen (2011): *Reading Machines: Towards an Algorithmic Criticism*, Urbana & Springfield.

Sahle, Patrick (2013): *Digitale Editionsformen: Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, 3 Bde., Norderstedt.

Jockers, Matthew L. (2013): *Macroanalysis: Digital Methods & Literary History*, Urbana & Springfield.

Kulturelle Evolution. Zur Kritik der literaturhistorischen Methode

Lauer, Gerhard

gerhard.lauer@unibas.ch
Universität Basel, Schweiz

Zu den Provokationen der Literaturwissenschaft durch die Digital Humanities gehört ihre Arbeit an großen Datenmengen. Nicht das besondere Buch, der Kanon oder der Großschriftsteller, sondern die vielen Bücher und Literaturen sind der ‚andere' Gegenstand der computergestützten Literaturwissenschaft. Geradezu typisiert werden Digital Humanities und Distant Reading zusammen genannt. Tatsächlich war (Moretti, 2000) als Kritik an der herkömmlichen Literaturwissenschaft angelegt, genauer an ihrer methodischen Beschränkung, die Vielfalt der Literaturen methodisch in den Griff zu bekommen. Die Kritiker Morettis haben sein Anliegen konzediert, ohne konkrete Vorschläge zu machen, wie mit dem ‚Mengenproblem' in der Literaturgeschichtsschreibung besser umgegangen werden könnte (Ross, 2014). Ein radikaler Ansatz, die Geschichte der Literatur anders als bisher zu modellieren, ist der Ansatz der kulturellen Evolution (Lewens, 2013). Meine These lautet: Evolutionäre Modelle und Theorien sind für die Beantwortung literaturhistorische Fragestellung geeignet, besonders um das ‚Mengenproblem' in den Griff zu bekommen. Im Folgenden skizziere ich Theorie und Methodologie eines solchen Ansatzes und frage nach den Folgen für ein Fach wie die Literaturgeschichte.

Wie der Name schon andeutet, verschiebt der Ansatz den Akzent von der Geschichte auf die Evolution. In den Blick rückt nicht weniger als die Menschheitsgeschichte. Der Ansatz kommt denn auch aus der biologischen Anthropologie, nicht aus den geisteswissenschaftlichen Fächern. Das leitende Paradigma ist Darwins Theorie der Evolution. Die Theorie der kulturellen Evolution überträgt dieses Modell auf die Kulturgeschichte der Menschheit. Sie geht von der These aus, dass die Entwicklung der menschlichen Kulturen der gleichen evolutionären Entwicklungslogik folgt, der auch die Natur unterliegt (Mesoudi, 2016). Dabei geht es nicht um eine Analogie, vielmehr lautet die Dual-Heritance-These dieser Theorie, dass die Kultur die Natur des Menschen ist, kulturelle Entwicklungen die Natur des Menschen bis in seine genetische und biologische Veranlagung beeinflussen, wie umgekehrt die Naturgeschichte des Menschen seine Kultur bestimmt (Henrich & McElreath, 2007). So generell angelegt versteht sich kulturelle Evolution als eine Supertheorie, die verspricht, die Sozial- und Geisteswissenschaften mit den biologischen Wissenschaften zusammenzuführen (Mesoudi, Whiten & Laland, 2006; Laubichler & Renn, 2015). Aus der Sicht der Digital Humanities ist die Theorie der kulturellen Evolution daher am extremen Ende des Distant Reading angesiedelt. Statt das einzelne Buch versucht die Theorie die Kultur der Menschheit in den Blick der Untersuchung zu nehmen.

Kulturelle Evolution ist aber neben dem ‚Distant Reading' auch noch aus einem zweiten Grund von Interesse für Digital Humanities. Sie arbeitet mit computergestützten Modellen. Bereits die Arbeiten aus den 70er und 80er Jahren haben computer-basierte Modelle genutzt, als (Cavalli-Sforza & Marc Feldman, 1981), dann auch (Boyd & Richerson, 1985) begonnen haben, die Evolutionsbiologie für das Verstehen von kulturellen Prozessen zu nutzen. Ihr Modell der kulturellen Evolution geht von Populationen aus, die ihrerseits aus Gruppen von Individuen bestehen, von denen jedes Individuum über variierende kulturelle Eigenschaften verfügt. Soziale Transmission von Informationen ist der wesentliche kulturelle ‚Vererbungs'-Mechanismus zwischen Individuen und Populationen. Nur dann, wenn man annimmt, dass durch Prozesse des kulturellen Lernens Wissen vertikal, aber auch horizontal zwischen gleichzeitig lebenden Generationen weitergegeben wird, etwa das Lernen von Sprachen durch Kinder, kann man verstehen, warum sich die horizontale Weitergabe langfristig auch auf die vertikale, genetische Weitergabe von Eigenschaften auswirken kann und sich so etwas wie komplexe Sprachen entwickeln konnten. Hier kommen mathematische Ansätze und Com-

putersimulationen ins Spiel, um die langfristigen Veränderungen mikroevolutionärer Veränderungen in Populationen, die Rate der Ausbreitung und räumlichen Verteilung neuer kultureller Eigenschaften zu modellieren. Quantitative Ansätze waren für Cavalli-Sforza, Feldman, Boyd und Richerson trotz der damals noch bescheidenen Speicherraten notwendig geworden, weil die Prozesses, die Veränderungen in der kulturellen Variationen verursachen, so vielfältig sind, dass sie mit herkömmlichen Methoden nicht mehr gehandhabt werden konnten.

Die Linguistik hat die Theorie der kulturellen Evolution rasch adaptiert, um die Evolution der Sprache untersuchen zu können, wie etwa die Befunde zur Diversität von Sprachen, nämlich dass größere Populationen ein größeres Inventar an Wörtern besitzen, ihre Sprachen stärker grammatikalisiert sind, mehr Phoneme haben, aber ihre Morphologie zugleich einfacher ist und ihre Wörter kürzer sind (Atkinson, 2011; Nettle, 2012). Auch kann ein solcher Ansatz zeigen, wie sich wärmere Klimata auf das Klangspektrum der in einer Sprache genutzten Laute auswirken (Munroe, Fought & Macaulay, 2009) oder wie zerklüftete Landschaften bestimmte Distanzsprachen wie etwa Zeichen- oder Pfeifsprachen präferieren (Meyer, 2015). Immer geht es dabei um Einsichten in die Struktur der Kultur und die Prozesse ihrer Veränderung der langen Dauer, die nicht der Granularität etablierter historischer Beschreibungen entsprechen.

In der Literaturgeschichte sind kulturevolutionäre Modelle nicht etabliert. Das hat zunächst damit zu tun, dass generell der Aufbau von Korpora und die formale Modellierung von Fragestellungen innerhalb des historisch-hermeneutischen Paradigmas keine Rolle spielen und kaum eine Tradition haben. Linguisten dagegen wie (Labov, 1963) haben variationstheoretische und funktionalistischen Methoden genutzt, um die Systematizität in der sozialen und individuellen Variation des Sprachgebrauchs zu verstehen. Neuere Arbeiten in der Linguistik fragen danach, ob selbst solche fundamentalen Unterscheidungen wie die Unterscheidung der Wortklassen Nomen und Verben nicht in allen Sprachen zu finden sein könnte, die Wortordnung viel variabler als bislang angenommen sein dürfte, sprachliche Register höchst unterschiedlich gebraucht werden, Sprachen in einer höchst unterschiedlichen Interaktion zwischen Kindern und Eltern erworben werden (Evans, 2013; Lieven, 2013). Wenn aber Sprache in ihrer Entwicklungsgeschichte selbst in ihren Grundkategorien diverser sein könnten, als lange angenommen, und die Evolutionsrate ihrerseits je nach Sprache und Umwelt stark zu variieren scheint (Gray et al., 2013), müsste nicht Ähnliches auch für die Literaturen der Welt und ihre Entwicklung gelten, so dass man annehmen könnte, dass die Rate der Evolution von literarischen Formen mit Faktoren wie Gruppengröße, Dichte des sozialen Netzwerks, Menge der geteilten Informationen, soziale Stabilität und das Niveau des Austausches mit anderen Gruppen korreliert (Trudgill, 2011)? Das wäre eine innovative Forschungsagenda.

Die wenigen literaturhistorischen Arbeiten, die bislang vorliegen, nutzen verschiedene statistische Modelle und Methoden, um eine kulturevolutionäre Literaturgeschichte zu untersuchen. Besonders naheliegend ist der Ansatz einer Weiterentwicklung von Stemma zur Rekonstruktion von Manuskript-Kulturen. 1998 druckte die Zeitschrift *Nature* einen kurzen Forschungsbericht über die phylogenetische Verwandtschaft der mehr als 80 überlieferten Manuskripte der *Canterbury Tales* ab (Barbrook et al., 1998). Als Äquivalent zum genetischen Code wurden Fehler in den Abschriften und deren Weitergabe als Maß der Ähnlichkeit verwendet und etablierte Verfahren der Kladistik-Analyse (PAUP 3.1.1.) für die Erstellung von Bäumen genutzt. Dieselbe Forschergruppe hat inzwischen weitere Untersuchungen zur Verwandtschaft von Manuskripten erarbeitet, die durch Tagging verschiedene Variationen in den Manuskripten wie Wortvarianten, Wortergänzungen, kleine Wortzusätze, fehlende Zeilen, Veränderungen im Reim u.a. in die Erstellung der auf SplitsTree-Verfahren beruhenden Analysen einbeziehen (Howe et al., 2001). In einer Zusammenarbeit von Biologen und Philologen wurden die Methoden inzwischen um verschiedenen statistische Ansätze wie Maximum Parsimony und NeighborNet u.a. erweitert und ermöglichen die Abhängigkeiten von Manuskripten von ihren Vorlagen und die Regionen der Veränderung genau zu ermitteln (Windram et al., 2008).

Ein weiterer kulturevolutionärer Ansatz nutzt die Klassifikation von Motiven, wie sie in der historisch-geographischen Schule der Märchenforschung mit dem Aarne-Uther-Thompson-Index vorliegt. Mit Methoden der Kladistik (Most Parsimonious Trees), Bayesian und phylogentischen Netzwerkanalysen konnten (2013; Tehrani & d'Huy, 2106) die weltweite Verwandtschaft des Rotkäppchen-Märchens bestimmen. Von Interesse sind dabei auch Zusammenhänge mit der Populationsstruktur (Ross, Greenhill & Atkinson, 2013), wenn dabei gezeigt werden kann, wie geographische Distanz, Genetik und Variation der Märchenmotive zusammenhängen (Bortolini et al., 2017). Die Annahme dabei ist, dass genetische, linguistische und motivliche Distanz korrelieren.

Statt über textuelle Merkmale wie Abschriftenfehler oder Motive zu gehen nutzen andere

Ansätze innerhalb dieses Paradigmas Spielexperimente, wie sie besonders in der Verhaltensökonomie gängig sind. So wählen Probanden aus mehr als 60 Geschichten diejenigen aus, die ihnen besonders erzählenswert erscheinen und schreiben eine der Geschichten in kurzen Abschnitten (120-160 Buchstaben) weiter, bevor anderen Probanden die Geschichte weiterschreiben. Jeder Proband kennt jeweils nur die ‚Eltern' der Geschichte, die sie gerade fortschreiben (Cuskley et al., 2016). Gemessen werden qualitativ und quantitativ narrative Innovationen, um so experimentell die Ausbreitung von Geschichten zu messen. Ein weiterer Ansatz nutzt das Konzept der ‚Minimally Counterintuitive Narratives', demzufolge Geschichten dann eher geteilt und weitergegeben werden, wenn sie eine realistische Ontologie leicht durchbrechen, wie das etwa bei Märchen der Fall ist (Porubanova-Norquist, Shaw & Xygalatas, 2013). Auch hier werden phylogenetische Methoden verwendet, um zu berechnen, welche Texteigenschaften in welcher Umwelt evolutionär vorteilhaft sind und daher eher geteilt und weitergegeben werden (Stubbersfield & Tehrani, 2013). Es liegt auf der Hand, dass noch sehr viele andere Dimensionen von (literarischen) Texten eine Rolle im Prozess der Evolution spielen dürften.

Systematisch gewendet heißt das: Eine kulturevolutionäre Literaturgeschichtsschreibung hat ein anderes Gegenstandsfeld, nicht den Viktorianischen Roman, sondern die evolutionäre Logik seiner Entwicklung und die Schreibmuster dieses Genres. Literatur ist kein Werk, sondern Teil von Populationen. Sie benutzt zweitens andere Methodensets als die herkömmliche Literaturgeschichte, aber teilweise auch andere als sie sonst in den Digital Humanities gängig sind. Drittens leistet der Ansatz anderes. Er ist auf Erkenntnisse zur Logik der langen Dauer ausgerichtet. Und viertens rückt der Ansatz die Literaturgeschichte nahe an die Biologie heran mit Folgen für die Theoriebildung und Methodenentwicklung. Die Theorie der kulturellen Evolution ist nicht weniger als ein Ansatz, Digital Humanities als Teil eines größeren Forschungsprogramms zu betreiben. Mein Vortrag ist ein Plädoyer für ein solches Forschungsprogramm einer Literaturgeschichte der langen Dauer.

Bibliographie

Atkinson, Quentin D. (2011): "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa", in: *Science* 332,6027: 346-349.

Barbrook, Adrian / Howe, Christopher / Blake, Norman / Robinson, Peter (1998): "The Phylogeny of *The Canterbury Tales*", in: *Nature* 394: 839, https://www.nature.com/nature/journal/v394/n6696/full/394839a0.html [letzter Zugriff 16. September 2017].

Bortolini, Eugenio / Pagani, Luca / Crema, Enrico / Sarno, Stefania / Barbieri, Chiara / Boattini, Alessio / Sazzini, Marco / Graça da Silva, Sara / Martini, Gessica / Metspalu, Mait / Pettener, Davide / Luiselli, Donata / Tehrani, Jamshid (2017): "Inferring Patterns of Folktale Diffusion Using Genomic Data", in: *Proceedings of the National Academy of Sciences of the United States of America* 114,34, http://www.pnas.org/content/114/34/9140 [letzter Zugriff 16. September 2017].

Boyd, Robert / Richerson, Peter (1985): *Culture and the Evolutionary Process*. Chicago.

Cavalli-Sforza, Luigi / Feldman, Marcus (1973): "Models for Cultural Inheritance. Within Group Variation", in: *Theoretical Population Biology* 4: 42-55.

Cuskley, Christine / Monechi, Bernardo / Gravino, Pietro / Loreto, Vittorio (2016): "The Evolution Of Collaborative Stories", in: S.G. Roberts / C. Cuskley / L. McCrohon / L. Barceló-Coblijn / O. Fehér / T. Verhoef (Eds.): *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, http://evolang.org/neworleans/papers/133.html [letzter Zugriff 16. September 2017].

Evans, Nicholas (2013): "Language Diversity as a Resource for Understanding Cultural Evolution", in: Richerson, Peter / Christiansen, Morten (Eds.) **:** *Cultural Evolution. Society, Technology, Language, and Religion*. Cambridge/Mass.

Gray, Russell / Greenhill, Simon / Atkinson, Quentin (2013): "Phylogenetic Models of Language Change", in: Richerson, Peter / Christiansen, Morten (Eds.) **:** *Cultural Evolution. Society, Technology, Language, and Religion*. Cambridge/Mass.

Henrich, Joseph / McElreath, Richard (2007): "Dual Inheritance Theory. The Evolution of Human Cultural Capacities and Cultural Evolution", in: Dunbar, Robin / Barrett, Louise (Eds.): *Oxford Handbook of Evolutionary Psychology*. Oxford, 555-570.

Howe, Christopher / Barbrook, Adrian / Spencer, Matthew / Mooney, Linne (2001): "Manuscript Evolution", in: *Endeavour* 25,3: 121-126.

Labov, William (1963): "The Social Motivation of a Sound Change", in: *Word* 19,3: 273-309.

Lewens, Tim (2013): "Cultural Evolution", in: *The Standford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL http://plato.stanford.edu/archives/spr2013/

entries/evolution-cultural/ [letzter Zugriff 16. September 2017].

Laubichler, Manfred / Renn, Jürgen (2015): "Extended Evolution. A Conceptual Framework for Integrating Regulatory Networks and Niche Construction", in: Journal of Experimental Zoology 324B: 565-577.

Lieven, Elena (2013): "Language Acquisition as a Cultural Process", in: Richerson, Peter / Christiansen, Morten (Eds.) **:** *Cultural Evolution. Society, Technology, Language, and Religion*. Cambridge/Mass.

Mesoudi, Alex / Whiten, Andrew / Laland, Kevin N. (2006): "Towards a Unified Science of Cultural Evolution", in: *Behavioural and Brain Sciences* 29: 329-383.

Mesoudi, Alex (2016): "Cultural Evolution. Integrating Psychology, Evolution and Culture", in: *Current Opinion in Psychology* 7: 17-22.

Meyer, Julien (2015): *Whistled Languages. A Worldwide Inquiry on Human Whistled Speech*. Berlin, Heidelberg.

Moretti, Franco (2000): "Conjectures on World Literature", in: *New Left Review* 1: 54-68.

Munroe, Robert / Fought, John / Macaulay, Ronald (2009): "Warm Climates and Sonority Classes not Simply More Vowels and Fewer Consonants", in: *Cross-Cultural Research* 43,2: 123-133.

Nettle, Daniel (2012): "Social Scale and Structural Complexity in Human Languages", in: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367,1597: 1829-1836.

Porubanova-Norquist, Michaela / Shaw, Daniel / Xygalatas, Dimitris (2013): "Minimal-Counterintuitivenes Revisited. Effects of Cultural and Ontological Violations on Concept Memorability", in: *Journal for the Cognitive Science of Religion* 1,2: 181-192 https://pure.au.dk/ws/files/71907173/Porubanova_et_al.pdf [letzter Zugriff 16. September 2017].

Ross, Robert / Greenhill, Simon / Atkinson, Quentin (2013): "Population structure and cultural geography of a folktale in Europe", in: *Proceedings of the Royal Society B: Biological Sciences* 280,1756, http://rspb.royalsocietypublishing.org/content/280/1756/20123065 [letzter Zugriff 16. September 2017].

Ross, Shawna (2014): "In Praise of Overstating the Case. A Review of Franco Moretti, *Distant Reading*", in: *Digital Humanities Quarterly* 8,1 http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html [letzter Zugriff 16. September 2017].

Stubbersfield, Joseph / Tehrani, Jamshid (2013): "Expect the Unexpected? Testing for Minimally Counterintuitive (MCI) Bias in the Transmission of Contemporary Legends: A Computational Phylogenetic Approach", in: *Social Science Computer Review* 31,1: 90-102.

Tehrani, Jamshid (2013): "The Phylogeny of Little Red Riding Hood", in: PLoS ONE 8,11: e78871, https://doi.org/10.1371/journal.pone.0078871 [letzter Zugriff 16. September 2017].

Tehrani, Jamshid / d'Huy, Julien (2016): "Phylogenetics Meets Folklore. Bioinformatics Approaches to the Study of International Folktales", in: Kenna, Ralph / MacCaroon, Márían / MacCarron, Pádraig (Eds.): *Maths Meets Myths. Quantiative Approaches to Ancient Narratives*. Zürich: 91-114, https://link.springer.com/chapter/10.1007%2F978-3-319-39445-9_6 [letzter Zugriff 16. September 2017].

Trudgill, Peter (2011): *Sociolinguistic Typology. Social Determinants of Linguistics Complexity*. Oxford.

Windram, Heather / Shaw, Prue / Robinson, Peter / Howe, Christopher (2008): "Dante's *Monarchia* as a Text Case for the Use of Phylogentic Methods in Stemmatic Analysis", in: *Digital Scholarship in the Humanities* 23,4: 443-463.

Lexikographie: Explizite und implizite Verortung in den Digital Humanities

Lindemann, David
david.lindemann@uni-hildesheim.de
Universität Hildesheim, Deutschland

Kliche, Fritz
fritz.kliche@uni-hildesheim.de
Universität Hildesheim, Deutschland

Kutzner, Kristin
kutzner@uni-hildesheim.de
Universität Hildesheim, Deutschland

Zusammenfassung

Ziel der hier vorgestellten Studie ist eine Beschreibung der Schnittmenge von Diskursräumen in der Lexikographie bzw. Metalexikographie und den Digital Humanities (DH). Dabei geht es um die Bestimmung von explizit bzw. implizit als Teil der DH aufzufassenden Beiträgen zu lexikographischen Themen und, andersherum, von lexikographierelevanten Themen, die in den DH

diskutiert werden. Zur Bestimmung der Diskursräume, von Schnitt- und disjunktiven Mengen, werden Volltexte und Metadaten analysiert, bibliometrische Netzwerke (Autoren- bzw. Zitationsnetzwerke) verglichen und Topic Modelings vorgenommen.

Einleitung

Der Einzug digitaler Methoden und Werkzeuge in die Geistes- und Sozialwissenschaften, genauer: der als „computational turn“ (Berry 2011) bezeichnete methodisch-epistemologische Quantensprung, lässt sich in allen Disziplinen der Humanities beobachten. In der Sprachwissenschaft hat sich dieser Wandel bekanntermaßen besonders deutlich in der Etablierung der Computerlinguistik als eigene Disziplin niedergeschlagen. Neben computerlinguistischen Verfahren der Textanalyse sind eine maschinenlesbare Wissensrepräsentation und -organisation, sind Formate für digitale Editionen und komputationell erstellte Visualisierungen heute in allen textbasierten Disziplinen in Gebrauch.

Der angesprochene Wandel lässt sich ebenfalls in der Lexikographie feststellen. Die Lexikographie bzw. Metalexikographie, als solche bereits seit geraumer Zeit als Disziplin emanzipiert (Tarp 2008; Wiegand 2013), haben den Übergang zum digitalen Medium inzwischen vollzogen (cf. zum frühen Stand der Dinge De Schryver 2003) und sind beständig dabei, ihr komputationell informiertes methodisches Instrumentarium weiterzuentwickeln (Heid 2013). Als zentrale Aspekte gelten hier der Einzug korpuslinguistischer Verfahren in die Lexikographie (Hanks 2008; Heid 2008), komputationelle Methoden zur Datenrepräsentation (Spohr 2012), speziell auch für die digitale Edition historischer Wörterbücher (Lemnitzer u. a. 2013) und zur Implementierung funktionsgerichteter Benutzerschnittstellen (Heid 2014) sowie zur Wörterbuchbenutzungsforschung (Müller-Spitzer 2014).

In der hier vorgestellten Studie gehen wir der Frage nach, wie sich der gemeinsame Diskursraum als Schnittmenge von Lexikographie und Digital Humanities mit quantitativen Methoden definieren lässt. Nach einer nicht exhaustiven und von Hand durchgeführten Voruntersuchung folgen wir der Ausgangshypothese, der gemeinsame Diskursraum sei um ein Vielfaches größer als man annehmen könnte, folgte man allein denjenigen Themen, die als lexikographierelevant gelten können und die in Publikationen diskutiert werden, die explizit zum Bereich der DH gehören (vgl. Abb. 1).

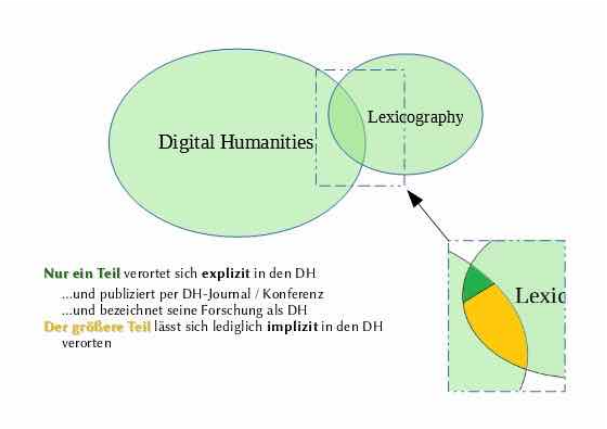


Abb. 1: Ausgangshypothese: Explizite und implizite Schnittmengen.

Diejenigen Arbeiten, die im DH-Kontext veröffentlicht werden und explizit einem Thema der Lexikographie zugeordnet werden, sind recht leicht über relevante Schlüsselwörter bestimmbar. Dazu tritt die Gruppe jener Publikationen, die zur eingangs skizzierten Schnittmenge zu zählen sind, ohne dass sie sich selbst ausdrücklich den Digital Humanities zuordnen. Es ist das Ziel dieser Untersuchung, zu bestimmen, welche in der Lexikographie diskutierten Themen und welche Autoren zu dieser Gruppe gerechnet werden und also eine Zurechnung zu den Digital Humanities *implizieren* können.

Voruntersuchung und Zwischenergebnis: Explizite Verortung in den DH

Als Voruntersuchung zum benannten Gegenstand haben wir eine Recherche in den Archiven bedeutender englischsprachiger Zeitschriften der Digital Humanities [1] sowie in den Proceedings der ADHO-Jahreskonferenzen [2] durchgeführt. Über die Suchbegriffe „Lexicography“ und „Dictionary“ finden sich in den genannten Archiven 31 englischsprachige Beiträge, die sich mit lexikographischen Themen befassen, und die sich qua Erscheinen in DH-Medien zu denjenigen Publikationen zählen lassen, die sich explizit in den Digital Humanities verorten.

Eine manuelle Zuordnung lexikographierelevanter Schlüsselwörter zu den genannten 31 Beiträgen ergibt das in Tabelle 1 wiedergegebene Bild; dabei sind mehrfache Zuordnungen möglich. Zunächst lässt sich ohne Verwunderung feststellen, dass in allen Beiträgen die digitale Repräsentation lexikalischer Daten eine Rolle spielt,

allerdings mit unterschiedlichen Fragestellungen, Herangehensweisen und Zielsetzungen. Die drei größten Themencluster haben wir hier, in dieser Reihenfolge, mit den Schlagwörtern „e-Wörterbücher / Visualisierung lexikalischer Daten“, „Historische Lexikographie“ und „Korpuslinguistik“ bezeichnet. Ersteres benennt Fragen der Produktion digitaler Wörterbücher einschließlich neuer Methoden der Visualisierung, letzteres die Erstellung und Nutzung elektronischer Textkorpora zu einer Reihe lexikographischer Zwecke. Beide Bereiche sind durch die Heraufkunft digitaler Methoden überhaupt erst möglich geworden und haben die Lexikographie revolutioniert. Die Historische Lexikographie kann als philologische Disziplin gelten, die sich mit der Edition historischer lexikalischer Datensammlungen befasst; die diesem Schlüsselwort zugeordneten Beiträge befassen sich grundsätzlich mit Methoden digitaler Edition, einem Kernbereich der DH.

Schlüsselwort (Topic)	Zählung
Digitale Wissensrepräsentation / Formate	31
e-Wörterbücher / Visualisierung lexikalischer Daten	14
Historische Lexikographie	10
Korpuslinguistik	9
Wörterbuchnetz	4
NLP-Lexicon	2
Bilingual Dictionary Drafting	2
Autorenwörterbuch	2
Dialektologie	1

Tabelle 1: Schlüsselwörter, manuelles Clustering, manuelle Zählung

Unter den weniger häufig gewählten Schlüsselwörtern sticht das „Wörterbuchnetz“ hervor, das Strategien zur Vernetzung lexikalischer Ressourcen bezeichnet. Hinzu kommen noch lexikalische Datensammlungen zur Anwendung in der maschinellen Sprachverarbeitung („NLP-Lexicon“), Methoden zum Entwurf zweisprachiger Wörterbuchinhalte („Bilingual Dictionary Drafting“), das „Autorenwörterbuch“, also Extraktionen aus Korpora, die aus dem Schaffen jeweils einer Literatin oder eines Literaten bestehen, sowie in einem Fall dialektologische Forschung mit digitalen Methoden.

Methode für die Bestimmung implizit in den DH verorteter Arbeiten

Wir haben ein Textkorpus erstellt, das im Zeitraum 2000 bis zur Gegenwart (2018) erschienene englischsprachige Beiträge aus Zeitschriften, Kongressakten und Handbüchern zu den Digital Humanities (Subkorpus DH) und der Lexikographie (Subkorpus Lexicog) enthält; Tabelle 2 gibt die Titel der verarbeiteten Quellen wieder. Dabei wurden die ausgewählten Zeitschriften bzw. Sammelbände jeweils vollständig berücksichtigt; die Beiträge wurden zusammen mit Metadaten, u. a. Verfasser (Name und Affiliation), Datum, Textsorte, Umfang und Identifier (ISBN, DOI), im Tool Zotero [3] verwaltet. Die Volltexte wurden semiautomatisch bereinigt und zusammen mit Metadatensätzen in das Korpus aufgenommen. Darüber hinaus wurden die in den Volltexten enthaltenen bibliographischen Referenzen extrahiert (GROBID, Lopez 2009).

DH / 1.422 (41%)	Digital Humanities Quarterly: http://www.digitalhumanities.org/dhq / 284
	DSH (ex LLC): https://academic.oup.com/dsh / 886
	TEI Journal of the Text Encoding Initiative: http://jtei.revues.org / 63
	Digital Studies/Le champ numerique: https://www.digitalstudies.org/ / 152
	Blackwell Companion to DH: Schreibman et al. (ed.) 2004 / 37
Lexikog / 2.056 (59%)	IJL: http://ijl.oxfordjournals.org/ 282
	Lexikos: http://lexikos.journals.ac.za/pub / 376
	Dictionaries (Journal of the DSNA): https://muse.jhu.edu/journal/540 / 257
	Euralex: https://euralex.org/publications/ / 782
	eLex: https://elex.link/ / 202
	HSK 5/4: Gouws et al. (ed.) 2013 / 110
	The Routledge Handbook of Lexicography: Fuertes-Olivera (ed.) 2018 / 47

Tabelle 2: Quellen für das DH/Lexikog Textkorpus / Zahl der Volltexte

Topic Modeling

Unüberwachtes Topic Modeling (LDA, eingesetztes Tool: MALLET (McCallum 2002)) soll es uns ermöglichen, die in Abb. 1 grob skizzierten Mengen als sich überschneidende Diskursräume zu bestimmen und zu visualisieren. Unsere Ergebnisse zeigen die relative Relevanz in beiden Subkorpora von 50 durch den LDA-Algorithmus bestimmten, jeweils mit einer Reihe von Schlüssel-Tokens repräsentierten Topics. Eine Reihe von Anhaltspunkten spricht für das zuverlässige Funktionieren der Methode: Die Liste der Topics, die besonders DH-relevant seien, wird von den Tokens „digital humanities computing tools“ angeführt, die Liste der Lexikographie-Topics von „dictionary dictionaries english words word learners language“.

Bei der Ansicht der im Mittelfeld befindlichen Topics, also Themen, die in beiden Subcorpora als relevant bezeichnet sind, stellt sich heraus, dass sich hier nicht nur der digitale Wandel als Thema widerspiegelt, sondern dass darüber hinaus weitere Topics den gemeinsamen Diskursraum von Lexikographie und DH ausmachen. Inmitten von Zeilen, die, quasi erwartungsgemäß, Tokens wie „information model data structure process analysis“ oder „corpus words frequency texts word corpora table“ sowie eine ganze Reihe von Namen natürlicher Sprachen enthalten, ist hier etwa das mit den Tokens „women male female gender woman man people [...] black [...] girl feminist [...]“ repräsentierte Topic auffällig (40 der 100 für dieses Topic relevantesten Beiträge stammen aus dem DH-, 60 aus dem Lexikog-Subkorpus).

Abbildung 2 zeigt für 50 von MALLET bestimmte Topics (Spalten) die Verteilung der 100 jeweils relevantesten Texte über die Subcorpora (Ausgabe von MALLET; DH-Beiträge sind grün, Lexicog-Beiträge violett unterlegt). Es wird deutlich, dass ein Teil der Topics eindeutig einem der Subkorpora zuzurechnen ist, andere Topics dagegen eine starke Durchmischung aufweisen.

Abb. 2: Visualisierung des Topic Modeling

Citation Network

Für alle Artikel des Korpus (siehe Tabelle 2) untersuchten wir die Anzahl der auf Items innerhalb des Netzwerks gerichteten Zitationen. Es zeigten sich 2.431 Zitationen (31% DH, 69% Lexicog). Die Zitationen aus DH sind nur zu 2% auf items aus Lexicog gerichtet; die Zitationen aus Lexicog zu 1% auf items aus DH.

Ergebnisse und Schlussfolgerungen

Die vorgestellten korpuslinguistischen und bibliometrischen Untersuchungen bieten wie beschrieben Aufschluss über Schnitt- und disjunkte Mengen von Themen- und Autorenclustern der Lexikographie und der Digital Humanities. Visualisierungen dieser Cluster und Listen der relevanten Keywords und Autoren werden bereitgestellt. Topic Modeling und Zitationsnetzwerke bilden unterschiedlich große Schnittmengen zwischen beiden Disziplinen ab: Während einerseits deutlich wird, dass eine ganze Reihe von Themen in beiden Disziplinen relevant ist, zitiert man sich vergleichsweise selten gegenseitig.

Die gezeigten Ergebnisse können zunächst zu einer verbesserten gegenseitigen Wahrnehmung in Lexikographie und Digital Humanities beitragen sowie in der lexikographischen Community das Bewusstsein dafür stärken, ein Gutteil der Disziplin gehöre durch die inhaltliche Überschneidung de facto zum Einflussbereich der Digital Humanities. Dies wiederum kann in der Zukunft zu einer stärkeren expliziten Verortung relevanter lexikographischer Beiträge in den Digital Humanities führen.

Weiterhin haben wir mit den für diese Studie durchgeführten Arbeiten eine annotierte bibliographische Datensammlung angelegt und die dazugehörigen Volltexte mit korpuslinguistischen Methoden annotiert und analysiert. Wir beabsichtigen, diese Sammlung auch weiterhin zu pflegen und öffentlich zugänglich zu machen.

Fußnoten

1. Digital Humanities Quarterly, DSH, TEI Journal
2. http://adho.org
3. http://www.zotero.org

Bibliographie

Berry, David M. (2011): „The Computational Turn: Thinking About the Digital Humanities“. (The Computational Turn). In *Culture Machine* 12 (0).

De Schryver, Gilles-Maurice (2003): „Lexicographers' Dreams in the Electronic-Dictionary Age“. In *International Journal of Lexicography* 16 (2): 143–199.

Fuertes-Olivera, Pedro (ed.) (2018): *The Routledge Handbook of Lexicography*. London: Routledge.

Gouws, Rufus / Heid, Ulrich / Schweickard, Wolfgang / Wiegand, Herbert E. (eds.). (2013): *Dictionaries. An International Encyclopedia of Lexicography*. HSK 5/4. Berlin / Boston: De Gruyter Mouton.

Hanks, Patrick (2008). „The Lexicographical Legacy of John Sinclair“. In *International Journal of Lexicography* 21 (3): 219–229.

Heid, Ulrich (2008). „Corpus linguistics and lexicography“. In Anke Lüdeling / Merja Kytö (ed.) *Corpus Linguistics. An international Handbook*: 131–153. Berlin: Mouton de Gruyter.

Heid, Ulrich (2013): „The impact of computational lexicography“. In Gouws et al. 2013: 24–30.

Heid, Ulrich (2014): „Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries“. In *Proceedings of EURALEX 2012*: 47–61.

Lemnitzer, Lothar / Romary, Laurent / Witt, Andreas (2013): „Representing human and machine dictionaries in markup languages (SGML, XML)“. In Gouws et al. 2013: 1195–1209.

Lopez, Patrice (2009): „GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications“. In *Research and Advanced Technology for Digital Libraries*: 473–474. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu/.

Müller-Spitzer, Carolin (2014): „Methoden der Wörterbuchbenutzungsforschung“. In *Lexicographica* 30 (1): 112–151.

Schreibman, Susan / Siemens, Ray / Unsworth, John (2004): *A Companion to Digital Humanities*. Oxford: Blackwell

Spohr, Dennis (2012): *Towards a Multifunctional Lexical Resource, Design and Implementation of a Graph-Based Lexicon Model*. Lexicographica Series Maior, 141. Berlin / Boston: De Gruyter.

Tarp, Sven (2008): *Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography*. Lexicographica, Series Mayor 134. Tübingen: Niemeyer.

Wiegand, Herbert E. (2013): „Lexikographie und Angewandte Linguistik“. In *Zeitschrift für angewandte Linguistik* 58 (1): 13–39.

Liebe und Tod in der Deutschen Nationalbibliothek Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft

Fischer, Frank
ffischer@hse.ru
Higher School of Economics, Moskau

Jäschke, Robert
r.jaschke@sheffield.ac.uk
Humboldt-Universität, Berlin

Einleitung

Der Sammelauftrag der Deutschen Nationalbibliothek (DNB) beginnt 1913 und bezieht sich auf »lückenlos alle deutschen und deutschsprachigen Publikationen« (»Wir über uns«, 16.03.2017). Der DNB-Katalog ist natürlich längst digitalisiert und die Arbeit mit ihm mittlerweile sehr komfortabel, da der Datendienst der DNB unter http://www.dnb.de/datendienst vierteljährlich einen Komplettabzug der Katalogdaten im RDF-Format bereitstellt, unter der freien Lizenz CC0 1.0. Momentan (Stand vom 23.06.2017) enthält er 14 102 309 Datensätze, also Metadaten zu von der DNB gesammelten Medien. Bisher gibt es aus geisteswissenschaftlicher Sicht nur wenige Versuche, diese Quelle nutzbar zu machen (eine Ausnahme bilden etwa Häntzschel u. a. 2009). Wir präsentieren ein einfaches Framework, mit dem verschiedene Aspekte des DNB-Katalogs untersucht werden können, seine Entwicklung über die knapp 105 Jahre seit Bestehen der Nationalbibliothek (vgl. auch Schmidt 2017, der für die Library of Congress einen ähnlichen Ansatz vorgestellt hat). Wir konzentrieren uns dabei auf Romane als Untersuchungsobjekt, von denen in der DNB rund 180 000

als solche rubriziert sind (dies entspricht nicht der Gesamtanzahl an Romanen, denn Nachauflagen und Übersetzungen zählen dort mit hinein – außerdem fehlen auch einige Romane, da sie nicht entsprechend verschlagwortet worden sind. Dieser Vortrag ist methoden-, nicht vorderhand ergebniszentriert, wobei wir an zwei Anwendungsszenarien aus der Praxis der digitalen Literaturwissenschaft demonstrieren, wie Katalogmetadaten bei der Bearbeitung konkreter Forschungsfragen behilflich sein können bzw. diese überhaupt erst ermöglichen.

Beschreibung des Frameworks

Die Titeldaten der DNB werden in typischen Linked-Data-Formaten (RDF/XML, JSON-LD usw.) angeboten. Der übliche Ansatz mit solchen Daten zu arbeiten ist, diese in eine geeignete Datenbank (Triple-Store) einzuladen und Anfragen mit Hilfe der entsprechenden Anfragesprache (i. A. SPARQL) zu stellen. Prinzipiell sind auch andere Systeme (z. B. relationale Datenbank, Suchmaschine) geeignet. Dies ermöglicht sehr flexible Anfragen und die leichte Einbindung weiterer Datenquellen. Da die Größe der Daten (unkomprimiert ca. 21 GB) jedoch gewisse Anforderungen an die Hardware stellt und die Konfiguration und Optimierung der Datenbank aufwendig ist, haben wir uns für eine andere, kompakte und leichter nachzuvollziehende Lösung entschieden. Langfristiges Ziel ist jedoch die Bereitstellung einer fertig konfigurierten Arbeitsumgebung in Form eines Docker-Containers, in der die Daten in einer Datenbank ad hoc verfüg- und analysierbar sind.

Der Titeldatensatz ist mit 14 102 309 Datensätzen und 227 212 707 Tripeln (»Fakten«) sehr umfangreich und enthält neben Angaben zu Büchern auch Angaben zu weiteren Medientypen wie etwa Zeitschriften. Neben den üblichen Metadatenfeldern wie Titel und Erscheinungsjahr ist bei Buchobjekten meist auch die Seitenanzahl sowie das Format vermerkt. Ganz im Sinne von Linked Data werden viele Angaben mit Hilfe von standardisierten Vokabularien (z. B. Dublin Core oder Bibo) beschrieben und ermöglichen so die Verlinkung mit weiteren Datensätzen. Insbesondere ermöglicht die Angabe der Autor*innen durch die numerische Kennung aus der Gemeinsamen Normdatei (GND) die Verknüpfung der Daten mit Wikidata, der (zukünftig) hinter Wikipedia stehenden Faktendatenbank. Wikidata verwendet ein auf Linked Data basierendes Datenmodell und ermöglicht, ähnlich wie Wikipedia, jedermann das Hinzufügen und Bearbeiten von Daten. Neben Angaben zu Städten und Ländern (z. B. Fläche, Einwohnerzahl) sind in Wikidata auch Daten zu zahlreichen Persönlichkeiten gespeichert, etwa deren Namen, Geburtsdaten, Berufe, Werke und, falls vorhanden, GND-Kennung (als Beispiel sei auf die Seite zu Johann Wolfgang von Goethe verwiesen: https://www.wikidata.org/wiki/Q5879).

Unser Framework umfasst derzeit vier Schritte, die im Folgenden beschrieben werden:

Vorverarbeitung und Konvertierung der Daten von RDF/XML zu JSON (rdf2json.py)

RDF/XML wird von den üblichen Softwaretools im Allgemeinen nicht als Datenstrom verarbeitet, sondern im Hauptspeicher abgelegt und dann weiterverarbeitet. Aufgrund der Größe der Daten scheidet diese Möglichkeit aus. Da jedoch alle wesentlichen Daten zu einem Medium typischerweise innerhalb eines XML-Tags "rdf:Description" abgelegt sind, können wir die Daten auch mit Hilfe eines SAX-Parsers als XML verarbeiten. Wir extrahieren die für die Analyse wesentlichen Metadaten (z. B. dcterms:contributor, dcterms:language, dc:title, dcterms:extent, rdau:P60493) und speichern diese als JSON ab. JSON ist im Allgemeinen platzsparender als RDF/XML und kann leicht in Elasticsearch eingeladen werden, was ein geplanter nächster Schritt ist.

Extraktion von Daten zu Autoren aus Wikidata (WKD-Toolkit)

Unser Ziel ist die Anreicherung der Autorenangaben im DNB-Datensatz mit Informationen aus Wikidata, beispielsweise Geburtsdatum- und -ort, Beruf und Verweis auf einen etwa vorhandenen Artikel in Wikipedia. Da die Python-Softwarebibliothek zur Verarbeitung von Wikidata-Datensätzen veraltet ist, greifen wir auf das Javabasierte Wikidata Toolkit zurück. Nach Herunterladen des aktuell (14.08.2017) 16 GB großen komprimierten Wikidata-Datensatzes extrahieren wir in zwei Durchgängen zunächst alle Elemente mit einer GND-Kennung einschließlich ausgewählter Merkmale und ergänzen im zweiten Durchlauf die Werte der Merkmale. Das Ergebnis speichern wir im JSON-Format.

Normalisierung und Anreicherung der Daten (json2json.py)

Unser Python-Skript implementiert eine Pipeline, die alle in den vorherigen Schritten extrahierten Daten einliest und mit Hilfe der GND-Kennung verknüpft, Metadatenangaben (wie z. B. Seitenanzahlen) extrahiert, vereinfacht und normalisiert, Datensätze mit fehlenden Angaben filtert und schließlich die gewünschten Datenfelder spaltenbasiert ausgibt. Die Vereinfachung umfasst vor Allem das Entfernen von Namespace-Präfixen (etwa http://id.loc.gov/vocabulary/iso639-2/ bei der Angabe der Sprache); Seitenanzahlen werden mit Hilfe eines regulären Ausdrucks extrahiert, der die häufigsten Fälle abdeckt; Jahreszahlen ebenso; Verlagsnamen können mit Hilfe einer Normtabelle normiert werden (dies ist nötig, da die Schreibung dieser Namen innerhalb des Katalogs nicht standardisiert ist).

Analyse der Daten (Shell-Skripte und -Tools wie awk, sort, datamash, gnuplot, ...)

Die entstandenen Dateien im TSV-Format können mit den üblichen Unix-Kommandozeilen-Werkzeugen wie awk, sort, uniq etc. leicht verarbeitet und analysiert werden; Visualisierungen wurden mit gnuplot erzeugt. Alle Schritte sind im GitHub-Repository dokumentiert.

Zeitliche Entwicklung über 105 Jahre DNB

Abbildung 1 zeigt die zeitliche Verteilung einiger Subdatensätze des Katalogs. Von den etwa 14,1 Mio. Objekten im originalen DNB-Datensatz weisen etwa 8,3 Mio. extrahierbare Seitenanzahlen auf (59 %). Beschränken wir diese Anzahl auf ›Romane‹ (über das Datenfeld "rdau:P60493"), bleiben 353 498 übrig, von denen wiederum 316 518 Umfangsangaben aufweisen und 180 219 einen Verfasser, der mindestens einen Wikipedia-Eintrag (in egal welcher Sprache) besitzt. Dieses Datenset ist die Grundlage für die unten folgenden Anwendungsszenarien.

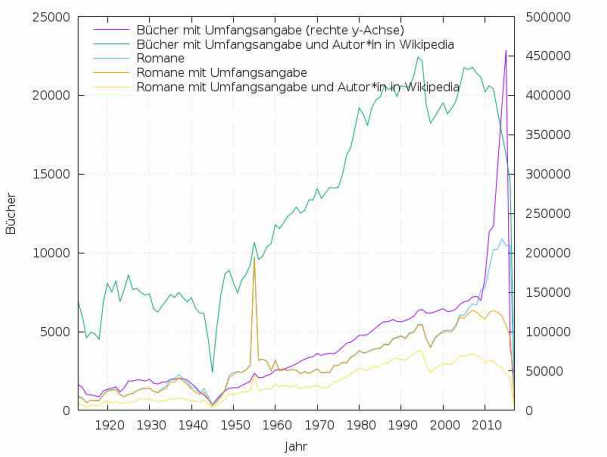


Abbildung 1: Fünf verschieden qualifizierte Subdatensätze des DNB-Katalogs in zeitlicher Verteilung.

Repräsentativität

Als möglicher Plausibilitäts- bzw. Repräsentativitätstest kann das Auszählen derjenigen Romanciers dienen, die mit den meisten Romanen im Katalog vertreten sind. Da der DNB-Katalog Vollständigkeit anstrebt, kann ein entsprechendes Ranking etwas über vergangene Realitäten auf dem deutschsprachigen Buchmarkt aussagen (Tab. 1), und tatsächlich stehen die Verfasser*innen von Romanbestsellern im Unterhaltungsbereich ganz oben (die Anzahl der Bücher umfasst von der DNB mitgesammelte Neuauflagen, Konsalik hat also nicht über 2 000 Romane geschrieben).

Autor*in	Romane
Heinz G. Konsalik	2232
Marie Louise Fischer	1264
Gert Fritz Unger	1013
Georges Simenon	783
Utta Danella	778
Edgar Wallace	654
Hedwig Courths-Mahler	647
Eleanor Hibbert	635
Pearl S. Buck	596
Alistair MacLean	582
Stephen King	577
Georgette Heyer	576
Agatha Christie	574
Theodor Fontane	565
Hans Ernst	563
Lion Feuchtwanger	501
Erich Maria Remarque	419
Hans Hellmut Kirst	411
Johannes Mario Simmel	403
Hans Fallada	396
Heinrich Mann	394
Fjodor Dostojewski	390
Barbara Cartland	390
Nora Roberts	381
Graham Greene	375
A. J. Cronin	370
Vicki Baum	366
Thomas Mann	359
Robert Ludlum	358
Gerd Hafner	357
Dean Koontz	354
Heinrich Böll	340
Alexandra Cordes	325
John le Carré	322
Marion Zimmer Bradley	321
Jason Dark	317
Willi Heinrich	313
Ludwig Ganghofer	311
Jack London	309
Joseph Roth	307
Danielle Steel	299
Johanna Lindsey	288
Erle Stanley Gardner	287
Siegfried Lenz	279
Jules Verne	277
Rosamunde Pilcher	274
Franz Kafka	271
Ernest Hemingway	271
Taylor Caldwell	269
Dorothy L. Sayers	269

Tabelle 1: Romanautor*innen geordnet nach Anzahl der Werke (inkl. Nachauflagen) im DNB-Katalog.

Anwendungsfall 1: Buchtitel

Die Verfügbarkeit großer digitalisierter Kataloge ermöglicht Large-Scale-Analysen bibliografischer Metadaten, etwa die Entwicklung von Romantiteln. Ein Vorläufer auf diesem Gebiet, Werner Bergengruens immer noch zu empfehlende Bibliothekarsfantasie »Titulus« von 1960, musste sich noch auf eine manuelle Sammlung des Autors stützen. Mittlerweile gibt es mit Franco Morettis Studie »Style Inc.« (2009) ein prominentes datengestütztes Beispiel (wobei sich Moretti bei seiner Analyse von um die 7 000 Romantiteln auf Fachbibliografien stützte, nicht auf Katalogdaten).

Um einen ersten Einblick in das Vokabular von Romantiteln zu bekommen, seien in Tabelle 2 die am häufigsten vorkommenden Substantive aufgelistet.

Substantiv	Frequenz
Liebe	3117
Mann	1906
Frau	1686
Tod	1537
Nacht	1505
Leben	1496
Welt	1188
Haus	1158
Zeit	1037
Schatten	1029

Tabelle 2: Häufigste Substantive in Romantiteln im gesamten DNB-Katalog.

Überzeitliche Konzepte – Liebe, Tod usw. – dominieren das Feld. Und nebenbei bemerkt: Ein wenig erinnert diese Liste an Jan Böhmermanns satirischen Song »Menschen, Leben, Tanzen, Welt«, mit dem auf die Beliebig- und Austauschbarkeit kontemporärer deutschsprachiger Liedproduktion angespielt wird (vgl. Pandzko/Böhmermann 2017), ein Befund, der sich analog auch auf Romantitel projizieren ließe.

Diese Anfragetechnik kann – wie beim Google Ngram Viewer – auf n-Gramme ausgedehnt werden, die Top-10 der häufigsten Trigramme findet sich in Tabelle 3.

Trigramm	Frequenz
Das Geheimnis der	238
Das Haus der	224
Der Mann der	189
Das Geheimnis des	175
Die Tochter des	160
Im Schatten des	128
Der Mann im	128
Das Lied der	125
Die Frau des	124
Die Reise nach	108

Tabelle 3: Häufigste Trigramme in Romantiteln im DNB-Katalog.

Ebenfalls analog zum Ngram Viewer lässt sich die zeitliche Entwicklung von n-Gramm-Frequenzen darstellen. Die unterschiedlichen Darstellungen in absoluten (Abb. 2) und relativen Zahlen (Abb. 3) kann etwa zeigen, dass sich zwischen Mitte der 1970er-Jahre und Mitte der 1990er-Jahre die Zahl an Romanen mit »Liebes«-Titeln zwar nahezu verdoppelt, dass sich diese Titel aber in relativen Zahlen nicht großartig vermehren.

Für genauere Analysen auf Grundlage dieser Extraktions- und Visualisierungsmethoden stellt das von uns vorgestellte Framework eine ideale Basis dar.

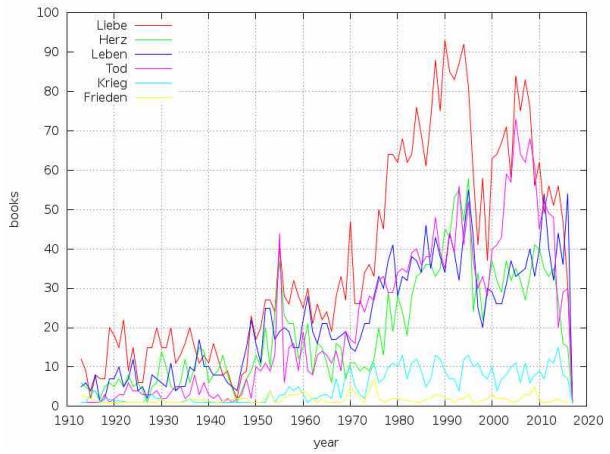


Abbildung 2: Vorkommen ausgewählter Wörter in Romantiteln im zeitlichen Verlauf (absolut).

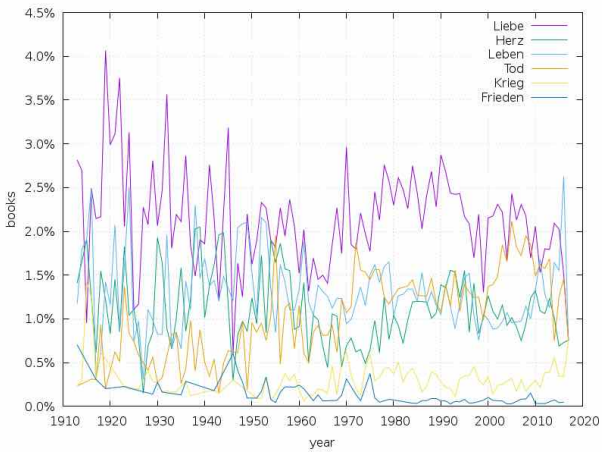


Abbildung 3: Vorkommen ausgewählter Wörter in Romantiteln im zeitlichen Verlauf (relativ).

Anwendungsfall 2: Textumfang

Unser zweites Anwendungsszenario betrifft die Erforschung des literarischen Textumfangs. Abbildung 4 zeigt die durchschnittliche Seitenanzahl von Romanen im Katalog der DNB.

Als Zuarbeit zu einer Theorie des literarischen Textumfangs haben wir mit dem von uns hier vorgestellten Framework in einer umfangreicheren Studie untersucht, wie sich der Umfang von Romanen etwa auf die Kanonbildung auswirkt (längere Romane, speziell solche von mehr als 1 000 Seiten Umfang, haben es leichter, in Kanonlisten zu landen). Außerdem ist es uns gelungen zu zeigen, wie umfangreiche Romane die DNA von Verlagen bestimmen können (vgl. Fischer/Jäschke 2018).

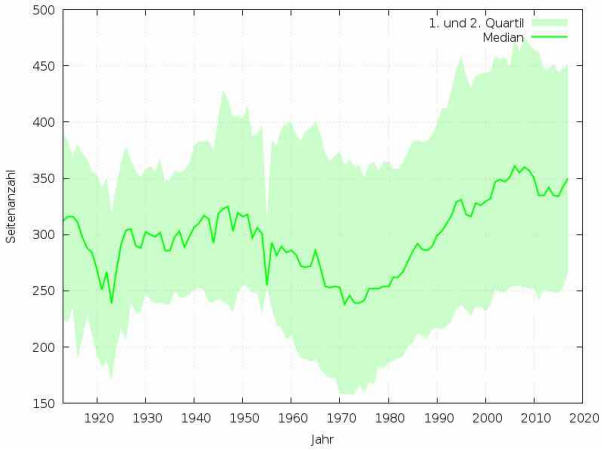


Abbildung 4: Entwicklung der mittleren Seitenanzahl pro Jahr seit 1913.

Fazit

Katalogdaten als Untersuchungsobjekt der quantifizierenden Literaturwissenschaften sind keine sich selbst erklärende Quelle, sondern ein über Jahrhunderte gewachsenes, überaus komplexes System. Die bibliothekarische Betreuung dieser Daten zielt nicht per se auf literaturwissenschaftliche Anwendungsfälle. Die Verschlagwortung kann lückenbehaftet sein, bestimmte Angaben wie etwa zum Textumfang können Fehler aufweisen. Die literaturwissenschaftliche Beschäftigung mit Katalogdaten setzt deren Explorier- und Kontrollierbarkeit voraus, wozu das hier vorgestellte Framework einen ersten Beitrag leisten soll. Zwei konkrete Anwendungsfälle sollten als Praxisbeispiele und ausdrücklich als Anreiz für weitere Szenarien dienen.

Bibliographie

Das **Arbeitsrepositorium** ist unter < https://github.com/weltliteratur/dnb > zu finden.

Bergengruen, Werner (1960): Titulus. Das ist: Miszellen, Kollektaneen u. fragmentar., mit gelegentl. Irrtümern durchsetzte Gedanken zur Naturgeschichte d. dt. Buchtitels oder unbetitelter Lebensroman e. Bibliotheksbeamten. Zürich: Verlag der Arche.

DNB (2017): »Wir über uns«, Stand 16.03.2017. URL: < http://www.dnb.de/DE/Wir/wir_node.html >.

Fischer, Frank; Jäschke, Robert (2018): Ein Quantum Literatur. Empirische Daten zu einer Theorie des literarischen Textumfangs. DFG-Symposium »Digitale Literaturwissenschaft«. Villa Vigoni, 9.–13. Oktober 2017. (Entsprechender Sammelband erscheint demnächst.)

Häntzschel, Günter; Hummel, Adrian; Zedler, Jörg (2009): Deutschsprachige Buchkultur der 1950er Jahre. Fiktionale Literatur in Quellen, Analysen und Interpretationen. Wiesbaden: Harrassowitz 2009. URL: < https://books.google.com/books?id=t88xc3CzK60C >.

Moretti, Franco (2009): Style Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). In: Critical Inquiry, Vol. 36, No. 1 (Autumn 2009), S. 134–158.

Pandzko, Jim ; Böhmermann, Jan (2017): Menschen Leben Tanzen Welt [Musikvideo]. In: Neo Magazin Royale, 05.04.2017. URL: < https://youtu.be/h8MVXC_hqNY >.

Schmidt, Ben (2017): A brief visual history of MARC cataloging at the Library of Congress. In: Sapping Attention [Blog], 16.05.2017. URL: < http://sappingattention.blogspot.de/2017/05/a-brief-visual-history-of-marc.html >.

Modellieren durch mediale Transformation: Das Theater Brechts in der virtuellen Realität

Wieners, Jan
jan.wieners@uni-koeln.de
Universität zu Köln, Deutschland

Schubert, Zoe
zoe.schubert@uni-koeln.de
Universität zu Köln, Deutschland

Eide, Øyvind
oeide@uni-koeln.de
Universität zu Köln, Deutschland

Einleitung

Die Konstruktion und Vorstellung einer artifiziell-medial vermittelten virtuellen Realität als Erweiterung, gar Gegenstück zur individuell psychisch- und physisch konstruierten Wahrnehmungswelt findet ihre umfangreiche narrative Beschreibung zunächst in Science-Fiction Erzählungen des ausgehenden zwanzigsten Jahrhunderts mit Konzepten wie Stanislaw Lems “Periphere Phantomatik” (Lem: 1981), William Gibsons “Cyberspace” (Gibson: 1995) oder virtueller “Metaversen” (Stephenson: 1992).

Wenige Jahrzehnte nach der vielversprechenden Darstellung virtueller Realitäten in der Mitte der 1990er Jahre lässt sich eine Begeisterung für sowohl kommerzielle VR-Lösungen wie Nintendos *Virtual Boy* (McFerran 2017) als auch geisteswissenschaftliche VR-Entwürfe konstatieren, um kulturelles Erbe innovativ erfahrbar zu machen. Gegenwärtige elaborierte Interaktionstechnologien wie *Oculus Rift* (Oculus VR 2018), *HTC Vive* (HTC Corporation 2018) oder *Google Cardboard* (Google 2018) machen virtuelle Welten erfahr- und interagierbar. Performante Spiel-Engines sowie Entwicklungsschnittstellen wie *Unity* (Unity Technologies 2018), *WebVR* (WebVR 2018), A-Frame (A-Frame 2018) und *ReactVR* (React VR 2018) stellen das leicht zugängliche Handwerkszeug bereit, um virtuelle Welten zum Leben zu erwecken und mit Inhalten zu füllen.

Fand die Weiterentwicklung von VR-Hardware und VR-Software in den vergangenen Jahren in gar schwindelerregender Geschwindigkeit statt, so steckt die theoretische Durchdringung virtueller Welten in ihren Kinderschuhen: Auf welche Art und Weise transformieren Anwendungen der virtuellen Realität Narrationen, Geschichte und Geschichten, um sie Anwenderinnen und Anwendern (virtuell) erfahrbar zu machen? Vermögen es virtuelle Welten, Inhalte und Fragestellungen digitaler Geisteswissenschaften adäquat bereitzustellen und zu formulieren? Welche Interaktions- und Wahrnehmungsmechanismen bieten virtuelle Welten?

Das Institut für Digital Humanities der Universität zu Köln führt seit mehreren Semestern Lehrveranstaltungen durch, in denen die Theorie um Medialität im Kontext virtueller Welten reflektiert und von Studierenden des vom Institut angebotenen Verbundsstudienganges Medieninformatik / Medienkulturwissenschaften praktisch erprobt wird. Ein zentrales Ziel ist es, dabei den Einfluss von Virtual Reality (VR) auf die Modellierung von Inhalten und Narrationen zu ergründen. Anhand der durchgeführten Lehrveranstaltungen offenbarten sich grundlegende Forschungsfragen für die digitalen Geisteswissenschaften in der Beschäftigung mit Modellierung und der Transformation geisteswissenschaftlicher Inhalte.

Die nachfolgenden Ausführungen geben zunächst einen kurzen Überblick über die durchgeführten Lehrveranstaltungen und die eingesetzten Technologien. Anschließend wird mit dem Brechtschen Theater der Gegenstand der Lehrveranstaltungen eigens vorgestellt und über besondere Herausforderungen bei der Transformation durch Implementation eines *virtuellen* epischen Theaters informiert. Schließlich wird der Prozess der medialen Transformation und der Modellierung reflektiert.

Virtual Reality und Digital Humanities - Ein Lehrangebot an der Universität zu Köln

Das Institut für Digital Humanities bot in der jüngsten Vergangenheit und aktuell eine Folge von unterschiedlichen Lehrveranstaltungen [1] an, die von den Studierenden äußerst positiv wahrgenommen, zahlreich belegt und durch praktische Umsetzungen von VR Anwendungen erfolgreich abgeschlossen wurden. Durch den Fokus auf zeitgemäße neue Frameworks wie dem im Folgenden vorgestellten WebVR-Framework *A-Frame* und der Nutzung von *Cardboards* wurde die Einstiegshürde für die Implementation eigener VR-Umgebungen reduziert.

Boten frühe Entwicklungen und Standards wie die 1994 vorgestellte *Virtual Reality Modeling Language (VRML)* (Techopedia Inc. 2018) oder ihr 2004 zum ISO-Standard erhobener Nachfolger *X3D* (Brutzman 2018) basale Möglichkeiten in der Umsetzung browserbasierter virtueller Welten, so intendiert das seit 2016 entwickelte und kostenfrei bereitgestellte WebVR-Framework *A-Frame*, den praktischen Zugang zu browserbasierter VR zu erleichtern, indem es die Nutzung neuartiger Browsertechnologien wie *WebVR* und *WebGL* (WebGL Public Wiki contributors 2018) vereinfacht. So sind einzig HTML- und JavaScript-Kenntnisse vonnöten, um erste basale VR-Umgebung zu realisieren.

Inhalte für virtuelle Realitäten

Die Vorstellung, durch das Erleben der Virtuellen Realität vollständig in eine andere Welt einzutauchen und diese als real zu empfinden, wird mit dem Begriff der Immersion beschrieben. Das Konzept der Immersion, wie es im Anwendungsfeld virtueller Welten definiert ist, unterscheidet sich dabei vom Film, der in der Lage ist, eine Szene durch die Aufnahme bewegter Bilder zu spiegeln und die Wirklichkeit für den Zuschauer abzubilden. [2] Der erfahrbare Raum der virtuellen Realität muss nicht notwendig der Erfahrungs- und Wahrnehmungswelt des Anwenders entsprechen. Er ergibt sich auch nicht unmittelbar aus einem möglichst genauen virtuellen Abbild der realen Welt, wenn er für den Nutzer ein immersives Erlebnis erzeugen soll. Eine starke immersive Erfahrung für den Nutzer muss nicht bedeuten, dass dieser den Eindruck gewinnt, sich an einem anderen Ort oder zu einem anderen Zeitpunkt in der bekannten Realität aufzuhalten. Es soll vielmehr der Eindruck entstehen in einer alternativen Realität präsent zu sein, die eben nicht der tatsächlichen Umgebung, der Welt der physischen Gegeben- und Anwesenheit entspricht und diese damit erfahrbar zu machen. Die Immersion ergibt sich einerseits durch Interaktionsmöglichkeiten der NutzerInnen mit der virtuellen Welt, andererseits aus einer vollständigen Beschlagnahmung all ihrer visuellen und gegebenenfalls auch akustischen Eindrücke.

Derzeitige VR Hard- und Software zeigt die Grenzen der Immersion durch VR schnell auf. Während beeindruckende Fortschritte gemacht wurden, bleibt das “Gehirn im Tank” vorerst ein nicht erlebbares Gedankenexperiment. Erlebbar

sind hingegen Simulationen, Rekonstruktionen, Demonstrationen in Form von Spielen und dreidimensionale Modelle, die für unterschiedliche Zwecke erstellt wurden. Derzeit werden viele unterschiedliche Anwendungskontexte erprobt und erschlossen. Eine zentrale Frage ist dabei: Welche Inhalte könnten sich noch für eine Virtuelle Darstellung eignen und was passiert, wenn Narration oder Geschichten übertragen werden sollen (vgl. Ryan 2015)?

Bemüht man an dieser Stelle erneut den Vergleich zum Film und betrachtet Produktionen des noch jungen Mediums zu Beginn des 20. Jahrhunderts, so finden sich Theaterverfilmungen, die belegen, dass massiv gekürzte Stücke als Stofflieferant herhalten mussten. Kino und Theater, beides Orte der darstellenden Kunst, treten im Zeitalter der technischen Reproduzierbarkeit in Konkurrenz zueinander. Sie grenzen sich dabei im Laufe der Zeit einerseits voneinander ab, können sich aber andererseits auch ergänzen, wie Theaterproduktionen zeigen, in denen Film eingesetzt wird. Der Hintergrund dafür war weniger die inhaltliche Übertragung (Theaterstücke als Stoff für das Kino) als die Tatsache, dass der Film damals, im Gegensatz zum Theater, Handlungen und Menschen in ihrer unmittelbaren, “wirklichen” Umgebung zeigen kann (Lang 2006).

Bertolt Brecht entwickelte in diesem Kontext des Theaters des 20. Jahrhunderts eine vollständige Theaterästhetik. In seinem Theater, das als episches Theater bekannt wurde, soll das Spiel im Theater immer als inszeniertes Spiel für den Zuschauer erkennbar sein. Die Wirklichkeit soll nicht als Illusion abgebildet werden, denn für Brecht sagt die einfache Wiedergabe der Realität weniger denn je über die Realität aus. Gerade das macht die Theaterstücke Brechts in Anbetracht der Immersion als Grundlage für Inhalte der virtuellen Realität interessant, da die Immersion nicht durch eine realistische Abbildung der realen Welt und Umgebung geschaffen werden soll. Die Fragestellung, was geschieht, wenn VR zeigt, dass etwas gezeigt wird, verspricht neue Erkenntnisse über diese Form der Darstellung selbst.

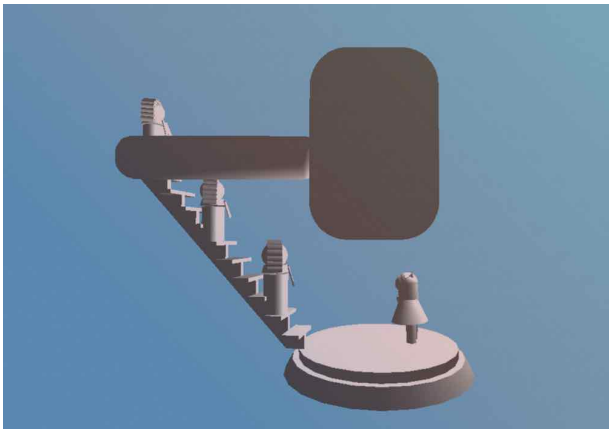
Die ästhetische Wirkung des Brechtschen Theaters ist augenscheinlich nicht immersiv. Allerdings kann ein Realitätsgefühl nur dann gebrochen werden, wenn es sich zuvor etabliert hat. Daraus ergibt sich die Frage, ob das Herstellen von Immersion möglicherweise auch in solchen Kunstrichtungen möglich erscheint. Die Überlegung dabei ist, dass Handlung und distanzierte Darbietung eine neue kritische Haltung des Zuschauers entstehen lassen können (Lang: 2006).

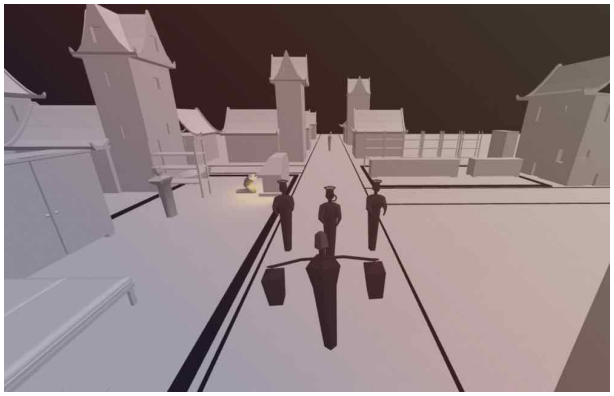
Anlässlich dieser Überlegungen und der geschilderten neuen Möglichkeiten durch die Entwicklungen der Technologie für VR wurden die Lehrveranstaltungen konzipiert. Dabei sollten von Teilnehmerinnen und Teilnehmern der Veranstaltungen VR-Anwendungen entwickelt werden, die Theaterstücke von Brecht in eine neue audiovisuelle virtuelle Welt übertragen. Die Beispiele, Erfahrungen und die Reflexion hinsichtlich der umgesetzten Projekte waren unterschiedlich und überraschend und sollen in diesem Vortrag genauer geschildert werden.

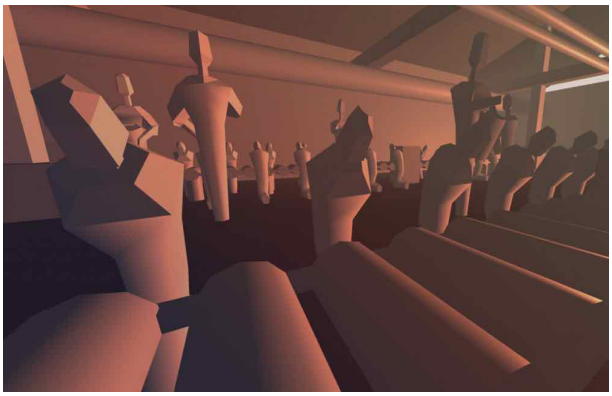
Fazit

Als Grundlage für die Modellierung der Projekte und für die Erkenntnisse, die sich aus den Lehrveranstaltungen ergaben, wurde zunächst die zentrale Frage gestellt, welche Interaktionsmöglichkeiten und Darstellungsmöglichkeiten sich mit und in virtuellen Realitäten bieten. Das Ziel war die Überprüfung der These, dass räumliche Illusionsbildung soweit gesteigert werden kann, dass Nutzerinnen und Nutzer mit dem virtuellen Raum interagieren. Anschließend wurden Kategorien Langs umfassender Studie zu Verfilmungen Brechtscher Theaterstücke betrachtet, um diese in einem weiteren Schritt nicht einzig auf Theater und Film zu beziehen, sondern um virtuelle Realität(en) zu erweitern. Die vorliegende Ausarbeitung intendiert demnach, die Gegenüberstellung der Prozesse des Vorführens vs. des Abbildens zu analysieren, indem adäquate Differenzierungen vorgenommen und untersucht werden:

- Ausgangssituation: live vs. vorproduziert
- Textgrundlage: aussprechen vs. zeigen
- Szenerie: funktional vs. illusionistisch
- Licht: Sichtbarkeit vs. Stimmung
- Bild: eigener Blick vs. festgelegte Einstellungsfolge
- Ton: unmittelbar vs. synthetisch
- Montage: sichtbarer Einschnitt vs. unsichtbarer Schnitt







Die Ergebnisse der Veranstaltung werden in diesem Vortrag aus zwei unterschiedlichen Perspektiven analysiert. Einerseits wird die Theorie der Modellierung angewendet, wie sie maßgeblich in den Digitalen Geisteswissenschaften entwickelt wurde, basierend auf einer semiotisch-analytischen Grundlage (Kralemann und Lattmann 2014; Ciula und Eide 2017). Andererseits wird das Konzept der „media transformations" (Elleström 2014) herangezogen, um zu analysieren, wie auf unterschiedliche Weise eine virtuelle Realität, basierend auf einem Theaterstück auf der Achse zwischen „media representation" und „transmediation" (Elleström 2014), verstanden werden kann. Im Kontext des Vortrags wird verdeutlicht, wie diese beiden zusammenwirken können, um zu zeigen, welche Prozesse wir in den Lehrveranstaltungen beobachten konnten.

Fußnoten

1. Visualisierung mit JavaScript", "Media Transformation I – Theater als Virtual Reality (VR) Erfahrung" und "Media Transformation II - Social VR and Interactive Storytelling in Virtual Reality", durchgeführt im Wintersemester 2016 /

2017 und im anschließenden Sommersemester 2017
2. Es existiert eine komplexe Beziehung zwischen Immersion und ästhetischen Konzepten unterschiedlicher künstlerischer Arbeiten, auf die in diesem Kontext nicht ausführlicher eingegangen werden kann.

Bibliographie

A-Frame (2018). “A-Frame - Make WebVR.” Mozilla Corporation. Letzter Zugriff Januar 12, 2018. *https://aframe.io/* .

Brecht, Bertolt (1989): “Der gute Mensch von Sezuan”, in: *Werke.* Große kommentierte Berliner und Frankfurter Ausgabe, herausgegeben von Werner Hecht, Jan Knopf, Werner Mittenzwei und Klaus-Detlef Müller, Band 6: Stücke 6, bearbeitet von Klaus-Detlef Müller. Frankfurt am Main: Suhrkamp Verlag: 175-279, 280 f.

Brutzman, Don (2018). “Applications, Players and Plugins for X3D / VRML Viewing”. Letzter Zugriff Januar 12, 2018. *http://www.web3d.org/x3d/content/examples/X3dResources.html* .

Ciula, Arianna / Eide, Øyvind (2017): “Modelling in Digital Humanities: Signs in Context”, in: *Digital Scholarship in the Humanities* 32: i33–i46. 10.1093/llc/fqw045

Elleström, Lars (2014): *Media Transformation. The Transfer of Media Characteristics among Media.* Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Gibson, William (1995): “Burning chrome”, in *“Burning chrome and other stories*”. London: Voyager. First published: *Omni*, 1982.

Google (2018). “Google Cardboard.” Google VR. Letzter Zugriff Januar 12, 2018. *https://vr.google.com/cardboard/* .

HTC Corporation (2018). “Vive”. Letzter Zugriff Januar 12, 2018. *https://www.vive.com/de/* .

Kralemann, Björn / Lattmann, Claas (2013): “Models as icons: modeling models in the semiotic framework of Peirce’s theory of signs.” *Synthese,* 190: 3397–3420.

Lang, Joachim (2006): “Episches Theater als Film”. Würzburg: Verlag Königshausen & Neumann GmbH, 2006.

Lem, Stanislaw (1981): “ *Summa technologieae*”. Berlin: Suhrkamp Verlag.

McFerran, Damien. 2017. “The Nintendo Virtual Boy Could Be Getting New Software.” nintendolife.com. Letzter Zugriff Januar 12, 2018. *http://www.nintendolife.com/news/2017/10/the_nintendo_virtual_boy_could_be_getting_new_software*.

Oculus VR, LLC (2018). “oculus rift”. Letzter Zugriff Januar 12, 2018. *https://www.oculus.com/rift/* .

React VR (2018). “React VR: Build VR websites and interactive 360 experiences with React.” Facebook. Letzter Zugriff Januar 12, 2018. *https://aframe.io/* .

Ryan, Marie-Laure (2015): *Narrative as Virtual Reality 2*. Baltimore: John Hopkins University Press.

Stephenson, Neal (1992): *Snow Crash.* New York: Bantam Books.

Techopedia Inc (2018). “Virtual Reality Modeling Language (VRML)”. Letzter Zugriff Januar 12, 2018. *https://www.techopedia.com/definition/4808/virtual-reality-modeling-language-vrml* .

Unity Technologies (2018). “Unity - Game Engine.” Letzter Zugriff Januar 12, 2018. *https://unity3d.com/de/* .

WebGL Public Wiki contributors (2018). "Getting Started," WebGL Public Wiki. Letzter Zugriff Januar 12, 2018. *http://www.khronos.org/webgl/wiki_1_15/index.php?title=Getting_Started&oldid=371* .

WebVR (2018). “Bringing Virtual Reality to the Web”. Letzter Zugriff Januar 12, 2018. *https://webvr.info/* .

Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora

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

Vertan, Cristina
fsha060@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

In den letzten Jahren wurden verschiedene Werkzeuge zur Annotation von Dokumenten entwickelt, z.B. *WebAnno* (Eckart de Castilho et al. 2016), *Cor A* (Bollmann et al. 2014) und *CATMA* (Meister et al. 2016). Diese Werkzeuge bieten zahlreiche Funktionalitäten, die für viele Sprachen und Anwendungen ausreichend sind. Die Unterstützung verschiedener Skripten erweckt den Ein-

druck, dass diese Werkzeuge für die Annotation aller Sprachkorpora erfolgreich eingesetzt werden können, was solange korrekt ist, wie man sehr flache Annotation durchführt.

Tiefere Annotationen, insbesondere für Sprachen außerhalb der Europäischen Sprachfamilie oder für historische Sprachen, dagegen gestalten sich problematischer. Häufig jedoch wird die Verwendung bereits etablierter Annotationstools trotzdem bevorzugt, weil diese eine reibungslose Integration mit weiteren Analyse- und Visualisierungs-Tools wie z.B. *ANNIS* (Krause & Zeldes 2016) versprechen. In diesem Beitrag werden wir zeigen, dass die Entwicklung dedizierter Annotationswerkzeuge dann als Lösung in Betracht gezogen werden kann, wenn gleichzeitig Schnittstellen (z.B. zu Analyse-Tools) entwickelt werden, die die Neuentwicklung wiederum an vorhandene Software und Infrastrukturen anbinden. Auf diese Weise können die Nachteile einer Neuentwicklung unter Gesichtspunkten der nachhaltigen Entwicklung von Forschungssoftware gegenüber der Anpassung bestehender Tools - bei gleichzeitiger Ermöglichung eines spezialisierten Anwendungsfalles - weitgehend abgeschwächt werden. Der Vorteil eines solchen Verfahrens ist die Realisierung eines Annotationsmodells, das exakt den Besonderheiten der Sprache entspricht. Insbesondere für diachrone Analysen ist es häufig nötig, dass man eine sehr detaillierte Modellierung der morphologischen, syntaktischen und semantischen Merkmale vornimmt, da die Unterschiede häufig nur im Detail reflektiert sind. Im Einzelnen werden wir die Entwicklung des *GeTa*-Annotationstools, eines Mehrebenenannotationswerkzeugs für die Altäthiopische Sprache, und dessen Integration mit dem Such- und Visualisierungsframework *ANNIS* darstellen.

Spezielle Anforderungen für die Altäthiopische Sprache

Das Altäthiopische (Geʿez), ist eine südsemitische Sprache. Es war bis ins 19. Jahrhundert hinein die bedeutendste Schriftsprache des christlichen Äthiopien. Die reiche christlich-äthiopische Literatur ist zunächst von Übersetzungen - anfangs aus dem Griechischen und später aus dem Arabischen - geprägt, bevor sich eine mannigfaltige indigene Literatur mit ganz eigenen Zügen entwickelt. Insbesondere Texte, die einzig im Altäthiopischen vollständig überliefert, und deren Textzeugen in anderen Sprachen entweder vollständig verloren, oder von denen nur Fragmente erhalten sind (z.B. das Henochbuch) erlangen dabei ganz besondere Bedeutung (Vertan et. al. 2016).

Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf; außerdem werden die Vokale vollständig geschrieben. Das äthiopische Silbenalphabet bringt dabei mit sich, dass Morphemgrenzen in der Schrift nicht darstellbar sind, sodass beispielsweise ein einzelner Vokal als Bestandteil einer Silbe eine eigenständige Wortart darstellen kann und tokenisiert werden muss; z.B. ist im zweisilbigen Wort ቤቱ: /be·tu/ das /u/ als pronominales Suffix zu bet- *u* (*sein* Haus) zu segmentieren. Eine solche Annotation kann nur auf der Transkription erfolgen. Annotationen auf anderen Ebenen (z.B. Seiten- Spaltenumbrüche, Textkorrekturen) müssen auf dem Fidal-Skript realisiert werden. Dies bedeutet, dass Original und Transkription synchronisiert im Annotationswerkzeug dargestellt werden müssen. Eine korrekte Darstellung einer Transliteration benötigt die Transkription. Während diese regelbasiert automatisch durchgeführt werden kann, ist korrekte Transliteration (z.B. Konsonantendopplung) in vielen Fällen nur zusammen mit morphologischer Analyse möglich. Daher muss das Annotationstool Korrekturen am Basistext während der Annotation zulassen und zuverlässig verarbeiten können. Keins der existierenden Annotationswerkzeuge erfüllt diese Voraussetzungen. CorA arbeitet mit Listen von Wertkombinationen von Attributen. Die benötigte morphologische Annotation im Fall der altäthiopischen Sprache umfasst 30 Merkmale, mit je mindestens drei möglichen Werten. Die Handhabung solch einer umfangreichen Liste macht manuelle Annotation praktisch unmöglich. Weder WebAnno noch CATMA ermöglichen die Korrektur des zugrunde liegenden Textes während der Annotation. Die Implementierung einer solchen Funktion auf Basis eines dieser Werkzeuge würde tief in die Architektur der Software eingreifen, was innerhalb der Laufzeit des Projektes nicht realisierbar war. Auch die semi-automatischen Annotationsmöglichkeiten unter Supervision des Benutzenden sind für diese beiden Werkzeuge nicht leicht erweiterbar.

Das GeTa Annotationstool

Im TraCES-Projekt [1], das als Hauptziel die Entwicklung eines diachronen, tief annotierten Korpus für die Altäthiopische Sprache hat, implementieren wir eine neuartige Architektur, die sowohl

Änderungen im Text als auch eine Mehrebenenannotation ermöglicht.

Wir betrachten als Grundtext den Originaltext in der altäthiopischen Schrift. Die Transliteration bildet die erste, die morphologische Annotation die zweite Ebene, wobei die Transliteration und der Originaltext bei allen Arbeitsschritten synchronisiert bleiben. Im folgenden Abschnitt beschreiben wir das Datenmodell, das diese Architektur ermöglicht.

Die Basiseinheit in unserem System ist ein Wort, das eine einmalige ID zugewiesen erhält (graphische Einheit). Ein Wort hat folgende Komponenten:

- Eine Liste der einzelnen Fidal [2]-Objekte, wobei ein Fidal-Objekt aus einer ID und einem Label (dem Fidal-Buchstaben) besteht.
- Eine Liste einzelner Silben-Objekte, wobei ein Silben-Objekt aus einer ID und einer Liste von einzelnen Buchstaben-Objekten besteht.
- Ein Buchstaben-Objekt hat immer eine ID und ein Label (das graphische Symbol).

Graphische Einheiten können in mehreren Tokens geteilt werden. Die Tokens sind getrennt gespeichert (mit eigenenen IDs) und verlinkt mit der graphischen Einheit. Elemente der Textstruktur werden durch selbständige Objekte dargestellt, die mit den beinhalteten Tokens verlinkt werden. Editorische Annotationen werden mit den graphischen Einheiten verlinkt. Eine derartig stark vernetzte komplexe Struktur ist mit XML schwierig zu modellieren, insbesondere bei manueller Annotation. Auch die maschinelle Verarbeitung einer derartig tief vernetzten TEI-Struktur wäre sehr kompliziert. Alternativ bietet JSON für dieses Modell eine bessere Handhabbarkeit. Wir haben uns daher entschieden, die Daten in JSON zu speichern und zusätzlich XML-Export zu implementieren. Die Datenstruktur ist in vier JSON-Dateien gespeichert: eine für die graphischen Einheiten, eine für die Tokens, je eine für weitere Annotationsebenen (Textstruktur und Edition). Diese JSON-Dateisammlung wird via Konverterschnittstelle nach *ANNIS* importiert.

corpus-tools.org: *ANNIS*, *Pepper*, *Salt*

ANNIS (Krause & Zeldes 2016) ist eine Such- und Visualisierungsplattform für linguistische Daten, die mehrebenenfähig ist, d.h., verschiedene Annotationsebenen eines Korpus darstellen kann und dabei verschiedenste Annotationsarten berücksichtigt. Dies macht *ANNIS* zu einem hervorragenden Analysetool für die Daten des TraCES-Projekts. Eine besondere Rolle dabei spielt die Mächtigkeit der *ANNIS*-eigenen Abfragesprache AQL (ANNIS Query Language), die neben Freitextsuche und regulären Ausdrücken sich vor allem dadurch auszeichnet, linguistische Strukturen über mehrere Ebenen suchen zu können.

Durch *ANNIS*‘ beschriebene Eigenschaften und seine Konfigurierbarkeit ist die Software für die Verwendung in den unterschiedlichsten Analyseszenarien geeignet und wird in der Tat bereits in verschiedenen Communities verwendet, z.B. in historischer Linguistik, Typologie, Sprachdokumentation u.v.m.

Diese Eignung wird verstärkt durch eine hohe Kompatibilität mit vielen unterschiedlichen linguistischen Datenformaten, die erreicht wird durch Einsatz von *Pepper* (Zipser et al. 2011), einem Konvertierungsframework für liguistische Daten. *Pepper* nutzt ein generisches, graphbasiertes Zwischenmodell, *Salt* (Zipser & Romary 2010), das unterschiedlichste, auch über mehrere Ebenen eines Korpus verteilte, Annotationsarten aufnehmen kann und so weiterverarbeitbar macht. *Pepper* zeichnet sich weiterhin durch eine Plugin-Architektur aus, die es ermöglicht mit geringem Aufwand sowohl Import- als auch Manipulations- und Exportmodule zu entwickeln, die die Quelldaten in ein *Salt*-Modell im Hauptspeicher übertragen, das dort manipuliert werden und anschließend in ein Zielformat überführt werden kann.

ANNIS, *Pepper* und *Salt* sind Komponenten der Softwarefamilie *corpus-tools.org* (Druskat et al. 2016).

Schnittstelle zwischen *GeTa* und *ANNIS*

Diese Infrastruktur ermöglicht es ohne Weiteres, die in *GeTa* annotierten Daten über ein dediziertes *GeTa*-Importmodul für *Pepper* und die Verwendung der bereits existierenden *ANNIS*-Module für *Pepper* – hier das Exportmodul – in *ANNIS* verarbeitbar zu machen. Mit *GeTaModules* (Druskat 2018) haben wir einen solchen Importer entwickelt, dessen Funktionsweise hier kurz beschrieben werden soll. *GeTaModules* ist in Java implementiert und wird als OSGi-Bundle ausgeliefert, das von *Pepper*s OSGi-Plattform verwaltet werden kann. *GeTaModules* ist Open Source unter der Apache License, Version 2.0.

Um Performanz und Speicherökonomie zu optimieren nutzt der *GeTaModules*-Importer eine Kombination aus Streaming und Object-Mapping-Methoden, um die aus *GeTa* exportierten JSON-Dateien einzulesen. Dabei werden die

kleinsten Einheiten der Transliteration auf den graphischen Einheiten genutzt, um eine Tokenisierung in *Salt* zu erstellen und einen virtuellen Primärtext aufzubauen. Auf den so erstellten Modelltokens werden rekursiv die Silben- und Fidal-Objekte als Spannen aufgebaut. Im Anschluss werden die linguistischen Annotationen der *GeTa*-Tokens sowie die weiteren Annotationen aus den JSON-Objekten per Identifikator auf die entsprechenden Einheiten projiziert.

In diesem Zustand hält *Pepper* einen kompletten *Salt*-Dokumentgraphen für das entsprechende Korpusdokument im Hauptspeicher. In einem weiteren Schritt werden Ordnungsrelationen jeweils zwischen den Knoten der Fidal und der transliterierten Wortebene erstellt, damit diese später in *ANNIS* als Segmentierungsgrundlage angezeigt werden können.

Im dritten Schritt werden mit Hilfe des *ANNIS*-Exportmoduls [3] die für den Import in *ANNIS* benötigten Dateien im nativen Format geschrieben. Diese können nun in *ANNIS* über dessen graphische Benutzeroberfläche importiert werden.

Die Konfiguration der Visualisierungen erfolgt durch Anpassung einer während des Exportvorgangs generierten Konfigurationsdatei. Im Fall der *GeTa*-Daten müssen hier lediglich die anzuzeigenden Annotationsebenen und ihre Reihenfolge eingestellt werden. In den *ANNIS*-Daten wird weiterhin ein dedizierter HTML-Visualisierer konfiguriert, der Wortartenannotationen auch graphisch Fidalwörtern zuordnet.

Mit *Pepper* erfolgt die Konvertierung auf der Kommandozeile und die Konfiguration in Textdateien. Da dies für einige potenzielle Anwendergruppen unbekanntes Terrain bedeuten könnte haben wir den Prototypen einer Desktopanwendung für *Pepper* für die Verwendung mit *GeTaModules* angepasst, die eine grafische Oberfläche für die Verwendung von *Pepper* bietet: *Pepper Grinder* (Druskat 2017). *Pepper Grinder (TraCES Edition)* bietet so den Workflow für die Konvertierung der *GeTa*-Daten nach *ANNIS* quasi per Knopfdruck an. *Pepper Grinder* wird in Zukunft in eine vollfunktionale Anwendung ausgebaut, die Modi für verschiedene Anwendergruppen anbieten wird.

Neben der projektinternen Nutzung von *GeTaModules* für die Konvertierung in das *ANNIS*-Format eröffnet die Software weitere Nachnutzungsszenarien für die annotierten Daten. Durch die Existenz von Manipulator- und Export-Modulen für *Pepper* für die verschiedensten Datenformate können die Daten mit anderen Werkzeugen nachgenutzt und etwa um zusätzliche Annotationsebenen angereichert, oder mit anderen Visualisierungen zu weiteren Forschungsfragen analysiert werden. Gleichzeitig leistet die Nachnutzung und Anpassung von *Pepper* durch *GeTaModules* einen Beitrag zur Vernachhaltigung dieses Frameworks. Durch die Ermöglichung des Imports von GeTa-Daten in ANNIS wird aus dem GeTa-Modell eine durchaus attraktive Modellierungsalternative für andere Sprachen mit ähnlichen Merkmalen. Derzeit wird das Modell in laufenden Projekten zu Mayasprachen und Jiddisch erprobt.

Zusammenfassend stellt unser Ansatz eine Alternative dar zu Tendenzen der „Format-Hoheit“, also der Modellierung auf Grundlage eines – eventuell de facto standardisierten – Datenformats im Gegensatz zur Modellierung auf Grundlage der Forschungsfrage und der vorliegenden Daten. Die Einrichtung von Schnittstellen, wie z.B. *GeTaModules*, die eine Nachnutzbarkeit der Daten auch über die ursprüngliche Erstellung hinaus gewährleisten können, ermöglicht eine optimierte Modellierung und die Entwicklung spezifischer, den Bedürfnissen der Forschung und ihrer Daten hochgradig angepasster Werkzeuge, wie etwa *GeTa*.

Fußnoten

1. https://www.traces.uni-hamburg.de/ .
2. "Fidal" ist der Terminus technicus für das äthiopische Silbenalphabet.
3. https://github.com/korpling/pepperModules-ANNISModules/ .

Bibliographie

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014): „CorA: A web-based annotation tool for historical and other non-standard language data“, in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Gothenburg, Sweden: 86-90.

Druskat, Stephan (2018): „GeTaModules (Version 0.9.0)“. Zenodo. DOI: 10.5281/zenodo.1146985. http://doi.org/10.5281/zenodo.1146985

Druskat, Stephan (2017): „Pepper Grinder (Version 0.1.7)“. Zenodo. DOI: 10.5281/zenodo.1041735. https://doi.org/10.5281/zenodo.1041735

Druskat, Stephan / Gast, Volker / Krause, Thomas / Zipser, Florian (2016): "corpus-tools.org: An Interoperable Generic Software Tool Set for Multi-layer Linguistic Corpora", in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016):* 23–28.

Eckart de Castilho, Richard / Mújdricza-Maydt, Éva / Yiman, Seid Muhie / Hartmann, Silvana / Gurevych, Iryna / Frank, Anette / Biemann, Chris (2016): „A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures“, in: *Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan:* 76-84.

Krause, Thomas / Zeldes, Amir (2016): "ANNIS3: A new architecture for generic corpus query and visualization", in: *Digital Scholarship in the Humanities*: 118-139.

Meister, J.C. / Petris, M. / Gius, E. / Jacke, J. (2016): CATMA 5.0 [Software for text annotation and analysis] http://www.catma.de [letzter Zugriff 10.01.2018]

Vertan, Cristina / Ellwardt, Andreas / Hummel, Susanne (2016): "Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte", in: *Proceedings der DHd-Konferenz 2016* http://www.dhd2016.de/abstracts/vortr%C3%A4ge-061.html [letzter Zugriff 25.09.2017].

Zipser, Florian / Romary, Laurent (2010): "A model oriented approach to the mapping of annotation formats using standards", in: *Proceedings of the Workshop on Language Resource and Language Technology Standards (LREC 2010)* https://hal.inria.fr/inria-00527799/ [letzter Zugriff 25.09.2017].

Zipser, Florian / Zeldes, Amir / Ritz, Julia / Romary, Laurent / Leser, Ulf (2011): "Pepper: Handling a multiverse of formats", Poster, *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, Göttingen.

Neue Wahlverwandtschaften

Althof, Daniel

daniel.althof@bbaw.de
BBAW, Deutschland

Digitalität in den Geisteswissenschaften

Die Ausgangslage:

Eine *natürliche* Verwandtschaft besteht zwischen der Geisteswissenschaft (GW) und dem Text sowie zwischen der Naturwissenschaft (NW) und der Zahl. Gründe hierfür lassen sich in einer Ausdifferenzierung der Wissenschaften finden, die mit dem Ende der großen Systeme der klassischen deutschen Philosophie begonnen hat bzw. intensiviert wurde. Im Zuge dieser Ausdifferenzierung haben sich NW und GW aus erkenntnis- und wissenschaftstheoretischen Gründen *in Abgrenzung* zueinander entwickelt. Zugleich hat das Erkenntnisideal der NW einen Rechtfertigungsdruck auf die GW ausgeübt. Denn mit dem Siegeszug der modernen NW wurde die empirisch-experimentelle Wissenschaft zum leitenden Paradigma für objektive Erkenntnis. Wie begründen nun aber die sich jeweils entwickelnden Selbstverständnisse solche natürlichen Verwandtschaftsverhältnisse? Moderne NW bricht mit der Tradition, auch mit der Reflexion des Tradierten, und baut in kritischer Absicht auf Induktion. [1] Eine moderne Variante ist der Positivismus, der Erkenntnis auf empirisch überprüfbare Methoden stützt und traditionelle Forschung unter Sinnlosigkeitsverdacht stellt. [2] Durch Entwicklungen in Mathematik und Logik rückt im logischen Positivismus noch mehr die Formalisierungsbestrebung von Erkenntnis ins Zentrum. [3] Die Grenze zwischen Wissen und Unsinn verläuft damit immer mehr da, wo Erkenntnis formal beschreibbar und empirisch verifizierbar bzw. im kritischen Rationalismus nicht falsifizierbar [4] ist. Die Beschränkung des Sagbaren auf analytische Sätze oder empirisch fundierte Theorie über quantifizierbare und nomologisch reformulierbare Gegenstände erhebt die formalisierte Idealsprache [5] bzw. die Formel zur Wahrheit. Das wirkte auch in den GW. Diese Entwicklung fand Befürworter (Helmholtz 1903, Mill 1974) , aber provozierte bereits früh Kritik und führte zur Ausbildung eines Selbstverständnisses der GW in Opposition zum mathematisch-naturwissenschaftlichen Paradigma. Im Verweis auf die Besonderheit der geisteswissenschaftlichen Gegenstände wurden in logisch-positivistischen Verfahren formulierte Kriterien für Erkenntnis abgelehnt: Nicht die Entdeckung von allgemeinen Gesetzen, sondern die Betrachtung des Individuellen in geschichtlicher Gestalt leitet die GW (Rickert 1986). Die NW sind nomothetisch, die GW ideographisch (Windelband 1924) , haben es mit kulturell vermittelten Werten (Rickert 1986) und erlebnisbasiertem Sinnverstehen (Dilthey 1992) zu tun, das auch immer eine Selbsterkenntnis mit einschließt (Ritter 1974). In den GW wird nicht Bezug genommen auf die Konstanz der Gesetze (Natur), sondern auf ein gemeinsames sprachlich vermitteltes Tun (Kultur) (Cassirer 1961). So spitzt sich die Unterscheidung von Natur- und Geisteswissenschaften darin zu, dass Objektivierungen von Lebensäußerungen *verstanden* werden, indem eigenes Erleben herangezogen wird, Naturerscheinungen aber über Gesetze, Statistik und Verallgemeinerung *erklärt*

werden (Dilthey 1992). Sinn rekonstruierende Hermeneutik auf Grundlage der überlieferten Tradition als Weiterführung eines Diskurses ist somit in den GW (Gadamer 2010) , logisch-mathematische Formalisierung basierend auf erhobenen Daten in den NW die naheliegende Heuristik. Die Verwandtschaft von GW und Text bzw. von NW und der Zahl ist so gesehen noch viel strikter, insofern die Heuristik der NW die (natürliche) Sprache, die der GW aber die Formalisierung prinzipiell ausschließt.

Das Problem:

Eine *natürliche* Verbindung besteht somit in den (überlieferungsbezogenen) GW zwischen Hermeneutik und dem Forscher, in den NW zwischen den aus Experimenten gewonnenen Daten und dem Computer. Während der Einsatz von Forschern in den NW begrüßt wird, verhält es sich mit dem Einsatz von Computern in den GW ganz anders. Das lässt sich aus der obigen Skizze auch einfach nachvollziehen. Mathematik und Logik bilden die Grundlagendisziplinen sowohl der NW als auch der Informatik. Aufgrund dessen stehen sie auf gleicher Basis und bilden quasi lediglich Spezialgebiete desselben Paradigmas. Aus der Skizze ist aber auch klar, dass aus dem gewonnenen Selbstverständnis der GW der Computer die Antithese zum erklärten Erkenntnisideal bildet. Der Versuch der Digital Humanities (DH), den Computer in den Hoheitsgebieten der GW zum Einsatz zu bringen, stellt sich für viele als eine naive und sinnlose Unternehmung dar, weil algorithmische Formalisierung ein grober Reduktionismus sein muss. Das ist gegenwärtig ein starker Vorbehalt in geisteswissenschaftlichen Institutionen. Die breite Diskussion um die Angemessenheit der neuen Methoden schlägt sich beispielhaft im Thema ‚distant reading' nieder. [6] Generell wird dabei die Reichweite der Quantifizierung und Gefahren wie Verdinglichung problematisiert, die in der zugrundeliegenden Ontologie der Comuterisierung unvermeidlich angelegt sind (Berry 2011) . Reduktionismus und Simplifizierung (in der Auslegung) literarischer Texte bilden einen Aspekt, der in den GW den Grundkonflikt (aus dem Themenkomplex KI) zwischen Mensch und Maschine wiederspiegelt (Searle 1993). Neben der Kritik am Umgang mit dem Gegenstand gibt es auch Kritik am Erkenntniswert der aufwändig gewonnenen Daten (Kirsch 2012) . Ein weiterer, die Disziplin selbst betreffender Aspekt ist die fehlende Selbstreflexion durch alleinige Konzentration auf die Entwicklung von Tools. Sowohl die Methodenreflexion (Liu 2008) als auch das kritische Potential der Digital Humanities (Liu 2012) werden nicht ausgeschöpft. Die DH, so die Linie der Kritik, zeigen sich weder dem speziellen Gegenstand der GW, nämlich (hauptsächlich) dem Text, noch den Kernaufgaben der GW, nämlich der Auslegung und der Kritik auch der eigenen Methoden, nicht angemessen.

Der Vortrag:

Die Liste der Einwände könnte sicherlich detailliert fortgeführt werden. Der Vortrag soll jedoch nicht noch einen Grund mehr zu dieser Liste hinzufügen. Vielmehr soll aus der Kernkompetenz der DH, nämlich aus der Digitalisierung, eine intime Verbindung mit den Anliegen der GW herausgearbeitet werden. Es ist also nicht Ziel des Vortrages, die Kluft zwischen den Paradigmen (aus Geistes- und Naturwissenschaften) vermittels einer Rückbesinnung auf eine Kompetenz der GW (äußerlich) zu überbrücken (wie es (Liu 2012) oder auch (Berry und Fagerjord 2017) tun). Es ist die Ambition, eine Wesensverwandtschaft in den Heuristiken zu enthüllen, die sich *aus dem Charakter der Digitalität* selbst entwickeln lässt. Hermeneutik und Computation [7] müssen auf diese Weise also nicht extern (über pragmatische Argumente) verklammert, sondern können vermittelt über Digitalität immanent ineinander überführt werden. Auf diese Weise entsteht eine neue *Wahlverwandtschaft* zwischen Hermeneutik und Computation. Und so wird eine kleine Theorie der Digitalität zu entwickeln sein, die die Limitation der mathematisch-logische Formalisierung transzendiert. Das Ergebnis ist damit ein Verständnis von Digitalisierung, das den Einsatz von Computern in den GW nicht nur als gewinnbringendes Unterfangen zeigt, sondern sogar als folgenrichtigen Schritt ausweist.

Fußnoten

1. Das reicht zurück bis Francis Bacon, der mit der Scholastik gebrochen hat. Dass Induktion für die Rechtfertigung einer Theorie nicht hinreichend ist und im weiteren Verlauf der Wissenschaftsgeschichte weiter ausdifferenziert werden wird, sei hier der Vollständigkeit halber angemerkt.
2. Vgl. hier die exemplarischen Positionen von August Comte (Comte 1830-1842) und John Stuart Mill (Mill 1974) .
3. Vgl. den Wiener Kreis, aber auch den frühen Ludwig Wittgenstein (Wittgenstein 1998).
4. So z.B. bei Karl Popper (Popper 2005).
5. Vgl. wiederum den Wiener Kreis und besonders Bertrand Russel.

6. Gewichtige Argumente gegen die Sinnhaftigkeit eines solchen Vorhabens versammelt (Marche 2012)
7. Ich werde diesen terminus technicus hier verwenden, um damit die computerseitigen Prozesse der Digitalisierung abzubilden, die eine sozio-kulturelle Umwälzung ist. Datenverarbeitung als Alternativ-Begriff wäre zu eng gefasst.

Bibliographie

Berry, D. M. (2011): *Philosophy of Software. Code and Mediation in the Digital Age,* Basingstoke, Hampshire: Palgrave Macmillan UK.

Berry, D. M. & Fagerjord, A. (2017): *Digital Humanities: Knowledge and Critique in a Digital Age,* London: Polity Press.

Cassirer, E. (1961): Naturbegriffe und Kulturbegriffe. *Zur Logik der Kulturwissenschaften. Fünf Studien.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Comte, A. (1830-1842): *Cours de la philosophie positive,* Paris: Bachelier.

Dilthey, W. (1992): *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften,* Stuttgart/Göttingen: Teubner/Vandenhoeck & Ruprecht.

Gadamer, H.-G. (2010): *Hermeneutik I. Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik,* Tübingen: Mohr Siebeck.

Helmholtz, H. L. F. V. (1903): *Über das Verhältnis von Naturwissenschaften zu der Gesammtheit der Wissenschaften,* Braunschweig: Vieweg.

Kirsch, A. (2012): Technology Is Taking Over English Departments. The false promise of the digital humanities. Available from: https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch [Accessed 22. September 2017].

Liu, A. (2008): *Local Transcendence: Essays on Postmodern Historicism and the Database,* Chicago/London: The University of Chicago Press.

Liu, A. (2012): Where is Cultural Criticism in the Digital Humanities? *In:* Gold, M. K. (ed.) *Debates in the Digital Humanities.* Minneapolis: The University of Minnesota Press.

Marche, S. (2012): Literature Is not Data: Against Digital Humanities. *Los Angeles Reviews of Books* [Online]. [Accessed 22. September 2017].

Mill, J. S. (1974): *A System of Logic Ratiocinative and Inductive,* London/Toronto: Routledge and Kegan Paul/University of Toronto Press.

Moretti, F. (2005): *Graphs, Maps, Trees. Abstract Models for a Literary History,* London/New York: Verso.

Popper, K. (2005): *Logik der Forschung,* Tübingen: Mohr Siebeck.

Rickert, H. (1986): *Kulturwissenschaft und Naturwissenschaft,* Stuttgart: Reclam.

Ritter, J. (1974): Die Aufgabe der Geisteswissenschaften in der modernen Gesellschaft. *Subjektivität. Sechs Aufsätze.* Frankfurt/M.: Suhrkamp.

Searle, J. (1993): *Die Wiederentdeckung des Geistes,* München: Artemis & Winkler.

Windelband, W. (1924): Geschichte und Naturwissenschaft. *Präludien. Aufsätze und Reden zur Philosophie und ihrer Geschichte.* Tübingen: Mohr.

Wittgenstein, L. (1998): *Logisch-philosophische Abhandlung. Tractatus logico-philosophicus - Kritische Edition,* Frankfurt/M.: Suhrkamp.

Objekte im Netz – Die Digitalisierung der Sammlungen der Universität Erlangen-Nürnberg als Gegenstand und Methode.

Wagner, Sarah
s.wagner@gnm.de
Germanisches Nationalmuseum, Deutschland

Scholz, Martin
martin.scholz@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Andraschke, Udo
udo.andraschke@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Die Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) entwickelt in Zusammenarbeit mit dem Germanischen Nationalmuseum Nürnberg (GNM) eine gemeinsame Dokumentations- und Digitalisierungsstrategie für die Sammlungen der FAU, um ihre Sicht- und Nutzbarkeit zu erhöhen und sie als bedeutende und noch immer zu wenig genutzte Infrastrukturen für Forschung und Lehre auszubauen. Digitalisierung kommt dabei nicht nur als Methode und praktische Anwendung zum Einsatz, sondern wird ebenso als kritisch zu befragender Gegenstand

untersucht. Der Beitrag stellt Hintergrund, Ziele und bisherige Ergebnisse der Zusammenarbeit vor und skizziert Vorgehen und Methodik.

Universitäre Sammlungen als (digitale) Forschungsinfrastrukturen

Universitäre Sammlungen erleben in den letzten Jahren eine beachtliche Renaissance. Allein in Deutschland existieren rund 1.000 solcher Sammlungen an über 80 Universitäten, die eine Vielzahl an Dingen und Disziplinen umfassen. Das wissenschaftliche Potenzial universitärer Sammlungen ist enorm, weshalb ihr Ausbau zu Forschungsinfrastrukturen in einer Empfehlung des Wissenschaftsrats von 2011 dringend empfohlen wurde (Wissenschaftsrat 2011). Allerdings sind längst nicht alle Sammlungen und Bestände erfasst, nur gut ein Drittel ist digital zugänglich. Die systematische Erfassung und Erschließung wissenschaftlicher Sammlungen sind jedoch grundlegende Voraussetzungen, um sie möglichst effektiv in Forschung und Lehre einsetzen zu können. Die Vorteile einer digitalen Dokumentation sind dabei inzwischen unbestritten. Der Relevanz universitärer Sammlungen für die Forschung und dem Bedarf nach ihrer Nutz- und Verfügbarkeit stehen häufig nicht nur unzureichende monetäre Mittel und fehlendes Personal entgegen, sondern auch das Fehlen von angemessenen Software-Lösungen und Know-how für eine flächendeckende Digitalisierung und Online-Präsenz. Die Situation der digitalen Dokumentation universitärer Sammlungen hat sich in den vergangenen Jahren zwar erheblich verbessert, die Fortschritte sind allerdings meist auf einzelne Sammlungen begrenzt. Übergreifende Strukturen wurden hingegen kaum entwickelt und haben sich auch nicht ergeben. Zahlreiche Sammlungen wurden ad hoc inventarisiert, jedoch ohne hinreichend standardisierte Erfassung oder weiterführende Digitalisierungsstrategie. Das vom Bundesministerium für Bildung und Forschung geförderte Projekt „Objekte im Netz" [1] setzt hier an und entwickelt im Verbund mit dem GNM eine gemeinsame Digitalisierungsstrategie für die Sammlungen der FAU.

Objekte im Netz - Die Sammlungen der Universität Erlangen-Nürnberg als Testlandschaft

Die FAU besitzt über 20 Sammlungen [2] aus den verschiedensten Fachbereichen. Ihre Vielfalt spiegelt sich nicht nur in ihren Objekten, Entstehungskontexten oder Funktionen wider, sondern auch im überaus divergenten Stand ihrer Erfassung sowie in den verschiedenartigen Verfahren ihrer Dokumentation. Sie stellen somit eine ideale Testlandschaft für die exemplarische Entwicklung von fächer- und sammlungsübergreifenden Konzepten zur digitalen Erfassung und Erschließung dar, wie sie der Wissenschaftsrat in seinen Empfehlungen gefordert hat.

Das Forschungsprojekt „Objekte im Netz" zielt auf die Entwicklung einer digitalen Infrastruktur, die langfristig eine gesicherte Erfassung und Vernetzung der Bestände der FAU erlaubt sowie deren weitere Sicht- und Nutzbarkeit befördert. Von zentraler Bedeutung sind dabei gemeinsame Erfassungsstandards und -formate sowie eine gemeinsame Software-Lösung und Webpräsenz. Als Ergebnis des Vorhabens werden weiterhin ein allgemeines Konzept zur digitalen Dokumentation sowie eine dazu passende Software zur Verfügung gestellt. Damit dient das Projekt längst nicht nur der infrastrukturellen Verbesserung und Dynamisierung der hiesigen Bestände, sondern bietet weit darüber hinaus auch anderen wissenschaftlichen Sammlungen Werkzeuge und Workflows an, mit deren Hilfe sich heterogene Bestände erfassen, erforschen und vernetzen lassen. Nicht zuletzt strengt das Projekt einen kritischen Dialog über die Herausforderungen, Hindernisse und Folgen der Digitalisierung wissenschaftlicher Sammlungen an und trägt damit zu einem notwendigen Diskurs bei, der bislang nur wenig entwickelt ist und kaum in die digitale Sammlungspraxis einfließt.

Vorgehen und Vernetzung

Das Projekt „Objekte im Netz" versteht sich als gemeinschaftliches Vorhaben sämtlicher Projektbeteiligter und wird daher im überaus engen Austausch vorangetrieben. Für die Entwicklung eines standardisierten Erfassungsschemas für die universitären Sammlungen wurden insgesamt sechs Teilbestände der FAU ausgewählt, die mit ihren heterogenen Beständen und ihrem unterschiedlichen Stand der Erschließung, Digitalisierung sowie den dazu eingesetzten Methoden und Werkzeugen die Bandbreite universitärer Sammlungen repräsentieren sollen: Die Graphische

Sammlung, die Medizinische Sammlung, die Geowissenschaftliche Sammlung, die Schulgeschichtliche Sammlung, die Ur- und Frühgeschichtliche Sammlung sowie die Studiensammlung Musikinstrumente und Medien der Universität Würzburg mit den Beständen des ehemaligen musikwissenschaftlichen Seminars der FAU. Ausgehend von und anhand dieser Auswahl wurde ein Metadatenschema unter Berücksichtigung bestehender Dokumentationsstandards entwickelt, das mit Blick auf die übrigen Sammlungen grundlegende Aspekte, aber auch sammlungsspezifische Eigenheiten zu berücksichtigen hat.

Vom Ziel einer gemeinsamen digitalen Erfassung, Datenspeicherung und Präsentation leiten sich auch die zentralen Forschungsfragen nach der Tragfähigkeit und Umsetzbarkeit eines gemeinsamen Datenmodells sowie einer abgestimmten Terminologie zur Abbildung der in ihrer Materialität, Funktion und Provenienz heterogenen Sammlungsbestände ab. Dabei werden die Erkenntnisse ausgehend von den oben erwähnten Sammlungen konsequent auf ihre Generizität hin geprüft, um die Anwendbarkeit auf weitere universitäre Sammlungen hin zu gewährleisten.

Als technische Grundlage dient die virtuelle Forschungsumgebung „WissKI“ (Wissenschaftliche KommunikationsInfrastruktur), die im Hinblick auf die spezifischen Anforderungen und Eigenarten universitärer Sammlungen – insbesondere in Bezug auf die semantische Wissensrepräsentation (Görz 2011, Hohmann 2011, Hohmann/Schiemann 2013) – anzupassen und auszubauen ist. [3] Ihr Einsatz erlaubt die Vernetzung unterschiedlichster Bestände und komplexer Bestandsinformationen sowie weiterer digitaler Ressourcen, aus der sich neue Forschungsfragen und erhebliche Erkenntnispotentiale ergeben können.

WissKI erweitert die Idee und Konzepte des Wiki zu einer webbasierten virtuellen Forschungsumgebung, die insbesondere auf die Belange und Besonderheiten der Forschung und Dokumentation von kulturellem Erbe ausgerichtet ist. Das System setzt auf offene Datenformate und Standards, die den Langzeiterhalt der verwalteten Datenbestände sichern. Dazu werden die Technologie des Linked Open Data und des Semantic Web benutzt. Eine Schlüsselrolle kommt hierbei dem dem Conceptual Reference Model [4] (CRM) des ICOM-CIDOC als formaler Referenzontologie zu, gleichermaßen um fach- und sammlungsspezifische Anwendungsontologien erweitert werden kann.

Die notwendige Weiterentwicklung von WissKI hinsichtlich der Spezifika universitärer Sammlungsbestände sowie den Anforderungen einer objektbezogenen Forschung und Lehre erfolgt dabei weniger technik- als sammlungsgetrieben. Dem entspricht das durchaus aufwendige Vorgehen im Projekt: die avisierte digitale Infrastruktur wird von Sammlungsbeauftragten und Informatikern gemeinsam aufgebaut. Auf diese Weise finden die Belange der einzelnen Sammlungen unmittelbar Eingang in die Entwicklung der Software und benötigten Funktionalitäten. Den Sammlungen kann somit ein digitales Werkzeug zur Verfügung gestellt werden, das nicht bloß die Erfassung ihrer Bestände fördert, sondern darüber hinaus spezifische forschungs- und lehrrelevante Anwendungen ermöglicht sowie betriebliche Abläufe des Sammlungsalltags unterstützt.

Das gewählte Vorgehen setzt allerdings eine hohe Bereitschaft der beteiligten Sammlungsbeauftragten voraus, sich eingehend mit oft fachfremden Werkzeugen und Techniken auseinanderzusetzen, im Gegenzug festigt es das Verständnis für die eingesetzten Technologien und Verfahren. Nicht zuletzt bauen die Sammlungen über die beteiligten Mitarbeiterinnen und Mitarbeiter eigene Kompetenzen im Bereich der digitalen Dokumentation auf. Die Herausforderungen der ontologischen Modellierung sowie die fächer- und sammlungsübergreifende Anlage des Projekts führen dabei zu einer vertieften Reflexion über die eigenen Bestände und Sammlungslogiken, aber auch zur Einsicht in die notwendige Standardisierung von Begrifflichkeiten, Werkzeugen und Workflows, wie sie eine übergreifende Digitalisierungs- und Dokumentationsstrategie unbedingt zu berücksichtigen hat.

Ansätze zu einer reflexiven Digitalisierung

Die neuen Möglichkeiten der Informationstechnologien und die allerorten wachsenden digitalen Dingarchive verändern die Arbeit, Forschung und Lehre an und mit den Objektbeständen. Das Thema Digitalisierung wird im Rahmen des Projektes deshalb nicht alleine aus Sicht methodischer und technischer Aspekte der digitalen Dokumentation behandelt, sondern im Sinne einer reflexiven Digitalisierung auch hinsichtlich der Herausforderungen und Folgen des Aufbaus digitaler Infrastrukturen sowie des Einsatzes digitaler Mittel und Methoden. In welchem Verhältnis stehen analoge und digitale Bestände, Original und Digitalisat? Welchen technischen, rechtlichen und nicht zuletzt epistemologischen Problemen hat sich die Anwendung digitaler Methoden in der Sammlungspraxis zu stellen und welche Herausforderungen muss sie bewältigen? Welche Kompetenzen erfordert sie und welchen Wandel

erfahren Sammlungspraxis und sammlungsbezogene Forschung dadurch? Inwiefern verändern also digitale Sammlungen die kustodiale und wissenschaftliche Arbeit an und mit den Beständen? Zur Beantwortung dieser Fragen arbeitet das Projekt eng mit dem Interdisziplinären Zentrum für digitale Geistes- und Sozialwissenschaften der FAU (IZdigital) zusammen und bietet darüber hinaus ein fächer- und sammlungsübergreifendes Lehrangebot im Bereich der Digital Humanities und der Museologie an. Die Ergebnisse und Erkenntnisse der innerhalb des Projekts geführten Diskurse finden wiederum Eingang in die Sammlungspraxis und digitale Dokumentation der Bestände. Der Einsatz digitaler Werkzeuge und Praktiken erscheint aus dieser Sicht gleichermaßen als zu untersuchender Gegenstand und angewandte Methode. Für die Entwicklung einer an der Arbeit mit und an wissenschaftlichen Sammlungen orientierte Digitalisierungsstrategie gilt es beide Perspektiven angemessen zu berücksichtigen und möglichst zusammenzuführen.

Fußnoten

1. Das Projekt wird von 2017 bis 2020 gefördert. Mehr Informationen unter: http://objekte-im-netz.fau.de/projekt/ [letzter Zugriff 10.01.2018].
2. Informationen zu den Sammlungen der FAU unter https://www.fau.de/universitaet/das-ist-die-fau/sammlungen-der-fau/ [letzter Zugriff 10.01.2018].
3. Einen Einblick in WissKI gibt die Webseite http://wiss-ki.eu/ [letzter Zugriff 10.01.2018].
4. Das CIDOC CRM wurde vom International Committee for Documentation als Teil des International Council of Museums (ICOM) als formale Referenzontologie erarbeitet und ist seit 2006 als ISO Norm (ISO 21127) anerkannt. In der „Erlangen CRM“ (URL: http://erlangen-crm.org/ [letzter Zugriff 10.01.2018]) auf Basis von OWL liegt eine maschinenlesbare Version vor.

Bibliographie

Görz, Günther : „WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung” in: Kunstgeschichte, Open Peer Reviewed Journal, urn:nbn:de:bvb:355-kuge-167-7 [letzter Zugriff 10.01.2018].

Hohmann, Georg (2011): „Die Anwendung von Ontologien zur Wissensrepräsentation und -kommunikation im Bereich des Kulturellen Erbes” in: Schomburg, Silke u.a. (eds.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Köln: Hochschulbibliothekszentrum NRW 33-39.

Hohmann, Georg / Schiemann, Bernhard (2013): „An Ontology-Based Communication System for Cultural Heritage. Approach and Progress of the WissKI Project” in: Hans Bock u.a. (eds.): Scientific Computing and Cultural Heritage. Berlin: Springer 127-135.

Wissenschaftsrat (2011): *Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen*. Berlin https://www.wissenschaftsrat.de/download/archiv/10464-11.pdf) [letzter Zugriff 10.01.2018].

Perspektiven kritischer Interfaces für die Digital Humanities im 3DH-Projekt

Kleymann, Rabea
Rabea.Kleymann@uni-hamburg.de
Universität Hamburg, Deutschland

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

Stange, Jan-Erik
jan-erik.stange@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

Eine Haupttätigkeit von LiteraturwissenschaftlerInnen ist die Interpretation von literarischen Texten. Unter „Interpretation“ wird „ein Sprechen oder Schreiben über Texte [verstanden], in dem ihnen auf methodische und argumentierende Weise Bedeutungen zugeschrieben werden“ (Albrecht et. al. 2015: 1). Mit dem Einzug digitaler Methoden in die literaturwissenschaftliche Praxis werden dabei insbesondere die Formen der Bedeutungszuschreibung und -produktion sowie die damit verbundene Frage nach einem Textverstehen zur Disposition gestellt (vgl. Rockwell/Sinclair 2016). Eine wichtige Rolle in der Gestaltung von digitalen Zugängen zur Textinterpretation spielen in diesem Kontext User Interfaces im Allgemeinen und interaktive Visualisierungen im Speziellen. Unter „Interfaces“ verstehen wir die gesamte visuelle Struktur und Bedienlogik ei-

ner Software, mit „Visualisierungen“ bezeichnen wir interaktive Teile des Interfaces, die Nutzern auf Daten (in diesem Fall Textdaten) basierende visuelle Repräsentationen zugänglich und diese manipulierbar machen.

Die Frage, wie solche Interfaces und Visualisierungen an den Bedürfnissen interpretierender LiteraturwissenschaftlerInnen orientiert sein sollten, um eine Textanalyse und -interpretation sinnvoll zu unterstützen, ist Teil des Forschungsanliegens des Projekts 3DH – *Dreidimensionale dynamische Datenvisualisierung und Exploration für Digital Humanities-Forschungen* der Universität Hamburg. [1]

Vor dem Hintergrund, dass Interfaces und Visualisierungen in den Geisteswissenschaften eine wachsende Relevanz zugesprochen wird, plädieren wir für eine designbasierte kritische Perspektive. Darunter verstehen wir die Einbindung von Methoden aus dem nutzerzentrierten Design in die Tool-Entwicklung. Diese erlauben eine an Nutzerbedürfnissen und -erwartungen orientierte Gestaltung von Interfaces und Visualisierungen. Dieser Ansatz zielt darauf, die Grenzen der klassischen Verfahren der Datenerhebung und -verarbeitung, wie sie bereits von bestehenden Annotationstools (z. B. CATMA; siehe http://catma.de) unterstützt werden, neu auszuloten.

Zunächst stellen wir die aus unserer Sicht notwendigen normativen Anforderungen an ein kritisches Interface- und Visualisierungskonzept vor, um dann darzulegen, welche Strategie wir in der Gestaltung und Entwicklung solcher Interfaces und Visualisierungen verfolgen. Im Anschluss werden beispielhaft einzelne Wireframes diskutiert.

Grundlagen eines kritischen Interface- und Visualisierungskonzeptes

Eine kritische Bestandsaufnahme aktueller Visualisierungstools und -metaphern und deren Anwendung in den Geisteswissenschaften (vgl. Bradley 2008; Drucker 2011, 2014; Gibbs/Owens 2012) führte im 3DH-Projekt zur Formulierung von vier konzeptionellen Postulaten. [2] Definiert werden damit die Prinzipien, die einem Interface bzw. einer Visualisierung zugrunde gelegt werden müssen, sofern diese literaturwissenschaftliche Arbeitsprozesse sinnvoll unterstützen sollen. Es handelt sich um die Postulate

1. des „Two-Way-Screens“: Das Interface soll vom Renderer zum bidirektionalen Instrument werden;
2. der „Qualität“: Die Visualisierung soll nicht nur der Repräsentation von Daten fungieren, sondern Möglichkeiten bieten, qualitative Beobachtungen und Aussagen zu tätigen;
3. der „Parallaxe“: Die visuelle und epistemische Multiperspektivität einer Visualisierung soll gestärkt werden;
4. der „Diskursivität“: Durch geeignete visuelle Erscheinungsformen und -verfahren soll die Generierung, Diskussion und Kritik von Hypothesen gefördert werden.

Die normativen Anforderungen an ein reflektiertes Interface- und Visualisierungskonzept rekurrieren auf die spezifischen Eigenschaften eines literaturwissenschaftlichen Arbeitsprozesses. Dieser folgt keinem finiten Set von methodischen Regeln oder Verfahrensweisen. Vielmehr tritt die Textanalyse und -interpretation als komplexer epistemisch und ästhetisch geformter sowie zirkulär organisierter Prozess auf, der sich als in hohem Maße subjektiv und kontextsensitiv erweist (vgl. Winko 2015: 486).

Für die praktische Entwicklung literaturwissenschaftlicher Interfaces und Visualisierungen gehen wir von zwei Annahmen aus.

Die Komplexität des Interpretationsprozesses kann erstens auf eine Reihe von grundlegenden „routineförmige[n] Tätigkeiten“ (Albrecht et al. 2015: 2) bzw. „scholarly primitives“ (Unsworth 2000) zurückgeführt werden, die iterativ von LiteraturwissenschaftlerInnen zur Bedeutungszuweisung durchgeführt werden. Diese routineförmigen Tätigkeiten sind „oftmals nicht vollständig durch explizierbare Regeln oder Methoden bestimmt, sondern [beruhen] in hohem Maße auf implizitem Wissen und Können“ (Albrecht et al. 2015: 2).

Zweitens nehmen wir an, dass nutzerzentrierte Gestaltung als etablierter Prozess, dessen Methodenrepertoire sich sowohl aus dem an der Praxis orientierten Design speist, als auch aus der eher wissenschaftlich orientierten Human-Computer-Interaction (HCI), als Vorgehen besonders geeignet ist, implizites Wissen und Können von Nutzern herauszufinden. In diesem Sinne steht die Entwicklung eines Interfaces nicht am Ende „as an afterthought, thrown together after completing the core functionality” (Gibbs/Owens 2012), sondern ist Teil der Ausdifferenzierung und Analyse von Textinterpretation. Mit der Integration von Designmethoden, so die Argumentation des 3DH-Projekts, begegnen wir der von Gibbs und Owens beklagten mangelnden Nutzerzentriertheit bestehender DH-Annotationstools.

Annäherung an die literaturwissenschaftliche Arbeitspraxis über nutzerzentrierte Szenarien

Mithilfe von Szenarien werden besagte ‚routineförmige Tätigkeiten' untersucht und erfassbar gemacht. [3] Kollaborativ wurden bislang drei generische Szenarien konstruiert, die das Spektrum typischer interpretierender Forschungstätigkeiten repräsentieren. [4]

Die Konzepte erstrecken sich vom klassischen „Close Reading"-Beispiel mit Fokus auf noch nicht taxonomisch gelenkte freie Annotationen und Kommentare bis zum „Distant Reading"-Beispiel unter Anwendung des probabilistischen Verfahrens „Topic Modeling".

Analog zur Konstruktion der Szenarien ist auch der von uns verfolgte nutzerzentrierte Gestaltungsprozess typischerweise durch ein iteratives Vorgehen gekennzeichnet. [5]

Abbildung 1 zeigt eine Auswahl von Varianten sogenannter „Wireframes", d.h. von Designskizzen zu einem Interface für die Tätigkeit „Kommentieren" in dem Szenario „Exploration freier Annotationen zwecks Schärfung der literaturwissenschaftlichen Fragestellung".

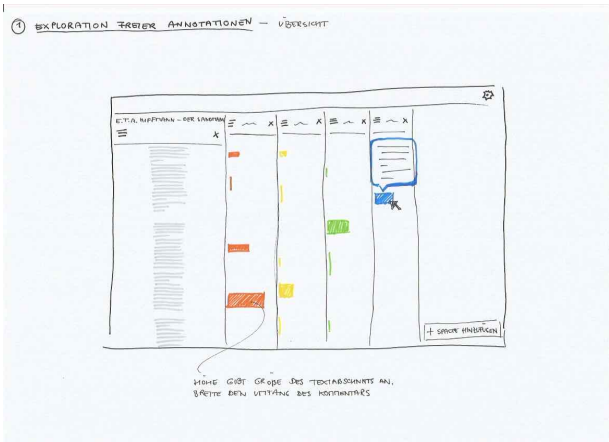


Abb.1: Wireframes für ein Interface zur Funktion „Kommentieren" in dem Szenario „Exploration freier Annotationen zwecks Schärfung der literaturwissenschaftlichen Fragestellung".

In der Abbildung wird der Entwicklungsstand der visuellen Gestaltung dieser Tätigkeit zu zwei verschiedenen Zeitpunkten sichtbar. Der untere Wireframe präsentiert einen fortgeschrittenen Stand. Die Wireframes illustrieren die Verschränkung einzelner Tätigkeiten, die im Szenario beschrieben werden. Hier sieht man etwa den Zusammenhang zwischen den Tätigkeiten „Annotieren" und „Kommentieren" dargestellt, in dem Sinne, dass ein Kommentar in der Regel (wenn auch nicht immer) einer Annotation zugeordnet ist, die eine diskrete Position im Text besitzt, also häufig nicht isoliert von der Tätigkeit des Annotierens betrachtet werden kann.

Erste Ergebnisse und Schlussfolgerungen

Bereits in diesem grob umrissenen Beispiel von Varianten einer visuellen Unterstützung der Tätigkeit „Kommentieren", manifestieren sich zwei der wesentlichen Herausforderungen an die Erarbeitung und Implementierung eines kritischen Interface- und Visualisierungskonzeptes:

Überlagerung und Wechselwirkung routinemäßiger Tätigkeiten im Interface

Aufgabe der Designiterationen von Szenarien und Wireframes wird es sein, für die Überlagerung und Wechselwirkung der routinemäßigen Tätigkeiten in der Gestaltung des Interfaces eine adäquate Entsprechung zu finden.

Auf Grundlage der bisher erstellten Wireframes für die drei Szenarien ließen sich zwei Modi des literaturwissenschaftlichen Interpretationsverfahrens ausmachen. Zum einen besteht der Interpretationsprozess aus Tätigkeiten, die primär die analytische, investigative Erforschung

des Textes adressieren (Annotieren, Sammeln, Kommentieren). [6] Komplementär dazu verhalten sich zum anderen die argumentativen, vornehmlich synthetisierenden Tätigkeiten der Textinterpretation (Gruppieren, Ordnen, Strukturieren). Um den literaturwissenschaftlichen Arbeitsprozess adäquat unterstützen zu können, so unsere Annahme, müssen Interfaces beide Modi in enger Verschränkung umfassen und stetigen Wechsel zwischen diesen ermöglichen. Abbildung 2 zeigt einen ersten Versuch, die Ebene der Bedeutungsproduktion – von uns als „semantic plane“ bezeichnet – d.h. die Ebene, die analytische und synthetisierende Tätigkeiten umfasst und miteinander verknüpft, in die Interface-Entwürfe zu integrieren.

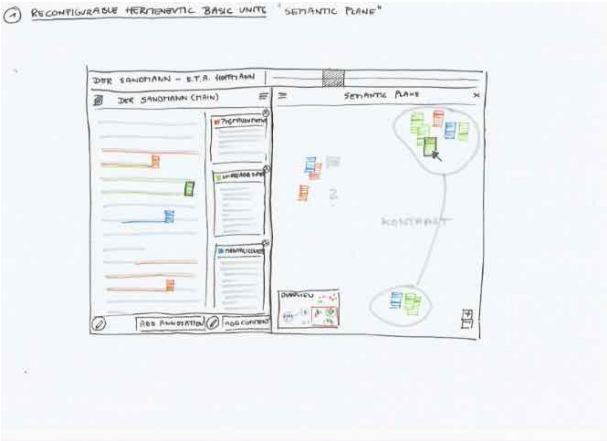


Abb.2: Versuch der Integration der „semantic plane“ über die vom User vollzogene Positionierung von Elementen.

Flexibilität und Manipulierbarkeit von Visualisierungen

Um eine flexiblere Unterstützung des literaturwissenschaftlichen Arbeitsprozesses im Interface zu ermöglichen, ist ferner der Einsatz unterschiedlicher Visualisierungen sinnvoll (vgl. Cheema et al. 2015). Diese sollten vergleichbar sein in ihren Resultaten und ihre Darstellungsform, Interaktivität und der jeweils repräsentierte Datenausschnitt sollte individuell bestimmbar sein.

Im Interesse einer Justierbarkeit der Visualisierungen orientieren wir uns an den Visualisierungssystematiken, wie sie Wilkinson mit seinem „Grammar-of-Graphics“-Ansatz vorschlägt (vgl. Wilkinson et al. 2005). Es existieren bereits verschiedene Implementierungen dieses Ansatzes, die es erlauben, die oben genannten Parameter einer Visualisierung mithilfe einer Grammatik beschreibbar und manipulierbar zu machen (vgl. Satyanarayan et al. 2017). In unserem Kontext ist die deklarative Sprache VEGA [7] interessant, die auf der verbreiteten Javascript-Visualisierungsbibliothek D3 [8] beruht und es Nutzerinnen und Nutzern ermöglichen würde, ihre Visualisierungsspezifikationen in JSON zu formulieren.

Im Sinne einer nutzerzentrierten Gestaltung ist es wichtig, diese Funktionalitäten an die unterschiedlichen Vorkenntnisse der Nutzerinnen und Nutzer anpassen zu können. Weniger fortgeschrittene Nutzerinnen und Nutzer sollen stets auf Standardvisualisierungen zurückgreifen können. Unserem Anspruch gemäß steht die Unterstützung des interpretierenden Prozesses als Ganzheit mit dem Interface im Vordergrund.

Fazit

Ein reflektiertes Interface- und Visualisierungskonzept ist nach unserer Einschätzung besonders relevant, um eine Durchdringung und Neuausrichtung der Verfahren und Zugänge der Datenverarbeitung und -exploration der Textanalyse und –interpretation zu unterstützen. Die Integration von Methoden aus dem nutzerzentrierten Design stellt in diesem Kontext einen noch kaum erprobten Weg dar, um neue Tools in den DH zu entwickeln. So dient unsere designbasierte kritische Perspektive, die im Sinne einer Interface- und Visualisierungsutopie zu verstehen ist, letztlich auch der „epistemologischen Selbstaufklärung“ (Albrecht et al. 2015, 10) der literaturwissenschaftlichen Disziplin. Denn wie im 3DH-Projekt bereits deutlich wird, trägt ein reflektiertes Interface- und Visualisierungskonzept zur interdisziplinären Methodendiskussion in der Literaturwissenschaft bei.

Um den Mehrwert eines Interface- und Visualisierungskonzepts praktisch belegen zu können, wird eine sorgfältige Überprüfung unserer auf den Szenarien beruhenden Annahmen notwendig sein. Während das informelle Feedback von Expertinnen und Experten zwar hilfreich sein kann, so erlaubt es doch noch keine Rückschlüsse auf die tatsächliche Eignung und Aufgabenangemessenheit der Interfaces und Visualisierungen in der Praxis. Vorgesehen ist deshalb, im nächsten Schritt das bisher nur in Form statischer Wireframes umgesetzte Konzept in Form von funktionalen Interaktions-Prototypen mit Nutzern zu testen, um belastbarere Aussagen über seine Validität zu erhalten.

Fußnoten

1. Das Projekt *3DH – Dreidimensionale dynamische Datenvisualisierung und Exploration für Digital Humanities-Forschungen* wird in der ersten Projektphase (02/2016 – 01/2019) von der Behörde für Wissenschaft, Forschung und Gleichstellung gefördert. Für weitere Informationen vgl. *www.threedh.net* [letzter Zugriff: 07.09.2017].
2. Einige Digital-Humanities-Tools und Forschungsansätze, die auf die Unterstützung von hermeneutischen Prozessen zielen: Pliny und TEASys.
3. Der Begriff „Szenario“ wird im Projekt in seiner Bedeutung als Methode und essentieller Bestandteil eines nutzerzentrierten Designprozesses nach Rosson/Carroll (2009) verwendet: „Scenarios of use reconcile concreteness and flexibility. [...] Initial scenarios are often extremely rough. They specify a possible design by specifying the tasks users can carry out, but without committing to lower-level details describing how the tasks will be carried out, or how the system will present the functionality for the tasks.”
4. Szenarien:(1) Exploration freier Annotationen zwecks Schärfung der literaturwissenschaftlichen Fragestellung(2) Themenexploration in einem größeren Textkorpus – Topic Modeling(3) Exploration taxonomiebasierter Annotationen zur Auswertung strukturierter Textanalysen- Während die ursprüngliche Fassung der Szenarien auf den Erfahrungen literaturwissenschaftlicher Forschungstätigkeit im eigenen Team basierte, haben wir diese im Rahmen eines Workshops (5.-7. August 2017) mit Johanna Drucker, Geoffrey Rockwell, Marian Dörk und Evelyn Gius überarbeitet.
5. Grundsätzlich lassen sich vier Phasen unterscheiden nach DIN EN ISO 9241-210:1. Analyse, Beschreibung und Verständnis des Nutzungskontextes2. Spezifikation der Nutzungsanforderungen3. Gestaltungslösungen entwerfen4. Evaluation der entworfenen Gestaltungslösungen aus NutzerperspektiveMit jeder Iteration steigt die Komplexität und/oder der Grad ästhetischer Qualität der Entwürfe. Während zu Beginn mit einfachen Handskizzen von Abläufen im Interface gearbeitet wird (Wireframes), kommen später Papierprototypen zum Einsatz, Screendesigns und schließlich auch interaktive Mock-Ups oder sogar Prototypen. Die Evaluation kann ebenfalls unterschiedliche Formen annehmen. Während in frühen Stadien Entwürfe eher informell diskutiert werden oder in Workshops mit Nutzern und Experten aller beteiligten Fachrichtungen ergänzt und erweitert werden, können spätere interaktive Entwurfsformen auch mit Probanden in qualitativen Nutzertests getestet werden.
6. Vgl. auch Winkos Ausführungen zur Differenzierung von Textanalyse und -interpretation (2003: 598): “Textanalysen basieren auf sprachlichen, formalen und strukturellen Informationen und greifen in erster Linie auf intratextuelle sowie auf denjenigen Typ extratextueller Kontexte zurück, der zum primären Verständnis vorauszusetzen ist; Interpretationen dagegen setzen Ergebnisse der Textanalyse voraus, und es dominiert tendenziell der Rekurs auf weitergehende, insbesondere inter- und extratextuelle Kontexte. Darüber hinaus sind Hypothesen über die Bedeutung eines Textes in Interpretationen komplexer als in Textanalysen.”
7. VEGA – A Visualization Grammar: *https://vega.github.io/vega/*
8. D3 – Data-Driven Documents: *https://d3js.org/*

Bibliographie

Albrecht, Andrea / Danneberg, Lutz / Krämer, Olav / Spoerhase, Carlos (Hg.) (2015): *Theorien, Methode und Praktiken des Interpretierens.* Berlin: de Gruyter.

Bradley, John (2008): “Thinking about interpretation: Pliny and scholarship in the humanities”, in: *Literary and linguistic computing*, 23(3), 263-279.

Bühler, Axel (2003): “Interpretieren - Vielfalt oder Einheit?”, in: Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (Hg.): *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte.* Berlin: De Gruyter (Revisionen, 1) 169-181.

Cheema, Muhammad F. / Jänicke, Stefan / Franzini, Greta / Scheuermann, Gerik (2015): “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges“, in: *Eurographics Conference on Visualization* (EuroVis) - STARs 83-103. *https://www.informatik.uni-leipzig.de/~stjaenicke/Survey.pdf* [letzter Zugriff: 10.09.2017]

Drucker, Johanna (2011): “Humanities Approaches to Graphical Display”, in: *DH Quarterly*, 5(1). *http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html* [letzter Zugriff: 20.09.2017]

Drucker, Johanna (2014): *Graphesis: Visual Forms of Knowledge Production*. MetaLABprojects. Cambridge, Massachusetts: Harvard University Press.

Gadamer, Hans-Georg (1960, 1990): *Hermeneutik I. Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik.* Tübingen: Mohr.

Gibbs, Fred / Owens, Trevor (2012): “Building better digital humanities tools”, in: *DH Quarterly*, 6(2).

Grinstein, Georges (2012): “New Grand Challenges in Information Visualization: New Theories, New Devices, and New Capabilities“, in: *Keynote address at iV2012*.

Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (2003): “Der Bedeutungsbegriff in der Literaturwissenschaft. Eine historische und systematische Skizze”, in: dies. (Hg.): *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin: De Gruyter (Revisionen, 1) 3-32.

Kindt, Tom / Köppe, Tilmann (2008): “Einleitung”, in: dies. (Hg.): *Moderne Interpretationstheorien. Ein Reader.* Göttingen: Vandenhoeck & Ruprecht 7-26.

Nünning, Ansgar (Hg.) (2008): *Metzler Lexikon. Literatur- und Kulturtheorie. Ansätze - Personen - Grundbegriffe.* Stuttgart: Metzler 281-284.

Nünning, Vera (2010): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen.* Stuttgart: Metzler 1-29.

Rosson, Mary B. / Carroll, John M. (2009): *Scenario based design. Human-computer interaction.* Boca Raton, FL, 145-162.

Satyanarayan, Arvind / Moritz, Dominik / Wongsuphasawat, Kanit / Heer, Jeffrey (2017): “Vega-Lite: A Grammar of Interactive Graphics”, in: I *EEE transactions on visualization and computer graphics* 23(1), 341–350. DOI: 10.1109/TVCG.2016.2599030. [letzter Zugriff: 18.09.2017]

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.* King’s College London, 2000. *http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html* .[letzter Zugriff: 18.09.2017]

Wilkinson, Leland (2006): *The grammar of graphics*. New York: Springer Science & Business Media.

Winko, Simone (2003): “Textanalyse”, in: Georg Braungart / Harald Fricke / Klaus Grubmüller / Friedrich Vollhardt / Klaus Weimar (Hg.): *Reallexikon der deutschen Literaturwissenschaft*. Bd. 2.Berlin, New York: de Gruyter 2003 597-601.

Winko, Simone (2015): “Zur Plausibilität als Beurteilungskriterium literaturwissenschaftlicher Interpretationen”, in: Andrea Albrecht / Lutz Danneberg / Olav Krämer / Carlos Spoerhase (Hg.): *Theorien, Methoden und Praktiken des Interpretierens.* Berlin: de Gruyter 483-511.

3DH *http://www.threedh.net*

CATMA *http://www.catma.de*

D3 – Data-Driven Documents: *https://d3js.org/*

forTEXT *http://www.fortext.net*

hermA *https://www.korpuslab.uni-hamburg.de/projekte/herma.html*

heureCLÉA *http://www.heureclea.de*

Pliny http://pliny.cch.kcl.ac.uk/

TEASys http://www.annotation.es.uni-tuebingen.de/?page_id=200

VEGA – A Visualization Grammar: *https://vega.github.io/vega/*

http://www.procontext.com/aktuelles/2010/03/iso-9241210-prozess-zur-entwicklung-gebrauchstauglicher-interaktiver-systeme-veroeffentlicht.html [letzter Zugriff: 18.09.2017]

Positivistischer Methodenfetischismus als Anathema der digitalen Geisteswissenschaften

Arnold, Eckhart

arnold@badw.de

Bayerische Akademie der Wissenschaften, Deutschland

Entsprechend dem Thema der diesjährigen DhD-Konferenz "Kritik der digitalen Vernunft" möchte ich mich in meinem Vortrag der Kritik fehlgeleiteter digitaler Geisteswissenschaften widmen, und dazu erstens anhand von Beispielen zeigen, wie sich fehlgeleiteter Technikeinsatz in den Geistes- und Sozialwissenschaften darstellt und zweitens Kriterien dafür vorschlagen, nach denen beurteilt werden kann, wann der der Technikeinsatz in den Geistes- und Sozialwissenschaften sinnvoll und wann fehlgeleitet ist. Das dient - so hoffe ich - der besseren Aufklärung unseres digital untersützten Erkenntnisvermögens über sich selbst.

Es steht außer Frage, dass Computer nützliche Werkzeuge sind, und dass man digitale Methoden in den Geisteswissenschaften auf vielfache Weise gewinnbringend einsetzen kann. Dabei wäre es aber eine Fehleinschätzung zu glauben, dass die Digitaltechnik für die Geisteswissenschaften lediglich ein Hilfsmittel darstellt, die uns erlaubt, bestimmte Forschugsaufgaben in den Geisteswissenschaften schneller und besser zu erledigen, die Ergebnisse leichter zu verbreiten etc. Viel-

mehr ändert der massive Einsatz von Digitaltechnik auch den Charakter der Wissenschaften: Neue Themenfelder werden erschlossen, andere Fragestellungen als relevant empfunden, andere Verfahrensweisen als mustergültig oder auch nicht (mehr) akzeptabel empfunden und andere Fertigkeiten und Kenntnisse von den Wissenschaftlerinnen und Wissenschaftlern erwartet.

Diese Veränderungen können durchaus an den Identitätskern dessen rühren, was ein geistes- oder gesellschaftswissenschaftliches Fach bisher ausgemacht hat. Das allein bedeutet noch nicht, dass diese Veränderungen schlecht sind, denn dass wissenschaftliche Fächer und ihre Untersuchungsgegenstände einem historischem Wandel unterliegen ist unvermeidlich. Es wirft aber die Frage auf, wann der Einsatz bestimmter technischer Mittel und die dadurch herbeigeführten Veränderungen legitim und zum Nutzen der Wissenschaft sind und wann zum Schaden der Wissenschaft, so dass man besser Abstand davon nehmen sollte.

Auf einer höheren Ebene schließt sich daran die weitergehende Frage an, nach welchen Kriterien dies beurteilt werden kann. Eine offensichtlich unzureichende Antwort ist die, dass der Einsatz technischer Mittel dann legitim ist, wenn er es erlaubt, irgendwelche (bestehenden) wissenschaftlichen Probleme besser zu lösen, denn diese Antwort lässt die Frage unbeantwortet, wie technische Mittel zu beurteilen sind, die vor allem neue Problemfelder erschließen, nicht alte Probleme besser lösen.

Unzureichend wäre es aber auch, den Einsatz technischer Mittel in den Geisteswissenschaften auf jeden Fall gut zu heißen, wenn dadurch bekannte Probleme besser gelöst werden oder neue Problemfelder erschlossen werden. Diese - nur oberflächlich plausible – Antwort ignoriert nämlich, dass die Problemauswahl wertgesteuert erfolgt, und sie verkennt zudem, wie wissenschaftliche Verdrängungsprozesse realiter ablaufen. Sie öffnet darüber hinaus einem positivistischen Methodenfetischismus Tür und Tor, der - grob gesprochen – darin besteht, dass die Methode zum Kriterium der Relevanz erhoben wird und die Phänomene nicht mehr (oder nur noch durch den Blickwinkel einer bestimmten Methode) zur Kenntnis genommen werden.

Am Beispiel der analytischen Philosophie, (und konkret dem Einsatz spieltheoretischer Modelle in der Analystischen Philosophie), möchte ich vor Augen führen, wie Verdrängungsprozesse zwischen wissenschaftlichen Schulen ablaufen. Solche strategischen Verdrängsprozesse, die man in Abgrenzung zum natürlichen Paradigmenwechsel als "imperialistisch" bezeichnen kann, leisten natürlich nicht nur besonders technische Ansätze, aber mit technischen Ansätzen funktioniert das aus verschiedenen Gründen besonders gut. Die typischen Etappen eines wissenschaftlichen Verdrängungsprozesses sind diese:

1. Zunächst wird eine neue wissenschaftliche Methode propagiert, in diesem Fall die Spieltheorie, mit der man angeblich sozialphilosophische Probleme wissenschaftliche viel genauer (nämlich mit mathematischer Präzision) behandeln kann als das zuvor der Fall war.
2. Anfangs funktioniert das noch nicht besonders gut. Aber jede wissenschaftliche Methode muss sich ja erst einmal die Chance bekommen, sich zu entwickeln. Speziell bei technischen Ansätzen erweist sich der anscheinend unbezwingbare Mythos des Vorsprungs durch Technik als sehr hilfreich für "die Sache". Zudem tun sich Kritiker aus dem Fach schwer einen Ansatz anzugreifen, dessen hochtechnische Einzelheiten sie nicht begreifen, selbst wenn sie die Dürftigkeit der Ergebnisse nicht übersehen können.
3. Der entscheidende Durchsetzungstrick besteht nun darin, dass - anders als die oben beschriebene zweite Antwort stillschweigend unterstellt - ein fairer Vergleich mit den früheren Ansätzen hinsichtlich der vermeintlich überlegenen Problemlösungskapazität gar nicht mehr stattfindet. Hat die neue wissenschaftliche Schule erst einmal eine kritische Masse erreicht (d.h. haben genug Leute siche ihr angeschlossen, um eigene Fachzeitschriften herauszugeben, die Gutachten dazu beizusteuern, und schließlich auch die Besetzung von Professuren entscheidend zu beeinflussen), dann braucht sie sich dem Wettbewerb mit dem Verlierer entweder gar nicht mehr zu stellen, oder sie kann die Regeln bestimmten, nach denen er geführt wird. (Im Fall der analytischen Philosophie wird dazu die Verlierposition gerne für unwissenschaftlich, sprachlich unklar oder frei von Argumenten eklärt, was viel bequemer ist als die Argumente zu kritisieren.)
4. Der wesentliche Unterschied zwischen einem legitimen Paraidgmenwechsel - wie von Kuhn beschrieben - und einem imperialistischen Durchsetzungsprozess besteht darin, dass im letzteren Fall die Vergrößerung der Problemlösungskapazität (ein wesentliches Merkmal des Kuhn'schen "Paradigmenwechsels", der ansonsten auch teilweise irrational verläuft) nur scheinbar stattgefunden hat. Der Prozess ist dann vollendet, wenn eine neue Generation von Wissenschaftlern erzogen worden ist, die mit dem erfolgreich verdrängten Ansatz gar

nicht mehr in Berührung gekommen sind und daher den relativ defizitären Charakter der Verdränger-Schule auch nicht bemerken können.

5. Speziell im Fall der Spieltheorie in der Sozialphilosophie ist zu beobachten, dass:
   - a) der Kontakt zur Empirie, insbesondere der Feldforschung verloren gegangen (oder nie recht hergestellt worden) ist. M.a.W.: Die Phänomene werden nicht mehr zur Kenntnis genommen (das zweite Hauptmerkmal des positivistischen Methodenfetischismus).
   - b) die Ergebnisse purer Modellstudien, insbesondere von sehr leicht anzufertigenden Computersimulationen ohne empirischen Bezug als publikationswürdig angesehen werden. M.a.W.: Die Methode wird zum Kriterium der Relevanz gemacht (das erste Hauptmerkmal des positivitischen Methodenfetischismus).

Der Vorgang gleicht einer Gehirnamputation, gegen die sich der Patient zunächst vielleicht sträubt. Ist sie aber erst einmal vollzogen, so ist er gerade auf Grund dieser Operation gar nicht mehr in der Lage zu registrieren, dass irgendetwas vorgefallen ist. (An dieser Stelle kann man natürlich fragen, wie man denn dann solche Verdrängungsprozesse überhaupt als solche identifizieren und kritisieren kann, wenn es doch die Fachleute selbst am Ende nicht mehr können? Man kann, wenn man bereit ist, sich der Mühe zu unterziehen, die historische ältere verdrängte mit der historisch neueren verdrängenden Position unvoreingenommen zu vergleichen. Das setzt allerdings historische Bildung oder zumindest die Befähigung dazu voraus.)

Inwiefern betrifft dies nun die digitalen Geisteswissenschaften? Bestimmte Bereiche der digitalen Geisteswissenschaften – vornehmlich solche, wo der Computer tatsächlich vor allem Hilfsmittel ist, wie z.B. beim elektronischen Publizieren oder beim transkribieren und editieren - sind davon nicht betroffen. Aber es gibt andere Bereiche, wo der Technik-Einsatz in den digitalen Geisteswissenschaften viel tiefer in den Prozess des Verstehens- und Erklärens eingreift. Sofern man die auf die Geisteswissenschaften angewandte Spieltheorie nicht als digitale Geisteswissenschaften gelten lassen will (warum wird sie eigentlich nicht dazu gezählt?), dann könnte man hier das Distant-Reading oder auch die Netzwerkanalysen in der Literaturwissenschaft anführen (was nicht heisst, dass diese noch jungen Ansätze sich schon so sehr verrant haben, wie die analytische Philosophie mit der Spieltheorie).

Wenn man bereit ist zuzugestehen, dass der Technik-Einsatz in der beschriebenen Weise schief gehen kann - ähnlich wie eine Revolution ja auch auf die Weise scheitern kann, dass sie zwar die Herrscher erfolgreich beseitigt, dann aber noch schlimmere an ihre Stelle setzt - dann bleibt immer noch die Frage anhand welcher Kriterien man den legitimen und sinnvollen Technikeinsatz von einem fehlgeleiteten unterscheiden kann. Diese Frage ist deshalb nicht pauschal zu beantworten, weil ihre Antwort von dem Fachgebiet, der konkreten Technologie und - wie oben angedeutet - auch von Relevanzfragen d.h. Fragen denach abhängt, was ein legitimes und wichtiges Problem in einem Wissenschaftsfach ausmacht. Relevanzfragen sind aber immer auch Wertfragen und können so gesehen vielleicht gar nicht objektiv entschieden werden.

Dennoch kann man womöglich Faustregeln dafür aufstellen:

1. Aus welchen Gründen auch immer eine bestimmte wissenschaftliche Problemstellung als relevant erachtet wird, es sollte nicht bloß aus dem Grund sein, dass sie die Verwendung einer interessanten Technologie erlaubt. Das klingt trivial, ist es in der Praxis aber nicht immer, besonders dann nicht, wenn der Technikeinsatz mit Forschungsgeldern prämiert wird. Eine zuzugestehende Ausnahme sind Forschungsprojekte, die rein der Methodenentwicklung dienen.
2. Technik ist in den Geisteswissenschaften nicht das Wesentliche. Wenn sie die Arbeit nicht einfacher macht oder ihr Einsatz nicht zu einem Surplus an Erkenntnis führt, der den Aufwand rechtfertigt, dann ist sie Fehl am Platze. (Auch trivial, aber nicht immer wird die Rechnung aufgemacht.)

Bibliographie

Beispiele methodenfixierter Forschung in der analytischen Philosophie und Spieltheorie:

Brian Skyrms: The Stag Hunt and the Evolution of Social Structure, Cambridge University Press 2004.

Rudolf Schüssler: Kooperation unter Egoisten: Vier Dilemmata, Scientia Nova Oldenbourg 1998.

Zur Kritik der methodenfixierter, "scientistischer" Ansätze in der analytischen Philosophie und in den Sozialwissenschaften:

John Dupré: Human Nature and the Limits of Science, Clarendon Press 2002.

Ian Shapiro: The Flight from Reality in the Human Sciences, Princeton University Press 2009.

Eckhart Arnold: How Models Fail. A Critical Look at the History of Computer Simulations of the Evolution of Cooperation, in: Catrin Missel-

horn (Ed.): Collective Agency and Cooperation in Natural and Artificial Systems, Springer 2015, p. 261-279, https://t1p.de/HMF

James Urmson: Philosophical Analysis, Clarendon Press 1960. (*Ein Klassiker!*)

Zu möglichen Beispielen und Kritik im Bereich Digital Humanities

Alan Kirkness: Es leben die Riesenschildkröten! Plädoyer für die wissenschaftlich-historische Lexikographie des Deutschen, Lexicographia 32, 2017, S. 17-137. (*bezieht sich kritisch auf https://www.dwds.de*)

David M. Berry / Anders Fagerjord: Digital Humanities. Knowledge and Critique in a Digital Age, polity press 2017. (*kritisiert im 4. Kapitel die unreflektierte Verwendung von Visualisierungen am Beispiel von:* http://www.the-everyday.net/)

Zu Paradigmenwechseln und "wissenschaftlichem Imperialismus"

Thomas S. Kuhn: The Essential Tension. Selected Studies in Scientific Tradition and Change, University of Chicago Press, 1977.

Uskali Mäki / Adrian Walsh / Manuela Fernández Pinto: Scientific Imperialism: Exploring the Boundaries of Interdisciplinarity, Routledge 2017.

Praktische Tagger-Kritik. Zur Evaluation des POS-Tagging des Deutschen Textarchivs

Herrmann, J. Berenike
berenike.herrmann@unibas.ch
Universität Basel, Schweiz

Einleitung

Der vorliegende Beitrag leistet eine Tool- und Methoden-Kritik der automatischen Auszeichnung von Wortarten (Part of Speech-, bzw. POS-Taggern) an literarischen Texten des 19. und frühen 20. Jahrhunderts. Er geht über eine rein intellektuelle Reflektion hinaus, indem er erste Schritte einer empirischen Evaluation des POS-Tagging des Deutschen Textarchivs (DTA, Berlin-Brandenburgische Akademie der Wissenschaften) und seiner praktischen Verbesserung vorlegt.

Aus der Perspektive der Digitalen Literaturstilistik und des Distant Reading sind Wortarten besonders interessante lexiko-grammatikalische Merkmale, ist ihre Verteilung doch ein wichtiger Indikator für Dimensionen wie Autorstil, Gattung und Register (z.B. Biber / Conrad 2009). POS sind vergleichsweise leicht und scheinbar valide zu bestimmen, gilt doch in der Computerlinguistik das Problem der automatischen Wortartenannotation als gelöst – auch für das Deutsche, wo eine durchschnittliche Erkennungsgenauigkeit bei 95-97% liegt (vgl. Giesbrecht / Evert 2009). Für DH-Anwender scheint es also nahe zu liegen, ihre Korpora komfortabel mit out-of-the-box-Taggern zu annotieren, oder sich bereits annotierter Korpora zu bedienen, wie zum Beispiel des DTA.

Ein genauerer Blick zeigt jedoch, dass Korpora der Geisteswissenschaft, historische wie literarische, von der sprachlichen Varietät abweichen, die den Sprachmodellen der verfügbaren Tagger zugrunde liegt, also Zeitungtexten der Gegenwart (der für das Deutsche frei verfügbare Goldstandard ist derzeit TIGER, ein Korpus von 900.000 Wörtern aus der Frankfurter Rundschau, vgl. Brants et al. 2004). In Nichtstandardvarietäten sinkt die Genauigkeit des POS-Taggings rapide (vgl. z.B. Scheible et al. 2011), und teilweise sind Aussagen über die Annotationsgenauigkeit mangels Referenzstandards gar nicht möglich. Dies betrifft auch das DTA, dessen POS-Tagging bislang nicht systematisch evaluiert wurde. Insgesamt ist die DH-Community also noch recht weit von einem Goldstandard für historische literarische narrative Texte des Zeitraums entfernt.

Unser Beitrag leistet hier einen wichtigen Schritt, indem er erste Ergebnisse zur Einschätzung der Qualität ebenso wie zur Verbesserung des Annotationstools vorlegt. Ausgehend von dem Ziel unser Korpus der Literarischen Moderne (KOLIMO http://kolimo.uni-goettingen.de/) valide mit POS auszuzeichnen, haben wir eine Stichprobe (N= 9.065) des DTA manuell nachannotiert. Unsere Methode verbindet einen Tagger-Vergleich mit einer händischen Analyse. Dabei werden folgende Ziele verfolgt:

- Eine erste Evaluation des POS-Tagging des DTA für den Zeitraum 1800-1930 im Vergleich mit der gegenwärtigen Generation der POS-Tagger;
- Der heuristische Aufweis von interessanten Fällen, die Forschungsdesiderate für Linguistik und Literaturwissenschaft aufzeigen;
- Die Verbesserung des Sprachmodells und so eine Domänen-Adaptation der Tagger.

Studie

Prozedur

Die Evaluation des POS-Taggings wurde durchgeführt auf einer randomisierten Stichprobe des DTA, die aufgrund unseres Forschungsinteresses auf narrative Texte mit Publikationsdatum ab 1800 beschränkt war, wobei sowohl fiktionale wie auch nicht-fiktionale Texte berücksichtigt wurden (ausschlaggebend waren die Metadaten zur Erstveröffentlichung und Gattung im Header des DTA). Die Grundgesamtheit der aus dem DTA entnommenen Stichprobe umfasste N= 64.924.458 Tokens, die der händisch annotierten Tokens umfasste n= 9.065 Tokens/POS-Tags, also 0,014%). Die Stichprobe wurde in ihrer tokenisierten und normalisierten Form aus dem DTA übernommen (vgl. DTA). Der Taggervergleich nutzte neben dem DTA-Tagger *moot* (Jurish / Würzner 2013) den TreeTagger (Schmid 1994), MarMoT (Müller et al. 2013) sowie den Perceptron-Tagger (Rosenblatt 1958), also solche Tagger, die in der digitalen Textanalyse häufig verwendet werden. Input war für alle Tagger dieselbe Stichprobe aus dem DTA.

Das Tagging wurde durch vier studentische Hilfskräfte besorgt, wobei iterative Analysen und finale Annotation durch die PI betreut wurden. Mit einem Skript wurden csv-Tabellen erstellt, die die Tokens (fortlaufende Wortformen, inklusive Interpunktion) und POS-Tags in einem *Keyword-in-Context*-Format präsentieren. Abbildung 1 zeigt einen Ausschnitt der Ansicht des Annotationstools: jede Zeile enthält neben dem Token (Wort) die jeweiligen POS-Tags, das Lemma, den linken und rechtem Satzkontext, sowie einen größeren Satzkontext, Angaben zu Werktitel, Autor, und Erscheinungsdatum. Bei der Analyse wurde jeweils nur die Abweichungen zum (von *moot* zugewiesenen) DTA-Tag händisch in eine gesonderte Zelle (*newtag*) eingefügt, ebenso wie ein fakultativer Kommentar des Coders.

DTA-Tag	TT-Tag	MM-Tag	Perceptron-T	Lemma	newtag	Prefix	Wort	Suffix	Sentence_ne	sentence_inc	index	filename	Kommentar
$(	ADJA	XY	ADJD	--		Das GeprÃ¤	â€"	wer mÃ¶chte	ein Gedicht i		82	9 laube_europa0101_1833.t	
$.	$.	$.	$.	,		Versuch	,	X. 2. 416 ff .	Moser â€™s		3155	1 klueber_voelkerrecht01_1	
$.	$.	$.	$.	,		Das GeprÃ¤	,	â€" wer mÃ¶	ein Gedicht i		82	8 laube_europa0101_1833.t	
$.	$.	$.	$.	,		Trotzdem bÃ¼	,	so lange er e	Unterkauflei		2121	5 berg_ostasien01_1864.txt	
$.	$.	$.	$.	.			Versuch . X. .				3155	6 klueber_voel flasch Token	
$.	$.	$.	$.	.		Auf den Endc	.		kein Schub Ã		1722	13 wanderley_bauconstructi	
$.	$.	$.	$.	.		Das GeprÃ¤	.		ein Gedicht i		82	20 laube_europa0101_1833.t	
$.	$.	$.	$.	.		Trotzdem bÃ¼	.		Unterkauflei		2121	11 berg_ostasien01_1864.txt	
$.	$.	$.	$.	.		8. Preis 1 Rth	.		, was Noth lÃ		261	6 hc_1671710_1812.txt	
ADJA	ADJA	ADJA	ADJA	8.			8.	Preis 1 Rthlr.	, was Noth lÃ		261	0 hc_1671710_1812.txt	
ADJD	ADJD	ADJD	ADJD	eintrÃ¤glich		Trotzdem bÃ¼	eintrÃ¤glich	blieb .	Unterkauflei		2121	9 berg_ostasien01_1864.txt	
ADJD	ADJD	ADJD	ADJD	gefÃ¤hrlich		Das GeprÃ¤	gefÃ¤hrlich	. â€" wer mÃ¶	ein Gedicht i		82	7 laube_europa0101_1833.t	
ADV	ADV	ADV	ADV	2.	CARD	Versuch . X. .	2.	416 ff .	Moser â€™s		3155	3 klueber_voelkerrecht01_1	
ADV	ADV	ADV	ADV	lange		Trotzdem bÃ¼	lange	er eintrÃ¤gli	Unterkauflei		2121	7 berg_ostasien01_1864.txt	
ADV	ADV	ADV	ADV	so		Trotzdem bÃ¼	so	lange er eint	Unterkauflei		2121	6 berg_ostasien01_1864.txt	
APPR	APPR	APPR	APPR	auf			Auf	den Enden d	kein Schub Ã		1722	0 wanderley_bauconstructi	
APPR	APPR	APPR	APPR	in		Das GeprÃ¤	in	die Gefahr ei	ein Gedicht i		82	12 laube_europa0101_1833.t	
APPRART	APPRART	APPRART	APPRART	zur		Auf den Endc	zur	UnterÃ¿tÃ¾t	kein Schub Ã		1722	9 wanderley_bauconstructi	
ART	ART	ART	ART	d			Das	GeprÃ¤nge c	ein Gedicht i		82	0 laube_europa0101_1833.t	
ART	ART	ART	ART	d		Auf	den	Enden dieÃ¿	kein Schub Ã		1722	1 wanderley_bauconstructi	
ART	ART	ART	ART	d		Das GeprÃ¤	den	Glanz wegwe	ein Gedicht i		82	17 laube_europa0101_1833.t	
ART	ART	ART	ART	d		Auf den Endc	der	SparrenroÃ¿	kein Schub Ã		1722	11 wanderley_bauconstructi	
ART	ART	ART	ART	d		Das GeprÃ¤	der	TÃ¤tuÃ¿chun	ein Gedicht i		82	2 laube_europa0101_1833.t	
ART	ART	ART	ART	d		Trotzdem bÃ¼	der	Schleichhanc	Unterkauflei		2121	2 berg_ostasien01_1864.txt	
ART	ART	ART	ART	d		Auf den Endc	die	beiden Fette	kein Schub Ã		1722	6 wanderley_bauconstructi	
ART	ART	ART	ART	d		Das GeprÃ¤	die	SchauÃ¿spiele	ein Gedicht i		82	5 laube_europa0101_1833.t	
ART	ART	ART	ART	d		Das GeprÃ¤	die	Gefahr einge	ein Gedicht i		82	13 laube_europa0101_1833.t	
CARD	CARD	CARD	CARD	1		8. Preis	1	Rthlr. 12 Gr	, was Noth lÃ		261	2 hc_1671710_1812.txt	
CARD	CARD	CARD	CARD	12		8. Preis 1 Rth	12	Gr .	, was Noth lÃ		261	4 hc_1671710_1812.txt	
CARD	CARD	CARD	CARD	416		Versuch . X. .	416	ff .	Moser â€™s		3155	4 klueber_voelkerrecht01_1	

Abbildung 1: Screenshot POS-Annotationstool (Ausschnitt)

Dem automatischen wie händischen Tagging lag das Tagset des STTS (Schiller et al. 1999) zugrunde. Wo angezeigt, wurden im Projekt zusätzliche Regeln für der Handhabung des STTS-Manuals vereinbart und dokumentiert. Hierzu gehört auch u.a. die systematische Einbindung eines korpusbasierten Wörterbuchs (http://www.duden.de/) bei Eigennamen und Fremdwörtern.

Das Tagging wurde in drei Phasen durchgeführt. Primäres Ziel des Taggings war es, die Genauigkeit von *moot* auf der genannten Stichprobe gegen manuelles und automatisches Tagging (TT, MarMot, Perceptron) zu evaluieren. Die Phase I diente neben der Erarbeitung und Ergänzung des Tagging-Manuals und dem Einarbeiten der Coder einer ersten quantitativen und qualitativen Analyse des *moot*-Taggings. Die untersuchte Stichprobe umfasste N=100 randomisiert extrahierte Sätze (N=3.635 Tokens). Jedes Token wurde bezüglich des zugewiesenen POS-Tag durch alle Coder unabhängig überprüft und ggf. korrigiert.

In Phase II wurden dieselben Coder, Tagger und dieselbe Software eingesetzt, ebenso wie die in Phase 1 erarbeiteten Tagging-Guidelines. Anders als in der ersten Phase wurden jedoch nicht ganze Sätze, sondern jeweils einhundert Token pro POS-Kategorie aus dem DTA extrahiert. Dadurch wurde eine Ungleichverteilung der einzelnen POS-Kategorien, welche in natürlichen Sätzen gegeben ist (vgl. Evert 2006; Kilgarriff 2005), vermieden. Für fünfundfünfzig POS-Kategorien des STTS wurden jeweils n=100 Wort/Token-POS-Paare sortiert nach STTS-Tag annotiert. Dies entspricht einer Grundgesamtheit von N=5.500 Tokens. Jedes Token wurde von zwei Codern unabhängig annotiert.

In der anschließenden Diskussionsphase wurden die strittigen Fälle besprochen und finale Annotationen erarbeitet. Zu diesem Zweck wurden Statistiken für die Tags (über Coder und Tagger) analysiert und Nichtübereinstimmung der vergebenen Tags identifiziert. Für die statistische Evaluation wurde die Interrater-Reliabilität als *Agreement* und Cohen‘s Kappa berechnet (Package „irr“ in R Version 3.3). Darüber hinaus wurde für die erste Tagging-Phase ein Fleiss’-Kappa für die Coder berechnet (dies steht für Phase 2 noch aus).

Ergebnisse

Die Ergebnisse in Tabelle 1 basieren für Phase I auf den finalen Annotationen und für Phase II zum momentanen Zeitpunkt auf etwa der Hälfte der finalen Annotation. Sie zeigen für *moot* verglichen mit dem Referenzstandard eine Gesamtgenauigkeit von 90,16% (TreeTagger 80,88%; MarMoT 83,99%; Perceptron 79,75%). Dies ist eine niedrige Gesamtgenauigkeit gemessen an 98,6%

zur modernen Standardvarietät (Brants 2000), entspricht aber in etwa den von Scheible et al. (2011) für das Frühneuhochdeutsche erhobenen 91,6%.

		moot	Tree-Tagger	Mar-MoT	Percept-ron
Phase I	Mittel-wert Coder	93,44	84,75	86,43	86,23
	Refe-renz-stan-dard	92,93	85,47	87,18	86,35
Phase II	Mittel-wert Coder	84,93	67,87	72,32	62,34
	Refe-renz-stan-dard	82,84	68,73	72,29	62,33
Kombi-nierter Wert Phase I & II*	Mittel-wert Coder	89,17	75,92	79,11	73,72
	Refe-renz-stan-dard	90,16	80,88	83,09	79,76

Tabelle 1: *Genauigkeit der Tagger bzgl. Coder (Mittelwert über Coder vor Diskussion) und Referenzstandard (nach Diskussion) in Prozent*

*Es handelt sich nicht um den Mittelwert von Phase I und II sondern um eine gewichtete Statistik, die die unterschiedlichen Stichprobengrößen (zum Zeitpunkt der Rechnung) einbezieht.

Die Übereinstimmung zwischen Codern vor der Diskussion und Referenzstandard ist hingegen vergleichsweise hoch, auch wenn sie in der zweiten Tagging-Phase etwas abfällt (*Agreement* Phase I = 95,47 – 98,13%, *Agreement* Phase II = 95,56 – 96,22%, Cohen‘s Kappa Phase I = 0,95 – 0,98, Cohen‘s Kappa Phase II = 0,95 – 0,96). Gleiches gilt für die Interrater-Reliabilität (Übereinstimmung zwischen Codern vor Diskussion) obwohl die Differenz zwischen den beiden Phasen größer ist (*Agreement* Phase I = 94,14 – 95,20%, *Agreement* Phase II = 89,45 – 92,64%, Cohen‘s Kappa Phase I = 0,94 – 0,95, Cohen‘s Kappa Phase II = 0,89 – 0,92). Das Fleiss’ Kappa weist mit 0,94 einen hohen Wert auf. Die Coder annotieren in beiden Phasen also genauer als *moot*.

Obwohl die Gesamtergebnisse noch ausstehen, könnte der Unterschied zwischen den Phasen tentativ damit erklärt werden, dass Phase II mehr problematische Tags annotiert, die eine niedrigere Distribution haben und im per-Satz-Tagging seltener auftreten. Für die einzelnen POS-Kategorien variiert die Genauigkeit zwischen 0% und 100%, wobei *moot* die höchste mittlere Genauigkeit (Mittelwert = 88,65%) und niedrigste Streuung (Standardabweichung = 17,25) aufweist (TreeTagger = 67,05 ± 28,52, MarMoT = 73,41 ± 27,49, Perceptron = 62,66 ± 31,37). Eine detaillierte Analyse der einzelnen POS-Tag-Kategorien zeigt, dass *moot* in den meisten, aber nicht allen, POS-Kategorien die besten Ergebnisse erzielt (vgl. Tabelle 2).

	moot	TreeTag-ger	MarMoT	Percept-ron
ADJA	94,5	93	93,5	95
ADJD	85,37	79,67	78,05	76,42
ADV	81,2	72,93	75,56	68,42
NE	75,25	61,87	63,55	87,63
NN	92,81	93,46	92,32	91,67

Tabelle 2: *Genauigkeit einiger STTS-Tags über Tagger (in Prozent)*

Diskussion

In den Gruppendiskussionen konnten Probleme identifiziert werden, die vornehmlich bei den Taggern lagen (z.B. bei Abkürzungen, Relativpronomen). Es traten aber auch Fälle auf, in denen die STTS-Guideline nicht präzise genug ist (z.B. bei Vergleichspartikeln, Possessivpronomen, Indefinitpronomina). Dabei war die Analyse der Disagreements eine produktive Heuristik, um (computer-)linguistisch und literarturwissenschaftlich interessante Fälle aufzuwerfen. So scheint gerade in literarischen Fällen eine Ambiguität (etwa zwischen Adjektiv und Verb bei Partizipien) geradezu intentional. Ähnlich *und in „Bravo! Warum denn nicht? Bravo! Und wieder Bravo!“ (Kafka, Der Prozess)*, welches als Konjunktion, aber auch als Diskurspartikel interpretiert werden kann.

Insgesamt zeigt unsere praktische Taggerkritik, dass auch eine scheinbar gelöste NLP-Aufgabe wie die Wortartenauszeichnung kein Solitär ist, auf den geistes- und literaturwissenschaftliche Projekte ohne genauere Prüfung bauen sollten. Unsere Ergebnisse zeigen trotz der hochqualitativen Vorverarbeitung des DTA eine Fehlerrate von ca. 9% an, die allerdings stark nach POS-Tag variiert. Die diachrone wie synchrone Heterogenität des literarischen Diskurses führt generische POS-Tagger bislang fast zwangsläufig an ihre Grenzen, durch historische Sprachformen, aber auch die Vielfalt der Gattungen, Erzähltechniken und

kreative Lexik und Syntax. Zukünftig bieten sich hier wohl zwei Wege an: zum einen die fortlaufende Verbesserung von generischen Tools, zum anderen gerade aber auch die Feinabstimmung der Tools für spezifische Anwendungen, mit flexibel ansprechbaren Tagging- und Sprachmodellen. So haben wir unsere Annotation an *moot* zurückgespielt, um das spezifische Sprachmodell zu verbessern. Die Ergebnisse unseres Taggervergleich deuten zudem für bestimmte Tags auf die Nützlichkeit eines Ensemble-Taggings hin, bei dem verschiedene Algorithmen verschränkt werden (van Halteren et al., 2001).

Bibliographie

Biber, Douglas / Conrad, Susan (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Brants, Thorsten (2000): "Inter-annotator agreement for a German newspaper corpus", in: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000),* Athens, Greece. European Language Resources Association (ELRA). *http://www.lrec-conf.org/proceedings/lrec2000/pdf/333.pdf* .

Brants, Sabine / Dipper, Stefanie / Eisenberg, Peter / Hansen-Schirra, Silvia / König, Esther / Lezius, Wolfgang / Rohrer, Christian / Smith, George / Uszkoreit, Hans (2004) : "TIGER: Linguistic interpretation of a German corpus", in: *Research on Language and Computation*, 2(4): 597–620.

Berlin-Brandenburgische Akademie der Wissenschaften. *Deutsches Textarchiv.* http://www.deutschestextarchiv.de/ [Letzter Zugriff 13.01.2018].

Giesbrecht, Eugenie / Evert, Stefan (2009). "Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus", in: Alegria, I. / Leturia, I. / Sharoff, S. (eds.): *Proceedings of the 5th Web as Corpus Workshop* (WAC5), San Sebastian, Spain.

Evert, Stefan (2006): "How random is a corpus? The library metaphor", in: Zeitschrift für Anglistik und Amerikanistik 54.2: 177-190.

Jurish, Bryan / Würzner, Kay-Michael (2013). "Word and sentence tokenization with Hidden Markov Models", in: JLCL 28(2): 61-83.

Kilgarriff, Adam (2005): "Language is never, ever, ever, random", in: *Corpus linguistics and linguistic theory* 1(2): 263-276.

Müller, Thomas / Schmid, Helmut / Schütze, Hinrich (2013): "Efficient higher-order CRFs for morphological Tagging", in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Rosenblatt, Frank (1958): "The perceptron: A probabilistic model for information storage and organization in the brain", in: *Psychological Review*, 65(6): 386.

Scheible, S. / Whitt, R. J. / Durrell, M. / Bennett, P. (2011): "A gold standard corpus of Early Modern German" in: *Proceedings of the 5th Linguistic Annotation Workshop* 124–128.

Schmid, Helmut (1994). "Probabilistic part-of speech Tagging using decision trees", in: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Halteren, H. / Daelemans, W. / Zavrel, J. (2001): "Improving accuracy in word class tagging through the combination of Machine Learning systems", in: *Computational Linguistics, 27.*

Principles Aiding in Reading Abbreviations in OldGeorgian and Latin

Hoenen, Armin
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Samushia, Lela
samushia@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Introduction

A project on Georgian Epigraphy [1] in the context of which this article is grounded brought our attention to two phenomena which seem inseparably intertwined with the epigraphic record: abbreviations and gaps. In Old Georgian, abbreviations are very prominent and yet very heterogeneous. That is more so than in other (historical) languages. One finds many abbreviations for the same word, compare Boeder (1987). In this article, we assess the research question why this is so by trying to highlight properties of Old Georgian abbreviations, which can also be used as an aid to read them. We use Old Georgian abbreviations in inscriptions and manuscripts from the Titus website, Gippert (1995) under more entered within the Georgian National Corpus [2]. In order to take a somewhat wider perspective and enable data hungry technologies to aid in the analyses of the aforementioned phenomena [3], we decided to com-

pare the Georgian record with one of the largest digital epigraphic records available, classical Latin, for which we use data from the Epigraphic database Heidelberg [4]: last accessed December 2016. The Latin data is several orders of magnitude larger than the Georgian one. While Latin contains roughly 70,000 inscriptions, Georgian features 91, so we decided to additionally analyse abbreviations in Old Georgian manuscripts. We ended up with roughly 4,000 abbreviations for Old Georgian (roughly 1,100 from the inscriptions) and roughly 170,000 for Latin.

There are several partly overlapping typologies of abbreviations, compare Marchand (1969); Kreidler (1979); McArthur (1988); McArthur and McArthur (1992); Rúa (2004); Driscoll (2009). (Carroll, 2003, p.205) remarks: ”Unfortunately, universally accepted standards for many abbreviations and acronyms do not exist”.

Position

The first investigation concerns the position of the letters in the extension, which are maintained in the abbreviation. That is <precip> as an abbreviation for <precipitation>, will be converted into 1-2-3-4-5-6. We look at the type and token levels for Old Georgian and Latin, see Table 1.

Text Type	p_0=first	p_l=last	Suspension	Contraction
Old Georgian Inscriptions	0.998	0.895	0.037	0.158
Old Georgian Manuscripts	1.0	0.814	0.185	0.344
Old Georgian Inscriptions (types)	0.998	0.902	0.012	0.076
Old Georgian Manuscripts (types)	1.0	0.979	0.016	0.166
Latin Inscriptions	0.998	0.037	0.424	0.003
Latin Inscriptions (types)	0.987	0.183	0.12	0.01

Table 1: Proportions of abbreviations starting in the first letter (first column), ending in the last (second column), being suspensions (third column) or contractions (fourth column). 'types' here counts each abbreviation - expansion tuple uniquely.

We find that abbreviations usually start in the first letter although this might be part of a prefix in both Latin and Old Georgian. This allows for keeping parafoveal preview information intact, see for instance Slattery et al. (2011); Rayner et al. (2012). Chanceaux et al. (2013) summarizing Dandurand et al. (2011) find that ”initial letters provide more information with respect to word identity than any other letter position”. Initital letters are together with the last letter the ”most visible” letters of a word, Dandurand et al. (2011), which means under more that due to the adjacent spaces, they can more easily be recognized.

Secondly, Old Georgian uses more contractions (abbreviations by first and last letter, a Christian abbreviation tradition which entails the question of how to spiritually correctly contract affixed words inducing a dilemma once many affixes are present), Latin suspensions (abbreviation by the first n letters. Figure 1 shows the patterns of occurrence of positions, given a certain word length. The y-axis represents the percentage of abbreviations (regardless of their lengths) that contain the position specified on the x-axis. In a normalized plot, all wordlengths from 4 to 11 are plotted together. For Old Georgian there is a gap after some first letters. We conjecture that the end of the word stem is most unlikely to occur in abbreviations. While in Latin, suffixes are often left out, in Georgian, which has an agglutinative morphology, this would lead to considerable difficulties in relating the actual word to the context since some suffixes carry information which in Latin are expressed by independent pre- or postpositions.

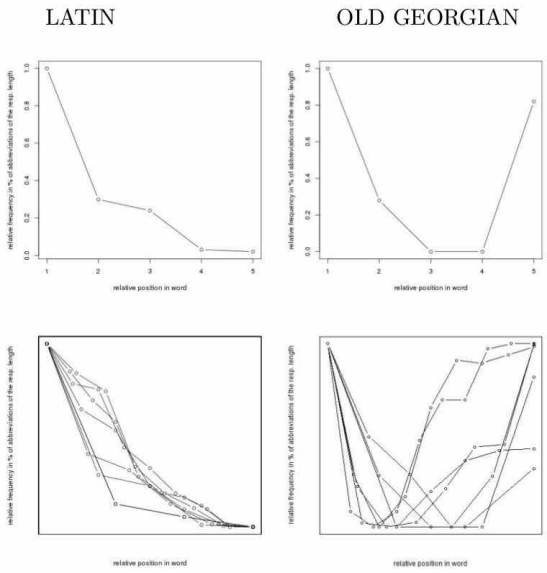


Figure 1: Letter position incidence in abbreviations. First row: for words of length 5 (24, 569 data points in Latin and 401 in Old Georgian). Second row: for words of length 4 to 11 (overlay). In the upper left graphic, for all abbreviated words of length 5 in Latin, the first letter was always present in the abbreviation, the second in roughly 30% of the cases, the third in slightly less cases, the fourth and fifth almost never. In Old Georgian, an alternation with 2 dominant patterns appears: one for shorter words with probably one suffix and one for words with 2 suffixes which then have more letters towards the relative middle.

Vowels & Consonants

Now, we look at how many vowels and consonants are retained in abbreviations, see Table 2. As one can see, in all contexts the ratio of vowels in the abbreviations is clearly lower than in the extensions. Information theory, see for instance Shannon (1948) provides a stable basis for the interpretation of these results. a) the class of vowels is smaller than the class of consonants and b) vowels are much more frequent not only because of a) but also because they occur in syllable nuclei and consequently a single vowel, yet not a single consonant can constitute a syllable. [5] [6] Thus, vowels are less informative and hence more easily guessable than consonants. Finally, we analysed our larger Latin data in more detail. We extract the frequency based rankings of letters and test their correlation with the ranking through probability of being retained in an abbreviation. The Spearman rank correlation suggests, that there is a significant correlation between the frequency of a consonant and its probability to be retained in an abbreviation (p-value = 0.002655, ρ 0.622807). For vowels p is above 0.01, and ρ is negative at -0.2 suggesting that for vowels, which are all quite frequent, there is no clear effect. Carreiras et al. (2009) find that vowels do not yield priming effects for recognition in opposition to consonants, which would nicely align with this finding. Thus, an inf requent consonant grapheme is more probably retained. One may deduct that consonants that are left out - if any - are more frequent than the ones retained. Knowing this and knowing that the first portion of the word root is probably represented in the abbreviation decreases the number of possible interpretations ideally by this token helping to decipher abbreviations.

Text Type	**Proportion** of vowels in extension	**Proportion** of vowels in abbreviation
Old Georgian Inscriptions	0.408	0.276
Old Georgian Manuscripts	0.42	0.275
Old Georgian Inscriptions (types)	0.41	0.271
Old Georgian Manuscripts (types)	0.412	0.347
Latin Inscriptions	0.466	0.337
Latin Inscriptions (types)	0.446	0.392

Table 2: Proportions of vowels in extensions and abbreviations. Note, that the inverse (1-value) is the proportion of consonants. Here, vowels and consonants are measured as vowel and consonant characters, which for both Latin and Old Georgian largely coincides with their phonemic class due to both writing systems being rather shallow.

Possibilities for the Abbreviator

Reading abbreviations may be understood even better when taking their generation into account. One hypothesis assumes, that someone, who wants to abbreviate a word, holds in mind all possible abbreviations and chooses from them. We try to validate this, counting all possible abbreviations per word length given suspension and contraction. This can also help estimate algorithmic complexity in abbreviating.

As we have empirically observed but not explicitly stated so far and implicitly assumed, a condition for a valid abbreviation may be that the indices of the letters must maintain ascending order, or, in other words, the original sequence of the letters is not permuted. If so, for each word length w and abbreviation length a there are $\{ w \mid choose\ a \}$ [7] possible abbreviations, since for each distinct combination only one can maintain the ascending order of elements. Now in order for understanding how many possible abbreviations there are for each word we need to add up values for all different a, where $a < w$. [8]

By simply using this binomial coefficient one can compute the numbers of possible combinations (for numbers until 10, see Table 3 leftmost numbers in cells) which gives the inner portion of Pascal's triangle. However, the outer 1s are missing,

since the extension itself is no abbreviation and neither is a sequence of zero letters, the sum is thus 2^w - 2. The increase in possibilities is linear and quick, for a word of length 15, there are 32,766 possibilities how it could be abbreviated.

If we additionally fix the first letter, we count only all combinations containing element '1', which must then be{ *w-1* \choose *a-1* }, since we have fixed the first letter and from the remaining *w - 1* letters, we can choose any *a - 1* remaining elements of the abbreviation. This restricts the possible numbers already considerably. Overall, we halve possibilities, so 2^{w-1} - 1 becomes the sum formula, which would still leave us with 16,383 possibilities for a word of 15 letters.

In case of a contraction, one fixes the first and the last letter. Then again, results are halved with one letter long abbreviations excluded, hence the sum relates to *w* by 2^{w-2} - 1. For a word of length 15 still 8,191 possibilities would be left.

Numbers remain so high towards the end of the scale that it seems improbable that someone who abbreviates a word be aware of all of the possibilities simulateneously at decision time. One also sees that contraction is quite effective in restricting possible abbreviations and might thus considerably speed-up abbreviating and decipherment/reading of abbreviations.

	1	2	3	4	5
2	2/1/0				
3	3/1/0	3/2/1			
4	4/1/0	6/3/1	4/3/2		
5	5/1/0	10/4/1	10/6/3	5/4/3	
6	6/1/0	15/5/1	20/10/4	15/10/6	6/5/4
7	7/1/0	21/6/1	35/15/5	35/20/10	21/15/10
8	8/1/0	28/7/1	56/21/6	70/35/15	56/35/20
9	9/1/0	36/8/1	84/28/7	126/56/21	126/70/35
10	10/1/0	45/9/1	120/36/8	210/84/28	252/126/56

	6	7	8	9	Σ
2					2/1/0
3					6/3/1
4					14/7/3
5					30/15/7
6					62/31/15
7	7/6/5				126/63/31
8	28/21/15	8/7/6			254/127/63
9	84/56/35	36/28/21	9/8/7		510/255/127
10	210/126/70	120/84/56	45/36/28	10/9/8	1022/511/255

Table 3: Numbers of possible abbreviations per word length / when first letter is fixed / when first and last letters are fixed. Rows have word length, columns abbreviation length.

Discussion and Conclusion

Various analyses (not only the presented) conducted have shown two properties of abbreviations which can help understand and read Old Georgian and Latin abbreviations.

1. the overwhelming majority of abbreviations contains the first letter of the extension
2. morphological type is an important factor for abbreviating, for Old Georgian the end of the word stem is likely to disappear whereas suffixes tend to be represented, the last letter being likely for the contraction principle
3. for Latin we found that consonants are more likely to occur in an abbreviation than vowels, less frequent consonants more than frequent ones

Combinatorial evidence suggests that keeping in mind all possible abbreviations even in case of contraction is unlikely for longer words. Considering that Old Georgian as an agglutinative language produces long words this may partly explain why there is larger variety in the abbreviation landscape, since in each abbreviation process another possible abbreviation may have surfaced.

Fußnoten

1. https://www.cedifor.de/en/cedifor/current-pilot-projects/pilot-projects/digitale-erschliessung-epigraphischer-denkmaeler

2. http://titus.fkidg1.uni-frankfurt.de/texte/etcg/cauc/ageo/inscr/carcera/carce.htm/ and http://gnc.gov.ge/gnc/page
3. See also Hoenen & Samushia (2016)
4. http://edh-www.adw.uni-heidelberg.de/home
5. In principle, all but sign languages should make use of vowels, at least empirically in the Word atlas of language structures, Dryer and Haspelmath (2013) (see wals.info), there are no languages with less than 2 vowel qualities. Furthermore the lowest consonant vowel ratio in 563 languages has still more consonants (10) than vowels (9), see Maddieson (2013b,a).
6. A third observation on vowels could be related: vowels are synchronically (dialects) and diachronically (language change) less stable.
7. We write formulas in LaTeX syntax in order to avoid for simple formulae to be given too much space.
8. Note, that an abbreviation can be understood as a k-skip-n-gram, Guthrie et al. (2006). The first numbers in the summing column of the table then coincide with the numbers (sum of) of all possible k-skip-n-grams for w (and all possible n and k).

Bibliography

Boeder, W. (1987). Versuch einer sprachwissenschaftlichen Interpretation der altgeorgischen Abkürzungen. In: *Revue des études géorgiennes et caucasiennes* 3:33–81.

Carreiras, M., Duñabeitia, J. A., and Molinaro, N. (2009). Consonants and vowels contribute differently to visual word recognition: Erps of relative position priming. I: *C erebral Cortex* 19(11):2659–2670.

Carroll, J. (2003). *Oxford handbook of computational linguistics*. Oxford University Press.

Chanceaux, M., Mathôt, S., and Grainger, J. (2013). Flank to the left, flank to the right: Testing the modified receptive field hypothesis of letter-specific crowding. In: *Journal of Cognitive Psychology* 25(6):774–780.

Dandurand, F., Grainger, J., Duñabeitia, J. A., and Granier, J.-P. (2011). On coding non-contiguous letter combinations. In: *Frontiers in psychology* 2:136.

Driscoll, M. (2009). Marking up abbreviations in old norse-icelandic manuscripts. In: *Medieval Texts–Contemporary Media*. Ibis.

Dryer, M. S. and Haspelmath, M., (eds) (2013). WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gippert, J. (1995). TITUS. Das Projekt eines indogermanistischen Thesaurus (”TITUS. The project of an Indo-European thesaurus”). In: *LDV- Forum* 12(2):35–47.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In: *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006):* 1–4.

Hoenen, A. and Samushia, L. (2016). Gepi: An Epigraphic Corpus for Old Georgian and a Tool Sketch for Aiding Reconstruction, In: *JLCL* 31 (2):25–38.

Kreidler, C. W. (1979). Creating new words by shortening. In: *Journal of English Linguistics* 13(1):24–36.

Maddieson, I. (2013a). WALS, Consonant-Vowel Ratio. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Maddieson, I. (2013b). WALS, Vowel Quality Inventories. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Marchand, H. (1969). The categories and types of present-day English word formation: A synchronic-diachronic approach. Beck.

McArthur, T. (1988). The cult of abbreviation. In: *English Today* 4(03):36–42.

McArthur, T. B. and McArthur, F. (1992). The Oxford companion to the English language. Oxford University Press.

Rayner, K., Pollatsek, A., Ashby, J., and jr. Clifton, C. (2012). Psychology of Reading. Psychology Press, New York/Hove.

Rúa, P. L. (2004). Acronyms & co.: A typology of typologies= acrónimos y cía: una tipología de tipologías. In: *Estudios Ingleses de la Universidad Complutense* 12:109–129.

Shannon, C. E. (1948). A mathematical theory of communication. In: *Bell System Technical Journal* 27:379–423.

Slattery, T. J., Schotter, E. R., Berry, R. W., and Rayner, K. (2011). Parafoveal and foveal processing of abbreviations during eye fixations in reading: making a case for case. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(4):1022.

Quantitatives „close reading“? Vier mikroanalytische Methoden der digitalen Dramenanalyse im Vergleich.

Krautter, Benjamin
Benjamin.Krautter@gmail.com
Universität Stuttgart, Deutschland

Einführung

Jüngste Ergebnisse der computergestützten Forschung legen nahe, dass Romanfiguren – gemessen an ihrer Figurenrede – von den jeweiligen Autoren stilistisch distinktiv angelegt werden können (Hoover 2017; Fields, Bassist, Roper 2017). Versierte Autoren könnten ihren Figuren also sogenannte „distinctive voices“ einschreiben, die sich stilometrisch identifizieren lassen. Anders als bei Autorschafts-, Gattungs- oder Epochensignalen handelt es sich hierbei um ein *intratextuelles* Unterscheidungskriterium. Untersuchungsgegenstand ist somit nicht ein großes Textkorpus verschiedener Autoren, sondern ein einzelner literarischer Text. David Hoover benennt dieses Vorgehen der Textselektion und -aufbereitung ‚microanalysis‘. Er setzt sich damit nicht nur von Schlagwörtern wie ‚big data‘ ab, er betont trotz vergleichbarer quantitativer Methoden auch die Unterschiede zu Konzepten wie ‚macroanalysis‘ (Jockers 2013) und ‚distant reading‘ (Moretti 2000; 2005).

Erstaunlicherweise beschränken sich die Studien zur stilistischen Differenzierung von Figurenrede größtenteils auf Romane. Dabei ist es doch gerade die Struktur dramatischer Texte, die eine quantitative Untersuchung der Figurenrede plausibel erscheinen lässt – die Rede wird nicht von einem Erzähler sortiert, kommentiert und in einen Rahmen gebettet. Auch erste Forschungsansätze sind durchaus vorhanden: John Burrows und Hugh Craig zeigen etwa, dass einzelne Dramenfiguren sehr wohl erfolgreich einem Autorsignal zugeordnet werden können (Burrows, Craig 2012). Sie reagieren damit interessanterweise auf Kritiker, die die erfolgreiche Autorschaftsattribution von Dramentexten aufgrund der vielen verschiedenartigen Stimmen – weil es also keinen Erzähler gibt – in Frage stellen (Masten 1997).

Nachfolgend soll geprüft werden, inwieweit sich Hoovers Vorgehen (2017) zur Ermittlung distinktiver Figurenrede auch auf dramatische Texte übertragen lässt. Die Ergebnisse der stilometrischen Untersuchung werden im Anschluss durch drei weitere quantitative Analyseverfahren kontextualisiert und zugleich kritisch hinterfragt.

Distinktive Figurenrede im Drama?

Abbildung 1 [1] zeigt eine hierarchische Clusteranalyse der wichtigsten Figuren aus Gotthold Ephraim Lessings *Minna von Barnhelm, oder das Soldatenglück*. [2] Die Abbildung setzt die Redeanteile der wichtigsten Dramenfiguren, die gemäß der Aktgrenzen segmentiert wurden, stilometrisch in Relation. Grundlage der Analyse sind die Wortfrequenzlisten, die den Redeäußerungen der einzelnen Figuren entnommen werden. Mit Hilfe von ‚Cosine Delta‘, das zuverlässigere Ergebnisse als ‚Burrows’s‘ oder ‚Argamon’s Delta‘ erzielen sollte (Evert u.a. 2017), wird aus den Wortfrequenzen die relative stilistische Ähnlichkeit der Textpassagen berechnet. [3] Anders als bei Hoover erfolgt die Unterteilung der Figurenrede jedoch nicht nach Segmenten zu je 1500 Wörtern. Stattdessen werden die bereits gegebenen Aktgrenzen des Dramas zur Einteilung herangezogen. [4] Dieses Vorgehen ist weniger artifiziell, da keine künstlich normalisierten Grenzen zu setzen sind. Es ist zugleich hilfreich, um die stilometrischen Ergebnisse anhand der Bedingungen ihres Zustandekommens, etwa der Kopräsenz von Figuren, zu interpretieren. Das Vorgehen hat jedoch zum Nachteil, dass die Segmente keine einheitliche Länge aufweisen und zu kurze Abschnitte aufgrund ihrer geringen Wortzahl aus dem Korpus gestrichen werden mussten.

Die stilometrische Analyse von *Minna von Barnhelm* zeigt, dass die Ergebnisse, die Hoover sowie Fields, Bassist und Roper für ausgewählte englischsprachige Romane erzielen, nicht unmittelbar auf Lessings Drama übertragbar zu sein scheinen. Es gibt zwar vereinzelte Anzeichen für stilistisch distinktiv angelegte Figurenrede – von der männlichen Hauptfigur Tellheim gruppieren sich Akt 1 und Akt 3 in unmittelbarer Nähe, auf Paul Werner trifft das für Akt 3 und Akt 5 zu. Die Mehrheit der Redesegmente scheint sich allerdings nach einem anderen Kriterium anzuordnen. Besonders deutlich wird dies im obersten Abschnitt der Grafik: Die Redeanteile von Major

Tellheim, Minna von Barnhelm, ihrer Kammerjungfrau Franziska und die des Wirts gruppieren sich auf einem zusammenhängenden Ast. Die vier genannten Redesegmente entstammen allesamt dem zweiten Akt des Dramas und ihre Anordnung signalisiert eine gegenseitige Ähnlichkeit in Relation zum Vergleichskorpus. Es gibt weitere Beispiele, die die Aktgrenzen als wichtigen Faktor der Analyseergebnisse plausibilisieren. Die prägnantesten sind diejenigen von Tellheim und Minna in Akt 4 und 5.

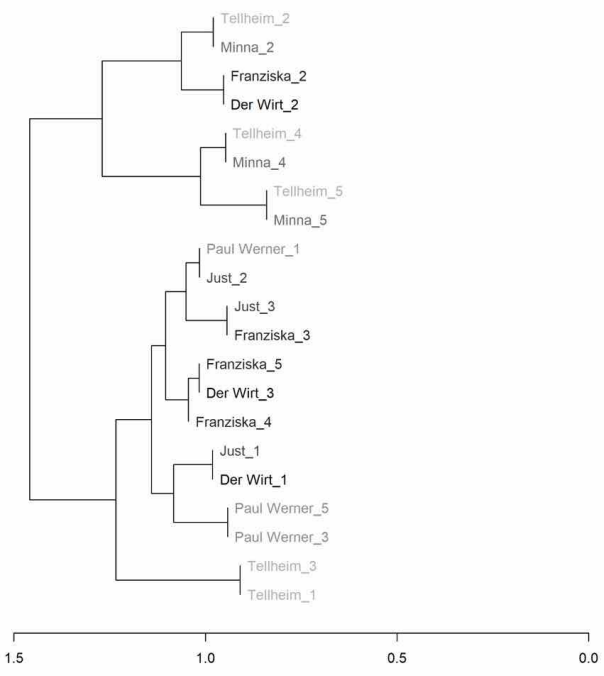


Abbildung 1: Minna von Barnhelm. Dendrogram, 1000 n-mfw, Cosine Delta, kein Culling, Ward-Clustering.

Ein einzelnes Dendrogram darf für die Bewertung der Hypothese jedoch nicht mehr sein als ein erstes Indiz, zumal die Segmentierung nicht gemäß einer festen Länge mit vorgegebener Wortanzahl, sondern nach den Aktgrenzen vorgenommen wurde.

Um ein potentielles ‚Cherry Picking'-Problem an dieser Stelle zu vermeiden, ergänzt *Abbildung 2* die Analyse um fünf weitere Dramen, drei von Friedrich Schiller und zwei weitere von Lessing. [5] Betrachtet man die hierarchische Struktur, wird deutlich, dass sowohl das Autorsignal als auch die Texteinheit klar erkennbar bleiben und die Anordnung dahingehend durch die relativ geringen Wortumfänge der Redesegmente nicht negativ beeinflusst wird.

Abbildung 2: Die Räuber, Die Verschwörung des Fiesco zu Genua, Maria Stuart, Minna von Barnhelm, Emilia Galotti, Miß Sara Sampson. Dendrogram, 1000 n-mfw, Cosine Delta, kein Culling, Ward-Clustering.

Ein Ähnlichkeitssignal, das Figurentypen, etwa den zärtlichen Vater in den bürgerlichen Trauerspielen Lessings, über das einzelne Drama hinweg verbinden würde, ist zumindest auf diese Weise nicht auszumachen. Die Vermutung liegt nahe, schreibt Lessing seine Dramen doch dezidiert für die am Theater üblichen Rollenfächer des 18. Jahrhunderts (Harris 1992). Sie scheint jedoch nicht auf diese Weise stilometrisch operationalisierbar zu sein.

Kopräsenz, Wortfeldsemantik und Sentiment

Stilometrische Analysen sind nicht das einzige Verfahren, um relative Ähnlichkeiten innerhalb eines Textkorpus zu bestimmen. Inwieweit sie geeignet sind, offene Fragestellungen – im Gegensatz etwa zur Autorschaftsattribution – zu erörtern, ist überhaupt noch zu prüfen. Sollten

Parameter wie Distanzmaß, Wortumfang oder Culling tatsächlich je nach Textkorpus neu zu bestimmen sein, wären ‚Cherry Picking'-Probleme der Methode inhärent (Schöch 2014; Jannidis 2014; Eder 2013).

Nachfolgend ist es deshalb geboten, die bisherigen Beobachtungen weiteren quantitativen Verfahren gegenüberzustellen. Dazu dienen Analysen der Kopräsenz, der Figurensemantik und der Empfindung, sogenannte Sentiment-Analysen.

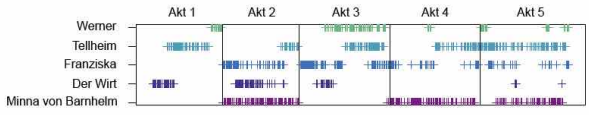


Abbildung 3: Redeäußerungen und deren Position, Minna von Barnhelm. [6]

Die Tabelle in *Abbildung 3* listet die fünf wichtigsten Figuren aus Lessings *Minna von Barnhelm* und markiert ihre Redeanteile in zeitlicher Abfolge. Im Zentrum der Untersuchung stehen die beiden Hauptfiguren des Dramas, namentlich Tellheim und Minna. Im zweiten, insbesondere aber im vierten und fünften Akt agieren Tellheim und Minna häufig gemeinsam auf der Bühne. Sie sind kopräsent. Diese Strukturdaten korrelieren mit den Beobachtungen aus *Abbildung 1*. Die Redesegmente der genannten Akte gruppieren sich dort in unmittelbarer Nähe zueinander, während Tellheims Redeanteile in den Akten 1 und 3 davon deutlich separiert abgetragen sind. In diesen beiden Akten stehen Tellheim und Minna nicht zur gleichen Zeit auf der Bühne. Die stilistische Ähnlichkeit der Figurenrede scheint also in Zusammenhang mit einem strukturellen Merkmal, der gemeinsamen Bühnenpräsenz der Figuren, zu stehen.

Um diesen Befund weiter zu spezifizieren, soll eine semantische Wortfeldanalyse die thematische Konzeption der Figurenrede operationalisieren (Willand, Reiter 2017).[7] *Abbildung 4* schlüsselt die Äußerungen der zentralen Figuren nach den Themen ‚Liebe', ‚Krieg', ‚Familie', ‚Ratio' und ‚Religion' auf. Die Häufigkeiten zeigen deutlich, wie ähnlich die Figurenrede von Minna und Tellheim hinsichtlich der Wortfelder konzipiert ist. Gerade die für Lessings Drama zentralen Themen ‚Liebe' und ‚Ratio' korrelieren merklich – auch verglichen mit den übrigen Figuren. Die Heatmap in *Abbildung 5* veranschaulicht jedoch, dass dieses thematische Ähnlichkeitsverhältnis der Figurenrede von Minna und Tellheim nur eingeschränkt auf einzelne Akte heruntergebrochen werden kann. Es scheint eher der Fall zu sein, dass die anhand der Wortfelder ablesbaren Rollen von Minna und Tellheim je nach Akt differieren und zugleich weitergegeben werden können. Ein Beispiel hierfür ist der wechselnde Zweifel an einer gemeinsamen Zukunft, der Dialoge von Minna und Tellheim bis zum fünften Akt prägt. Basis der Darstellung sind die die Häufigkeiten der Wortfelder in der Figurenrede auf Ebene einzelner Akte. Die Häufigkeiten werden über die Bestimmung der euklidischen Distanz in ein Ähnlichkeitsverhältnis gesetzt. Demnach sind die nach Wortfeldern ähnlichsten Segmente die Figurenrede von Tellheim aus dem vierten Akt und diejenige von Minna aus dem dritten (euklidische Distanz: 0,001467). Es folgen die Paare Tellheim 5 und Minna 2 (0,002705) sowie Tellheim 1 und Minna 4 (0.003758).

Während also *Abbildung 4* ein Indiz dafür liefert, dass Stil, Thema und Präsenz der Figuren und ihrer Äußerungen zusammenhängen, zeigt *Abbildung 5* eine gegensätzliche Tendenz – zumindest, wenn man nur die Redesegmente von Tellheim und Minna auf Ebene der einzelnen Akte vergleicht. Ein Grund dafür ist die Skalierung. Fügt man der Distanzberechnung weitere Figurensegmente bei, wird erkennbar, dass etwa auch die Segmente des fünften Akts von Minna und Tellheim thematisch relativ ähnlich erscheinen (0.006024).[8]

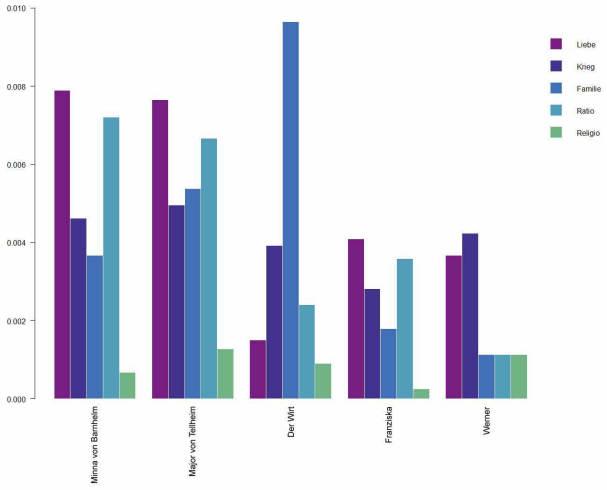


Abbildung 4: Semantische Wortfelder in Minna von Barnhelm, normalisiert nach Länge der Figurenrede.

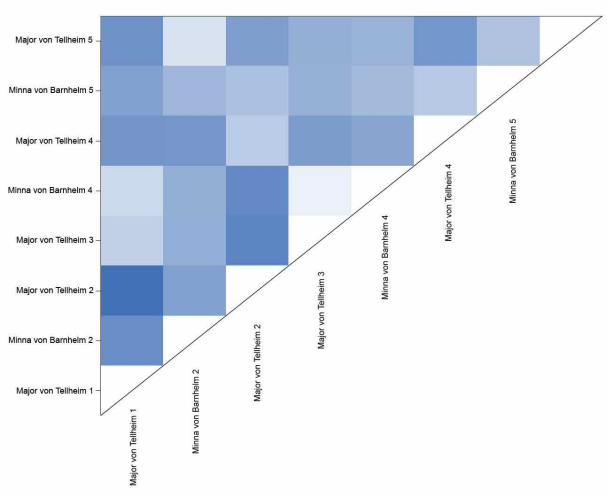


Abbildung 5: Euklidische Distanz der semantischen Wortfelder als Heatmap, Figurensegmente von Minna und Tellheim. Umso heller die Flächen dargestellt sind, umso größer ist die Übereinstimmung der Wortfelder.

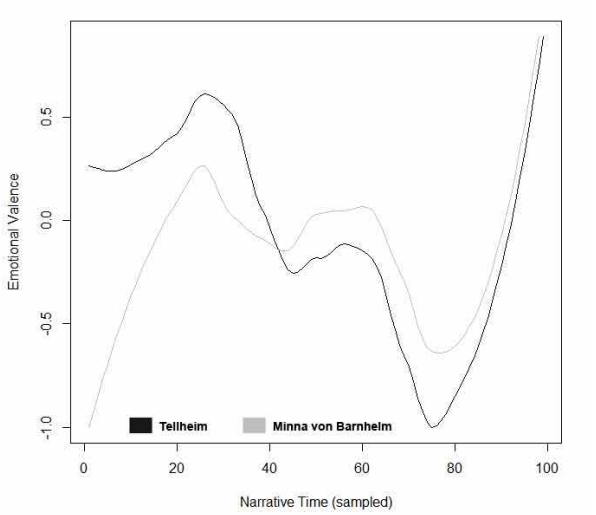


Abbildung 6: kumulative Sentiment-Analyse, Minna von Barnhelm Akt 5, Valenz der Redeanteile von Minna und Tellheim.[10]

Die Sentiment-Analyse in *Abbildung 6* gibt den letzten der vier mikroanalytisch genutzten Zugänge zur Dramenbetrachtung wieder. In der Forschung genutzt, um etwa die unterschiedliche Verwendung von Emotionswörtern in Märchen und Romanen zu analysieren (Mohammad 2011), archetypische Stimmungskurven in Romanhandlunge sichtbar zu machen (Jockers 2015) oder zwischenmenschliche Beziehungen in Shakespeares Stücken zu untersuchen (Nalisnick, Baird 2013), wird die Analyse hier eingesetzt, um die Empfindung von Tellheim und Minna anhand ihrer Valenzwerte im fünften Akt zu vergleichen. Die Grafik muss dabei als Annäherung aufgefasst werden, da zwei Dramenfiguren immer versetzt voneinander sprechen. Normalisierung nach Länge der Figurenrede, fast dauerhafte Kopräsenz im fünften Akt und ein gleichzeitiges Abtreten im 14. Auftritt legitimieren die Annäherung gleichwohl. Die eingangs zu vernehmende Diskrepanz der beiden Kurven kann dabei tatsächlich auf das Drama rückbezogen werden: Erst im Verlauf des fünften Akts durchschaut Tellheim das missglückte Trickspiel Minnas, das beide abwechselnd an einer glücklichen Zukunft zweifeln lässt. Im fünften Akt wird diese Spannung gelöst, Minna und Tellheim finden als liebende Partner zusammen. Die Übereinstimmung der beiden Sentiment-Kurven und die positiven Werte zum Ende des fünften Akts sind somit ein bestätigender Befund.[9]

Fazit und Ausblick:

Die nähere Betrachtung der Figurenrede in *Minna von Barnhelm* konnte aufzeigen, dass es sinnvoll ist, die Möglichkeiten verschiedener Analysemethoden zu kombinieren und so die jeweiligen Stärken in die Untersuchung einzubringen. Ergebnisse können zusätzlich validiert und zugleich für breitere Fragestellungen geöffnet werden. Kopräsenz scheint in den gewählten Dramentexten – das Fehlen eines Erzählers könnte hierbei eine zentrale Rolle spielen – einen stärkeren Einfluss auf die Figurenrede zu haben als im (englischsprachigen) Roman. Exemplarisch dafür stehen die Redeäußerungen von Minna und Tellheim im fünften Akt von Lessings *Minna von Barnhelm*. Für die Betrachtung der Figurenentwicklung im Verlauf der Handlung eines Dramas scheint es entscheidend zu sein, Erkenntnisse aus Strukturdaten sowie semantische und stilistische Analyseverfahren kritisch gegeneinander zu stellen.

Die vorliegende Arbeit wurde im Rahmen des Projekts „Quantitative Drama Analytics“ (QuaDramA) durchgeführt, das von der VolkswagenStiftung finanziert wird.

Fußnoten

1. Die Zahlen geben den Akt des Dramas wieder, dem die Figurenrede entnommen ist. Abbildung 1 und Abbildung 2 wurden mit Hilfe des ‚stylo'-Pakets für R angefertigt (Eder, Kestemont, Rybicki 2013).

2. Alle untersuchten Dramen entstammen dem Textgrid Repository.
3. Inwieweit der ohnehin schwer zu fassende Begriff ‚Stil' eine gute Wahl ist, um quantitative Methoden wie das Auszählen und vergleichen von Wortlisten zu beschreiben, ist zu diskutieren. Zumal die Vorstellung eines stilistischen Fingerabdrucks falsch zu sein scheint und ‚Autorstil' eher auf vielen kleinen Signalen fußt (Jannidis 2014).
4. Zum Vergleich: Fields, Bassist und Roper nutzen Blöcke von jeweils nur 200 Wörtern (2017).
5. Schillers Texte sollen einen Gegenpol zu den Dramen Lessings bieten. Während Lessing dezidiert für die üblichen Rollenfächer des 18. Jahrhunderts schreibt (Harris 1992), entwirft Schiller seine Figuren eher konträr zu den Rollenfächern (Detken 2014).
6. Die Abbildungen 3 und 4 wurden mit Hilfe des ‚DramaAnalysis'-Pakets für R erstellt (Reiter, Willand).
7. Zu diesem Zweck wurden fünf Wörterbücher mit 75 bis 105 Wörtern zu den Themen ‚Liebe', ‚Krieg', ‚Familie', ‚Ratio' und ‚Religion' erstellt, die dem jeweiligen Wortfeld zugehörig sind und in Dramen zwischen 1770 und 1830 verwendet wurden (Willand, Reiter 2017).
8. Distanzwerte von mehr als 0,01 sind die Norm.
9. Andere (hier nicht aufgeführte) Auswertungen zeigen allerdings, dass Ähnlichkeitssignale von Sentiment-Analysen nicht mit stilometrischen Analysen oder Wortfeldsemantiken korrelieren müssen. Sentiment-Analysen geben keine Themen wieder, sondern Werte der Empfindung, mit der Figuren über Themen sprechen. Trotz gleicher Themen können somit ganz unterschiedliche Sentiment-Werte entstehen.
10. Erstellt mit Hilfe des ‚syuzhet'-Pakets für R, das die Valenz auf Ebene des Satzes bestimmt (Jockers). https://github.com/mjockers/syuzhet. Die Auswertung erfolgt mit der deutschen Version des NRC Word-Emotion Association Lexicon. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Bibliographie

Burrows, John / Craig, Hugh (2012) „Authors and characters“, in: *English Studies* 93(3): 292–309.

Detken, Anke (2014): „Die Figur und ihr Fach: Konzeptionelle Überlegungen am Beispiel von Lessing und Schiller“, in: *Zeitschrift für Literatur- und Theatersoziologie* 11: 36–53.

Eder, Maciej (2013): „Computational Sylistics and Biblical Translation: How Reliable Can a Dendrogram Be?“, in: Piotrowski Tadeusz / Grabowski, Łukasz. *The Translator and the Computer*. Breslau: WSF Press: 155–170.

Eder, Maciej / Kestemont, Mike /Rybicki, Jan (2013): „Stylometry with R: a Suite of Tools“, in: *Digital Humanities 2013: Conference Abstracts*: 487–89.

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*: ii4–ii16 https://doi.org/10.1093/llc/fqx023 [letzter Zugriff 15. September 2017].

Fields, Paul J. / Bassist, Larry / Roper, Matt (2017): „Characters in 19th Century Novels Display Distinctive Voices as Seen by Stylometric Analysis“, in: *DH2017 Conference Abstracts* https://dh2017.adho.org/abstracts/494/494.pdf [letzter Zugriff 15. September 2017]. Harris, Edward P. (1992): „Lessing und das Rollenfachsystem. Überlegungen zur praktischen Charakterologie im 18. Jahrhundert“, in: Bender, Wolfgang F.: *Schauspielkunst im 18. Jahrhundert: Grundlagen, Praxis, Autoren.* Stuttgart: Steiner: 221–235.

Hoover, David (2017): „The Microanalysis of Style Variation“, in: *Digital Scholarship in the Humanities* https://doi.org/10.1093/llc/fqx022 [letzter Zugriff 15. September 2017].

Jannidis, Fotis (2014): „Der Autor ganz nah. Autorstil in Stilistik und Stilometrie“, in: *Theorien und Praktiken der Autorschaft*. Schaffrick, Matthias / Willand, Marcus. Berlin, Boston: De Gruyter: 169–195.

Jockers, Matthew L. (2015): *Revealing Sentiment and Plot Arcs with the Syuzhet Package* http://www.matthewjockers.net/2015/02/02/syuzhet/ [letzter Zugriff 15. September 2017].

Jockers, Matthew (2013): *Macroanalysis: Digital Methods and Literary History*. Urbana u.a.: Topics in the Digital Humanities.

Masten, Jeffrey (1997): *Textual Intercourse: Collaboration, Authorship and Sexualities in Renaissance Drama*. Cambridge: 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 (2000): „Conjectures of World Literature“, in: *New Left Review* 1: 54–68.

Moretti, Franco (2005): *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso. 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.

Schöch, Christof (2014): „Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“, in: *Literaturwissenschaft im digitalen Medienwandel. Beihefte zu Philologie im Netz* 7: 130–157. Willand, Marcus / Reiter, Nils (2017): „Geschlecht und Gattung: Digitale Analysen von Kleists ‚Familie Schroffenstein‘, in: *Kleist Jahrbuch* 2017: 177–195.

Realität programmieren? Zum Einfluss von Algorithmen auf die Wirklichkeit

Pfeiffer, Jasmin
jasmin_pfeiffer@gmx.de
FAU Erlangen, Deutschland

Im Mai erregte ein Artikel der Bloggerin Lisa Ringen Aufsehen, der auf die Benachteiligung weiblicher Personen durch den Suchalgorithmus des Netzwerks Xing hinwies: Gibt ein potentieller Auftraggeber bei seiner Suche nach Freiberuflern nicht explizit die weibliche Form der entsprechenden Berufsbezeichnung ein, so werden ihm ausschließlich Profile von Männern in der Ergebnisliste angezeigt. Möchte man eine Freiberuflerin finden, so muss man etwa nach einer „Fotografin“ oder einer „Entwicklerin“ suchen – eine Idee, auf die vermutlich die wenigsten User kommen werden. Wie Ringen richtig schreibt, werden Frauen hierdurch nicht nur benachteiligt, sondern es entsteht auch „der unterschwellige Eindruck, Männer seien die erfolgreicheren, kompetenteren Fotografie-, Beratungs- und Grafik-Spezialisten“ (Ringen 2017). Bezeichnend ist das Statement des Xing-Sprechers Kopka, der, wie die SZ berichtete, betonte, dass ihm das Problem nicht bewusst gewesen sei (Holzki 2017).

Die Affäre um Xing legt ein Problem offen, das bisher noch unzureichend erforscht und thematisiert worden ist: Computerprogramme stellen keine neutralen mathematischen Gebilde dar, sondern sind, wie Lemire betont, zu wirkungsmächtigen Agenten im sozialen Raum geworden, die die uns umgebende Wirklichkeit beeinflussen und verändern können: „In any case, we have to accept software as an active agent that helps shape our views and our consumption rather than a mere passive tool.“ (Lemire 2016) Die Annäherung der Informationswissenschaften an die Disziplinen der Mathematik und der Naturwissenschaften, die u. a. im Begriff der MINT-Fächer evident wird, hat den Blick auf diese Formen der Einflussnahme verstellt – Tech-Größen wie Mark Zuckerberg und Alexander Nix preisen die angebliche Vorurteilslosigkeit und Neutralität der Algorithmen, und Unternehmen verkaufen die von ihnen entwickelten Computerprogramme als Meilensteine auf dem Weg zu einer neuen Objektivität der durch sie übernommenen Prozesse (vgl. Maschewski / Nosthoff 2017). Auch die Klassifikation von Programmiersprachen als formale Sprachen, die im Vergleich zu natürlichen Sprachen als präziser und eindeutiger betrachtet werden, verstärkt diesen Eindruck.

In Abgrenzung hierzu möchte ich in meinem Vortrag Programmiercode als deklarativen Sprechakt beschreiben und auf diese Weise den Blick auf die von Lemire diagnostizierte realitätskonstituierende Dimension von Computerprogrammen lenken. Ich werde mich dabei auf objektorientierte Sprachen konzentrieren, da der Bezug zwischen Code und Wirklichkeit bei diesen besonders evident ist.

In einem ersten Teil werde ich aufzeigen, inwiefern in objektorientierten Sprachen verfasster Programmcode als deklarativer Sprechakt beschrieben werden kann. Searle definiert Deklarationen als Sprechakte, die eine Übereinstimmung von Wirklichkeit und Worten herstellen: „It is the defining characteristic of this class that the successful performance of one of its members brings about the correspondence between the propositional content and reality; successful performance guarantees that the propositional content corresponds to the world.“ (Searle 1975: 358) Deklarative Sprechakte beschreiben folglich etwas, was durch diese Beschreibung zur Wirklichkeit wird.

Dies trifft auch auf in objektorientierten Sprachen verfassten Programmiercode zu: Er beschreibt Klassen, Instanzen und deren Eigenschaften sowie Methoden, die dadurch zugleich innerhalb des Codes existieren und, wenn das Programm kompiliert wird, die Ausführung der von ihnen beschriebenen Operationen in der Hardware des Computers veranlassen.

Jedoch bleibt der Bezug zur Wirklichkeit im objektorientierten Paradigma nicht auf die Erzeugung von elektronischen Spannungen in den Bauteilen beschränkt. Vielmehr wird objektorientierter Code im Software Engineering als Repräsentation der Wirklichkeit konzipiert. Joachim Goll und Cornelia Heinisch beispielsweise beschreiben Klassen in ihrer Java-Einführung als Entsprechungen von Gegenständen der realen Welt: „Eine […] Klasse entspricht einem Typ ei-

nes Gegenstands der realen Welt.“ (Goll / Heinisch 2014: 36 f.) Die Idee der objektorientierten Programmierung besteht laut Goll und Heinisch darin, realweltliche Entitäten in Repräsentationen im Programmcode zu überführen: „Der Ansatz der Objektorientierung basiert darauf, Objekte der realen Welt mit Hilfe softwaretechnischer Mittel als Entity-Objekte abzubilden.“ (Goll / Heinisch 2014: 37) Bei dieser Überführung werden nur die für das Computerprogramm relevanten Aspekte der realen Entitäten im Code dargestellt: „ Bei der Abbildung einer Entität der Realität auf ein Objekt muss aber eine Abstraktion stattfinden, bei der das Unwesentliche weggelassen wird. Die typischen Eigenschaften bleiben übrig. Die unwesentlichen Eigenschaften werden ignoriert (Abstraktion).“ (Goll / Heinisch 2014: 37)

Ziel des objektorientierten Programmierens ist also, Objekte und Zusammenhänge der Wirklichkeit in Klassen, Instanzen und Methoden im Code darzustellen. Diese Überführung stellt einen selektiven und damit notwendigerweise reduktionistischen Akt der Zurichtung von Wirklichkeit dar. Obgleich in den verschiedenen Ratgebern zum Software-Engineering versucht wird, objektivierbare Kriterien für die Abbildung von Realität in Code zu entwickeln, wird diese letztlich immer vom Erfahrungshorizont und den *world versions* des zuständigen Software-Entwicklers beeinflusst. Die Ergebnisse der Operationen, die der Algorithmus auf den Repräsentationen der realen Entitäten im Code vollzieht, werden von den Benutzern des Programms potenziell in die Realität rückgeführt und können als Ausgangspunkt realer Handlungen fungieren. Dies zeigt sich deutlich am Beispiel des in Washington D.C. implementierten Systems IMPACT zur Evaluierung der Leistungen der Lehrer von Washingtons Schulen: Die Ergebnisse der vom Algorithmus durchgeführten Berechnungen dienten hier als Grundlage für die Entscheidung, welche Lehrer entlassen und welche befördert wurden. Die deklarativen Sprechakte des Codes verändern folglich, so die These, die Wirklichkeit nicht nur auf der Ebene der Einsen und Nullen, sondern können auch Einfluss auf Diskurse nehmen und somit, wie Lemire konstatiert, zu aktiven, die Wirklichkeit verändernden Agenten werden. Ich möchte den Akt des Programmierens daher ähnlich wie natürliche Sprechakte als eine Art Einschreibung in den Diskurs und als wirklichkeitsschaffend beschreiben.

Möchte man diese Formen der Einflussnahme auf die Realität verstehen und adäquat beschreiben, so ist eine genaue Analyse des den Computerprogrammen zugrunde liegenden Codes unabdingbar. Hier setzt die in den frühen 2000er-Jahren unter anderem von Matthew Fuller und Lev Manovich propagierte Strömung der Software Studies an, in der sich dieser Vortrag situieren möchte. Daher werde ich im zweiten Teil meines Vortrags ein *Close Reading* eines objektorientierten Algorithmus vorstellen. Hierzu werde ich ein Beispiel aus dem Bereich des Machine Learning wählen. Einige in der jüngeren Vergangenheit erschienene Paper machen darauf aufmerksam, dass bestehende kulturelle und soziale Meinungen Systeme künstlicher Intelligenz in starkem Maß beeinflussen können. Für den Bereich der *Structured prediction* beispielsweise haben Zhao, Wang u. a. gezeigt, dass die in den zum Training verwendeten Datenmengen enthaltenen Geschlechterstereotypen nicht nur vom Model übernommen, sondern sogar verstärkt wurden.

Was die Maschine lernt und welche Konzeption der Realität sie hat, ist folglich offensichtlich in starkem Maße beeinflusst von den zum Training verwendeten Datensets, aber auch vom Wirklichkeitsverständnis des Programmierers, der die ML-Algorithmen erschafft. Dies möchte ich in meinem Vortrag am Beispiel der Entscheidungsbäume näher erläutern, da diese einerseits leicht zu verstehen sind und andererseits die oben umrissenen Probleme besonders deutlich demonstrieren. Entscheidungsbäume liefern zu Objekten, die durch Mengen von Attribut-Wert-Paaren beschrieben sind, jeweils eine Entscheidung darüber, welcher Klasse das betreffende Objekt zuzuordnen ist. Welche Eigenschaften eine Klasse hat, wird dabei vom Programmierer festgelegt und somit unmittelbar von dessen Wirklichkeitsverständnis beeinflusst.

In Hinblick auf das Tagungs-Thema, „Kritik der digitalen Vernunft“, verfolgt der Vortrag zweierlei Ziele: Einerseits soll ein möglicher methodischer Ansatzpunkt zur Analyse von Programmiersprachen und Algorithmen entwickelt werden und aufgezeigt werden, dass die genaue Analyse von Code und seiner Funktionsweise unabdingbar für das Verständnis digitaler Phänomene ist. Andererseits soll zu einer gewissen Vorsicht gegenüber digitalen Methoden in den Geisteswissenschaften aufgerufen werden: Die Benutzung digitaler Mittel ist eine große Bereicherung für die Geisteswissenschaft, erfordert aber auch eine eingehende Reflexion der Beziehung zwischen Realität und Programmcode und des Einflusses der Algorithmen auf den betrachteten Gegenstand.

Bibliographie

Alpaydin, Ethan (2016): *Machine Learning*, Cambridge, Massachusetts: MIT Press.

Berry, David M. (2011): „The Computational Turn: Thinking about the Digital Humanities“, in: *Culture Machine* 12: 1–22.

Berry, David M. (2011): *The Philosophy of Software. Code and Mediation in the Digital Age* , New York: Palgrave Macmillan.

Fuller, Matthew (2008): *Software Studies. A Lexicon* , Cambridge, Massachusetts: MIT Press.

Holzki, Larissa (2017): „Frauen sind im Netz schwieriger zu finden“, http://www.sueddeutsche.de/karriere/frauen-und-karriere-frauen-sind-im-netz-schwieriger-zu-finden-1.3492732 , Stand: 24.09.17.

Lemire, Daniel (2016): „Is software a neutral agent?“, https://lemire.me/blog/2016/04/29/is-software-a-neutral-agent/ , Stand: 24.09.17.

Manovich, Lev (2001): *The Language of New Media* , Cambridge, Massachusetts: MIT Press.

Maschewski, Felix / Nosthoff, Anna-Verena (2017): „Das Netz ist nie neutral“, https://www.nzz.ch/feuilleton/kuenstliche-intelligenz-digitale-technik-ist-nie-neutral-ld.1302959 , Stand: 24.09.17.

O’Neil, Cathy (2016): *Weapons of Math Destruction* , New York: Broadway Books.

Ringen, Lisa (2017): „Dummer Algorithmus: Als Frau bei Xing gefunden werden“, https://www.marketing-madam.de/2017/05/02/dummer-algorithmus-als-frau-bei-xing-gefunden-werden/92706211/ , Stand: 24.09.17.

Searle, John R. (1975): *Expression and Meaning. Studies in the Theory of Speech Acts* . Cambridge: Cambridge University Press.

Zhao, Jieyu / Wang, Tainlu / Yatskar, Mark / Ordonez, Vicente / Chang, Kai-Wei (2017): „Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints“, https://arxiv.org/abs/1707.09457 , Stand: 01.01.18.

SANTA: Systematische Analyse Narrativer Texte durch Annotation

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

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

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

Willand, Marcus
marcus.willand@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

In diesem Beitrag wollen wir ein Vorhaben zur Diskussion stellen, das an zwei zentralen Herausforderungen in den Digital Humanities ansetzt: Der Erstellung adäquater Annotationsrichtlinien für geisteswissenschaftlich relevante textuelle Konzepte und der Schnittstelle in der Kooperation zwischen beteiligten Wissenschaftlerinnen und Wissenschaftlern aus Geisteswissenschaft und Informatik. Für DH-Projekte sind Kooperationen unerlässlich, wenn fortgeschrittene Techniken zur Textanalyse eingesetzt werden und/oder es um eine Zusammenführung von Konzepten oder Zugangsweisen geht, die bereits intradisziplinär als komplex gelten. Dabei wird ein signifikanter Anteil der Projektlaufzeit auf die Entwicklung einer “gemeinsamen Sprache” und die Identifikation der exakten, gemeinsamen wissenschaftlichen Fragestellung verwendet. Dies ist zweifellos ein produktiver Prozess, dessen erfolgreiche Durchführung allerdings voraussetzt, dass auf beiden Seiten Forscherinnen und Forscher beteiligt sind, die sich auf das interdisziplinäre Vorgehen voll einlassen und auch den nötigen Zeitaufwand tragen.

Methodisch-technisch ist ein substanzielles Nadelöhr bei der Entwicklung automatischer Werkzeuge das Fehlen von annotierten Goldstandards, an/auf denen Werkzeuge trainiert, verglichen und feinjustiert werden können. Das Fehlen der Goldstandards ist jedoch eigentlich ein nachgelagertes Problem, wie sich z.B. in narratologisch orientierten Projekten zeigt (heureCLÉA: Bögel et al., 2015; Propp annotation: Fisseni et al., 2014): Die Umsetzung narratologischer Theorien als Annotationen ist alles andere als trivial, da narratologische Konzepte nicht im Hinblick auf Annotation entwickelt wurden. Leerstellen in den Definitionen müssen gefüllt, Voraussetzungen geklärt und Unterkategorien geklärt werden. Die Annotation solcher Kategorien ist also kein reiner Umsetzungs- oder Implementierungsprozess, sondern einer bei dem sich tiefe, konzeptionelle Fragen stellen. Als Ergebnis solcher Prozesse stehen dann Annotationsrichtlinien, die die Brücke zwischen Theorie und Praxis schlagen. Erst wenn Anno-

tationsrichtlinien für ein Phänomen (oder eine Gruppe von Phänomenen) etabliert sind, können größere Annotationsprojekte mit Aussicht auf Erfolg durchgeführt werden.

Das von uns vorgeschlagene Vorgehen erlaubt den Beteiligten Forscherinnen und Forschern ihre Expertise einzubringen, ohne in einem gemeinsamen Projektkontext zu arbeiten. Die Schnittstelle zwischen D und H wird hierbei von annotierten Daten und Annotationsrichtlinien gebildet, wobei die Richtlinien ohne Kompromisse bezüglich möglicher Automatisierungen erstellt werden. Das Vorhaben gibt somit auch narratologisch/literaturwissenschaftlich anspruchsvoller Konzeptentwicklung und damit Theoriebildung einen Rahmen. Verfügbare annotierte Daten wiederum erlauben Informatikerinnen und Informatikern ohne Expertise in narratologischen Fragen die Entwicklung von Werkzeugen für komplexe technische Probleme.

Ein *shared task* zur Erstellung von Annotationsrichtlinien

Shared Tasks sind in der Computerlinguistik weit verbreitet und haben für viele Bereiche gezeigt, dass sie ein geeignetes Instrument sind, Forschungsbemühungen verschiedener Gruppen zum gleichen Thema zu bündeln und zu verstärken. In einem *shared task* versuchen verschiedene Arbeitsgruppen mit verschiedenen Methoden dieselbe, klar definierte Aufgabe zu lösen, z.B. Word Sense Disambiguation (z.B. Mihalcea et al., 2004), Sentiment Analysis (z.B. Nakov et al., 2013) oder Named Entity Recognition (z.B. Sang and/De Meulder 2003). Auch wenn bisweilen im Rahmen von NLP-shared tasks Annotationsstandards neu entwickelt werden, liegt der Fokus hier auf der Verbesserung der Vorhersagequalität automatischer Systeme. Damit in einem solchen Vorgehen literaturwissenschaftlich relevante und interessante Konzepte und Phänomene bearbeitet werden, muss literaturwissenschaftliche Expertise bei der Erstellung der Annotationsrichtlinien einfließen.

Als Rahmen für die Entwicklung von Annotationsrichtlinien organisieren wir einen shared task der sich genau auf dieses Ziel konzentriert (Phase 1: Erstellung von Guidelines). Sind die Richtlinien etabliert, kann anschließend ein großes Korpus annotiert werden, das wiederum in einem NLP-shared task eingesetzt werden kann, um Verfahren zu erproben, die die annotierten Phänomene automatisch finden (Phase 2: Automatisierung).

Als Phänomen haben wir uns dabei auf Erzählebenen in englischen und deutschen Texten festgelegt, da diese für zahlreiche, komplexere literaturwissenschaftliche Fragestellungen hilfreich sind, ohne selbst (für einen ersten *shared task*) zu komplex zu sein. Zudem sind sie als Phänomen omnipräsent: Praktisch jeder narrative Text enthält mehr als eine Erzählebene, und sie sind auch in nicht-textuellen Medien wie z.B. Filmen verbreitet. Die Existenz verschiedener Theorien zur Analyse von Erzählebenen in literarischen Texten belegt, dass es dabei auch konzeptionellen, theoretischen Entwicklungsbedarf gibt. Erzählebenen bilden darüber hinaus eine wichtige Segmentierungsstufe für die weitere automatische semantische Verarbeitung von Texten: z.B. sollte Koreferenzresolution von der vorher erfolgten Erkennung von Erzählebenen profitieren,da Koreferenzketten in heterodiegetischen eingebetteten Erzählungen nicht ebenenübergreifend sein sollten.

Während Details zum Gesamtaufbau des Shared Tasks bereits in einem anderen Artikel beschrieben wurden (Reiter et al., 2017), fokussieren wir uns in diesem Beitrag auf die genauere Beschreibung der ersten Phase des *shared tasks*.

Geplanter Ablauf

Erstellung von Annotationsrichtlinien

(bis Mitte Juni 2018)

Im ersten Schritt wird allen Teilnehmerinnen und Teilnehmern ein *development corpus* bestehend aus ca. 20 Texten zugänglich gemacht. Die Texte liegen auf deutsch und englisch vor und decken verschiedene Genres und Epochen ab. Die Texte enthalten verschiedene Arten von Erzählebenen, gemäß eines etwas vagen Vorverständnisses.

Die Texte können und sollen von den Teilnehmerinnen und Teilnehmern benutzt werden, um Richtlinien für die Annotation von Erzählebenen zu entwickeln und zu testen. Ob die Texte in einer oder in beiden Sprachen verwendet werden, ist dabei den Teilnehmerinnen und Teilnehmern überlassen. Sie sollten dabei das Ziel verfolgen, eine möglichst breite Anwendbarkeit der Richtlinien sicherzustellen (auch jenseits des *development corpus*). Außerdem sollen die Richtlinien vollständig und selbsterklärend sein, so dass kein Expertenwissen zur Anwendung vorausgesetzt wird. Um mehrsprachige Anwendung zu ermöglichen, sollen die Richtlinien auf Englisch formuliert sein, sie dürfen aber sprachspezifische Beispiele enthalten.

Wie genau die Gruppen dabei vorgehen, bleibt ihnen überlassen. In vergangenen Annotations-

projekten (mit und ohne Bezug zu Literaturwissenschaft bzw. literarischen Texten) hat sich aber ein iterativer Prozess als fruchtbar erwiesen. Sobald eine erste Version der Richtlinien erstellt wurde, werden sie auf neuen Texten getestet, um ihre Definitionslücken oder Vagheiten zu identifizieren. Aus dem Schließen der Lücken ergibt sich dann eine weitere Version der Richtlinien, die wiederum auf Texten getestet werden können.

Anwendung eigener Guidelines

(bis Ende Juni 2018)

Im zweiten Schritt sollen die Arbeitsgruppen ihre eigenen Richtlinien auf neuen Texten testen. Nach dem Einreichen ihrer Richtlinien erhalten die Teilnehmerinnen und Teilnehmer hierzu sechs neue literarische Texte, die vom Organisationsteams des *shared tasks* ausgesucht wurden. Die Annotation dieser Texte muss dabei in einem Web-basierten, frei zugänglichen, von den Organisatoren bereitgestellten Annotationstool durchgeführt werden, um die automatisierte Auswertung der Annotationen und ihren Vergleich zu ermöglichen.

Anwendung von Guidelines anderer Teilnehmer

(bis Mitte Juli 2018)

Im dritten Schritt erhält jede teilnehmende Gruppe Richtlinien anderer Gruppen, auf deren Basis Erzählebenen in den sechs Texten erneut annotiert werden, wobei alle Richtlinien von uns zuvor anonymisiert werden. Zusätzlich wird auch eine vom Organisationsteam betreute Gruppe von studentischen Hilfskräften alle eingereichten Annotationsrichtlinien auf den sechs Texten anwenden.

Evaluation aller vorgeschlagener Guidelines

(August/September 2018)

Im letzten Schritt der ersten Phase des *shared tasks* werden alle eingereichten Annotationsrichtlinien verglichen und evaluiert. Dafür treffen sich die Teilnehmerinnen und Teilnehmer zu einem Workshop, auf dem sie ihre eigenen Richtlinien vorstellen und gemeinsam Qualität und Komplexität bewertet werden. Das Ziel des Workshops ist außerdem, basierend auf der Diskussion und den Informationen bezüglich der Inter-Annotator-Agreements im Plenum und möglichst konsensual die Annotationsrichtlinien zu bestimmen, die dann in der zweiten Phase des *shared tasks* verwendet werden. Auf deren Basis werden dann Methoden und Systeme entwickelt, die automatisch Erzählebenen in Texten identifizieren können.

Zur vergleichenden Evaluation von Annotationsrichtlinien sind bisher Ansätze aus der Computer- und Korpuslinguistik zur quantitativen Messung des Inter-Annotator-Agreement (IAA) bekannt (vgl. Artstein, 2017), die im Bereich der Digital Humanities angewendet wurden und werden. Da es aber bei der Erstellung von Annotationsrichtlinien für narratologische Phänomene eben nicht *nur* um die Umsetzung und Erklärung einer klar spezifizierten Theorie geht, sondern eben *auch* um die (Weiter-)Entwicklung narratologischer Konzepte, bedarf es eines weitergehenden Blickes. Dabei sollen drei Aspekte Berücksichtigung finden: Die **Anwendbarkeit** von Annotationsrichtlinien kann durch quantitatives IAA gemessen werden. Hier stellen sich durch möglicherweise unterschiedliche theoretische Zugänge vor allem Fragen der Vergleichbarkeit. Der Aspekt der begrifflichen **Abdeckung** bezieht sich darauf, welche (bekannten) narratologischen Ebenenkonzeptionen in der konkreten Ausgestaltung vollständig oder teilweise enthalten sind. Dies wird sich nur durch qualitative Analyse und wissenschaftliche Diskussion basierend auf theoretischen Vorstudien klären lassen, für die der Workshop einen Rahmen bieten soll. Die **Nützlichkeit** von Annotationsrichtlinien kann bei narrativen Ebenen dahingehend bewertet werden, ob sie interpretativ wertvolle Beschreibungen erlauben. Leitgedanke ist hier, dass narratologische Annotationen eine deskriptive Basis für literaturwissenschaftliche Interpretationen liefern sollen. Unterschiedlichen Annotationsrichtlinien zu folgen hieße also zu unterschiedlichen Text-Deskriptionen zu kommen, die wiederum unterschiedliche Interpretationen zulassen.

Conclusions

Im Rahmen des Vortrags wollen wir insbesondere zwei der o.g. Aspekte in den Fokus rücken und diskutieren: Die iterative Entwicklung von Annotationsrichtlinien als verteiltes, kollaboratives Projekt sowie die Evaluation und Vergleichbarkeit von Annotationsrichtlinien für literarische Phänomene.

Bibliographie

Artstein, Ron (2017): “Inter-annotator Agreement”, in: Ide, Nancy / Pustejovsky James (eds.):

Handbook of Linguistic Annotation. Dordrecht: Springer. DOI 10.1007/978-94-024-0881-2.

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (2015): "Collaborative text annotation meets machine learning: heurecléa, a digital heuristic of narrative", in: DHCommons 1.

Fisseni, Bernhard / Kurji, Aadil / Löwe, Benedikt (2014): "Annotating with Propp's morphology of the folktale: Reproducibility and trainability", in: *Literary and Linguistic Computing* 29(4):488–510, 1093/llc/fqu050

Mihalcea, Rada / Chklovski, Timothy / Kilgarriff, Adam (2004): "The Senseval-3 English Lexical Sample Task". In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

Nakov, Preslav / Rosenthal, Sara / Kozareva, Zornitsa / Stoyanov, Veselin / Ritter, Alan / Wilson, Theresa (2013): "SemEval-2013 Task 2: Sentiment Analysis in Twitter". In *Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13*, Atlanta, Georgia, USA.

Reiter, Nils / Gius, Evelyn / Strötgen, Jannik / Willand, Marcus (2017): "A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts ". In *Digital Humanities 2017: Conference Abstracts*, Montreal, Canada.

Sang, Erik F. Tjong Kim / de Meulder, Fien (2003): "Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition", in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*.

Sentimentanalyse in unstrukturierten Texten (am Bsp. literaturgeschichtlicher Rezeptionsanalyse)

Mellmann, Katja
katja.mellmann@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Du, Keli
keli.du@stud-mail.uni-wuerzburg.de
Universität Göttingen, Deutschland

Ausgangspunkt

Die fortschreitende Retrodigitalisierung von Kulturzeitschriften und anderen Publikationsformen mit literaturkritischen Inhalten eröffnet der literaturgeschichtlichen Rezeptionsanalyse die Möglichkeit, mit historisch repräsentativen Korpora zu arbeiten. Dabei stellt sich jedoch das Problem, dass insbesondere Zeitschriftendigitalisate in der Regel nicht als edierte Texte von standardisierter Qualität vorliegen, sondern mit 'schmutzigen Texten', also Texten mit fehlerhafter OCR und ohne linguistische Strukturierung gearbeitet werden muss. Wir wollen im Rahmen des Themas "Kritik der digitalen Vernunft" einen konstruktiven Umgang mit diesem Problem vorstellen. Wir unterscheiden dazu grundsätzlich zwei Zielperspektiven:

1. Korpusanalyse als Untersuchung mit validen Ergebnissen und
2. Korpusanalyse als Heuristik zur vorläufigen Trenddarstellung.

Die erste Perspektive ist auf qualitativ hochwertige Textkorpora angewiesen, um statistisch aussagekräftige Ergebnisse zu erzielen. Sie ist die Standardperspektive, wenn Korpusanalysen als Forschungsinstrument eingesetzt werden. Wir nehmen hingegen die zweite Perspektive ein: In ihr werden qualitativ minderwertige Textkorpora nicht als bedauerliche Abweichung vom eigentlich Gewünschten aufgefasst, sondern als Forschungsgegenstand eigener Art, der auch eine Methodik eigener Art erfordert. Diese Methodik hebt auf vorläufige Trenddarstellungen ab, die nicht als (noch unvollkommene) Vorstufe einer validen Korpusanalyse, sondern als eigenständige Heuristik zur Identifikation potentieller Ereignisse in einem diachronen Korpus aufgefasst werden. Die Methode soll sozusagen grobe Bewegungsprofile liefern, von denen aus anschließend wieder gezielte hermeneutische Tiefensondierungen unternommen werden können. Digitale und hermeneutische 'Vernunft' stehen hier also in einem komplementären Verhältnis; nicht versucht wird, die eine durch die andere möglichst perfekt nachzubilden.

Beispielprojekt: Sentimentanalyse in historischer Literaturkritik

Bei den angezielten Bewegungsprofilen handelt es sich im Rahmen unseres Forschungspro-

jekts [1] um Zäsuren in der Bewertung literarischer Autoren. Wir untersuchen in einem Pilotprojekt die Rezeption von Literatur in literaturkritischen Zeitschriften des ausgehenden 19. Jahrhunderts mittels einer Sentimentanalyse der Textumgebung von Autorerwähnungen. An einem Testkorpus optimieren wir durch den Vergleich automatisierter mit manuellen Analysen eine an das historische Genre 'Literaturkritik um 1900' angepasste Sentimentwortliste.

1. Literaturwissenschaftliche Fragestellung

Historische Rezeptionsanalyse (Mellmann/Willand 2013) rekonstruiert die Aufnahme literarischer Werke durch das originale zeitgenössische Publikum. Dazu zählen insbesondere (a) Inhaltsverständnis, (b) ästhetische Wertung und (c) Kontextualisierung mit außerliterarischen zeitgenössischen Wissensformationen. Wir befassen uns in unserer Studie ausschließlich mit der ästhetischen Wertung (b).

Diese ist symptomatisch für einen umfassenden literaturgeschichtlichen Wandel in der zweiten Hälfte des 19. Jahrhunderts: Auf dem Übergang vom Bürgerlichen Realismus zur Klassischen Moderne verlieren ehemals reputierte Autoren noch während ihrer aktien Schaffenszeit ihren führenden Status; neue Stile aus dem Bereich des gesamteuropäischen Naturalismus und Ästhetizismus gewinnen an Reputation. Für einzelne Autoren und insbesondere für poetologische Programmatiken wurde dies bereits vielfach gezeigt. Was noch aussteht, ist eine Einschätzung, wie repräsentativ die in Einzelstudien ermittelten Entwicklungen für die Gesamtentwicklung sind, insbesondere unter Einschluss auch der wenig erforschten nichtkanonischen Literatur. Abhilfe schaffen könnte hier die Analyse eines großen Korpus repräsentativer Literaturzeitschriften, die diachrone Wertungsprofile zu symptomatischen Autorengruppen liefert.

2. Korpusbildung

Langfristiges Ziel ist eine Analyse von sieben repräsentativen Zeitschriften über einen Erscheinungszeitraum von ca. 1860 bis 1900: "Die Grenzboten", "Die Gegenwart", "Deutsche Rundschau", "Nord und Süd", "Blätter für literarische Unterhaltung", "Westermanns Monatshefte" und "Magazin für Literatur". Als Volltext verfügbar ist derzeit nur die Zeitschrift "Die Grenzboten". Sie dient uns als erster Anwendungsfall nach den überwachten Optimierungsläufen an einem auf der Basis einer Anthologie (Kreuzer 2006, Bd. I und II) erstellten Testkorpus.

Literaturkritisches Schrifttum im ausgehenden 19. Jahrhundert ist ein außergewöhnlich stark rhetorisiertes Textgenre, das durch einen hohen Grad an Ironie, Intertextualität, Euphemismus und Understatement besondere Herausforderungen an die Methode der digitalen Sentimentanalyse stellt, zumal in unstrukturierten Texten, die keine Berücksichtigung von grammatischen Komplexitäten (wie z.B. doppelter Verneinung, Konjunktiven, indirekter Rede) zulassen. Die Optimierung der Sentimentwortliste ist deshalb weniger auf eine Verfeinerung als auf eine Vergröberung hin ausgelegt. Die Korpusanalyse soll vor allem eklatante Veränderungen identifizieren.

3. Erste Testläufe digitaler Analysen

Der Volltext des Testkorpus wurde durch NLTK Punkt Sentence Tokenizer [2] tokenisiert. Mittels des Namensregisters der Anthologie wurden anschließend alle Sätze mit Erwähnung eines Autornamens extrahiert. Jeder Satz wurde in einem manuellen Rating als positiv, neutral oder negativ annotiert. Das Testkorpus umfasst 1731 1-Satz-Textsnippets. Nach der manuellen Annotation sind 505 davon positiv, 909 neutral und 317 negativ. Ausgangsbasis unserer Sentimentwortliste war die deutschsprachige Ressource "SentimentWortschatz" (Remus et al. 2010), die manuell um den im Rating auffällig gewordenen Wortschatz erweitert und um offenkundig unbrauchbare oder überflüssige Wörter gekürzt wurde. Danach erfolgte ein erster Testlauf:

Der Grundwert jedes Sentimentworts wurde mit 1 angesetzt. Die Polarität eines Satzes wurde in zwei Schritten festgelegt: Zuerst wurde das Vorkommen der positiven und negativen Sentimentwörter im Satz gezählt. Anschließend wurde die „Aggregation function" (Abb. 1) für die Berechnung des Sentiment-Werts des Satzes verwendet: „ *se* j is a sentiment expression in sentence *s*, *dist(se* j *, a* i *)* is the word distance between aspect *a* i and sentiment expression *se* j in *s*, and *se* j *.ss* is the sentiment score of *se* j" (Liu 2015).

$$\text{score}(a_i, s) = \sum_{se_j \in s} \frac{se_j.ss}{\text{dist}(se_j, a_i)},$$

Abb. 1: Aggregation function

Ist der Sentiment-Wert größer als 0, gilt der Satz als positiv; kleiner als 0 gilt als negativ; ist der Wert gleich 0 (z.B., weil kein Sentimentwort im Satz auftaucht), gilt der Satz als neutral. Die im Satz ermittelten Sentimentwörter wurden in die Ergebnisdarstellung übernommen, um die digitale Analyse anschließend manuell überprüfen und die Sentimentwortliste optimieren zu können. Wörter, die sich als überwiegend dysfunktional erweisen, werden von der Sentimentwortliste gelöscht, fehlende Sentimentwörter werden ergänzt.

Im ersten Testlauf wurden nur ca. 47.4% der Sätze richtig erkannt (Tab. 1). Unsere Wortliste konnte vor allem die negativen Sätze schwer identifizieren. Ca. 77% der positiven Sätze wurden richtig identifiziert. Aber auch der Anteil der fälschlich als positiv klassifizierten Sätze war sehr hoch. Außerdem war der Sentiment-Wert von vielen als neutral eingestuften Sätzen nicht gleich 0.

Dimension	richtig identifizierte Sätze	Precision	Recall	F1 score	Manuelle Annotation
Positiv	387	0.4	0.77	0.52	505
Neutral	322	0.7	0.35	0.47	909
Negativ	112	0.33	0.35	0.34	317

Tab. 1: Ergebnis des ersten Testlaufs

In einem zweiten Testlauf haben wir eine automatische Klassifikation ausprobiert. Dabei wurden die Anzahl der Sentimentwörter und der Sentiment-Wert eines Satzes als Feature verwendet. Im Verhältnis 80% zu 20% wurden die Daten in einen Trainings- und einen Testdatensatz aufgespalten. Es wurde ein Support Vector Machine (SVM) Modell trainiert; die Evaluation erfolgte als 10-fache Kreuzvalidierung (Cross-Validation). Dadurch verbesserte sich das Ergebnis um ca. 10%: Die Trefferquote der Klassifikation lag bei 57% (+/- 7%).[3] Wenn man eine Klassifikation nur zwischen positiven und negativen Sätzen durchführt, beträgt die Trefferquote 68% (+/- 10%).[4]

Für einen dritten Testlauf wurde die Sentimentwortliste bearbeitet und die Textsnippets wurden einem zweiten manuellen Rating mit mehr als nur 3 Kategorien unterzogen. Insbesondere sollte zwischen tatsächlich neutralen Sätzen (z.B. “X wurde 1826 in Berlin geboren.” = 0) und Sätzen mit (einander ausgleichenden) positiven und negativen Bewertungen (z.B. “Trotz dieser erheblichen Schwächen ist X ein Werk gelungen, das …” = 0,###) unterschieden werden. Auch Problemfälle (wie z.B. erwartbare Artefakte durch Zitation oder Ironie) wurden separiert, um die Analyseergebnisse gesondert evaluieren zu können. Von den 1687 eindeutig positiven, negativen oder neutralen Sätzen wurden 65,6% richtig erkannt (Tab. 2).

Dimension	richtig identifizierte Sätze	Precision	Recall	F1 score	Manuelle Annotation
Positiv	599	0.65	0.83	0.73	718
Neutral	375	0.78	0.55	0.65	677
Negativ	133	0.47	0.46	0.46	292

Tab. 2: Ergebnis des dritten Testlaufs

In unserer Präsentation werden wir die ausführliche Ergebnisevaluation des dritten Testlaufs vorstellen und die sich stellenden Probleme im Hinblick auf die eingangs dargestellte Zielsetzung diskutieren. Außerdem soll ein erster Probelauf über das inzwischen provisorisch erstellte Satzkorpus aus den “Grenzboten” präsentiert werden, der die angezielte Methodik der diachronen Trenddarstellung illustriert.

Fußnoten

1. “Historische Rezeptionsanalyse” (gefördert von der Volkswagenstiftung), Teilprojekt “SentiLitKrit” (Göttingen, 2015-2018).
2. http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt
3. Es wurden zusätzlich Lineare SVM und Gaussian Naive Bayes ausprobiert. Die Trefferquote lag bei 0.56 (+/- 0.05) bzw. 0.56 (+/- 0.07).
4. Es wurden zusätzlich Logistic Regression, Lineare SVM und Gaussian Naive Bayes ausprobiert. Die Trefferquote lag bei 0.70 (+/- 0.10), 0.70 (+/- 0.11) bzw. 0.70 (+/- 0.11).

Bibliographie

Kreuzer, Helmut (Hg.) (2006): Deutschsprachige Literaturkritik 1870-1914. Eine Dokumentation. Unter Mitarbeit von Doris Rosenstein. 4 Bde. Frankfurt am Main: Lang.

Liu, Bing (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Mellmann, Katja / Willand, Marcus (2013): Historische Rezeptionsanalyse. Zur Empirisierung von Textbedeutungen. In: P. Ajouri, K. Mellmann & C. Rauen (Hg.): Empirie in der Literaturwissenschaft. Münster: Mentis, S. 263–281.

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): SentiWS - a Publicly Available German-Language Resource for Sentiment Analysis.

In: Proceedings of the 7th International Language Ressources and Evaluation, pp. 1168-1171.

van Bellen, Maurits (2010): Sentiment Analysis on historical book reviews with a Bayesian Classifier. Bachelor-Arbeit. University of Amsterdam.

„Software Aging“ in den DH: Kritik des reinen Forschungswillen

Bürgermeister, Martina
martina.buergermeister@uni-graz.at
ZIM-ACDH, Universität Graz, Österreich

Schneider, Gerlinde
gerlinde.schneider@uni-graz.at
ZIM-ACDH, Universität Graz, Österreich

Makowski, Stephan
stephan.makowski@uni-koeln.de
CCeH, Universität Köln, Deutschland

Jeller, Daniel
daniel.jeller@icar-us.eu
ICARUS, Wien, Österreich

Bigalke, Jan
JBigalke@smail.uni-koeln.de
CCeH, Universität Köln, Deutschland

Theisen, Christian
ctheise1@smail.uni-koeln.de
CCeH, Universität Köln, Deutschland

Vogeler, Georg
georg.vogeler@uni-graz.at
ZIM-ACDH, Universität Graz, Österreich

Einleitung

Dieser Beitrag behandelt die Frage, warum in den DH entwickelte und angewandte Software häufig schnell altert. Jede Software altert relativ zu der Umgebung, in der sie eingesetzt wird, unabhängig von der Qualität am Beginn ihrer Verwendung (Engels et al. 2009: 393). Wandeln sich Hardware, Infrastruktur oder Anforderungen an die Software, wird sie, um weiter brauchbar zu sein, angepasst. Je nach Beschaffenheit können sich diese Anpassungen positiv, oftmals aber auch negativ auf die Lebensdauer und Fitness einer Software auswirken.

Aus der Praxis behaupten wir, dass kontextuelle und inhaltliche Spezifika von DH-Software dazu führen, dass eine langfristige Lauffähigkeit und Brauchbarkeit erschwert werden. Unser Beitrag bringt allgemein die Bedeutung und Relevanz des Themas „Software Evolution“ (2) nahe, beschreibt Spezifika der Software Evolution aus der DH-Praxis (3) und zeigt welche konkreten Maßnahmen im Projekt *monasterium.net* (4) dahingehend gesetzt werden.

Software Evolution

Software Evolution umfasst alle Aktivitäten und Prozesse, die Software verändern (Godfrey/German 2008). Änderungen der Hardware, der Infomationsübermittlung sowie der Anforderungen sind Kräfte die auf diesen Evolutionsprozess wirken. Softwareentwicklungsprozesse werden seit den 1970er Jahren definiert und systematisiert, um die Qualität von Software zu steigern. Aus dieser Zeit stammt auch das Konzept des sogenannten Software Lifecycles und die Idee, diesen Zyklus zu managen (Lehman 1980). Unterschiedliche Methoden und Techniken dazu haben sich seither für alle Phasen im Lebenslauf von Softwaresystemen etabliert. Dank der intensiven Auseinandersetzung mit der Qualitätssteigerung in der Softwareentwicklung wurden die Fehlerquoten gesenkt (Thaller 2000: 6). Hochwertige Software ist nicht nur (nahezu) fehlerfrei, sondern auch kompatibel zur ihrer Umgebung. Verläuft die Evolution einer Software nicht in diesem Sinne, spricht man vom „Software Aging“ beziehungsweise sogar von deren Verfall (Parnas 1994). Demeyer et al. (2013: 4f.) fassen die Symptome veralteter Software wie folgt zusammen: Unvollständige oder keine Dokumentation, fehlende Tests, Ausstieg ursprünglicher Entwickler, verlorengegangenes Insiderwissen, fehlender Überblick über Gesamtsystem, zeitintensive Anpassungen, ständige Fehlerkorrekturen und Wartung damit verbundener Abhängigkeiten, lange „Build“-Zeiten und schlechter Code.

Parnas (1994: 280) erkennt zwei Hauptfaktoren für das Altern von Software. „Lack of movement“, also keine Änderungen an der Software vorzunehmen, und „Ignorant surgery“: Aus der Praxis weiß man, dass bei dringenden Korrekturen am Programmcode, die formale Kriterien für gute Software oftmals nicht eingehalten werden. Ein Beispiel ist das unreflektierte Copy-and-paste aus *Stack Overflow* [1]. Kurzfristige werden den

besten Lösungen vorgezogen. Derartige Eingriffe und nicht-systematisches Vorgehen beschleunigen den Prozess der Softwarealterung. Es wird immer aufwendiger, Änderungen an der Software vorzunehmen.

Demzufolge werden Ideen zur systematischen und automatisierten Verjüngung von Software erforscht und erprobt: Refactoring-Tools, beispielsweise für Java in *Eclipse*, *Python Rope*, oder aber auch für HTML und CSS (Mazinanian/Tsantalis 2017, Harold 2008), wurden entwickelt. Sogenannte „Prediction"- Modelle werden ermittelt, um Softwareevolution besser verstehen zu können und vor allem dem Problem der „Legacy software" zu begegnen (Goltz et al. 2015, Paech et al. 2016).

Software Herausforderungen in der DH Praxis

Diese teilweise schon seit Jahrzehnten bekannten Erkenntnisse aus dem Software Engineering haben für die DH eine besondere Relevanz, da die Projekte hier wesentlich kleinere Budgets, oftmals kurze Projektlaufzeiten und andere Unsicherheiten haben. Aus unserer Erfahrung wird Softwareentwicklung in den DH häufig sehr informell gehandhabt. Diesbezüglich nachhaltiger zu werden, haben unter anderem Czmiel (2017), Schrade (2017) oder Kasper/Grüntgens (2017) gefordert. Nicht nur der Entwicklungsprozess von DH-Software muss längerfristig gedacht werden (Hattrick 2016), auch der Kontext, in dem die Software entsteht und besteht, beeinflusst deren Entwicklung und Veränderung.

Erstens ist es nicht ungewöhnlich, dass Projekte in den DH von einer einzigen Person technisch umgesetzt werden, wie es etwa im Falle von Dissertationsprojekten typisch ist. Der entstandene Code ist bei Projektende lauffähig, es kann aber nicht vorausgesetzt werden, dass dieser auf einen langfristigen Einsatz ausgelegt ist und entsprechend gewissenhaft programmiert und dokumentiert ist. Forschungsergebnisse sind im Projektkontext meist wichtiger als die Qualität der entwickelten Software. Generell bedeutet ein Projektende nicht die Übergabe eines Produktes an einen Kunden, es bedeutet vielmehr: Die Finanzierung läuft aus und der/die Entwickler/in verlässt das Projekt. Was zurückbleibt, ist Software, die von anderen gewartet werden muss. Dazu ist es notwendig, die Dokumentation und Systemarchitektur zu verstehen, sich in den Fremdcode einzuarbeiten. Veränderungen am Code können oft nicht mehr ihrer ursprünglichen Intention entsprechend vorgenommen werden. Die Wartung wird aufwendig und zeitintensiv. Das heißt, die Organisationsstrukturen des Forschungsbetriebes beeinflussen die Alterung von Software.

Zweitens bringen die komplexen Anforderungen der Forschungsdaten nicht-klassische Lösungsansätze mit sich. Mit diesen Ansätzen vertraute Entwickler/innen sind schwer zu finden und zu halten, Einarbeitungsphasen dauern lange. Besonders augenfällig wird das am in den DH weit verbreiteten Gebrauch von X-Technologien. Sie werden immer mehr zur Nischenanwendung. Während die Definitionen von XSLT 1.0 und XPath 1.0 noch von einer größeren Breite von Softwareprodukten implementiert wurden, sogar Teil der Browser wurden, gibt es nur noch wenige Implementationen der Weiterentwicklungen XSLT 2.0 und 3.0. Auch die Menge verwendbarer XML-Datenbanksysteme ist heute geringer als noch vor einigen Jahren. In den DH entwickelte Softwarelösungen sind also speziell auf die Bedürfnisse des Gegenstandes ausgelegt und stellen keine Standardlösungen dar. Sie brauchen spezifisches Know-how, um gewartet werden zu können. Fehlt dieses, beziehungsweise ist es nur mangelhaft vorhanden, droht die Software zum unbrauchbaren Altsystem zu verkommen.

DH-Software verlangt drittens besondere Zuwendung, wenn der Code gleichzeitig die Forschungsergebnisse interpretiert. Wenn die Forschungsleistung also nicht allein in den Daten liegt, braucht es individuelle Wartungslösungen. Eine Digitale Edition kann beispielsweise als die Gesamtheit von Daten, Systemarchitektur, Anwendung und GUI verstanden werden (Andrews/Zundert 2018). Diese Interpretationsleistung als Teil der Forschung muss bei allen Phänomenen der Veränderung an der Edition mitbedacht werden. Die Gefahr ist groß, dass nach einiger Zeit das Argument durch Softwareanpassungen verwässert oder im schlimmsten Fall nicht mehr nachvollziehbar ist und für die Forschung unbrauchbar wird.

Zusammenfassend sehen wir in der nicht langfristigen Finanzierung, der hohen Fluktuation an Personen, der Notwendigkeit von Speziallösungen und im Forschungsgegenstand selbst erhöhten Bedarf an Maßnahmen, um unsere Softwareprojekte lauffähig zu halten.

Anti-Aging Maßnahmen im Projekt monasterium.net

Seit 2008 basiert die Urkundenplattform *monasterium.net* auf *eXist-db* als Applikationsserver und Datenbank. Die Plattform wurde hauptsächlich

von drei aufeinanderfolgenden Hauptentwicklern programmiert. Um die Software zu modularisieren, wurde seit 2011 das *mom-ca*-Framework entwickelt, eine Webapplikation in XRX-Architektur (XQuery, REST, XForms). Die Architektur galt damals in Verbindung mit XML-Datenbanken als Empfehlung, wird allerdings in der modernen Webentwicklung kaum mehr eingesetzt. Mit Auslaufen eines Projektes 2014 verließ der letzte Entwickler mit Überblick über das Gesamtsystem das Projekt. Zuvor wurde der Gesamtcode in ein öffentliches Repository überführt. Wissen und Intentionen gingen jedoch verloren. Wir, als das aktuelle, größtenteils projektfinanzierte Entwicklerteam, beschäftigen uns nun aktiv damit, wie der derzeitige Code-Bestand unter unsteten Umständen wartbar und aktuell gehalten werden kann. Im Folgenden beschreiben wir vier Anti-Aging-Maßnahmen, die einerseits Refactoring (das Überarbeiten des Codes), aber auch ganz grundsätzliche Umstellungen des Entwicklungsworkflows betreffen.

Softwareverwaltung durch *Git* und Nutzung der Services von *GitHub*.

Sowohl Entwicklung als auch Dokumentation erfolgen über ein öffentliches *GitHub*-Repository[2]. Die dadurch verfügbaren Möglichkeiten der Versionsverwaltung, des Bugtracking und des Code Review werden genutzt, um die Qualität des Codes zu verbessern und diesen transparent und nachvollziehbar zu entwickeln.

Einrichtung einer Testumgebung.

Jede Neuentwicklung wird, vor ihrer Übernahme in das Produktivsystem anhand eines festgelegten Testszenarios evaluiert. Durch die Spiegelung des Livesystems auf einem Testserver soll reales Systemverhalten reproduziert werden. Fehler können so vorzeitig entdeckt und behoben werden.

Refactoring von HTML und CSS.

Die Verwendung eines auf den Konzepten von Material Design[3] basierenden CSS-Frameworks garantiert ein konsistentes Gesamtdesign von *monasterium.net*. Teile des Benutzerinterfaces werden dadurch modularisiert und leichter anpassbar. Die Verwendung eines Präprozessors und das Einführen einer Namenskonvention sollen die Wartbarkeit, das Auffinden von Fehlern und die Umsetzung neuer Features erleichtern.

Entwicklung einer RESTful API zwischen Client und Datenbank.

Die zukünftige Kommunikation zwischen Client und Datenbank übernimmt eine neudefinierte REST-API. Die Datenabfrage aus der XML-Datenbank erfolgt noch per XQuery, zurückgeliefert werden wahlweise in XML oder JSON serialisierte Daten. Diese Form des Reengineerings gewährt eine definierte, standardisierte Verarbeitungsweise sowie die Weiternutzung und Kombination multipler Datenquellen. Die Abstraktion von Datenbank, Programmlogik und Benutzeroberfläche erleichtert so in Zukunft deren entkoppelte Anpassung oder Austausch.

Fazit

Softwarealterung ist nicht nur in der Softwareindustrie eine aktuelle und fordernde Problematik. Auch für DH-Forschungsinfrastrukturen ist diesbezüglich ein gezielter Umgang gefragt, um Software fit zu halten. Unwissenheit hinsichtlich der Wartung einer Software kann schlimmstenfalls zu einer zukünftigen Unbrauchbarkeit der Forschungsergebnisse führen. Eine dahingehende Bewusstseinsbildung kann über die empirische Betrachtung vorhandener Praktiken und Lösungswege geschehen.

Anhand von *monasterium.net* haben wir exemplarisch mögliche Verjüngungsmaßnahmen dargestellt. Das Projekt eignet sich als Fallbeispiel, da seine Software eine über zehnjährige Laufzeit aufweist. Geringes Projektbudget und häufiger Personalwechsel mit daraus resultierenden Wissensverlusten haben die Codebasis gezeichnet. Das Projekt zeigt, dass Nachvollziehbarkeit des Entwicklungsprozesses, systematisches und standardisiertes Vorgehen, Modularisierung von Softwarekomponenten sowie kontinuierliches Testing in die Evolution von Software gewinnbringend eingreifen können.

Die Verantwortung kann allerdings nicht allein bei den Entwickler/innen liegen. Um Wissensverluste vorzubeugen, müssen langfristige Strukturen aufgebaut und finanziell abgesichert werden. Es muss Teil der Förderungspolitik werden, die Unausweichlichkeit der Softwarealterung zu bedenken. Sollen Entwicklungen auch nach fünf Jahren noch benutzbar sein, muss der Aufwand der nachhaltigen Entwicklung und Wartung in der Antragsplanung verankert werden.

Fußnoten

1. Stack Overflow ist eine Online Community, zur gegenseitigen Unterstützung und zur Wissensgenerierung bei Fragen zur Softwareentwicklung: stackoverflow.com
2. github.com/icaruseu/mom-ca
3. material.io/guidelines/

Bibliographie

Andrews, Tara / Zundert, Joris van (2018): "What are you Trying to Say? The Interface as an Integral Element of Argument", in: Bleier, Roman et al. (eds.): *Digital Scholarly Editions as Interfaces* (=Schriften des Instituts für Dokumentologie und Editorik). Norderstedt: Books on Demand.

Czmiel, Alexander (2017): "Funktionalität Digitaler Editionen", in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts.* Bern 138-141. http://www.dhd2017.ch/wp-content/uploads/2017/02/Abstractband_ergaenzt.pdf [letzter Zugriff 24. September 2017].

Demeyer, Serge / Ducasse, Stéphane / Nierstrasz, Oscar (2013): *Object-Oriented Reengineering Patterns*. Bern: Square Bracket Associates. http://scg.unibe.ch/download/oorp/OORP.pdf [letzter Zugriff 24. September 2017].

Engels, Gregor et al. (2009) "Design for Future: Legacy-Probleme von morgen vermeidbar?", in: *Informatik Spektrum* 32, 5: 393-397. https://doi.org/10.1007/s00287-009-0356-3 [letzter Zugriff 24. September 2017].

Godfrey, Michael W. / German, Daniel M. (2008): "The Past, Present, and Future of Software Evolution", in: *Proceedings of the 2008 Frontiers of Software Maintenance.*
New York: IEEE 129-138. https://doi.org/10.1109/FOSM.2008.4659256 [letzter Zugriff 24. September 2017].

Goltz, Ursula et al. (2015): "Design for future: managed software evolution", in: *Computer Science - Research and Development 30, 3-4*: 321-331. https://doi.org/10.1007/s00450-014-0273-9 [letzter Zugriff 24. September 2017].

Harold, Rusty Elliotte (2008): *Refactoring HTML. Improving the Design of Existing Web Applications*. Upper Saddle River, NJ: Addison-Wesley.

Hattrick, Simon (2016): Research Software Sustainability. Report on a Knowledge Exchange Workshop. JISC: http://repository.jisc.ac.uk/6332/1/Research_Software_Sustainability_Report_on_KE_Work shop_Feb_2016_FINAL.pdf [letzter Zugriff 24. September 2017].

Kasper, Dominik / Grüntgens, Max (2017): "Nachhaltige Konzeptionsmethoden für Digital Humanities Projekte am Beispiel der Goethe-Propyläen", in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts.* Bern 165-168. http://www.dhd2017.ch/wp-content/uploads/2017/02/Abstractband_ergaenzt.pdf [letzter Zugriff 24. September 2017].

Lehman, Meir M. (1980): "Programs, life cycles, and laws of software evolution", in: *Proceedings of the IEEE 68, 9: 1060-1076.* https://doi.org/10.1109/PROC.1980.11805 [letzter Zugriff 24. September 2017].

Mazinanian, Davood / Tsantalis, Nikolaos (2017): "CCSDev: Refactoring duplication in Cascading Style Sheets", in: *Proceedings of the 39th International Conference on Software Engineering Companion.* New York: IEEE 63-66. https://doi.org/10.1109/ICSE-C.2017.7 [letzter Zugriff 24. September 2017].

Paech, Barbara et al. (2016): "Empirische Forschung zu Software-Evolution", in: *Informatik Spektrum 39, 3*: 186-193. https://doi.org/10.1007/s00287-015-0910-0 [letzter Zugriff 24. September 2017].

Parnas, David L. (1994): "Software Aging", in: *Proceedings of 16th International Conference on Software Engineering.* New York: IEEE 279-287. https://doi.org/10.1109/ICSE.1994.296790 [letzter Zugriff 24. September 2017].

Schrade, Torsten (2017): "Nachhaltige Softwareentwicklung in den Digital Humanities. Konzepte und Methoden", in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts.* Bern 168-171. http://www.dhd2017.ch/wp-content/uploads/2017/02/Abstractband_ergaenzt.pdf [letzter Zugriff 24. September 2017].

Thaller, Georg E. (2000): ISO 9001: *Software-Entwicklung in der Praxis.* Hannover: Heise.

Sprachliche Variation in der Germanistik: eine n-Gramm-basierte Stilanalyse

Andresen, Melanie
Melanie.Andresen@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

An zahlreichen Universitäten werden die wissenschaftlichen Disziplinen Linguistik und Literaturwissenschaft in einem gemeinsamen Studiengang angeboten, der beispielsweise „Germanistik“ oder „Deutsche Sprache und Literatur“ heißt. Dies suggeriert eine große fachliche Nähe dieser Disziplinen, die jedoch im Selbstverständnis der meisten Wissenschaftler/innen dieser Fächer keine Entsprechung hat. Linguistik und Literaturwissenschaft unterscheiden sich in ihrem Erkenntnisinteresse, ihren Methoden und auch in ihrer Sprache, wie punktuell bereits beschrieben wurde: So stellt Haggan (2004) bei der Untersuchung von Titeln wissenschaftlicher Publikationen in Linguistik, Literatur- und Naturwissenschaft fest, dass die Sprache der Literaturwissenschaft sich (auch) an ästhetischen Prinzipien orientiert. Afros und Schryer (2009) kommen zu einem ähnlichen Ergebnis bezüglich der Verwendung von „promotional metadiscourse“ und attestieren sogar verschwimmende Grenzen mit den literarischen Texten selbst (305). Die Studierenden von Studiengängen wie „Germanistik“ finden sich also mit (mindestens) zwei unterschiedlichen Fachkulturen und Sprachen konfrontiert, deren Erwerb überwiegend auf dem Weg der Imitation erfolgt (Graefen 1999). Dieser Beitrag hat das Ziel, die stilistischen Unterschiede zwischen den beiden Fächern mithilfe einer n-Gramm-Analyse zu beschreiben und damit bei Lehrenden und Studierenden zu einem höheren Bewusstsein für die damit verbundenen Herausforderungen beizutragen.

Methode

Die hier verwendete Methode ist eine n-Gramm-Analyse, die (fast) rein datengeleitet funktioniert und keine spezifischen Hypothesen erfordert. Ein n-Gramm ist eine Sequenz aus *n* Elementen, im einfachsten Fall aus Wörtern. N-Gramm-basierte Verfahren sind insbesondere in der Computerlinguistik verbreitet, wenn es um einfach zu berechnende Modellierungen von Sprache geht (Jurafsky und Martin 2009). Auch für die linguistische Interpretation wurden n-Gramme bereits genutzt: Scharloth und Bubenhofer (2012) zeigen bei der Analyse von Tonbandprotokollen zweier 68-Kommunen, dass auf diese Weise charakteristische Muster identifiziert werden können, die mit außersprachlichen Merkmalen der beiden Gruppen in Verbindung gebracht werden können. Mahlberg (2013) nutzt ein ähnliches Vorgehen zur Charakterisierung der Prosa Charles Dickens’, Biber et al. (2004) beschreiben unterschiedliche Formen der Wissenschaftssprache anhand sog. lexical bundles. Mit Ausnahme von Scharloth und Bubenhofer (2012) wird in diesen Ansätzen nur die Tokenebene einbezogen. Im Rahmen des hier präsentierten Vorhabens sollen die Potentiale zusätzlicher syntaktischer Informationen ermittelt werden.

Die Datengrundlage der folgenden Analyse ist ein Korpus aus 60 deutschen Dissertationen (30 pro Fach, ca. 3,5 Mio. Token) aus dem Zeitraum von 2003 bis 2016, die an 15 unterschiedlichen deutschen Universitäten eingereicht wurden und über universitäre Server online zur Verfügung stehen. Im Rahmen der Datenaufbereitung wurden semiautomatisch Textelemente ausgeschlossen, die nicht zur Zielvarietät gehören (Zitate), nicht aus Fließtext bestehen (Tabellen, Abbildungen,...) oder den Textfluss unterbrechen (Fußnoten). Die Texte wurden außerdem automatisch mit Informationen zu Lemma, Wortart und syntaktischen Dependenzstrukturen annotiert. [1]

Aus diesen Daten wurden n-Gramme der Größe $n = 1$ bis 5 generiert, die aus Token bzw. Wortartentags bestehen. Neben traditionellen, linearen n-Grammen, die der Reihenfolge der Wörter an der Textoberfläche folgen, wurden zusätzlich syntaktische n-Gramme generiert, die der Dependenzstruktur im Satz folgen (siehe Abbildung 1, beschrieben von Sidorov et al. 2012, Goldberg und Orwant 2013). Die Frequenzen aller n-Gramme mit mindestens zehn Vorkommen (rund 500.000) wurden signifikanzbasiert mit dem t-Test verglichen (siehe Empfehlungen in Lijffijt et al. 2014, Paquot und Bestgen 2009). Die Auswertung bezieht sich auf die n-Gramme mit den größten Unterschieden zwischen den linguistischen und literaturwissenschaftlichen Texten.

Abbildung 1: Lineare und syntaktische n-Gramme im Vergleich an einem Beispielsatz (vgl. Andresen und Zinsmeister 2017)

Ergebnisse

Exemplarisch werden hier die Ergebnisse zu den Wortarten-Unigrammen sowie den linearen und syntaktischen Token-Trigrammen präsentiert. Abbildung 2 zeigt die 20 Wortarten [2] mit den größten Unterschieden zwischen den beiden Disziplinen. Je weiter außen sich die Wortart befindet, desto größer ist der Unterschied. Wortarten auf der rechten Seite sind in der Literaturwissenschaft, die auf der rechten in der Linguistik häufiger. Der mit Abstand größte Unterschied zeigt sich in den attributiv gebrauchten Possessivpronomen (PPOSAT, z. B. *seine Existenz* (Lit_Stu_30 [3])). Ver-

wandt hiermit sind Unterschiede in den Reflexivpronomen (PRF) und Personalpronomen (PPER). Zusammen mit einer ebenfalls deutlich höheren Frequenz von Eigennamen (NE) spiegelt sich hier, dass sich literaturwissenschaftliche Texte in weitaus höherem Maße als die Linguistik mit Personen beschäftigen, seien es reale Autor/inn/en oder literarische Figuren, zum Beispiel:

(1) So bildet ihr autobiographisches Werk eine Brücke zwischen Tradition und Moderne. (Lit_Stu_30)

Weitere Unterschiede zeigen sich im Zusammenhang mit der Verbverwendung: Bei den finiten Verben der literaturwissenschaftlichen Texte handelt es sich eher um Vollverben (VVFIN), während finite Modal- und Auxiliarverben [4] (VMFIN, VAFIN) in der Linguistik frequenter sind. Korrespondierend dazu sind Auxiliarverben im Infinitiv (insb. *werden* in Passivkonstruktionen mit Modalverb) und Vollverben in ihrer Partizipform (VVPP) in der Linguistik häufiger. Lediglich die Form des passiven Perfekts ist in der Literaturwissenschaft häufiger, wie das Tag VAPP zeigt. Insgesamt lässt sich sagen, dass die Sprache der Linguistik im Vergleich mit der Literaturwissenschaft durch komplexe Verbkonstruktionen gekennzeichnet ist.

In der Linguistik zeigt sich außerdem eine höhere Frequenz von attribuierenden Indefinitpronomen (PIAT). Darunter sind die Lemmata *kein*, *aller* und *beide* am häufigsten. Dies kann damit in Verbindung gebracht werden, dass die Linguistik in stärkerem Maße auf Generalisierungen abzielt. Die höhere Frequenz von Zahlen (CARD) in der Linguistik überrascht nicht, da quantitative Verfahren hier deutlich häufiger zum Einsatz kommen als in der Literaturwissenschaft. Das Tag PRELS (Relativpronomen) weist auf eine häufigere Verwendung von Relativsätzen in der Literaturwissenschaft hin.

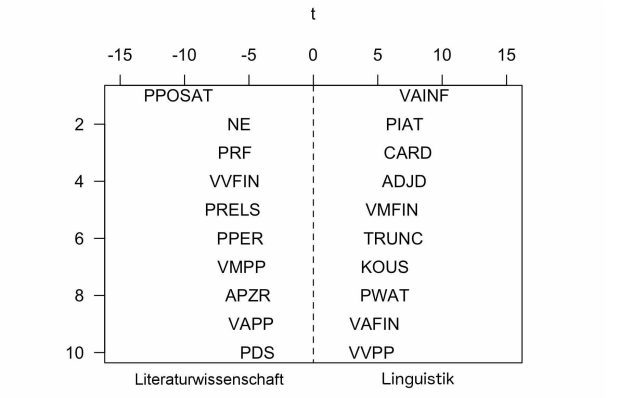


Abbildung 2: Die distinktivsten Wortarten, visualisiert anhand der Teststatistik t. Visualisierung inspiriert durch das R-Paket *stylo* (Eder et al. 2016).

Ergänzend werden in Tabelle 1 Informationen auf Wortebene herangezogen, zunächst ohne zusätzliche syntaktische Information. Gezeigt werden die zehn distinktivsten linearen Trigramme (inkl. Interpunktion). Hier spiegeln sich viele der Phänomene, die bereits auf Ebene der Wortarten erkennbar waren. Die für die Literaturwissenschaft charakteristischen Muster scheinen mehrheitlich aus Relativsätzen zu stammen. Dabei ist zu bedenken, dass der Anfang von Nebensätzen besonders leicht durch eine n-Gramm-Analyse erfasst werden kann, da hier nur ein begrenztes Maß an Variation möglich ist. Das klarste Trigramm für die Linguistik hingegen weist auf die bereits beschriebene häufigere Verwendung von Passiv und Modalverben hin, hier speziell in Kombination miteinander. Interessant ist das n-Gramm *die bei der*, das in 36 von 43 Fällen aus einem Relativsatz stammt, aber in der Linguistik häufiger ist. Mit dem Relativpronomen *die* kann es sich auf Feminina beziehen, in der Mehrzahl handelt es sich im Korpus aber um Substantive im Plural. Das passt zu der bereits oben genannten Annahme, dass die Literaturwissenschaft sich tendenziell exemplarisch mit konkreten Einzelphänomenen beschäftigt, die Linguistik hingegen in stärkerem Maße Generalisierungen anstrebt, die den Plural wahrscheinlich machen.

Rang	n-Gramm	Häufiger im Fach
1	, der sich	Literaturwissenschaft
2	, der seine	Literaturwissenschaft
3	, das sich	Literaturwissenschaft
4	werden können .	Linguistik
5	, die ihm	Literaturwissenschaft
6	, der die	Literaturwissenschaft
7	, dass eine	Linguistik
8	, vor dem	Literaturwissenschaft
9	, die er	Literaturwissenschaft
10	die bei der	Linguistik

Tabelle 1: Die distinktivsten linearen Token-Trigramme

Tabelle 2 zeigt die häufigsten syntaktischen Trigramme, die zusätzlich Informationen zur Depen-

denzstruktur im Satz nutzen. „>“ zeigt hier ein syntaktisches Dominanzverhältnis an. Die Relativsatzmuster sind hier nicht vorhanden, da ihre gute Erkennbarkeit in der Analyse vermutlich primär auf der linearen Abfolge von Interpunktion, Relativpronomen und folgendem Wort beruht. Das Muster *können>werden>.* ist höher gerankt als das Gegenstück in der linearen Analyse, da hier nicht nur unmittelbar aufeinanderfolgende Instanzen erfasst werden, sondern auch solche mit Distanzstellung:

> (2) Einige Substantive können nicht eindeutig einer Geschlechtskategorie zugeordnet werden [...]. (Lin_Bam_01)

Zusätzlich tauchen Kombinationen von Passiv mit dem Modalverb *müssen* und *können* mit *sein* auf. Viele der für die Literaturwissenschaft charakteristischen Muster haben eine direkte lineare Entsprechung: So steht das syntaktische n-Gramm *für>Leben>das* für die lineare Abfolge *für das Leben*. Allerdings umfasst das syntaktische n-Gramm zusätzlich Instanzen, in denen beispielsweise das Substantiv noch durch Attribute modifiziert wird, z. B. *für das eigene Leben* (Lit_Jen_19).

Rang	n-Gramm	Häufiger im Fach
1	können>werden>.	Linguistik
2	in>Regel>der	Linguistik
3	und>können>werden	Linguistik
4	in>Vorstellung>der	Literaturwissenschaft
5	müssen>werden>,	Linguistik
6	in>Darstellung>der	Literaturwissenschaft
7	für>Leben>das	Literaturwissenschaft
8	ist>Wunsch>der	Literaturwissenschaft
9	mit>Realität>der	Literaturwissenschaft
10	können>sein>.	Linguistik

Tabelle 2: Die distinktivsten syntaktischen Token-Trigramme

Fazit

In der hier präsentierten Analyse konnten deutliche stilistische Unterschiede zwischen Linguistik und Literaturwissenschaft gezeigt werden. Die Linguistik zeichnet sich demzufolge durch komplexe Verben (Passiv und Modalverben), die stärkere Verwendung von Zahlen sowie Mustern der Generalisierung aus. In der Literaturwissenschaft finden sich mehr Bezüge auf Personen und komplexe Nominalphrasen mit Relativsätzen.

Die verwendete Methode hat stark explorativen Charakter, sodass viele der hier angebotenen Interpretationen zunächst als Hypothesen betrachtet werden sollten und einer sorgfältigen Prüfung in Folgestudien bedürfen. Zusätzlich ergibt sich mit der Methode eine Beschränkung auf Phänomene, die sich auf konstante Weise auf der sprachlichen Oberfläche niederschlagen.

Im Rahmen dieses Beitrags wurden nur besonders hoch gerankten n-Gramme betrachtet und interpretiert. Andresen und Zinsmeister (2017) präsentieren ergänzend ein Annotationsexperiment, in dessen Rahmen insgesamt 420 Token- und Wortarten-n-Gramme auf die enthaltenen linguistischen Informationen hin ausgewertet wurden. In Folgearbeiten gilt es das Potential unterschiedlicher n-Gramm-Typen für eine Stilanalyse zu erforschen. Der Fokus wird dabei auf stärker syntaktisch informierten Formen liegen, die beispielsweise Informationen zu Token und Wortart kombinieren oder die syntaktischen Dependenzrelationen zwischen den Elementen einbeziehen.

Fußnoten

1. mit einem auf der Dependenzversion des TIGER-Korpus (Seeker und Kuhn 2012) trainierten Modell für MATE (Bohnet 2010)
2. Die Wortartentags stammen aus dem STTS (Schiller et al. 1999).
3. Die Bezeichnung der Korpustexte setzt sich aus einem Kürzel für die Disziplin, die Universität und einer fortlaufenden Zahl zusammen.
4. Bei den Verben *haben*, *sein* und *werden* wird nicht zwischen einer Verwendung als Auxiliar- oder Vollverb unterschieden.

Bibliographie

Afros, Elena / Schryer, Catherine F. (2009): „Promotional (meta)discourse in research articles in language and literary studies“, in: *English for Specific Purposes*. 28 (1), 58–68, doi: 10.1016/j.esp.2008.09.001.

Andresen, Melanie / Zinsmeister, Heike (2017): „Approximating Style by N-gram-based Annotation“, in: *Proceedings of the Workshop on Stylistic Variation*. Copenhagen, Denmark: Association for Computational Linguistics 105–115.

Biber, Douglas / Conrad, Susan / Cortes, Viviana (2004): „If you look at…: Lexical bundles in university teaching and textbooks“, in: *Applied linguistics*. 25 (3), 371–405.

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.

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

Goldberg, Yoav / Orwant, Jon (2013): „A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books“, in: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA. 241–247.

Graefen, Gabriele (1999): „Wie formuliert man wissenschaftlich?“, in: Barkowski, Hans; Wolff, Armin (Hrsg.) *Alternative Vermittlungsmethoden und Lernformen auf dem Prüfstand. Wissenschaftssprache - Fachsprache. Landeskunde aktuell. Interkulturelle Begegnungen - interkulturelles Lernen*. Regensburg: Fachverband Deutsch als Fremdsprache (Materialien Deutsch als Fremdsprache), 222–239.

Haggan, Madeline (2004): „Research paper titles in literature, linguistics and science: dimensions of attraction“, in: *Journal of Pragmatics*. 36 (2), 293–317, doi: 10.1016/S0378-2166(03)00090-0.

Jurafsky, Dan / Martin, James H. (2009): *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2. Aufl. Upper Saddle River, London: Pearson (Prentice Hall series in artificial intelligence).

Lijffijt, Jefrey / Nevalainen, Terttu / Säily, Tanja / Papapetrou, Panagiotis / Puolamäki, Kai / Mannila, Heikki (2014): „Significance testing of word frequencies in corpora“, in: *Digital Scholarship in the Humanities*. 1–24, doi: 10.1093/llc/fqu064.

Mahlberg, Michaela (2013): *Corpus stylistics and Dickens's fiction*. New York: Routledge (Routledge advances in corpus linguistics).

Paquot, Magali / Bestgen, Yves (2009): „Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction“, in: Jucker, Andreas H.; Schreier, Daniel; Hundt, Marianne (Hrsg.) *Corpora: Pragmatics and Discourse*. Brill 247–269, doi: 10.1163/9789042029101_014.

Scharloth, Joachim / Bubenhofer, Noah (2012): „Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse“, in: Felder, Ekkehard; Müller, Marcus; Vogel, Friedemann (Hrsg.) *Korpuspragmatik: thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin [u.a.]: De Gruyter (Linguistik - Impulse & Tendenzen), 195–230.

Schiller, Anne / Teufel, Simone / Thielen, Christine / Stöckert, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Stuttgart, Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [letzter Zugriff am 11.01.2018].

Seeker, Wolfgang / Kuhn, Jonas (2012): „Making Ellipses Explicit in Dependency Conversion for a German Treebank“, in: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey 3132–3139.

Sidorov, Grigori / Velasquez, Francisco / Stamatatos, Efstathios / Gelbukh, Alexander / Chanona-Hernández, Liliana (2012): „Syntactic Dependency-Based N-grams as Classification Features“, in: Batyrshin, Ildar; Mendoza, Miguel González (Hrsg.) *Advances in Computational Intelligence*. Springer (Lecture Notes in Computer Science), 1–11, doi: 10.1007/978-3-642-37798-3_1.

The 'Tiroler Soldaten-Zeitung' and its Authors. A Computer-Aided Search for Robert Musil

Salgaro, Massimo
massimo.salgaro@univr.it
University of Verona, Italy

Rebora, Simone
simone.rebora@univr.it
University of Verona, Italy

Lauer, Gerhard
gerhard.lauer@unibas.ch
University of Basel, Switzerland

Herrmann, J. Berenike
berenike.herrmann@unibas.ch
University of Basel, Switzerland

Robert Musil, one of the most important authors of the twentieth-century German-written literature, fought in the Austrian army at the Italian front. During the First World War, between 1916 and 1917, Musil was chief editor of the *Tiroler Soldaten-Zeitung* in Bozen. This activity has always been a philological problem for Musil scholars,

who have not been able to attribute with certainty a range of texts to the author. However, their identification is fundamental in the study of his political thinking. With this paper, we present a new approach, that combines historical and philological research with stylometric methods.

The starting point for the determination of possible authorship is the screening of previous attempts. The number of articles attributed to Musil has so far varied extensively:

Attribution proposed by	Number of TSZ articles attributed to Musil
(Dinklage 1960)	3
(Roth 1972)	19
(Corino 1973, 2003, and 2010)	8
(Arntzen 1980)	22
(Fontanari / Libardi 1987)	36
(Amann *et al.* 2009)	36

We have limited our test set to the 38 TSZ articles listed by (Schaunig 2014), for which Musil's authorship has been proposed at least once. The major problem for carrying out a stylometric analysis on this corpus is text length. As demonstrated by recent research, the minimum length for a reliable authorship attribution is around 5,000 words (see Eder 2015). However, the average length of the 38 disputed TSZ articles is slightly below 1,000 words (see Figure 1). As a possible solution for this issue, we decided to develop a combinatory design that analyzes longer chunks composed by the juxtaposition of single texts. To reduce the number of combinations, we excluded the nine shortest texts (below 500 word), together with the only text attributed to Musil on solid philological ground (see Corino 1973). This leaves us with a corpus of 28 texts, already digitized by (Amann *et al.* 2009). The optimal configuration was obtained by combining groups of 6 texts. This permutation generated 376,740 text chunks with an average length of N=6,963 words and a standard deviation of 909 words.

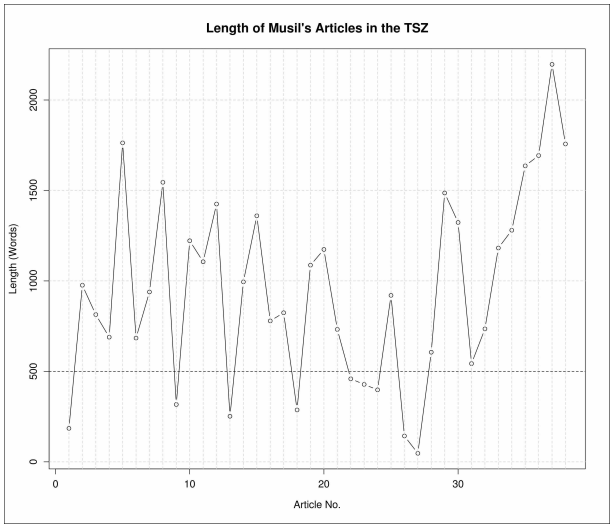


Figure 1. Test set composition

As for the composition of the training set, we drew both on the "impostors method" (see Koppel / Winter 2014) and on historiographical research. Following (Juola 2015), we fixed the number of "impostors" to a minimum of three: Franz Blei, Franz Kafka, and Stefan Zweig. Subsequently, we selected three authors suggested by (Urbaner 2001) as possible TSZ collaborators: Marie delle Grazie, Hugo Salus, and Albert Ritter (his texts were not available in digitized format, so we OCRed and manually refined them). The training set was then completed by a selection of articles published by Musil in various journals between 1911 and 1919. For each author, the retrieved material was subdivided in three text chunks with a length comprised between 6,000 and 8,000 words: the training set was thus composed by 21 text chunks (see Figure 2).

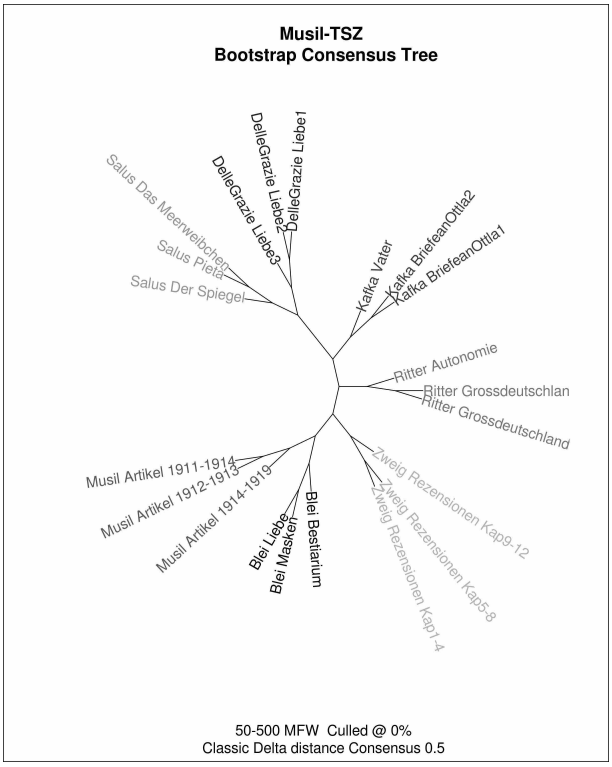


Figure 2. Training set

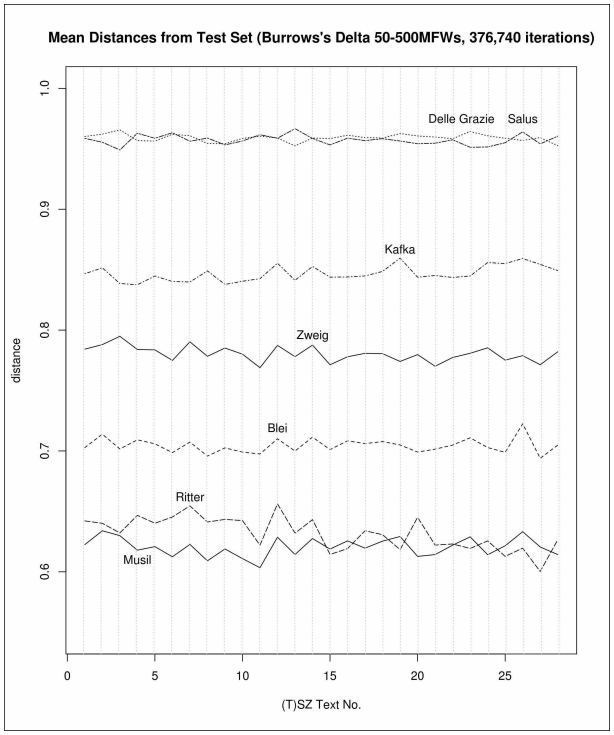


Figure 3. Combinatory design results

The analysis was carried out using the R package *Stylo* (see Eder / Rybicki / Kestemont 2016). For each iteration, the distances between test set and training set were saved in the tabular form provided by the package. At the end of the process, mean values were calculated by sub-grouping the combinations by each TSZ text. Notwithstanding the employment of a high-standard computational power (provided by GWDG, University of Göttingen), a first experiment using 50–500 most frequent words (MFW) and Burrows's Delta distance took more than one week to be completed (see Figure 3). However, when repeating the experiment with only one-tenth of the combinations (i.e. 37,674 iterations, randomly selected), results were rather identical (see Figure 4) and the process took less than one day. When the experiment was repeated without any combination, results were extremely noisier (see Figure 5), thus confirming that the combinatory design was able to better discern authorial signals.

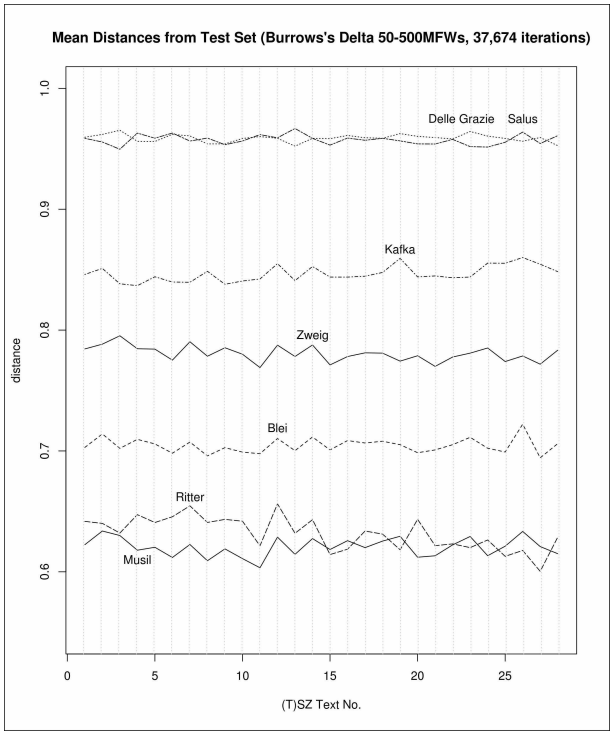


Figure 4. Combinatory design results (one-tenth iterations)

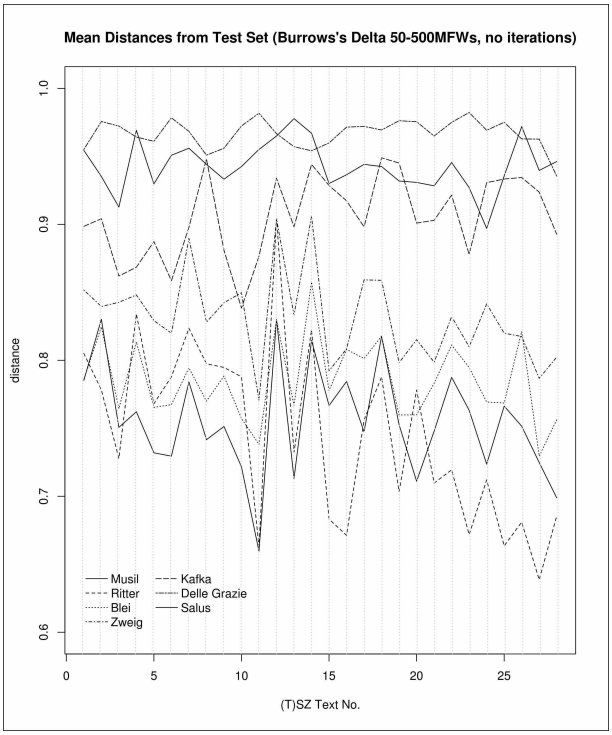
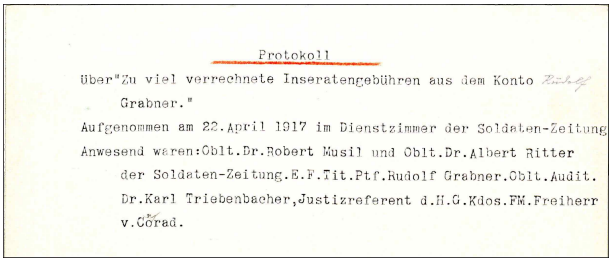


Figure 5. Results without combinatory design

To validate the results, the experiment has been repeated with 16 different configurations, by combining Eder's Delta, Burrow's Delta, Canberra, and Cosine distances with 10–100, 20–200, 50–500, and 100–1,000 MFW. In all configurations, Ritter and Musil are the only authors disputing the authorship of the TSZ articles. This evidence has been corroborated by the discovery of a document in the *Kriegsarchiv* in Wien, which confirms that Albert Ritter was part of the TSZ editorial team (see Figure 6).


Protokoll

Über"Zu viel verrechnete Inseratengebühren aus dem Konto
Grabner."
Aufgenommen am 22.April 1917 im Dienstzimmer der Soldaten-Zeitung
Anwesend waren:Oblt.Dr.Robert Musil und Oblt.Dr.Albert Ritter
der Soldaten-Zeitung.E.F.Tit.Ptf.Rudolf Grabner.Obit.Audit.
Dr.Karl Triebenbacher,Justizreferent d.H.G.Kdos.FM.Freiherr
v.Conrad.

Figure 6. Source: Kriegsarchiv, Wien

Final results have been synthetized here:

TSZ articles' titles and dates of publication	Agreement between classifiers on Musil's authorship
1. „Kameraden arbeitet mit!“ (6. 8. 1916)	100,00%
2. „Bin ich ein Österreicher?“ (20. 8. 1916)	87,50%
3. „Herr Tüchtig und Herr Wichtig” (27. 8. 1916)	81,25%
4. „Das Schlagwort” (27. 8. 1916)	100,00%
5. „Die Erziehung zum Staat” (3. 9. 1916)	100,00%
6. „Bauernleben” (1. 10. 1916)	100,00%
7. „Sonderbare Patrioten” (15. 10. 1916)	100,00%
8. „Noch einmal Bauernleben” (29. 10. 1916)	100,00%
9. „Opportunität” (12. 11. 1916)	100,00%
10. „Eine gute persönliche Beziehung” (26. 11. 1916)	100,00%
11. „Eine österreichische Kultur” (10. 12. 1916)	100,00%
12. „Der Nörgler und der neue Österreicher” (17. 12. 1916)	100,00%
13. „Das Kompromiß” (24. 12. 1916)	100,00%
14. „Heilige Zeit” (31. 12. 1916)	100,00%
15. „Zentralismus und Föderalismus” (7. 1. 1917)	68,75%
16. „Föderalismus oder Zentralismus” (14. 1. 1917)	68,75%
17. „Zu Milde und zu Wilde” (11. 2. 1917)	93,75%
18. „Neu-Altösterreichisches” (25. 2. 1917)	87,50%
19. „Ist die »österreichische Frage« schwierig?” (4. 3. 1917)	62,50%
20. „Seiner Hochwohlgeboren!” (4. 3. 1917)	100,00%
21. „Luxussteuern” (4. 3. 1917)	93,75%
22. „Positive Ziele” (11. 3. 1917)	81,25%
23. „Der Frieden versprochen!” (18. 3. 1917)	68,75%
24. „Das Staatsprogramm der Deutschen” (18. 3. 1917)	87,50%
25. „Wehe dem Staatsmann!” (25. 3. 1917)	68,75%

26. „Der Frieden und die Zukunft" (1. 4. 1917)	62,50%
27. „Presse und Krieg" (8. 4. 1917)	68,75%
28. „Vermächtnis" (15. 4. 1917)	100,00%

A general trend is evident: while, for the articles published in 1916, Musil's authorship is almost unquestionable, many more doubts emerge with the articles published in 1917. In no case, however, Ritter's signal becomes dominant. Notwithstanding the high margins of uncertainty, these results are to be considered as significant for multiple reasons. First, the combinatory design, while having shown the dominance of Musil's signal throughout the test set, may have overshadowed different, minor signals. Second, it should be considered the fact that Musil, in the role of chief editor, may have altered many articles in the journal, thus intermixing his authorial signal with those of others. All this considered, further research is advisable, while the focus should be shifted towards the texts on which classifiers disagree.

Among possible future developments of the research, is the definition of new training sets to validate the results and an expansion of the test set. Both these developments, however, will require an extensive digitization effort: most of the useful texts, in fact, are not available in a clean plain-text format. In addition, other software should be tested on the already defined corpus, from JGAAP (see Juola *et al.* 2008) to the CLEF/PAN software (see Stamatatos *et al.* 2014), focusing specifically on different methods for authorship attribution, from lower-level features such as character n-grams (see Halvani *et al.* 2016), to higher-level features such as syntactic labels (see Hirst / Feiguina 2007), taking into consideration also machine-learning techniques (see Jockers / Witten 2010). With our study, we hope to have cast the groundwork for a research that can have long-lasting consequences on the history of German literature, confirming at the same time how quantitative methods are not in opposition, but complementary to qualitative analysis (see Herrmann 2018).

Bibliography

Amann, Klaus / **Corino, Karl** / **Fanta, Walter** (2009): *Robert Musil, Klagenfurter Ausgabe.* Klagenfurt: Robert Musil-Institut der Universität Klagenfurt.

Arntzen, Helmut (1980): *Musil-Kommentar sämtlicher zu Lebzeiten erschienener Schriften außer dem Roman "Der Mann ohne Eigenschaften"*. München: Winkler.

Corino, Karl (1973): "Robert Musil, Aus der Geschichte eines Regiments", in: Studi Germanici 11: 109–115.

Corino, Karl (2003): *Robert Musil: eine Biographie*, Reinbek bei Hamburg: Rowohlt.

Corino, Karl (2010): "Klaviersonnen über Schluchten des Gemüts. Robert Musil und die Musik", in: *Das Plateau* 120: 4–21.

Dinklage, Karl (1960): *Robert Musil. Leben, Werk, Wirkung*, Zürich.

Eder, Maciej / **Kestemont, Mike** / **Rybicki, Jan** (2016): "Stylometry with R: a package for computational text analysis", in: *R Journal* 8(1): 107–121.

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

Fontanari, Alessandro / **Libardi, Massimo** (1987): *La guerra parallela*. Trento: Reverdito.

Halvani, Oren / **Winter, Christian** / **Pflug, Anika** (2016): "Authorship verification for different languages, genres and topics", in: *Digital Investigation* 16: 33–43.

Herrmann, J. Berenike (2018): "In test bed with Kafka. Introducing a mixed-method approach to digital stylistics", in: *Digital Humanities Quarterly* [in press].

Hirst, Graeme / **Feiguina, Ol'ga** (2007): "Bigrams of syntactic labels for authorship discrimination of short texts", in: *Literary and Linguistic Computing* 22(4): 405–417.

Jockers, Mattew / **Witten, Daniela** (2010): "A comparative study of machine learning methods for authorship attribution", in: *Literary and Linguistic Computing* 25(2): 215–223.

Juola, Patrick / **Noecker, John** / **Ryan, Mike** / **Zhao, Mengjia** (2008): "JGAAP3.0 – authorship attribution for the rest of us", in: *Digital Humanities 2008: Book of Abstracts*: 250–251.

Juola, Patrick (2015): "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions", in: *Digital Scholarship in the Humanities* 30: 100–113.

Koppel, Moshe / **Winter Yaron** (2014): "Determining if two documents are by the same author", in: JASIST 65(1): 178-187.

Roth, Marie-Louise (1972): *Robert Musil. Ethik und Ästhetik*. München: List.

Schaunig, Regina (2014): *Der Dichter im Dienst des Generals. Robert Musils Propagandaschriften im ersten Weltkrieg*. Klagenfurt: Kitab.

Stamatatos, Efstathios / **Daelemans, Walter** / **Verhoeven, Ben** / **Potthast, Martin** / **Stein, Benno** / **Juola, Patrick** / **Sanchez-Perez, Miguel A.** / **Barrón-Cedeño, Alberto** (2014): "Overview of

the Author Identification Task at PAN 2014 ", in: *CLEF 2014 (Working Notes)*: 877-897.

Urbaner, Roman (2001): "'... daran zugrunde gegangen, daß sie Tagespolitik treiben wollte'? Die ‚(Tiroler) Soldaten-Zeitung' 1915-1917", in: *eForum zeitGeschichte* 3/4. www.eforum-zeitgeschichte.at [accessed 14.09.2017]

Vagheit hoch Zweifel plus Kritik! Die Bewertung von Widersprüchen in einer digitalen Entzifferungsarbeit der Maya-Hieroglyphen

Gronemeyer, Sven
sgronemeyer@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universtität Bonn, Deutschland; La Trobe University Melbourne, Australien

Diehr, Franziska
diehr@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Prager, Christian
cprager@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universtität Bonn, Deutschland

Diederichs, Katja
katja.diederichs@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universtität Bonn, Deutschland

Wagner, Elisabeth
ewagner@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universtität Bonn, Deutschland

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

Grube, Nikolai
ngrube@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universtität Bonn, Deutschland

Die Hieroglyphenschrift des Klassischen Maya

Die logosyllabische Hieroglyphenschrift der Maya umfasst rund 1000 Zeichen und wurde etwa zwischen 350 v. Chr. und 1550 n. Chr. im südlichen Mesoamerika zur Aufzeichnung der Hochsprache des Klassischen Maya verwendet. Die Anzahl der Graphe ist raum-zeitlich betrachtet mit rund 3000+ Formen weitaus höher, da Zeichen gleichzeitig mehrere Graphvarianten aufweisen können, die von einer Vollform abgeleitet sind.

Einzelne Graphe werden in einem meist mit einem Wort oder morphemischem Verbund identischen Hieroglyphenblock arrangiert, ähnlich dem koreanischen Hangul. Allerdings offenbart das Maya aufgrund seines Variantenreichtums eine wesentlich größere kalligraphische Freiheit als die einzelnen Varianten nur durch simples Aneinanderreihen im Block zu schreiben. Je nach Platzbedarf und Ästhetik können Graphe etwa miteinander verschmelzen, infigiert oder gedreht werden.

Bei der Katalogisierung der Graphe muss weiterhin mehr als ein Jahrtausend paläographischer Entwicklung berücksichtigt werden, ebenso unterschiedliche Stile in skulptierten oder gemalten Texten. Wegen all dieser Eigenheiten und der herausfordernden Struktur widersetzt sich die Maya-Schrift aktuell, Teil des Unicode-Standards zu werden.

Zeichenkataloge als Hilfskonstrukte der Epigraphik

Bis in die 1950er Jahre war die Maya-Schrift nicht entziffert und blieb es in großen Teilen bis in die 1980er Jahre, als eine Reihe bahnbrechender Erkenntnisse einen Kaskadeneffekt in Gang setzte, etwa Stuart (1987). Bis heute kennt man ebensowenig die genaue Zahl der Zeichen und ihrer graphischen Repräsentationen, da alle bisher publizierten Verzeichnisse aufgrund des Entzifferungsprozesses unvollständig und unzulänglich sind. Über 300 Zeichen sind bis heute nur vage oder gar nicht entziffert. Für viele dieser Fälle existieren konkurrierende Entzifferungsvorschläge, die vielleicht nur in ausgewählten Kontexten valide sind, sich aber wegen möglicher Po-

lyvalenz nicht gegenseitig ausschließen müssen. Es gilt nicht nur, das Zeicheninventar vollständig zu erfassen, sondern existierende Entzifferungsvorschläge kritisch im Textzusammenhang zu prüfen, ob sie verifizierbar sind, und wo sie sich als falsch oder nicht überprüfbar herausstellen und somit nicht weiter berücksichtigt werden müssen.

Die elf bisher publizierten Zeicheninventare, vor allem Thompson (1962), weisen viele Schwachstellen auf, besonders problematisch sind dabei Mehrfachinventarisierungen von Allographen als verschiedene Zeichen. Ein weiterer offensichtlicher Nachteil der traditionellen Zeichenkataloge ist die unveränderbare Natur einer gedruckten Fassung. Dies verhindert, dass gegebene Mehrfach- und Fehlklassifikationen korrigiert oder neue Beziehungen zu Zeichen und zu verschiedenen Zeichenfunktionen erstellt werden können, wobei zudem neue Entzifferungen nicht berücksichtigt werden können.

Um einen Überblick über die bisher geleistete Zeichenklassifikationsarbeit der Mayaschriftforschung zu schaffen und den Vergleich der Inventare zu ermöglichen, fließen die bisher publizierten Kataloge in unseren digitalen Zeichenkatalog als Konkordanz ein. Die fehlerhaften Klassifikationen werden durch unseren Katalog korrigiert, bleiben aber nachvollziehbar dokumentiert, da wir die konkordanten Katalogeinträge beim jeweiligen Graph, das optional einem Zeichen zugeordnet werden kann, erfassen.

Ein digitales Zeichen- und Graphinventar für das Klassische Maya

Als Resultat interdisziplinärer Arbeit, bei dem die Modellierung und Verarbeitung der Daten auf Grundlage epigraphischer Prinzipien und Forschungsfragen erfolgte, ist unser digitaler Zeichenkatalog so konzipiert, dass er sowohl bisherige Forschungsergebnisse kritisch abbilden als auch noch zu erwartende Erkenntnisse flexibel einbinden kann.

Der Katalog basiert auf einem innovativen Konzept der flexiblen Zuordnung von Zeichen zu ihren Graphen: Zeichen als Träger sprachlicher Informationen und Graphe als Form ihrer schriftlichen Realisierung werden getrennt erfasst, und erst die Verbindung eines Zeichens mit seinen Graphen macht es zu dessen Allograph, deren Gesamtheit bildet das Graphem eines Zeichens, das phonemisch KV-Silben oder freie und gebundene Morpheme, Diakritika und Zahlen wiedergibt.

Die ontologisch-vernetzte Struktur des Modells und dessen Implementierung in RDF erlauben es, semantische Relationen über persistente URIs zwischen eindeutig referenzierbaren Entitäten herzustellen. Dadurch ist es möglich, Zuordnungen flexibel anzupassen und Allographe durch neue Verknüpfungen hinzuzufügen oder Falschzuweisungen zu korrigieren, etwa wenn ein Graph in Wirklichkeit aus zwei Graphen verschiedener Zeichen besteht.
Sobald die Neuinventarisierung abgeschlossen ist (voraussichtlich Mitte 2018), wird der Katalog auf unserem zukünftigen Projektportal (*https://classicmayan.org/*) publiziert und die RDF-Daten über einen SPARQL-Endpoint zugänglich gemacht. Zusätzlich werden die Daten auch im TextGrid Repository (*https://textgridrep.org/*) veröffentlicht, wo sie mittels einer OAI-PMH Schnittstelle auch für externe Nutzer abrufbar sind. Die Dokumentation des digitalen Zeichenkatalogs ist unter *http://idiom-projekt.de/catalogue* erreichbar.

Bewertung und qualitative Einstufung von Lesungshypothesen

Bei einer noch nicht vollständig entschlüsselten Schrift machen Epigraphiker zwangsläufig verschiedene Annahmen zur phonemischen Lesung von Zeichen. Es ist notwendig, alle plausiblen und nicht eindeutig widerlegten Entzifferungsvorschläge zu dokumentieren, vor allem aber, deren Qualität nach formalen Kriterien bewertbar zu machen. Dazu haben wir ein neutrales, transparentes System modelliert, das anhand formaler Kriterien eine qualitative Einstufung von Lesungshypothesen ermöglicht und deren Plausibilität im Textkorpus überprüfbar macht.

Die Zeichen werden Zeichenklassen (syllabisch, morphographisch, diakritisch und/oder numerisch) zugewiesen, die jeweils einen breiten Transliterationswert ohne allophonische Varianz haben, z.B. “ku” für das Silbenzeichen 528. Aufgrund der Polyvalenz hat das Zeichen noch zwei logographische Lesungen mit distinktem Transliterationswert, statt des normalen “TUN” wird “CHAHUK” gelesen, wenn es als Tagesname verwendet wird. Für die Konfidenz eines Transliterationswertes werden jene Kriterien ausgewählt, auf die er sich stützt (siehe Abb. 1). Für jede Art der Zeichenfunktion wurde ein eigenes Kriterien-Set basierend auf Kelley (1962) und Houston (2001) entwickelt, das sich vor allem am graphematischen und sprachlichen Nutzungskontext orientiert, z.B. hat die Übereinstimmung mit einer

bestimmten Wortart durch das Syntagma für ein Silbenzeichen keine Bedeutung, jedoch für Logogramme - hier besonders auch der Nachweis in modernen Maya-Sprachen.

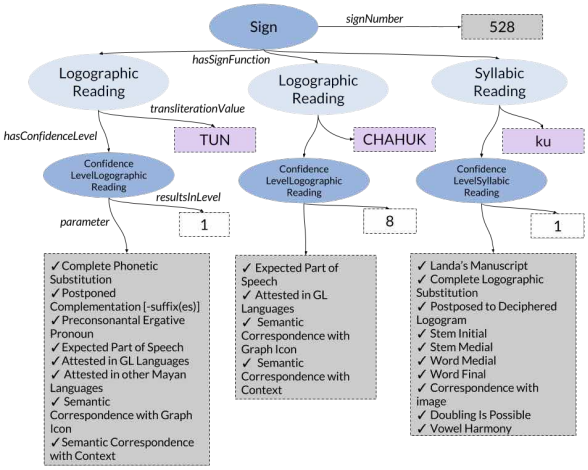


Abb. 1: Polyvalentes Zeichen mit drei Lesungen und jeweils unterschiedlichen Konfidenzen.

Die Kriterien sind mittels Aussagenlogik so miteinander in Bezug gesetzt, dass je nach Kombination eine qualitative Einstufung vorgenommen wird. Dabei steht "1" für die höchstmögliche Konfidenz und eine evidente Lesung. Die Anzahl der Konfidenzstufen je Zeichenfunktion ist unterschiedlich: während Logogramme eine granulare Einteilung benötigen, brauchen Silbenzeichen weniger Stufen, da deren Permutationen im Kontext eines Wortes recht eindeutig sind.

Das Wortzeichen "CH'AM" etwa taucht mit phonemischen Komplementen auf, die entsprechenden Kriterien ergeben Stufe 2. Damit liegt eine wahrscheinliche, aber ohne funktional äquivalente syllabische Substitution noch keine gesicherte Entzifferung vor. Die Kriterien sind bewusst streng gehalten, um für jede Lesungsaussage eine kritische Evaluation gegenüber den Zeichenvorkommen im von uns TEI-kodierten Textkorpus durchführen zu können und damit unserem Wörterbuch eine hohe Zuverlässigkeit für den Nutzer zu geben. Lesungen unterhalb einer bestimmten Konfidenz werden nämlich nicht aufgenommen, so dass in unserem Wörterbuch bestimmten Entzifferungsvorschlägen der normative Charakter genommen wird, den sie vielleicht in der epigraphischen Forschung durch Zitierung gewonnen haben.

Zusammenspiel von Zeichenkatalog, Textkorpus und linguistischer Analyse

Ohne eine sichere Identifizierung aller Graphe und mit vielen Zeichen unbekannter oder umstrittener Lesung kann das TEI-kodierte Textkorpus weder aus einem festen Schriftzeichensatz wie Unicode noch aus phonemisch transliterierten Werten bestehen. Für ein flexibles Korpus, das auf neue Entzifferungen und verschiedene Lesungshypothesen reagieren kann, nutzen wir den Zeichenkatalog als eine Art "Grundbaukasten" bei der Korpuserstellung.

Im kodierten Text wird jede Hieroglyphe mittels Katalognummer und einer Referenz zur URI im Zeichenkatalog erfasst. Mittels einer Software zur linguistischen Analyse, Teil unserer virtuellen Arbeitsumgebung, wird in einem weiteren Prozessierungsschritt die numerische Transliteration (Katalognummern) zunächst in eine graphemische Transliteration (Transliterationswerte) überführt. Dies geschieht wegen der Polyvalenz semi-automatisch, wenn der Epigraphiker nach Verwendungskontext eine Entscheidung fällen muss - erst damit wird der Text "menschenlesbar".

Die qualitative Einstufung der Entzifferungsvorschläge im Zeichenkatalog wird jetzt relevant: Die Analyse kann anhand hoher oder niedriger Konfidenzstufen durchgeführt werden. Die Entzifferungsaussagen können in ihrem Verwendungskontext überprüft werden, was idealerweise zu neuen Erkenntnissen für die Entzifferung führen kann, aber auch der Vorbereitung der zweiten Stufe der phonemischen Transliteration dient. Jetzt werden die quasi als Container genutzten Transliterationswerte aus dem Zeichenkatalog kontextuell der korrekten sprachlichen Lesung angepasst. So besitzt das Zeichen "CHAN" üblicherweise die Lesung "chan" - "Himmel", syllabische Substitutionen in Nordwest-Yukatan zeigen aber, dass das Zeichen dort in einem vernakularen Kontext "káan" ausgesprochen wurde. Der Einfluss lokaler Maya-Sprachen (relevant sind drei Sprachfamilien) auf die Schriftsprache ist noch nicht systematisch erforscht und mit Ergebnissen der historischen Linguistik abgeglichen worden.

Das Tool zur linguistischen Analyse ermöglicht darüber hinaus die Anlage paralleler, als gleichwertig anzusehende Textanalysen, damit wird den verschiedenen Entzifferungsvorschlägen Rechnung getragen. Durch die Verbindung von Zeichenkatalog, Textkorpus und linguistischer Analyse entsteht letztendlich ein dynami-

scher Text, der je nach Forschungsfrage individuell generiert werden kann. Dieser Ansatz der ontologischen Vernetzung der Komponenten dürfte auch für die Erforschung weiterer nicht entzifferter Schriften von Interesse sein.

Neue Perspektiven für die Maya-Epigraphik

Auch ein digitaler Zeichenkatalog des Klassischen Maya kann nur so gut sein wie die epigraphische Forschung, und ist vor allem vom Grad der kritischen Selbstreflektion abhängig. Konkurrierende Entzifferungsvorschläge werden erst einmal als gleichwertig aufgenommen und erst dann anhand formaler Kriterien kategorisiert. Die Kriterienvergabe folgt den Argumenten der Hypothese und ist damit faktisch. Die Aussagelogiken zur Festlegung der Konfidenzstufen sind dabei eine kritische Zusammenführung fast 70-jähriger epigraphischer Forschungspraxis, auch im Vergleich mit den Methoden bei der Entzifferung anderer nicht-alphabetischer Schriftsysteme. Die Konfidenz eines Entzifferungsvorschlags ist damit weit mehr als ein Wahrscheinlichkeitsbegriff. Das Datenmodell ist dabei dem noch nicht gefestigten Erkenntnisstand zum Mayaschriftsystem angepasst.

Die Arbeit mit digitalen Methoden hat so manche Kritik an der eigenen Forschungstradition erst in Gang kommen lassen, lenkt diese aber auch in eine vorher nicht denkbare Richtung. Die Möglichkeit eines digitalen Zeichenkatalogs und eines digitalen Textkorpus erlaubt erstmals, den gesamten Schriftschatz des Klassischen Maya nach Entzifferungskontexten zu durchsuchen, anstatt sich auf sein “cerebrales” Textkorpus verlassen zu müssen. Erst die Verbindung von geisteswissenschaftlichen und digitalen Methoden erlaubt es, die Maya-Epigraphik in eine neue Phase eintreten zu lassen.

Bibliographie

Houston, Stephen (2001): *The Decipherment of Ancient Maya Writing*. Norman: University of Oklahoma Press: 3–19.

Kelley, David H. (1962): Review of “A Catalog of Maya Hieroglyphs, by J. Eric S. Thompson. Pp. xiv + 458, Including pls 16. University of Oklahoma Press, Norman, in Cooperation with the Carnegie Institution of Washington, 1962.” *American Journal of Archaeology* 66(4): 436–438.

Stuart, David (1987): *Ten Phonetic Syllables*. Research Reports on Ancient Maya Writing 14. Center for Maya Research, Washington, D.C.

Thompson, J. Eric S. (1962): *A Catalog of Maya Hieroglyphs*. Norman: University of Oklahoma Press.

Wichmann, Søren (2006): Mayan Historical Linguistics and Epigraphy: A New Synthesis. *Annual Review of Anthropology* 35, 279–294.

Wahrnehmung und digitale Mustererkennung am Beispiel antiker Terrakottastatuetten

Böttger, Lucie
Lucie.boettger@stud.uni-goettingen.de
Georg-August-Universität Göttingen,
Deutschland - GCDH/Archäologisches Institut

Zeckey, Alexander
alexander.zeckey@stud.uni-goettingen.de
Georg-August-Universität Göttingen,
Deutschland - GCDH/Archäologisches Institut

Langner, Martin
mlangne@gwdg.de
Georg-August-Universität Göttingen,
Deutschland - GCDH/Archäologisches Institut

Sowohl die Informatik als auch die Bild- und Objektwissenschaften nutzen für die Klassifizierung von Objekten und der Bestimmung ihres Ähnlichkeitsgrades Klassifizierungsverfahren und Methoden der Mustererkennung. Während allerdings die Informatik darauf abzielt durch Mustervergleich die Klassifizierung unbekannter Objekte zu automatisieren, dienen Typologien in der Archäologie als Ordnungskriterien für soziokulturelle Fragen, um Informationen über Funktion, Bedeutung oder Produktion zu erfahren.

Die in der Archäologie angewandten qualitativen Analysen basieren auf einem wissenschaftlichen Konstrukt von Klassifikationskriterien [Adams – Adams 2008]. In Fällen, in denen eine große Anzahl von Artefakten eine ganz ähnliche Form hat, sich aber in gewissen Einzelheiten deutlich unterscheidet, wie bei seriell hergestellten Terrakotta-Figuren, die nachträglich überar-

beitet wurden, hat der Begriff der Typologie jedoch seine Grenzen erreicht. In Bezug auf die Wahrnehmung und den Wert der Figuren gibt es zu viele verschiedene Kriterien, die eine Bedeutung tragen. Nur ein statistischer Ansatz, der die Hauptmerkmale in Kombination mit archäologischen Quellen und den intrinsischen ästhetischen Werten (wie Farbe oder Stil) berücksichtigt, kann das Problem lösen.

Die Methoden der 3D-Mustererkennung basieren in der Regel auf Konzepten der kognitiven Psychologie zur Objekterkennung. Die Form eines Objektes wird in geometrische Grundelemente zerlegt und statistisch nach Teilen und Teilsegmentierungen analysiert. Maschinelle Lernalgorithmen helfen, diesen Prozess zu automatisieren. Für die Klassifizierung von Artefakten können diese Methoden jedoch nur grobe Näherungswerte liefern. Auf der Grundlage archäologischer Kategorien kann eine rechnerische Merkmalsextraktion bisher nur manuell durch qualitativen Formvergleich durchgeführt werden. Darüber hinaus könnten in Bereichen, in denen archäologische Methoden keine entsprechenden Typologien schaffen konnten, Methoden der digitalen Formerkennung (shape recognition) hilfreich sein, um geeignete archäologische Kategorien zu definieren.

Deshalb sollen qualitative und quantitative Klassifizierungsmethoden kombiniert werden, um die Typologie von Artefakten zu überarbeiten und somit Vorarbeit zur Erschließung großer technischer bzw. mentaler Bildcorpora zu leisten. Hier werden die Methoden der Informatik (Objekterkennung und Formvergleich) und Archäologie (Typologie und Kopienkritik) von einander profitieren, um die oben genannten Mängel zu überwinden.

Im Vordergrund steht die Frage nach der computergestützten Analysefähigkeit und ihren Grenzen in der Adressierung von Binnenstrukturen sowie der Möglichkeit, neuartige Analyseverfahren zu entwickeln.

Als Materialgrundlage dienen die kleinformatigen, seriell aus Modeln genommenen Tonfiguren. [Burn 2012; Erlich 2015]. Diese bieten sich für Fragen der Klassifizierung zum einen aufgrund ihrer hoch überlieferten Anzahl, zum anderen wegen ihrer großen Formenvarianz an.

Die antiken Terrakotten weisen untereinander verschiedene, vom Archäologen präzis definierbare Grade der Ähnlichkeit auf: So existieren die aus derselben Matrize genommenen Figuren, die eine exakte Übereinstimmung verbindet, die aus denselben Patrizen gewonnenen Figuren, die sich nur in der Größe von ihren ansonst genauen Ebenbildern unterscheiden und die ebenfalls aus derselben Matrize genommenen Figuren, die aber nachträglich noch durch zusätzliche Additionen und Abänderungen ein verändertes Erscheinungsbild aufweisen. Zudem die Terrakotten, die nicht derselben Matrize entstammen, sich aber in Haltung und Drapierung der Tracht untereinander ähneln. Zwar kann man auf der handwerklichen Ebene konstatieren, dass zwei Terrakotten aus derselben Produktion stammen, ist dies aber nicht der Fall, fehlen bislang geeignete Kriterien zur Bestimmung der Ähnlichkeit und ihrer Abstufungen.

Grundlage einer Interpretation ist in der Archäologie die visuelle Beschaffenheit der Artefakte. Ein Großteil dieser Eigenschaften wie Größe, Form, Farbe oder Material lässt sich genau vermessen, somit verbal erfassen und dient vielen Digitalen Corpora der Klassischen Archäologie als Schlagworte zur Einordnung der Artefakte – eine zeit- und resourcenaufwändige Methode.

Ein dem Textmining vergleichbares Objectmining ist in der Klassischen Archäologie bislang noch nicht erprobt. Hier setzt unser Projekt an, das sowohl anwendungsbezogen als auch methodenreflektierend vorgehen möchte, denn es sollen sowohl Verfahren der automatisierten Corpusbildung durch 3D-Mustererkennung entwickelt werden als auch die damit verbundenen Schematisierungen und ihr wissenschaftlicher Nutzen reflektiert werden. In mehreren Schritten werden die Ergebnisse wiederholt evaluiert und die Verfahren feiner kalibriert. Eine systematische Untersuchung formaler Elemente könnte als Schlüssel zur Entwicklung eines Konzepts der Materialisierung von Wissen und Anschauung dienen.

Drei zentrale Leitfragen haben sich bisher herauskristallisiert:

Lassen sich Figurentypen mit digitalen Methoden der Musterkennung nonverbal erfassen und in welcher Exaktheit?

In welchem Umfang ist sprachlogische Begrifflichkeit zur sinnvollen Ausdifferenzierung der Typen notwendig?

Sind die von der archäologischen Stilforschung entwickelten Kategorien zur Beschreibung von Typen auch für digitale Verfahren nutzbar oder müssen an ihre Stelle neue diakritische Verfahren treten?

Ende des 19. Jhs. erstellte Franz Winter einen Katalog antiker Terrakotten, der ihre „Typen“ möglichst vollständig erfassen sollte [Winter 1903]. Seine Materialanordnung kann als frühe Form der archäologischen Mustererkennung gelten, denn er präsentierte die “Typen” in vereinfachten Umzeichnungen. Für eine weitreichende Erforschung der Terrakotten reichte dies jedoch kaum aus, denn die von Winter erstellten „Typen“ sind weder betitelt noch verbal definiert. Unter

diesen somit alleinig durch die Zeichnung definierten „Figurentypen“ subsumierte er vermeintliche Wiederholungen, die keineswegs Typen im Sinne der bereits damals in der Skulpturenforschung etablierten Terminologie und schon gar nicht Terrakotten aus derselben Werkstatt oder gar Form bezeichneten [Heilmeyer 2008; Anguissola 2015]. Zudem wurden Abweichungen vom angezeigten „Typus“ zwar in wenigen Fällen erwähnt, es wurde allerdings nie auf den genauen Ähnlichkeitsgrad der zusammengefassten Terrakotten eingegangen.

Aufgrund dieser evidenten Schwächen wurde kein vergleichbarer Versuch unternommen, die Gesamtheit der antiken Terrakotten in Typen zu unterteilen. Deshalb hat sich die archäologische Forschung den Fundkontexten der Terrakotten [Rotroff 1987; Graepler 1997; Rumscheid 2006] oder semiotischen Ansätzen [Haase 2003] zugewandt. Erst unter Einfluss der Material Culture Studies kehrte die archäologische Forschung zur morphologischen und ästhetischen Wirkung der Figurinen zurück [Bailey 2005; Lesure 2011]. Dieser Ansatz erscheint vielversprechend, vor allem, wenn der methodische Rahmen der Material Culture Studies [Berger 2009; Malafouris – Renfrew 2010; Gerritsen – Riello 2015] mit einer Überprüfung von Konzepten zur Typologie kombiniert wird [Koortbojian 2002; Mattusch 2015].

In diesem Projekt wird der Versuch Winters aufgegriffen eine „Typologie“ der Terrakotten zu erstellen. Um den Typus einer Figur besser klassifizieren zu können, wurde sie in die drei Ebenen Haltung, Gewanddrapierung und Oberflächengestaltung untergliedert. Für diese wurden erste vereinfachte Schemata erstellt.

Eine ökonomische Massendigitalisierung von 3D-Artefakten stellt weiterhin ein ungelöstes Problem dar. Obwohl die semantische Anreicherung von 3D-Daten anspruchsvoll ist, sind Methoden zur Verwendung der Geometrie der 3D-Form für das Data Mining ein aktives Forschungsgebiet [De Luca et al. 2014; Aggarwal 2015]. Verschiedene Verfahren der 3D-Object-Recognition sind seit vielen Jahren bekannt: CAD-Modelle, datengesteuerte geometrische Grundelemente, Oberflächen-Klassifizierung auf Grundlage des Gaußschen Image [Amann 1990; Taylor – Kleeman 2006] und digitaler Bildvergleich [Hueting et al. 2015]. Sie basieren meistens darauf, Grundformen automatisch aus Bereichsdaten zu extrahieren und bekannten Mustern zuzuweisen, um unbekannte Objekte zu klassifizieren. Die Formanalyse wird in der Regel statistisch durchgeführt [Dryden – Mardia 1998]. Statistische Werte, die geometrische Eigenschaften ähnlicher Formen beschreiben, werden mit der Hauptkomponentenanalyse (PCA) [Jolliffe 2002] ausgewertet, um die Formvariabilität zu analysieren. Darüber hinaus sind aber auch partielle Formanpassungsmethoden weit verbreitet [Funkhouser – Shilane 2006] Daneben läuft ein Umrissvergleich eines oder mehrerer Scheiben des 3D-Modells [Tal 2014] und mit bildbasierten 3D-Rekonstruktionsansätzen und formalisierten Grundelementen, um eine Elementbibliothek durch die einfache Deklaration einer Sequenz von Formteilen zu erzeugen [De Luca et al. 2014]. Im Allgemeinen ist es viel einfacher, die Form eines konzentrischen Feststoffs abzurufen als die einer komplexen Struktur. Die verfügbaren Methoden und Technologien bieten keine endgültige Lösung für diese. So entstehen Forschungsfragen in der inhaltsbasierten 3D-Objekt-Retrieval-Adressenabfrage und -klassifizierung auf texturierten 3D-Modellen, Bereichs-Scans-basierte 3D-Formwiedergewinnung, Formabfrage auf nicht starren 3D-Wasserdicht-Meshes, umfangreiches 3D-Formular-Retrieval und 3D-Objekt-Retrieval mit multimodalen Ansichten. Diese verschiedenen algorithmenbasierten Ansätze klassifizieren 3D-Modelle nur in Form von Grundinstanzen (wie Frau, Hund, Becher usw.).

Eine schnelle Teilzerlegung in einfache geometrische Formen ist daher unzureichend. Vielmehr müssen geeignete Mustererkennungsverfahren entwickelt werden, die den Grad der Simplifizierung und Abstraktion an menschliche recognition and dissemination patterns knüpft und so die Klassifikation unbekannter Objekte schrittweise evaluiert und kalibriert.

Zur automatisierten Erfassung von Artefakten kamen diese Methoden bislang kaum zum Einsatz, obwohl Experimente mit Kurvenerkennung und Entlastungserkennung bereits mit archäologischen Artefakten durchgeführt wurden [Tal 2014]. Die Ursache liegt zum einen darin, dass keine hinreichende Zahl an Bildwerken als 3D-Modell vorliegt, um diese Verfahren in signifikantem Ausmaß auf ihre Anwendbarkeit zu überprüfen. Zum anderen stellen Kunstwerke (anders als z.B. Bauteile) wegen ihrer hohen Variabilität eine große Herausforderung an jede computergestützte Klassifikation dar. Die Zuordnung einer spezifischen Instanz zu einer allgemeineren Klasse fällt hier weitaus schwerer, da sie sich untereinander in ihrer Form, Größe und Farbe erheblich unterscheiden können.

Daher wurde ein einfacher rechnerischer Formvergleich für "best fit" von Archäologen verwendet, um die Ähnlichkeit von zwei Artefakten zu analysieren [z.B. Beenhouwer 2008]. Best-Fit-Prozesse sind in Engineering und ähnlichen Branchen etabliert und es gibt zahlreiche Software-Lösungen. Diese Tests sind eher qualitativ als quantitativ und wurden bereits für die

Toleranz basierte Pass/Fail Shape Analysis antiker Skulptur verwendet [z.B. Www.digitalsculpture.org/laocoon/index.html; Lu et al. 2013; Frischer 2014].

Infolgedessen reicht es nicht aus, die Modelle in einfache geometrische Formen zu zerlegen. Ein vielversprechender Ansatz für Formerkennungsverfahren wird derzeit anhand der neu erstellten Terrakotta-Schemata unter Verwendung verschiedener Verfahren der shape comparison im Bereich 2/3D und machine learning entwickelt und evaluiert. Ziel ist es den Grad der Vereinfachung und Abstraktion nicht nur mit den menschlichen Erkennungs- und Verbreitungsmustern zu verknüpfen, um die unbekannten Objekte inkrementell zu bewerten und zu klassifizieren, sondern auch die in der Archäologie und Kunstgeschichte entwickelten Kategorisierungen zu verwenden. Die 3D-Mustererkennung der Hauptkomponenten (Form, Größe und Farbe) muss daher mit archäologischer Subkategorisierung und geeigneten Formen des maschinellen Lernens einhergehen [Bishop 2006].

Bibliographie

W. Adams – E. Adams, Archaeological typology and practical reality. A dialectic approach to artifact classification and sorting (Cambridge 2008)

C. Aggarwal, Data Mining: The Textbook. Heidelberg: Springer comprehensively discusses a wide variety of methods (Cham 2015)

H. Amann, 3D object recognition based on surface representations (Neuchâtel 1990)

A. Anguissola, Styles and genres. "Idealplastik" and the relationship between Greek and Roman sculpture. In: E. A. Friedland, M. Grunow Sobocinski, and E. K. Gazda (Hrsg.), The Oxford handbook of Roman sculpture (Oxford 2015) 240-259

D. Bailey, Prehistoric Figurines: Representation and Corporeality in the Neolithic (London 2005) Routledge

J. de Beenhouwer, Datamanagement for moulded ceramics and digital image comparison. A case study of Roman terra cotta figurines. In: Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2-6, 2007 (Bonn 2008) 160- 163

A. Berger, What objects mean. An introduction to material culture (Walnut Creek 2009)

C. Bishop, Pattern Recognition and Machine Learning (Berlin 2006)

L. Burn, Terracottas. In: T. J. Smith and D. Plantzos (Hrsg.). A companion to Greek art (Malden, Mass. 2012) 221-234

L. De Luca, D. Lo Buglio, Geometry vs Semantics. Open Issues on 3D Reconstruction of Architectural Elements, In: Ioannides M. and E. Quak (Hrsg.), 3D Research Challenges in Cultural Heritage. A Roadmap in Digital Heritage Preservation (Heidelberg / New York 2014) 36-49

I. Dryden – K. Mardia, Statistical Shape Analysis (New York 1998)

A. Erlich, Terracottas. In: E. A. Friedland, M. G. Sobocinski with E. K. Gazda (Hrsg.), The Oxford handbook of Roman sculpture (Oxford 2015) 155-172

B. Frischer, 3D data capture, restoration and online publication of sculpture. In: F. Remondino and S. Campana (Hrsg.). 3D Recording and Modelling in Archaeology and Cultural Heritage (Oxford 2014) 137–144

T. Funkhouser – P. Shilane. 2006. Partial matching of 3d shapes with priority-driven search, In: Proceedings of the fourth Eurographics symposium on Geometry processing, ser. SGP '06 (Aire-la-Ville 2006) 131–142

A. Gerritsen – G. Riello (Hrsg.), Writing material culture history (London 2015)

D. Graepler, Tonfiguren im Grab : Fundkontexte hellenistischer Terrakotten aus der Nekropole von Tarent (München 1997)

M. Haase, Votivbilder als Werbemedien? Votivterrakotten aus Gravisca als Zeichenträger in Prozessen symbolischer Interaktion. In: U. Veit, T. L. Kienlin, C. Kümmel et al. (Hrsg.). Spuren und Botschaften. Interpretationen materieller Kultur (New York 2003) 369-383

W. Heilmeyer, Kunst und Serie. In: K. Junker, A. Stähli and C. Kunze (Hrsg.), Original und Kopie. Formen und Konzepte der Nachahmung in der antiken Kunst. Akten des Kolloquiums in Berlin, 17. - 19. Februar 2005 (Wiesbaden 2008) 243-251

M. Hueting – M. Ovsjanikov – N. J. Mitra, CrossLink. joint understanding of image and 3D model collections through shape and camera pose variations. ACM Transactions on Graphics (TOG) 34(6), (New York 2015) 233

I. Jolliffe, Principal Component Analysis, 2nd ed. (Heidelberg 2002)

M. Koortbojian, Forms of attention. Four notes on replication and variation. In: E. K. Gazda (Hrsg.), The ancient art of emulation. Studies in artistic originality and tradition from the present to classical antiquity (Ann Arbor 2002) 173-204

R. Lesure, Interpreting Ancient Figurines: Context, Comparison, and Prehistoric Art (Cambridge 2011)

M. Lu – Y. Zhang – B. Zheng et al. 2013. Portrait sculptures of Augustus: Categorization via local shape comparison. In: 2013 Digital Heritage International Congress (Marseille 2013)

L. Malafouris – C. Renfrew (Hrsg.), The cognitive life of things. Recasting the boundaries of the mind (Cambridge 2010)

C. Mattusch, 2015. Repeated images. Beauty with economy. In: J. Daehner – K. Lapatin (Hrsg.), Power and pathos. Bronze sculpture of the hellenistic world. Exhibition from March 14 to June 21, 2015 (Los Angeles 2015) 111-125

S. Rotroff, Three centuries of hellenistic terracottas. Preface. A chronological commentary on the contexts. In: H. A. Thompson and D. B. Thompson (Hrsg.). Hellenistic pottery and terracottas (Princeton 1987) 183-194

F. Rumscheid, Die hellenistischen Wohnhäuser von Priene. Befunde, Funde und Raumfunktionen, in: Annete Haug – Dirk Steuernagel (Hrsg.), Hellenistische Häuser und ihre Funktionen. Internationale Tagung Kiel, 4. bis 6. April 2013 (Bonn 2014) 143–160

A. Tal, 3D Shape Analysis for Archaeology. In: Ioannides M. and E. Quak (Hrsg.). 3D Research Challenges in Cultural Heritage. A Roadmap in Digital Heritage Preservation (Heidelberg / New York 2014) 50-63

G. Taylor – L. Kleeman, Visual Perception and Robotic Manipulation: 3D Object Recognition, Tracking and Hand-Eye Coordination (Heidelberg 2006)

F. Winter, Die Typen der figürlichen Terrakotten, Die antiken Terrakotten III (Berlin 1903) (http://digi.ub.uni-heidelberg.de/diglit/winter1903)

Was Lesende denken: Assoziationen zu Büchern in Sozialen Medien

Beck, Jens
jens_beck@gmx.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Willand, Marcus
marcus.willand@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Reiter, Nils
Nils.Reiter@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Einleitung

Im vorliegenden Abstract stellen wir eine Methode sowie erste Ergebnisse der Analyse von Entitäten-Assoziationen realer Leserinnen und Leser vor.

Literaturwissenschaftliche Rezeptions-, Lese- und Lesertheorien gehen seit ihren hermeneutischen und wirkungsästhetischen Anfängen (Schleiermacher 1838, insb. 309f.; Iser 1976) von professionellen (Dijkstra 1994), informierten (Fish 1970, 86), Modell- (Eco 1979) oder sogar idealen (Schmid 2005) Lesern aus (vgl. Willand, 2014). Diesen wird die Kompetenz zugeschrieben, idealerweise sämtliche Textmerkmale referentialisieren zu können, wobei je nach literaturtheoretischer Provenienz unterschiedliche Kontexte die Grundlage der Zuschreibungen an den Text bilden. Dazu gehören u.a. Informationen über den Autor oder über die sozialhistorischen Bedingungen der Textproduktion, über die Rezeptionsbedingungen, über Vorgänger- oder zeitgenössische Texte oder über Wissen aus dem Bereich der Literaturwissenschaftlerin bzw. des Lesers selbst.

An bestimmte Wissensbestände dieser *realen* Leserinnen und Leser literarischer Texte können wir uns durch eine computergestützte empirische Analyse von Rezeptionszeugnissen aus sozialen Medien annähern. Konkret ist unser Ziel die Rekonstruktion und Analyse der von literarischen Texten ausgelösten Assoziationen. Dabei beschränken wir uns auf die Assoziationen, die reale oder fiktive Entitäten betreffen, also etwa Personen des öffentlichen Lebens oder Figuren aus fiktionalen Werken.

Die Plattform Goodreads bietet Leserinnen und Lesern die Möglichkeit des freien schriftlichen Austauschs über literarische Texte in einer großen Community. 55 Mio. Mitglieder haben bis 2017 über 50 Mio. Reviews geschrieben, wobei die Besprechungen die Inhalte der Bücher selbst und nicht - wie etwa bei Verkaufsplattformen wie Amazon - die Distribution, den Preis o.ä. fokussieren (Piper et al. 2015).

Verarbeitung

Als Grundlage unserer Analysen wurden die Reviews in einer lokalen Datenbank gespeichert.

Die Datenbank enthält 1,3 Millionen englischsprachige Reviews zu 5.481 besprochenen Texten. Die Reviews umfassen insgesamt etwa 150 Mio. Tokens, d.h. uns steht eine große Datenmenge zur Extraktion der Entitäten zur Verfügung. In einem ersten Schritten wurden die Reviews bereinigt: HTML-Tags wurden entfernt und Wiederholun-

gen von mehr als dreimal dem gleichen Zeichen oder Wort auf drei reduziert.

Zur Extraktion der Entitäten aus den Reviews haben wir den Stanford Named Entity Recognizer (Finkel et al., 2005) verwendet. Der Tagger klassifiziert die gefundenen Entitäten in mehrere Klassen. Für uns ist die Klasse „PERSON“ relevant, da diese alle gefundenen Entitäten von Personen enthält.

Im nächsten Schritt disambiguieren wir die extrahierten Entitäten, da z.B. ein Name wie “Harry” auf viele mögliche Träger des Namens verweisen kann. Mit Hilfe von UKB (Agirre et al., 2009) und UKB-wiki (Agirre et al., 2015) können den Entitäten Wikipedia-Seiten zugeordnet werden, welche die möglichen Entitäten repräsentieren. Für diese Disambiguierung verwendet UKB den PageRank-Algorithmus (Page et al. 1999), der Dokumente nach ihrem Verlinkungsgrad bewertet. Sobald Namen wie „Ron“ und „Dumbledore“ im selben Kontext erwähnt werden, wird die Wahrscheinlichkeit größer, dass mit “Harry” *Harry Potter*, mit “Ron” *Ron Weasly* und mit “Dumbledore” *Albus Dumbledore* aus der Romanreihe *Harry Potter* gemeint sind, weil diese Entitäten aus dem selben Kontext kommen und dies in der Wissensbasis Wikipedia durch Verlinkungen explizit ablesbar und quantifizierbar ist.

UKB-wiki stellt einen herunterladbaren Graphen zur Verfügung, der Wikipedia-Seiten und Links auf andere Wikipedia-Seiten repräsentiert. In einem mitgelieferten Wörterbuch sind Entitäten mit allen möglichen Entitäten (Verweise auf Wikipedia Seiten) aufgeführt.

Die auf diese Weise gewonnen Wikipedia-Einträge wurden anschließend hinsichtlich des ontologischen Status der referenzialisierten Entität kategorisiert, also ob es sich um eine reale Person oder fiktionale Figur handelt. Dazu wurde die Wissensbasis DBpedia [1] verwendet, die die Daten aus Wikipedia strukturiert und maschinenlesbar kodiert. Da die Disambiguierung Wikipedia-Einträge liefert, können wir anhand dieser die auf den zugehörigen DBpedia-Eintrag zugreifen. Über DBpedia lassen sich neben ontologischen Kategorien auch andere Eigenschaften extrahieren, die für eine Analyse ggf. interessant sind, etwa Geschlecht oder Relationen zu anderen Figuren.

Die extrahierten Daten werden zunächst als Tabelle gespeichert und erlauben somit eine flexible weitergehende Verarbeitung, etwa in einem Netzwerk. Eine Zeile der Tabelle besteht aus dem Werktitel, der disambiguierten Entität (Verweis auf Wikipedia Seite), einer Liste der extrahierten Entitäten aus den Reviews, einer Liste von Review-IDs, um nachvollziehen zu können in welchen Reviews der Name erwähnt wird, der Anzahl der Erwähnungen und der Angabe ob es sich um eine Figur handelt oder nicht.

Titel	Disambiguierte Entität (Verweis auf Wikipedia Seite)	Extrahierte Entität	Review IDs	Anzahl der Erwähnungen	Handelt es sich um eine fiktionale Figur?
The Hound of the Baskervilles	Agatha_Christie	christie, agatha_christie, agatha_christy	470784120, ..., 188608568	20	False
The Hound of the Baskervilles	Spock	spock	42971473	1	True
The Hound of the Baskervilles	Robert_Downey,_Jr.	robert_downey_jr, robert_downey	1077543609, ..., 1250976986	18	False
The Hound of the Baskervilles	Ann_Radcliffe	ann_radcliffe	435380655	1	False

Tabelle 1: Auszug aus den extrahierten Daten. Die extrahierten Entitäten stammen aus den Reviews zu *The Hound of the Baskervilles*.

Zwischenergebnisse

Um ein exemplarisches Resultat zu präsentieren, haben wir Reviews zu “The Hound of the Baskervilles” (deutsch: “Der Hund von Baskerville”) analysiert. Unter den häufig erwähnten Entitäten finden sich erwartungsgemäß Sherlock Holmes, Dr. Watson, sowie der Autor Arthur Conan Doyle. Weitere häufig erwähnte Figuren aus der fiktionalen Welt des Sherlock Holmes' sind James Mortimer und Charles Baskerville. Aber auch Professor Moriarty wird häufig erwähnt, obwohl er in diesem Buch der Sherlock-Reihe gar nicht auftaucht. Das System erzeugt jedoch auch Fehler. Beispielsweise wird der Antagonist Stapleton zwar sehr oft erwähnt, da zu ihm aber kein eigener Wikipedia-Eintrag existiert, wird er fälschlicherweise mit dem Footballspieler Frank Stapleton verknüpft. Henry Baskerville, der Sohn von Charles und Erbe des Anwesens, wird im Buch

fast durchgehend als Sir Henry bezeichnet, und kommt mit diesem Namen ebenfalls häufig in den Reviews vor. Da auch für ihn kein eigener Wikipedia-Eintrag existiert und der Name Henry extrem mehrdeutig ist, werden eine Reihe klar falscher Entitäten verknüpft: Henry II. von Frankreich; Henry County (Alabama); oder Henry I. von England.

Bemerkenswert sind insbesondere jedoch die referenzialisierten extra-textuellen Entitäten, also diejenigen, die nicht aus der fiktionalen Welt Sherlocks stammen. Es finden sich etwa *Hercule Poirot* und *Agatha Christie* unter den erwähnten Entitäten, was als klares Zeichen dafür gesehen werden kann, dass die Leserinnen und Leser den Text vor dem Hintergrund eines starken Gattungsbewusstseins rezipieren. Dafür spricht auch, dass mit *Benedict Cumberbatch*, *Robert Downey, Jr.*, *John Barrymore* und *Jeremy Brett* gerade die Schauspieler unter den assoziierten Referenzen vertreten sind, die in einer der vielen Verfilmungen die Rolle des Sherlock Holmes verkörpert haben.

Fehleranalyse

Das am häufigsten auftretende Problem ist das Fehlen eines Wikipedia-Eintrages für eine Figur. In der englischsprachigen Wikipedia sind fiktionale Figuren zwar nicht per se davon ausgeschlossen -- Richtschnur hier ist deren “Notability”. Viele Figuren sind jedoch nur auf den Einträgen des entsprechenden Werks erwähnt. Da der Algorithmus nicht in der Lage ist, *keinen* Eintrag zu liefern, wird in solchen Fällen eben ein anderer Eintrag verwendet, auch wenn dieser relativ weit entfernt sein mag. Eine technische Lösung wäre sicher, nur ab einem gewissen Schwellwert eine Disambiguierung vorzunehmen, und die nicht-disambiguierten Einträge zumindest als solche erkennen zu lassen. Eine andere Möglichkeit läge in der (zusätzlichen) Verwendung von Literaturlexika, die (womöglich) eine größere Abdeckung zu fiktionalen Figuren aufweisen. Beide Optionen werden wir in zukünftigen Arbeiten genauer untersuchen.

Da es sich bei den Reviews letztlich um Inhalte aus einem sozialen Medium handelt, kommt es auch vor, dass Namen falsch geschrieben werden oder gar der gesamte Text schriftsprachliche Konventionen übergeht. Prima vista sind diese Fälle im Vergleich zu Buchrezensionen zwar häufig anzutreffen, wir können das Problem aber umgehen, indem wir nur diejenigen Erwähnungen berücksichtigen, die mehr als einmal vorkommen. Festzuhalten bleibt aber ebenfalls, dass die Texte im Vergleich zu z.B. Twitter-Daten deutlich sauberer sind.

Eine weitere mögliche (jedoch noch nicht tatsächlich beobachtete) Fehlerquelle liegt in der Natur des PageRank-Algorithmus: Wenn eine Figur in einem Werk existiert, ein Leser oder eine Leserin jedoch explizit z.B. eine Person des öffentlichen Lebens mit dem gleichen Namen erwähnt, wird der Algorithmus diese Erwähnung eher der Figur zuschlagen, da diese dichter mit anderen Figuren verknüpft ist.

Auswertung als Netzwerk

Die oben extrahierten Daten erlauben Auswertungen auf vielfältige Weise. Exemplarisch konzentrieren wir uns hier auf eine Form, in der von Lesern zugeschriebene Gemeinsamkeiten zwischen literarischen Texten untersucht werden. Die Texte und die ihnen zugeschriebenen Assoziationen werden dabei als Knoten in einem Netzwerk repräsentiert. Ein Text ist also mit allen ihm zugeschriebenen Assoziationen verbunden, wobei das Gewicht der Kante die Anzahl der Reviews angibt, in denen eine bestimmte Assoziation auftaucht.

Durch diesen Aufbau ergeben sich Kerneigenschaften des Netzwerkes, die bei der Analyse zu beachten sind: Ein Teil der erwähnten Entitäten sind *intra*textuelle Referenzen, d.h. Figuren aus dem jeweiligen Text selbst (Veldhues, 1995). Auch wenn diese keine *inter*textuellen Assoziationen und damit nur sekundäres Extraktionsziel sind, behandeln wir sie als gleichwertige Assoziationen [2].

Figuren, die in mehr als einem Werk auftauchen (z.B. *Sherlock Holmes* oder *Harry Potter*) bilden eine hoch gewichtete Verbindung zwischen den Texten einer literarischen Reihe, wobei Reihen durch die von ihnen geteilte fiktionale Welt markiert sind. Als gemeinsamer Assoziationsraum sind sie aufgrund der hohen Gewichtung auch angemessen im Netzwerk repräsentiert.

Durch die gemeinsame Darstellung der Werke und assoziierten Entitäten ergeben sich -- bei Auswahl eines geeigneten Layout-Algorithmus z.B. in Gephi [3] -- eng zusammenhängende Gruppen von Werken. Das hier exemplarisch angeführte Resultat eines engen Zusammenhangs repräsentiert jedoch nicht bestimmte Texteigenschaften selbst, sondern lediglich von Leserinnen und Lesern gemeinsam gemachte Zuschreibungen an diese Texte.

Das hier beschriebene Netzwerk wird im Zuge der Konferenz frei zugänglich gemacht.

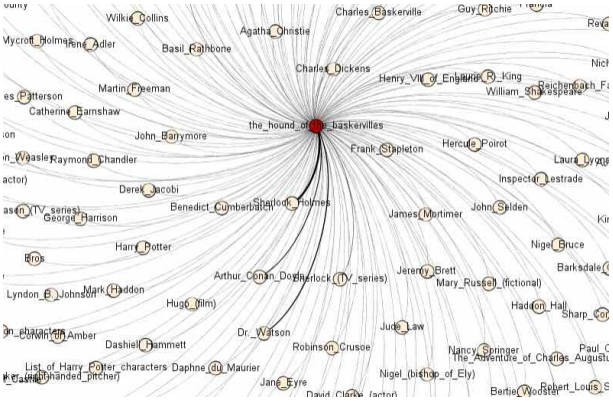


Abbildung 1: Assoziationen zu Conan Doyles The Hound of the Baskervilles, extrahiert aus den Reviews von Benutzern. Die Abbildung zeigt zur Illustration sämtliche assoziierte Entitäten, unabhängig von der Häufigkeit.

Nächste Schritte

Durch den Zugriff auf bisher undenkbar große Rezeptionsdatenmengen erhält die empirische Leseforschung einen sie fundamental erweiternden Impetus, war sie methodisch betrachtet bisher überwiegend auf Fragebögen [4] und peripheriephysiologische Messungen angewiesen, jüngst gestützt durch bildgebende Verfahren. Computerlinguistische Methoden der Sprach- und Korpusverarbeitung versprechen nicht nur die Analyse unlesbarer Mengen an Rezeptionszeugnissen, sondern auch eine Modellierung leseratttribuierter Kontexte literarischer Texte und somit einen ersten Einblick in die bisher unbeantwortete Frage, mit welchem Vorwissen echte Leser eigentlich lesen.

In diesem Sinne präsentiert das eingereichte Paper erste, jedoch bereits substanzielle Ergebnisse.

Die nächsten Schritte leiten sich direkt aus der oben diskutierten Fehleranalyse ab. Zum einen soll die Wissensbasis um fiktionale Figuren aus den Werken erweitert werden (was z.B. über *named entity recognition* über den Volltexten machbar wäre). Zum anderen soll der Algorithmus in die Lage versetzt werden bestimmte (fehlerhafte) Zuweisungen zurückzuweisen, etwa mit einem zu definierenden *threshold*.

Fußnoten

1. http://wiki.dbpedia.org/
2. Das Filtern von innertextuellen Figuren ist technisch möglich (Beck, 2017), aber zeitaufwändig und für die hier vorgestellte Nutzung als Explorationswerkzeug letztlich unnötig.
3. https://gephi.org
4. Groeben 1979; Baurmann 1981; Funke 2003; Christmann u. Schreier 2003; Wübben 2009 u.v.m.

Bibliographie

Agirre, Eneko / Soroa, Aitor (2009): “Personalizing PageRank for Word Sense Disambiguation”, in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece.

Agirre, Eneko / Barrena, Ander / Soroa, Aitor (2015): Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. http://arxiv.org/abs/1503.01655

Beck, Jens (2017): How do People Read Literature? - Detection and Identification of Names in Book Reviews. Bachelor’s thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Baurmann, Jürgen (1981). „Textrezeption empirisch. Wege zu einem ziel, behelfsbrücken oder holzwege?". Rezeptionspragmatik. Beiträge zur Praxis des Lesens. Uni-Taschenbücher. Band 1026. Hrsg. v. Gerhard Köpf, 201–218. München.

Christmann, Ursula / Margrit Schreier (2003). „Kognitionspsychologie der Textverarbeitung und Konsequenzen für die Bedeutungskonstitution literarischer Texte". Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte. Revisionen. Hrsg. v. Fotis Jannidis, Gerhard Lauer, Matías Martínez & Simone Winko, 246–284. Berlin.

Dijkstra, Katinka (1994): Leseentscheidung und Lektürewahl. Empirische Untersuchungen über Einflussfaktoren auf das Leseverhalten. Berlin.

Dimitrov, Stefan / Zamal, Faiyaz / Piper, Andrew / Ruths, Derek (2015): “Goodreads vs Amazon: The Effect Of Decoupling Book Reviewing And Book Selling", in: International Conference on Web and Social Media (ICWSM-14).

Eco, Umberto (1979): The Role of the Reader. Explorations in the Semiotics of Texts. Bloomington, IN.

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),* pp. 363-370.

Fish, Stanley E. (1970): „Literature in the Reader: Affective Stylistics“, in: *New Literary History* 1(2): 123–162.

Funke, Mandy (2003). „Das Abenteuer der Fragebögen. Aspekte zur empirischen Wirkungsforschung in der DDR". Wissenschaft und Systemver-

änderung. Rezeptionsforschung in Ost und West – Eine konvergente Entwicklung? Euphorion. Band 44. Hrsg. v. Wolfgang Adam, Holger Dainat & Gunther Schandera, 119–126. Heidelberg.

Groeben, Norbert (1979). „Zur Relevanz empirischer Konkretisationserhebungen für die Literaturwissenschaft". Empirie in Literatur- und Kunstwissenschaft. Grundfragen der Literaturwissenschaft. Hrsg. v. Siegfried J. Schmidt, 43–82. München.

Iser, Wolfgang (1976). Der Akt des Lesens. Theorie ästhetischer Wirkung. Band 636. München.

Page, Lawrence / Brin, Sergey / Motwani, Rajeev / Winograd, Terry (1999): "The PageRank Citation Ranking: Bringing Order to the Web" , technical Report. Stanford InfoLab.

Schleiermacher, Friedrich (1838): Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament. Aus Schleiermachers handschriftlichem Nachlasse und nachgeschriebenen Vorlesungen herausgegeben von Friedrich Lücke. In: Friedrich Schleiermacher's sämmtliche Werke. Berlin: Reimer.

Schmid, Wolf (2005): Elemente der Narratologie. Narratologia. Band 8. Berlin.

Veldhues, Christoph (1995): "Gleich- und Gegenüberstellung".Intratextuelle und intertextuelle Bedeutung in der Literatur. Zeitschrift für französische Sprache und Literatur 40/3 (1995), 243-267.

Willand, Marcus (2014): Lesermodelle und Lesertheorien. Historische und systematische Perspektiven. Narratologia. Band 41. Berlin.

Wübben, Yvonne (2009). „Lesen als Mentalisieren? Neuere kognitionswissenschaftliche Ansätze in der Leseforschung". Literatur und Kognition. Bestandsaufnahmen und Perspektiven eines Arbeitsfeldes. Poetogenesis. Band 6. Hrsg. v. Martin Huber & Simone Winko, 29–44. Paderborn.

Wenn der Funke überspringt – Word Embeddings im Dienst der Wissenschaftsgeschichte

Hellrich, Johannes
johannes.hellrich@uni-jena.de
Graduiertenkolleg „Modell Romantik", Friedrich-Schiller-Universität Jena, Jena, Deutschland; Jena University Language & Information Engineering Lab (JULIE Lab), Friedrich-Schiller-Universität Jena, Jena, Deutschland

Stöger, Alexander
alexander.stoeger@uni-jena.de
Graduiertenkolleg „Modell Romantik", Friedrich-Schiller-Universität Jena, Jena, Deutschland

Hahn, Udo
udo.hahn@uni-jena.de
Jena University Language & Information Engineering Lab (JULIE Lab), Friedrich-Schiller-Universität Jena, Jena, Deutschland

Einleitung

Das moderne Verständnis von Elektrizität fußt auf wissenschaftlichen Entdeckungen des 17. und 18. Jahrhunderts, die die technische Nutzbarmachung ab dem 19. Jahrhundert ermöglichten. Unsere Studie wendet state-of-the-art Methoden der distributionellen diachronen Semantik an, um dies anhand des damit einhergehenden Wandels der Wortsemantik von *Elektrizität*, *electricity*, *elektrisch*, *electrical*, *Funken* und *spark* nachzuvollziehen.[1] Im Fokus liegt die Entwicklung der menschlichen Wahrnehmung eines Phänomens, das zuerst als nur schwer begreifliche Naturerscheinung galt, bevor es durch wissenschaftliche Experimente zu einer erklärten Kraft und einem unverzichtbaren Teil des menschlichen Alltags wurde. Wir sind von anderen Studien zur Entwicklung wissenschaftlicher Konzepte (Hall u.a. 2008; Mimno 2012; Fankhauer u.a. 2016; Schumann 2016) inspiriert, die die Entwicklung von Wortgruppen beschreiben (vgl. etwa Topic Modelling nach Blei u.a. (2003)). Dagegen untersuchen wir einzelne Wörter, um differenzierte Aussagen zur Entwicklung des Konzepts Elektrizität

machen zu können, das insbesondere in seinem frühen Untersuchungsstadium noch stark fluktuiert.

Methoden

Die von uns verwendeten Verfahren der distributionellen Semantik basieren auf der strukturalistischen Annahme, dass die Semantik von Wörtern über die typischerweise in ihrer Nähe stehenden Wörter approximiert werden kann, prägnant zusammengefasst: “You shall know a word by the company it keeps!” (Firth 1957: 11). Zur Durchführung unserer Experimente nutzen wir JeSemE[2] (Hellrich u. Hahn 2017b). JeSemE basiert auf Word Embeddings, einer state-of-the-art distributionellen Methode zur Messung von Wortähnlichkeit. Word Embeddings sind niedrigdimensionale[3] Vektorrepräsentationen, die die Semantik von Wörtern anhand aller Kontexte repräsentieren, in denen Wörter in einem Korpus beobachtet wurden. Somit benötigen sie keinerlei manuell erstellte Wissensbasis (Wörterbücher oder Ontologien), sondern nur Korpora. Der populärste Algorithmus zur Erzeugung von Word Embeddings ist word2vec (Mikolov u.a. 2013), jedoch kommt in JeSemE stattdessen SVD $_{\text{PPMI}}$ (Levy u.a. 2015) zur Anwendung. Grund für diese Entscheidung ist dessen hohe Reliabilität, wohingegen mit word2vec durchgeführte Experimente nur bedingt wiederholbar sind (Hellrich u. Hahn 2017a). Durch Word Embeddings, die mit Texten aufeinanderfolgender Zeitabschnitte trainiert wurden, kann die diachrone semantische Entwicklung von Wörtern nachvollzogen werden, indem man die zum jeweiligen Zeitpunkt ähnlichsten Wörter ermittelt (Kim u.a. 2014; Kulkarni u.a. 2015; Hamilton u.a. 2016; Hellrich u. Hahn 2016; Jo 2016). Außerdem kann JeSemE anhand der statistischen Maße PPMI (Bullinaria u.a. 2007) und χ^2 (vgl. etwa Manning u. Schütze (1999)) die spezifischsten Kontextwörter eines Worts identifizieren, also Wörter, die häufiger in seiner Nähe stehen, als auf Grund zufälliger Prozesse zu erwarten wäre. Letzteres dient als Plausibilitätscheck für die mittels Word Embeddings ermittelte Ähnlichkeit.

Korpora

Die beiden von uns untersuchten Korpora sind das Kernkorpus des Deutschen Textarchivs (DTA; Geyken (2013)), das deutschsprachige Sachtexte und literarische Texte des 17. bis 19. Jahrhunderts sammelt, und das Royal Society Corpus (RSC; Kermes u.a. (2016)), das die ersten beiden Jahrhunderte der „Philosophical Transactions of the Royal Society of London“ (1665–1869, älteste wissenschaftliche Zeitschrift in englischer Sprache) beinhaltet. Tabelle 1 zeigt die Kerndaten der von JeSemE prozessierten Korpora, dabei wurden die Korpora in Zeitabschnitte mit ähnlichen Textmengen unterteilt.[4] Beide Korpora sind mit Informationen zur orthographischen Normalisierung und Lemmatisierung annotiert, die von uns genutzt wurde.

Korpus	Zeitraum	Zeitabschnitte	Wörter im Korpus	Modellierte Wörter
DTA	1751–1900	jeweils 30 Jahre	$79{,}6 \times 10^6$	5.388
RSC	1750–1869	50, 50 und 29 Jahre	$24{,}7 \times 10^6$	3.080

Tabelle 1: Kerndaten der verwendeten Korpora.

Ergebnisse

Ausgehend vom wissenschaftshistorischen Forschungsstand ist mit einer semantischen Entwicklung von *Elektrizität* und *electricity* zu rechnen, die weg von beobachteten Naturphänomenen und hin zu kontrollierten Experimenten und systematisch erfassten Grundsätzen führt (Home 2016: 368–71). Wir testeten diese Hypothese, indem wir mittels JeSemE die Ähnlichkeit der beiden Wörter zu anderen bestimmten, die im historischen naturwissenschaftlichen Diskurs relevant waren.

Die Abbildungen[5] 1 und 2 zeigen diese Entwicklung für *electricity* im Englischen und *Elektrizität* im Deutschen. Die Ergebnisse entsprechen der Hypothese, *electricity* wird den mit Experimentalaufbauten verbundenen Wörtern *spark* [‚Funke‘] und *conductor* [‚Leiter‘] im Lauf der Zeit immer ähnlicher, während die Ähnlichkeit zu *lightning* [‚Blitz‘] abnimmt. Analoge Entwicklungen sind auch für *Elektrizität* im DTA erkennbar, besonders deutlich für die Ähnlichkeit zu *Blitz*. Dem zunehmenden theoretischen Verständnis entspricht die steigende Ähnlichkeit zu *magnetism* [‚Magnetismus‘] beziehungsweise *Magnet* — die Entdeckung des Elektromagnetismus im 19. Jahrhundert führte zu einer Vielzahl wissenschaftlicher Publikationen durch Forscher wie Michael Faraday, einem Mitglied der Royal Society.

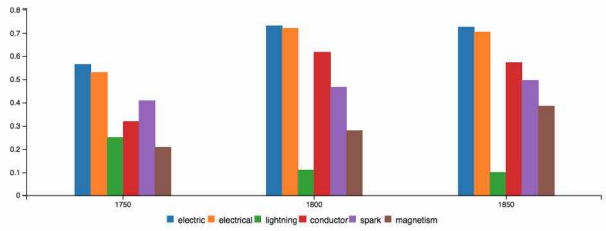


Abbildung 1: Ähnlichkeit ausgewählter Wörter zu ‚electricity' im RSC.

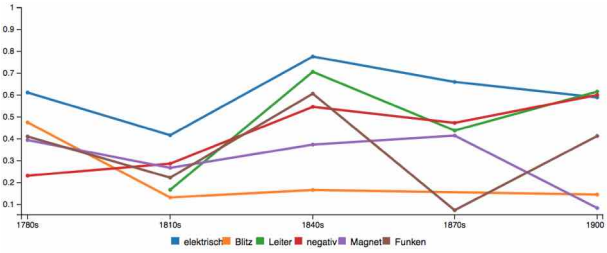


Abbildung 2: Ähnlichkeit ausgewählter Wörter zu ‚Elektrizität' im DTA.

Die Analyse spezifischer Kontextwörter (mittels normalisiertem x^2) weist auf einen engen Zusammenhang mit elektrischen Ladungen in DTA und RSC hin, wie in Abbildungen 3 für das RSC gezeigt. Dabei werden *vitreous* [‚gläsern'] und *resinous* [‚harzig'] in den historischen Texten entsprechend ihrer elektrischen Eigenschaften synonym zu *negative* und *positive* verwendet (siehe etwa Lichtenberg u. Erxleben (1787: 434)).

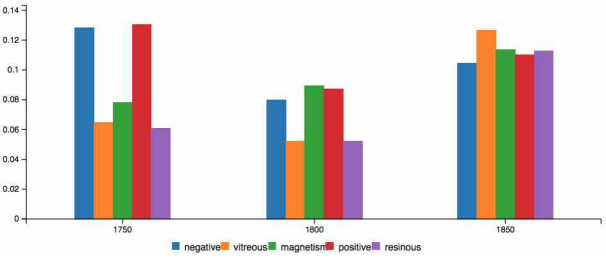


Abbildung 3: Spezifische Kontextwörter für ‚electricity' im RSC.

Ein weiterer klarer Hinweis auf das steigende Verständnis elektrischer Phänomene ist die Entwicklung der Adjektive *electrical* und *elektrisch.* Für diese ist ein starker Rückgang der Spezifität von *Materie,* respektive *matter*, erkennbar, wie in Abbildung 4 für das DTA gezeigt. Dies entspricht der bis ins 19. Jahrhundert verbreiteten Idee einer *electrical matter* [‚Elektrisches Fluidum'], einer abstrakten und unspezifizierten Kraft, die auch als Erklärung für Lebensvorgänge verwendet wurde (Steigerwald 2013). Auffällig ist die unterschiedliche Entwicklung von *Funken* und *spark* – während die Spezifizität von ersterem stetig sinkt, ist die von letzterem vergleichsweise konstant. Funkenbildung war Teil vieler populärer Schauexperimente zur Elektrizität, so beispielsweise bei der „Elektrischen Apotheose", bei der ein künstlicher Heiligenschein erzeugt wurde (Hochadel 2006: 528).

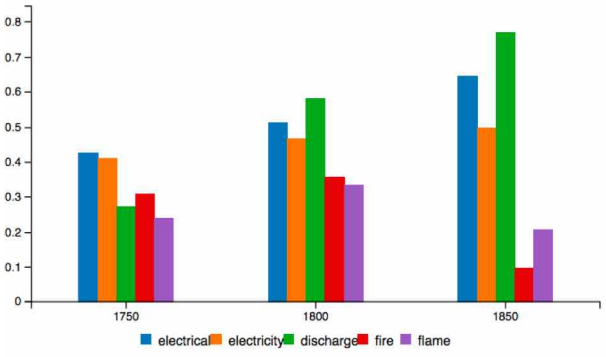


Abbildung 4: Ähnlichkeit ausgewählter Wörter zu ‚spark' im RSC.

Diesem Unterschied zwischen DTA und RSC entspricht, dass das Wort *Funken* eine vergleichsweise konstante Ähnlichkeit zu *Feuer* und *Flamme* hat, während die Ähnlichkeit von *spark* und *fire* bzw. *flame* ab Mitte des 19. Jahrhunderts abnimmt, wie in Abbildung 5 gezeigt. Mögliche Erklärungen sind sowohl sprachspezifische semantische Unterschiede, als auch die unterschiedliche Zusammensetzung der beiden Korpora, wie etwa eine Wechselwirkung der im DTA enthaltenen literarischen Texte (vgl. etwa „Wenn er dich mit seinem Auge electrisirt, fühl es, daß es ein königlicher Funke sey" (Hippel 1781: 28)). Wissenschaftshistorische Studien legen aber auch nahe, dass die im 18. Jahrhundert mit Elektrizität durchgeführten Experimente bis ins 19. Jahrhundert länger und öfter wiederholt und erweitert wurden, als das im deutschen Sprachraum der Fall war. Während die Erscheinung dort zum Ausgang des 18. Jahrhunderts abnahm und rasch zum Elektromagnetismus umgedeutet wurde, verlieren die Elektrizitätsexperimente in Großbritannien nicht an Beliebtheit (vgl. Morus (1998)).

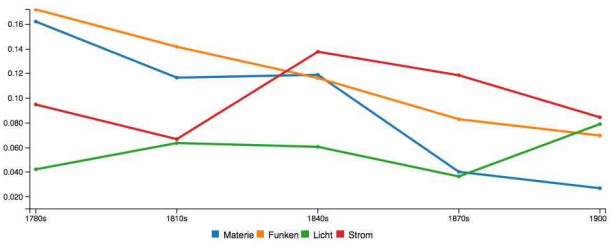


Abbildung 5: Spezifische Kontextwörter für ‚elektrisch' im DTA.

Schluss

Durch die Anwendung von state-of-the-art Methoden der distributionellen diachronen Semantik konnten wir unsere Hypothese zur Entwicklung des Verständnisses von Elektrizität, weg von einem in der Natur beobachteten Phänomen und hin zu etwas im Labor Erzeugtem, das mit anderen Phänomenen wie Magnetismus kombiniert untersucht wird, bestätigen. Dazu erkundeten wir die semantische Entwicklung der Wörter *Elektrizität, electricity, elektrisch, electrical, Funken* und *spark* mittels des SVD $_{\text{PPMI}}$ Word Embedding Verfahrens und des statistischen Maßes x^2. Die Bestätigung unserer Erwartung, die auf dem wissenschaftsgeschichtlichen Forschungsstand basiert, sehen wir als Hinweis für die methodologische Eignung unserer Herangehensweise.

Wir konnten JeSemE erfolgreich zur Bestätigung bereits bestehender Thesen, die auf close-reading-Methoden aufbauen, einsetzen. Wir hoffen, darauf aufbauend Arbeitsabläufe zu entwickeln, die einen gezielten Zugang zu forschungsrelevanten Texten innerhalb umfangreicher Textkorpora wie historische Fachzeitschriften ermöglichen und somit close und distant reading (vgl. Moretti (2013)) verbinden. Die Analyse der Bedeutungsentwicklung über lange Zeiträume und mehrere internationale Korpora hinweg eröffnet neue Forschungsperspektiven, durch die bisher unzugängliche Eigenheiten oder Veränderungen erkennbar werden. Offen bleibt die Frage, wie unsere fokussierte Betrachtung von Einzelwörtern am besten mit der abstrakteren Analyse von Konzepten — dem in der aktuellen Forschung dominierenden Ansatz — verbunden werden kann.

Danksagung

Die beschriebenen Arbeiten wurden von der Deutschen Forschungsgemeinschaft im Rahmen des Graduiertenkollegs „Modell 'Romantik'" (GRK 2041/1) gefördert.

Fußnoten

1. Im Folgenden wird *kursiv* verwendet um untersuchte Wörter vom sonstigen Text abzusetzen.
2. www.jeseme.org
3. Typischerweise einige wenige hundert Dimensionen, in JeSemE kommen 500 zum Einsatz.
4. Grundsätzlich decken beide Korpora längere Zeiträume ab als in den in JeSemE genutzten Versionen, das ein Mindestmaß an Text für die angemessene Modellierung der Wortsemantik benötigt.
5. Die Darstellungsarten für DTA und RSC wurden entsprechend der jeweils für die Analyse nutzbaren Zeitabschnitte gewählt. Alternative Dartellungen wie in Kulkarni u.a. (2015) oder Hamilton u.a. (2016), die Bewegungen innerhalb einer Art Wordcloud nutzen, wurden verworfen, da sie fälschlicherweise eine statische Semantik der Vergleichswörter implizieren.

Bibliographie

Blei, David M. / **Ng, Andrew Y.** / **Jordan, Michael I.** (2003): "Latent Dirichlet Allocation", in: *The Journal of Machine Learning Research* 3: 993–1022.

Bullinaria, John A. / **Levy, Joseph P.** (2007): "Extracting semantic representations from word co- occurrence statistics: A computational study", in: *Behavior Research Methods* 39(3):510–26.

Fankhauser, Peter / **Knappen, Jörg** / **Teich, Elke** (2016): "Topical diversification over time in the royal society corpus", in: *Digital Humanities 2016* 496–500.

Firth, John Rupert (1957): "A synopsis of Linguistic Theory, 1930–1955", in: *Studies in linguistic analysis* 1–32.

Geyken, Alexander (2013): "Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv", in: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie* 221–34.

Hall, David / **Jurafsky, Daniel** / **Manning, Christopher D.** (2008): "Studying the History of Ideas Using Topic Models", in: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* 363–71.

Hamilton, William L. / **Leskovec, Jure** / **Jurafsky, Dan** (2016): "Diachronic Word Embeddings reveal sta- tistical laws of semantic change", in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 1489–501.

Hellrich, Johannes / Hahn, Udo (2016): "Measuring the Dynamics of Lexico-Semantic Change Since the German Romantic Period", in: *Digital Humanities 2016* 545–7.

Hellrich, Johannes / Hahn, Udo (2017a): "Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood", in: *Digital Humanities 2017* 250–2.

Hellrich, Johannes / Hahn, Udo (2017b): "Exploring Diachronic Lexical Semantics with JeSemE", in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics – System Demonstrations* 31–6.

Hippel (1781): *Lebensläufe nach Aufsteigender Linie* . Meines Lebenslaufs Dritter Theil. Zweyter Band.

Hochadel, Oliver (2006): „The Business of Experimental Physics: Instrument Makers and Itinerant Lecturers in the German Enlightenment“, in: *Science & Education 2007* 525–37.

Home, R. W. (2016): „Experimental Physics“, in: Porter, Roy: *The Cambridge History of Science, Volume 4: Eighteenth-Century Science* 363–71.

Jo, Eun Seo (2016): "Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP", in: *Digital Humanities 2016* 582–5.

Kermes, Hannah / Degaetano-Ortlieb, Stefania / Khamis, Ashraf / Knappen, Jörg / Teich, Elke (2016): "The royal society corpus: From uncharted data to corpus", in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* 1928–31.

Kim, Yoon / Chiu, Yi-I / Hanaki, Kentaro / Hegde, Darshan / Petrov, Slav (2014): "Temporal analysis of language through neural language models", in: *Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014* 61–5.

Kulkarni, Vivek / Al-Rfou, Rami / Perozzi, Bryan / Skiena, Steven (2015): "Statistically significant detection of linguistic change", in: *Proceedings of the 24th International Conference on World Wide Web – Technical Papers* 625–35.

Levy, Omer / Goldberg, Yoav / Dagan, Ido (2015): "Improving distributional similarity with lessons learned from Word Embeddings", in: *Transactions of the Association for Computational Linguistics* 3:211–25.

Lichtenberg, Georg Christoph / Erxleben, Johann Christian Polykarp (1787): *Anfangsgründe der Naturlehre* . Dietrich, 4. Auflage.

Manning, Chris / Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing.* MIT Press, Kapitel 5: Collocations.

Mikolov, Tomas / Chen, Kai / Corrado, Gregory S. / Dean, Jeffrey (2013): "Efficient estimation of word representations in vector space", in: *Workshop Proceedings of the International Conference on Learning Representations* https://arxiv.org/abs/1301.3781

Mimno, David (2012): "Computational historiography: Data mining in a century of classics journals", in: *Journal on Computing and Cultural Heritage* 5(1): Article 3.

Moretti, Franco (2013): *Distant Reading* . Verso.

Morus, Iwan Rhys (1998): *Frankenstein's Children. Electricity, Exhibition, And Experiment in Early-Nineteenth-Century London.* Princeton University Press.

Schumann, Anne-Kathrin (2016): "Brave new world: Uncovering topical dynamics in the ACL anthology reference corpus using term life cycle information", in: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL 2016* 1–11.

Steigerwald, Joan (2013): "Rethinking organic vitality in Germany at the turn of the nineteenth century", in: Normandin, Sebastian / Wolfe, Charles T.: *Vitalism and the Scientific Image in Post Enlightenment Life Science, 1800-2010* , Springer, 51-75.

What do you do with 5 million posts? Versuche zum distant reading religiöser Online-Foren

Pfahler, Lukas
lukas@wandelt-pfahler.de
Technische Universität Dortmund

Elwert, Frederik
frederik.elwert@rub.de
RUB Bochum, Deutschland

Tabti, Samira
Samira.Tabti@ruhr-uni-bochum.de
RUB Bochum, Deutschland

Morik, Katharina
katharina.morik@tu-dortmund.de
Technische Universität Dortmund

Krech, Volker
volkhard.krech@rub.de
RUB Bochum, Deutschland

Einleitung

Religiöse Kommunikation als Teil moderner Gesellschaften findet zunehmend auch über internetbasierte Medien statt. Dabei sind es nicht nur liberale Gruppen, die diese neuen Medien nutzen, sondern gerade auch neo-konservative Gemeinschaften wie etwa Evangelikale oder Salafisten. Vor diesem Hintergrund nehmen wir ein spezielles Segment gegenwärtiger Religiosität in den Blick: Neo-konservative christliche und islamische Bewegungen (etwa Evangelikale oder Anhänger der Salafiyya) haben in den letzten Jahren mit eigenen Online-Foren Kommunikationsplattformen geschaffen, in denen sie jeweils eigene Auslegungen in Theologie und Fragen der Lebensführung diskutieren (Becker 2009: 9, Neumaier 2016).

Bei allen Unterschieden zeichnen sich diese Bewegungen durch zwei Merkmale aus: a) eine Universalisierung von Religion im Sinne einer Ablösung „reiner“ Religion von Kultur und Politik, u. b) eine religiöse Durchdringung aller Lebensbereiche, die sich insbesondere durch eine umfassende Regulierung der Lebensführung ausdrückt (Roy 2010: 57). Die Analyse dieser Online-Communities erlaubt es, Rückschlüsse über die Entwicklung und Verbreitung bestimmter Vorstellungen, aber auch über die Genese sozialer Strukturen und neuer Autoritäten zu ziehen.

Ein besonderes Augenmerk legen wir auf die diskutierten Inhalte. Themen und ihre zeitliche Entwicklung werden über Topic Models, Keyword-Analysen und ähnliche Verfahren untersucht. Damit lassen sich thematische Konjunkturen und religiöse Traditionseinflüsse identifizieren.

Datenerhebung und -aufbereitung

Als Grundlage unserer Analysen dienen vier Online-Foren: Zwei christliche (jesus.de seit 2009 online/Christianchat.com seit 2012 online) und zwei muslimische (ahlu-sunnah.com von 2008-2016 online/Ummah.com seit 2002 online), wobei jeweils eins überwiegend deutschsprachig und eins überwiegend englischsprachig ist. Mithilfe eines Web-Crawlers wurden erhebliche Teile der Foren heruntergeladen und für die Analyse zur Verfügung gestellt. Aus den erhobenen HTML Daten werden alle Formatierungen entfernt, sodass der reine textuelle Inhalt vorliegt. Standardtechniken zur digitalen Verarbeitung natürlichsprachlicher Daten werden angewandt, um den Text weiter zu normalisieren. Dazu gehören Tokenisierung, Konvertierung aller Buchstaben zu Kleinbuchstaben, Entfernen von Sonder- und Satzzeichen, etc. Wir entfernen Wörter, die insgesamt seltener als 10 mal verwendet werden. So erhalten wir insgesamt über 5,52 Mio. Posts in über 260,000 Threads oder mehr als 470 Mio. Wörter an Daten.

Des Weiteren verwenden wir domänenspezifische Ersetzungsregeln, um verschiedene Schreibweisen von Referenzen auf externe Quellen wie Koran und Bibel oder externe religiöse Autoritäten wie Schriftgelehrte zu normalisieren. Dies erlaubt es uns ganze Foren hinsichtlich ihrer religiösen Ausrichtung zu untersuchen, da sich je nach Ausrichtung die vornehmlich referenzierten Gelehrten und Textpassagen unterscheiden.

Methoden

Ein häufiges Problem automatischer Textverarbeitung ist, dass zwei Texte zum selben Thema vollständig verschiedene Wörter verwenden können. Ein zentraler Analyseschritt ist aber das automatische Gruppieren ähnlicher Dokumente, genannt Clustering. Clustering eignet sich zum einen, um einen Überblick über die vorherrschenden Themen in Online-Foren zu verschaffen, andererseits eignet es sich auch als Sampling-Instrument um fokussiert Teilmengen für eine manuelle Inhaltsanalyse auszuwählen. Wie aber erkennt man thematische Ähnlichkeiten, wenn ein Vergleich der Mengen der Wörter nicht ausreicht?

Das populäre Topic-Modeling-Verfahren Latent Dirichlet Allocation (LDA) (Blei et al. 2003) berechnet in einem Schritt latente Repräsentationen und Gruppierungen: Dokumente werden einem oder mehreren Topics zugewiesen, die Vektoren der Topic-Zugehörigkeiten dienen als latente Repräsentation. Stellt sich bei der Auswertung der Ergebnisse heraus, dass die Anzahl der Themen zu unpassend gewählt wurde, muss die Berechnung mit veränderten Parametern wiederholt werden. Dabei ist nicht garantiert, dass sich genau die Topics vereinigen oder aufspalten, an denen der Anwender festgemacht hat, dass die Anzahl falsch gewählt wurde. Weiterhin ist es rechenaufwendig, die Granularität der Analyse zu verändern, da die volle Berechnung mit den veränderten Parametern wiederholt werden.

Statt LDA-Repräsentationen zu berechnen, verwenden wir die konzeptionell einfacheren Document Embeddings nach Le und Mikolov (2014). Die latente Repräsentation x ist hier ein niedrig-dimensionaler, reellwertiger Parametervektor einer diskreten kategorischen Verteilung über Wörter in einem Dokument $P(w \mid x, u) \sim \exp(u' x)$. Diese Parametervektoren sowie der Para-

metervektor der Wortverteilungen *u* werden so gewählt, dass die Likelihood der Daten maximiert wird. Dieses Optimierungsproblem betrachten wir als Matrix-Faktorisierungsproblem; statt über die *x* und *u* zu optimieren, optimieren wir über die Matrix der jeweiligen Skalarprodukte *u*' *x*.

Mithilfe eines numerischen Optimierungsverfahrens können die latenten Repräsentationen berechnet werden.

In einem zweiten Schritt wird das Clustering der Dokumentenkollektion auf Basis der latenten Repräsentationen berechnet. Hierzu verwenden wir das Agglomerative Hierarchische Clustering. Das Verfahren berechnet einen Cluster-Baum, an dessen Blättern die einzelnen Dokumente liegen; durch sukzessive Vereinigung der zwei ähnlichsten Cluster entsteht ein Baum. Dieser erlaubt uns beliebige Anzahlen von Gruppen zu identifizieren, indem der Baum auf einer festgelegten Höhe abgeschnitten wird. Soll die Clusteranzahl verändert werden, müssen die latenten Repräsentationen der Dokumente nicht neu berechnet werden, es muss lediglich der vollständige Baum anders abgeschnitten werden. Dazu gibt es verschiedene Möglichkeiten: Einzelne Cluster können in ihre zwei Untercluster aufgespalten werden oder andersrum wiedervereinigt werden oder es kann eine andere feste Gesamtanzahl angegeben werden. So kann interaktiv und dynamisch ein Clustering der Dokumente erarbeitet werden; jede Änderung der Cluster-Hierarchie ist in wenigen Sekunden berechnet.

Damit das Clustering interpretiert werden kann, werden Repräsentanten der Cluster berechnet. Wie die Topics bei LDA-Modellen handelt es sich auch hierbei um Wortlisten, die prinzipiell angeben, wie oft die jeweiligen Wörter in jedem Cluster vorkommen.

Besser interpretierbare Ergebnisse werden erzielt, wenn statt der Anzahl der TFiDF-Score (Robertson 2004) der Wörter angegeben wird; dieser setzt die Anzahlen ins Verhältnis zur Anzahl der Dokumente, in denen die jeweiligen Wörter verwendet werden.

Vorläufige Ergebnisse

Das oben vorgestellte Verfahren erlaubt es, die ausgewählten Korpora für die weitere qualitative Feinanalyse gezielt aufzuarbeiten. Die folgende Tabelle zeigt exemplarisch 6 verschiedene Themen, die wir in den Foren-Diskussionen identifiziert haben. In diesen semantischen Feldern finden die unterschiedlichen Diskurse immer im Rahmen religiöser Argumentation und Legitimation statt. Beginnend mit Diskussionen über das Weltgeschehen, besonders im islamischen Raum (Themen 4,5), zur Lebensführung, in diesem Fall Ernährung (Thema 1), über religiöse Rituale oder Handlungen wie Heirat oder Gottesdienst (Thema 2,3) bis zu den unmittelbaren theologischen Diskussionen wie bspw. Diskussionen religiöser Schriften (Thema 6).

Thema 1	Thema 2	Thema 3
halal	moscheen	heirat
fleisch	gebetet	heiraten
enthalten	beten	wali
alkohol	verrichten	ehe
trinken	freitagsgebet	verheiratet
essen	raum	scheidung
tier	gebete	ehemann
blut	isha	vater
haraam	stadt	mahram

Thema 4	Thema 5	Thema 6
staat	soldaten	sonne
demokratie	politik	wohlgefallen
gesellschaft	spiegel	himmel
krieg	afghanistan	paradies
ländern	taliban	erklärung
gesetze	krieg	überlieferung
länder	israel	erde
regierung	regierung	sallallahu
staaten	usa	berichtete

Die Frage der Lebensführung in den neokonservativen Religionsgemeinschaften ist für unsere Forschung zentral. Die Identifizierung von Themenbereichen und Schlüsselwörtern, die mit Lebensführungsfragen in Verbindung stehen, können erste Hinweise darüber geben, was für die salafistisch korrekte Lebensführung besonders stark diskutiert wird. Auch interessiert uns, wie stark diese Themenkomplexe mit religiösem Vokabular durchsetzt sind.

Für verschiedene Bereiche der Lebensführung lassen sich diesbezüglich folgende Beobachtungen machen:

Ernährung und Lebensmittel werden hier besonders im Zusammenhang mit haram/halal (erlaubtes/nicht erlaubtes), also insbesondere aus einer religiösen Perspektive, diskutiert. Daneben finden sich aber auch eher auf Gesundheitsfragen ausgerichtete Diskussionen.

Musik als Thema wird auffällig häufig diskutiert und taucht in unterschiedlichen thematischen Clustern auf: Einmal im Zusammenhang mit Musikinstrumenten und der Bewertung als haram/halal, aber auch im Themenfeld Medien (gemeinsam mit Film und Fotografie) sowie im Zusammenhang mit negativ konnotierten Verhaltensweisen (Alkohol, Glücksspiel).

Im nächsten Schritt sollen Themen und religiöse Quellen/Autoritäten miteinander in Beziehung gesetzt werden. Dabei interessiert uns die Frage, auf welche Schriftquellen und welche Gelehrten sich die AkteurInnen berufen, wenn sie bestimmte Themen diskutieren.

Am Beispiel ahlu-sunnah.com zeigt sich, dass bestimmte salafistische Schriftgelehrte eine bedeutendere Rolle spielen als andere (siehe Abbildung 1). Die Zwischenergebnisse lassen erste Rückschlüsse auf die religiösen Richtungen und Referenzpraktiken zu: welche salafistischen Gelehrten finden mehr Zustimmung als andere und welche salafistischen Traditionsschulen lassen sich identifizieren (Wahabiya, Ad-Da'wa As-Salafiya, Madchalia etc.). So sind z.B. die Gelehrten Ibn Taymiyyah und al-Albani wichtige Referenzgrößen. Ibn Taymiyyah (13. Jhd.) als "Vater der Salafiyya" findet große Verehrung bei den puristischen sowie auch bei den politisch-aktivistischen Gruppierungen der Salafiya-Bewegung. Al-Albani (20. Jhd.) dagegen findet eher in der puristischen Bewegung Zuspruch und wird auch kontrovers diskutiert. Vor dem Hintergrund der thematischen Analyse lässt sich dann noch weiter untersuchen, welche Quellen verstärkt für welche Themenkomplexe

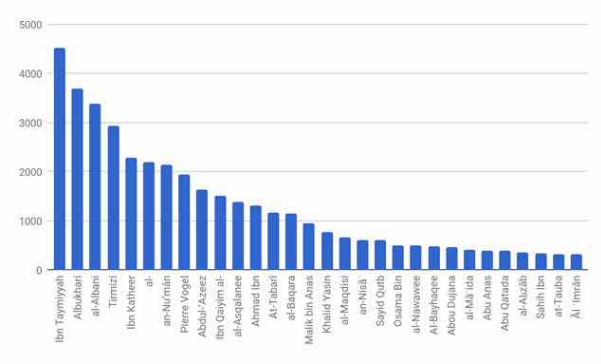


Abbildung : Quellenreferenzen in ahlusunnah.com

Bibliographie

Becker, Carmen (2009): "Gaining Knowledge: Salafi Activism in German and Dutch Online Forums" in: Masaryk University Journal of Law and Technology 3 (1): 79–98.

Blei, David M./Andrew Y. Ng/Michael I. Jordan (2003): "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, no. 4–5, pp. 993–1022.

Le,Quoc V./ Tomas Mikolov (2014): "Distributed Representations of Sentences and Documents" in: International Conference on Machine Learning - ICML 2014, vol. 32, pp. 1188–1196.

Neumaier, Anna (2016): Religion@home? Religionsbezogene Online-Plattformen Und Ihre Nutzung: Eine Untersuchung Zu Neuen Formen Gegenwärtiger Religiosität. Religion in Der Gesellschaft 39.

Pang-Ning, Tan/Michael, Steinbach/Vipin, Kumar (2006): Introduction to data mining

Pfahler, Lukas/Katharina, Morik/Frederik, Elwert/Samira, Tabti/Volkhard, Krech (2017): "Learning Low-Rank Document Embeddings with Weighted Nuclear Norm Regularization" in: Proceedings of the 4th DSAA

Robertson, Stephen (2004) "Understanding inverse document frequency: on theoretical arguments for IDF," in: J. Doc., vol. 60, no. 5, pp. 503–520.

Roy, Olivier (2010): Heilige Einfalt: Über Die Politischen Gefahren Entwurzelter Religionen. München.

Wissenschaft ohne Geist: Herausforderungen der Digital Humanities am Beispiel der Korpuslinguistik

Bubenhofer, Noah
bubenhofer@cl.uzh.ch
Universität Zürich, Schweiz

Alle, die mit maschinellen Methoden Sprache analysieren, erleben momentan einen tiefgreifenden methodologischen Wandel.[1] Einerseits erfreut man sich vielleicht als Geisteswissenschaftler/in am immer stärkeren Interesse der Ingenieurtechniken und der Informatik für Sprache. Dies kann durchaus als Erfolg der Linguistik betrachtet werden, zurückgehend auf den Linguistic Turn, der viele andere Disziplinen schon seit Jahrzehnten beeinflusst. Unternehmen interessieren sich für ihre Reputation im massenmedialen Diskurs oder sind der Überzeugung, ihr in unzähligen Dokumenten versprachlichtes Wissen besser verwalten zu können, wenn sie es nach sprachlichen Kriterien neu ordnen. Das Geschäftsmodell von Internetunternehmen basiert ganz erheblich darauf, sprachliche Kommunikation maschinell zu verarbeiten um daraus Wissen aufzubauen und Vorhersagen über das Handeln

von Kunden zu machen. Auch in der Politik ist die Analyse von Sprachgebrauch ein wichtiger Faktor, um Wahlkämpfe zu gewinnen.

Andererseits beschert einen dieses Interesse eine Vielzahl von neuen Methoden für die maschinelle Analyse von Text, die auch für geisteswissenschaftliche Fragestellungen interessant sind. Die Digital Humanities sind ein Beispiel für eine Disziplin, die sich den Experimenten mit diesen Methoden verschrieben hat. Auch die Korpuslinguistik profitiert maßgeblich von diesen neuen Methoden.

Aktuell erfahren in der Computerlinguistik und generell im Data Mining neuronale Netze großen Zuspruch, die den Prozess des maschinellen Lernens nach dem Modell des menschlichen Gehirns gestalten. Solche Systeme, „Deep Learning"-Systeme genannt, sind in der Lage, Muster in den Daten zu erkennen, ohne dass vorher explizit die Eigenschaften festgelegt werden, die getestet werden sollen. Zudem findet das Lernen auf mehreren verborgenen Ebenen statt, so dass das Lernen nicht beobachtet und damit auch die Frage, welche Eigenschaften nun welchen Einfluss auf das gelernte Modell haben, kaum beantwortet werden kann.

In der Computerlinguistik wurden bereits für viele Probleme Deep-Learning-Algorithmen eingesetzt, meist mit Erfolg. Erfolg bedeutet, dass die statistischen Modelle besser den Goldstandard voraussagen können, aber nicht, dass das grundlegende linguistische Problem (z.B.: Sentiment-Analyse: wie werden Gefühle und Meinungen ausgedrückt; Textklassifikation: wie drückt sich Stil, Autorschaft, Textsorte, Thema etc. aus) besser gelöst wäre.

Überall wo Sprachgebrauch quantitativ und maschinell analysiert wird, gibt es einen starken Trend, möglichst ohne linguistischen Kategorien und Theorien auszukommen und Black-Box-Systeme zu verwenden. Das ist nachvollziehbar, da es in den meisten Fällen darum geht, ein System zu bauen, das eine klar definierte Aufgabe sehr gut lösen kann. Obwohl diese Ansätze natürlich auch für die Korpuslinguistik interessant sind, genügen sie linguistischen Forschungsinteressen eigentlich nicht, da sie keinen Beitrag dazu leisten, sprachliche Phänomene zu verstehen und erklären zu können.

Viel dramatischer ist jedoch, dass die Linguistik offensichtlich nicht in der Lage ist, einen nützlichen Beitrag zur Lösung der Probleme der maschinellen Textanalyse zu leisten. Die Linguistik scheint für die quantitative Analyse von Text weitgehend bedeutungslos zu werden.

Um der Bedeutungslosigkeit zu entgehen, muss die Linguistik ein kritisches Verhältnis zur Forschungslogik in den ingenieurstechnischen Disziplinen pflegen und auf zwei Prinzipien bestehen: 1) Mehr linguistische Theorie. 2) Ergebnisse von quantitativen Analysen müssen gedeutet werden.

Zu 1): Nicht nur für die Linguistik, sondern für alle geistes- und sozialwissenschaftlichen Disziplinen gilt: Eine theoretische Fundierung der Analysekategorien ist essentiell. Dafür werden valide Analysekategorien benötigt, die deutbar sind. Dieses Prinzip richtet sich jedoch keinesfalls gegen datengeleitete Verfahren, im Gegenteil: Sie sind es, die die theoretischen Modelle herausfordern und schärfen können. Aber das Ziel aller Analysen muss darin liegen, ein Puzzleteil zu einem besseren Verständnis sprachlicher Strukturen, von Sprachgebrauch oder gesellschaftlichen und kulturellen Bedeutungen von Sprache zu führen. Wir benötigen White-Box-, nicht Black-Box-Systeme.

Das Problem der fehlenden Validität zeigt sich z.B. im Feld der sog. „Authorship Attribution", also der Zuordnung eines Textes X zu einem Autor A, B, C, …. Um dies zu tun, stehen Texte zur Verfügung, von denen die Autorschaft bekannt ist. Die Frage ist dann also, ob über die sprachlichen Merkmale des Textes X automatisch bestimmt werden kann, wer der Autor/die Autorin (aus der Menge der möglichen Autoren/innen) von Text X ist. Genauer lautet die Frage aber, ob und wie sich persönlicher Schreibstil sprachlich niederschlägt.

Besonders erfolgreich für diese Aufgabe sind Methoden maschinellen Lernens, die das Problem als Klassifikationsaufgabe auffassen und anhand von Trainingskorpora typische sprachliche Merkmale der Texte der jeweiligen Autor/innen lernen. Dabei zeigt sich, dass „low-level features like character n-grams are very successful for representing texts for stylistic purposes" (Stamatatos, 2009, S. 24). Das bedeutet, solche Modelle, die auf der Distribution von Buchstaben-N-Grammen beruhen, sind, gemessen an einem Goldstandard, am erfolgreichsten. Allein: Solche Modelle lassen sich nicht linguistisch deuten, da völlig unklar ist, was sie eigentlich messen. Ist es Stil, Thema, Textsorte, …? Es handelt sich also weder um eine valide, noch um eine deutbare Kategorie (insbesondere, wenn das statistische Modell nicht einsehbar ist). Für spezifische Aufgaben der Autorschaftsattribution mag das ausreichend sein, aber bereits für forensische Anwendungen, beispielsweise vor Gericht, ist eine solche Modellierung fragwürdig und gefährlich. Und für eine linguistische Deutung des Phänomens Autorschaftsstil ist sie gänzlich unbrauchbar.[2]

Die Kritik geht jedoch nicht nur in Richtung des Textminings und der Computerlinguistik, manchmal nicht-valide Kategorien einzusetzen (was zudem oft für die dortigen Zwecke auch sinnvoll ist), sondern auch in die Richtung der Linguis-

tik: Die Computer- und die Korpuslinguistik zeigen beide gleichermaßen, wie wichtig es ist, auch abstrakte Kategorien so zu definieren versuchen, dass überhaupt eine Chance besteht, sie für eine quantitative Analyse operationalisierbar zu machen. Wenn eine linguistische Kategorie so vage ist, dass sich selbst (geschulte) Menschen uneinig darüber sind, wenn sie an authentischem Sprachegebrauch angewendet werden, scheitert die quantitativ-maschinelle Lösung unweigerlich.

2) Die Ergebnisse von quantitativen Analysen sind nicht Antworten auf Fragestellungen, sondern neue Daten, die vor einem geistes- und sozialwissenschaftlichen Hintergrund genauso hermeneutisch gedeutet werden müssen, wie einzelne Texte. Das ist vielleicht das größte Missverständnis, wenn Textminer und Computerlinguistinnen mit Korpuslinguistinnen zusammenarbeiten: Erstere wollen, dass ein Werkzeug ein Ergebnis hervorbringt, das an einem Goldstandard evaluiert werden kann. Das Ergebnis ist dann im Einzelfall richtig oder falsch und in der Gesamtheit genügend präzise oder nicht. Das Ergebnis ist dann auch im Idealfall die Lösung der Forschungsfrage. Bei den meisten geistes- und sozialwissenschaftlichen Fragestellungen beginnt auf der Grundlage dieser Ergebnisse jedoch ein Interpretationsprozess, um (meist in Kombination mit weiteren Analysen) eine plausible Deutung zu ermöglichen – eine vorläufige Deutung. Die Stärke der Geistes- und Sozialwissenschaften liegt dabei ja gerade darin, dass in ihrer Methodologie ein Zweifeln inhärent ist, mit dem die „gegenwärtig besiegelten Bedeutungen jeweils eingeklammert oder angezweifelt [werden], um zu prüfen, inwiefern sich nach rationalem Ermessen nicht bessere Lösungen, überlegenere Interpretationen oder zustimmungsfähigere Regelungen finden lassen“ (Honneth, 2016, S. 312).

Neben der Suche nach validen Analysekategorien und dem Hochhalten geisteswissenschaftlicher Prinzipien der Deutung sehe ich einen weiteren Aspekt, der helfen sollte, der Korpuslinguistik eine deutliche linguistische Prägung zu verleihen. Es ist der Versuch, korpuslinguistisches Arbeiten als „diagrammatisches Operieren“ aufzufassen. Mit dem Diagramm-Begriff folge ich Krämer (2016), die deutlich macht, dass Diagramme als Formen der Visualisierung von Daten „Denkzeuge“ sind, mit denen operiert wird: Ich kann Daten in einem Diagramm darstellen (auf einer Karte, in einem Netzwerkgraph, einem Punkteplot, …) und danach damit operieren, um neue Erkenntnisse daraus zu ziehen. Wenn man einem breiten Diagramm-Begriff folgt, wird deutlich, dass auch Listen, Tabellen und dergleichen diagrammatischen Charakter haben (Siegel, 2009; Steinseifer, 2013). Dies sind nun aber Formen, die in der Korpuslinguistik zentral sind: Die Keyword in Context-Liste (zurückgehend etwa auf Zettelkästen im 16. Jahrhundert) etwa kann als Keimzelle eines völlig neuen Textverständnisses angesehen werden, mit dem die Einheit des Textes zerstört wird, um eine neue Sicht auf Textdaten zu gewinnen. Viele weitere Formen der Anordnung von Textdaten spielen ebenfalls wichtige Rollen, entscheidend etwa die Überführung von Textdaten in den Vektorraum, in dem operiert werden kann (z.B. in Form geometrischer Operationen – Lagen von Vektoren und ihren Winkeln zueinander). Aber auch die Erfindung der Partiturdarstellung bei Gesprächstrankripten, mit der überhaupt erst eine moderne Gesprächslinguistik möglich wurde, zeigt die Kraft von diagrammatischen Umformungen, um Daten neu lesbar zu machen. Hinter diesen diagrammatischen Umformungen stecken diagrammatische Grundfiguren (Bubenhofer im Druck b), die in den Geisteswissenschaften generell wirkmächtig sind.

Ich meine, es lohnt sich, korpuslinguistisches Arbeiten unter diagrammatischer Perspektive zu reflektieren, um die Mechanismen und Möglichkeiten der Gegenstandskonstitution besser zu verstehen. Die Digitalität der Daten und Methoden erlaubt dabei neue Transformationen und macht Daten, egal welcher Modalität, miteinander verrechenbar. Aber die diagrammatischen Grundfiguren führen zu unterschiedlichen Gegenständen: Repräsentiert in einem Vektorraum geben die gleichen Daten einen völlig anderen Gegenstand ab als dargestellt in einer Keyword in Context-Liste. Und es müsste vordringliches Ziel sein, noch ganz andere Formen der diagrammatischen Darstellung von Text zu finden, um damit andere Gegenstandskonstitutionen und Fragestellungen zu ermöglichen. Die algorithmische Repräsentation der Daten folgt dabei ebenfalls den diagrammatischen Transformationen (Beispiel Vektorraum) und kann daher nicht unabhängig davon gedacht werden. Für eine hermeneutische Deutung brauchbare Analysekategorien zu erarbeiten, bedeutet deshalb auch, die damit verbundenen diagrammatischen Operationen zu reflektieren. Dafür nötig sind semiotische und natürlich auch wissenschaftstheoretische Überlegungen, die für alle Disziplinen, die mit maschineller Textanalyse befasst sind, relevant sein müssten.

Fußnoten

1. Dieses „extended Abstract“ ist eine verkürzte und angepasste Fassung des stärker linguistisch ausgerichteten Beitrages von Bubenhofer (im Druck a).

2. Vgl. für eine aktuelle linguistisch motivierte Diskussion von stilometrischen Messmethoden für die Autorschaftsattribution Büttner et al. (2017).

Bibliographie

Bubenhofer, Noah (im Druck a): Wenn „Linguistik“ in „Korpuslinguistik“ bedeutungslos wird. Vier Thesen zur Zukunft der Korpuslinguistik. In: Osnabrücker Beiträge zur Sprachtheorie (OBST).

Bubenhofer, Noah (im Druck b): Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In: Bubenhofer, Noah / Kupietz, Marc (Hg.): Visualisierung sprachlicher Daten. Heidelberg: HeiUP.

Büttner, Andreas / Dimpel, Friedrich Michael / Evert, Stefan / Jannidis, Fotis / Pielström, Steffen / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (2017): »Delta« in der stilometrischen Autorschaftsattribution. In: Zeitschrift für digitale Geisteswissenschaften. text/html Format. DOI: 10.17175/2017_006.

Honneth, Axel (2016): Denaturierung der Lebenswelt. Vom dreifachen Nutzen der Geisteswissenschaften. In: Panteos, A./Rojek, T. (Hrsg.): *Texte zur Theorie der Geisteswissenschaften, Reclams Universal-Bibliothek*. Stuttgart : Reclam, S. 283–315

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis: Grundlinien einer Diagrammatologie.* Frankfurt/Main: Suhrkamp Verlag.

Nakov, Preslav/Ritter, Alan/Rosenthal, Sara/Stoyanov, Veselin/Sebastiani, Fabrizio (2016): SemEval-2016 Task 4: Sentiment Analysis in Twitter. In: *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval ’16*. San Diego, California : Association for Computational Linguistics.

Siegel, Steffen (2009): *Tabula: Figuren der Ordnung um 1600*. Berlin / Boston : Akademie-Verlag.

Stamatatos, Efstathios (2009): A Survey of Modern Authorship Attribution Methods. In: *J. Am. Soc. Inf. Sci. Technol.* Bd. 60, Nr. 3, S. 538–556

Steinseifer, Martin (2013): Texte sehen – Diagrammatologische Impulse für die Textlinguistik. In: *Zeitschrift für germanistische Linguistik* Bd. 41, Nr. 1, S. 8–39

Zur Weiterentwicklung des “cognition support”: Sammlungs-visualisierungen als Austragungsort kritisch-kulturwissenschaftlicher Forschung

Windhager, Florian
florian.windhager@donau-uni.ac.at
Donau-Universität Krems, Österreich

Glinka, Katrin
k.glinka@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz, Deutschland

Mayr, Eva
eva.mayr@donau-uni.ac.at
Donau-Universität Krems, Österreich

Schreder, Günther
Guenther.Schreder@donau-uni.ac.at
Donau-Universität Krems, Österreich

Dörk, Marian
doerk@fh-potsdam.de
Fachhochschule Potsdam

Einleitung

Interfaces und Methoden der Informationsvisualisierung dienen insbesondere in Bezug auf abstrakte und komplexe Gegenstände der Unterstützung, Verstärkung und Augmentierung der menschlichen Kognition (Arias-Hernandez et al., 2012). Sammlungen des kulturellen Erbes (Galerien, Bibliotheken, Archive und Museen) sind Paradebeispiele für solche komplexen Gegenstände: sie organisieren und bereiten tausende Objekte auf und stellen diese gemeinsam mit assoziierten Informationen für Forschung und Öffentlichkeit bereit. Viele dieser Sammlungen sind mittlerweile digitalisiert im Netz zugänglich, womit lokale Sammlungs-Interfaces und große Aggregatoren zu Portalen von neuen Informationsräumen werden, in denen Kultur erlebbar und verhandelbar wird.

Ausgangspunkt unseres Vortrags ist eine aktuelle Studie zu Sammlungsinterfaces, die Methoden der Informationsvisualisierung nutzen um kognitive Operationen wie Exploration, Navigation und Analyse auf verschiedenen Ebenen einer Sammlung zu unterstützen.[1] Im Vortrag werden wir einige der durch die Studie gewonnenen Erkenntnisse vertiefen und die Frage ins Zentrum stellen, wie Visualisierungsinterfaces auch jene kognitiven Operationen fördern können, die im kulturwissenschaftlichen Kontext als "kritische" tradiert werden. Analog zu existierenden Definitionen (vgl. Jaeggie & Wesche, 2009; Foucault, 1990; und Butler, 2001) verstehen wir Kritik als jene Form der Kognition, die einen Gegenstand - und das ihn konstituierende Forschungssystem - in seiner Umwelt kontextualisiert und einer Bewertung unterzieht. Dabei bündelt die Operation i) analytisches und exploratives Wissen über Struktur und Dynamik ihres Gegenstands, ii) eine Bewertung im Sinne einer differenzierten Vermessung von Aktualität und Potentialität des Gegenstands, iii) eine Offenlegung und Argumentierung der instrumentalisierten Maßstäben und Normen, sowie oftmals eine iv) Ableitung von Handlungsoptionen zur (Selbst)berichtigung und (Selbst)Steuerung des fokussierten und des fokussierenden Systems.

Mit direktem Bezug auf Fragestellungen des Calls entwickeln wir ein Analyseschema, um dieses Potenzial für Visualisierungssysteme zu kulturellen Sammlungen zu erkunden und diskutieren Designstrategien, um entsprechende Funktionen zu stärken.

Visuelle Sammlungsinterfaces

Interfaces zu kulturellen Sammlungen vermitteln zwischen großen Mengen von digitalen Objekten und den darauf bezogenen Absichten und Interessen von ExpertInnen oder BesucherInnen. Abbildung 1 zeigt eine schematische Aufreihung der wichtigsten Komponenten eines entsprechenden Systems, in dem das visuelle Interface (Mitte) einen kognitiven Mehrwert auf BenutzerInnenseite (rechts) schaffen soll (vgl. Card et al., 1999, S. 17). Dieser kognitive Mehrwert umfasst beispielsweise einen Erkenntnisgewinn bezüglich einer digitalen Sammlung, welche oftmals auf physische Objektsammlungen und letztlich auf Ursprungskulturen oder Sammlungskontexte verweisen (links).

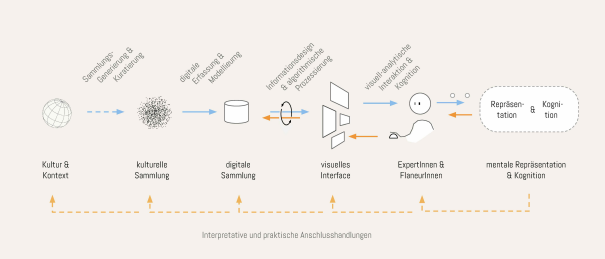


Abb.1: Komponenten eines visualisierungsgestützten HCI-Systems, das Daten von kulturellen Sammlungen (links) in visuellen Interfaces repräsentiert (Mitte), um damit diverse kognitive Operationen (wie Exploration, Navigation, Analyse, und mentale Repräsentation) zu unterstützen (rechts) .

Interfacesysteme - als zunächst meist opake Ensembles von Datenbank, Algorithmik und Benutzeroberfläche - entscheiden in solchen Konstellationen über eine zentrale Passage der Informationsverarbeitung und haben nicht zuletzt deshalb kritische Aufmerksamkeit verdient. Dies betrifft sowohl die Skepsis, ob quantitative und algorithmische Visualisierungsverfahren überhaupt nicht-verfremdend oder epistemologisch "a-trojanisch" (Drucker, 2011) auf Feldern von geistes- und kulturwissenschaftlichen Fragestellungen eingeführt und genutzt werden können, wie auch den konstanten internen Diskurs der Methodenkritik, der die Konferenzen und Reader der Digital Humanities bestimmt. So wie alle mediale Technologien beeinflussen Interfacesysteme den Aufbau der mentalen Repräsentationen ihrer Gegenstände über multiple Stationen der Übertragung und Übersetzung (Vektoren in blau, Abb. 1) und begünstigen oder erschweren in der Folge gewisse interpretative oder praktische Anschlussoperationen (Vektoren in orange).

In der diesem Beitrag zugrundeliegenden Studie erfassten wir den Gestaltungsraum solcher medialen Systeme im Sammlungskontext und leiteten Anforderungen an zukünftige Interfaces, darunter *Generosität* (Dörk, 2011; Whitelaw, 2015), *Serendipität* (Thudt et al., 2012), *Narrativität* (Davis et al., 2016) und *kontextuelle Konnektivität* (Hooland et al., 2014), als wichtige Gestaltungsprinzipien ab.

Strategien des "Critical Cognition Support" durch Sammlungsinterfaces

Eine weitere der identifizierten Anforderung beschreibt den Aspekt des *"Critical Cognition Sup-*

port”, welcher im Kontext dieses Calls eine analytische Vertiefung verdient. Darunter verstehen wir “ *kritik-unterstützende”* Funktionen von visuellen Interfaces, die über Exploration und Analyse hinaus eine differenzierte Beurteilung und gegebenenfalls eine (Selbst)berichtigung des visuell gestützten Forschungssystems ermöglichen - inklusive einer Kritik des kulturellen Gegenstands selbst. Auf medientheoretischer Basis ist bereits evident, dass die Gegenstände der Digital Humanities durch Interface- und Forschungssysteme entscheidend mit konstituiert werden, was einer konstanten Reflexion bedarf. Abgesehen von der Diskussion dieser medialen und technologischen Bedingtheit von digitalen Gegenständen besteht jedoch ebenso die Notwendigkeit, sich kritisch mit den vermittelten „Dingen von Belang“ - den Sammlungen, ihren Objekten, Kulturen und Kontexten - und deren historischen Bedingtheiten auseinanderzusetzen. Die Balancierung von medienkritischen Praktiken mit der kritischen Reflexion der modellierten Realitäten und ihrer prämedialen Verfasstheit eröffnet auch eine produktive Strategie, um ein mögliches Abdriften in die Selbstbezüglichkeit des digitalen Methodendiskurses zu verhindern (cf. Latour, 2004). Sammlungsvisualisierungen sollten vor diesem Hintergrund so oft wie möglich über scheinbar objektive oder rein deskriptive Abbildungen ihrer Gegenstände hinausgehen, da diese immer produktiv bezüglich ihrer Position und Funktion in gesellschaftlichen, sozio-ökonomischen, epistemischen oder normativen Kontexten befragt werden können und müssen (vgl. Calhoun, 1995; Swartz, 2012).

Die relevantesten Dimensionen der kritischen Kognition innerhalb eines Forschungssystems lassen sich in der Folge durch multiple Vektoren der Systemkritik fokussieren (siehe Abb. 2 oben): a) Kritik und Evaluation der Kultur und des Kontexts einer Sammlung, b) Kritik der Sammlungs- und Kuratierungspraxis, c) Kritik der digitalen Modellierung, sowie d) Kritik der Informationsvisualisierung.

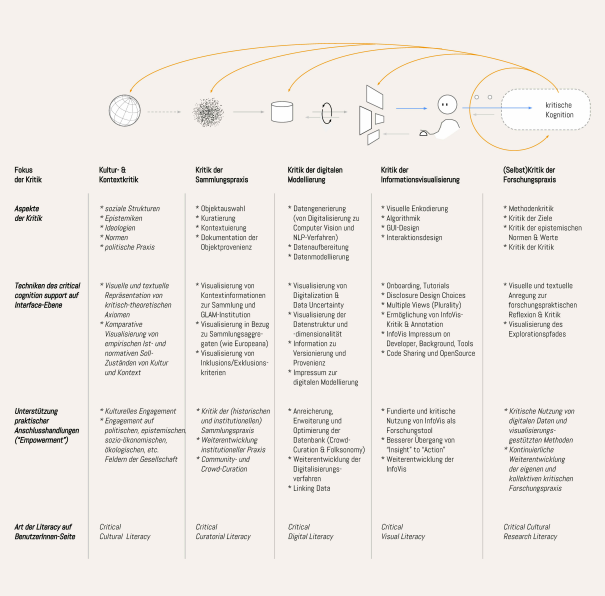


Abb.2. Kritische Kognition kann durch das Zusammenspiel von Designstrategien für visuelle Interfaces mit Bezug auf multiple Systemkomponenten gefördert werden. Wir unterscheiden in der Folge pro Komponente Aspekte der Kritik, Techniken der Kritik-Unterstützung, praktische Relevanz und individuelle “Critical Literacies”.

Aus multipler Motivation (als AnwenderInnen, EntwicklerInnen und AnalytikerInnen von Interfacesystemen) gehen wir davon aus, dass sich die Kritik und Transparenz dieser Systemkomponenten durch Designprinzipien von Interfaces selbst fördern lassen, und dass sich die kritische Interpretation der Komponenten oft nur schrittweise und “rückwärts” (d.h. mit steigender kulturwissenschaftlicher Relevanz von der Interfacebis zur Kuratierungs- und Kulturkritik) entfalten lässt. Wir summieren Aspekte und mögliche Gestaltungsoptionen zur Unterstützung dieser Komponenten in Abbildung 2 und beleuchten ein paar Facetten in der abschließenden Diskussion.

a) Kritik der Informationsvisualisierung: Sammlungsinterfaces können ihre eigenen technologischen Funktionsprinzipien zur kritischen Ermächtigung ihrer NutzerInnen via onboarding-Techniken, Tutorials, und Offenlegung von design choices transparent machen (vgl. Dörk et al., 2013). Akteure hinter Interfaceprojekten können gemeinsam mit ihren Strategien, Interessen und technologischen Präferenzen (Bubenhofer, 2016) über Kontextinformation kenntlich gemacht werden. Weiterhin kann die Kritik von Visualisierungen durch die Implementierung von multiple views gefördert werden, die Pluralitäten betonen und den Vergleich von Stärken und Schwächen der jeweiligen Darstellungen ermög-

lichen. Transparenz kann weiterhin durch Code Sharing und Open Source Dissemination unterstützt werden.

b) Kritik der digitalen Modellierung: Methoden und Prozesse der Datengenerierung, -aufbereitung und -modellierung spielen eine konstitutive Rolle für die spezifische Medialität des Systems, seine analytischen Möglichkeiten, Grenzen, aber auch Unbestimmtheiten und Unschärfen. Dies gilt für die Modellierung von quantitativen und kategorialen Metadaten aus bestehenden analogen Beständen, sowie umso mehr für Systeme die Verfahren des Natural Language Processings und der Computer Vision in die Datengenerierung einbeziehen. Solche Informationen zur Provenienz von Daten sollten in visuelle Repräsentation ebenso einfließen wie Informationen zu Akteuren, Institutionen und Konventionen der digitalen Modellierung.

c) Kritik der Kurations- und Sammlungspraxis: Datenbanken und Metadaten beruhen häufig auf historischen Sammlungskatalogen und Erfassungssystemen, deren Selektivität bei der Entwicklung von kritisch informierten Sammlungsrepräsentationen reflektiert werden muss. Dies kann beispielsweise in einer Fokussierung von Visualisierungen auf Themen, Objekte oder Akteure einer Sammlung geschehen, die aufgrund von historischen Kanonisierungen, kuratorischer Selektion, soziokulturell und historisch bedingten Strukturen von Exklusion oder institutionellem Bias nicht (oder wenig) repräsentiert sind (Glinka et al., 2015). Im Sinne einer Diversifizierung von Narrativen im Anschluss an institutionskritische Interventionen ermöglichen Sammlungsvisualisierungen die Verhandlung von kritischen Interpretationen und Analysen der institutionellen Selbstbeschreibungen. Zu den kritikunterstützdenden Funktionen von Interfaces zählen beispielsweise auch Designs, welche institutionelle Wissensbestände und Interpretationen in Form von außer-institutioneller Teilhabe über Folksonomy, Crowd-Curation oder Community Co-Creation anreichern, in Frage stellen oder ergänzen.

d) Kulturkritik von Sammlungen: Sammlungen sind in der Regel reichhaltige Quellen der Information über assoziierte Kulturen, deren heterogene und herrschende Intentionen, Interessen, Epistemiken, Ideologien und sozio-ökonomischen Praktiken zum Kerngebiet kulturwissenschaftlicher Forschung gehören. Wir gehen zudem davon aus, dass die kritische Interpretation von Sammlungen teilweise nur über eine differenzierte Bewertung ihres Kontextes gewonnen werden kann. Wir sehen Möglichkeiten, solche Kontextualisierungen schon auf Ebene des Interfacedesigns herzustellen und ihre Interpretation mit Axiomen aus dem Bestand der kritischen Kulturtheorien zu verknüpfen. Die Veranschaulichung von historischen kulturellen Soll-Werten oder normativen Bewertungskategorien kann bei deren Neubewertung unterstützen. Die direkte Aktivierung von kritischem Engagement im gesellschaftlichen Kontext kann nicht zuletzt bei Interfaces zu zeitgenössischen Sammlungen eine zentrale Rolle spielen.

Resümee und Ausblick

Wir präsentieren ein modulares und multifokales Schema für die Beurteilung und Entwicklung von Designstrategien, die kritische Kognition in digitalen Systemen der visuellen Analyse von Kulturdaten unterstützen. Über die oftmals technikgeleitete Entwicklung von analytischen und explorativen Systemen hinaus soll dies dabei helfen, den Anschluss der digitalen Tool-Entwicklung an kultur- und geisteswissenschaftliche Grundströmungen zu suchen. Auf Seiten des Interfacedesigns ergibt diese Skizze ein breites Arbeits- und Entwicklungsprogramm, das auch schon durch partielle Implementierungen neue Akzente gesetzt hat - sowohl im Bereich der Evaluierung von Sammlungsinterfaces, als auch in der Gestaltung und Umsetzung. Wir gehen davon aus, dass Aspekte dieser Designprinzipien auch auf andere Forschungssysteme der Digital Humanities übertragbar sind, jedoch von kritischen Kompetenzen (Literacies) der Akteure, sowie durch ihre konstitutive Selbstkritik als ForscherInnen oder BeobachterInnen ergänzt werden müssen, um ein transparent ineinandergreifendes und sich selbst korrigierendes und entwickelndes Ensemble von Vermittlungen zu erreichen.

Danksagung

Die Arbeit wurde zum Teil durch den Wissenschaftsfonds FWF P.Nr. P28363 gefördert.

Fußnoten

1. Diese Studie erscheint im Journal “Transactions of Visualization and Computer Graphics” unter dem Titel “Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges" (Windhager F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E.).

Bibliographie

Arias-Hernandez, R., Green, T. M., & Fisher, B. (2012). From cognitive amplifiers to cognitive prostheses: Understandings of the material basis of cognition in visual analytics. *Interdisciplinary Science Reviews*, *37*(1), 4–18.

Bubenhofer, N. (2016): Drei Thesen zu Visualisierungspraktiken in den Digital Humanities. In *Rechtsgeschichte - Legal History*, *24*, S.351-355.

Butler, J. (2001). Was ist Kritik? Ein Essay über Foucault's Tugend. *EIPCP, 5* (2001).

Calhoun, C. (1995). *Critical Social Theory: Culture, History, and the Challenge of Difference*. Cambridge, Mass: Wiley-Blackwell.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). Using Vision to Think, chapter 1: Information Visualization, pages 1–34. Morgan Kaufmann.

vDörk, M., Carpendale, S., & Williamson, C. (2011). The information flaneur: A fresh look at information seeking. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1215–1224). ACM.

Dörk, M., Feng, P., Collins, C., & Carpendale, S. (2013). Critical InfoVis: exploring the politics of visualization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2189–2198). ACM.

Dörk, M., Pietsch, C., & Credico, G. (2017). One view is not enough. High-level visualizations of a large cultural collection. *Special Issue of Information Design Journal, 23(1):39–47*.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, *5*(1), 1–21.

Foucault, M. (1992). *Was ist Kritik?* Berlin: Merve.

Glinka, K., Meier, S., & Dörk, M. (2015). Visualising the» Un-seen «: Towards Critical Approaches and Strategies of Inclusion in Digital Cultural Heritage Interfaces. In C. Busch & J. Sieck (Eds.), *Kultur und Informatik (XIII) - Cross Media* (pp. 105–118). Berlin: Werner Hülsbusch.

Hooland, S. van, & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata.* Facet Publishing.

Jaeggi, R., & Wesche, T. (2013). *Was ist Kritik?* Frankfurt: Suhrkamp.

Jones, A. L. (1993). Exploding Canons: The Anthropology of Museums. *Annual Review of Anthropology*, *22*(1), 201–220.

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, *30*(2), 225–248.

Swartz, D. (2012). *Culture and power: The sociology of Pierre Bourdieu*. University of Chicago Press.

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, *9*(1).

Windhager F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2017). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *Transactions of Visualization and Computer Graphics* (eingereicht)

Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches' Semantic Web

Nantke, Julia
nantke@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Schlupkothen, Frederik
schlupko@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Kontext und Zielsetzung

Die Modellierung textueller und transbiblionomer Relationen mithilfe von Semantic Web-Technologien bildet mittlerweile eines der zentralen Forschungsfelder der Digital Humanities. Die Struktur und Funktionsweise literarischer Texte erfordern in Bezug auf die formale Beschreibung, Erklärung und Kategorisierung von semantischen Strukturen ein besonders differenziertes Vorgehen: Interne und externe textuelle Beziehungen bestehen in Form komplexer, häufig ambiger Zeichenrelationen, die plurale, sich auf verschiedenen Ebenen überlagernde Bedeutungsangebote stiften. Letztere können zudem nicht auf verort-

bare Ereignisse, stabile Relationen zwischen (bibliografischen, historischen) Artefakten oder ein konkretes argumentatives Ziel bezogen werden. Die Kategorisierung literarischer ‚Daten' erfolgt deshalb systematisch im Spannungsfeld zwischen dem Text als linguistisch-materieller Zeichenformation und deren interpretierender Auffassung. Die Aktualisierung bestimmter Codes hängt hierbei immer auch von kulturellen und historischen Faktoren, methodologischen Vorannahmen sowie der Kontingenz interpretativer Schlussfolgerung ab. Die daraus resultierende Bedingtheit und potentielle Vielfalt semantischer Zuschreibungen muss daher in einer Modellierung transparent abbildbar sein. [1] Dieser Herausforderung begegnet das Projekt, welches in dem Beitrag vorgestellt wird, indem auf der Basis eines situationstheoretischen Formalismus (Barwise/Perry 1983, Devlin 1990) ein mehrstufiges Modell zur Abbildung und Beschreibung komplexer intertextueller Relationen zwischen literarischen Texten entwickelt wird.

Das Projekt möchte damit in zweierlei Hinsicht einen Beitrag zur methodologischen Reflexion leisten: Zum einen streben wir mit dieser zunächst auf das genauere Verständnis intertextueller Phänomene gerichteten stufenweisen Modellierung [2] einen literaturtheoretisch reflektierten Einsatz digitaler Methoden an. Zum anderen trägt das Vorgehen bei der Modellierung ebenso zu einer Schärfung der literaturwissenschaftlichen Perspektive auf Intertextualität und zur fundierten Beschreibung hierbei wirksamer Faktoren bei. Ziel des Projekts ist eine maschinenlesbare Systematisierung intertextueller „Schreibweisen" (Verweyen/Wittig, S. 38) [3] sowie der Kriterien zur Isolierung und Charakterisierung der Schreibweisen. Diese soll eine computergestützte Erschließung literarischer Intertextualität ermöglichen.

Der Beitrag diskutiert zum einen konkrete Probleme, welche sich im Spannungsfeld zwischen literarischer Polysemie, der Literaturwissenschaft inhärenter Perspektivenvielfalt und technischer Normierung ergeben, denn das Projekt dient nicht zuletzt auch der Reflexion der Möglichkeiten zur Formalisierung literaturwissenschaftlicher Erkenntnisse sowie dem Ausloten der Grenzen für den Einsatz formaler Beschreibungssprachen im Hinblick auf literaturwissenschaftliche Forschungsfragen.

Zum anderen wird dargestellt, wie durch die spezifische Anlage der Modellierung auf verschiedenen Ebenen Desideraten bisheriger Ansätze begegnet und gleichzeitig ein Beitrag zum literaturtheoretisch fundierten Einsatz von DH-Methoden geleistet werden kann.

Erste Ergebnisse werden in dem vorgeschlagenen Beitrag anhand konkreter Beispiele wie etwa des intertextuellen Netzes um Matthias Claudius' *Rheinweinlied* präsentiert, welches aufgrund der dem Netz inhärenten Vielzahl intertextueller Phänomene bei gleichzeitig relativ kurzen Texten hierfür besonders geeignet erscheint.

Stand der Forschung und Abgrenzung

In vielen bisherigen Projekten zur Erfassung literarischer Beziehungen schränkt die Konzentration auf automatisierbare Analysevorgänge das heuristische Potential digitaler Modellierung für die literaturwissenschaftliche Forschung in verschiedener Hinsicht ein:

Erstens werden Modelle zur Beschreibung (innertextueller) literarischer Strukturen an stark Plot-lastigen Texten entwickelt, [4] was die Herausforderung der vielfältigen Bedeutungsebenen komplexerer literarischer Texte deutlich reduziert. Die Modelle erscheinen deshalb kaum auf den Großteil der literaturwissenschaftlich relevanten Beispiele übertragbar.

Zweitens erfolgt eine Modellierung intertextueller Beziehungen anhand einer „historische[n] Positivtät von Kontext-Dokumenten" (Wagner/Mehler/Biber 2016, S. 90 mit Bezug auf das Projekt Wikidition), deren Verknüpfungen auf linguistischer Ebene modelliert werden. Auf diese Weise werden zwar viele Probleme im Hinblick auf die Intersubjektivierbarkeit der Modellierung und die Differenzen zwischen verschiedenen Intertextualitätskonzepten vermieden, gleichzeitig wird aber aus literaturwissenschaftlicher Sicht die Aussagekraft der Ergebnisse stark eingeschränkt, indem der literaturwissenschaftlich relevante Fokus auf die Kategorisierung, Funktion und Wirkung von Intertextualität und die hierbei produktiven Schreibweisen und Markierungen (vgl. Kocher 2007, 179) zugunsten einer eher enzyklopädischen Perspektive verloren geht. [5]

Unser Projekt richtet sich hingegen weder auf die automatisierte Textanalyse noch beschränkt es sich auf die linguistische Ebene konkreter Wortäquivalenz. Vielmehr steht die Entwicklung eines formalisierten Vokabulars zur semantischen Repräsentation literarischer Intertextualität im Zentrum des Forschungsinteresses. Die Isolierung und formale Beschreibung intertextueller Phänomene dient der Beobachtung und Darstellung des Zusammenwirkens jener Schreibweisen und Markierungen bei der Erzeugung von Intertextualität. Die intertextuellen Beziehungen werden dabei also nicht deduktiv im Sinne einer

Qualifizierung als Parodie, Kontrafaktur, Nachahmung, Hommage etc. modelliert, da derartige ‚Gattungszuschreibungen' in systematischen literaturwissenschaftlichen Untersuchungen zur Typisierung von Intertextualität oftmals den Blick für die spezifischen, bei der Erzeugung von Intertextualität wirksamen Faktoren verstellen. Die angeführten literarischen Beispiele dienen dann eher der selektiven Untermauerung der jeweils präfigurierten Typologie (vgl. einschlägig Broich/Pfister 1985; Genette 1993). Im Gegensatz dazu bildet die Modellierung von Beziehungen zwischen konkreten Schreibweisen den Ausgangspunkt unseres Projekts, welcher im Anschluss Schlussfolgerungen über die Relationen der verschiedenen Ebenen, auf denen intertextuelle Verknüpfungen stattfinden, sowie über die jeweils erzeugten Wirkungen ermöglichen soll.

Methodik

Für das angestrebte Forschungsziel stellt die mehrstufige Modellierung einen expliziten heuristischen Gewinn gegenüber bisherigen Ansätzen der Systematisierung dar, indem die umfassende induktive Erfassung und Beschreibung sowie die anschließend abgeleiteten strukturellen Konstanten nicht unverbunden nebeneinander stehen, sondern Mikro- und Makrostrukturen durch das Modell in ihrem Zusammenhang beobachtbar gemacht werden (s. Abbildung 1).

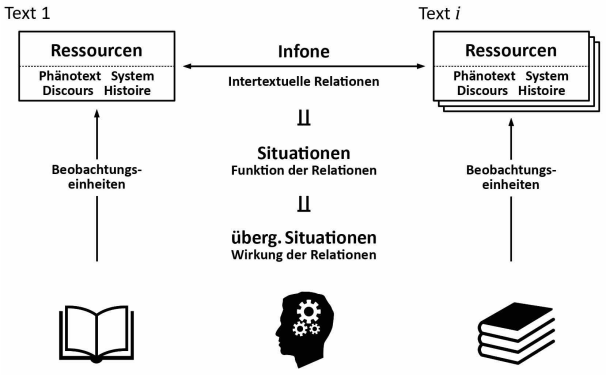


Abbildung 1: Konzeptuelle Darstellung des Modells

Indem zunächst Einzeltextphänomene modelliert, darauf aufbauend deren Funktionen im Sinne ihres Beitrags zur „Bedeutungskonstitution" (Hempfer 1991, S. 19) erfasst und daraus übergreifende Kategorien abgeleitet werden, können zwei bislang getrennt voneinander verhandelte Bereiche der Untersuchung von Intertextualität in einer ganzheitlichen Modellierung verbunden werden: die ‚Entwirrung' des intertextuellen Gefüges eines einzelnen Textes (vgl. hierfür exemplarisch Bauer Lucca 2001; Dudzik 2017) sowie die übergeordnete Suche nach gemeinsamen Strukturen und Funktionsweisen intertextueller Verweise. Das vorgestellte Modell verknüpft also die in der Literaturwissenschaft seit den 1980er Jahren unternommenen Bestrebungen zur Typologisierung intertextueller Strukturen und Funktionsweisen mit einer umfassenden Detailuntersuchung literarischer Texte. Die Differenzierung in Phänomenbeschreibung und -bewertung, welche der eingesetzte Formalismus unterstützt (s. u.), sieht explizit die Modellierung funktionaler Überlagerungen und alternativer Forschungsmeinungen vor, sodass im Rahmen der Formalisierung sowohl der Multifunktionalität intertextueller Schreibweisen (vgl. Kocher 2010, S. 179) als auch der maßgeblich auf produktivem Dissens basierenden Dynamik des literaturwissenschaftlichen Diskurses Rechnung getragen wird.

Als Ausgangspunkt zur formalen Beschreibung intertextueller Phänomene dient ein situationstheoretischer Ansatz, welcher die Brücke zwischen literaturwissenschaftlicher Analyse und technischer Modellierung darstellt. Die Situationstheorie bietet sich an, da sie im Sinne der angestrebten Beschreibung der Intertextualität einen mehrstufigen Formalismus zur Verfügung stellt, welcher Informationen und Informationsflüsse in Kontextabhängigkeit beschreibt: Basale Phänomene werden durch basale Informationseinheiten (sog. „Infone") beschrieben, welche wiederum „Situationen" als Phänomene einer höheren Ordnung aus der Perspektive eines oder mehrerer „Agenten" zu bilden erlauben.

Somit kann formal unterschieden werden zwischen der Modellierung konkreter (sprachlicher, inhaltlicher, stilistischer) Texteigenschaften und -relationen (beschrieben als Infone) und der Klassifizierung der modellierten Informationseinheiten im Sinne ihrer Funktion sowie einer durch sie indizierten, Kontext-abhängigen Wirkung (beschrieben als Situationen). Im Gegensatz zu technischen Beschreibungssprachen (wie etwa RDF oder OWL) liefert die Situationstheorie einen Formalismus, welcher zunächst frei von umsetzungsspezifischen Einschränkungen (wie etwa festgelegten Datentypen oder Objekthierarchien) ist, die sich ungewollt perspektivierend auf die Modellierung auswirken können. [6] Für die Beschreibung literarischer Texte erweist sich der situationstheoretische Formalismus also als besonders geeignet, da er Modellierungsfreiheit mit der für die technische Umsetzung notwendigen formalen Strenge vereint.

Ausblick

Das Modell wird sukzessive unter Einbezug literarischer Texte verschiedener Textsorten und Publikationszeiträume getestet und weiterentwickelt. An diese sukzessive formale Strukturierung anknüpfend wird geprüft, inwieweit etablierte Beschreibungssprachen bei einer technischen Umsetzung des Modells Anwendung finden können. Insbesondere etablierte Sprachen aus dem Umfeld elektronischer Publikation sollen auf ihre Anwendbarkeit bzw. Möglichkeiten der Erweiterung hin betrachtet werden. Dies sind im Rahmen der durch das W3C beschriebene Standards für Verweisstrukturen Sprachen wie XPath, XLink oder XPointer (vgl. einschlägig Wilde/Lowe 2003), für semantische Auszeichnungen die Sprachen des Semantic Web wie RDF oder OWL. Dies schließt – unabhängig von der konkreten Sprache – die Berücksichtigung unterschiedlicher Auszeichnungskonzepte wie bspw. Standoff- in Abgrenzung zu Inline-Markup ein (vgl. Banski 2010). [7]

Fußnoten

1. Meister 2012, S. 112 betont aufgrund der Dynamik und historisch-kulturellen Dependenz von Sprache zurecht, dass „[t]he problems posed by the interpretation of literary texts are thus not an exceptional, but rather an exemplary case“. Dennoch erfolgt literarische (bzw. allgemein künstlerische) Kommunikation im Rahmen einer spezifischen „Kommunikationslogik“, die sich von der alltagssprachlicher Kommunikationssituationen unterscheidet (vgl. Spoerhase 2007, S. 414–418).
2. Vgl. zur Unterscheidung zwischen „modeling for understanding“ und „modeling for production“ Eide 2014.
3. „Schreibweisen“ ist hierbei nicht intentionalistisch, sondern im Sinne von Textstrukturen mit Verweisfunktion zu verstehen.
4. Vgl. hierzu u. a. die Ausführungen zur visuellen Analyse in John u. a. 2016. Einen komplexeren, narratologisch ausgerichteten Ansatz verfolgt das Projekt heureCLÉA (vgl. hierzu Gius/Jacke 2015). Das Projekt ist daher in seiner Orientierung an konkreten literaturwissenschaftlichen Methoden beispielhaft, verfolgt aber mit seiner Ausrichtung auf innertextuelle Strukturen ein grundsätzlich anderes Ziel als unser Projekt.
5. Wagner/Mehler/Biber 2016, S. 90 verstehen ihr Projekt daher auch eher im Sinne einer Vorstufe für die „Erschließung des intertextuellen Potentials eines je gegebenen literarischen Texts“. Ihr methodischer Ansatz „zielt nicht auf die *Implementierung* literarischer Intertextualität“.
6. Im Gegensatz dazu ist bspw. der in Heßbrüggen-Walter 2015 dargestellte Ansatz unmittelbar RDF-basiert gedacht.
7. Zahlreiche rechnergestützte, aber proprietäre Anwendungen wurden im Verlauf der Entwicklung der Situationstheorie vorgestellt (vgl. einschlägig etwa Tin/Akman 1994). Neuere Ansätze schlagen die Verwendung etablierter Sprachen, wie insb. durch das Semantic Web gegeben, vor (Kokar/Matheus/Baclawski 2009).

Bibliographie

Banski, Pjotr (2010): “Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless”, in: *Proceedings of Balisage: The Markup Conference* 2010.

Barwise, Jon / Perry, John (1983): *Situations and Attitudes*. Cambridge: Bradford Book, MIT Press.

Bauer Lucca, Eva (2001): *Versteckte Spuren: eine intertextuelle Annäherung an Thomas Manns Roman “Doktor Faustus”*. Wiesbaden: Deutscher Universitätsverlag.

Broich, Ulrich / Pfister, Manfred (eds.) (1985): *Intertextualität. Formen, Funktionen, anglistische Fallstudien*. Tübingen: Niemeyer.

Devlin, Keith J. (1990): Logic and Information. Cambridge: Cambridge University Press.

Dudzik, Yvonne (2017): *Geschichten bereichern die Geschichte: Intertextualität als Untersuchungskategorie in Uwe Johnsons “Jahrestage”*. Göttingen: V&R unipress.

Eide, Øvind (2014): “Ontologies, Data, and TEI”, in: *Journal of the Text Encoding Initiative* 8: 1–22.

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

Gius, Evelyn / Jacke, Janina (2015): “Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse”, in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1) 10.17175/sb001_006.

Hempfer, Klaus W. (1991): „Intertextualität. Systemreferenz und Strukturwandel: Die Pluralisierung des erotischen Diskurses in der italienischen und französischen Renaissance-Lyrik (Ariost, Bembo, Du Bellay, Ronsard)“, in: Titzmann, Michael (ed.): *Modelle des literarischen Strukturwandels*. Tübingen: Niemeyer, 7–43.

Heßbrüggen-Walter, Stefan (2015): “What People Said: The Theoretical Foundations of a Minimal Doxographical Ontology and Its Use in the History of Philosophy”, in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten*

der Digital Humanities. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). DOI: 10.17175/sb001_001

John, Markus / Lohmann, Steffen / Koch, Steffen / Wörner, Michael / Ertl, Thomas (2016): „Visual Analytics for Narrative Texts. Visualizing Characters and their Relationships as Extracted from Novels", in: *Proceedings of the 7th International Conference on Information Visualization Theory and Applications*. Rom, Italien: SciTePress [Preprint: http://www.visualdataweb.org/publications/2016_IVAPP_VA-for-Narrative_preprint.pdf].

Kocher, Ursula (2007): "Im Gewirr der Fäden: Intertextualitätstheorie und Edition", in: Falk, Rainer / Mattenklott, Gert (eds.): *Ästhetische Erfahrung und Edition*. Tübingen: Max Niemeyer, 175–185.

Kokar, Mieczyslaw M. / Matheus, Christopher J. / Baclawski, Kenneth (2009): „Ontology-based situation awareness", in: *Information Fusion* 10, 1. St. Louis, MO, USA: Elsevier, 83–98.

Tin, Erkan / Akman, Varol (1994): „Computational Situation Theory", in: Hoebel, Louis J. / Powers, David (eds.): *SIGART Bulletin* 5, 4. New York, NY, USA: ACM, 4–17.

Meister, Jan-Christoph (2012): „Crowdsourcing ‚True Meaning': A Collaborative Markup Approach to Textual Interpretation", in: Deegan, Marilyn (ed.): *Collaborative Research in the Digital Humanities. A Volume in Honour of Harold Short, on the Occasion of his 65th Birthday and his Retirement*. Farnham: Ashgate 2012, 106–122.

Spoerhase, Carlos (2007): *Autorschaft und Interpretation. Methodische Grundlagen einer philologischen Hermeneutik*. Berlin/New York: De Gruyter.

Verweyen, Theodor / Wittig, Gunther (2010): *Einfache Formen der Intertextualität: Theoretische Überlegungen und historische Untersuchungen*. Paderborn: mentis.

Wagner, Benno / Mehler, Alexander / Biber, Hanno (2016): "Transbiblionome Daten in der Literaturwissenschaft. Texttechnologische Erschließung und digitale Visualisierng intertextueller Beziehungen digitaler Korpora", in: *Konferenzabstracts DHd 2016 Modellierung, Vernetzung, Visualisierung. Die Digital Humanties als fächerübergreifendes Forschungsparadigma* http://dhd2016.de/boa.pdf , 88–94.

Wilde, Erik / Lowe, David (2003): *XPath, XLink, XPointer, and XML: A Practical Guide to Web Hyperlinking and Transclusion*. Boston, MA: Addison-Wesley.

Poster

Ambraser Heldenbuch: Transkription und wissenschaftliches Datenset

Sojer, Claudia
claudia.sojer@uibk.ac.at
Universität Innsbruck, Österreich

Tratter, Aaron Rudolf
aaron.tratter@student.uibk.ac.at
Universität Innsbruck, Österreich

Seit Januar 2017 arbeitet eine Forschungsgruppe an der Universität Innsbruck unter der Leitung von Mario Klarer an dem ÖAW-go!digital-Projekt »Ambraser Heldenbuch: Transkription und wissenschaftliches Datenset«. Das Forschungsprojekt setzt sich zum Ziel, bis zum Jahr 2019 – dem 500. Todestag von Kaiser Maximilian I. – das *Ambraser Heldenbuch (AHB)* (Wien, Österreichische Nationalbibliothek, Cod. Ser. nova 2663) zur Gänze zu transkribieren und als Forschungsdatenset online und offline öffentlich zugänglich zu machen.

Eine Reihe bedeutender Texte der mittelalterlichen deutschen Literatur ist ausschließlich in frühneuhochdeutscher Sprache im *AHB* überliefert (wie etwa Hartmanns von Aue *Erec*, *Kudrun*, *Moriz von Craûn*, etc.). Das *AHB* wurde zu Beginn des 16. Jahrhunderts von Kaiser Maximilian I. als Prunkhandschrift in Auftrag gegeben und vom Bozner Zollschreiber Hans Ried auf ca. 500 großformatigen Pergamentseiten ausgeführt. Die meisten Editionen dieser unikal überlieferten Texte sind jedoch Rückübersetzungen in ein standardisiertes Mittelhochdeutsch, wodurch die sprachliche Form des einzigen Textzeugens in den Hintergrund gerät. Aus diesem Grund mehren sich seit vielen Jahren Stimmen, die einer Gesamttranskription des *AHB* höchste Priorität zusprechen (z. B. Leitzman 1935; Gärtner 2006; Mura 2008).

Das *AHB* ist mit 25 wichtigen mittelalterlichen literarischen Erzähltexten in einer Hand, wovon 15 als Unikate ausschließlich im *AHB* überliefert sind, der **umfangreichste Kodex** (ca. 500.000 Wörter) seiner Art.

In einer Hand bzw. von einem **einzelnen Schreiber** verfasst stellt dieses Textkorpus eine exzellente Materialbasis für weitere literaturhistorische und sprachwissenschaftliche Untersuchungen dar.

Die **Sprachform** der von Hans Ried niedergeschriebenen Texte deckt sich nicht mit dem (standardisierten) Mittelhochdeutsch seiner Vorlagen aus dem 12. und 13. Jahrhundert. Im *AHB* manifestiert sich eine offensichtlich hybride literarische Kunstsprache des frühen 16. Jahrhunderts, die sich von den anderen überlieferten Autographen Rieds (aus einem dezidiert nicht literarischen Kontext) abhebt.

Gerade für die unikal im *AHB* überlieferten Texte sehen Homeyer und Knor (2015) das große Potential einer digitalen Gesamttranskription: „[F]ehlt doch bisher die Gesamtschau auf den Schreibusus Rieds im Rahmen seiner Abschrift des ‚Ambraser Heldenbuches', um mögliche Vorlagenreflexe von Texteingriffen, Wortschatzwandel oder individuellen Schreibgewohnheiten zu trennen" (98). Damit unterstützt die Transkription **Editionsbemühungen von Einzeltexten** des *AHB*.

Aufgrund der oben angeführten Gründe ermöglicht eine **allographische Transkription**, die Richtlinien anwendet, die im Rahmen des Projektes „Ambraser Heldenbuch: Transkription und wissenschaftliches Datenset" an der Universität Innsbruck eigens erstellt wurden, um die Schreibweisen Hans Rieds möglichst präzise wiedergeben zu können, eine Untersuchung der transkribierten Texte in vielfältiger Weise. So können etwa Idiosynkrasien Hans Rieds leichter identifiziert und schnelle Abfragen einzelner Grapheme durchgeführt werden. Zusätzlich wird jedoch auch eine **diplomatische Transkription** des gesamten *AHB* erarbeitet und online veröffentlicht. Bei der allographischen Transkription werden möglichst viele verschiedene Varianten der Grapheme unterschieden. Ein wichtiges Kriterium für die Zuordnung zu einer Variante sowohl bei den Buchstaben als auch bei den Superskripta spielt die Federführung Hans Rieds. Nur sehr wenige Transkriptionen bilden die verschiedenen Varianten der Grapheme in den Texten ab, sodass in diesem Bereich der Forschung noch Aufholbedarf besteht.

Da die Transkription nicht nur auf Buchstabenebene, sondern auch auf Zeichenebene erfolgt, entsteht dadurch ein differenziertes und möglichst präzises Abbild des Manuskriptes. Die Grapheme werden durch unterschiedliche Schriftzeichen des Unicode-Zeichensatzes dargestellt. In der folgenden Abbildung sieht man einen Ausschnitt aus dem Nibelungenlied. In den Wörtern *Ross*, *des* und *Seyfrids* tritt mit einem langen s eine kontextsensitive Variante des Buchstabens *s* auf. Außerdem treten drei unterschiedliche graphische Varianten des runden s auf.

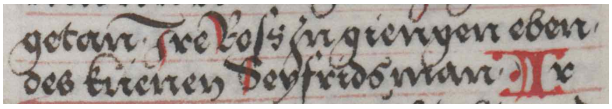


Abb. 1: f. 95vc ll. 7–8 ab imo

getan / Jre Roſs jn giengen eben /
deꝫ küenen Seyfridσ man / Ir

Abb. 2: Transkription nach den neuesten Transkriptionsrichtlinien

Die einzelnen Varianten der Buchstaben weisen eine relativ geringe Varianz auf. Anders verhält es sich jedoch bei den diakritischen Zeichen bzw. Superskripta. Es werden vier verschiede Superskripta unterschieden: Trema, Breve, Superskriptum o und Superskriptum a. Vor allem das Breve und das Superskriptum o ähneln sich teilweise recht stark, sodass rein graphisch nicht immer bestimmt werden kann, um welches Superskriptum es sich handelt. Neben der Federführung des Schreibers werden andere Kriterien zur Entscheidung herangezogen, beispielsweise phonetische Merkmale.

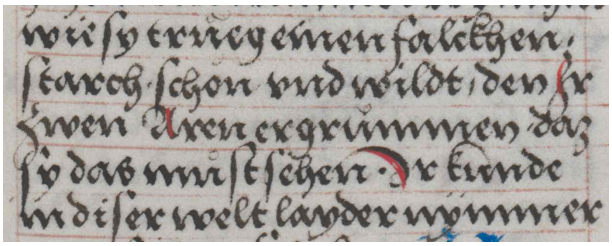


Abb. 3: f. 95ra ll. 17–21 ab imo

wie ſy trŭeg einen Falckhen /
ſtarch / ſchŏn vnd wildt / deŋ jr
zwen Aren ergrŭmmeŋ / daȝ
ſy daꝫ můſt ſehen · Ir kunde
in diſer welt layder nÿmmer

Abb. 4: Transkription nach den neuesten Transkriptionsrichtlinien

Durch **Verlinkung** des Manuskriptbildes mit dem Transkript auf Zeilenebene wird ermöglicht, dass die Überprüfbarkeit der Transkription bei späteren Forschungsprojekten stets gegeben ist. Man kann sich selbst ein Urteil über die vorliegende Arbeit bilden und diese dann gegebenenfalls revidieren oder zusätzliche Möglichkeiten der Interpretation aufzeigen.

Neben der Transkription erfolgt mittels **Annotation** die Auszeichnung übergeordneter Strukturen wie Verse und Strophen. Es werden aber auch Initialen, Lombarden und Rubrizierungen ausgewiesen, da diese in der Transkription nicht ausreichend dargestellt werden können. Darüber hinaus werden auch Zweifelsfälle getaggt, um darauf hinzuweisen, dass in diesem Fall die Schrift nicht eindeutig identifiziert werden konnte. So wird Transparenz und Überprüfbarkeit der Transkription gewährleistet.

Bibliographie

Gärtner, Kurt et al., eds. Hartmann von Aue. *Erec*. 7th ed. Tübingen: Niemeyer, 2006.

Homeyer, Susanne, and Inta Knor. „Zu einer umfassenden Untersuchung der Schreibsprache Hans Rieds im Ambraser Heldenbuch.“ *Zeitschrift für Deutsche Philologie* 134.1 (2015): 97-103.

Leitzmann, Albert. „Die Ambraser Erecüberlieferung.“ *Beiträge zur Geschichte der deutschen Sprache und Literatur* 59 (1935): 143-234.

Mura, Angela. „Spuren einer verlorenen Bibliothek: Bozen und seine Rolle bei der Entstehung des Ambraser Heldenbuchs.“ *Cristallîn Wort. Hartmann-Studien. Rahmenthema: Das Ambraser Heldenbuch*. Ed. Waltraud Fritsch-Rößler. Wien: Lit-Verlag, 2008. 59-128.

Annotationen anhand der Gemeinsamen Normdatei aus einer anwendungsorientierten Perspektive historischer Forschung

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

Mache, Beata
mac@steinheim-institut.org
SUB Göttingen, Deutschland

Gemeinsame Normdatei

Die Anwendung von Normdaten und kontrollierten Vokabularen, die über einfache Glossare hinausgehen, gewinnt für die Geisteswissenschaften zunehmend Bedeutung: digitale Taxonomien, Thesauri, Ontologien. Die Posterpräsentation erläutert die Linked-Data-Strategie des Steinheim-Instituts (STI): die Annotation aller digitalen Angebote mittels der Gemeinsamen Normdatei (GND), die dafür entwickelten Tools, den Mehrwert der Teilnahme an der sich herausbildenden dezentralen Infrastruktur, das Weiterentwicklungspotenzial aus Praxisperspektive, schließlich die Notwendigkeit und Möglichkeit des schreibenden Zugriffs von Projektmitarbeitern auf die Normdatei sowie erste Erfahrungen damit.

Die GND ist eine Sammlung von Normdatensätzen zu Personen, Körperschaften, Konferenzen, Geografika, Sachschlagwörtern und Werktiteln. Sie wird im Bibliotheksbereich kooperativ erstellt und gepflegt und ist dort insbesondere im Zusammenhang der Katalogisierung etabliert. Der Zugriff auf die GND und ihre Integration in digitale Anwendungen ist schrankenlos möglich: Sie steht unter der Lizenz CC0, wird in verschiedenen Formaten angeboten (u.a. RDF) und ist über unterschiedliche Schnittstellen ansprechbar (u.a. OAI-PMH). Zudem stehen weitere Services wie BEACON Findbuch und Entity Facts bereit.

Identifikation und Disambiguierung durch Verschlagwortung mittels der GND einschließlich ihrer internen Relationen erlauben, entsprechende Schnittstellen vorausgesetzt, digitale Anwendungen als Linked Data anzubieten. Dies wird zunehmend in Webanwendungen von Bibliotheken, Kultureinrichtungen und eben auch Forschungsprojekten genutzt. Auch außerhalb der engeren Bibliothekssphäre findet man längst eine ausgeprägte und erfolgreiche Praxis – allerdings selten besprochen und diskutiert.

Verteilte Infrastruktur

Im Umfeld der Wikipedia ist eine dezentrale Recherche-Infrastruktur entstanden, die ein solches Linked-Data-Netz aufspannt. Eine Projektseite der deutschen Wikipedia (Wikipedia:BEACON) dient als ‚Collection Registry', in die entprechend ausgerüstete Angebote mit ihren BEACON-Dateien eingetragen werden. Diese BEACON-Dateien können frei genutzt werden (Public Domain), und erlauben die Vernetzung untereinander. Der Aggregator BEACON Findbuch enthält 488 (Stand 14.01.2018) dieser (noch stark biografieorientierten) Sammlungen. Deren „Fact Sheets" liefern üblicherweise Basisinformationen zur recherchierten Entität sowie weiterführende Links.

Für verschiedene Webanwendungen des STI wurde ein eigener Service entworfen und implementiert, der diese Funktionalität fachspezifisch und nutzerzentriert bündelt und nicht nur für Personen, sondern das gesamte Spektrum der GND Antworten liefert (Abb.1).

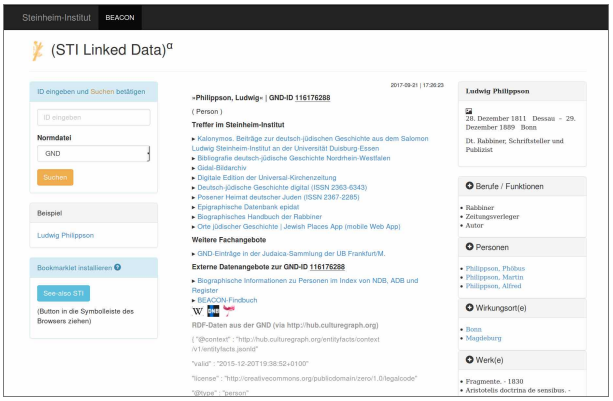


Abbildung 1: STI Fact Sheet

Ein weiteres Tool dient dem Vergleich von BEACON-Dateien zur Ermittlung ihrer Schnittmenge hinsichtlich gemeinsamer IDs (Abb.2).

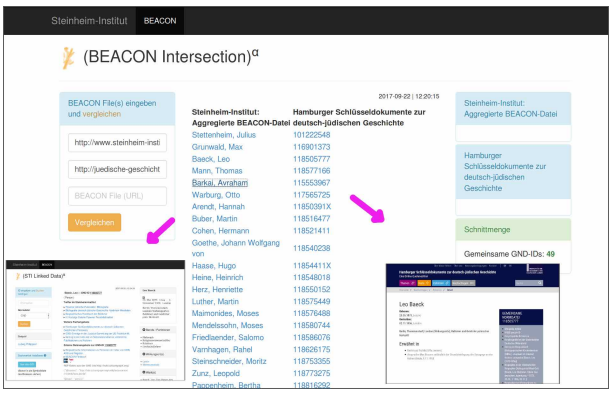


Abbildung 2: Vergleich von BEACON-Dateien

Use Case

Nach der retrospektiven GND-Annotation unterschiedlicher digitaler Webangebote des STI wurde dies Verfahren bei dem Projekt „Posener Heimat in der Literatur und Publizistik deutscher

Juden 1918–1938“ schon bei Antragstellung eingeplant. Die Erschließungsdatenbank zur Zeitschrift „Posener Heimatblätter“ und das Projektblog werden zunächst konsequent anhand interner IDs annotiert und dann, soweit verfügbar, via GND-Verschlagwortung untereinander und extern vernetzt, ebenso die abschließende digitale Publikation.

Schon die laufende Projektarbeit profitiert von diesen Annotationen und den hierdurch erweiterten Recherchemöglichkeiten. Zudem werden durch intensive Recherche in deutschen und polnischen Archiven Informationen gehoben, anhand derer eine substanzielle Ergänzung der GND wünschenswert und möglich ist, durch Neuanlage, Ergänzung und Korrektur von Datensätzen.

Illustrierende Beispiele: Zu der jüdischen Kinderbuchautorin und Dichterin Frieda Mehler, die mit ihren Publikationen nach 1933 aus den deutschen Bibliotheken verbannt wurde, konnten im Projekt alle für einen GND-Personen-Datensatz notwendigen Informationen erhoben werden (Abb.3). Und zu dem Publizisten, Historiker und Kunstsammler Arthur Kronthal fand sich in der GND zwar ein Personen-Datensatz, aber ohne Lebensdaten und ohne Verknüpfung mit seinen gleichwohl katalogisierten Publikationen. Kronthal verbrachte seine letzten Lebensjahre in einem Siechenheim und verstarb im Jüdischen Krankenhaus in Berlin 1941. Er wurde fälschlich als *Adolph* Kronthal im Sterberegister verzeichnet. Durch wissenschaftliche Archivrecherche und intellektuellen Abgleich der Daten gelang seine für die GND relevante Identifikation.

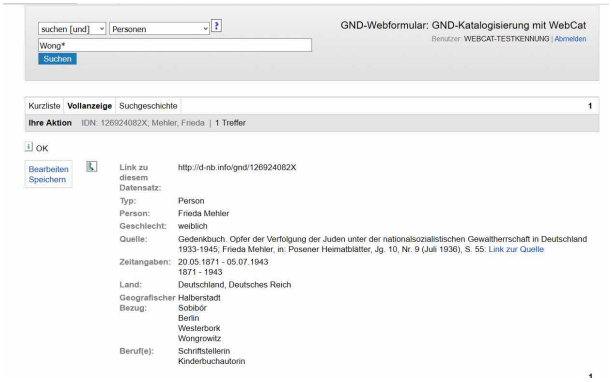


Abbildung 3: Eintrag zu Frieda Mehler im GND-Webformular

Folgerungen aus der Projektpraxis

Die bisher übliche Anwendung des GND/BEACON-Verfahrens ist schwerpunktmäßig biografie- und metadatenbezogen. Darüber hinaus erscheint jedoch die ausweitende Verschlagwortung etwa mit „Sachbegriffen“ oder „Organisationen“ sowie die tiefe Erschließung, d.h. die durchgängige Annotation von Volltexten (und nicht nur ihrer Metadaten) für die historische Forschung sehr attraktiv.

Die projektbezogene, zur bestmöglichen Verlinkung sinnvolle Auswertung aller oder passender BEACON-Dateien könnte unterstützt werden durch eine maschinenlesbare(re) ‚Collection Registry‘. Entsprechend wären Ansätze zu fördern, die darauf zielen, die „Weiternutzung der Daten zu vereinfachen“ (Wikipedia:BEACON). Auch die seitens DARIAH-DE angebotene „Collection Registry“ ist durch ihre Unterstützung von BEACON-Dateien geeignet.

Der hohe Stellenwert von Kooperation in den Digital Humanities wird an solchen Linked-Data-Projekten besonders deutlich. Je mehr Forschende mitmachen, desto größeren Nutzen ziehen alle Beteiligten aus diesem Netzwerk. Das gilt auch für die GND: Erst ein auch schreibender Zugriff auf die GND erlaubt einen ‚runden‘ Workflow und die adäquate Veröffentlichung des Projektwissens – und alle Normdatei-Anwender profitieren davon. Auch hier sind Fortschritte zu berichten: Nach Vereinbarung einer Kooperation mit einer Redaktionsstelle und der Registrierung bei der Deutschen Nationalbibliothek ist externen Akteuren die Neuanlage und Änderung von Personeneinträgen über das „GND-Webformular“ möglich. (Hartmann 2017)

Bibliographie

BEACON Findbuch http://beacon.findbuch.de/ [letzter Zugriff 14. Januar 2018]

Gemeinsame Normdatei (GND) http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html [letzter Zugriff 14. Januar 2018]

Danowski, Patrick / Pohl, Adrian (Hg.) (2013): *(Open) Linked Data in Bibliotheken* (Bibliotheks- und Informationspraxis 50), Berlin.

Hartmann, Sarah (2017): GND-Webformular – eine neue Schnittstelle für die GND https://wiki.dnb.de/download/attachments/125420735/2-4_%20GND-Webformular_Hartmann.pdf?version=1&modificationDate=1501514388000&api=v2 [letzter Zugriff 14. Januar 2018]

Informationsseite zur GND https://wiki.dnb.de/display/ILTIS/Informationsseite+zur+GND [letzter Zugriff 14. Januar 2018]

Lordick, Harald (2015): »BEACON – »Leuchtfeuer« für Online-Publikationen«, in: *Deutsch-jüdische Geschichte digital,* 17. Mai 2015, https://

djgd.hypotheses.org/672 [letzter Zugriff 14. Januar 2018]

Lordick, Harald (2016): Fachspezifische und nutzerzentrierte Perspektiven — Quellen vernetzen mit der Gemeinsamen Normdatei, in: *Deutsch-jüdische Geschichte digital,* 27. November 2016, https://djgd.hypotheses.org/1181 [letzter Zugriff 14. Januar 2018]

Mache, Beata (2015): *Digitale Edition und Erschließung eines interreligiösen Periodikums aus dem Vormärz als editionsphilologische Aufgabe: Universal-Kirchenzeitung für die Geistlichkeit und die gebildete Weltklasse des protestantischen, katholischen und israelitischen Deutschlands (1837),* Duisburg, Essen, Univ., Diss., 2015 urn:nbn:de:hbz:464-20150327-080454-5

Plum, Nathalie Madeleine (2017): Ein Thesaurus für den Naturschutz. Erstellung eines vernetzten Vokabulars für die Literaturdatenbank DNL-online, in: *Natur und Landschaft. Zeitschrift für Naturschutz und Landschaftspflege,* 92 (2017), Nr. 8, S. 356–364.

Wikipedia:BEACON https://de.wikipedia.org/wiki/Wikipedia:BEACON [letzter Zugriff 14. Januar 2018]

Aufdecken von “versteckten” Einflüssen: Teil-Automatisierte Textgenetische Prozesse mit Methoden der Computerlinguistik und des Machine Learning

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

Bruder, Daniel
dmb77@cam.ac.uk
Universität Cambridge, Vereinigtes Königreich

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

Einleitung

Am Beispiel von Ludwig Wittgensteins Nachlass (Pichler et al. 2009; Wittgenstein 1996) wird ein Tool vorgestellt und erläutert, welches die Digital Humanities durch einen computer-unterstützten Prozess für textgenetische Aufgaben bereichern soll.

Ludwig Wittgensteins Nachlass umfasst etwa 20.000 Seiten, in welchen er eine Menge Zitate aus der Weltliteratur verwendet, diese aber nicht notwendigerweise explizit als solche kennzeichnet. Es scheint, als verzichte Wittgenstein auf explizite Quellenangaben bei Autoren, von welchen er annimmt, sie seien Teil eines „allgemeinen” „kulturellen Horizonts“, und lässt bei seinen Zitaten in Fällen wie z.B. Goethe die Namen der Autoren weg.

In einer sinnvollen Zusammenarbeit im Rahmen der Digital Humanities kann die Informatik im Allgemeinen und die Computerlinguistik im Speziellen der Philologie unterstützende Werkzeuge anbieten, die dem Geisteswissenschaftler helfen sollte, derartige mühevolle Prozesse der Zitat-Aufdeckung teil-automatisieren zu können.

In diesem Poster wird ein Tool vorgestellt, welches dem Philologen mit Methoden des Machine Learning beim Aufspüren von Einflüssen aus Zitaten in textgenetischen Prozessen unterstützen kann, indem es erlaubt, „ähnliche” Textabschnitte, d.h. potentielle Zitate, vorzufiltern und zu sortieren, um diese im nächsten Schritt einer genaueren Untersuchung zu unterziehen.

Bezug zum aktuellen Forschungsstand

Die Plagiatsaufdeckung ist die Bestimmung von „ähnlichen“ Texten, d.h. die Aufdeckung von (ungekennzeichneten) Zitaten. Eine Möglichkeit zur Aufdeckung von Ähnlichkeiten ist dabei der Vergleich von syntaktischen Merkmalen zweier oder mehrerer Texte. Diese Charakteristika umfassen beispielsweise die Berücksichtigung von Part-of-Speech Tags, Lemmata und Wortpositionen (Ekbahl et al. 2012). Um zusätzlich modifizierte, aber dennoch semantisch gleiche Texte zu identifizieren, müssen auch Synonyme in Betracht gezogen werden (Abdalgader et al. 2010). Eine Vorverarbeitung, welche u.a. das Tokenisieren der Texte, die Entfernung von Stoppwörtern und eine Sprachidentifizierung beinhaltet, hilft zudem, redundante Wörter zu ignorieren und damit genauere Ergebnisse zu erzielen. Weitere Methoden zur Erkennung von ähnlichen Texten beinhalten u.a.

subjektbasierte Graphen (Tomita et al., 2004) und *Document Fingerprinting* (Sadowski and Lewin, 2007; Kent et al. 2010). Überraschenderweise wurde bislang kein Versuch unternommen, die oben genannten linguistischen Features aus dem Bereich der Syntax und der Semantik zu kombinieren, um die Performanz der Ähnlichkeitssuche weiter zu verbessern. Dass eine solche Kombination sinnvoll ist und besonders gute Ergebnisse leisten kann, zeigt Ullrich (2017).

Forschungsproblem

Solche Ansätze könnten – durch entsprechende Umwidmung und in koordinierter Zusammenarbeit mit Philologen – in den Digital Humanities sinnvoll zur Anwendung gebracht werden, um textgenetische Prozesse teil-automatisiert zu unterstützen und Zeit für intensivere Analysen zu gewähren.

Die Idee ist, das bestehende Tool zur Ähnlichkeitsbestimmung aus Ullrich (2017) so umzufunktionieren, dass nicht nur die Ähnlichkeit *einer* gegebene Texteingabe mit *einer* anderen gegebenen Eingabe bestimmt werden kann, sondern eine *Sortierung* (*ranking*) von ähnlichen Texten vorgenommen werden kann. Zunächst werden dafür die Texte in kürzere Abschnitte – bei Wittgenstein „Bemerkungen“ genannt – geteit. Nun wird eine *Sammlung* dieser Abschnitte mit einer anderen *Sammlung* an Abschnitten verglichen, um dann die potentiell Ähnlichsten in einer Art Hitliste auszugeben. Eine derartige Vorsortierung könnte es dem Philologen besonders erleichtern, potentielle Zitate, Einflüsse und Verweise eines Autors innerhalb seines Werkes und im Bezug auf die Literatur seiner Zeit aufzuspüren.

Es muss betont werden, dass mit einem derartigen Werkzeug die Arbeit des Philologen lediglich unterstützt aber keinesfalls ersetzt werden kann. Im Wesentlichen erlaubt eine teil-automatisierte Vorfilterung von „ähnlichen” Textstellen eine drastische Reduktion des „Suchraums“.

Um die Leistung eines Philologen wie Hans Biesenbach (2014) zu illustrieren: Wittgensteins Nachlass umfasst 20.000 Seiten. Rechnet man mit 5 Bemerkungen pro Seite und die angestrebte Ähnlichkeitssuche beschränkt sich auf 20.000 Seiten der „Weltliteratur” mit ebenfalls 5 Abschnitten pro Seite, dann gäbe es mathematisch 100.000 x 100.000, also 10 Milliarden mögliche Beeinflussungen, die manuell zu prüfen wären. Selbst für besonders leistungsfähige Rechner werden hier Grenzen erreicht, die nach geeigneten NLP Methoden verlangen.

Angewendete Methode

Mit Hilfe computerlinguistischer Methoden berechnet man für jeden Abschnitt eines Textes seinen „charakteristischen” Vektor oder, intuitiv gesprochen, seinen linguistischen „Fingerabdruck”. Dieser automatisierte Prozess kann unabhängig im Voraus berechnet werden um spätere Prozesse zu vereinfachen und beschleunigen. Dieser „Fingerabdruck“ beinhaltet die oben genannten syntaktischen, sowie semantischen Informationen.

Wird eine Suchanfrage zum Auffinden ähnlicher Abschnitte gestartet, lässt sich der charakteristische Vektor des eingegebenen Abschnitts berechnen und daraufhin die Vektoren mit dem geringsten Abstand im multi-dimensionalen Raum bestimmen. Diese Vektoren verweisen auf die ähnlichsten Textabschnitte, die dann dem Philologen zur genaueren Prüfung in einer Hitliste vorgeschlagen werden.

Die bereits erfolgreiche Ähnlichkeits *bestimmung* in Ullrich (2017) soll zu einem Ähnlichkeits *rankingtool*weiterentwickelt werden, um sie vor allem für textgenetische Prozesse in digitalen Editionen nutzbar zu machen. Sobald diese Weiterentwicklung abgeschlossen ist, soll sie die Wittgenstein Advanced Search Tools (Hadersbeck et al. 2014) in der Suchmaschine WiTTFind (siehe: http://wittfind.cis.lmu.de) erweitern, welche am Centrum für Informations- und Sprachverarbeitung der Universität München entwickelt wurde.

Bibliographie

Abdalgader, Khaled / Skabar, Andrew(2010): “Short-text similarity measurement using word sense disambiguation and synonym expansion.”, in: *Australasian Joint Conference on Articial Intelligence*, Springer 435-444.

Biesenbach, Hans (2014): *Anspielungen und Zitate im Werk Ludwig Wittgensteins*, Sofia University Press.

Ekbal, Asif / Saha, Sriparna / Choudhary, Gaurav (2012): “Plagiarism detection in text using vector space model.”, in: *Hybrid Intelligent Systems* (HIS), 2012 12th International Conference on, pages 366-371. IEEE.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Øyvind, L. (2014): *Wittgenstein‘s Nachlass: WiTTFind and Wittgenstein Advanced Search Tools (WAST).* DATeH. Madrid.

C.K. Kent / N. Salim (2010): „Features based text similarity detection.“, in: *Journal of Computing*, 2 (1).

Pichler, Alois / Krüger, Heinz W. / Smith, D. / Bruvik, Tone / Lindebjerg, Anne / Olstad, Ve-

mund (Hrsg.) (2009): *Wittgenstein Source Bergen Facsimile (BTE)*. Wittgenstein Source Bergen.

Sadowski, Caitlin / Levin, Greg (2007): *Simhash: Hash-based similarity detection*.

Tomita, Junji / Nakawatase, Hidekazu / Ishii, Megumi (2004): „Calculating similarity between texts using graph-based text representation model", in: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 248-249. ACM.

Ullrich, Sabine (2017): *Evaluation of Existing Plagiarism Research for the Optimisation of NLP-based Similarity Detection using Ludwig Wittgenstein's Remarks,* Bachelor thesis, Ludwig-Maximilians-Universität München.

Wittgenstein, Ludwig / Nedo, Michael (Hrsg.) (1996): *Wiener Ausgabe*. Band 1-5.

Aus erster Hand – 3000 Jahre Kursivschrift der Pharaonenzeit digital analysiert

Gerhards, Simone
gerhards@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland; Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Gülden, Svenja A.
sguelden@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland; Technische Universität Darmstadt, Deutschland; Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Konrad, Tobias
tokonrad@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland; Technische Universität Darmstadt, Deutschland; Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Leuk, Michael
michael.leuk@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Verhoeven-van Elsbergen, Ursula
verhoeve@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland; Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland; Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Das Projekt *Altägyptische Kursivschriften* (AKU 2015) an der Akademie der Wissenschaften und der Literatur Mainz unter der Leitung von Prof. Dr. Ursula Verhoeven-van Elsbergen (JGU Mainz) in Kooperation mit Prof. Dr. Andrea Rapp (TU Darmstadt) besteht seit April 2015. Ziel ist es, in verschiedenen Modulen im Verlauf von maximal 23 Jahren eine digitale Paläographie zum Hieratischen und zu den Kursivhieroglyphen zu erstellen sowie verschiedene Aspekte der Kursivschrift-Kultur systematisch unter Einbeziehung digitaler Methoden zu untersuchen.

Im Alten Ägypten gab es neben den monumentalen und detailliert ausgeführten Hieroglyphen auch kursive (Hand-)Schriften, die als Hieratisch, Kursivhieroglyphen, Kursivhieratisch und Demotisch bezeichnet werden. Sie wurden mit Pflanzenstengeln und Rußtusche auf Papyrus, Leinen, Leder, Holz, Ton oder Stein geschrieben oder eingeritzt. Die Kursivschriften spielten unter Gelehrten, Priestern, Beamten und Schreibern eine wesentliche Rolle in den Bereichen der Kommunikation und Verwaltung, aber auch in der Dichtung, den Wissensgebieten sowie religiösen und funerären Texten. Das Hieratische war über 3000 Jahre lang in Gebrauch und wurde von den Schülern als erste Schriftart noch vor den Hieroglyphen erlernt.

Bis heute ist die knapp 100 Jahre alte *Hieratische Paläographie* von Georg Möller das Standardwerk für die ägyptologische paläographische Forschung (Möller 1909–1912). Er hat aus nur 32 gut datierten Schriftzeugnissen (vor allem Papyri) alle identifizierbaren Grapheme faksimiliert und in übersichtlichen Listen erfasst, die die Zeitspanne von der 5. Dynastie (ca. 2500 v. Chr.) bis zur römischen Kaiserzeit (3. Jh. n. Chr.) abdecken; die drei Bände bestehen aber zusammengenommen aus nur etwa 220 Seiten. Da ihm für manche Epochen nur sehr wenige oder gar keine Schriftquellen zur Verfügung standen, sind einige Zeiträume nicht oder nur unzureichend dokumentiert. Möller selbst betrachtete diese Listen als

Vorarbeiten für weitergehende Untersuchungen, was er aber aufgrund seines frühen Todes nicht realisieren konnte. Erst ca. 70 Jahre später formulierte Posener (1973) seine Anforderungen an eine zukünftige paläographische Forschung und einen *nouveau Möller*. Mit den damaligen technischen Voraussetzungen hätten die komplexen Anforderungen in Verbindung mit der Materialfülle selbst von einer Forschergruppe nicht erfüllt werden können. So erklären sich die zahlreichen Teilpaläographien, die in den nachfolgenden Jahrzehnten im Rahmen ägyptologisch-paläographischer Forschung entstanden sind (z. B. Goedicke 1988, Verhoeven 2001, Allen 2002, Lenzo 2011). Diese halten sich bis heute an das Prinzip von Möller, ordnen die Zeichen allerdings nach der Standardliste (*Sign-list*), die Gardiner in seiner *Egyptian Grammar* (Gardiner 1927, 31973: 438-548) publiziert hat. Da bei Gardiner aber nicht alle hieroglyphischen Entsprechungen zu den *Hieratogrammen* (Verhoeven 2001: 1) zu finden sind, kam es in den verschiedenen Teilpaläographien zu diversen Erweiterungen, die keinem einheitlichen Prinzip folgen und somit nicht eindeutig referenzierbar sind (Gülden 2016: 3).

Digitale Ansätze zur Analyse von Handschriften beschreiben beispielsweise Stokes (2009) für das europäische Mittelalter und Quirke (2011) für die hieratische Schrift des Alten Ägypten. Das AKU-Projekt entwickelt erstmals eine Paläographiedatenbank, in der nach und nach das gesamte Zeichenrepertoire der altägyptischen Kursivschriften erfasst wird: ca. 600 Grapheme ¬– sowohl Laut- als auch Deutzeichen, Zahlen, Maße und Korrekturzeichen sowie Ligaturen, Zeichengruppen und besondere Orthographien (Verhoeven 2015: 32). Für die Auswertung des Datenmaterials sind zudem umfassende Metadaten der Schriftträger (z. B. Herkunft, Datierung, Genre, Materialität, Beschreibstoffe und Schreibgerät) notwendig, die ebenfalls in der Datenbank erfasst werden (Gülden, Krause, Verhoeven 2017 und dies. im Druck).

Während für alphabetische Schriften bereits zahlreiche Vorarbeiten im Bereich der Handschriftenerkennung vorliegen, muss dies für die Handschrift des Alten Ägypten erst entwickelt werden, um eine Grundlage für automatisierte Prozesse bei der Zeichenerkennung, -erfassung und -auswertung dieser komplexen Schrift zu ermöglichen.

Zunächst werden die einzelnen Schriftzeichen (*Hieratogramme*) auf der Basis hochauflösender Digitalisate der Textträger faksimiliert (umgezeichnet). Diese werden sowohl als Vektor- und Rastergrafiken gespeichert. In der Datenbank, die in den nächsten Jahren als *open access online tool* zur Verfügung stehen soll, werden sie kategorisiert und annotiert. Dadurch soll die Auswertung mit unterschiedlichen Verfahren (z. B. *shape matching*, *image retrieval* und *pattern recognition*) ermöglicht werden.

Für die Hieratistik sind das vor allem Fragen zu Entwicklung und Diversität der Kursiven, Bezügen zur Hieroglyphenschrift sowie kontextuellen und funktionellen Anpassungen. Hinzu kommen Aspekte zur Schriftökonomie, Schreibrichtung, zu Abkürzungen, Diakritika und Ligaturen sowie zum Layout. Mithilfe von Clusteranalysen können Datierungen, Schreiberpersonen und regionale Unterschiede identifiziert werden.

Langfristig erhofft sich das Projekt mit den erfassten Zeichen zudem eine Basis für weitere automatisierte Verfahren (*machine learning*) zu schaffen, damit auch umfangreiche Textträger (bspw. verfügt ein 23 m langen Papyrus hochgerechnet über 140.000 Einzelzeichen) analysiert werden können. Für ein Repositorium, aber auch für Auswertungen, Visualisierungen und *linked open data*, sollen die Daten in verschiedene Formate übertragen werden, z. B. in ein TEI konformes XML-Schema und als csv-Files.

Bibliographie

Allen, James P. (2002): *The Heqanakht papyri.* Publications of the Metropolitan Museum of Art Egyptian Expedition 27. New York: Metropolitan Museum of Art.

AKU (2015): *Altägyptische Kursivschriften. Digitale Paläographie und systematische Analyse des Hieratischen und der Kursivhieroglyphen (AKU).* Akademie der Wissenschaften und der Literatur Mainz http://aku.uni-mainz.de [letzter Zugriff 22. September 2017].

Gardiner, Sir Alan (1927 ^{3}1973): *Egyptian Grammar*. Oxford 1927. Third edition. Oxford: Oxford University Press, ^{3}1973.

Goedicke, Hans (1988): *Old Hieratic Paleography*. Baltimore: Halgo.

Gülden, Svenja A. (2016): „ *Ein ‚nouveau Möller'? Grenzen und Möglichkeiten. Ein working paper zum gleichnamigen Vortrag".* Hieratic Studies Online 1 urn:nbn:de:hebis:77-publ-557584.

Gülden, Svenja A. / Krause, Celia / Verhoeven, Ursula (2017): „Prolegomena zu einer digitalen Paläographie des Hieratischen" in: Fischer, Franz / Sahle, Patrick / Busch, Hannah (eds.): *Kodikologie & Paläographie im digitalen Zeitalter 4*. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand 253-273 kups.ub.uni-koeln.de/7774/.

Gülden, Svenja A. / Krause, Celia / Verhoeven, Ursula (im Druck): „Digital Palaeography of Hie-

ratic“, in: Davies, Vanessa / Laboury, Dimitri (eds.): *Oxford Handbook of Epigraphy and Palaeography*, Oxford: Oxford University Press.

Lenzo, Giuseppina (2011): „Paleografia“, in: Roccati, Alessandro: *Magica Taurinensia. Il grande papiro magico di Torino e i suoi duplicati*. Analecta Orientalia 56. Roma: Gregorian & Biblical Press 193–255.

Möller, Georg (1909–1912): *Hieratische Paläographie. Die Aegyptische Buchschrift in ihrer Entwicklung von der fünften Dynastie bis zur Römischen Kaiserzeit* I–III. Leipzig: J. C. Hinrichs, 1909–1912. I–IV: Leipzig: J. C. Hinrichs, [2]1927–1936. Neudruck Osnabrück: Otto Zeller, 1965.

Posener, Georges (1973): „L’écriture hiératique“, in: *Textes et langages de l’Égypte pharaonique, cent cinquante années de recherches* I , *1822–1972*. Bibliothèque d’Étude 64, 1. Le Caire: Institut français d’archéologie orientale 25-30.

Quirke, Stephen (2011): „Agendas for Digital Palaeography in an Archaeological Context: Egypt 1800 BC“, in: Fischer, Franz et al. (eds.): *Kodikologie und Paläographie im digitalen Zeitalter*. Schriften des Instituts für Dokumentologie und Editorik 3. Norderstedt: Books on Demand 279-294 kups.ub.uni-koeln.de/4354/.

Stokes, Peter (2009): „ *Computer-Aided Palaeography, Present and Future*“, in: Rehbein, Malte et al. (eds.): *Kodikologie und Paläographie im digitalen Zeitalter.* Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand GmbH 310-338 kups.ub.uni-koeln.de/2978/.

Verhoeven, Ursula (2001): *Untersuchungen zur späthieratischen Buchschrift*. Orientalia Lovaniensia Analecta 99. Leuven: Peeters.

Verhoeven, Ursula (2015): „Stand und Aufgaben der Erforschung des Hieratischen und der Kursivhieroglyphen“, in: Verhoeven, Ursula (ed.): *Ägyptologische „Binsen“-Weisheiten I–II, Neue Forschungen und Methoden der Hieratistik, Akten zweier Tagungen in Mainz im April 2011 und März 2013*. Abhandlungen der Akademie der Wissenschaften und der Literatur Mainz, Einzelveröffentlichungen Nr. 14. Mainz und Stuttgart: Franz Steiner Verlag 23–63 urn:nbn:de:hebis:77-publ-547544.

BeWeB-3D – Zur Digitalisierung interaktiver Buchobjekte

Hug, Marius
marius.hug@googlemail.com
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Deutschland

Im Rahmen des Projekts BeWeb-3D [1]und in Kooperation mit dem Zentrum für digitale Kulturgüter in Museen (ZEDIKUM [2]) wurde an der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz ein Konzept zur Digitalisierung sogenannter Bewegungsbücher erarbeitet. Dabei handelt es sich um Buchobjekte, die bewegliche Teile enthalten und damit eine Interaktion erfordern, die über das reine Umblättern der Seiten hinausgeht (vgl. Schmitz-Emans 2016). Populäre Formen dieses Buchtyps sind Klappbilderbücher – bereits im 16. Jahrhundert gab es anatomische Lehrbücher mit Papierklappen – oder Bücher mit Volvellen, i.e. drehbare Papierscheiben, die bspw. für kalendarische Berechnungen genutzt wurden. Einen richtigen Boom erfuhr das Medium Bewegungsbuch, als im Laufe des 19. Jahrhunderts das sogenannte Spielbilderbuch die (gutbürgerlichen) Kinderzimmer Mitteleuropas eroberte. In diesen aufwendig und von Hand produzierten Kinderbüchern finden sich alle Formen des angesprochenen Papierdesigns wieder, was im direkten Kontext zu den industriellen Fortschritten in den Bereichen der Papierproduktion und Buchdrucktechnologie (Lithographie und v.a. Chromolithographie) zu sehen ist. Eine besondere Ausprägung sind die in den 1880er und 90er Jahren in Perfektion vom Münchner Buchkünstler Lothar Meggendorfer produzierten Ziehbilderbücher (siehe Abb. 1), die historisch in enger Verwandtschaft zu den damals populären Marionetten- und Schattentheatern einerseits und Laterna Magica Vorführungen andererseits rezipiert werden müssen.

Abb. 1: Vier stills des Anglers bei der Arbeit. Aus: Lothar Meggendorfer: Bewegliche Schattenbilder, 1886.

All diesen Buchtypen – seien es wissenschaftliche Bücher aus der Zeit des 12. Jahrhunderts oder Kinderbücher ab dem Anfang des 19. Jahrhunderts – ist eines gemein: Sie verfügen über eine zusätzliche (räumliche oder zeitliche) Dimension und verweigern sich so gängigen Digitalisierungspraktiken aus dem Bereich der Buchdigitalisierung. Elena Pierazzo beschreibt diese Objekte als „notable exceptions“ (Pierazzo 2015, 32) und verweist damit einerseits auf deren Exklusivität, andererseits aber auch auf die unbestrittene Relevanz des Gegenstands.

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt schlägt einen Digitalisierungsworkflow vor, der der Komplexität des Gegenstands gerecht wird und entsprechend aufwendig ist. Unterschiedliche Ausgabeformate (Bild, Film, 3D) sollen in einer additiv-synoptischen Präsentation zugänglich gemacht werden. 1) Für die 2D-Digitalisierung bedeutet die spätere Datenverarbeitung, dass unterschiedliche Zustände der beweglichen Bilder aufgenommen werden müssen (Stichwort Keyframing). 2) Um eine entsprechende metrische Präzision des Digitalisats zu erreichen, wird das aus dem Bereich der Kulturgutdigitalisierung bekannte Structure from Motion-Verfahren (SfM) eingesetzt. 3) Um schließlich die Interaktion mit dem Digitalisat zu ermöglichen, kommen bei der Datenvisualisierung Game und/oder Physik Engines aus dem Computerspielebereich (z.B. Unity) zum Einsatz. Das Ziel ist ein multimodales Replikat bestehend aus Bild, Film, 3D-Objekt und einem eigenen Metadatenmodell, in welchem die bereits erwähnte Möglichkeit der Interaktion abgebildet ist.

Im Rahmen der Posterpräsentation werden verschiedene bis dahin replizierte Buchtypen unter Berücksichtigung der jeweiligen Digitalisierungstechnologien vorgestellt. Die digitalisierten Spielbilderbücher stehen browserbasiert oder als Augmented Reality *hands-on* zur Verfügung. Der so durch BeWeB-3D zwischen datenbasiertem Poster und Android-App buchstäblich aufgespannte Raum soll zu einem Raum der kritischen Befragung konkreter Möglichkeiten und Grenzen von Digital Humanities werden. Das *Bewegungsbuch* als Teil einer „ergodic edition“ (Vanhoutte 2015, 141f) eignet sich dafür nicht zuletzt aufgrund der noch immer fehlenden Standards im Bereich der (interaktiven) 3D-Digitalisierung.

Fußnoten

1. http://staatsbibliothek-berlin.de/die-staatsbibliothek/projekte/beweb-3d/.
2. http://www.zedikum.de/

Bibliographie

Meggendorfer, Lothar (1886): *Bewegliche Schattenbilder, 1. Vorstellung,* München: Braun & Schneider. Staatsbibliothek zu Berlin, Kinder- und Jugendbuchabteilung, Signatur: 53 BB 500600-1 R.

Pierazzo, Elena (2015): *Digital Scholarly Editing: Theories, Models and Methods*. Farnham: Ashgate.

Schmitz-Emans, Monika (2016): "Modellierungen, Inszenierungen, Transgressionen. Zu Geschichte, Spielformen und Poetik des beweglichen Buchs", in: Bachmann, Christian A./Emans, Laura/Schmitz-Emans, Monika (eds.): *Bewegungsbücher. Spielformen, Poetiken, Konstellationen,* Berlin: Ch. A. Bachmann Verlag, S. 85–123.

Vanhoutte, Edward (2010): "Defining Electronic Editions: A Historical and Functional Perspective", in: McCarty, W. (ed.): *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Open Book Publishers, Cambridge, S. 119–144.

Biographik in den Digital Humanities – Kritische Bestandsaufnahme und quantitative Analysemöglichkeiten am Beispiel des Österreichischen Biographischen Lexikons 1815–1950

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Bernád, Ágoston
agoston.bernad@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Kaiser, Maximilian
maximilian.kaiser@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Lejtovicz, Katalin
katalin.lejtocicz@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Rumpolt, Peter
peter.rumpolt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Quantitative Auswertungen gewinnen zunehmend auch in den Geisteswissenschaften an Bedeutung (z. B. Harber 2011). Nicht selten wird mittlerweile auch der Begriff „Big Data“ in Zusammenhang mit den Humanities ganz allgemein verwendet (z. B. gral 2016). Für solche Analysen wird des Öfteren auf einheitliche Korpora zurückgegriffen (z. B. Wikipedia: Russo et al. 2015), die nicht nur einen standardisierten Zugriff auf tausende Datensätze, sondern auch homogene Metadaten und harmonisierte Vokabulare bieten. Während diese Korpora formal geradezu prädestiniert für Analysen mit den neuen digitalen Tools scheinen, sind die grundlegenden Daten durch die Entstehungsgeschichte der Datensammlungen nicht selten unausgewogen und können mitunter zu falschen Ergebnissen führen. Insbesondere trifft dies auf über einen längeren Zeitraum entstandene Nationalbiographien zu. Trotz dieses offensichtlichen Dilemmas fokussieren Studien meist auf die Zuverlässigkeit der digitalen Tools selbst und weniger auf die Verlässlichkeit der Quellen (z. B. Reinert et al. 2015; Stotz et al. 2015).

Die Posterpräsentation bietet eine exemplarische Korpusanalyse der digitalen Fassung des 1946 gegründeten und seit 1954 (seit 2009 auch online) erscheinenden *Österreichische Biographische Lexikon 1815–1950*. Dieses enthält in 14 Bänden (68 Lieferungen) über 18.000 Biographien. Es wird dadurch die oftmals vernachlässigte Unausgewogenheit einer auf den ersten Blick einheitlich wirkenden Quelle diskutiert.

Analyse der vorläufigen Ergebnisse

1. Informationsdichte

Anders als ursprünglich angenommen, bleibt die Informationsdichte (Named Entities pro tokens) mit steigender Länge der Biographietexte etwa gleich hoch. Biographien werden inhaltlich bereits ab Anfang der 1960er-Jahre umfangreicher. Ab den 1980er-Jahren zeigt die Kurve eine sehr deutliche Steigerung (Abb. 1.).

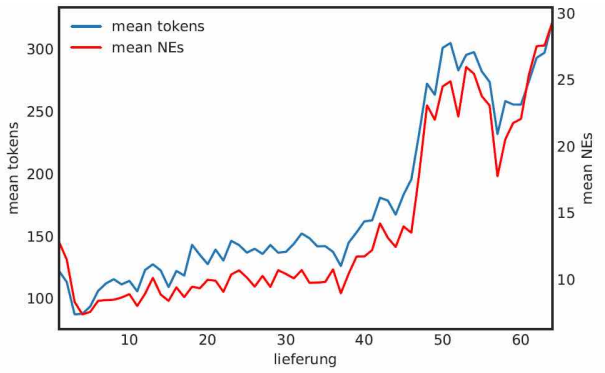


Abbildung 1: Durchschnitt Tokens und Named Entities nach Lieferung

Diese Entwicklung lässt sich anhand der Werkgenese bzw. der Geschichte des Lexikons erklären. Nach den ursprünglichen Plänen sollte das ÖBL drei bis vier Bände an Kurzbiographien umfassen. Die ab 1946 verfassten Biographien der ersten fünf Lieferungen – gesammelt im ersten Band, 1957 erschienen – sind allesamt nach diesem Konzept entstanden. Aus methodischen Gründen wurde der Ursprungsplan jedoch bereits in den 1950er-Jahren als unbrauchbar verworfen. Die Österreichische Akademie der Wissenschaften entschied sich für die Gestaltung eines umfangreichen Nachschlagewerkes (Reitterer 1998) wodurch die durchschnittliche Länge der Biographien deutlich anstieg.

2. Die geographische Verteilung der Geburtsorte der biographierten Personen

Die geographische Verteilung der Biographierten nach Geburtsorten ist nicht ausgewogen. Verglichen wurden die erste Lieferung des ÖBL (1954) mit den Lieferungen 62-66 (2010–2015); beide Stichproben enthalten etwa gleich viele Biographien (Abb. 2. und 3.).

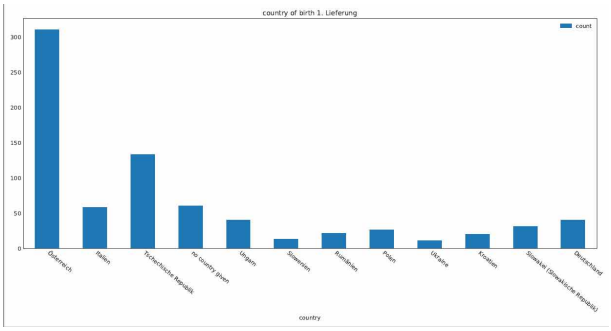


Abbildung 2: Verteilung der Geburtsorte in der 1. Lieferung des ÖBL (1954)

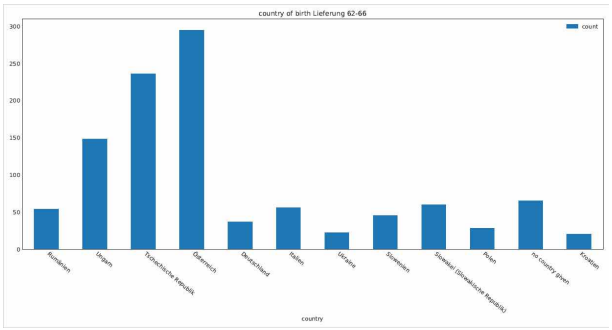


Abbildung 3: Verteilung der Geburtsorte in den Lieferungen 62–66 des ÖBL (2010–2015)

Die Kronländer der westlichen Reichshälfte der Monarchie (u. a. die Nachfolgestaaten Österreich und Tschechien) sind in der ersten Lieferung gegenüber der östlichen Reichshälfte (auf den Abbildungen die Nachfolgestaaten Ungarn, Slowakei, Rumänien [Siebenbürgen], Kroatien) deutlich bevorzugt. Die Verteilung der Geburtsorte in den Lieferungen 62–66 zeigt hingegen ein wesentlich ausgeglicheneres Bild. Die Diskrepanz resultiert aus den ursprünglichen Aufnahmekriterien des ÖBL, die zeitlich und räumlich, einerseits nach kulturhistorischen Gesichtspunkten, andererseits jedoch staatsrechtlich definiert wurden (Obermayer-Marnach 1957). Bei der Aufnahme von Personen fokussierte man zwar auf den gesamten Raum des ehemaligen Habsburgerreiches, allerdings mit der Einschränkung, dass nach dem Ausgleich (1867) auf dem Gebiet der Länder der Ungarischen Krone geborene Personen – von wenigen Ausnahmen abgesehen – nicht in das Lexikon aufgenommen wurden.

3. Anteil weiblicher Persönlichkeiten im ÖBL

Der Anteil weiblicher Biographierter hat sich über die mehr als siebzig Jahre lange Geschichte des Lexikons kaum verändert (Abb. 4.).

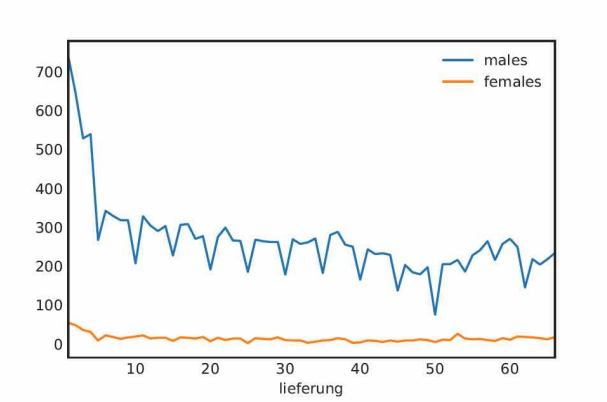


Abbildung 4: Männliche und weibliche Personen im ÖBL nach Lieferung

Unsere ursprüngliche Annahme, die ab den 1970er- und 1980er-Jahre sich entfaltende Frauen- und Geschlechtergeschichte hätte in diesem Zusammenhang eine positive Auswirkung auf die Anzahl der in das Lexikon aufgenommenen Frauen gehabt, konnte nicht bestätigt werden. Der geringe Anstieg ab der 50. Lieferung lässt sich auf den relativ hohen Frauenanteil in der Hauptberufsgruppe “Musik und darstellende Kunst” zurückführen (Abb. 5.).

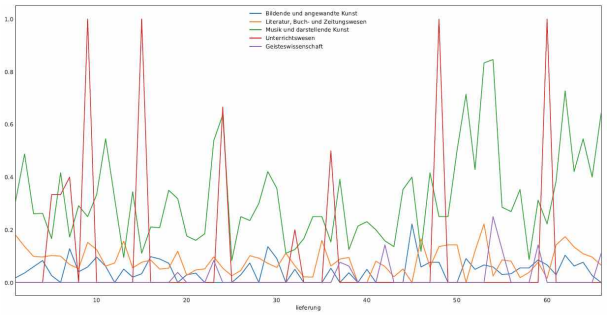


Abbildung 5: Verteilung weiblicher Personen im ÖBL nach Hauptberufsgruppen und Lieferung

Der relativ geringe Anteil an Frauen lässt sich durch den Berichtszeitrahmen des ÖBL, allen voran jedoch durch den Schwerpunkt 19. Jahrhundert erklären. Als Beispiel sei hier lediglich darauf verwiesen, dass in der Monarchie Frauen die Möglichkeit des Studiums großteils verwehrt blieb (Heindl, Tichy 1993). Anteilsmäßig findet man die meisten Frauen in der Berufsgruppe „Diverse“, in der u. a. die Mitglieder des Kaiserhauses erfasst werden (Abb. 6.).

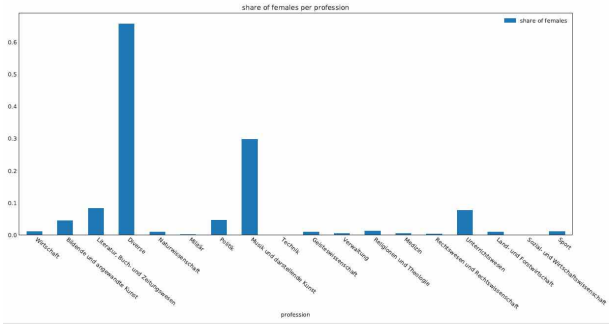


Abbildung 6: Anteil der Frauen in den einzelnen Hauptberufsgruppen des ÖBL

Conclusio

Die vorliegende erste Analyse zeigt exemplarisch die erheblichen Variationen innerhalb der Biographien des ÖBL (Informationsdichte, Herkunftsländer, Geschlecht). Nationalbiographien wie das ÖBL können zu einer wertvollen Ressource für die quantitativen Geschichtswissenschaften - Stichwort "BigData" - werden, wenn die Herausforderungen ihrer heterogenen Zusammensetzung gelöst werden.

Bibliographie

Österreichisches Biographisches Lexikon 1815–1950. Wien: Böhlau Verlag, Verlag der Österreichischen Akademie der Wissenschaften, 1954ff.; Online-Ausgabe: www.biographien.ac.at , Zugriff: 14.12.2017.

Haber, P. (2011): „Neue – digitale – Wege für die Geschichtswissenschaft?: Die Rückkehr der Zahlen und Daten". In: *Neue Zürcher Zeitung*, 2.1.2011.

gral (2016): „Big Data in den Geschichtswissenschaften". In: *Die Presse*, 4.4.2016.

Russo, I.; Caselli, T.; Monachini, M. (2015): „Extracting and Visualising Biographical Events from Wikipedia". In: S. ter Braake et al. (eds.), *BD2015. Biographical Data in a Digital World 2015. Proceedings of the 1st Conference on Biographical Data in a Digital World*, Amsterdam, 111–115.

Reinert, M.; Schrott, M.; Ebneth, B. et al. (2015): „From Biographies to Data Curation – The Making of www.deutsche-biographie.de". In: S. ter Braake et al. (eds.), *BD2015*, 13–19.

Stotz, S.; Stuß, V.; Reinert M.; Schrott M. (2015): „Interpersonal Relations in Biographical Dictionaries. A Case Study". In: Serge ter Braake et al. (eds.), *BD2015*, 74–80.

Gruber, Ch.; Feigl, R. (2009): „Von der Karteikarte zum biografischen Informationsmanagementsystem: Neue Wege am Institut Österreichisches Biographisches Lexikon und biographische Dokumentation". In: M. Schattkowsky, F. Metasch (eds.), *Biografische Lexika im Internet: Internationale Tagung der „Sächsischen Biografie" in Dresden (30. und 31. Mai 2008)* , Bausteine aus dem Institut für Sächsische Geschichte und Volkskunde 14, Dresden, 55–75.

Reitterer, H. (1998): „Österreichisches Biographisches Lexikon und biographische Dokumentation". In: Csendes P., Lebensaft E. (eds.), *Traditionelle und zukunftsorientierte Ansätze biographischer Forschung und Lexikographie* , ÖBL – Schriftenreihe 4, Wien, 42–46.

Obermayer-Marnach E. (1957): „Einleitung". In: *Österreichisches Biographisches Lexikon 1815–1950*, Bd. 1, X–XV.

Latzke W. (1960): „Österreichisches Biographisches Lexikon 1815–1950, Bd. 1. und 2.". In: *Archivalische Zeitschrift* 56, 1960, 143–145.

Heindl, W.; Tichy, M, eds. (1993): „ *»Durch Erkenntnis zu Freiheit und Glück ...«. Frauen an der Universität Wien (ab 1897)* , Schriftenreihe des Universitätsarchivs, Universität Wien 5, Wien.

Chancen und Grenzen von Digitalen Methoden zur Analyse der politischen Meinungsbildung in Sozialen Medien

Guhr, Svenja
svenjasimone.guhr@stud.uni-goettingen.de
Universität Göttingen, Deutschland

Pannach, Franziska
franziska.pannach@stud.uni-goettingen.de
Universität Göttingen, Deutschland

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

Knauth, Jürgen
jknauth@uni-goettingen.de
Universität Göttingen, Deutschland

Kauf, Carina
carina.kauf@stud.uni-goettingen.de
Universität Göttingen, Deutschland

Sporleder, Caroline
csporled@gwdg.de
Universität Göttingen, Deutschland

Soziale Medien spielen weltweit im politischen Meinungsbildungsprozess eine immer wichtigere Rolle. Sowohl Onlineangebote von Zeitungen als auch Twitteraccounts von Organisationen oder Personen können quasi in Echtzeit über aktuelle Geschehnisse informieren. Die Kommentar- und Replyfunktionen bieten zudem einen digitalen Ort für den öffentlichen Austausch. Damit können auch ‚normale' Nutzer ganz gezielt Nachrichten wie auch persönliche Kommentare oder Gerüchte in einem Ausmaß verbreiten wie es im vor-digitalen Zeitalter kaum möglich war. Das Entstehen eines vollkommen neuen digitalen Kommunikationsraumes, der sowohl Grenzen überschreitet als auch potenziell neue Grenzen schafft („Echokammern") kann zum einen positiv im Sinne einer Demokratisierung öffentlicher Meinungsbildung gewertet werden (Mossberger et al. 2007), birgt aber auch Risiken (Mancini, 2013, Sarcinelli, 2014).

Die Analyse der Rolle von Sozialen Medien im öffentlichen Meinungsbildungsprozess ist inzwischen ein aktives Forschungsfeld (z.B. Törnberg & Törnberg, 2016, Eilders, 2013). Für eine umfassende Auswertung des Datenmaterials sind jedoch mächtige Analyseverfahren notwendig (Sentimentanalyse, Netzwerkanalyse, Diskursanalyse, Bot-Erkennung etc.), die zur Zeit noch nicht in adäquatem Maß zur Verfügung stehen (s. Mohammad et al., 2015). So ist zum Beispiel die Sentimentanalyse relativ gut erforscht, beschränkt sich aber in der Regel auf das Englische sowie auf die Textsorte „Rezension". In drei Pilotstudien haben wir untersucht, mit welchem Aufwand sich Methoden der Sentimentanalyse und Bot-Erkennung auf neue Sprachen anpassen lassen und wo mögliche Grenzen dieser Verfahren liegen. Im weiteren Projektverlauf, soll ergründet werden, inwieweit sich Prozesse der öffentlichen Meinungsbildung in sozialen Medien mit den zur Zeit zur Verfügung stehenden Verfahren nachvollziehen lassen. Die Pilotstudien untersuchen den öffentlichen Diskurs im Kontext von Wahlen und decken zwei verschiedene Sprachen (französisch, deutsch) und Textsorten (Kommentare auf Nachrichtenseiten, Tweets) ab und variieren hinsichtlich der ausgewerteten Datenmenge und Herangehensweise (gemischt qualitativ-quantitative stilistische Auswertung vs. primär quantitative Polaritätsanalyse vs. Bot-Erkennung).

Pilotstudie 1: Sentimentanalyse zur Bundestagswahl

In der ersten Studie stehen Tweets zur Bundestagswahl (BTW) im Vordergrund, für die ein Sentimenttagger entwickelt wurde, um das in den Tweets ausgedrückte Sentiment im Hinblick auf Themen, Parteien und Personen zu analysieren. Zwischen Mai und September 2017 wurden mehr als 5.4 Millionen Tweets gesammelt und nach Schlagworten und Hashtags zur BTW und zu den Parteien gefiltert. 600 Tweets wurden von 3 AnnotatorInnen manuell als `positiv', `negativ', 'neutral' oder 'irrelevant' (kein Bezug zur Bundestagswahl) klassifiziert. Um die Anforderungen eines (bisher nicht verfügbaren) Twitter-Sentimenttaggers für das Deutsche im Hinblick auf die Fragestellung zu ermitteln, wurden in einer Vorstudie 100 Tweets von fünf AnnotatorInnen auf ihr Sentiment hin untersucht. Dabei wurde deutlich, dass das Sentiment zum Teil aus dem Kontext inferiert werden muss (z.B. angehängte Bilder), was besonders die automatische Analyse erschwert. Zudem werden oft mehrere Sentiments ausgedrückt werden (s. Mohammad, 2016). Das Inter-Annotator-Agreement für alle 3 AnnotatorInnen lag bei 61% (94,5% bei Übereinstimmung von 2 AnnotatorInnen). Eine Sichtung der annotierten Daten zeigt, dass Tweets zur BTW überwiegend sentiment-behaftet sind (91%) und negatives Sentiment vorherrscht (81% neg., 10% pos, 9% neut). Für den Sentimenttagger wurden verschiedene überwachte maschinelle Lernverfahren auf den annotierten Daten trainiert und getestet (10-fache Kreuzvalidierung). Verwendet wurden dabei Unigramme sowie weitere Informationen wie das Vorkommen von Emoticons oder bestimmten Satzzeichen. Der beste Klassifikator (Naive Bayes) erreichte 71% F-Score (69% Prec., 75% Rec.). Besonders indikativ für negatives Sentiment in unserem Datenset sind die Unigramme „Schulz" und „Merkel", während positives Sentiment durch „!" angezeigt wird. Hier ist sicher eine weitergehende Analyse notwendig.

Pilotstudie 2: Französische Präsidentschaftswahl

In einer gemischt qualitativ-quantitativen Analyse der französischen Präsidentschaftswahl wurden Kommentare unter Onlineartikeln der französischen Tageszeitung *Le Monde* im Zeitraum zwischen dem ersten (23.04.2017) und dem zwei-

ten (07.05.2017) Wahltag manuell und automatisch analysiert. Hierfür wurden zunächst themenähnliche Artikel ausgewählt, die sich mit den zwei Präsidentschaftskandidaten der zweiten Wahlrunde befassten. Für die weitere Analyse wurden sechs repräsentative Onlineartikel und die zugehörigen Userkommentare ausgewählt. Diese ausgewählten Textdateien wurden mithilfe der Topic Modelling Software MALLET auf ihre Hauptthemen hin analysiert. Als Trainingsdaten für das Topic Modelling wurden ähnliche Onlineartikel anderer Zeitungen und Wahlprogramme der zwei Hauptparteien der zweiten Wahlrunde verwendet. Da es auch für das Französische keinen frei verfügbaren Sentimenttagger gibt, wurde ein regelbasiertes System entwickelt, dass auf einer Kombination von verschiedenen Sentimentlexika beruht, u.a. auch multi-linguale Ressourcen (NRC Emotion Lexicon, Mohammad & Turney, 2013), die auf der einen Seite die Datengrundlage vergrößern, auf der anderen Seite aber auch potenziell Fehler (z.B. fehlerhafte Übersetzungen, Lesartenambiguitäten) beitragen.

Pilotstudie 3: Erkennung von Social Bots

In der dritten Studie geht es um die Unterscheidung von menschlichen und künstlichen Akteuren. Um die potenziellen Auswirkungen von Social Bots auf die politische Meinungsbildung zu quantifizieren ist es notwendig, künstliche Agenten automatisiert erkennen zu können. Bisherige Verfahren (z.B. Varol et al., 2017) können simple Bot-Accounts zuverlässig erkennen, scheitern aber an fortgeschrittenen Bots, die sich nicht mehr offensichtlich von echten Menschen unterscheiden. In einem weiterentwickelten Verfahren analysieren wir speziell Accounts dieser Art. Als Datengrundlage hierfür wird der MIB-Datensatz verwendet (Cresci et al., 2017). Zur Bot-Erkennung wird mit überwachten machinellen Lernverfahren experimentiert.

Bibliographie

Cresci, Stefano / Di Pietro, Roberto / Petrocchi, Marinella / Spognardi, Angelo / Tesconi, Maurizio (2017): „The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race.“ in: *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion).*

Eilders, Christiane (2013): „Öffentliche Meinungsbildung in Online-Umgebungen. Zur Zentralität der normativen Perspektive in der politischen Kommunikationsforschung.“ In: Karmasin, M. et al. (Eds.): *Normativität in der Kommunikationswissenschaft*, Wiesbaden: Springer, 329-351.

Mancini, Paolo (2013): "Media Fragmentation, Party System, and Democracy." *The International Journal of Press/Politics* 18 (1): 43-60.

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)*, June 2014, San Diego, California.

Mohammad, Saif / Kiritchenko, Svetlana / Zhu, Xiaodan / Martinet, Joel (2015): „Sentiment, Emotion, Purpose, and Style in Electoral Tweets“, in: *Information Processing & Management,* 51:4, 480–499.

Mohammad, Saif / Turney, Peter (2013): Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence, 29 (3), 436-465, 2013.

Mossberger, Karen / Tolbert, Caroline J. / McNeal, Ramona S. (2007): *Digital Citizenship. The Internet, Society, and Participation,* Cambridge MA/London: MIT Press.

Sarcinelli, Ulrich (2014): „Von der Bewirtschaftung der Aufmerksamkeit zur simulativen Demokratie?“ in: *Zeitschrift für Politikwissenschaft 24 (3), 329 - 339.*

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

Törnberg, Anton / Törnberg, Petter (2016): „Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum“, in: *Discourse and Society* 27:4, 401-422.

Varol, Onur / Ferrara, Emilio / Davis, Clayton A. / Menczer, Filippo / Flammini, Alessandro (2017): „Online Human-Bot Interactions: Detection, Estimation, and Characterization.“ in: *Proceedings of ICWSM'17*

CLARIN Legal Information Plattformen und Legal Helpdesk

Kamocki, Pawel
pawel.kamocki@gmail.com
WWU Münster, Germany; IDS Mannheim; ELDA, France

Ketzan, Erik
eketzan@gmail.com
Birkbeck, University of Lonfon

Wildgans, Julia
j.wildgans@googlemail.com
IDS Mannheim; Universität Mannheim

Witt, Andreas
witt@ids-mannheim.de
IDS Mannheim; Universität zu Köln

Wissenschaftler im Bereich der Digital Humanities sind ständig auf einen Zugang zu vertrauenswürdigen und zuverlässigen rechtlichen Informationen angewiesen. Die entscheidenden rechtlichen Herausforderungen stellen sich vor allem im Immaterialgüterrecht (insbesondere in Bezug auf das Urheberrecht, das sui generis-Recht für Datenbanken und das verwandte Schutzrecht für den Verfasser von wissenschaftlichen Ausgaben) und im Datenschutzrecht. Daher ist es sinnvoll, beides bereits in der Anfangsphase jedes Projekts zu berücksichtigen, um rechtliche Probleme in späteren Projektphasen und das Scheitern von Forschungsprojekten zu vermeiden.

Allerdings erscheint eine ständige Information über die rechtlichen Rahmenbedingungen vor dem Hintergrund der ständigen Änderungen der Gesetze, die die neuen Technologien betreffen, sehr schwierig. Auch in 2018 wird es sowohl im deutschen als auch im europäischen Datenschutzrecht wesentliche Änderungen geben, die Auswirkungen auf die Erhebung, den Zugang und die Verwendung von Forschungsdaten haben werden. Darüber hinaus wird derzeit über den Entwurf einer neuen Richtlinie im Urheberrecht diskutiert, die möglicherweise schon bald verabschiedet wird. All diese Änderungen im Blick zu behalten erfordert jedenfalls regelmäßigen Zugang zu aktuellen rechtlichen Informationen.

Daher haben Pawel Kamocki und Erik Ketzan im Jahr 2012 die CLARIN-D Legal Information Plattform für DH Forscher in Deutschland aufgesetzt: Sie ist sowohl in deutscher als auch in englischer Sprache verfügbar. 2016 folgte die CLARIN Legal Information Plattform für Wissenschaftler aus den übrigen CLARIN Consortium Ländern, die bisher lediglich in englischer Sprache abrufbar ist. Beide Webseiten stellen in verschiedenen Artikeln und Formaten (derzeit insgesamt ca. 25.000 Wörter) rechtliche Informationen für den Bereich der Digital Humanities bereit und streben dabei danach, die umfangreichste und aktuellste Wissensressource für Wissenschaftler zu sein.

Sie enthalten Erklärungen zu den grundlegenden rechtlichen Prinzipien und Konzepten im Bereich des Urheberrechts (Gegenstand, Rechteinhaberschaft, Umfang und Reichweite des Schutzes und Schrankenregelungen insbesondere für wissenschaftliche Zwecke) und des sui-generis-Rechts für Datenbanken, zur Lizenzierung (einschließlich der Nutzung öffentlicher Lizenzen für Daten und Software) und zum Datenschutz. Darüber hinaus werden Wissenschaftler bei Bedarf auch zu praktischen Lizenzauswahlinstrumenten weitergeleitet, wie z.B. dem “Public License Selector” (http://ufal.github.io/public-license-selector/), der 2014 im Rahmen einer Kooperation zweier CLARIN-Zentren von Kamocki, Stranak und Sedlak entwickelt wurde. Zusätzlich bieten die Plattformen Zugriff auf die CLARIN Legal Issues Committee (CLIC) White Paper Series, die eine Open Access Publikation von Kommentaren und Forschungsergebnissen bezüglich rechtlicher Fragestellungen im Bereich der Sprachwissenschaft unter der redaktionellen Leitung des CLIC ermöglichen.

Das Legal Helpdesk ist der “direkte Draht” zu einer persönlichen Hilfestellung: Dieses ermöglicht eine Kontaktaufnahme mit einem Teammitglied des CLARIN-Teams, das Wissenschaftler zu hilfreichen Ressourcen und Informationen bezüglich ihrer Forschungsfrage leiten kann.

Die Plattformen sind frei im Internet verfügbar und werden in regelmäßigen Abständen aktualisiert. Beide werden häufig im Rahmen von CLARIN-D und CLARIN-EU-Projekten genutzt.

Unser Poster wird diese hilfreichen CLARIN Ressourcen anhand von Graphiken und Text vorstellen und aktuelle Updates darstellen, die der DH Community möglicherweise noch unbekannt sind.

Delta vs. N-Gram-Tracing: Wie robust ist die Autorschafts-attribuierung?

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

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

Die Autorschaftsattribuierung, also die Zuweisung von Texten unbekannter oder umstrittener Autorschaft zu ihrem wahren Autor, hat vielfältige Anwendungen beispielsweise in der Literatur- und Geschichtswissenschaft oder der forensischen Sprachwissenschaft. Eine populäre Methode zur Autorschaftsattribuierung ist die Anwendung von Deltamaßen (Burrows 2002; Argamon 2008) wie zum Beispiel Cosine-Delta (Smith und Aldridge 2011). Deltamaße verwenden die n häufigsten Wörter im Korpus, standardisieren die Frequenzen auf z-Werte und wenden ein Abstandsmaß, im Fall von Cosine-Delta den Kosinusabstand, an. Typischerweise schließt sich die Anwendung eines (hierarchischen) Clusterverfahrens an, das Texte des selben Autors zusammengruppiert.

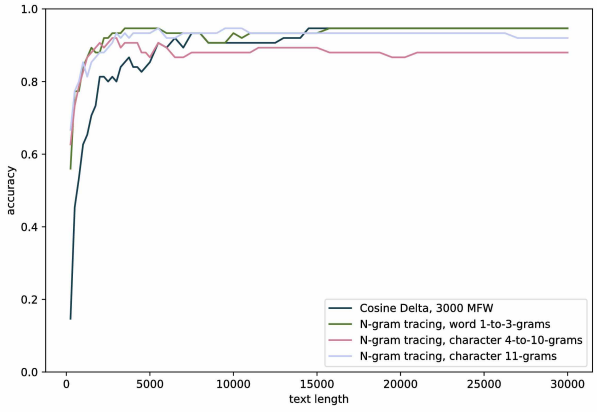
Eine neue Methode zur Autorschafts-attribuierung ist das sogenannte N-Gram-Tracing (Grieve et al., in Begutachtung). Hierbei werden aus dem zu klassifizierenden Text alle Wort- oder Buchstaben-N-Gramme einer bestimmten Länge extrahiert. Der Text wird dann dem Autor zugewiesen, der im Vergleichskorpus die meisten dieser N-Grammtypen verwendet. Die Häufigkeit der N-Gramme spielt dabei keine Rolle, es geht nur darum, wie viele N-Gramme aus dem zu klassifizierenden Text auch im Vergleichskorpus auftauchen.

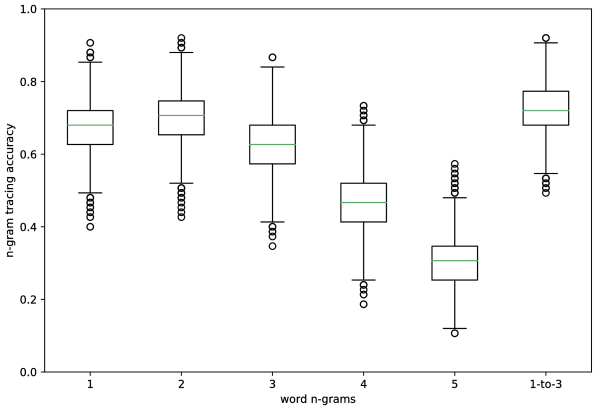
Wenn Methoden zur Autorschaftsattribuierung angewandt werden sollen um tatsächlich eine strittige Autorschaftsfrage zu klären, ist es sehr wichtig die Zuverlässigkeit und Robustheit der Verfahren abschätzen zu können, schließlich gibt es eine ganze Reihe von Einflussfaktoren. Kritisch sind zum Beispiel die folgenden Fragen: Welchen Einfluss haben die Länge des zu klassifizierenden Textes und die Größe des Vergleichskorpus auf die Genauigkeit der Autorschaftsattribuierung? Gibt es für die beiden Verfahren eine Mindesttextlänge, die nicht unterschritten werden sollte? Wie stark werden die Verfahren durch autor- und werkspezifische Eigenheiten beeinflusst? Ist die Genauigkeit der Autorschaftsattribuierung robust in Bezug auf die Zusammensetzung des Vergleichskorpus oder kann die Auswahl der Autoren und Texte das Ergebnis beeinträchtigen?

Um diese Fragen zumindest teilweise beantworten zu können, führen wir eine Reihe von Evaluationsexperimenten durch. Um die Ergebnisse des N-Gram-Tracings besser mit denen von Delta vergleichen zu können, führen wir auf den Deltaabständen zwischen den Texten kein Clustering sondern eine nearest-neighbor-Klassifikation durch, d.h. wir weisen den zu klassifizierenden Text dem Autor des Textes mit dem geringsten Abstand zu. Im Einzelnen handelt es sich um zwei Kürzungs- und zwei Samplingexperimente. Datengrundlage für die Kürzungsexperimente sind die deutschen, englischen und französischen Romankorpora, die unter anderem von Jannidis et al. (2015) und Evert et al. (2017) verwendet wurden. Jedes Korpus besteht aus je drei Romanen von 25 Autoren, also aus 75 Romanen. Für das erste Kürzungsexperiment wird die Größe des Vergleichskorpus stabil gehalten und nur der zu klassifizierende Text gekürzt. Für Delta wird zusätzlich die Anzahl der verwendeten häufigsten Wörter variiert. Im zweiten Kürzungsexperiment werden sowohl der zu klassifizierende Text als auch das Vergleichskorpus gekürzt. Über ein leave-one-out-Verfahren werden alle Texte im Korpus klassifiziert um die Genauigkeit der Verfahren zu ermitteln.

Für die Samplingexperimente verwenden wir eine Sammlung von 1018 deutschen Romanen aus dem langen 19. Jahrhundert. Alle Texte wurden von Muttersprachlern verfasst. Für das erste Samplingexperiment ziehen wir 5000 zufällige Stichproben von 25 Autoren und je drei Romanen (die Zusammensetzung der einzelnen Stichproben ist also vergleichbar mit den oben erwähnten Romankorpora). Für das zweite Samplingexperiment beschränken wir uns auf die 25 Autoren, die in unserer Sammlung mit den meisten Romanen vertreten sind, und ziehen 5000 zufällige Stichproben von je drei Romanen pro Autor (also ebenfalls 25×3 Texte). Für jede Stichprobe ermitteln wir über ein leave-one-out-Verfahren die Genauigkeit der beiden Verfahren.

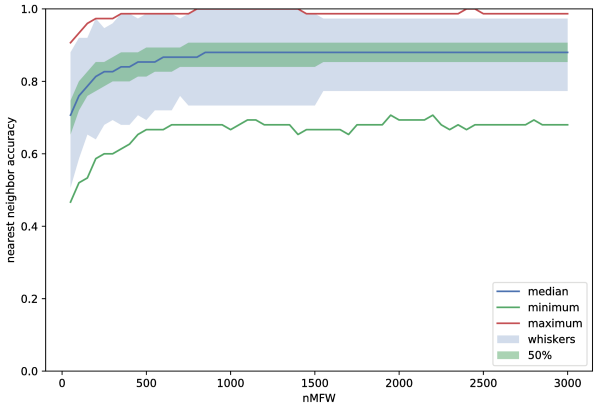
Aus Platzgründen berichten wir an dieser Stelle nur knapp die Ergebnisse des ersten Kürzungsexperiments und des ersten Samplingexperiments und beschränken und dabei auf auf die deutschen Daten. Die Ergebnisse des ersten Kürzungsexperiments, in dem nur der zu klassifizierende Text gekürzt wird, sind in der folgenden Abbildung dargestellt:





Wir vergleichen Cosine-Delta auf Basis der 3000 häufigsten Wörter mit N-Gram-Tracing auf Basis von Wort-1-bis-3-Grammen, Zeichen-4-bis-10-Grammen und Zeichen-11-Grammen. Bis zu einer Textlänge von 5000 Tokens liefern alle N-Gram-Tracing-Varianten bessere Ergebnisse als Delta, für längere Texte funktionieren Delta und N-Gram-Tracing auf Wort-N-Grammen am besten. Bei weniger als 2000 Tokens brechen die Ergebnisse für Delta ein, bei weniger als 1000 Tokens auch die für N-Gram-Tracing.

Die Ergebnisse des ersten Samplingexperiments zeigen, dass die Klassifikationsgenauigkeit bei beiden Verfahren großen Schwankungen unterworfen ist. Hier die Ergebnisse für Cosine-Delta:



Die Grafik zeigt, dass ab ca. den 1000 häufigsten Wörtern zwar im Mittel eine Klassifikationsgenauigkeit von rund 85% erreicht wird, allerdings mit enormen Schwankungen zwischen knapp über 60% und knapp unter 100%.

Die Ergebnisse für N-Gram-Tracing auf Basis von Wort-N-Grammen sehen ähnlich aus:

Durch die Kombination von Wort-1- bis Wort-3-Grammen wird zwar eine mittlere Klassifikationsgenauigkeit von über 70% erreicht, aber auch hier mit enormen Schwankungen.

Die Ergebnisse zeigen, dass N-Gram-Tracing auf kurzen Texten besser funktioniert als Cosine-Delta, allerdings werden für beide Verfahren längere Texte benötigt, als häufig verwendet werden. Die Wahl der Autoren im Vergleichskorpus und auch, wie das Poster zeigen wird, die Wahl der einzelnen Werke haben einen enormen und schwer vorhersehbaren Einfluss auf die Qualität der Autorschaftszuschreibung, deren Genauigkeit ohne weiteres um 20 Prozentpunkte schwanken kann. Im Licht dieser Erkenntnisse ist es durchaus fraglich, wie valide und generalisierbar bisherige Forschungsergebnisse auf dem Gebiet der Autorschaftsattribuierung sind.

Bibliographie

Argamon, Shlomo (2008): „Interpreting Burrows' delta: Geometric and probabilistic foundations“. In: *Literary and Linguistic Computing* 23/2: 131–47. https://doi.org/10.1093/llc/fqn003

Burrows, John (2002): „'Delta'—A measure of stylistic difference and a guide to likely authorship“. In: *Literary and Linguistic Computing* 17/3: 267–87. https://doi.org/10.1093/llc/17.3.267.

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*. https://doi.org/10.1093/llc/fqx023.

Grieve, Jack / Carmody, Emily / Clarke, Isobelle / Gideon, Hannah / Heini, Annina / Nini, Andrea / Waibel, Emily (in Begutachtung): „Attributing the Bixby Letter using n-gram tracing“. Eingereicht bei *Digital Scholarship in the Humanities* am 26. Mai 2017.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Improving Burrows' Delta – An empirical evaluation of text distance measures". In: *Digital Humanities 2015: Conference Abstracts*. http://dh2015.org/abstracts.

Smith, Peter W. H. / Aldridge, W. (2011): „Improving authorship attribution: Optimizing Burrows' delta method". In: *Journal of Quantitative Linguistics* 18/1: 63–88. https://doi.org/10.1080/09296174.2011.533591.

Denkmalpflege in der DDR. Analoge Netzwerke digital – Chancen und Möglichkeiten

Klemstein, Franziska
f.klemstein@gmail.com
Technische Universität Berlin, Deutschland

Die klassische Kunstgeschichte verzichtet noch heute weitestgehend auf die Möglichkeiten, die unsere digitale Welt uns bietet. Zwar werden digitale Werkzeuge bereits vielfältig genutzt, jedoch bisher häufig ohne ausreichende Reflexion und Rückkopplung in die Lehre. [1]

Innerhalb meines Dissertationsprojektes zum Thema „Denkmalpflege zwischen System und Gesellschaft – Netzwerke der Denkmalpflege im Sozialismus" habe ich es mir zum Ziel gesetzt sowohl eine technikgeschichtliche Methode zur Darstellung von Handlungen und Strukturen zu nutzen als auch analoge Netzwerke digital abzubilden.

Die Zielsetzung ist es, zum einen die Komplexität der denkmalpflegerischen Aufgaben abbilden zu können und zum anderen – und dies ist das Ziel des gesamten Dissertationsprojektes – unzutreffende Verkürzungen und Verallgemeinerungen in Bezug auf die Denkmalpflege in der DDR zu vermeiden, in dem die unterschiedlichen Akteure und Zeitphasen im Zusammenhang mit den konkreten Objekten und der Darstellung des Erfolgs oder Misserfolgs der Denkmalpfleger und Denkmalpflegerinnen innerhalb der DDR erfasst, dargestellt und abgefragt werden können. Fragekomplexe, die hierbei in den Blick genommen werden, sind u.a.: Welche Akteure haben an welchen denkmalpflegerischen Projekten gearbeitet oder waren involviert? Welche Bauaufgaben wurden zu welchen Zeiten besonders stark gefördert, diskutiert, unter Schutz gestellt oder zum Abriss freigegeben? Welche Akteure konnten zu welchem Zeitpunkt erfolgreich Belange der Denkmalpflege umsetzen?

Die technikhistorische Methode basiert auf dem von Wolfgang König entwickelten Akteur-Struktur-Modell (ASM), das eine Kombination von Handlungs- und Strukturtheorie darstellt. Innerhalb der Kunstgeschichte und Denkmalpflege fand diese Methode bisher jedoch kaum Beachtung. Die Anwendung dieses Modells innerhalb einer architekturhistorischen Arbeit, soll den Blick auf einen Themenbereich weiten, der bislang häufig nur auf Teilaspekte oder regionale Entwicklungen beschränkt wurde. Das Akteur-Struktur-Modell stellt dabei den Versuch dar, Handlungen und Strukturen strikt symmetrisch zu behandeln, da Strukturen aus Handlungen hervorgehen und Handlungen aus Strukturen. (König 2013a : 514) Dabei wird zwischen verschiedenen Handlungsebenen (Makro-, Meso-, Mikroebene) unterschiedenen. Strukturen stehen hingegen „für Tradition und für Dauer, für soziokulturelle Verfasstheiten, in denen sich die Akteure bewegen und bewegen müssen." (König 2013a : 512) Strukturen bilden somit den Handlungsrahmen oder Spielraum der handelnden Personen (Mikroebene), Organisationen (Mesoebene) oder auch der Regierungen (Makroebene), wobei deren Handlungen bestehende Strukturen sowohl stabilisieren als auch destabilisieren können.

Zugleich sollen mit der Anwendung des Modells auch seine Grenzen und Probleme aufgezeigt werden, die sich ergeben, wenn ein Modell aus einem anderen Wissenschaftsbereich für die Kunstgeschichte nutzbar gemacht wird. Obwohl Wolfgang König auf den Begriff des Netzwerkes verzichtet, möchte ich diesen innerhalb meines Dissertationsprojektes verwenden und folge hierbei Christoph Hubig, welcher vorschlägt, die Dynamik zwischen Akteuren und Strukturen mit Hilfe der Netzwerkmetapher zu modellieren. (König 2013b : 605 und Hubig 2013 : 546f.) Dies erscheint sinnvoll, da die Protagonisten der Denkmalpflege der DDR formale Beziehungen[2] miteinander unterhalten haben, welche die Dynamik innerhalb der scheinbar festen Strukturen, welche das sozialistische System geprägt beziehungsweise festgelegt hat, überhaupt erst möglich werden ließ. In diesem Sinne möchte ich auch die graphbasierte Datenbank neo4j nutzbar machen und den Netzwerkbegriff nicht nur als Metapher verwenden.

Jeder Teil unseres Lebens wird von zahlreichen Verbindungen geprägt, so auch die institutionellen wie auch persönlichen Netzwerke innerhalb

der Denkmalpflege der DDR. Es reicht mir jedoch nicht aus, diese Netzwerke lediglich aufzuzeigen. Vielmehr sollen die vernetzten Informationen (Personen, Dinge, Orte usw.) abgespeichert und durch unterschiedliche Abfragen in ihrer Vielschichtigkeit analysierbar sein. Die Informationen innerhalb eines Netzwerkes sollen dem entsprechend weder ignoriert noch in irgendeiner Weise zusammengefasst, sondern in ihrer Detailliertheit erfasst werden. Mehr noch als „gephi“ [3], das vor allem als Visualisierungstool von Netzwerken geeignet ist, bietet die Graph-Datenbank „neo4j“ [4] die Möglichkeit wenig strukturierte Daten, die stark vernetzt sind, darzustellen. Im Mittelpunkt sollen dabei die Beziehungen zwischen den Akteuren und den Objekten stehen, sodass hier ein weiterer Vorteil im Hinblick auf relationale Datenbanken zu sehen ist. Auch die flexible Datenmodellierung sowie die relativ benutzerfreundliche Abfragesprache „Cypher“, sind weitere Vorteile im Vergleich zu etablierten RDBM-Systemen.

Der Einsatz von neo4j erfolgt derzeit im Hinblick auf die konkreten und innerhalb der Dissertation umfassender dargestellten Fallbeispiele. Auf diese Weise werden sowohl der Mehrwert als auch die Besonderheiten des Einsatzes von „digital tools“ exemplarisch aufgezeigt.

Hierzu war es zunächst notwendig, die vorhandenen Quellen eingehend zu recherchieren, um sowohl die verschiedenen Akteure zu identifizieren, als auch Fallbeispiele auswählen zu können, die derzeit in neo4j erfasst, dargestellt und analysiert werden. Der derzeitige Stand umfasst dabei den ersten Untersuchungszeitraum vom 1952 bis 1961.

Langfristiges Ziel bzw. Wunsch ist es, neo4j nicht allein auf das Dissertationsprojekt zu beschränken, sondern im Kontext von Architektur, Städtebau und Denkmalpflege in der DDR zu „vervollständigen“, um die Verwendung von graphbasierten Datenbanken (oder auch anderen digitalen Techniken) nicht nur als Zusatz zum Forschungsprozess zu verstehen, sondern um aufzuzeigen, dass durch die Verwendung Forschungsergebnisse sichtbar werden, die bislang nicht oder nicht in diesem Umfang aufgezeigt werden konnten.

An seine Grenzen kommt neo4j (nach bisherigem Stand) bei der Darstellung der verschiedenen (politischen bzw. fachspezifischen) Positionen der einzelnen Akteure. Im Verlaufe der Zeit veränderten sich die Strukturen innerhalb derer die unterschiedlichen Akteure (Politiker, Kunsthistoriker, Historiker, Architekten usw.) agieren konnten. Dadurch veränderten sich auch die Handlungsmöglichkeiten des einzelnen, ebenso wie seine/ihre Position zu bestimmten Denkmalen oder früheren Entscheidungen und Entscheidungsfindungsprozessen. Auch Beziehungen unterlagen dadurch einem ständigen Wandel, die innerhalb von neo4j nicht abbildbar sind.

Anhand meines Posters möchte ich einerseits die Möglichkeiten aufzeigen, die sich durch die Nutzung digitaler Werkzeuge (gephi und neo4j), besonders im Hinblick auf die Darstellung und Analyse von Netzwerken, ergeben und andererseits einen Vergleich zwischen analogen und digitalen Methoden anstellen.

Fußnoten

1. Eine Ausnahme bildet in diesem Bereich der „Arbeitskreis digitale Kunstgeschichte“, dessen Mitglieder sich engagiert für einen reflektierten Einsatz digitaler Methoden einsetzen und dies bereits selbst umsetzen.
2. Formale Beziehung meint soziale Beziehungen. König lehnt den Netzwerkbegriff ab, da er teilweise „realistisch“ und teilweise modellistisch verwendet wird und es bei der modellistischen Verwendung nichts gibt, was sich nicht in Netze integrieren ließe. Allerdings ist die Netzmetapher bei Akteuren, die formale Beziehungen – also „echte“ Beziehungen im Sinne von sozialen Beziehungen – unterhalten, Königs Ansicht nach durchaus gerechtfertigt, weshalb ich den Netzwerkbegriff innerhalb meines Dissertationsprojekts durchaus für sinnvoll erachte.
3. https://gephi.org/
4. https://neo4j.com/

Bibliographie

Hubig, Christoph (2013): Strukturdynamik und/oder Netzdynamik – Die Rolle der Akteure, in: EWE 24/4: 545-547.

König, Wolfgang (2013a): Strukturen und Akteure – Ein Vorschlag zur Konzeptualisierung technisch-historischer Entwicklung, in: EWE 24/4: 505-516.

König, Wolfgang (2013b): Technik und Geschichte. Interdisziplinarität, Theorien und Modelle, in: EWE 24/4: 605-616.

Deutsche Geschichte-Digital: Ergebnisse der TEI-Konvertierung und Integration in Pilotprojekten

Hiebert, Matthew
hiebert@ghi-dc.org
German Historical Institute Washington

Lässig, Simone
laessigs@ghi-dc.org
German Historical Institute Washington

Witt, Andreas
andreas.witt@uni-koeln.de
Universität zu Köln

Das Deutsche Historische Institut in Washington (DHI) befindet sich in der Entwicklungsphase von " Deutsche Geschichte- *Digital* / German History- *Digital*" (DG-D), einer transatlantischen digitalen Initiative, um die wissenschaftlichen Bedürfnisse von HistorikerInnen im Kontext neuer historiografischer und technologischer Herausforderungen zu bewältigen. DG-D ist eine neue Infrastruktur zur Erleichterung der transnationalen historischen Wissensschöpfung für eine große Wissenschaftsgemeinschaft und eine wachsende Zahl von "Citizen Scholars", die bereits auf digitale Ressourcen des DHI angewiesen sind. Im Poster stellen wir zwei zentrale GH-Digital-Pilotprojekte und deren Integration in die DG-D "Knowledge Creation Environment" vor. Das erste ist "Deutsche Geschichte in Dokumenten und Bildern", unterstützt von der Deutschen Forschungsgemeinschaft (DFG), und das zweite ist Deutsche Geschichte *Intersections*, unterstützt durch das Europäische Wiederaufbauprogramm (ERP).

Die Planung für DG-D umfasste die Befragung von mehr als vierhundert WissenschaftlerInnen, die bereits mit digitalen Ressourcen arbeiten, welche vom DHI produziert wurden. Die umfassendste dieser Ressourcen ist die im Jahr 2003 gestartete digitale Quellensammlung "Deutsche Geschichte in Dokumenten und Bildern / German History in Documents and Images" (GHDI), die an deutsch- und englischsprachigen Universitäten weitläufig genutzt wird. Derzeit wird GHDI in Verbindung mit DG-D einem technischen und konzeptionellen Umbau unterzogen. Es enthält Tausende Seiten englischsprachiger Übersetzungen deutscher historischer Texte sowie Bilder und Karten, die von ca. 5.000 Besuchern pro Tag genutzt werden. Unsere Planung für DG-D beinhaltet auch weiterhin Konsultationen und Workshops mit ExpertInnen aus den Geschichtswissenschaften und den digitalen Geisteswissenschaften sowie die Gründung von Partnerschaften mit Institutionen und großen Initiativen, die unser Interesse für die Zukunft der Geschichte im digitalen Zeitalter teilen.

Die Deutsche Geschichte- *Digital*-Plattform befasst sich mit den Bedürfnissen der digitalen Forschung durch fünf Ziele und integrierte Arbeitspakete, die auf diese Ziele abgestimmt sind: Entdeckung, Analyse, Produktion, Bewahrung und Gemeinschaft. DG-D beinhaltet die Entwicklung eines Peer-Review-Index von wissenschaftlichen digitalen Objekten mit Dublin Core (DC) und CLARINs Component MetaData Infrastructure (CMDI) Standards über einen angepassten Blacklight (http://projectblacklight.org/) Technologie-Stack.

Für WissenschaftlerInnen, die historische digitale Projekte in Nordamerika entwickeln, gibt es keine interinstitutionelle Infrastruktur für die Speicherung ihrer Daten oder die Bereitstellung von Open Access. CLARIN-D, Teil der europäischen Forschungsinfrastruktur CLARIN, berät und unterstützt das DG-D-Projekt zur Erstellung eines Portals für CLARIN in Nordamerika am DHI Washington. Im Mittelpunkt dieses Prozesses steht die Implementierung eines Repository-Systems, das eine nachhaltige Speicherung des Inhalts und die Einbindung in eine digitale Umgebung ermöglicht, um den Zugriff, die Suche und die interoperablen Datenformate zu erleichtern. Unsere Partnerschaft mit CLARIN fördert den freien Zugang, die offene, kooperative Wissenschafts- und Wissenserzeugung im nordamerikanischen Kontext und ist ein wichtiger Bestandteil der gesamten Strategie der digitalen Geisteswissenschaften des DHI. Als Institut der Max Weber Stiftung sind wir auch in Partnerschaft mit DARIAH-DE. DARIAH-DE wird Web-Hosting und langfristige Bewahrung von DHI-Digitalprojekten in ihrer Gesamtheit, einschließlich der ersten Version der Webseite Deutsche Geschichte in Dokumenten und Bildern, zur Verfügung stellen.

Als kooperative Wissensplattform wird DG-D Redakteure, Forscher und Citizen Scientists bei der Entwicklung weiterer innovativer Online-Projekte zusammenbringen. Drei solcher Pilotprojekte befinden sich derzeit in der Entwicklung und beinhalten TEI und unsere Internationalisierung der Scalar 2.0 Plattform. DG-D verwendet Scalar 2.0 für das Baseline Content Management System, insbesondere aufgrund seiner Schnitt-

stellenfunktionen, Unterstützung für Resource Description Framework (RDF), Konnektivität zu externen Repositorien, Dublin Core (DC) Unterstützung, Hypothes.is Integration und sein Mehrfachpfad-Navigationssystem.

HistorikerInnen nutzen zunehmend digitale Geisteswissenschaften, um Daten zu analysieren und ihre Forschungsergebnisse darzustellen. Ein weiterer Vorteil der Speicherung von TEI-Digitalobjekten in einem CLARIN-Repository ist, dass eine ganze Palette von korpus-linguistischen analytischen Werkzeugen von WissenschaftlerInnen auf Textinhalte angewendet werden kann. In diesem Zusammenhang werden wir unsere Entwicklung von Scalar Adapters für den Anschluss an deutsche Repositorien und virtuelle Forschungsumgebungen (VREs) diskutieren.

Die DG-D-Plattform integriert die Blog-Aggregation, ein erweitertes Diskussionssystem, Community-orientierte Tools und Social Media, um miteinander kooperierende Wissensgemeinschaften zu erleichtern und Forschung zu öffnen. Dies ist ein wegweisender Aspekt unseres Projektes, das die Annahme von sozialen und gemeinschaftlichen digitalen Instrumenten durch HistorikerInnen in ihren Forschungsaktivitäten untersuchen wird. Wir wollen hierbei auch die einzigartige Rolle nutzen, die das DHI als Drehscheibe des transatlantischen wissenschaftlichen Dialogs und als eines großen Knotenpunkts in einem internationalen Netzwerk von HistorikerInnen spielt, um die Zusammenhänge zwischen verschiedenen Wissenschaftsgemeinschaften zu erleichtern.

Deutsche Geschichte- *Digital* Projekte bietet ein Modell für neue, quellenbasierte methodische Ansätze in den Geschichtswissenschaften. Die Initiative zielt darauf ab, durch digitale Instrumente, Standards und Methoden zur argumentbasierten Forschung beizutragen. Es fördert transnationale Ansätze in der historischen Forschung durch das Verfügbarmachen einer transnationalen technischen Plattform, die auf TEI und anderen aufkommenden Standards in den digitalen Geisteswissenschaften wie DC und RDF gründet.

Es unterstützt narratologische Komplexität in der Geschichtsschreibung und vermeidet redaktionelle Ansätzen, die eine singuläre Erzählung oder „ *Master Narrative*“ ergeben.

Bibliographie

Cohen, Daniel J., and Roy Rosenzweig (2006): *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press.

Cohen, Daniel J., Michael Frisch, Patrick Gallagher, Steven Mintz, Kirsten Sword, Amy Murrell Taylor, William G. Thomas III, and William J. Turkel (2008): “Interchange: The Promise of Digital History.” *Journal of American History*, volume 95, issue 2, 452-491.

“Diskussionsforum: Historische Grundwissenschaften und die digitale Herausforderung.” From H-Soz-Kult, November 15, 2015 (http://www.hsozkult.de/text/id/texte-2890).

Fiedler, Norman and Werthmann, Antonina and Stuehrberg, Maik and Schonefeld, Oliver and Bingel, Joachim and Witt, Andreas (2014): Research infrastructures in non-university research facilities. Research paper. Mannheim: Institute for German Language, 2014. 116 S.

Hiebert, Matthew, Lässig, Simone and Witt, Andreas (2017): German history-digital: A platform for transnational historical knowledge co-creation. In: Digital Humanities 2017, Conference Abstracts, McGill University & Université de Montréal Montréal, Canada August 8-11, 2017. Montréal: McGill University & Université de Montréal, 2017. p. 269-271

Hiebert, Matthew, Bowen, W. R., and Siemens, R.G (2015): “Implementing a social knowledge creation environment.” Scholarly and Research Communication, 6(3).

Mandić, Slobodan (2005): *Computerization and Historiography 1995-2005*. Belgrade: Belgrade Historical Society.

McCullough, Kelly, and James Retallack (2013): “Digital History Anthologies on the Web: German History in Documents and Images.” *Central European History*, volume 46, 346-361.

Ngai, Mae M (2012): “The Future of the Discipline: The Promise and Perils of Transnational History.” *Perspectives on History*, December 2012 (https://www.historians.org/publications-and-directories/perspectives-on-history/december-2012/the-future-of-the-discipline/promises-and-perils-of-transnational-history).

Patel, Kiran Klaus (2010): “Transnational History.” In *European History Online* (EGO), published by the Institute of European History (IEG). Mainz, 2010 (http://www.ieg-ego.eu/patelk-2010-en).

Patel, Kiran Klaus (2011): “Zeitgeschichte im digitalen Zeitalter: Neue und alte Herausforderungen.” *Vierteljahrshefte für Zeitgeschichte*, volume 59, issue 3, July, 331-51.

Putnam, Lara (2016): “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast.” *The American Historical Review*, volume 121, issue 2, April, 377-402.

Sahle, Patrick (2013): *Digital Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Volume 3 (Norderstedt).

Thomas, William G., III (2016): “Renegotiating the Archive: Scholarly Practice in the Digital Age.” (http://railroads.unl.edu/blog/?p=1195).

DH-Toolvergleich im Hinblick auf Texte historischer Sprachstufen

Aehnlich, Barbara
barbara.aehnlich@uni-jena.de
FSU Jena, Deutschland

Seidel, Henry
hnrseidel@gmail.com
HU Berlin, Deutschland

Mittlerweile versprechen zahlreiche Tools eine mehr oder minder problemlose Lemmatisierung und Annotierung mit Part-of-Speech-Tags von Texten; viele sollen auch für historische (oder andere nicht-standardisierte) Sprachdaten nutzbar sein. [1] Dabei birgt die Verarbeitung historischer Sprachdaten des Deutschen zahlreiche Probleme aufgrund des hohen Grads an Variation, insbesondere auf den Ebenen Phonologie und Graphematik, aber auch in den Bereichen der Morphologie, Syntax und Lexik. Bei einer automatischen Verarbeitung solcher Daten stellen vor allem die Variationen in Phonologie, Graphematik und Morphologie ein besonderes Hindernis dar.

Das Poster stellt verschiedene Werkzeuge überblicksartig vor und befasst sich genauer mit zwei Tools, deren Anwendung auf Texte nicht-standardisierter Sprachstufen exemplarisch anhand zweier Textsorten aufgrund bestimmter Kriterien verglichen wurde. Zum einen handelt es sich um gedruckte deutschsprachige Rechtstexte der Frühen Neuzeit, also aus der Rezeptionszeit des römischen Rechts in Deutschland, deren Sprache in einem Projekt erforscht werden soll, zum anderen um Fürstinnen-Briefe aus einem an der Friedrich-Schiller-Universität Jena erstellten digitalen Korpus. [2] In beiden Fällen weisen die Quellen frühneuhochdeutschen Sprachstand auf. Anhand ausgewählter Beispiele aus den vorliegenden Texten sollen zwei gängige elektronische Werkzeuge miteinander verglichen werden – EXMARaLDA und LAKomp.

EXMARaLDA wurde für das computergestützte Arbeiten mit überwiegend mündlichen Korpora entwickelt, wird aber regelmäßig auch für schriftliche Sprachdaten verwendet, so auch bei den *Frühneuhochdeutschen Fürstinnenkorrespondenzen im mitteldeutschen Raum*. Das Tool besteht im Wesentlichen aus einem Transkriptions- und Annotationseditor, einem Werkzeug zum Verwalten von Korpora und einem Such- und Analysetool. [3]

Das Werkzeug LAKomp [4] wurde im Projekt SaDA (Semiautomatische Differenzanalyse von komplexen Textvarianten) [5] entwickelt und dient der Aufbereitung eines historischen Korpus. Nach der Transkription können die Texte hier lemmatisiert und annotiert werden. Aufgrund der Besonderheiten bei frühneuhochdeutschen Handschriften und Drucken wird der Lemmatisierungs- und Annotationsvorgang komplett manuell durchgeführt. Dabei ist dem Benutzer mit LAKomp ein Werkzeug an die Hand gegeben, das ihn sehr schnell und präzise große Textmengen bearbeiten lässt und damit den Mehraufwand händischer Annotation nahezu ausgleicht.

Damit wurden zwei Werkzeuge ausgewählt, bei denen manuell annotiert werden muss, die aber dennoch bestimmte Unterschiede aufweisen, die für Nutzerinnen und Nutzer, die mit nicht-standardisierten Sprachdaten arbeiten, je nach Arbeitsziel vor- oder nachteilig sein können. So sind etwa bei LAKomp die halbautomatische Annotation auf der Grundlage des DWB und die Ausgabefunktion besonders gelungen, leider kann hier aber bisher nur lemmabasiert annotiert werden; bei EXMARaLDA ermöglichen die flexiblen Annotationskriterien eine besondere Breite von möglichen Annotationen, eine automatische oder halbautomatische Annotation des frühneuhochdeutschen Textmaterials ist jedoch bislang auch mit Hilfsmitteln wie dem Treetagger [6] nicht ohne weiteres möglich.

Die beiden genannten Tools werden auf dem Poster hinsichtlich ihrer Anwendbarkeit auf Texte historischer Sprachstufen anhand folgender Kriterien verglichen: Funktionalitäten, Nutzerfreundlichkeit (Technischer Support, Qualität des Handbuchs, Verständlichkeit der Benutzeroberfläche, Verfügbarkeit eines Editors, Umgang mit Metadaten, Exportmöglichkeiten …) und Nachnutzbarkeit. Das Poster wird diesen Tool-Vergleich anhand ausgewählter Beispiele aus einem Rechtsbuch sowie einem Fürstinnenbrief aus der Mitte des 16. Jahrhundert präsentieren und stellt somit Überblick und Evaluation der Werkzeuge gleichermaßen dar.

Fußnoten

1. In Auswahl: CATMA, CorA, EXMARaLDA, GATE, LAKomp, WebAnno.

2. https://archive.thulb.uni-jena.de/hisbest/content/below/index.xml?XSL.DisplayComponentBrowse=true ; http://www.laudatio-repository.org/repository/corpus/LAUDATIO%3AFuerstinnenkorrespondenz/TEI-header_version4_Schema7_2017-03-06T08%3A38%3A 26%3A247Z
3. Für genauere Informationen vgl. http://exmaralda.org/de/ueber-exmaralda/ und http://exmaralda.org/de/exmaralda-nutzer/ .
4. LAKomp steht für **L**emmatisierung, **A**nnotation, **Komp**aration.
5. http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/
6. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Bibliographie

B. Aehnlich / **S. Kösser** (2016): „Das Tool LAKomp und seine Anwendung auf Texte nichtstandardisierter Sprachstufen“. In: DHd 2016. Konferenzabstracts, Leipzig, S. 263–264.

V. Faßhauer (2017): „Compilation, Transcription, Multi-Level Annotation and Gender-oriented Analysis of a Historical Text Corpus:Early Modern Ducal Correspondences in Central Germany“. In: Advances in Digital Scholarly Editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp, hg. v. Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini & Dirk van Hulle. Leiden, S. 269–274.

A. Medek (*Gießler) / **M. Pöckelmann** / **T. Bremer** / **H.-J. Solms** / **P. Molitor** / **J. Ritter** (2015): „Differenzanalyse komplexer Textvarianten - Diskussion und Werkzeuge“. In: Informationsmanagement für Digital Humanities, hg. v. G. Heyer und A. Henrich. In: Datenbank-Spektrum 2015. http://dx.doi.org/10.1007/s13222-014-0173-y.

A. Leipold / **J. Ritter** / **H.-J. Solms**: „Neue Wege zu Textzeugenvergleich und Edition am Beispiel der Wundarznei des Heinrich von Pfalzpaint“. In: Jahrbuch für Germanistische Sprachgeschichte 2014, Band 5, Heft 1, S. 335-358.

D. Prutscher / **H. Seidel** (2012): „Mehrebenenannotation frühneuzeitlicher Fürstinnenkorrespondenzen“. In: G. Brandt (Hg.): Bausteine weiblichen Sprachgebrauchs. X. Texte – Zeugnisse des produktiven Sprachhandelns von Frauen in privaten, halböffentlichen und öffentlichen Diskursen vom Mittelalter bis in die Gegenwart. Stuttgart, S. 109–124.

Die illustrierte Postkarte und die digitalen Geisteswissenschaften – (Kulturerbe)objekt oder (Nachrichten)text

Koch, Carina
carina.koch@uni-graz.at
Universität Graz, Österreich

Die Postkarte als Forschungsobjekt

Denkt man an Postkarten, denkt man zunächst an Urlaub und Landschaftsmotive aus fernen Ländern. Die aktuelle Nutzung der Postkarte unterscheidet sich jedoch stark von der vor über hundert Jahren. Postkarten waren ein weit verbreitetes Kommunikationsmittel, vergleichbar mit Kurznachrichtendiensten von heute. Sie wurden aber auch gesammelt und getauscht. So vielfältig wie ihr Gebrauch ist auch ihre Nutzung in der geisteswissenschaftlichen Forschung. Postkarten werden seit den 1980er Jahren, als die Ethnologie sie zunächst für Kolonialismusforschungen vermehrt nutzte, als Quelle unterschiedlicher geisteswissenschaftlicher Disziplinen herangezogen. Mittlerweile dient das Massenmedium Postkarte dazu, Fragen aus der Geschichtswissenschaft, Sozial-, wie Kunst- und Fotografiegeschichte, den Literaturwissenschaften oder der Linguistik zu beantworten. Wurde zunächst in erster Linie bildwissenschaftlich gearbeitet, wird in den letzten zwanzig Jahren auch vermehrt auf die Relevanz der Postkarte als Bild-Text-Medium hingewiesen. Doch nicht nur die Motive dieses Hybridmediums, das zugleich Bild und Text beinhaltet, sondern auch die Natur des Massenmediums an sich, die Herstellung, der Vertrieb, die Zirkulation im politisch-gesellschaftlichen Kontext, Drucktechniken oder Auflagen sind für Forschungsfragen relevant (vgl. Holzer 2007, Tropper 2014).

Insbesondere im Hinblick auf die letzten beiden Aspekte ermöglichen die digitalen Geisteswissenschaften neue Forschungsperspektiven. Aufbauend auf digitalisierte und strukturiert erfasste Konvolute können anhand digitaler Methoden statistische Auswertungen, quantitative Analysen

und qualitative Untersuchungen durchgeführt werden.

Das Projekt

Das Projekt “Postkartensammlung GrazMuseum Online” zeigt, welche Wege für die Postkartenforschung denkbar sind, da im Projekt der Quellenwert der Postkarte an sich und die Aufbereitung für unterschiedliche Disziplinen im Mittelpunkt stehen. Zentrales Anliegen ist es, bildbasierte ebenso wie textbasierte Analysen zu gewährleisten und die Materialität der Postkarte - als Objekt, das produziert und gebraucht wurde, ernst zu nehmen. In diesem Sinne werden im Unterschied zu den meisten Postkarten, die online gezeigt werden (vgl. AKON, Bildarchiv ETH Schweiz), nicht nur die Bild-, sondern auch die Textseite präsentiert. Dadurch wird über bildwissenschaftliche Fragestellungen hinaus ein ganzheitlicher Blick auf das Medium und seine Verwendungsweisen ermöglicht. Für eine inhaltliche Recherche sind die Postkarten mit Schlagworten angereichert, wodurch thematische Zugänge - etwa zu einer Geschichte des Tourismus oder des Verkehrs möglich sind. Ebenso erfasst sind die Reproduktionsverfahren, die für Untersuchungen zu Druck- oder Fototechniken herangezogen werden können und auch Rückschlüsse auf Auflagenhöhe erlauben. Daneben sind auch die Produzenten der Karten erfasst. Damit ist erstmals die Möglichkeit gegeben, zielgerichtet nach Herstellern zu suchen und so auch einen Einblick in deren Sortimente und Produktionsweisen zu erhalten.

Für jede Postkarte gibt es einen Datensatz, der die zugehörigen Digitalisate und Metadaten beinhaltet und im digitalen Archiv GAMS (GAMS; Steiner/Stigler 2017) langfristig verfügbar gemacht wird.

Der Zugang zum Bestand erfolgt über klassische Suchmöglichkeiten (Volltext- und Facettensuche), Schlagwortwolken, geographische Karten (DARIAH GeoBrowser) und virtuelle Rundgänge (StoryMapJS). Die digitalen Faksimiles können stufenlos gezoomt und gedreht werden (OpenSeadragon Viewer), wodurch unterschiedliche Leserichtungen oder etwa Details im Bildmotiv betrachtet werden können. Anhand der Faksimile der Text- oder Adressseite werden auch Informationen vermittelt, die bei der Aufnahme der Metadaten im Zuge des Projektes nicht erfasst werden konnten (z.B. Schriftbilder oder Postwertzeichen).

Herausforderungen in der Datenmodellierung

Bei der Datenmodellierung steht man vor der Entscheidung, aus welcher Richtung man sich der Beschreibung der Postkarte nähern will. Auf der einen Seite steht die Mitteilung, die Text beinhaltet, der transkribiert, annotiert und angereichert werden kann. Auf der anderen Seite steht das Objekt, das unter spezifischen Gesichtspunkten eingeordnet, bewertet und inventarisiert wird. Dafür bieten sich zwei etablierte Standards an: LIDO und TEI. Letzteres eignet sich speziell für den Zugang über textuelle Inhalte. [1] LIDO ist ein Standardformat, das eigens für die Beschreibung von Kulturerbeobjekten, ihre Klassifikation, Identifikation und Einbindung von in Bezug stehenden Ereignissen (Entstehung, Nutzung, Archivierung, ...), entwickelt wurde. Für tiefere Texterschließungen ist dieser Standard nicht geeignet. [2] Da im Projekt die Beschreibung der Postkarte aus museologischer Sicht im Vordergrund steht und sich das Modellierungskonzept der Events (insb. für Postläufe) anbietet, sind die Daten nach LIDO modelliert und annotiert. Für Transkriptionen handschriftlicher Mitteilungen, die im skalierbaren Modell für spätere Bearbeitungen angedacht sind, ist das Schema um TEI Lite erweitert. Sollte künftig Text detaillierter kodiert werden, als TEI Lite es zuläßt, ist die Textedition in einem separaten TEI-Dokument angedacht.

Fazit

Anhand digitaler Methoden eröffnen sich neue Wege für das Hybrid- und Massenmedium “Postkarte”. Will man bei der Datenmodellierung dem ganzheitlichen Blick gerecht werden, erkennt man bald, dass die Einbeziehung mehrerer Standards nötig ist. Denn auch wenn man zunächst dazu tendiert sich an die Restriktionen eines Standards zu halten, bieten genau die digitalen Methoden den enormen Vorteil, dass man sich von mehreren Richtungen dem Objekt nähern kann. Auch wenn spätestens die Europeana zu der Entscheidung IMAGE oder TEXT zwingt.

Fußnoten

1. Zur weiteren Beschreibung von Postkarten mit der TEI s. Correspondance-SIG, Burnard, Lou 2014.
2. Obwohl das Element <inscriptionsWrap> für die strukturierte Erfassung von Transkriptionen

und Beschreibungen von Markierungen oder Texten etc. definiert ist, ist das nur sehr oberflächlich möglich.

Bibliographie

AKON - Ansichtskarten Online der Österreichischen Nationalbibliothek, https://akon.onb.ac.at/ , [letzter Zugriff 2017-08-29].

Bildarchiv ETH Schweiz, http://ba.e-pics.ethz.ch , [letzter Zugriff 2017-08-29].

Burnard, Lou (2014): TEI ODD workshop - Case Study: designing a TEI customization for a corpus of postcards, http://tei.it.ox.ac.uk/Talks/2014-10-odds/talk-03-cards.xml , [letzter Zugriff 2017-09-22].

GAMS - Geisteswissenschaftliches Asset Management System, http://gams.uni-graz.at , [letzter Zugriff 2017-08-29].

Holzer, Anton (2007): Naserümpfend am Postkartenstand. Was Ansichtskarten erzählen können (wenn man sie lässt), in: Hochreiter, Otto / Otti, Margareth (eds.): Hier ist es schön. Grazer Ansichtskarten. Ausstellungskatalog Stadtmuseum Graz, Salzburg: Fotohof (= Fotohof edition 93) 75-87.

LIDO - Lightweight Information Describing Objects, http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/ , [letzter Zugriff 2017-09-22].ademic press 10-41.

Postkartensammlung GrazMuseum Online, http://gams.uni-graz.at/gm , [letzter Zugriff 2017-08-29].

Steiner, Elisabeth / Stigler, Johannes (2014): GAMS and Cirilo Client: Policies, documentation and tutorial, latest update 2017-04-10, http://gams.uni-graz.at/docs , [letzter Zugriff 2017-09-22].

TEI - Text Encoding Initiative, http://www.tei-c.org/index.xml, [letzter Zugriff 2017-08-29].

Tropper, Eva (2014): Illustrierte Postkarten - ein Format entsteht und verändert sich, in: Tropper, Eva / Starl, Timm (eds.): Format Postkarte - Illustrierte Korrespondenzen, 1900 bis 1936, Wien: new academic press.

Die Macht der Daten - vom konsequenten Umgang mit Forschungsdaten

Gálffy, Andreas
agalffy@smail.uni-koeln.de
Universität zu Köln, Deutschland

Kamphausen, Julian
julian.kamphausen@uni-koeln.de
Universität zu Köln, Deutschland

Kronenwett, Simone
simone.kronenwett@uni-koeln.de
Universität zu Köln, Deutschland

Wieners, Jan G.
jan.wieners@uni-koeln.de
Universität zu Köln, Deutschland

Theorie: Lehrveranstaltung in Kooperation mit nestor

Der vorgeschlagene Beitrag in der Kategorie "Poster" veranschaulicht ausgewählte Inhalte und Ergebnisse, welche im Rahmen einer Lehrveranstaltung zum Thema "Forschungsdatenmanagement und Langzeitarchivierung" (FDM und LZA) von den TeilnehmerInnen erarbeitet und präsentiert wurden. Die Übung wurde im Sommersemester 2017 am Institut für Digital Humanities (Historisch-Kulturwissenschaftliche Informationsverarbeitung) an der Universität zu Köln in Kooperation mit nestor (Network of Expertise in long-term Storage and availability of digital Resources in Germany), dem Kompetenznetzwerk für Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen in Deutschland, durchgeführt (Lehrveranstaltung 2017; Nestor 2017).

Praxis: Experten aus dem FDM- und LZA-Sektor

In weiterer Zusammenarbeit mit insgesamt sieben externen FDM- und LZA-Experten aus der Praxis (u.a. von GESIS, Hochschulbibliothekszen-

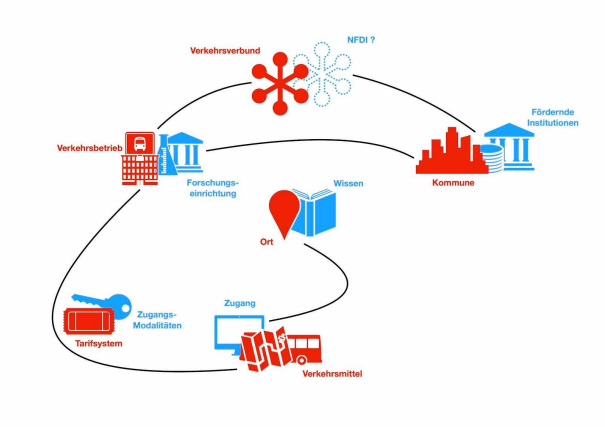
trum NRW (HBZ), Data Center for the Humanities (DCH), Forschungsdatenzentrum IANUS, Stiftung Rheinisch-Westfälisches Wirtschaftsarchiv zu Köln (RWWA)) wurden anhand konkreter Beispiele und Anwendungsfälle Fragen erörtert und Probleme dargestellt, wie Forschungsdaten nachhaltig bereitgestellt und langfristig gesichert werden können und welche Arbeitsabläufe dabei notwendig sind. In diesem Kontext wurde besonders ein Blick auf die sich derzeit etablierenden Data Professional Berufe geworfen (z.B. Data Librarian, Data Curator, Data Archivist, Data Manager, Data Scientist, Data Journalist etc.) sowie ihre jeweiligen Aufgabenprofile mit den eingeladenen Fachleuten diskutiert.

Aufgabe: Kurzartikel für neue nestor-Publikationsreihe

Vor diesem Hintergrund wählten die TeilnehmerInnen im Laufe der Lehrveranstaltung ein theoretisches oder praxisnahes Thema aus dem FDM- bzw. LZA-Bereich und verfassten dazu einen Beitrag, welcher in der neu entstandenen studentischen Kurzartikelreihe von nestor veröffentlicht werden wird. Von den insgesamt 20 Artikeln, deren Themenbreite von der digitalen Sicherung archäologischer Rekonstruktionen bis hin zur Transformation XML-basierter Daten reicht, werden im vorgeschlagenen Tagungsbeitrag ausgewählte Beispiele vorgestellt und visualisiert.

Beispiel: ”Informationserlebnis als Reiseerlebnis? Ein Vergleich eines Informationsinfrastrukturverbundes mit einem öffentlichen Personennahverkehrsverbund”

Sehr deutlich wird der derzeitige Zustand des Forschungsdatenmanagements, wenn man ihn mit einem öffentlichen Personennahverkehrsverbund (wie der Verkehrsverbund Rhein-Sieg, www.vrsinfo.de) vergleicht. Bis zum 1. September 1987 existierten im Nahverkehr um Köln redundante Verbindungen, die nur teilweise aufeinander abgestimmt waren - viele Verkehrsbetriebe leisteten nebeneinander ihren Dienst. Mit dem Verkehrsverbund war es möglich, Angebote aufeinander abzustimmen und Ortsverbindungen besser zu koordinieren. Eine ähnliche Entwicklung im Forschungsdatenmanagement wird in diesem Aufsatz vorgeschlagen, mit Verweis auf die Nationale Forschungsdateninfrastruktur (NFDI) (Rat 2016).



Bildquelle: Eigene Darstellung

Zusammenfassung

Es bleibt festzuhalten, dass die Digital Humanities ein sehr weites und schwer ab- und eingrenzbares Feld darstellen. Hiervon sind Forschungsdatenmanagement und Langzeitarchivierung nur zwei - augenscheinlich kleine - Themenbereiche, die ihrerseits heterogener kaum sein können. Diese Heterogenität begründet sich in der Mannigfaltigkeit der Daten und kommt in der Vielfalt der Themenbereiche und Tätigkeitsfelder zum Ausdruck. Dies wurde in der Lehrveranstaltung an zahlreichen Fallbeispielen deutlich und spiegelt sich erneut in den eingereichten Studierendenbeiträgen wider.

Des Weiteren lässt sich eine hohe Dynamik in diesen Feldern spüren. Schon allein in der stetig steigenden Zahl von Stellenanzeigen, die nach qualifizierten Fachkräften werben, aber auch in der Zahl der Ansätze, Normierungsversuche und Leitlinien, die zu einem bewusst nachhaltigen Umgang mit Daten anhalten sollen. Diese werden sowohl von oben herab (top-down) vorgeschrieben, als auch seitens der betroffenen Institutionen (bottom-up) betrieben. Das Ziel ist allen gemeinsam: der stetig steigenden Zahl von Forschungsdaten Herr zu werden, Forschungsdaten langfristig zu sichern und nachhaltig bereitzustellen.

Bibliographie

Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (Hrsg.) (2011): “Handbuch Forschungsdatenmanagement”, Bad Honnef: Bock + Herchen, http://www.forschungsdatenmanagement.de/ [zuletzt aufgerufen am 20.09.2017]

Lehrveranstaltung “Forschungsdatenmanagement und Langzeitarchivierung” (2017), Universität zu Köln, Institut für Digital Humanities (Historisch-Kulturwissenschaftli-

che Informationsverarbeitung), Sommersemester 2017, Homepage, http://www.lehre.jan-wiener-s.de/sosem17-fdm-lza/ [zuletzt aufgerufen am 20.09.2017]

Nestor (Network of Expertise in long-term Storage and availability of digital Resources in Germany) (2017), Homepage, http://www.langzeitarchivierung.de/Subsites/nestor/DE/Home/home_node.html , [zuletzt aufgerufen am 22.09.2017]

Neuroth, Heike / Strathmann, Stefan / Oßwald, Achim et al. (Hrsg.) (2012): "Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme", Version 1.0, Boizenburg: Verlag Werner Hülsbuch, http://nestor.sub.uni-goettingen.de/bestandsaufnahme/ [zuletzt aufgerufen am 20.09.2017]

Rat für Informationsinfrastrukturen (2016): "Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland", Göttingen, http://www.rfii.de/?wpdmdl=1998 [zuletzt aufgerufen am 20.09.2017]

Ray, Joyce M. (Hrsg.) (2014): "Research Data Management. Practical Strategies for Information Professionals", West Lafayette/Indiana

Die Max-Bense-Collection. Digitale Re-Publikation von Erstausgaben mit erweiterten Plattformfunktionen

Schlesinger, Claus-Michael
claus-michael.schlesinger@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Texte

Die Max-Bense-Collection versammelt rund 700 Texte von Max Bense in der Fassung der Erstdrucke, darunter viele, die nur schwer verfügbar sind. Der überwiegende Teil der Texte in der Sammlung besteht aus Zeitungs- und Zeitschriftenartikeln. Die Max-Bense-Collection schließt damit eine Lücke der Ausgaben mit ausgewählten Schriften (Bense 1965, 1997-1998 und 2000), die die Zeitungs- und Zeitschriftenartikel, nicht zuletzt aus Platzgründen, die gedruckte Bücher mit sich bringen, oftmals übergehen. Die Texte, die die Max-Bense-Collection zur Verfügung stellt, sind dabei auch deshalb interessant, weil in ihnen oft die argumentative und materiale Vorbereitung der umfangreicheren Schriften erkennbar ist und die publizistischen Auseinandersetzungen, etwa im Nachgang seiner Aesthetica (Bense 1965), weitere Überlegungen und ergänzende Antworten, etwa auf Rezensionen, provozieren.

Die Schriften von Max Bense, darunter auch die kürzeren Texte, die die Max-Bense-Collection zur Verfügung stellt, sind für die kritische historische Reflexion der Digital Humanities aus zwei Gründen relevant. Erstens war Max Bense seit den 1950er Jahren an der Entwicklung einer "Informationsästhetik" (Bense 1965, 1968, 1969) beteiligt, die einen wichtigen theoretischen Rahmen für den Einsatz und die kritische Reflexion quantitativer Methoden in den Geisteswissenschaften und in der Kunst lieferte. Die Verbindung von Informationstheorie und Ästhetik, die sich im Begriff der Informationsästhetik anzeigt und die unter diesem Label maßgeblich vollzogen wird, ist nicht zuletzt Bedingung für die Entstehung der Digital Humanities.

Das Projekt verfolgt drei Ziele: Erstens wird durch die Re-Publikation der Zugang zu den Texten für die Bense-Forschung und für kulturgeschichtliche Forschungen zur Informationsästhetik erleichtert. Die digitale Re-Publikation der Erstausgaben ist mit Blick auf die grafische und typografische Gestaltung der Texte forschungsrelevant, weil Gestaltung und Design für Benses Theoriebildung maßgebliche Konzepte sind. Zweitens ermöglicht die entwickelte Plattformlösung den beteiligten Forscher*innen und Projektpartner*innen eine erweiterte Nutzung der digitalisierten Texte für projekt- und ausstellungsbegleitende Präsentationen und Kommentierungen ausgewählter Texte auf eigenen, projektbasierten Unterseiten der Plattform. Drittens dient die Plattform damit der Vernetzung von Nutzer*innen aus Forschung und beteiligten Kulturerbeinstitutionen, weil die Nutzung des Materials nachvollziehbar wird.

Technik

Die Plattform nutzt die Software *Omeka*, die vom Roy Rosenzweig Center for History and New Media (CNHM) entwickelt wird, mit projektspezifischen Änderungen und Ergänzungen. Omeka ist eine Web-Publishing-Plattform mit einfachen Möglichkeiten zur Verwaltung der präsentierten Daten und Inhalte. Darüber hinaus bietet Omeka kuratorische Funktionen (siehe Abschnitt Funktionen). Die Omeka-Installation der Max-Bense-

Collection wird auf einem LAMP-Stack betrieben (Linux, Apache, MySQL, PHP). Für die Max-Bense-Collection werden eine Reihe verfügbarer Plugins genutzt sowie ein eigenes Theme (Schlesinger 2018), das auf dem "Berlin"-Theme der Omeka-Entwickler basiert und "Stuttgart" heißt, mit einem Augenzwinkern zur sogenannten Stuttgarter Schule, die sich 1950-1970 in Stuttgart entwickelte und an der Max Bense konzeptuell und praktisch (Herausgaben, Ausstellungen usw.) maßgeblichen Anteil hatte. (Klütsch 2012)

Die Metadaten der publizierten Texte sind in Dublin Core codiert, die Bild- und Textdaten liegen im Format PDF-A mit zusätzlichem Textlayer vor (automatische Texterkennung) und lagern auf einem hochverfügbaren und redundanten Speichersystem der Universitätsbibliothek Stuttgart. Die Langzeitverfügbarkeit der Daten ist damit zumindest technisch gesichert und hängt vor allem an einer institutionellen Stabilisierung von Personalverhältnissen.

Funktionen

Das Projekt zielt auf eine wissenschaftliche Nutzung der Texte und eine Stärkung des Kontakts zwischen Kulturerbeinstitutionen und Forschung. Die Texte werden in mehreren wissenschaftlichen Fächern bearbeitet, etwa in der Literaturwissenschaft, Kunstwissenschaft, Philosophie, Wissenschaftsgeschichte.

Neben Möglichkeiten zur Ansicht/Lektüre der Texte mit Suche in Textdaten und Metadaten bietet die Max-Bense-Collection kuratorische Funktionen, also die Präsentation und Kommentierung ausgewählter Texte/Objekte auf einer thematisch fokussierten Webseite.

An dieser Stelle wird der autorzentrierte Aufbau der Sammlung zum Problem, weil viele Forschungen nicht autorzentriert arbeiten, sondern etwa Diskurse, Konstellationen, soziale Zusammenhänge, Gattungen usw. bearbeiten. Die Abbildung solcher Zusammenhänge ist perspektivisch wünschenswert. Die Sammlung ist deshalb erweiterbar angelegt und soll in Zukunft auch Materialien aufnehmen, die nicht von Max Bense autorschaftlich gezeichnet sind, und die etwa über Forschungs- oder Ausstellungsprojekte mit einem sachlichen und/oder historischen Zusammenhang erschlossen werden.

Am Projekt beteiligt sind das Zentrum für Kunst und Medien Karlsruhe (ZKM), das Stuttgart Research Center for Text Studies (SRCTS) und die Abteilung Digital Humanities der Universität Stuttgart.

Bibliographie

Bense, M. (1965) Aesthetica. Einführung in die neue Ästhetik. Agis-Verlag, Baden-Baden.

Bense, M. (13.3.1968) "Ist Kunst berechenbar?". Frankfurter Allgemeine Zeitung, S 22.

Bense, M. (1969) Einführung in die informationstheoretische Ästhetik. Grundlegung und Anwendung in der Texttheorie. Rowohlt, Reinbek bei Hamburg.

Bense, M., Walther, E. (1997-1998) Ausgewählte Schriften in vier Bänden. J.B. Metzler, Stuttgart.

Bense, M. (2000) Radiotexte : Essays, Vorträge, Hörspiele. Winter, Heidelberg.

Klütsch, C. (2012) Information Aesthetics and the Stuttgart School. In: Higgins, H.: Mainframe experimentalism: early computing and the foundations of the digital arts, University of California Press, Berkeley, S 65-89.

Schlesinger, C. / Zittel, C. / Zentrum für Kunst und Medien Karlsruhe (2018): Max-Bense-Collection, http://max-bense.ub.uni-stuttgart.de, Zugriff: 15.1.2018

Schlesinger, C. (2018): Omeka-Theme "Stuttgart", https://github.com/cmcx/theme-stuttgart, Zugriff: 15.1.2018

Digital Dylan – Computergestützte Analyse der Liedtexte von Bob Dylan (1962 – 2016)

Sippl, Colin
Colin.Sippl@stud.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Fuchs, Florian
Florian.Fuchs@stud.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Burghardt, Manuel
manuel.burghardt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Hintergrund und Zielsetzung

Am 13. Oktober 2016 gab die Schwedische Akademie bekannt, dass sie den Nobelpreis in Literatur an Bob Dylan „für seine poetischen Neuschöpfungen in der großen amerikanischen Songtradition“ verleihen werde. Dass Dylan als Musiker und Songwriter den Literaturnobelpreis erhielt wurde mitunter sehr kontrovers diskutiert. Auf Kritik stieß z.B. die unzulässige Herauslösung von Dylans Texten aus der Musik und die Deutung seiner Lieder als Gedichte. Unbestritten ist nichtsdestotrotz Bob Dylans Rolle als einer der einflussreichsten Musiker des 20. Jahrhunderts.

Die (welt-)politischen Entwicklungen, die das Schaffen Dylans inspirierten, sind im Kontext seines Wirkens umfassend diskutiert worden, unter anderem in „Bob Dylan und die sechziger Jahre: Aufbruch und Abkehr“, und liefern noch immer Diskussionsstoff, wie etwa eine in jüngerer Zeit erschienene Arbeit von Taylor & Israelson (2015) über Dylans politische Einflüsse zeigt. Erfolgte die Beschäftigung mit Dylans Werk bislang allenfalls episodisch, so muss eine systematische Analyse des Gesamtwerks als Desiderat gelten, welches in gewisser Weise bereits von Bob Dylan selbst formuliert wurde:

> All these people who say whatever it is I’m supposed to be doing – that’s all gonna pass, because, obviously, I’m not gonna be around forever. That day’s gonna come when there aren’t gonna be any more records, and then people won’t be able to say ‘Well, this one’s not as good as the last one.’ **They’re gonna have to look at it all (eigene Hervorhebung).** And I don’t know what the picture will be, what people’s judgement will be at that time. I can’t help you in that area. – *Bob Dylan*

Dieser Beitrag erprobt, inwiefern mithilfe digitaler Methoden im Sinne des *Distant Reading*-Paradigmas ein neuer Zugang zu Dylans Gesamtwerk geschaffen werden kann. Bezugnehmend auf das Konferenzmotto einer „Kritik der Digitalen Vernunft“ soll untersucht werden, wo die Grenzen und Möglichkeiten eines solchen digitalen Analyseansatzes liegen, indem überprüft wird, ob sich bestehende qualitative Einteilungen von Dylans Werk in unterschiedliche Schaffensperioden auch anhand statistisch signifikanter Wörter (Rayson, Berridge & Francis 2004) und N-Gramme (Evert 2005) belegen lassen.

Stand der Forschung

Bereits vor der Auszeichnung Dylans mit dem Nobelpreis in Literatur, waren seine Texte Gegenstand wissenschaftlicher Betrachtungen im Sinne des *Close Reading* (vgl. etwa Brown 2014). Brown unterteilt Dylans Werk in einzelne Phasen und verknüpft diese jeweils mit allgemeinen, zeitgeschichtlichen Ereignissen sowie biographischen Meilensteinen des Künstlers. Taylor & Israelson (2015) gehen einen ähnlichen Weg, versuchen jedoch Dylans Werk abseits verbreiteter politischer Einordnungen zu betrachten. Etwas anders ausgerichtet ist die Arbeit von Wissolik et al. (1994): Hierbei handelt es sich um eine Art Wörterbuch, in dem Namen und Gegenstände, die in Dylans Texten auftauchen, erläutert werden.

Eine umfassende Untersuchung Dylans Werks mithilfe computerbasierter Methoden fand sich bis zum Abfassungszeitpunkt des vorliegenden Texts nicht. Allerdings sind quantitative Verfahren zur stilistischen und inhaltlichen Analyse von Liedtexten in den Digital Humanities durchaus verbreitet. So beschreiben etwa, wie mithilfe von N-Gramm-Modellen ein Liedtext-Korpus anhand der Merkmale Textlänge, Textstruktur, Wortschatz und Semantik analysiert werden kann, um eine automatisierte Genrezuordnung vornehmen zu können. Daneben existieren stilometrische Untersuchungen von Liedern oder Gedichten, die sich mit der Berechnung von autoren- und genrespezifischen Merkmalen befassen, wie z.B. Suzuki & Hosoya (2014), die japanische Pop-Songs analysieren.

Forschungsmethodik

Bezugsrahmen der Analyse: Schaffensphasen Bob Dylans

Den analytischen Bezugsrahmen dieser Studie stellt die phasenweise Einteilung von Dylans Schaffen nach Brown (2014) dar. Brown unterscheidet dabei neun unterschiedliche Phasen, die mit „Becoming Bob Dylan“ (1960-1964) beginnen und vorläufig mit „Bob Dylan Revisited“ (2000-2012) enden. Diese Stilphasen umfassen z.B. Dylans Hinwendung zum Christentum oder seine elektronische „Folk Rock“-Phase (vgl. Brown, 2014).

Korpus und Datenaufbereitung

In dieser Arbeit wurde ein Korpus bestehend aus 452 Liedtexten mit einem Umfang von 133.045

Tokens untersucht, die Bob Dylan zwischen den Jahren 1962 und 2016 auf Studio-Alben veröffentlicht hat. Die Liedtexte und Metainformationen wie etwa Titel, Album und Jahr stammen von der Plattform *LyricsWikia*[1]. Da es sich bei *LyricsWikia* um ein community-gestütztes Projekt handelt, erfolgte vorab ein stichprobenartiger Abgleich einzelner Lieder mit den offiziellen Texten nach, wobei keinerlei Abweichungen festgestellt werden konnten.

Das Korpus wurde weiterhin mit Methoden der Computerlinguistik aufbereitet, insbesondere unter Verwendung des *Python Natural Language Toolkits* (NLTK[2]). Die Verarbeitung des Korpus umfasst die grundlegende Lemmatisierung mit dem *WordNet-Lemmatizer* (Teil des NLTK) und eine Stoppwortbereinigung (NLTK-Stoppwortliste für Englisch mit eigener Erweiterung[3]) sowie die Wortartenannotation mithilfe des *Stanford Log-linear Part-of-Speech-Taggers*. Da Dylan in seinen Texten häufig umgangssprachliche Formulierungen wie etwa verkürzte Gerundformen (bspw. „savin“, „swimmin“) verwendet, wurde für den POS-Tagger ein Modell verwendet, welches auf der Grundlage von Twitter-Texten trainiert wurde und gute Ergebnisse für Texte mit nichtstandardisiertem Vokabular und Slang liefert.

Korpusvergleich – Assoziationsmaße und Referenzkorpus

Assoziationsmaße

Ein etabliertes Verfahren, um aus einem Korpus spezifische Wörter zu extrahieren, ist ein direkter Korpusvergleich mit dem *Log-Likelihood*-Test, der sich zum Vergleich von Korpora unterschiedlicher Größe besonders eignet (Rayson, Berridge, & Francis, 2004). Damit können Wörter, die im untersuchten Korpus mit einem signifikanten Frequenzunterschied zum Referenzkorpus auftreten, als Schlagworte betrachtet werden. Dies kann besonders aussagekräftige Ergebnisse liefern, wenn zusätzlich eine POS-Filterung erfolgt, womit sich beispielsweise signifikante Nomen oder Verben eines Korpus berechnen lassen. Darüber hinaus wurde eine Berechnung von N-Grammen in Form von Bi- und Trigrammen umgesetzt. Die berechneten N-Gramme lassen sich in der Web-App unter der Wahl eines Assoziationsmaßes, wie dem *Chi Quadrat-Test, dem Jaccard-Test, dem Poisson-Stirling-Test, dem Likelihood Ratio-Test sowie dem Pointwise Mutual Information-Test* anzeigen. Dabei liefert jedes Verfahren zur N-Gramm-Berechnung eigene spezifische Ergebnisse. Dieser Freiraum wird ganz bewusst erhalten, um die verschiedenen Facetten eines Texts, die ein Assoziationsmaße jeweils anzeigt, für die spätere Datenanalyse nutzen zu können.

Referenzkorpus

Als Referenzkorpus dient das mündliche Subkorpus des *Open American National Corpus* (OANC; American National Corpus Project), welches insgesamt 3.862.172 Tokens umfasst. Das Korpus enthält viele Belege aus der mündlichen Kommunikation und eignet sich dadurch in besonderer Weise als Vergleichskorpus für Dylans Texte, die wie bereits beschrieben einen hohen Anteil umgangssprachlicher Formulierungen und Slang-Ausdrücke enthalten.

Beim Korpusvergleich kann entweder das gesamte Dylan-Korpus mit dem Referenzkorpus verglichen werden, oder mit den jeweiligen Dylan-Subkorpora, also bspw. all seinen Texten aus den 1970er-Jahren oder aus der ersten Schaffensperiode „Becoming Bob Dylan“ (1960-1964). Ein Vergleich der einzelnen Dylan-Subkorpora zum Gesamtwerk ist ebenso möglich. Letztere Option wird z.B. genutzt, um anhand jeweils signifikanter Wörter die einzelnen Schaffensperioden nach Brown (2014) zu überprüfen und damit die grundsätzliche Eignung solch quantitativer Verfahren zur Identifikation thematischer Verschiebungen zu untersuchen. Die Ergebnisse dieses Korpusvergleichs sind, zusammen mit allen anderen Ergebnissen der angewandten Analyseverfahren, in einer interaktiven Webanwendung über unterschiedliche Visualisierungen (Balkendiagramm, *treemap, wordcloud, Tabelle*) für weitere Interpretationen zugänglich[4]. Wie schon bei den Assoziationsmaßen, so gilt auch hier, dass jede Visualisierungsform eine bestimmte Perspektive auf die Berechnungsergebnisse eröffnet.

Ergebnisse

Im direkten Vergleich des gesamten Dylan-Korpus (1962-2016) mit dem OANC-Referenzkorpus treten einige interessante, signifikant-häufige Wörter im Werk Dylans hervor. Die von Bob Dylan verwendeten Adjektive erzeugen in der Gesamtschau tendenziell eher eine bedrückende Stimmung (*blind, weary, lonely, drunken, scared, restless, ragged*). Bei den Substantiven mischen sich unter viele Personen- und Ortsnamen auch religiöse Begriffe (*soul, heaven, devil, eden, prayer, paradise*). Viele der übrigen Begriffe sind erwartungsgemäß typisch für Folk-Musik (*levee, rooster, train*), was sich wiederum durch die Wahl des

Referenzkorpus, das verschiedenartige mündliche Textquellen enthält, erklären lässt (Rayson, Berridge, & Francis, 2004: 8).

Die Analyse signifikant-häufiger Wörter für die einzelnen Schaffensphasen Dylans liefert Ergebnisse mit hoher Aussagekraft. So fällt etwa für die Phase „The Changing of the Guard" (1978-1981), in der sich Dylan dem Christentum hinwendet, auf, dass das Vokabular tatsächlich viele christliche Motive aufweist (*lord, Jesus, devil, altar, faith, confession, grace, power, serve*). Insgesamt nimmt der Anteil an „düsterem" Vokabular in dieser Phase ab, verschwindet jedoch nicht komplett (bspw. *shot, destruction*). Der Anteil an hoffnungsvollen Wörtern nimmt hingegen zu (bspw. *beginning, ready, arise, wake, thank*). Bei den übrigen Schaffensphasen fallen die Ergebnisse jedoch mitunter wesentlich weniger deutlich aus.

Ein differenziertes Bild ergibt sich für die N-Gramm-Analyse, was einerseits der Vielfalt an verfügbaren Methoden zur Berechnung und andererseits den unterschiedlichen N-Gramm-Längen geschuldet ist. Die Ergebnisse für Bigramme mit Hilfe des *Pointwise-Mutual-Information*-Tests (PMI) erschienen dabei am geeignetsten, um die thematischen Schwerpunkte von Dylans Schaffensphasen nach nachzuvollziehen. So findet das PMI-Verfahren im Subkorpus der Phase „The Changing of the Guard" (1978-1981) Bigramme wie *close prayer, name lucifer, jesus good, jesus bone oder arise upon*, die eindeutig religiöse Bezüge in Dylans Texten dieser Phase veranschaulichen. Generell fällt jedoch die Dominanz von Refrain-Versen in den Liedern bedeutend ins Gewicht (z.B. *knock heaven door*), was die Qualität der Ergebnisse insbesondere bei den Trigrammen beeinflusst.

Diskussion

Im Sinne einer Kritik der Digitalen Vernunft bleibt demnach festzuhalten, dass sich Methoden der computergestützten Textanalyse und des statistischen Korpusvergleichs grundsätzlich dafür eignen, einen inhaltlichen Gesamtüberblick zu einem Liedtext-Korpus zu erhalten. Es können damit diachrone Entwicklungen des Wortschatzes und Verlagerungen thematischer Schwerpunkte als grobe Tendenzen aufgezeigt werden, um das Bild des Gesamtwerks zu ergänzen. Ein solcher Ansatz eignet sich demnach gut für die initiale Thesengenerierung und kann in gewisser Weise die Funktion eines Empfehlungs- bzw. Hinweissystems für erklärungsbedürftige Stellen [5] in den Geisteswissenschaften übernehmen.

Die Identifikation konkreter Schaffensperioden, ausschließlich auf Basis signifikant häufiger Wörter ist aber – zumindest für das Werk Dylans – nicht ohne Weiteres erfassbar. Bei den N-Grammen zeigt sich, dass im Falle von Dylans Texten methodenübergreifend und mit zunehmender N-Gramm-Länge meist keine brauchbaren Ergebnisse erzielt werden konnten. Dies ist ein Hinweis darauf, dass die hier präsentierten Analysemethoden, die für andere Textsorten wie bspw. Parlamentsprotokolle bereits erfolgreich eingesetzt werden konnten (vgl. Sippl et al. 2016), auf Liedtexte nur eingeschränkt anwendbar sind. Ein möglicher Kritikpunkt am hier beschriebenen Vorgehen mag zudem das verwendete OANC-Referenzkorpus sein, welches trotz hoher Anteile mündlicher Kommunikation doch nur beschränkt vergleichbar mit der Textsorte „Liedtext" ist. Für künftige Vergleichsstudien böte sich ggf. ein Vergleich mehrerer unterschiedlicher Künstler und deren Liedtexte an, also bspw. Bob Dylan vs. Johnny Cash.

Fußnoten

1. http://lyrics.wikia.com, alle Hyperlinks dieses Dokuments wurden zuletzt abgerufen am 10.01.2018
2. Verfügbar unter http://www.nltk.org/
3. Filterung von Stoppwörtern, wie „hey", „ah", „yeah", und Verkürzungen, wie „'ve", „'s" etc.
4. https://www.colin-sippl.de/dylan (Klick auf den Analyse-Button rechts oben)
5. Diesen Gedanken äußerte Hubertus Kohle auf der #DigiCampus-Tagung im Juni 2017 in München, vgl. https://twitter.com/8urghardt/status/876725916487036928.

Bibliographie

American National Corpus Project (2015a): *American National Corpus. Frequency Data*. http://www.anc.org/data/anc-second-release/frequency-data/ [Letzter Zugriff 10. März 2017].

American National Corpus Project (2015b): *The Open American National Corpus (OANC)*. http://www.anc.org/ [Letzter Zugriff 10. März 2017].

Brown, Donald (2014): *Bob Dylan: American troubadour*. Lanham, Md. [u.a.]: Rowman & Littlefield.

Cott, Jonathan (2006): *Bob Dylan, the essential interviews*. New York: Wenner Books.

Derczynski, Leon et al. (2013): "Twitter part-of-speech tagging for all: Overcoming sparse

and noisy data", in: *Proceedings of the Recent Advances in Natural Language Processing* September, 198–206. http://www.derczynski.com/sheffield/papers/twitter_pos.pdf [Letzter Zugriff 9. März 2017].

Dylan, Bob (2016): *The lyrics: 1961-2012*. New York: Simon & Schuster.

Evert, Stefan (2005): "The Statistics of Word Cooccurrences, Word Pairs and Collocations", in: *Unpublished doctoral dissertation Institut fur maschinelle Sprachverarbeitung Universitat Stuttgart* 98: August 2004, 353. http://en.scientificcommons.org/19948039 [Letzter Zugriff 3. März 2017].

Fell, Michael / Sporleder, Caroline (2014): "Lyrics-based Analysis and Classification of Music", in: *International Conference on Computational Linguistics* 25: 23–29, 620–631.

Geisel, Sieglinde (2016): *Bob Dylan - Literaturnobelpreisträger wider Willen*. Deutschlandradio Kultur. http://www.deutschlandradiokultur.de/bob-dylan-literaturnobelpreistraeger-wider-willen.1005.de.html?dram:article_id=373494 [Letzter Zugriff 7. März 2017].

Rayson, Paul / Garside, Roger (2000): "Comparing corpora using frequency profiling", in: *Proceedings of the workshop on Comparing Corpora* 1–6.

Rayson, Paul / Berridge, Damon / Francis, Brian (2004): "Extending the Cochran rule for the comparison of word frequencies between corpora", in: *JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles:* 1–12.

Schmidt, Mathias R. (1983): *Bob Dylan und die sechziger Jahre: Aufbruch und Abkehr*. Frankfurt am Main: Fischer Taschenbuch Verlag.

Sippl, Colin / Burghardt, Manuel / Wolff, Christian / Mielke, Bettina (2016): Korpusbasierte Analyse österreichischer Parlamentsreden. In: *Netzwerke: Tagungsband des 19. Int. Rechtsinformatik Symposions IRIS 2016: 25.- 7. Feb. 2016, Univ. Salzburg, S. 139-148.*

Suzuki, Takafumi / Hosoya, Mai (2014): "Computational Stylistic Analysis of Popular Songs of Japanese Female Singer-songwriters", in: *Digital Humanities Quarterly* 8: 1, .

Svenska Akademien (2016): *Der Nobelpreis in Literatur des Jahres 2016*.

Taylor, Jeff / Israelson, Chad (2015): *The Political World of Bob Dylan: Freedom and Justice, Power and Sin*. New York: Palgrave Macmillan.

Toutanova, Kristina / Klein, Dan / Manning, Christopher D (2003): "Feature-rich part-of-speech tagging with a cyclic dependency network", in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03),* 252–259. http://nlp.stanford.edu/~manning/papers/tagging.pdf [Letzter Zugriff 3. März 2017].

Wissolik, Richard David / McGrath, Scott. / Colaianne, A. J. (1994): *Bob Dylan's words: a critical dictionary and commentary*. Greensburg, PA: Eadmer Press.

Digitale Wissenschaft – Eine Podcastreihe

Loebel, Jens-Martin
loebel@bitgilde.de
bitGilde IT Solutions UG, Deutschland

Hahn, Carolin
kontakt@carolinhahn.de
Beste Worte GmbH

Projektvorstellung

Die Methoden der Digital Humanities sind ebenso zahlreich wie die Fachbereiche, in denen sie zur Anwendung kommen. Ob Archäologie oder Literaturwissenschaft, Soziologie oder Geschichte: Unser Podcast will über Theorien und praktische Anwendungsbereiche der Digital Humanities informieren und diskutieren; er will neugierig machen und die Angst vor der Anwendung neuer, noch ungewohnter Forschungsmethoden nehmen.

Die Rezeptionsform „Podcast" soll Studierende auf eine noch weitgehend ungewohnte Weise ansprechen: Wir möchten sie in ihrem Alltag erreichen, sie zur Entdeckung neuer Forschungsmethoden anregen und gleichzeitig den kreativen Umgang mit ihrem je eignen Forschungsthema fördern.

Umgekehrt soll die Aufmerksamkeit für bereits existierende DH-Projekte und -Studiengänge erhöht werden, um den Forschungsnachwuchs neugierig zu machen und DH-Vorhaben mehr Reichweite zu ermöglichen. Dabei soll der Eindruck, dass die DH in weiten Teilen eine algorithmen- und werkzeuggetriebene Wissenschaft sei, dem kritischen Anspruch der Geisteswissenschaften gegenübergestellt werden. Wir führen u. a. Interviews mit Wissenschaftler/-innen und Praktikern und bereiten Forschungsdiskurse auf.

Themen

Wir möchten einen umfassenden Einblick in die Methoden und Anwendungsbereiche der Digital Humanities geben und die Forschungswerkstätten beleuchten. In Form von Interviews werden in den ersten Folgen Mitarbeitende verschiedener Universitäten und Institute interviewt und zu ihren aktuellen Forschungsprojekten befragt. Weitere Kooperationen sind in Planung. Doch auch "Praktiker" wie Editionswissenschaftler, Bibliothekare, Data Scientists etc. sollen nach ihren berufspraktischen Erfahrungen gefragt werden. Wie können hier digitale Methoden über die universitäre Forschungspraxis hinaus zum Einsatz kommen? Ziel ist zudem, laut über mögliche Kritikpunkte nachzudenken: Welche Potenziale haben digitale Methoden, welche Defizite des Analogen können damit ausgeglichen werden? Welche Gefahren ergeben sich? Geht der Methode überhaupt eine Theorie voraus?

Ergänzend zum Podcast bauen wir ein Glossar in Form des sogenannten "Wissensblogs" auf: Parallel zu den Audio-Beiträgen informieren wir hier ganz grundlegend über konkrete Werkzeuge und Methoden. Die hier gelieferten Informationen werden in den einzelnen thematisch passenden Folgen ggf. aufgegriffen und dem Publikum vorgestellt. Hier geben wir dem wissenschaftlichen Nachwuchs Tools an die Hand, die fachübergreifend nützlich sein können. So beleuchten wir zum Beispiel Fragen der Lizenzierung von Online-Veröffentlichungen, digitale Präsentationsmöglichkeiten, einheitliche Daten-Referenzsysteme etc.

Eine weitere Informationsquelle soll die Rubrik "Empfehlungen" darstellen, die ebenfalls über die Webseite zu erreichen ist. Hier werden hilfreiche Bücher und andere Publikationen rezensiert.

Der Podcast wird kostenlos über die üblichen Kanäle wie iTunes etc. und unsere Website verbreitet.

Das Poster soll eine Kostprobe geben und vor allem Lust aufs Zuhören machen. Ziel ist es, Aufmerksamkeit zu schaffen und ggf. neue Interviewpartner zu finden. Zudem sind kurze Hörproben geplant.

Key Facts

- Dauer pro Folge: 45 Minuten
- 6 Folgen pro Jahr
- Arbeitsaufwand pro Folge: ca. 25 Stunden
- Webauftritt: http://www.digitale-wissenschaft.de
- Twitter: @DigiWissen
- Verwendete Technik/Software: Blue Yeti Podcasting Mikrofon, Ultraschall
- Reichweite über: iTunes, Webseite, Twitter, Facebook, DH-Liste
- Finanzierung: Crowdfunding via Patreon

Aufbau

- Dialog zwischen Jens-Martin Loebel und Carolin Hahn
- 30 Minuten Interview mit einer dritten Person (je nach Rubrik)
- 15 Minuten Vor- und Nachbesprechung
- Einführung in ein zum Thema passendes Wissensgebiet im Bereich Digital Humanities, das im Blog ausführlicher nachgelesen werden kann

Zielgruppe

Als Informationsportal adressieren wir explizit Studierende und interessierte Laien.

Relevante Informationen auf dem Poster

- Key Facts, Beteiligte, Konzept, Distibutionskanäle und Zielgruppe
- Rubriken und Programm 2018
- Audio-Effekt: Einbau einer dezenten Tonspur im Poster per Lautsprecher, Hörproben zum Mitnehmen

Bibliographie

Geoghegan, Michael / Dan Klass (2007): Podcast Solutions. The Complete Guide to Audio and Video Podcasting. New York: Springer.

Hagedorn, Brigitte (2006): Podcasting: Konzept | Produktion | Vermarktung. Köln: mitp Verlags GmbH.

Podcast der Helmholtz-Gemeinschaft: https://resonator-podcast.de/ [letzter Zugriff 15. September 2017].

Podcast der Universität Wien: http://medienportal.univie.ac.at/uniview/podcast-audimax/ [letzter Zugriff 15. September 2017].

Physik-Podcast im ORF: http://www.physikalischesoiree.at/ [letzter Zugriff 15. September 2017].

Digital Medievalist: A Web Community for Medievalists working with Digital Media

Franzini, Greta
gfranzini@etrap.eu
Georg-August-Universität Göttingen

Fischer, Franz
franz.fischer@uni-koeln.de
Universität zu Köln

Kestemont, Mike
mike.kestemont@uantwerpen.be
Universiteit Antwerpen

Digital Medievalist is an international web community for medievalists working with digital media. Established in 2003 by a group of volunteers [1] and before the arrival of Facebook and Twitter [2], the goal of *Digital Medievalist* would also become that of the popular social networks: to connect people around the world providing them with an exchange platform. But where Facebook and Twitter were driven by relationships, *Digital Medievalist* was driven by interest.

The very first Digital Humanities disciplinary-focus community of practice, *Digital Medievalist* sought to meet the increasingly sophisticated demands faced by creators of digital projects working with medieval content. Among its initial community-building activities, *Digital Medievalist* started an homonymous open access scholarly journal, which is still active today, and commissioned the publication of short tutorials on its website to guide interested scholars through the basics of text encoding, web development and manuscript digitisation, to mention but a few. The benefits brought by *Digital Medievalist* became so evident that other, similar web communities began to emerge, such as the *Digital Classicist* [3] (Mahony, 2017) and *Digital Victorianist*.

Over time, and as new technologies rapidly developed, *Digital Medievalist*'s didactic component was superseded by free online courses, moodles and web tutorials, shifting its focus toward the dissemination of scholarly research to the widest possible audience. The *Digital Medievalist* community has continued to gather importance since its founding and today serves a number of disciplinary fields, including Digital Humanities, Medieval Studies and Auxiliary Sciences, Cultural Heritage, Archaeology, Literary Studies, History, Linguistics, and Museum and Archival Science.

Membership to *Digital Medievalist* is open to anyone with an interest in its subject matter, regardless of skill or previous experience in Digital Humanities or medieval studies. Participants range from novices contemplating their first project to many of the pioneers in the field. The entire *Digital Medievalist* community counts over 1,500 members worldwide.

The current activities and assets of *Digital Medievalist* include:

- The *Digital Medievalist* mailing list: 1,272 (as of Sept. 14th 2017) list members use this platform to ask for advice, discuss problems, and share any kind of information related to the field of medieval studies. The list's collegial atmosphere encourages a variety of conversations.
- The *Digital Medievalist* journal: The community's online, open access, refereed journal publishes original research and scholarship, notes on technological topics (standards, tools, software, etc.), commentary pieces discussing developments in the field, bibliographic and review articles, and project reports. The journal is funded in part through grants provided by the University of Lethbridge School of Graduate Studies and recently joined the Open Library of Humanities (OLH), a non-profit organisation dedicated to publishing open access scholarship with no author-facing article processing charges. Funded by an international consortium of libraries OLH has built a sustainable business model in order to make scholarly publishing fairer, more accessible, and rigorously preserved for the digital future.
- The *Digital Medievalist* website: The community's online presence provides comprehensive information about the organisation including membership, structure and bylaws. It also provides announcements and an up-to-date list of recent and upcoming conferences, colloquia, workshops and training events relevant to (digital) medieval studies. The website also invites members to write blog-posts for several thematic series.
- The *Digital Medievalist* Facebook group with over 1,600 members and a Twitter presence to widen the scope and impact of scholarly communication, and to disseminate best practice,

data and knowledge pertaining to digital medieval studies (Ross, 2012; Terras, 2012).

In 2017, *Digital Medievalist* joined the European Alliance for Social Sciences and Humanities (EASSH) as a learned society in order to increase the visibility of the *Digital Medievalist* community.

DHd 2018 provides the ideal venue to expose *Digital Medievalist* to a large German-speaking community of scholars. The *Digital Medievalist* website averages 2,760 views per day from Germany alone; the *Digital Medievalist* conference representatives are eager to speak to practitioners in Germany to better understand how *Digital Medievalist* is meeting their needs and how it can improve. Additionally, we think that *Digital Medievalist* can still serve as an example for community building. Its history and current state demonstrates how the interest in digital methods intersects with the use of digital communication tools, and is thus maybe an archetypical example of Digital Humanities.

The poster will outline the aforementioned activities and will serve as a conversation starter to establish connections with relevant initiatives, start reflection on the role of this and similar activities, collect feedback and continue fostering as wide a geographical coverage as possible.

Fußnoten

1. Daniel Paul O'Donnell, Peter Baker, James Cummings, Martin Foys, Murray McGillivray, Dot Porter, Roberto Rosselli Del Turco, and Elizabeth Solopova. *Digital Medievalist* was founded with direct financial support from the Faculty of Arts and Science at the University of Lethbridge, the Curriculum Redevelopment Centre (now the Teaching Centre) at the University of Lethbridge, the *Image, Text, Sound, and Technology (ITST)* programme of the Social Sciences and Humanities Research Council of Canada (SSHRC).
2. Facebook was founded in 2004 and Twitter in 2006.
3. For a discussion on the relationship between *Digital Medievalist* and *Digital Classicist*, see Bodard and O'Donnell (2008).

Bibliography

Bodard, G., O'Donnell, D. (2008) 'We are all together: On publishing a Digital Classicist issue of the Digital Medievalist journal', *Digital Medievalist*, 4. DOI: http://doi.org/10.16995/dm.18

Digital Medievalist website: https://digitalmedievalist.wordpress.com/

Digital Medievalist journal: https://journal.digitalmedievalist.org/

Digital Medievalist mailing list: https://digitalmedievalist.wordpress.com/mailing-list/

Digital Medievalist on Facebook: https://www.facebook.com/groups/49320313760/

Digital Medievalist on Twitter: https://twitter.com/digitalmedieval

European Alliance for Social Sciences and Humanities: http://www.eassh.eu/

Mahony, S. (2017) 'The Digital Classicist: Building a Digital Humanities Community', *Digital Humanities Quarterly*, 11(3). At: http://www.digitalhumanities.org/dhq/vol/11/3/000335/000335.html

Open Library of Humanities: https://www.openlibhums.org/

Open Scholarly Communities on the Web, ISCH COST Action A32. At: http://www.cost.eu/COST_Actions/isch/A32

Ross, C. (2012) 'Social media for digital humanities and community engagement', In C. Warwick, M. Terras and J. Nyhan (eds.) *Digital Humanities in Practice*. Facet Publishing, pp. 23-46.

Terras, M. (2012) 'The Impact of Social Media on the Dissemination of Research: Results of an Experiment', *Journal of Digital Humanities*, 1(3). At: http://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras/

Digital vs. Humanities. Didaktische Aufbereitung digitaler Methoden für die klassischen Geisteswissenschaften im Projekt forTEXT

Jacke, Janina
janina.jacke@uni-hamburg.de
Universität Hamburg, Deutschland

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

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

Einleitung

Computergestütztes Arbeiten kann geisteswissenschaftliches Forschen auf unterschiedlichste Weise befördern und bereichern. Dennoch müssen wir in unserem Arbeitsalltag und in Gesprächen mit Kolleginnen und Kollegen[1] immer wieder feststellen, dass viele traditioneller arbeitende Geisteswissenschaftler digitalen Methoden noch immer mit Skepsis begegnen.[2] Dies liegt nicht zuletzt daran, dass in den Geisteswissenschaften zahlreiche Methoden zum Einsatz kommen, von denen nur einigen wenigen eine derart formalisierte Arbeitsweise naheliegt, wie sie im Rahmen der Digital Humanities oft verfolgt wird.

Das lässt sich gut am Beispiel der Literaturwissenschaft illustrieren: Digitale Methoden werden bisher vornehmlich von Literaturwissenschaftlern genutzt, die an strukturellen oder anderen formalen Aspekten literarischer Texte interessiert sind (beispielsweise an narrativen Strukturen, Figurennetzwerken etc.).[3]

Ihrem traditionellen Selbstverständnis nach ist die Literaturwissenschaft allerdings zentral an komplexen und innovativen *Interpretationen* literarischer Texte interessiert[4] – und wie diese durch digitale Methoden der Textanalyse befördert werden können, ist nicht evident.

Damit digitale Methoden eine breitere Akzeptanz finden, ist es deswegen notwendig, den Nutzen dieser Methoden auch für stärker hermeneutisch ausgerichtete geisteswissenschaftliche Forschungsfragen zu reflektieren. Unserem (weiten) Verständnis von “Hermeneutik” entsprechend handelt es sich bei hermeneutischen Forschungsfragen um Fragen, die auf die (holistische) Auslegung bzw. Deutung von Texten gerichtet sind (vgl. bspw. Spörl 2004: 128). In literaturwissenschaftlichen Zusammenhängen spielen dabei insbesondere Fragen nach Funktion bzw. Wirkung bestimmter Textelemente oder des Gesamttextes eine Rolle, ebenso wie die In-Beziehung-Setzung des Textes mit bestimmten Kontexten. Diese Forschungsfragen sollten exponiert im Zusammenhang mit der Entwicklung von Tools, digitaler Forschungsumgebungen und vor allem didaktischer Konzepte zur Vermittlung von DH-Methoden berücksichtigt werden. Diese Forderungen werden bisher jedoch nicht in zureichendem Maße erfüllt.[5]

Wir möchten in diesem Beitrag das aktuelle Projekt forTEXT (2017–2020) vorstellen, das der Vermittlung, Aufbereitung und Bereitstellung von Mitteln zur computergestützten Textanalyse insbesondere für hermeneutisch arbeitende Geisteswissenschaftler gewidmet ist. Im Folgenden sollen in diesem Zusammenhang zunächst die unterschiedlichen konzeptionellen Dimensionen (Abschnitt 2) sowie anschließend erste inhaltliche Ergebnisse des Projekts präsentiert werden (Abschnitt 3).

Dimensionen des forTEXT-Projekts

Paradigmen

Das Anfang 2017 gestartete DFG-Projekt *forTEXT. Literatur digital erforschen* (http://www.fortext.net) hat die Entwicklung einer digitalen Forschungsumgebung zum Ziel, die im Rahmen der qualitativen Analyse und Interpretation von Texten genutzt werden kann. Das Augenmerk bei der Gestaltung dieser Umgebung liegt insbesondere auf zwei Aspekten:

(a) Orientierung an genuin geisteswissenschaftlichen Arbeitsweisen: Es geht, ganz im Sinne des geisteswissenschaftlichen Selbstverständnisses, um die Unterstützung der genuin *interpretativen* Auseinandersetzung mit Texten. In anderen Worten: forTEXTs Fokus liegt *nicht* ausschließlich auf der statistischen Auswertung von Texten, wie es sonst im Rahmen von DH-Methoden zur Textanalyse oft der Fall ist. [6] Auf diese Weise soll gewährleistet werden, dass traditioneller arbeitende Geisteswissenschaftler die Umgebung tatsächlich zur digitalen Unterstützung vertrauter Methoden der Textanalyse und -interpretation nutzen können und ihnen keine Hinwendung zur statistischen Textanalyse abverlangt wird.

(b) Niedrigschwelliger Zugang: Geisteswissenschaftler sollen die digitale Forschungsumgebung *intuitiv* und weitgehend ohne technische Vorkenntnisse nutzen können. Hierzu trägt zum einen die Tatsache bei, dass forTEXT ein individualisiertes Empfehlungssystem zur Verfügung stellt, das geisteswissenschaftlichen Nutzern Vorschläge unterbreitet, welche digitalen Ressourcen, Routinen und Tools für ihr Projekt hilfreich sein könnten (siehe auch Abschnitt 2.2). Nutzer werden also nicht einfach mit einem unüberschaubaren digitalen Angebot alleingelassen, dessen potenziellen Nutzen für die eigene Fragestellung sie sich erst noch selbst erschließen müssen. Zum anderen stellt forTEXT leicht verständliche Beschreibungen zu digitalen Methoden und Korpora zur Verfügung, ebenso wie intuitiv bedienbare Benutzeroberflächen für digitale Werkzeuge zur Textanalyse und -interpretation (siehe auch Abschnitt 2.3). Auf diese Weise können DH-Methoden ohne technisches Know-how sowie ohne das

aufwändige Studieren von Nutzerhandbüchern eingesetzt werden.

In den folgenden Unterabschnitten sollen sowohl forTEXTs Empfehlungssystem als auch die drei Komponenten des Informationsrepositoriums (*Routinen, Ressourcen und Tools*) kurz etwas detaillierter vorgestellt werden.

Individualisiertes Empfehlungssystem

Für Geisteswissenschaftler, die noch nicht wissen, auf welche Weise digitale Methoden der Textanalyse und -interpretation ihre eigene Forschung unterstützen können, bietet forTEXT ein individualisiertes Empfehlungssystem in Form eines digitalen Fragebogens an (siehe Abb. 1).

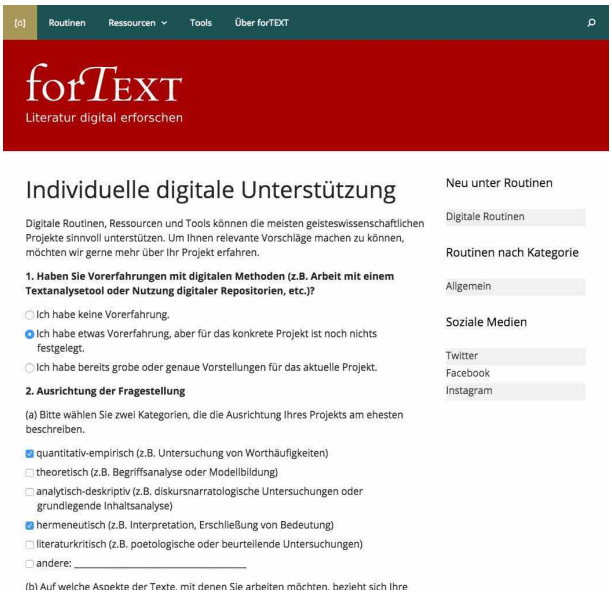


Abb. 1: forTEXTs individualisiertes Empfehlungssystem für digitale Textuntersuchung (Ausschnitt des Prototyps)

Hier können die Nutzer beispielsweise angeben, ob sie schon Vorerfahrungen mit digitalen Methoden der Textanalyse gemacht haben, in welchem Zustand sich ihr Textkorpus befindet, unter welcher Fragestellung sie ihre Texte untersuchen wollen und welcher literaturtheoretischen Schule sie sich zuordnen.[7] Als Output erhalten die Forscher, angepasst an die von ihnen gemachten Angaben, eine Liste mit Vorschlägen zu digitalen Korpora, Methoden und Werkzeugen, die zu ihrer Fragestellung und Arbeitsweise passen. Das Empfehlungssystem also eine individualisierte Kompilation aus forTEXTs Inhalten und Verzeichnissen, die im Folgenden kurz vorgestellt werden sollen.

Routinen, Ressourcen und Tools

forTEXTs digitale Forschungsumgebung ist in drei Bereiche gegliedert.

(a) Routinen: Im Teilbereich *Routinen* finden sich zum einen Informationstexte zu digitalen Methoden, die der Analyse und Interpretation von Texten dienen (bspw. zu taxonomiebasiertem Annotieren, zur Textanalyse mittels individualisierter Abfragen auf Text- und Annotationsdaten, zu Topic Modeling etc.), sowie zu vorbereitenden Prozeduren wie der Digitalisierung von Texten. In diesen Informationstexten finden sich darüber hinaus Links zu digitalen Tools (s.u.), mithilfe derer die fraglichen Methoden ausgeführt werden können.

Zum anderen werden unter *Routinen* auch didaktische Texte (d.h. Lerneinheiten und Lehrmodule) zur Verfügung gestellt. Die Lerneinheiten dienen der selbstständigen Aneignung bestimmter digitaler Methoden und Tools, während die Lehrmodule didaktisches Material für Lehrende zur Verfügung stellen, die auf 90-minütige Workshopsituationen zugeschnitten sind (siehe auch Abschnitt 3). Die Entwicklung neuer Lehrmodule wird sich dabei an den im Projektverlauf akquirierten Bedarfen der Nutzer orientieren.

(b) Ressourcen: Unter *Ressourcen* ist ein Verzeichnis digital nutzbarer Korpora zu finden. Hierunter fallen sowohl hochqualitativ digitalisierte Textkorpora als auch inhaltlich annotierte Korpora, die nachgenutzt werden können. Einige annotierte Korpora werden im Rahmen von forTEXT selbst bzw. affiliierten Projekten produziert.[8] Das Verzeichnis enthält darüber hinaus informative Beschreibungen der gelisteten Ressourcen.

(c) Tools: Im Bereich *Tools* findet sich schließlich eine kommentierte Liste digitaler Werkzeug-Suites bzw. Funktionskomponenten, mithilfe derer unterschiedliche textanalytische und -interpretatorische Operationen durchgeführt oder unterstützt werden können. Darüber hinaus sollen in forTEXT auch eigene Funktionskomponenten entwickelt werden. Hierzu gehört vornehmlich die Weiterentwicklung des Textannotations- und Analyseprogramms CATMA (http://www.catma.de) – aber auch die Entwicklung von graphischen Step-by-step-Benutzerschnittstellen für bestehende Tools.[9] Dies soll es technisch weniger versierten geisteswissenschaftlichen Nutzern ermöglichen, hochfunktionale Tools einzusetzen, ohne sich umfangreich einarbeiten zu müssen.

Alle Verzeichnisse und Einträge aus den drei forTEXT-Bereichen Routinen, Ressourcen und Tools können von Nutzern eigenständig durchsucht und aufgerufen werden – oder es erfolgt ein an-

geleiteter Zugriff durch die Nutzung des Empfehlungssystems.

Im verbleibenden Teil dieses Beitrags möchten wir etwas genauer auf einen Lehrmodulentwurf eingehen, das dem forTEXT-Bereich *Routinen* zuzuordnen ist. Anhand dieses Moduls soll beispielhaft gezeigt werden, wie in forTEXT die Paradigmen der Orientierung an einer genuin geisteswissenschaftlichen Arbeitsweise und des niedrigschwelligen Zugangs umgesetzt werden.

Lehrmodule Manuelles Annotieren

Zwei von forTEXTs 90-minütigen Lehrmodulen sind der Vermittlung der digitalen Methode des *manuellen Annotierens* gewidmet. Die Methode des Annotierens stellt einen guten Brückenschlag zur traditionelleren geisteswissenschaftlichen Arbeitsweise dar – schließlich gehört das Anbringen von Notizen in zu interpretierenden literarischen Texten seit jeher zur literaturwissenschaftlichen Arbeitspraxis.[10] In den Lehrmodulen zum *digitalen* Annotieren gilt es nun, zum einen an die bereits bekannte Praxis anzuknüpfen und zum anderen deutlich zu machen, inwiefern die digitale Unterstützung es ermöglicht, bekannte Arbeitsprozesse deutlich effektiver durchzuführen, oder gar ganz neue Arbeitsweisen eröffnet, die das jeweilige Forschungsziel befördern.

Um diese Anforderungen umzusetzen, sieht forTEXT zwei Lehreinheiten zum manuellen Annotieren vor, von denen wir die erste im Folgenden kurz vorstellen möchten.

In der Einheit *Taxonomiebasiertes Annotieren* wird schrittweise gezeigt, wie mithilfe des Annotations- und Analysetools CATMA freie Kommentare in literarischen Texten angebracht sowie analysiert und systematisiert werden können. Die Systematisierung freier Kommentare kann wiederum als Grundlage dienen, um eine Annotationstaxonomie zu entwickeln, die dann für eine noch feinere und zielgerichtetere Analyse des literarischen Textes genutzt werden kann.

Das verwendete Programm CATMA ist hierbei in zweifacher Hinsicht auf die forTEXT-Paradigmen abgestimmt: Es bietet eine intuitiv bedienbare Benutzeroberfläche und unterstützt den freien, undogmatischen und genuin interpretativen Zugang zu Texten, während es zugleich Optionen stärkerer Formalisierung bereithält (vgl. Abb. 2).[11]

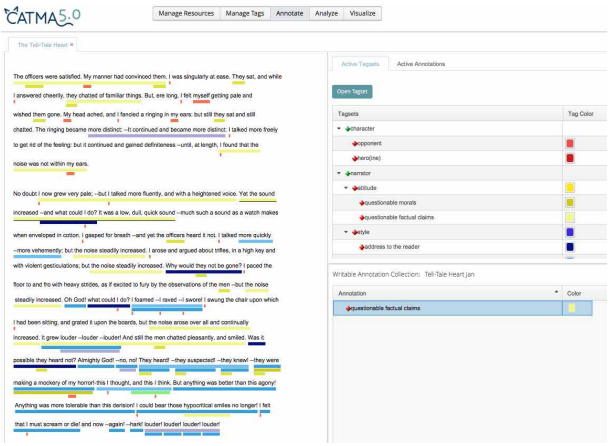


Abb. 2: Intuitives, nicht-deterministisches Annotieren in CATMA

Die im Rahmen des Lehrmoduls verfolgte didaktische Strategie hat mehrere Vorteile: Der erste Schritt, d.h. das digitale Anbringen freier Kommentare, stellt einen vollkommen explorativen und potenziell unstrukturierten Zugang zu literarischen Texten dar. Er erzwingt also kein formalistisches Umdenken und bildet die traditionellere geisteswissenschaftlich-hermeneutische Arbeitsweise gut ab. Im Vergleich zum analogen Arbeiten birgt er aber dennoch den Vorteil, dass die freien Kommentare durch digitale Unterstützung effektiver *nachgenutzt* werden können.

Als zusätzliches Angebot an Literaturwissenschaftler, die für eine etwas stärkere Formalisierung ihres Zugangs offen sind, zeigt das Lehrmodul, welche weiteren Vorteile und Optionen *taxonomiebasiertes* Annotieren mit sich bringt (bspw. bessere Vergleichbarkeit, detailliertere und vereinfachte Analyse, ggf. Reproduzierbarkeit etc.)[12] und wie dieses digital umgesetzt werden kann. Diese Herangehensweise kann in der Folge auch die Nützlichkeit noch stärker formalistisch anmutender DH-Techniken plausibilisieren – wie beispielsweise der kollaborativen, guidelinegestützten Annotation[13] oder der automatisierten Annotation bzw. Informationsextraktion.

Im Lehrmodul zum manuellen digitalen Annotieren sind also die Paradigmen umgesetzt, die auch die weitere Arbeit am forTEXT-Projekt bestimmen sollen: Durch stärkere Orientierung an der traditionell-geisteswissenschaftlichen Arbeitsweise und erleichterten Zugang können digitale Methoden einer breiteren Nutzergemeinschaft nähergebracht werden.

Fußnoten

1. Wir werden aufgrund besserer Lesbarkeit im Folgenden nur noch die männliche Form verwenden – unsere Ausführungen beziehen sich aber selbstverständlich dennoch auf alle Geschlechter.
2. Auf diese persistente Skepsis verweisen beispielsweise auch Fiedler und Weiß in ihrem Tagungsbericht zur DHd-Konferenz 2015 in Graz (vgl. Fiedler/Weiß 2015).
3. Beispiele hierfür sind u.a. die folgenden Projekte aus dem Bereich der digitalen Literaturwissenschaft: "heureCLÉA", ein Projekt zur automatischen Annotation von Zeitphänomenen in narrativen Texten (http://www.heurecle-a.de , vgl. Bögel et al. 2015); "Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse", ein Projekt zur automatisierten Annotation von Figurenrede (http://www.redewiedergabe.de , vgl. auch Brunner 2015); die "Computational Stylistics Group" (https://sites.google.com/site/computationalstylistics , vgl. Rybicki/Eder/Hoover 2016); das "Rhythmicalizer"-Projekt (http://www.rhythmicalizer.net , vgl. Baumann/Meyer-Sickendiek 2016), das ein Tool zur Erkennung von Prosodie in freien Versen entwickelt; ebenso wie Projekte zur Analyse von Figurennetzwerken wie etwa "Digital Literary Network Analysis" (https://dlina.github.io/about , vgl. Fischer et al. 2017).
4. So betonen beispielsweise Kindt und Köppe, dass literaturwissenschaftliche Interpretationen weniger auf strukturelle Aspekte oder bloßes Sprachverstehen gerichtet sind als vielmehr auf mannigfaltige, komplexe und oft nicht eindeutig umrissene Verstehensziele (vgl. Kindt/Köppe 2008: 12–14).
5. So sind beispielsweise bestehende Verzeichnisse für Tools zur Textanalyse aufgrund der vornehmlich technisch ausgerichteten Beschreibungen für Geisteswissenschaftler ohne DH-Vorwissen äußerst schwer zugänglich (z.B. TAPoR, http://www.tapor.ca). Einen vielversprechenden Ansatz zur nutzbaren Aufbereitung und Vermittlung digitaler Methoden – allerdings vornehmlich in schulischen Kontexten – stellt das Projekt "Digitalität in den Fachdidaktiken" dar (http://dhd-blog.org/?p=6812).
6. So basieren zahlreiche DH-Praktiken zur Textanalyse auf der automatischen Verarbeitung von Textoberflächendaten wie Wortfrequenzen, beispielsweise stilometrische Untersuchungen oder Topic Modeling (vgl. Brett 2012).
7. Der Fragebogen bietet Nutzern zudem die Option, auf fehlende Antwort- oder sogar Fragemöglichkeiten hinzuweisen. Auf diese Weise kann der Fragebogen – und mit ihm forTEXTs Empfehlungssystem – im Laufe des Projekts weiter optimiert werden.
8. Affiliierte Projekte sind beispielsweise "3DH" (http://www.threedh.net) und "SANTA: Shared Task on the Analysis of Narrative levels Through Annotation" (https://sharedtasksinthedh.github.io/), in deren Rahmen narrative Ebenen in Texten annotiert werden.
9. Ein Vorbild hierfür ist der bereits in CATMA integrierte "Query Builder", mithilfe dessen Nutzer komplexe Abfragen über Text- und Annotationsdaten laufen lassen können, ohne eine Abfragesprache lernen zu müssen.
10. Vgl. bspw. Bauer/Zirker 2015: Absatz 1.
11. So hat sich CATMA beispielsweise dem Konzept des *hermeneutischen Markups* verschrieben – darunter verstehen wir nach Piez Markup das bewusst interpretativ und flexibel ist (vgl. Piez 2010). Auf diese Weise erlaubt CATMA beispielsweise mehrfache und sogar "widersprüchliche" Annotationen derselben Textstelle – dennoch sind die Annotationen im standardisierten TEI-Format exportierbar und somit flexibel nachnutzbar. Diese Markup-Eigenschaften werden im Backend ermöglicht durch Standoff-Markup und die Nutzung von TEI Feature Structures (http://www.tei-c.org/release/doc/tei-p5-doc/de/html/FS.html).
12. Vgl. zu den Vorteilen von Klassifikationssystemen wie Taxonomien und Typologien auch Bailey 1994.
13. Zur Methode und zum Nutzen kollaborativen Annotierens, siehe auch Gius/Jacke 2015 und Gius/Jacke 2017.

Bibliographie

Bailey, Kenneth D. (1994): *Typologies and Taxonomies. An Introduction to Classification Techniques*. Thousand Oaks/London/New Delhi: Sage Publications.

Bauer, Matthias / Zirker, Angelika (2015): "Whipping Boys Explained. Literary Annotation and Digital Humanities", in: *MLA Commons. Literary Studies in the Digital Age* https://dlsanthology.mla.hcommons.org/whipping-boys-explained-literary-annotation-and-digital-humanities/ [letzter Zugriff 06. September 2017].

Baumann, Timo / Meyer-Sickendiek, Burkhard (2016): "Large-scale Analysis of Spoken Free-verse Poetry", in: *Proceedings of LT4DH-Workshop 2016* http://edoc.sub.uni-hamburg.de/informatik/volltexte/2016/228/pdf/

baumann_large_scale_analysis.pdf [letzter Zugriff 06. September 2017].

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (2015): "Collaborative Text Annotation Meets Machine Learning. heureCLÉA, a Digital Heuristic of Narrative", in: *DHCommons Journal* 1 http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic [letzter Zugriff 06. September 2017].

Brett, Megan R. (2012): "Topic Modeling. A Basic Introduction", in: *Journal of Digital Humanities* 2(1) http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ [letzter Zugriff 06. September 2017].

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe*. Berlin / Boston: de Gruyter (= Narratologia Bd. 47).

Fiedler, Maik / Weiß, Andreas (2015): "Von Daten zu Erkenntnissen. Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation. DHd-Jahrestagung 2015" http://www.hsozkult.de/conferencereport/id/tagungsberichte-6059 [letzter Zugriff 06. September 2017].

Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Kittel, Christopher / Trilcke, Peer (2017): "Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts", in: *Digital Humanities 2017. Conference Abstracts* https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf [letzter Zugriff 06. September 2017].

Gius, Evelyn / Jacke, Janina (2015): "Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse", in: *Zeitschrift für digitale Geisteswissenschaften*, Sonderband 1 http://www.zfdg.de/sb001_006 [letzter Zugriff 06. September 2017].

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.

Kindt, Tom / Köppe, Tilmann (2008): "Einleitung", in: dies. (eds.): *Moderne Interpretationstheorien. Ein Reader*. Göttingen: Vandenhoeck & Ruprecht 7–26.

Piez, Wendell (2010): "Towards Hermeneutic Markup. An Architectural Outline", in: *Digital Humanities 2010. Conference Abstracts* http://piez.org/wendell/papers/dh2010/ [letzter Zugriff 06. September 2017].

Rybicki, Jan / Eder, Maciej / Hoover, David (2016): "Computational stylistics and text analysis", in: Crompton, Constance / Lane, Richard J. / Siemens, Ray (eds.): *Doing Digital Humanities*. London / New York: Routledge 123–144.

Spörl, Uwe (2004): *Basislexikon Literaturwissenschaft*. 2., durchges. Aufl. Paderborn [u.a.]: Schöningh.

Websites

3DH (http://www.threedh.net)

CATMA (http://www.catma.de)

Computational Stylistics Group (https://sites.google.com/site/computationalstylistics)

Digital Literary Network Analysis (https://dlina.github.io/about)

Digitalität in den Fachdidaktiken. Projektpräsentation im DHd-Blog (http://dhd-blog.org/?p=6812)

forTEXT (http://www.fortext.net)

heureCLÉA (http://www.heureclea.de)

Redewiedergabe (http://www.redewiedergabe.de)

Rhythmicalizer (http://www.rhythmicalizer.net)

SANTA: Shared Task on the Analysis of Narrative levels Through Annotation (https://sharedtasksinthedh.github.io/)

TAPoR (http://www.tapor.ca)

TEI Feature Structures (http://www.tei-c.org/release/doc/tei-p5-doc/de/html/FS.html)

Digitized Inhumanities: Qualitative Inhaltsanalyse von Hexenprozessakten mit MAXQDA

Müller, Andreas

andreas.w.mueller@outlook.com
Universität Wien, Österreich

Forschungskontext:

In den Sozialwissenschaften ist die qualitative Inhaltsanalyse unter Anwendung moderner Analysesoftware wie MAXQDA etabliert. In den Geschichtswissenschaften gewinnen diese Zugänge unter dem Schlagwort der „Digital Humanities" erst langsam an Verbreitung. Meine Masterarbeit „Die Magie der Inhaltsanalyse: Entwurf einer Inhaltsanalyse für den Vergleich von Hexenprozessakten aus Rostock 1584 und Hainburg 1617/18" versucht daher diese Zugänge der Sozialwissenschaften in einem methodisch eher „traditionel-

len“ Forschungsbereich, der historischen Hexenprozessforschung, anzuwenden.

Ausgangspunkt:

Das Ausgangsmaterial bilden 37 „Geständnisse“ (Urgichten) aus Hexenprozessen in Hainburg 1617/18 und Rostock 1584. Die inhaltlich nach Befragungspunkten strukturierten Dokumente stehen am Ende des juristischen Prozesses vor der Hinrichtung der Angeklagten. Diese bilden eine Synthese aus den Ansichten des Gerichts, den Angaben von Zeugen und den unter Folter entstandenen Aussagen der Angeklagten.

Forschungsfrage(n):

F1: Wie unterscheiden sich die aus den Urgichten hervorgehenden Hexereiimaginationen in Hainburg 1617/18 und Rostock 1584?
F2: Wie spiegeln sich regionale Unterschiede (sozio-ökonomisch, konfessionell, politisch) in den Dokumenten wieder? Welche Elemente der Hexereivorstellung erweisen sich als starr, welche als flexibel?

Methodik:

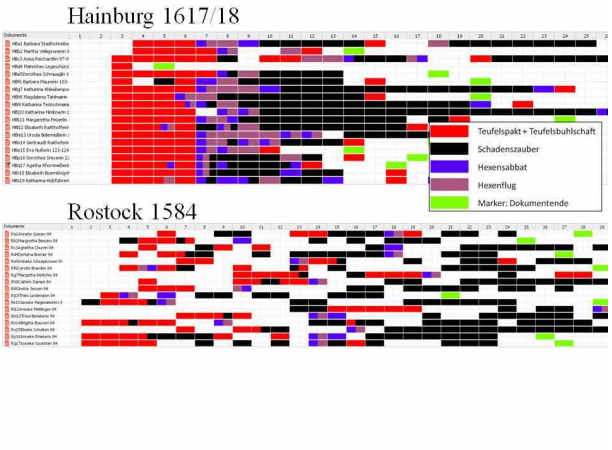
Methodische Grundlage bildet das Methodenangebot der qualitativen Inhaltsanalyse“ (Mayring 2015; Kuckartz 2014), die auf die spezifisch geschichtswissenschaftlichen Herausforderungen angepasst wurde. Als Analysetool wurde MAXQDA12 verwendet. Ein auf dem „elaborierten Hexereibegriff“ (Dillinger 2007) basierendes Kategoriensystem wurde deduktiv auf die Texte angewendet. Für die Analyse wurde ein Mixed-Methods Ansatz verfolgt, der die quantitative Auswertung des Kategoriensystems als „Kontrastmittel“ für die qualitative Analyse heranzieht.

Ergebnisse

Vor allem in der Imagination des Schadenszaubers traten deutliche Unterschiede hervor, welche die sozio-ökonomischen Kontexte widerspiegeln. Im Weinbaugebiet Hainburg findet sich vor allem der Vorwurf des Wetterzaubers gegen Weinreben, Obst und Getreide. In der Seehandelsstadt Rostock fehlen diese Delikte weitgehend und es steht vor allem die Verbreitung von Krankheiten durch Bettlerinnen im Zentrum. Die Vorstellungen vom Teufelspakt, Hexentanz und Flug sind deutlich homogener als der Schadenszauber, wenn sie auch verschiedene Schwerpunktsetzungen und Ausgestaltungen aufweisen. Darüber hinaus hat die quantitative Analyse überraschende Unterschiede in der Inhaltsstruktur zu Tage gefördert, die über den Rahmen der Arbeit offene Fragen aufwirft. Es entsteht dabei der Befund einer wesentlich stärker „integrierten“ bzw. „kohärenten“ Hexereiimagination in Hainburg, die sich in zahlreichen Kategorienüberschneidungen manifestiert, während in Rostock teils völlig isolierte Elemente und Narrative zu Tage treten.

Mehrwert und Problematik der digitalen Methodik

Durch die Anwendung der Software MAXQDA eröffnen sich neue analytische Möglichkeiten von großem Mehrwert. Durch die Zuordnung von verschiedenen Textabschnitten in analytische Kategorien wird es nicht nur möglich die Kategorien qualitativ auszuwerten, sondern auch Visualisierungen und Quantifizierungen zu erzeugen. Als ein Beispiel wird hier das Dokumentenvergleichsdiagramm herausgegriffen:



Die Zeilen bilden hier die einzelnen Dokumente ab, die Spalten die einzelnen Absätze des Textes. Hierdurch wird die thematische Struktur der Texte sichtbar, ähnlich einem „Topic Modelling“ wobei hier jedoch die Zuordnung manuell erfolgt. Ein Vorteil dieses Zugangs ist seine starke Nähe zum Forschungsgegenstand. Ein Klick auf eine der farbigen Flächen (z.B. eine thematische „Lücke“) führt in der Software sofort zurück in die entsprechende Textstelle. MAXQDA ist damit nicht nur eine Möglichkeit zur Ergebnispräsentation sondern auch ein analytisches Tool.

Das textnahe Arbeiten mit analytischen Kategorien ist vor allem für die qualitative Forschung ein

großer Mehrwert. Die Kategorien sowie ihre Zuordnung zum Text werden dabei vom Forschenden selbst festgelegt und ein Rückbezug in den Originaltext ist zu jedem Zeitpunkt möglich. Zu den problematischen Aspekten in der Arbeit mit einem Kategoriensystem zählt jedoch, dass Textteile, die außerhalb der eigenen analytischen Kategorien liegen, leicht aus dem Fokus geraten und der Blick für die Grundgesamtheit des Texts verloren gehen kann.

Diskussion:

Die Masterarbeit setzte sich als methodisches Ziel, den Vergleich zweier historischer Textkorpora unter Einsatz der qualitativen Inhaltsanalyse vorzunehmen. Angestrebtes Ziel einer nachfolgenden Dissertation ist es eine „vergleichende Inhaltsanalyse“ für den Einsatz insbesondere in den Geschichtswissenschaften zu entwerfen. Dafür scheint es angebracht diesen ersten Schritt zur kritischen Diskussion zu stellen und die Potenziale der qualitativen Inhaltsanalyse, aber auch die Möglichkeiten anderer Analyseverfahren, für den systematischen Textvergleich zu diskutieren.

Anliegen der Posterpräsentation:

Das Poster selbst wird die Ergebnisse des erfolgten Vergleichs in Form von Grafiken und Diagrammen (Codematrix Browser, Code Relation Browser, Dokumentenvergleichsdiagramm aus MAXQDA) in das Zentrum rücken. Hierfür werden vor allem die über das quantifizierende „Kontrastmittel“ deutlich gewordenen strukturellen Unterschiede der Vergleichsgruppen aufgezeigt und erläutert. Hiermit erfüllt das Poster zwei Zwecke: 1. Ermöglicht es die Diskussion über und Kritik an dem gewählten Analyseverfahren. 2. Können die Ergebnisse als Anregung dienen die Anwendbarkeit ähnliche Verfahren für das jeweils eigene Forschungsfeld zu reflektieren.

Bibliographie

Behringer, Wolfgang (1995): Weather, Hunger, and Fear. The Origins of the European Witch Prosecutions in Climate, Society, and Mentality. In: German History (13), S. 1–27.

Behringer, Wolfgang (1997): Hexenverfolgung in Bayern. Volksmagie, Glaubenseifer und Staatsräson in der Frühen Neuzeit. 3. Aufl. München: R. Oldenbourg.

Behringer, Wolfgang (2004a): Geschichte der Hexenforschung. In: Sönke Lorenz (Hg.): Wider alle Hexerei und Teufelswerk. Die Europäische Hexenverfolgung und ihre Auswirkungen auf Südwestdeutschland. Ostfildern: Thorbecke, S. 485–668.

Behringer, Wolfgang (2004b): Witches and witch-hunts. A global history. Cambridge, UK, Malden, MA: Polity Press.

Behringer, Wolfgang (2010a): A cultural history of climate. Cambridge: Polity Press.

Behringer, Wolfgang (Hg.) (2010b): Hexen und Hexenprozesse in Deutschland. 7. Aufl. München: Dt. Taschenbuch-Verl.

Berelson, Bernard (1952): Content analysis in communication research. Glencoe: Free Press.

Clark, Stuart (1999): Thinking with demons. The idea of witchcraft in early modern Europe. Oxford: Oxford Univ. Press.

Dillinger, Johannes (1999): "Böse Leute". Hexenverfolgungen in Schwäbisch-Österreich und Kurtrier im Vergleich. Trier: Spee.

Dillinger, Johannes (2007): Hexen und Magie. Eine historische Einführung. Frankfurt/Main: Campus-Verl.

Dillinger, Johannes (2013): Kinder im Hexenprozess. Magie und Kindheit in der Frühen Neuzeit. Stuttgart: Steiner.

Dorn-Haag, Verena (2015): Hexerei und Magie im Strafrecht. Historische und dogmatische Aspekte. Tübingen: Mohr Siebeck.

Flossmann, Ursula; Putschögl, Gerhard (1987): Hexenprozesse. Seminar zur Geschichte der Strafrechtspflege. Linz: Universitätsverlag R. Trauner.

Goodare, Julian (2016): The European witch-hunt. London, New York: Routledge Taylor & Francis Group.

Ignatieff, Nathalie (2009): Hexenprozesse in Hainburg 1617/18. Diplomarbeit. Universität Wien, Wien.

Kuckartz, Udo (2014): Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 2. Auflage. Weinheim, Basel: Beltz Juventa.

Landsteiner, Erich (1999): The Crisis of Wine Production in Late Sixteenth-Century Central Europe. Climatic Causes and Economic Consequences. In: Climatic Change (43), S. 323–334.

Landsteiner, Erich (2001): Trübselige Zeit? Auf der Suche nach den wirtschaftlichen und sozialen Dimensionen des Klimawandels im späten 16. Jahrhundert. In: Österreichische Zeitschrift für Geschichtswissenschaften 12 (2), S. 79–116.

Landsteiner, Erich; Weigl, Andreas (2001): „Sonsten finden wir die Sachen sehr übel aufm Landt beschaffen“. Krieg und lokale Gesellschaft in Niederösterreich (1618-1621). In: Benigna von Krusenstjern (Hg.): Zwischen Alltag und Katastro-

phe. Der Dreißigjährige Krieg aus der Nähe. 2. Aufl. Göttingen: Vandenhoeck & Ruprecht, S. 229–270.

Lang, Ines (2008): "Das zeichen hab er ihr mitt der prezen ins rechte wang vor 16 jahrn geben […]". zwei Hexenprozesse im Hainburg des Jahres 1624. Diplomarbeit. Universität Wien, Wien.

Levack, Brian (1995): Hexenjagd. Die Geschichte der Hexenverfolgungen in Europa. München: Beck.

Lorenz, Sönke (1982): Aktenversendung und Hexenprozeß. Dargestellt am Beispiel der Juristenfakultäten Rostock und Greifswald (1570/82 - 1630). Frankfurt am Main: Lang.

Lorenz, Sönke (2004): Der Hexenprozess. In: Sönke Lorenz (Hg.): Wider alle Hexerei und Teufelswerk. Die Europäische Hexenverfolgung und ihre Auswirkungen auf Südwestdeutschland. Ostfildern: Thorbecke, S. 131–154.

Mayring, Philipp (2015): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 12. Auflage. Weinheim, Basel: Beltz Juventa.

Midelfort, Hans Christian Erik (1968): Recent Witch Hunting Research, or Where Do We Go from Here? In: The Papers of the Bibliographical Society of America 62 (3), S. 373–420.

Midelfort, Hans Christian Erik (1972): Witch hunting in southwestern Germany, 1562-1684. The social and intellectual foundations. Stanford, Calif.: Stanford University Press.

Midelfort, Hans Christian Erik (1995): Alte Fragen und neue Methoden in der Geschichte des Hexenwahns. In: Sönke Lorenz (Hg.): Hexenverfolgung. Beiträge zur Forschung unter besonderer Berücksichtigung des südwestdeutschen Raumes. Würzburg: Königshausen und Neumann.

Moeller, Katrin (2007): Dass Willkür über Recht ginge. Hexenverfolgung in Mecklenburg im 16. und 17. Jahrhundert. Bielefeld: Verl. für Regionalgeschichte.

Münch, Ernst (2003a): Rostocks Aufstieg zur Stadtkommune. Von den Anfängen bis 1265. In: Karsten Schröder (Hg.): In deinen Mauern herrsche Eintracht und allgemeines Wohlergehen. Eine Geschichte der Stadt Rostock von ihren Ursprüngen bis zum Jahre 1990. Rostock: Koch, S. 12–28.

Münch, Ernst (2003b): Zwischen Reformation und Dreißigjährigem Krieg. 1532 bis 1648. In: Karsten Schröder (Hg.): In deinen Mauern herrsche Eintracht und allgemeines Wohlergehen. Eine Geschichte der Stadt Rostock von ihren Ursprüngen bis zum Jahre 1990. Rostock: Koch, S. 53–92.

Neugebauer-Wölk, Monika (2003): Wege aus dem Dschungel. Betrachtungen zur Hexenforschung. In: Geschichte und Gesellschaft (29), S. 316–347.

Raser, Dorothea (1987): Zauberei und Hexenprozesse in Niederösterreich. Diplomarbeit. Universität Wien, Wien.

Schild, Wolfgang (1994): Hexenglaube, Hexenbegriff und Hexenphantasie. In: Sönke Lorenz (Hg.): Hexen und Hexenverfolgung im deutschen Südwesten. Volkskundliche Veröffentlichungen des Badische Landesmuseums Karlsruhe. Ostfildern: Hatje Cantz Verl., S. 20–31.

Schormann, Gerhard (1981): Hexenprozesse in Deutschland. Göttingen: Vandenhoeck & Ruprecht.

Schulze, Winfried (1993): Untertanenrevolten, Hexenverfolgung und "kleine Eiszeit". Eine Krisenzeit um 1600? In: Bernd Roeck (Hg.): Venedig und Oberdeutschland in der Renaissance. Beziehungen zwischen Kunst und Wirtschaft. Sigmaringen: Thorbecke, S. 290–309.

Utz Tremp, Kathrin (2008): Von der Häresie zur Hexerei. "wirkliche" und imaginäre Sekten im Spätmittelalter. Hannover: Hahnsche Buchhandlung.

Venjakob, Judith (2017): Der Hexenflug in der frühneuzeitlichen Druckgrafik. Entstehung, Rezeption und Symbolik eines Bildtypus. Petersberg: Michael Imhof Verlag.

Voltmer, Rita (2006): Vom getrübten Blick auf die frühneuzeitlichen Hexenverfolgungen. Versuch einer Klärung. In: Gnostika. Zeitschrift für Wissenschaft und Esoterik 11, S. 45–58.

Voltmer, Rita (2015): Stimmen der Frauen? Gerichtsakten und Gender Studies am Beispiel der Hexenforschung. In: Johanna Blume, Jennifer Moos und Anne Conrad (Hg.): Frauen, Männer, Queer. Ansätze und Perspektiven aus der historischen Genderforschung. St. Ingbert: Röhrig Universitätsverlag, S. 19–46.

Wieden, Helge bei der (1981): Rostock zwischen Abhängigkeit und Reichsunmittelbarkeit. In: Roderich Schmidt (Hg.): Pommern und Mecklenburg. Beiträge zur mittelalterlichen Städtegeschichte. Köln: Böhlau, S. 111–132.

DISCO: Diachronic Spanish Sonnet Corpus

Ruiz Fabo, Pablo
pablo.ruiz@linhd.uned.es
UNED, Spanien

Martínez Cantón, Clara
cimartinez@flog.uned.es
UNED, Spanien

Calvo Tello, José
jose.calvo@uni-wuerzburg.de
Universität Würzburg

Introduction

This poster presents a corpus of 19th-century sonnets in Spanish in XML-TEI (685 authors, 2677 sonnets). It includes well-known authors, but also less canonized authors. Texts and authors are enriched with identifiers and metadata. See *https://github.com/pruizf/disco*

A fundamental difficulty for Digital Humanities studies on Spanish literature is a scarcity of digital resources (Agenjo, 2015).

Some resources do however exist. BiDTEA (Gago Jover et al, 2015), ADMYTE (Marcos Marin and Faulhaber, 1992), ReMetCa (González-Blanco and Rodríguez, 2014) and PoeMetCa (Escribano et al, 2016) have digitized Spanish Medieval texts. Navarro-Colorado et al. (2015) presented the Corpus of Spanish Golden-Age Sonnets.

Regarding 19th-century Spanish literature, available collections covering different genres are Textbox (Schöch et al., 2015), BETTE (Santa María Fernández, 2017), Aracne (Álvarez-Mellado and Martín-Fuertes, 2015), and Revistas Culturales 2.0 (Ehrlicher and Rißler-Pipka, 2015). Nevertheless, none of these projects are working on poetry.

DISCO complements this growing ecosystem by adding a meaningful representation of 19th-century sonnets, with more periods under validation, to be published shortly.

Corpus description

What is DISCO

Our corpus collects 2677 sonnets in Spanish from the 19th century, by 685 authors (Spain or Latin America). It intends to provide a wide sample, inspired by distant reading approaches (Moretti, 2005). The raw texts were extracted from Biblioteca Virtual Miguel de Cervantes (1999), which contains an anthology (Garcia, 2006) of 19th century sonnets covering both well-known and non-canonical authors. We used an anthology in order to have external scholarly criteria for the literary relevance of the corpus texts.

The texts have been encoded in XML-TEI P5, given this standard's benefits in terms of reuse, storage and retrieval. Author metadata have been extracted or inferred from unstructured content in the sources, and placed in the TEIheader (year, place of birth and death, and gender). Two versions of the texts are available: one collecting every sonnet per author, the other encoding a single sonnet per file. For corpus preparation, we closely followed the TEI guidelines and RIDE's criteria for Digital Text Collections (Henny and Neuber, 2017).

Additionally, authors have been assigned VIAF identifiers. This gives the corpus an entry-point to the linked open data cloud, enhancing its findability. The corpus is available on GitHub and saved in Zenodo, adopting good practices for data use, reuse, and conservation.

The metadata we added allow users to create subcorpora, such as "female authors from Cuba born in the first half of the 19th century".

We have also obtained sonnets from other centuries, since the 15th century to the present. These are under validation and will be published shortly within the DISCO corpus, which intends to give a wide perspective on the sonnet in Spanish diachronically.

Why sonnets

The sonnet has had great importance in European poetry; the relevance of the corpus for literary scholarship is guaranteed. It is a "manageable" form to treat computationally, obeying clear restrictions. Variability stays within bounds, making meaningful comparison across poems easier, regarding scansion or rhyme types. Besides, some digital collections of sonnets already exist (with different features to ours, as discussed below) as well as automatic analyses of this form. The sonnet has received attention from the computational linguistics community (Navarro-Colorado et al, 2015, 2016, 2017; Agirrezabal, 2017) including the ADSO project (Navarro-Colorado 2017). The DISCO corpus will also be useful for that audience. For these reasons, a new sonnet corpus allows us to engage in a dialogue with earlier work in traditional literary studies, in digital corpus development, and in computational poetry analyses.

Concerning digitally available sonnet corpora, Sonnet-Archiv (Elf Edition) is organized as a forum, and its coverage is less wide than ours. The "Sonnet Library" (Biblioteca Virtual Miguel de Cervantes, 2007) is organized alphabetically, rather than using meaningful criteria for literary scholarship, like periods. Both are traditional websites. Finally, the Corpus of Spanish Golden Age Sonnets (Navarro-Colorado et al., 2015) covers major authors from the 15th to the 17th century, with an automatic metrical annotation. Author metadata in these corpora are very limited and unavailable in a machine-readable format (see Calvo Tello, 2017, for discussion of related issues). With the DISCO corpus, we are offering a

wider period and author range, from major to minor authors, encoded in XML-TEI, available as repository, with richer structured and standard metadata.

Conclusion

With the DISCO corpus, while focusing on sonnets, we intend to increase available digital resources in Spanish poetry, by addressing additional periods, covering minor as well as canonical authors, and including materials from several Spanish-speaking countries. Choosing the sonnet complements existing work on this form, in traditional and computational literary studies. TEI was adopted in order to serve the large community using this format. The corpus can be made available as linked open data as it includes VIAF IDs. It is published at https://github.com/pruizf/disco and https://doi.org/10.5281/zenodo.1012567 .

Acknowledgements

The work was supported by Starting Grant ERC-2015-STG-679528 POSTDATA , PI Elena González-Blanco. We also thank Helena Bermúdez from the LINHD-UNED lab for later contributions to the corpus.

Bibliography

Agenjo, Xavier (2015): "Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos", in Ínsula: revista de letras y ciencias humanas 822: 12–15.

Agirrezabal, Manex (2017): Automatic Scansion of Poetry. PhD Thesis. University of the Basque Country.

Álvarez Mellado, Elena / Martín-Fuertes, Leticia (2015): Aracne Project [online]. Available at: http://www.fundeu.es/aracne/ [Accessed 22 Sep. 2017].

Biblioteca Virtual Miguel de Cervantes (1999): Biblioteca Virtual Miguel de Cervantes [online]. Available at: http://www.cervantesvirtual.com/ [Accessed 22 Sep. 2017].

Biblioteca Virtual Miguel de Cervantes (2007): Biblioteca del Soneto [Sonnet Library] [online]. Available at: http://www.cervantesvirtual.com/bib/portal/bibliotecasoneto/ [Accessed 22 Sep. 2017].

Calvo Tello, José. (2017). Review of Corpus of Spanish Golden Age Sonnets by Borja Navarro Colorado, María Ribes Lafoz and Noelia Sánchez (ed.), in RIDE, 6. Institut für Dokumentologie und Editorik, Köln. [Online]. Available at: http://ride.i-d-e.de/issue-6/corpus-of-spanish-golden-age-sonnets/ [Accessed 22 Sep. 2017].

Ehrlicher, Hanno / Rißler-Pipka, Nanette (2015). Revistas Culturales 2.0. Augsburg: Universität Augsburg. [Online]. Available at: https://www.revistas-culturales.de/es [Accessed 22 Sep. 2017].

Elf Edition: Sonett-Archiv [online]. Available at: http://sonett-archiv.com [Accessed 22 Sep. 2017].

Escribano, Juanjo / González-Blanco, Elena / Río Riande, Gimena del (2016). PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana. [Online]. Available at: http://poemteca.linhd.es [Accessed 22 Sep. 2017].

Gago Jover, Francisco (2015): "La biblioteca digital de textos del español antiguo (BiDTEA), in Scriptum Digital 4: 5–36.

García González, Ramón (2006): Sonetos del siglo XIX. Biblioteca Virtual Miguel de Cervantes, Alicante. [Online]. Available at: http://www.cervantesvirtual.com/obravisor/sonetos-del-siglo-xix--0/html/ [Accessed 26 Nov. 2017 **]**

González-Blanco, Elena / Rodríguez, José Luis (2014): "ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse", in Journal of the Text Encoding Initiative 8 [online]. Available at: https://jtei.revues.org/1274 [Accessed 22 Sep. 2017], doi:10.4000/jtei.1274.

Henny, Ulrike / Neuber, Frederike (2017): "Criteria for Reviewing Digital Text Collections, version 1.0". IDE, Institut for Dokumentologie und Editorik, [online]. Available at: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/ [Accessed 22 Sep. 2017].

Marcos Marín, Francisco / Faulhaber, Charles B. (coord.) (1992): ADMYTE. Archivo Digital de Manuscritos y Textos Españoles, in http://www.admyte.com/admyteonline/contenido.htm [Accessed 22 Sep. 2017].

Moretti, Franco (2005): Graphs, Maps, Trees: Abstract Models for a Literary History. London and New York: Verso.

Navarro-Colorado, Borja (2015): A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In ACL Workshop on Computational Linguistics for Literature 105.

Navarro-Colorado, Borja (2017): ADSO project – Análisis distante del soneto castellano de los Siglos de Oro [Distant analysis of the Spanish Golden Age sonnet] [online]. Available at: http://adso.gplsi.es/index.php/es/proyecto-adso [Accessed 22 Sep. 2017].

Navarro-Colorado, Borja / Ribes Lafoz, María / Sánchez, Noelia (2015): Corpus of Spanish Gol-

den-Age Sonnets. Alicante: University of Alicante [online]. Available at: https://github.com/bncolorado/CorpusSonetosSigloDeOro [Accessed 22 Sep. 2017].

Navarro-Colorado, Borja / Ribes Lafoz, María, / Sánchez, Noelia (2016): "Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation", in Proceedings of the Language Resources and Evaluation Conference [online]. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf [Accessed 22 Sep. 2017]

Navarro-Colorado, Borja (2017): "A metrical scansion system for fixed-metre Spanish poetry", in Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqx009 [Accessed 22 Sep. 2017]

Santa María Fernández, María Teresa / Jiménez Fernández, Concepción María (2017): Biblioteca Electrónica Textual Del Teatro Español, 1868-1936. Universidad Internacional de la Rioja, Spain.

Schöch, Christof / Henny, Ulrike / Calvo Tello, José / Popp, Stefanie (2015): The CLiGS Textbox. Würzburg: University of Würzburg. [Online]. Available at: https://github.com/cligs/textbox [Accessed 22 Sep. 2017]

Dramenquartett – Eine didaktische Intervention

Fischer, Frank
ffischer@hse.ru
Higher School of Economics, Moskau

Kittel, Christopher
contact@christopherkittel.eu
Karl-Franzens-Universität Graz; Open Knowledge Forum Österreich

Milling, Carsten
cmil@hashtable.de
Berlin

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam, Deutschland

Wolf, Jana
jana_a_wolf@hotmail.com
Mittelbayerische Zeitung Regensburg

Ziel dieses Posters ist es, anhand von 32 deutschsprachigen Dramen in die Netzwerkanalyse literarischer Texte einzuführen, eine didaktische Intervention für eine zwar mittlerweile etablierte Methode der literaturwissenschaftlichen Analyse, die aber nicht immer genügend reflektiert wird: Der Errechnung teils komplexer netzwerktheoretischer Maße entspricht nicht immer ein entsprechender Sprung zur Bedeutungsebene. Was bedeutet es zum Beispiel wirklich, dass die durchschnittliche Pfadlänge in Goethes »Faust. Der Tragödie erster Theil« genau 1,79 beträgt? Wenn man jedoch diesen Wert in Beziehung zu entsprechenden Werten anderer Stücke setzt, gewinnt er an komparatistischer Bedeutung. Die Anschaulichkeit der Wert und ihre spielerisch erfahrene Dimensionierung ermöglichen so die Einübung in die strukturalistische Betrachtung von Netzwerken am Beispiel von Dramen, wobei damit zugleich kulturelles Grundwissen über die Strukturation von Netzwerken – immerhin ubiquitäre Gegenstände der sozialen und technischen Welt – erworben werden kann.

Um den komparatischen Blick im Kontext der literaturwissenschaftlichen Netzwerkanalyse zu schulen, setzen wir mit unserem Poster auf einen Gamification-Ansatz. Anders als bei unserem ersten Experiment in dieser Richtung – der auf der DHd2016 präsentierten Android-App »Play(s)« (vgl. Göbel/Meiners 2016), in deren Mittelpunkt die spielerische Korrektur und Anreicherung unserer Korpusdaten stand –, handelt es sich diesmal um eine nicht-technische Anwendung, die auf spielerische Weise netzwerkanalytisches Datenmaterial explorierbar macht.

Dabei wird das Posterformat in zweierlei Hinsicht bespielt: Das Poster ist einerseits eine Datenvisualisierung auf Grundlage eines selbst gepflegten größeren Dramenkorpus. Andererseits ist es ein in 32 Teile zerlegbares Dramenquartett, das spielerisch mit den Bedeutungshorizonten verschiedener netzwerktheoretischer Größen bekannt macht und ein Bewusstsein für komparatistische Möglichkeiten trainiert. Dieser Ansatz ist in den Geisteswissenschaften nicht neu, verwiesen sei etwa auf das architekturgeschichtliche Quartettspiel »Plattenbauten. Berliner Betonerzeugnisse« (Mangold u. a. 2001), in dem technische Daten verschiedener Plattenbautypen gegenübergestellt werden (vgl. auch Richter 2006).

Die Didaxe des Dramenquartetts bezieht sich auf mehrere Dimensionen: eine literaturgeschichtliche, eine quantitative, eine netzwerktheoretische. Die 32 Stücke bilden einen Minimalkanon, der von der Zeit der Gottschedischen Theaterreformen bis in die Moderne reicht. Statt der lexikonartigen Beschreibung eines solchen Kanons (wie etwa im »Dramenlexikon des 18. Jahrhunderts«,

Hollmer/Meier 2001), besteht das Beschreibungsinstrument hier in visuellen und quantitativen Werten, die Vergleichbarkeit herstellen – erst dieser Umstand vereint die verschiedenen Karten zu einem kompetitiven Spiel.

Als visueller Catch der Quartettkarten dient eine Visualisierung des jeweiligen extrahierten sozialen Netzwerks (vgl. Fischer u. a. 2016). Die weiteren Informationen auf den Karten setzen sich aus (Kanonwissen präsentierenden) Metadaten (Autor*in – Titel – Untertitel – Genre – Jahr) und vor allem aus statischen und dynamischen Netzwerkdaten zusammen (Anzahl von Subgraphen – Netzwerkgröße – Netzwerkdichte – Clustering-Koeffizient – Durchschnittliche Pfadlänge – Höchster Degreewert und Name der entsprechenden Figur –), wie sie im Rahmen des dlina-Projekts berechnet wurden. [1] Das Deckblatt enthält eine Einführung zum Projekt und seinen Hintergründen sowie Kurzdefinitionen der auf den einzelnen Karten enthaltenen netzwerktheoretischen Maßzahlen, die damit nicht nur spielerisch erkundet, sondern auch konzeptuell verstanden werden können.

Das Poster wird mit unserer Python-Skriptsammlung ›dramavis‹ generiert, die in der neuen Version 0.4 eine entsprechende Funktion erhalten hat (Kittel/Fischer 2017). Für das Konferenzposter haben wir einen Fallback-Kanon zusammengestellt (Stücke von Johann Christoph und Luise Adelgunde Victorie Gottsched, von Gellert, J. E. Schlegel, Caroline Neuber, Klopstock, Lessing, Gerstenberg, Goethe, Lenz, Klinger, Schiller, Kotzebue, Kleist, Zacharias Werner, Müllner, Grillparzer, Grabbe, Büchner, Hebbel, Gustav Freytag, Anzengruber, Arno Holz, Wedekind, Schnitzler, Erich Mühsam). Über eine individualisierbare Kanon-Datei können aber auch eigene Quartette zusammengestellt werden, sodass sich etwa auch epochenspezifische Sets (Dramen der Aufklärung, Dramen der Klassik, Romantische vs. Klassische Dramen, Dramen des Sturm und Drang vs. Dramen des Naturalismus) oder gattungsspezifische Sets erstellen lassen.

Auf der Konferenz werden wir neben einem Poster, das das didaktisch-interventionistische Konzept veranschaulicht, auch diverse Quartett-Sets präsentieren.

Fußnoten

1. Vgl. das Blog https://dlina.github.io/ und das Github-Repo https://github.com/dlina .

Bibliographie

Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Kittel, Christopher / Trilcke, Peer (2016): “Distant-Reading Showcase. 200 Years of Literary Network Data at a Glance”, DHd2016, Leipzig. DOI: < https://dx.doi.org/10.6084/m9.figshare.3101203.v1 >.

Göbel, Matthias / Meiners, Hanna-Lena: Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus”. DHd2016, Leipzig.

Hollmer, Holmer / Meier, Albert (eds.) (2011): Dramenlexikon des 18. Jahrhunderts. München: C.H. Beck.

Kittel Christopher / Fischer, Frank: “dramavis v0.4” (September 2017). Repo: < https://github.com/lehkost/dramavis >.

Mangold, Cornelius u.a. (2001): *Plattenbauten. Berliner Betonerzeugnisse. Ein Quartettspiel*. Berlin

Richter, Peter (2006): *Der Plattenbau als Krisengebiet. Die architektonische und politische Transformation industriell errichteter Wohngebäude aus der DDR am Beispiel der Stadt Leinefelde*. Hamburg: Univ., Diss.

Ein Brief – zwei Perspektiven. Stellenkommentare in digitalen Briefeditionen über APIs austauschen

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

Vernetzung von Texten und anderen Ressourcen war und ist ein großes Versprechen des digitalen Zeitalters im Allgemeinen, aber auch der digitalen Editionsphilologie im Besonderen. Während man um 2000 darunter noch vorwiegend das manuelle Setzen von einzelnen Links verstand, wurde spätestens seit Beginn dieses Jahrzehnts die automatisierte Verknüpfung in den Blick genommen. So wurde in Portalen, Websites und auch digitalen Editionen (vornehmlich des deutschsprachigen Raums) die automatisierte Verlinkung von Personenregistereinträgen eingeführt, die mit Hilfe der Gemeinsamen Normdatei und von BEACON-Schnittstellen ermöglicht wird.

Diese Vernetzung auf Basis von BEACON-Schnittstellen gehört heute zum Standard digitaler Editionen und ist nicht mehr wegzudenken (Stadler 2012).

Mit den Fortschritten im Internet allgemein wuchs aber seit einigen Jahren der Wunsch, digitale Editionen über solche basalen Verlinkungen hinaus zu verknüpfen. Im Bereich der Briefeditionen z.B. wurde schon recht früh der Wunsch geäußert, Briefe editionsübergreifend durchsuchbar zu machen und zu vernetzen. Basis dafür sollten die in TEI-XML kodierten Briefmetadaten, also die wichtigsten Kopfdaten eines Briefes, sein. Mit Entwicklung des Correspondence Metadata Interchange Format (TEI Correspondence SIG 2015) und des da-rauf aufbauenden Webservices correspSearch (http://correspSearch.net) konnte dies realisiert werden. Nicht nur können Briefeditionen jetzt digitale Briefverzeichnisse bereitstellen, die durch correspSearch zentral recherchierbar gemacht werden. Digitalen Briefeditionen ist es hier nun möglich, über die API von correspSearch Briefmetadaten zu beziehen und passend zu einem – in der eigenen Edition vorliegenden – Brief zu präsentieren (Dumont 2018). Damit werden schon einige methodische Probleme der Briefedition überwunden, andere bleiben hingegen noch offen.

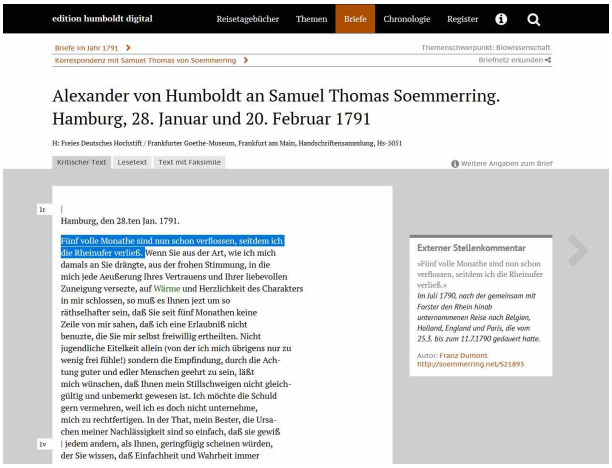
So kann es vorkommen, dass ein Brief mehrfach ediert vorliegt. Diese Situation kann dadurch entstanden sein, dass ältere Editionen (etwa des 19. Jahrhunderts) durch zeitgemäße abgelöst werden. Sie kann aber auch dadurch aufgekommen sein, dass der Brief nicht in einer Briefwechselausgabe, sondern in zwei Gesamtausgaben erschienen ist: Einerseits in derjenigen Gesamtausgabe, die die Korrespondenz des Autors ediert, andererseits in derjenigen, die die Korrespondenz des Empfängers ediert. In so einem Fall entstehen zum einen Kommentierungen, die in beiden Editionen ähnlich vorgenommen werden (z.B. Identifizierung von Personen etc.). Zum anderen werden aber auch Kommentierungen vorgenommen, die aus der spezifischen Perspektive der Edition entstehen – d.h. mit Fokus auf diejenige Person, deren Korrespondenz ediert wird. Dadurch entstehen zwei Annotationslayer, die sich im Idealfall ergänzen. Die Einzelstellen(-kommentierung) aus der jeweils anderen Edition kann also u.U. äußerst wertvoll für einen Nutzer sein.

In digitalen Briefeditionen besteht nun prinzipiell die Möglichkeit, Daten aus anderen digitalen Editionen bzw. Portalen zu integrieren. Anstelle eines einmaligen, manuellen Importes bzw. Übernahme dieser Daten, ist heutzutage deren automatisierter Abruf über ein Application Programming Interface (API) zu favorisieren. Daten aus externen Quellen können so schneller aggregiert, aktualisiert und zusammen mit den eigenen Forschungsdaten – hier: edierte Briefe – angeboten werden. Daher erscheint die Nutzung von APIs und standardisierten Datenformaten angeraten. Im Regelfall liegen Einzelstellenkommentare in digitalen Briefeditionen heutzutage als Bestandteil der TEI-XML-Dokumente vor (etwa im Element <note>). Sollen diese Kommentare allerdings über eine Schnittstelle zur freien Nachnutzung durch externe Editionen angeboten werden, müssen sie deutlich mehr Informationen tragen, als im Kontext der Edition – nämlich Metadaten (Autor des Kommentars, Bezugstext etc.). Auch eine Übermittlung des ganzen Briefes mitsamt seinen Einzelstellenerläuterungen ist nicht gewünscht, liegt der Text doch gerade im hier geschilderten Fall bereits vor.

In Betracht kommt daher ein Datenformat, dass dediziert für den Austausch und die Aggregation von Kommentaren, also vorwiegend textuell vorliegender Annotation, konzipiert ist. Zwar wäre es möglich ein solches Format in TEI-XML neu zu definieren, allerdings gibt es für diesen Bereich bereits einen Standard, der auch schon breite Anwendung findet: das Web Annotation Data Model (Sanderson, Ciccarese, und Young 2017). Es ist für genau dieses Nutzungsszenario – Austausch von Kommentaren – konzipiert und durch das W3C standardisiert. Es wird daher auch von DARIAH-DE empfohlen (Lordick u. a. 2016).

Das Poster stellt nun anhand eines Beispiels diesen Ansatz exemplarisch vor und demonstriert seinen Nutzen. Als Beispiel ist der überlieferte Briefwechsel zwischen Alexander von Humboldt und Samuel Thomas Soemmerring gewählt, der zum einen in *edition humboldt digital* (http://edition-humboldt.de) vorliegt, zum anderen in einer (derzeit noch unveröffentlichten) digitalen Edition zu Soemmerrings Biographica. Konkret sieht die Implementierung so aus: Einzelstellenerläuterungen aus der Soemmerring-Edition, die ursprünglich als TEI-XML vorliegen, werden über eine API im Web Annotation Data Model angeboten. Diese API wird von correspSearch abgefragt und die Daten aggregiert. Andere digitale Editionen, hier beispielhaft die *edition humboldt digital*, können anhand einer Kalliope-URI, also der eindeutigen und maschinenlesbaren Archivkennung, den Webservice correspSearch auf Annotationen zu einem bestimmten Brief hin abfragen. Im Erfolgsfall werden die Annotationen ausgeliefert und in der *edition humboldt digital* am Brief angezeigt (mit Quelle, Autor etc.).

Das Poster stellt nicht nur Konzept und Implementierung vor, sondern diskutiert auch die noch offenen Probleme einer solchen API, wie z.B. die korrekte und stabile Referenzierung des Bezugstextes (d.h. der kommentierten Textstelle).



Beispielhafte Implementierung in der Entwicklungsversion von edition humboldt digital

Bibliographie

Dumont, Stefan. 2018. „CorrespSearch – Connecting Scholarly Editions of Letters“. Journal of the Text Encoding Initiative 10. [Im Erscheinen].

Lordick, Harald, Rainer Becker, Michael Bender, Luise Borek, Canan Hastik, Thomas Kollatz, Beata Mache, Andrea Rapp, Ruth Reiche, und Niels-Oliver Walkowski. 2016. „Digitale Annotationen in der geisteswissenschaftlichen Praxis“. Bibliothek Forschung und Praxis 40 (2): 186–199. doi:10.1515/bfp-2016-0042.

Sanderson, Robert, Paolo Ciccarese, und Benjamin Young. 2017. „Web Annotation Data Model. W3C Recommendation“. W3C. https://www.w3.org/TR/annotation-model/.

Stadler, Peter. 2012. „Normdateien in der Edition“. editio 26: 174–83.

TEI Correspondence SIG. 2015. „Correspondence Metadata Interchange Format (CMIF)“. https://github.com/TEI-Correspondence-SIG/CMIF.

Eine Fallstudie zur Annotation von Vagheit in Werken Dimitrie Cantemirs

Vertan, Cristina
cristina.vertan@uni-hamburg.de
Universität Hamburg, Deutschland

von Hahn, Walther
vhahn@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Das Korpus

Das ausgewählte Korpus besteht aus zwei Hauptwerken Dimitrie Cantemirs, eines Universalgelehrten des 17. Jahrhunderts und Mitglied der „Kurfürstlich –Brandenburgischen Societät der Wissenschaften“. Die zwei Werke wurden ursprünglich auf Lateinisch verfasst, die Originale sind aber verloren aber Kopien davon wurden im späten 20. Jh. wiederentdeckt. Im Umlauf waren lange Zeit nur Übersetzungen ins Englische, Deutsche (Cantemir 1771), (Cantemir 1745), und Französische, die mindestens bis Mitte des 19. Jh. Referenzwerke für die Geschichte des osmanischen Reichs und der historischen Provinz Moldawien waren. Durch seinen langen Aufenthalt in Istanbul, hatte Cantemir Zugang zu vielen Quellen die er zitiert. Daneben zitiert er auch Sagen und Legenden und versucht immer durch geschickte sprachliche Redewendungen zu vermitteln, was seiner Meinung nach historisch gesichert ist. Daher ist das Korpus besonders illustrativ für das Problem der Vagheitsannotation.

Erste Schritte zur Vagheitsannotation

Das Projekt HerCore versucht durch gezielte Annotation von Vagheit drei geisteswissenschaftliche Fragestellungen in Bezug auf die Cantemir-Forschung zu lösen:

- Der erstmalige Vergleich aller historischen Übersetzungen, da seit geraumer Zeit die Vermutung formuliert wurde, dass diese relativ stark von den Originalen abweichen.
- Die Untersuchung der Zuverlässigkeit von Äußerungen Cantemirs. Hierbei werden vor allem Quellen von turkologischen Fachwissenschaftlern einbezogen.
- Die Konsistenz von Cantemir über dieselben Personen und Ereignisse in den zwei Werken.

Die Annotation von Vagheit wird auf drei Ebenen untersucht:

1. Linguistisch,
2. In Metadaten und Editorik,
3. im Fachwissen.

- Für die Linguistische Ebene wurde als Startpunkt die Klassifizierung von (Pinkal 1981)

benutzt. Für die Laufzeit des Projekts haben wir aus dem o.g. Schema wegen besonderer Angemessenheit für das zu analysierende Korpus folgende mögliche Vagheitsindikatoren ausgewählt:

Auf lexikalischer Ebene: Non-Intersectives, Adjektive, Hecken, inexakte Maße, Modalverben (Attitudes), Komplexe Quantoren, Zitiereinleitungen, zeitliche Ausdrücke.

Auf syntaktische Ebene: Subjunktiv-Konstruktionen

Zusätzlich werden Named Entities untersucht: Personen, Zeitangaben, Orte etc. und mit einem entsprechenden Vagheitsgrad versehen ("Konstantinopel" ist nur wahr zwischen 337 und 1930).

Als Vorbereitung wird das Korpus zuerst einer linguistischen Ananlyse unterzogen, um Lemmas und Wortarten, sowie die Textstruktur (Sätze, Paragraphen) zu markieren. Diese wird dann die Basis für die semi-automatische Annotation von Vagheitsausdrücken (Vertan et al 2017).

Die Annotation von vager Information wird dann in einem ersten Schritt manuell von Fachwissenschaftlern in einem Korpus-Ausschnitt vorgenommen. In einem zweiten Schritt wird versucht diese Annotation automatisch im Korpus zu propagieren. Ein dritter Schritt soll die Ergebnisse von Inferenzen zwischen vagen Ausdrücken erzeugen, um sich nicht dem Vorwurf auszusetzen man schreibe mit einem spezifischen Erkenntnis leitendem Interesse zunächst Annotationen in den Text um sie dann nur wörtlich wieder auszulesen.

Zusammenfassung

Der Beitrag wird die gesamte Systemarchitektur, sowie die einzelnen Schritte zur Annotation von Linguistischer Vagheit illustrieren.

Um dem Wissenschaftler am Ende eine hermeneutische Interpretation zu erlauben, muss ihm zu jedem annotierten Objekt ein Vagheits-Profil sowie Metadaten über Autoren, Genres und Inferenzergebnisse gezeigt werden können.

Hierzu eine Erweiterung von TEI und ein entsprechendes GUI zu entwickeln, sind wichtige Ziele des Projekts.

Außerdem muss die multilinguale Struktur der Cantemir-Texte mit Zitaten aus dem Griechischen, Lateinische und teilweise dem Türkischen annotiert und weiter erforscht werden.

Das Projekt soll zeigen, dass die Einbeziehung von Vagheit und Unschärfe in die Annotierung, in die Inferenz-Komponente und die hermeneutische Interpretationen durch den Wissenschaftler, einen erheblichen Gewinn an Funktionen und Glaubwürdigkeit für die DH bringt.

Bibliographie

Cantemir, Dimitrie, (1771) Beschreibung der Moldau, Faksimiledruck der Originalausgabe von 1771, Frankfurt und Leipzig

Cantemir, Dimitrie, (1745) Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen, 1745, Herold, Hamburg

Pinkal, Manfred,(1981) Semantische Vagheit: Phänomene und Theorien, Teil I/II. In: Linguistische Berichte Nr. 7/72, Wiesbaden 1980/1981.

Vertan, Cristina / von Hahn, Walther / Dinu, Anca (2017) On the annotation of vague expressions: a case study on Romanian historical texts, Proceedings of the first Workshop on Language Technology for Digital Humanities in Central and (South-) Eastern Europe, in association with RANL 2017, Varna

ELEXIS – Eine europäische Forschungsinfrastruktur für lexikographische Daten

Wissik, Tanja
Tanja.Wissik@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Krek, Simon
simon.krek@guest.arnes.si
Institut Josef Sefan, Slowenien

Jakubicek, Milos
milos.jakubicek@sketchengine.co.uk
Lexical Computing CZ s.r.o.,Tschechien

Tiberius, Carole
carole.tiberius@inl.nl
Instituut voor Nederlandse Lexicologie, Niederlande

Navigli, Roberto
navigli@di.uniroma1.it
Università degli Studi di Roma La Spienza, Italien

McCrae, John
john@mccr.ae
National Universtiy of Ireland, Galway Irland

Tasovac, Toma
ttasovac@humanistika.org
Centar za digitalne humanističke nauke, Serbien

Varadi, Tamas
varadi@nytud.hu
Magyar Tudományos Akadémia,Ungarn

Koeva, Svetla
svetla@dcl.bas.bg
Institute for Bulgarian Language, Bulgarien

Costa, Rute
rute.costa@fcsh.unl.pt
Universidade Nova de Lisboa, Portugal

Kernerman, Ilan
ilan@kdictionaries.com
K Dictionaries Ltd., Israel

Monachini, Monica
monica.monachini@ilc.cnr.it
Consiglio Nazionale delle Ricerche, Italien

Trap-Jensen, Lars
ltj@dsl.dk
Det Danske Sprog- og Litteraturselskab, Dänmark

Pedersen, Bolette S.
vnb282@ku.dk
Kobenhavns Universitet, Dänemark

Hildenbrandt, Vera
hildenbr@uni-trier.de
Universität Trier, Deutschland

Kallas, Jelena
jelena.kallas@eki.ee
Eesti Keele Instituut, Estland

Porta-Zamorano, Jordi
porta@rae.es
Real Academia Española; Spanien

Die Bedeutung von Wörterbüchern, seien es einsprachige, zweisprache oder mehrsprachige Wörterbücher, ist in der heutigen Informationsgesellschaft nicht zu unterschätzen. Sie geben nicht nur Auskunft über Wortbedeutungen und dazugehörige Übersetzungen, sondern sind selbst Bestandteil der Kulturgüter eines Landes und sie stellen bedeutende Ressourcen für Linked Open Data und Semantic-Web-Technologien dar. Obwohl in fast jedem Land Wörterbücher erstellt wurden und werden, seien es traditionelle Wörterbücher in gedruckter Form oder lexikographische Ressourcen in digitaler Form, waren die Bestrebungen nach Kooperation auf europäischer Ebene eher limitiert. Dies führte dazu, dass bis jetzt die Synergien zwischen traditioneller Lexikographie und maschineller Sprachverarbeitung nicht optimal genutzt werden konnten. Dies soll durch das Infrastrukturprojekt ELEXIS geändert werden.

Im Rahmen des ELEXIS Projekts soll eine Infrastruktur für lexikographische Daten entwickelt werden, die auf mehreren Ebenen ansetzt und den Bereich der traditionellen Lexikogrpahie mit dem Bereich der maschinellen Sprachverarbeitung verknüpft: Zum einen soll Kooperation und Austausch zwischen unterschiedlichen Forschungsgruppen, aber auch zwischen Forschungsgruppen und Verlagshäusern gefördert werden. Zum anderen soll an gemeinsamen Standards gearbeitet werden, um den Austausch und die Wiederverwendbarkeit von lexikographischen Daten in den unterschiedlichsten Szenarien zu fördern. Die Infrastruktur soll den Zugang zu Methoden, Tools und Daten ermöglichen und den bis jetzt nicht so weit verbreiteten Open-Access-Gedanken fördern.

Um diese Ziele umzusetzen, hat sich ein Konsortium von 17 Partnern gebildet. Unter den Partnern befinden sich Institutionen mit Expertise in Lexikographie, in maschineller Sprachverarbeitung, in Semantic-Web-Technologien und in den digitalen Geisteswissenschaften sowie nationale Sprachinstitutionen, Verlagshäuser und Partner mit Expertise im Bereich der Standardisierung und Normung. Nachfolgend die Liste der Partner (mit den Bezeichnungen in der jeweiligen Landessprache wie im Proposal angegeben): Institut Josef Sefan (Slowenien), Lexical Computing CZ s.r.o. (Tschechien), Instituut voor Nederlandse Lexicologie (Niederlande), Università degli Studi di Roma La Spienza (Italien), National Universtiy of Ireland, Galway (Irland), Österreichische Akademie der Wissenschaften (Österreich), Centar za digitalne humanističke nauke (Serbien), Magyar Tudományos Akadémia (Ungarn), Institute for Bulgarian Language (Bulgarien), Universidade Nova de Lisboa (Portugal), K Dictionaries Ltd (Israel), Consiglio Nazionale delle Ricerche

(Italy), Det Danske Sprog- og Litteraturselskab (Dänmark), Kobenhavns Universitet (Dänemark),Universität Trier (Germany), Eesti Keele Instituut (Estland), Real Academia Española (Spanien).

Die ELEXIS Infrastruktur wird zum einen Tools und Services zur Erstellung, Verarbeitung und Retrodigitalisierung von lexikographischen Daten und zum anderen den Zugang zu bereits existierenden lexikographischen Daten anbieten. Damit die zukünftigen Nutzer das Potential der Infrastruktur vollends ausnutzen können, sind die Entwicklung von online Trainingsmaterialien sowie die Abhaltung von Traininsworkshops geplant. Weiters sollen GastforscherInnen-Programme den Austausch zwischen den Forschungsgruppen aktiv fördern und für Forschungsvorhaben den Zugang zu Daten ermöglichen, die aus unterschiedlichen Gründen nicht Open Access zur Verfügung gestellt werden können.

ELEXIS wird eng mit den bereits existierenden Forschungsinfrastrukturen CLARIN und DARIAH zusammenarbeiten und auf den bereits vorhandenen Infrastrukturen aufbauen und zugleich diese beiden Infrastrukturen näher zusammenbringen.

In diesem Poster werden die Grundzüge des neuen europäischen Infrastrukturprojektes beschrieben sowie die Methoden und Maßnahmen, mit denen die oben genannten Ziele erreicht werden sollen, präsentiert. Weiters wird speziell auf den Nutzen und die Vorteile der ELEXIS Infrastruktur für die DH Community eingegangen.

Bibliographie

Schreibman, Susan / Ray Siemens, John Unsworth. ed. (2004): A Companion to Digital Humanites. Oxford: Blackwell.

Oldman, Dominic /Doerr, Martin/ Gradmann, Stefan (2016): "Zen and the Art of Linked Data: New Strategies for a Semantc Web of Humanist Knowledge", in: Susan Schreibman/Ray Siemens/John Unsworth e. (2016): A New Companion to Digital Humanites, 2nd Editon. 2012. Oxford: Wiley-Blackwell

Navigli Roberto / Ponzeto Simone P. (2012): "BabelNet: The Automatc Constructon, Evaluaton and Applicaton of a Wide-Coverage Multlingual Semantc Network", in: Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.

McCrae, John / Aguado-de-Cea, Guadalupe / Buitelaar, Paul / Cimiano, Philipp / Declerck, Thierry / Gómez-Pérez, Asunción / Gracia, Jorge / Hollink, Laura /Montel-Ponsoda, Elena / Spohr, Dennis / Wunner, Tobias (2012): "Interchanging lexical resources on the Semantc Web", in: Language Resources and Evaluaton, 46(6), pp 701-709, (2012).

Entitäten im Fokus am Beispiel von Captivity Narratives

Kessler, Linda
st150918@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Braun, Tamara
st151509@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Preuß, Tanja
st102459@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Eigennamenerkennung (NER) ist im Bereich der maschinellen Sprachverarbeitung bereits viel behandelt worden. Eine Übersicht hierzu findet sich bei Nadeau und Sekine (2007). In den Digital Humanities dient die Erkennung von benannten Entitäten der Identifikation zentraler Akteure und Elemente in Texten, welche unter anderem die Grundlage für tiefergehende Analysen bezüglich Beziehungen, Strukturen und Emotionen in diesen Texten bilden. Jannidis et al. (2015) thematisieren allerdings, dass die reine NER beispielsweise für eine Analyse von Figurennetzwerken in literarischen Texten unzureichend ist, da dabei nur Figurenreferenzen durch konkrete Namensnennung erfasst werden. Um spezifisch auf die Bedürfnisse von Textanalysen im Kontext der Digital Humanities einzugehen, wurden im „Center for Reflected Text Analytics“ (CRETA) (Kuhn et al. 2016) der Universität Stuttgart Annotationsrichtlinien entworfen, die über die Annotation reiner Eigennamen hinausgehen und sich auf verschiedenartige Entitätsreferenzen in deutschsprachigen Texten unterschiedlicher Genres fokussieren.[1] So wird beispielsweise das Appellativ *the indians* als Entität erfasst, obwohl die Referenz nicht mit Namen spezifiziert wird.

Ein Beispiel für den Mehrwert der Annotation solcher Entitätsreferenzen findet sich bei Blessing et al. (2017). Um die Übertragbarkeit der Richtlinien nicht nur zwischen verschiedenen Textsorten, sondern auch sprachübergreifend zu evaluieren, stellen wir unser Projekt mit dem Ziel der Annotation von Erzähltexten in englischer Spra-

che vor. Ausgehend von der durch CRETA geschaffenen Grundlage teilt sich unser Projekt in drei Phasen auf: die manuelle Annotation und Überprüfung der Übertragbarkeit der CRETA-Richtlinien auf die gegebene Textsorte, die Automatisierung der Entitätserkennung und die Einbindung der Entitäten in eine literaturwissenschaftliche Analyse.

Als Textgrundlage dient eine Sammlung von englischsprachigen Captivity Narratives.[2] Diese Erzählungen aus dem 18. Jahrhundert handeln von Erfahrungen weißer Siedler in Nordamerika, die in indianische Gefangenschaft geraten. Zunächst wurden in sieben Texten im Gesamtumfang von 71.526 Wörtern 5.163 Entitäten identifiziert und mit den von CRETA erarbeiteten Kategorien (Personen, Orte, Organisationen, Ereignisse, Werke und abstrakte Konzepte) annotiert. Im Verlauf dieser Annotationsphase wurden die CRETA Richtlinien an die speziellen Gegebenheiten der Textsorte angepasst, die in Bezug auf die Erwähnung von Personen und Orten einige Besonderheiten aufweist. Auffällig ist beispielsweise, dass Personen in vielen Fällen in Gruppen, oftmals auch ohne spezifische Namen, erwähnt werden. Um diese Nennungen dennoch zu erfassen, wird die Begrenzung auf Eigen- und Gattungsnamen aufgehoben und um Formen wie *some, others* oder *a few* erweitert. Zudem werden Orte häufig anhand von Landschaftsmerkmalen und nur selten mit konkreten Ortsnamen benannt. Dementsprechend bilden solche Nennungen (z.B. *the river* oder *the mountain*) den Großteil der annotierten Ortsentitäten. Die Erfassung von vollständigen Nominalphrasen als Entitäten erweist sich stellenweise als problematisch, da die Captivity Narratives verschachtelte Nominalphrasen enthalten, sodass sehr umfangreiche Entitäten zu annotieren sind.

Der so entstandene Goldstandard dient als Trainingsdatensatz zur Entwicklung eines maschinellen Lernverfahrens. Ein Naive Bayes Classifier wurde mit Features trainiert, die sich u.a. auf die äußere Gestalt (z.B. Großschreibung), die Wortart und die Zugehörigkeit zu Wortlisten (Namen und amerikanische Orte) beziehen. Im Kreuzvalidierungsverfahren kann damit ein Micro-Fscore von 0,29 erzielt werden. Für die am häufigsten im Trainingsmaterial vorhandene Klasse PER wurde ein Precision-Wert von 0,45 erzielt. Dies bedeutet, dass fast die Hälfte der automatisch mit PER annotierten Entitäten wirklich Personen sind. Der Recall von 0,3 zeigt, wie unvollständig die Erkennung mit einem knappen Drittel aller relevanten Personen noch ist. Eine Auswertung der Ergebnisse zeigt, dass die Länge und Verschachtelung vieler Entitäten die automatische Klassifizierung erschwert. Da sich im manuellen Annotationsprozess der Kontext häufig als Entscheidungshilfe herausstellte, sollte dieser bei der automatischen NER zukünftig berücksichtigt werden. Darüber hinaus könnte die Erweiterung der verwendeten Features durch syntaktische Informationen und die Verwendung einer größeren Menge an Trainingsdaten zu Verbesserungen führen.

Um den Mehrwert der Entitätsreferenzen für eine inhaltliche Fragestellung bezüglich der Captivity Narratives zu veranschaulichen, zeigen wir die textstatistische Analyse von Emotionen im Umfeld bestimmter Entitäten bzw. Entitätsgruppen. Basierend auf den manuell annotierten Texten, lassen sich Personenentitäten mithilfe von Clusteranalysen gruppieren. Anhand von positiven und negativen Wortlisten lassen sich zwei Gruppen bilden, die sich grob als *Indianer* und *Andere* gegenüber stehen (siehe Abbildung 1 und 2). Eine auf denselben Wortlisten basierende Sentiment Analyse ergab einen deutlich negativeren Emotionswert für Personenentitäten, die der Gruppe der Indianer zuzuordnen sind, als für die Gruppe der anderen Personen.

Abschließend lässt sich festhalten, dass auf Grundlage unserer Annotationen eine Abgrenzung der im Text auftretenden Gruppen anhand von emotionsgeladenen Wörtern möglich ist, die der erwarteten negativen Haltung der Verfasser gegenüber den Eingeborenen Nordamerikas entspricht.

Die von CRETA entwickelten Annotationsrichtlinien sind grundsätzlich auf die von uns analysierten Texte anwendbar, trotz abweichender Sprache und spezifischer Erzählweise. Um die Breite der enthaltenen Entitätsreferenzen vollständig abbilden zu können, bedarf es allerdings einzelner Spezifizierungen der Annotationsrichtlinien für diese Textsorte.

Abbildung 1: Ergebnisse der Cluster-Analyse: Indianer

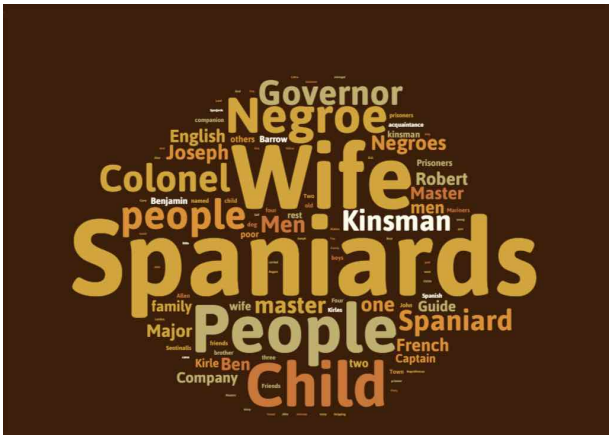


Abbildung 2: Ergebnisse der Cluster-Analyse: Andere

Fußnoten

1. Die Annotationsrichtlinien können eingesehen werden unter https://www.creta.uni-stuttgart.de/cute/datenmaterial/annotationsrichtlinien-1-1/ , abgerufen am 24.08.2017.
2. Wir danken Herrn Prof. Dr. Marc Priewe und den Mitarbeiter/innen der Professur für amerikanische Literatur und Kultur an der Universität Stuttgart

Bibliographie

Blessing, Andre / 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*, 57–67, Vancouver, Canada. Association for Computational Linguistics.

Jannidis, Fotis / Krug, Markus / Toepfer, Martin / Puppe, Frank / Reger, Isabelle / Weimer, Lukas (2015): "Automatische Erkennung von Figuren in deutschsprachigen Romanen". Abstract für die DHd 2015 in Graz.

Kuhn, Jonas / Alexiadou, Artemis / Braun, Manuel / Ertl, Thomas / Holtz, Sabine / Kantner, Catleen, . . . Zittel, Claus (2016): "CRETA (Centrum für reflektierte Textanalyse) – Fachübergreifende Methodenentwicklung in den Digital Humanities". Abstract für die DHd 2016 in Leipzig.

Nadeau, David / Sekine, Satoshi (2007): "A survey of named entity recognition and classification. *Linguisticae Investigationes", 30*, 3-26.

Entwicklungsstand im Projekt 'Digital Plato'

Kath, Roxana
roxana.kath@me.com
Universität Leipzig, Deutschland

Keilholz, Franz
franz.keilholz@tu-dresden.de
Technische Universität Dresden, Deutschland

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Rücker, Michaela
mruecker1@me.com
Universität Leipzig, Deutschland

Wöckener-Gade, Eva
woeckener-gade@uni-leipzig.de
Universität Leipzig, Deutschland

Yu, Xiaozhou
xiaozhou.yu@tu-dresden.de
Technische Universität Dresden, Deutschland

Einleitung

Das interdisziplinäre Forschungsprojekt *Digital Plato*[1] untersucht die Rezeption und Nachwirkung des platonischen Werkes in der griechischen Literatur bis in die Spätantike mit einem Fokus auf nicht-wörtlichen Referenzen. Der folgende Beitrag und das dazugehörige Poster geben einen Überblick über die wichtigsten Zwischenergebnisse, die während der zurückliegenden ersten Hälfte der Projektlaufzeit erzielt wurden. Neben der Organisation des Korpus sowie dessen linguistischer Anreicherung, dem Aufbau eines Wortnetzes und die Einbettung in einen Vektorraum, zählen dazu die Erfassung und Kategorisierung bekannter Referenzstellen, die Adaption der CTS zur wortgenauen Referenzierung, die theoretische Annäherung und schließlich die semi-automatische Suche nach neuen Paraphrasen.

Teilbereiche

Textgrundlage

Die Textgrundlage des Projekts ist der Thesaurus Linguae Graecae (TLG), der zunächst in ein XML-Format nach TEI-Standard überführtwurde. Da das Textkorpus als statisch anzusehen ist, wir im Laufe des Projektes aber mehrfach zusätzliche Annotationen hinterlegen und aktualisieren möchten, haben wir in einem zweiten Schritt Text und Annotationen nach dem Single-Source Prinzip voneinander getrennt. Dazu wurde der eigentliche Inhalt des Werks als unveränderliche Quelle (Single-Source) in Form einer einfachen fortlaufenden Textdatei angelegt, auf welche wiederum Dateien mit zusätzlichen Informationen (Standoff-Markup) referenzieren.Eine Referenz gibt dabei an, am wievielten Zeichen der Single-Source die Annotation startet und endet. Auch Texthervorhebungen und strukturelle Auszeichnungen des ursprünglichen TLG-Formats wurden so erfasst.

Für jede Form der Annotation wird eine eigene Datei mit Standoff-Markup angelegt, sodass eine Bearbeitung dieser keinen Einfluss auf die übrigen Auszeichnungen hat und sogar überlappende Annotationen ermöglicht werden.

Annotationen

Der TLG enthält hauptsächlich strukturelle Annotationen. Über eine Kombination verschiedener bestehender linguistischer Werkzeuge, wie Morpheus (Crane 1991) und Mate Tagger (Bohnet und Nivre 2012), werden dem Korpus nachträglich Lemmata und morphologische Informationen in Form von Standoff-Markup hinzugefügt. Ferner sollen auf dieser Basis auch die Nominalphrasen automatisch erkannt und ausgezeichnet werden. Solche Angaben helfen bei der Suche nach für das Projekt relevanten Textstellen.

Helleninet

Über die Verknüpfung diverser Wörterbücher wurde im Rahmen des Projekts ein Wortnetz für das Altgriechischegeneriert. Die Helleninet getaufte Struktur stellt ein weiteres wichtiges Standbein für die Einordnung der Relationen zwischen Wörtern und Textstellen dar.

Worteinbettung

Die Wörter eines Korpus können mit Hilfe statistischer Verfahren in einen Vektorraum eingebettet werden, sodass dieser semantische Beziehungen zwischen den Wörtern abbildet. Vorteilhaft hierbei ist, dass lediglich ein hinreichend großes Korpus benötigt wird, da das Verfahren auf den Kontexten der Wörter und nicht auf Vorwissen zur Sprache aufbaut. Das genutzte Verfahren word2vec (Mikolov et al. 2013) erlaubte uns dank seiner Performanz die Durchführung einer umfangreichen Evaluation, um eine für das Projekt möglichst optimale Einbettung zu finden.

Semi-automatische Rezeptionserkennung

Um weitere Referenzen auf Platon im Korpus aufzuspüren, verfolgen wir verschiedene (semi-)automatische Ansätze, die über die aus der Literatur bekannten Ansätze hinausgehen. Beim auf der DHd 2017 vorgestellten 'Rütteln' (Kath et al. 2017) handelt es sich um ein exploratives Verfahren, bei dem interaktiv von einer Textstelle Platons ausgehend einzelne Worte mit sinnverwandten Wörtern (bspw. Synonyme oder Übersetzungen) ersetzt werden und anschießend nach der modifizierten Textstelle im Korpus gesucht wird.

Ein ähnliches, aber systematisches Vorgehen stellt die n-Gramm-Suche dar. Nach einer umfangreichen Normalisierungwerden die n-Gramme verschiedener Längen für das gesamte Korpus indiziert. Anschließend können alle übereinstimmenden n-Gramme effizient ermittelt werden.

Ein drittes Verfahren basiert auf der Word Mover's Distance (Kusner et al. 2015), einem Distanzmaß für zwei Wortgruppen auf Grundlage einer Worteinbettung. Ausgehend von einer Textstelle wird das Korpus hierbei nach Textstellen mit möglichst geringer Distanz durchsucht(Pöckelmann et al. 2017).Die systematische Evaluation an Hand des im Projekt erstellten Goldstandard zeigt, dass dieses Verfahren zu sehr guten Ergebnissen führt.

Referenzierungssystem

Zur wortgenauen Referenzierung von Textstellen im Korpus wurden CTS-URNs adaptiert, d.h. *Uniform Resource Names* nach der Notation der *Canonical Text Services* (Blackwell und Smith 2014). Im Unterschied zum Standard werden die Wörter einer Zeile ebenfalls durchnummeriert, sodass in einer Subreferenz nicht das Wort selbst, sondern dessen Position genutzt werden kann. Um Probleme mit durch Zeilenumbruch getrenn-

ten Wörtern zu vermeiden, werden beide Teile in ihrer jeweiligen Zeile mitgezählt. Ein entsprechender Konverter bildet die CTS-URN auf Positionen in den Single-Sources sowie umgekehrt ab.

Goldstandard und Referenzannotierer

Mit Hilfe des im Projekt entwickelten, graphischen Werkzeugs - des Referenzannotierers - wurde eine zuvor erstellte Sammlung bekannter Rezeptionen mit Annotationen verschiedener Kategorien versehen. Der Goldstandard erlaubt durch diese umfassende Kategorisierung eine statistische Auswertung und hilft bei der Begriffsbildung. Zudem bildet er die Grundlage zur systematischen Evaluation der automatischen Suchverfahren und für einen umfassenden Thesaurus.

Begriffsbildung

Das Projekt arbeitet an einer theoretischen Ausdifferenzierung des Paraphrasenbegriffs: als konstitutiv werden hierbei Ähnlichkeiten von Textstellen zueinander angesehen, die sich auf der Wortebene abbilden. Hierbei nehmen wir wie anderen Ansätze die Notwendigkeit eines ‘Dritten’ an, um die Relation von Texten zueinander zu charakterisieren, nur verorten wir dies weniger stark im Bereich der Semantik, der schwer operationalisierbar ist. Vielmehr zielen wir auf eine fruchtbare Synthese dieser Theorie, die Paraphrasen ohne Annahme von Autorenintentionen oder der Bestimmung von Abhängigkeitsverhältnissen zwischen Texten beschreibbar macht, mit bestehenden Ansätzen zur Bestimmung von Ähnlichkeit zwischen Texten aus den DH.

Fußnoten

1. Gefördert durch die VolkswagenStiftung. Weitere Informationen auf der Projektseite unter: https://digital-plato.org/

Bibliographie

Blackwell, Christopher / Smith, Neel (2014): "The Canonical Text Service (CTS)" http://cite-architecture.github.io/cts/ [letzter Zugriff 18. September 2017].

Bohnet, Bernd / Nivre, Joakim (2012): "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing" in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: 1455-1465.

Crane, Gregory (1991): "Generating and parsing classical greek" in: *Literary and Linguistic Computing* 6(4):243–245.

Kath, Roxana / Keilholz, Franz / Klinker, Fabian / Pöckelmann, Marcus / Rücker, Michaela / Švitek, Mihael / Wöckener-Gade, Eva / Yu, Xiaozhou (2017): "Paraphrasenerkennung im Projekt Digital Plato" in: *Tagungsband der 4. Jahrestagung der Digital Humanities im deutschsprachigen Raum*: 266-270.

Kusner, Matt J. / Sun, Yu / Kolkin, Nicholas I. / Weinberger, Kilian Q. (2015): "From Word Embeddings To Document Distances" in: *Proceedings of the 32. International Conference on Machine Learning*: 957–966.

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg S. / Dean, Jeff (2013): "Distributed representations of words and phrases and their compositionality" in: *Advances in Neural Information Processing Systems* 26: 3111–3119.

Pöckelmann, Marcus / Ritter, Jörg / Wöckener-Gade, Eva / Schubert, Charlotte (2017): "Paraphrasensuche mittels word2vec und der Word Mover’s Distance im Altgriechischen"in: Digital Classics Online, Band 3, Ausgabe 3, S. 24-36.

erschließen - verknüpfen - finden: Forschungsdaten im Digitalen Wissensspeicher

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Grabsch, Sascha
grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Jürgens, Marco
juergens@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Maiwald, Anke
maiwald@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

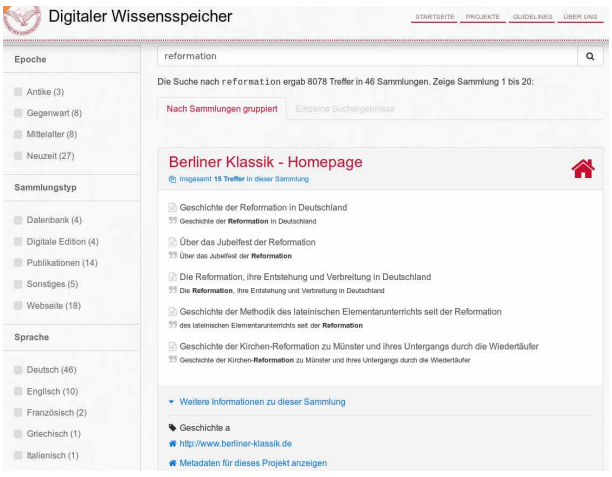
Willenborg, Josef
willenborg@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Die Auffindbarkeit und Sichtbarkeit digitaler Forschungsergebnisse in den Geisteswissenschaften leiden immer noch unter dem Mangel an einer zentralen Plattform für den wissenschaftlich fundierten Zugang, nicht nur zu den Metadaten, sondern auch zu den vollständigen digitalen, semantisch erschlossenen Forschungsdaten. Insbesondere die wichtigen Ergebnisse der geisteswissenschaftlichen Grundlagenforschung an den Akademien haben zu wenig Bekanntheit in den Fachcommunities und der breiteren Öffentlichkeit. Die Landschaft digitaler Forschungsergebnisse, wie Digitaler Editionen, Repositorien oder Forschungsdatenbanken, erscheint, trotz des vermehrten Einsatzes von Standards, technisch fragmentiert und wenig überschaubar. Dazu kommt eine mangelnde Vernetzung der verschiedenen Sammlungen, insbesondere institutionen- und fächerübergreifend, die zu dem aktuellen Zustand einzelner, isolierter digitaler Projekte geführt hat.

Im Rahmen des DFG-geförderten Projektes „Digitaler Wissensspeicher“ wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) seit 2012 ein zentraler Zugang für sämtliche digitale Forschungsdaten und Ressourcen der Akademie geschaffen. Hauptziel war dabei die vollständige Erfassung und Volltext-Indexierung der technisch sowie inhaltlich äußerst vielfältigen und heterogenen Ressourcen der BBAW. Gestützt auf einen Volltextindex (Apache Lucene) und ein anhand der Anforderungen der Akademie entwickeltes Metadenschema (basierend auf dem Metadatenstandard OAI-ORE) wurden über 230 Sammlungen aus 142 Projekten mit insgesamt mehr als 1 Mio. digitalen Ressourcen im Volltext und mit Metadaten erfasst. Durch den Einsatz von Sprachtechnologien (u. a. Donatus) ist eine morphologisch normalisierte Suche möglich. Über Text-Mining-Tools, wie DBpedia-Spotlight, werden die erfassten Ressourcen semantisch angereichert und vernetzt. Eine weitere Verknüpfung erfolgt über die manuell erfassten Metadaten zu den einzelnen Projekten und deren digitalen Sammlungen. Dies ermöglicht z.B. eine automatisierte Zuordnung semantisch ähnlicher Projekte.

Die durch das Textmining gewonnenen semantischen Annotationen ermöglichen eine Vielzahl weitergehender Nutzung. Beispielhaft wird dies auf der Website des Wissensspeichers anhand einer Kartenvisualisierung, die auch verschiedene Filtermöglichkeiten anbietet, demonstriert: die von DBpedia-Spotlight innerhalb der Ressourcen erkannten Orte werden projekt- und sammlungsübergreifend auf einer Karte referenziert. Sie bilden die Grundlage für einen visuellen, explorativen Zugang zu den indexierten Einzelressourcen.

Die Beschreibung der Metadaten erfolgt mittels eines flexiblen OAI-ORE-basierten Metadatenschemas, das die Abbildung aller, teilweise komplexen, Forschungsvorhaben der BBAW ermöglicht. Ähnlich den technisch niedrigschwelligen Anforderungen an die Volltext-Indexierung besteht auch für den Bereich der Metadaten die Möglichkeit, mit wenigen Pflichtfeldern weitere Sammlungen und Forschungsprojekte in den Wissensspeicher aufzunehmen. Die Metadaten zu allen Projekten werden über eine SPARQL-Schnittstelle für die weitere maschinelle Nachnutzung, z.B. die Integration in die Linked-Open-Data-Cloud bereitgestellt.



Der Digitale Wissensspeicher ist unter der Adresse http://wissensspeicher.bbaw.de für den öffentlichen Zugriff erreichbar. Damit bietet der Wissensspeicher eine exemplarische Lösung für einen Katalog heterogener, digitaler geisteswissenschaftlicher Ressourcen, die zentral aggregiert und über eine Volltextsuche verfügbar gemacht werden. Dass es sich dabei nicht nur um digitale Ressourcen der BBAW handeln muss, liegt auf der Hand. Damit auch Sammlungen anderer Projekte und Institutionen von der Entwicklung des Wissensspeichers profitieren können, wurden Guidelines mit strukturellen und inhaltlichen Mindestanforderungen, die Ressourcen und Metadaten für die Aufnahme erfüllen müssen bzw. Schnittstellenbeschreibungen für vom Wissensspeicher unterstützte Formate bereitgestellt. Damit werden niedrigschwellige Zielvorgaben für die (technische) Qualität der in den Wissensspeicher aufzunehmenden Ressourcen mit ihren Metadaten formuliert. Zu den unterstützten Formaten gehören u.a. HTML-Websites, XML-basierte Ressourcen, PDF-Dokumente und institutionelle Repositorien sowie jede Form von Daten, die in der vom Wissensspeicher definierten Form eines XML-Exports ausgeliefert werden können (z.B. relationale Datenbanken oder NoSQL-Datenbanken).

Das große Spektrum unterschiedlicher Erscheinungsformen digitaler Ressourcen in den Geisteswissenschaften kann somit vom Wissensspeicher verarbeitet und integriert werden. Durch die projektübergreifende Volltextsuche in allen verzeichneten Sammlungen sowie die miteinander vernetzten Metadaten verbessert der Digitale Wissensspeicher die Auffindbarkeit, Sichtbarkeit und damit die Nutzbarkeit von Digital-Humanities-Projekten und digitalen Forschungsergebnissen.

Chen, Ko-le, Marian Dörk und Martyn Dade-Robertson: „Exploring the Promises and Potentials of Visual Archive Interfaces.“ In IConference 2014 Proceedings, 735–41. iSchools, 2014. https://doi.org/10.9776/14348.

Glinka, Katrin, Christopher Pietsch und Marian Dörk: „Past Visions and Reconciling Views: Visualizing Time, Texture and Themes in Cultural Collections.“ digital humanities quarterly 11, Nr. 2 (27. Februar 2017). http://www.digitalhumanities.org/dhq/vol/11/2/000290/000290.html; abgerufen am 14.12.2017.

Horch, Andrea, Holger Kett und Anette Weisbecker: „Semantische Suchsysteme für das Internet. Architekturen und Komponenten semantischer Suchmaschinen.“ Stuttgart: Fraunhofer Verlag, 2013.

McMurry, Julie A., Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot u. a.: „Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data.“ PLOS Biology 15, Nr. 6 (29. Juni 2017): e2001414. https://doi.org/10.1371/journal.pbio.2001414.

Voß, Jakob: „Was sind eigentlich Daten?“ LIBREAS. Library Ideas, Nr. 23 (2013). http://libreas.eu/ausgabe23/02voss/; abgerufen am 14.12.2017.

Ward, David, Jim Hahn und Kirsten Feist: „Autocomplete as Research Tool: A Study on Providing Search Suggestions.“ Information Technology and Libraries 31, Nr. 4 (11. Dezember 2012). https://doi.org/10.6017/ital.v31i4.1930.

Woutersen-Windhouwer, Saskia (Hrsg.): „Enhanced Publications: Linking Publications and Research Data in Digital Repositories.“ Amsterdam: Amsterdam Univ. Press, 2009.

Formalisierung von Märchen

Declerck, Thierry
declerck@dfki.de
DFKI GmbH, Deutschland

Aman, Anastasija
aamann@coli.uni-saarland.de
Universität des Saarlandes, Deutschland

Grünewald, Stefan
stefang@coli.uni-saarland.de
Universität des Saarlandes, Deutschland

Lindemann, Matthias
malinux@t-online.de
Universität des Saarlandes, Deutschland

Schäfer, Lisa
lkschae@gmail.com
Universität des Saarlandes, Deutschland

Skachkova, Natalia
s9naskac@stud.uni-saarland.de
Universität des Saarlandes, Deutschland

Introduktion

Im Rahmen eines Softwareprojektes [1], das sich mit der automatisierten Analyse von Märchen in deutscher Sprache befasst, hat sich die Notwendigkeit ergeben, eine formale Repräsentation von Märchen zu bestimmen, damit die einzelnen Komponente des Systems miteinander integriert werden können.

Wir beschreiben in diesem Beitrag zum einen, welche Informationen in dieser formalen Repräsentation enthalten sind, und zum anderen, wie diese Informationen in XML bzw. Python konkret codiert werden.

Kodierte Information

Ein Märchen besteht im Sinne unseres Projektes aus den folgenden Bestandteilen:

- Eine Menge von Orten, an denen die Handlung spielt;
- Eine Menge von Charakteren, die an der Handlung beteiligt sind;
- Eine zeitliche Abfolge von Szenen, die jeweils an einem bestimmten Ort spielen und an denen jeweils eine Teilmenge der Märchencharaktere beteiligt ist;
- Jede Szene besteht ihrerseits aus einer zeitlichen Abfolge von Dialogakten zwischen den Märchencharakteren oder vom Erzähler zum Zuhörer. Zusammengenommen bilden diese Dialogakte den Märchentext.

Im Folgenden werden die verschiedenen Bestandteile, sowie ihre Eigenschaften und Beziehungen untereinander, näher beschrieben.

Orte, an denen das Märchen spielt, werden nur über ihren **Typ** (Attribut type) charakterisiert. Mögliche Ortstypen sind dabei z.B. Wald, Schloss oder Stall. Daneben existiert außerdem der Typ „Nirgendwo“ für Szenen ohne eindeutig bestimmbaren Ort (z. B. Abschnitte des Märchens, an denen nur der Erzähler beteiligt ist). Jeder Ort erhält eine spezifische ID der Form **loc1**, **loc2** etc.

Charaktere werden über eine Reihe von Eigenschaften beschrieben, welche zum einen inhärente demographische Eigenschaften (Name, Alter, Geschlecht, Typ), sowie zum anderen externe Eigenschaften (Einstellung, Propp-Archetyp – s. (Propp 1977)) beinhalten. Beim **Namen** (name) des Charakters handelt es sich um eine Zeichenkette, z.B. „Rapunzel.“ (Wird ein Charakter auf mehrere Arten gerufen, so wird die häufigste Bezeichnung gewählt.) **Alter** (age) des Charakters wird nicht in Zahlen, sondern in Stufen angegeben, da Märchen im Allgemeinen keine genauen Altersangaben enthalten; die möglichen Werte sind dabei „toddler“, „child“, „teenager“, „young adult“, „adult“ und „senior“. Das **Geschlecht** (gender) des Charakters wird den klassischen Vorstellungen folgend entweder mit „male“ oder „female“ angegeben. Zusätzlich gibt es den Wert „none“ für geschlechtlich unterspezifizierte Charaktere wie Tiere, Monster usw. Der **Typ** des Charakters unterscheidet z. B. zwischen „human“ oder „animal/monster“. Für „animal/monster“ unterscheiden wir zusätzlich nach **Subtypen**, z.B. für Tiere nach Größe, also „small“, „medium“ oder „big“, oder „witch“ und „demon“ für einen bestimmten Monstertyp. Eine binäre Feststellung der **Einstellung bzw. Gesinnung** des Charakters verortet diesen auf der Gut-/Böse-Achse: „evil“ oder „neutral“. Außerdem wird der **Propp-Archetyp** des Charakters angegeben: „hero“, „villain“ etc. (Propp, 1977). Jeder Charakter erhält eine spezifische ID von der Form **ch1**, **ch2** usw. Außerdem gehören zu jedem Märchen zwei „Dummy“-Charaktere für Erzähler und Zuhörer, welche stets die IDs **ch0** bzw. **ch-1** und die Typen „narrator“ bzw. „listener“ zugewiesen bekommen. Dies ist nötig, um auch Passagen darstellen zu können, welche vom Erzähler gesprochen werden, der selbst ja kein eigentlicher Charakter der Handlung ist. Dies ist notwendig, um ein automatisches „Vorlesen“ des Märchens zu implementieren.

Szenen werden im Hinblick auf Zeit, Ort, beteiligte Charaktere sowie Propp Funktionen (Propp, 1977) beschrieben. Der **Zeitpunkt** (time), zu dem die Szene spielt, wird anhand einer ID der Form **t1**, **t2** usw. angegeben, wobei die IDs den linearen Ablauf der Zeit darstellen. Der **Ort** (location), an dem die Szene spielt, wird als String in Großbuchstaben angegeben, ausgewählt aus einer Liste mit Möglichkeiten. Der **Übergang zur nächsten Szene** (transition) wird ebenfalls codiert, indem das Bewegungsverb, das den Übergang von einem Ort zum anderen beschreibt, oder die Phrase, die stattdessen den Szenenwechsel einleitet, angegeben wird. Die an der Handlung der Szene beteiligten **Charaktere** werden mit ihren IDs angegeben, also z. B. **ch2**, **ch3**, **ch5**. Dabei werden alle Charaktere berücksichtigt, die in der Szene zugegen sind, auch wenn diese bspw. nicht sprechen. Die Propp-Funktionen und -Subfunktionen der Szene werden mit ihrem Symbol (nach der englischen Ausgabe Propp (1977)) angegeben, also z. B. A4 – „theft of daylight“. Jede Szene erhält eine spezifische ID der Form **s1**, **s2** etc. Da die Märchenhandlung im Allgemeinen linear erzählt wird, ist der Index üblicherweise (aber nicht notwendigerweise) identisch mit demjenigen des Zeitpunkts der Szene, d. h. die Szene **s1** wird üblicherweise zum Zeitpunkt **t1** spielen usw. Jeder Szene sind **Dialogakte** un-

tergeordnet, denen der zu dieser Szene gehörige Text entspricht.

Dialogakte werden im Hinblick auf ihre Sprecher und Adressaten, ihren Inhalt sowie ihren Zeitpunkt beschrieben. Der **Zeitpunkt** (time), zu dem der Dialogakt geäußert wird, wird anhand einer ID angegeben, welche eine Spezifizierung der ID des Zeitpunkts der zugehörigen Szene darstellt. Spielt z. B. Szene **s5** zum Zeitpunkt **t5**, so haben die zugehörigen Dialogakte die Zeitpunkte **t5.1**, **t5.2** usw. Der **Sprecher** (speaker) des Dialogakts wird über seine ID angegeben. Der **Adressat** bzw. die **Adressaten** (receiver) des Dialogakts werden über eine Liste von Charakter-IDs angegeben, z.B. **ch2**, **ch4**, **ch6**. Passagen des Erzählers stellen dabei einen Spezialfall dar: Sie werden als Dialogakte des Erzählers mit dem Zuhörer bzw. Leser betrachtet, d. h. der „Dummy"-Charakter des Erzählers wird als Sprecher angegeben und der Dummy-Charakters des Zuhörers als Empfänger. Abgesehen davon werden sie behandelt wie Dialogakte zwischen Charakteren. Jeder Dialogakt erhält eine spezifische ID, die – unabhängig von der Szenestruktur – linear hochgezählt wird, also **d1**, **d2** usw.

XML-Repräsentation

Die oben beschriebenen Informationen lassen sich im XML-Format darstellen. Dabei wird eine XML-Baumstruktur genutzt, um die Hierarchie der verschiedenen Objekte zu repräsentieren. Das Wurzelelement des Dokuments hat stets den Bezeichner Tale und die Attribute „title" und „annotator", welche Titel und den Namen des Annotators des jeweiligen Märchens enthalten:

1: Struktur des Tale-Wurzelelements (Beispiel).

<Tale title="Froschkönig" annotator="Lisa Schäfer"> ... </Tale>

Diesem Element untergeordnet sind die Elemente Characters, Locations und Text. Das Characters-Element enthält Character-Subelemente, die jeweils die gesammelten Informationen für einen Charakter speichern:

2: Struktur des Characters-Elements (Beispiel).

<Character id="ch1" name="Frosch" age="adult" gender="male" type="animal_monster" subtype="small" attitude="neutral" archetype="hero"> </Character>

Analog dazu enthält das Locations-Element untergeordnete Location-Elemente, die jeweils einen Ort codieren:

3: Struktur des Locations-Elements (Beispiel).

<Location id="loc1" type="WALD"> </Location>

Das Text-Element enthält schließlich den eigentlichen Märchentext. Dieser ist auf die verschiedenen Szenen – repräsentiert durch Szene-Elemente – aufgeteilt, welche wiederum die verschiedenen Dialogakte (Dialog-Elemente) enthalten:

4: Struktur des Text- und Szene-Elemente (Beispiel).

<Text><Scene id="s2" time="t2" location="loc1" characters="ch1,ch2" propp_functions="d|e" propp_subfunctions="D7|E10" transition="gehen">

...

<Dialogue id="d5" time="t2.4" speaker="ch2" receiver="ch1"> Ach, du bist's, alter Wasserpatscher, </Dialogue>

...

</Scene></Text>

Beim Entwurf des XML-Schemas wurde besonders Wert auf Übersichtlichkeit und Leserlichkeit gelegt. Trotz der Vielzahl der kodierten Informationen sind die resultierenden XML-Dateien daher vergleichsweise kompakt; so besteht die XML-Repräsentation des (vergleichsweise langen) Märchens „*Hänsel und Gretel*" bspw. nur aus 226 Zeilen.

Diese XML Repräsentation basiert auf und erweitert das Annotation Schema, das in (Scheidel & Declerck, 2010) beschrieben wird.

Python-Repräsentation

Auf der Grundlage der oben beschriebenen XML-Struktur kann eine Python-Klassenstruktur aufgebaut werden, die ein Märchen sowie seine einzelnen Teile als Python-Objekte repräsentiert.

Neben einer Oberklasse Tale gibt es für jeden der oben beschriebenen Teile eine eigene Python-Klasse, d. h. die Klassen Location, Character, Scene und Dialogue. (Insgesamt bestehen die Dateien zur Märchen-Repräsentation aus 288 Zeilen Code.) Jede Klasse enthält dabei als Attribute die oben beschriebenen Eigenschaften, wobei diese auch Verweise auf andere Elemente darstellen können. So verweisen bspw. Dialogue-Objekte auf die Character-Objekte von Spre-

cher und Empfängern. Der Python-Code dient als Interface für drei Anwendungen. Erstens können Märchen aus bestehenden XML-Dateien eingelesen werden; zweitens können XML-Dateien anhand einer anderweitig (z. B. durch automatische Klassifizierung) erzeugten Python-Märchenstruktur generiert werden; und drittens kann anderer Python-Code auf die Märchen-Information zugreifen, was die Grundlage für Anwendungen wie Text-to-Speech oder Visualisierung bildet. Sowohl die XML Kodierung als auch die Python Objekte interagieren mit einer Märchen-Ontologie interagieren, die eine Erweiterung der in (Koleva et al., 2012) beschriebenen Ontologie ist.

Somit haben wir eine formale Repräsentation von Märchen, die in verschiedenen Anwendungen zum Tragen kommen kann.

Fußnoten

1. Mit Beiträgen von Anastasija Aman, Stefan Grünewald, Matthias Lindemann, Lisa Schäfer, Natalia Skachkova.

Bibliographie

Propp, Vladimir ; Scott, Laurence (Hrsg.): Morphology of the folktale. 2. überarbeitete Auflage. Austin, TX u.a., 1977

Antonia Scheidel and Thierry Declerck. 2010. Apftml - augmented proppian fairy tale markup language. In Sándor Dar´anyi and Piroska Lendvai, editors, First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts. Szeged University.

Nikolina Koleva, Thierry Declerck, and Hans-Ulrich Krieger. 2012. An ontology-based iterative text processing strategy for detecting and recognizing characters in folktales. In Jan Christoph Meister, editor, Digital Humanities 2012 Conference Abstracts, pages 467–470, Hamburg, 7. University of Hamburg, Hamburg University Press

hermA. Zur Rolle von Annotationen in hermeneutischen Prozessen

Adelmann, Benedikt
adelmann@informatik.uni-hamburg.de
Universität Hamburg

Andresen, Melanie
Melanie.Andresen@uni-hamburg.de
Universität Hamburg

Begerow, Anke
anke.begerow@haw-hamburg.de
Hochschule für angewandte Wissenschaften, Hamburg

Gaidys, Uta
uta.gaidys@haw-hamburg.de
Hochschule für angewandte Wissenschaften, Hamburg

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

Koch, Gertraud
gertraud.koch@uni-hamburg.de
Universität Hamburg

Menzel, Wolfgang
menzel@informatik.uni-hamburg.de
Universität Hamburg

Orth, Dominik
dominik.orth@uni-wuppertal.de
Bergische Universität Wuppertal

Topp, Sebastian
sebastian.topp@uni-hamburg.de
Universität Hamburg

Vauth, Michael
michael.vauth@tuhh.de
Technische Universität Hamburg

Zinsmeister, Heike
heike.zinsmeister@uni-hamburg.de
Universität Hamburg

Das Projekt hermA

Der Forschungsverbund „Automatisierte Modellierung hermeneutischer Prozesse“ (hermA) befasst sich im Rahmen einer interdisziplinären Zusammenarbeit von Literaturwissenschaft, Pflegewissenschaft, Kulturanthropologie, Computerlinguistik und Informatik mit der Frage, ob und inwieweit hermeneutisches Arbeiten im Bereich der sozial- und geisteswissenschaftlichen Textanalyse computergestützt automatisiert werden kann. Von der Auseinandersetzung mit dieser Frage sind zum einen Erkenntnisse über die Verwendung und Funktion von Annotationen in den jeweiligen hermeneutischen Prozessen zu erwarten, zum anderen sollen erste Ansätze zur Automatisierung des Analyseprozesses entwickelt werden, die die Auswertung größerer Textmengen unterstützen. Die fünf Teilprojekte von hermA folgen in ihrer hermeneutischen Arbeit an und mit Texten unterschiedlichen Forschungslogiken (deduktiv, induktiv und/oder abduktiv); sie arbeiten außerdem jeweils eigenständig zu einem Thema im Gesundheitsbereich und stellen damit thematisch verbundene, fachdisziplinäre Forschungsszenarien zur Evaluation der automatisierten Modelle zur Verfügung:

1. Das literaturwissenschaftliche Teilprojekt 1 „Annotationen und die Erkennung von Genremustern. Medizintechnik in literarischen Anti-Utopien“ untersucht ausgehend von der Rolle von Medizintechnik in dystopischen Welten, inwiefern sich Genremerkmale zu Analysekategorien für Dystopien operationalisieren lassen.
2. Das ebenfalls literaturwissenschaftliche Teilprojekt 2 „Gender und Krankheit“ betrachtet literarischen Figuren im Hinblick auf Zusammenhänge zwischen Gender und Zuschreibung von Krankheit.
3. Das pflegewissenschaftliche Teilprojekt 3 „Bedeutungszuschreibungen in krisenhaften gesundheitlichen Versorgungssituationen“ analysiert das Verständnis von sterbenden Menschen hinsichtlich der Entscheidungen bezüglich ihrer gesundheitlichen Versorgung auf Basis qualitativer Interviews.
4. Das kulturanthropologische Teilprojekt 4 „Akzeptanzproblematiken von Telemedizin“ untersucht die sich entwickelnden Kernproblematiken für die Akzeptanz der neuen Technologien der telematischen Medizin und ihre Konsequenzen für das soziale und kulturelle Zusammenleben.
5. Das computerlinguistische Teilprojekt 5 adaptiert maschinell erzeugte Modelle für domänenrelevante Interpretationen und unterstützt die übrigen Teilprojekte in der automatischen Textverarbeitung.

Hermeneutische Prozesse und Forschungslogiken

Die Teilprojekte im Projekt hermA decken die gesamte Bandbreite an hermeneutischen Vorgehensweisen ab, die sich auf deduktive, induktive und abduktive Schlussverfahren zurückführen lassen. Damit geht es um alle drei darauf basierende Forschungslogiken, die Charles Sanders Peirce folgendermaßen zusammenfasst: „Deduction proves that something *must be*; Induction shows that something *actually is* operative; Abduction merely suggests that something *may be*“ (Peirce, 1934, CP 5.171, Hervorhebungen im Original).

Es werden also jeweils bestimmte Strategien genutzt, um unterschiedliche Arten von Erkenntnissen zu erlangen:

- *deduktives Vorgehen:* Mithilfe von etablierten Regeln werden ausgehend von beobachteten Phänomenen Schlüsse gezogen.
- *induktives Vorgehen:* Auf Basis beobachteter, systematisch auftretender Phänomene werden Regeln formuliert.
- *abduktives Vorgehen:* Für die Erklärung neu beobachteter Phänomene werden neue Hypothesen über die Ursachen der Phänomene entwickelt.

Hinzu kommt: In jedem hermeneutischen Erkenntnisprozess werden laufend neue Erkenntnisse generiert. Wenn diese sich nicht in die jeweilige Forschungslogik integrieren lassen, müssen die entsprechenden Hypothesen und/oder Vorhersagen revidiert oder erweitert und anschließend erneut angewendet werden. Dabei ist es zum Teil nötig, auf andere Forschungslogiken zurückzugreifen (etwa durch eine induktive Herleitung einer neuen Regel, die im deduktiven Prozess angewendet werden kann).

Annotation in hermeneutischen Prozessen

Ein erstes Zwischenergebnis des Projekts hermA ist, dass die Rolle von Annotationen in her-

meneutischen Prozessen von der jeweilig zur Anwendung kommenden Forschungslogik abhängt. Diese Zusammenhänge zwischen den Forschungslogiken und dem Einsatz von Annotationen sollen auf dem vorgeschlagenen Poster mit Blick auf die von den Teilprojekten verfolgten Forschungsfragen dargestellt werden. Dabei geht es um folgende Aspekte:

- Annotationen dienen beim *deduktiven* Vorgehen dazu, Kategorien an Gegenstandstexten zu erproben und gegebenenfalls Modifikationen am Kategorienset vorzunehmen. Auf diesem Weg können die Analysekategorien bestenfalls auch für automatische Annotationen operationalisiert und durch die generierten automatischen Annotationen überprüft werden. In den literaturwissenschaftlichen Teilprojekten 1 und 2 geschieht das mit narratologischen, Genre- und Gender-Kategorien. Teilprojekt 5 nutzt den deduktiven Zugang für die Automatisierung von Analysen.
- Beim *induktiven* Forschungsansatz werden Annotationskategorien aus dem Forschungsobjekt abgeleitet, um nachhaltig durchsuchbare Daten zu generieren und diese deutend zu interpretieren. Dies ist insbesondere in Teilprojekt 3 der Fall, wo die zentralen Aspekte der Entscheidungssituationen sterbender Menschen herausgearbeitet werden. Teilprojekt 1 und 2 arbeiten ebenfalls induktiv an noch nicht etablierten Analysekategorien.
- Beim *abduktiven* Forschungsansatz ist die sukzessive Entwicklung des Annotationsschemas ein zentrales Mittel des hermeneutischen Zugangs zum Untersuchungsobjekt; der Aufbau des Schemas erfolgt parallel sowohl zur Datenerhebung als auch zur Anwendung des Annotationsschemas selbst. Dabei hängen die Struktur des Schemas und die Erkenntnis über das Untersuchungsobjekt direkt miteinander zusammen. Teilprojekt 4 arbeitet abduktiv, um relevante Analyseobjekte zu identifizieren und einen Zugang zu ihnen zu entwickeln.

Bei der Betrachtung der Rolle von Annotationen muss zusätzlich zwischen manuellen und automatischen Annotationen differenziert werden. Während manuelle Annotationen eher zur Entwicklung oder Überprüfung von Hypothesen genutzt werden, unterstützen die automatisierten Zugänge jene Aspekte der hermeneutischen Prozesse, die bereits klar definiert werden können – etwa die Erkennung bereits operationalisierter Phänomene oder die Identifikation relevanter Texte oder Textstellen für die weitere Analyse und Interpretation.

Bibliographie

Chilese, Vivian, und Heinz-Peter Preußer (Hrsg.). 2013. *Technik in Dystopien*. Jahrbuch Literatur und Politik 7. Heidelberg: Winter.

Gaidys, Uta, und Valerie Fleming. 2005. „Gadamers philosophische Hermeneutik in der Pflegeforschung? Eine Diskussion“. *Pflege* 18 (6): 389–95.

Gius, Evelyn, und Janina Jacke. 2017. „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis“. *International Journal of Humanities and Arts Computing* 11 (2): 233–54.

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

McCrae, Patrick, Wolfgang Menzel, und Maosong Sun. 2011. „A Computational Model of Concept Generalization in Cross-Modal Reference“. *Tsinghua Science & Technology* 16 (2): 113–20.

Moretti, Franco. 2013. „‘Operationalizing’. Or, the Function of Measurement in Literary Theory“. *New Left Review*, No. 84 (Dezember 2013): 103–19.

Peirce, Charles Sanders (1934): „Pragmatism and Pragmaticism”. Charles Hartshorne & Paul Weiss (Hrsg.): *Collected Papers of Charles Sanders Peirce*. Cambridge, Massachusetts: Harvard University Press.

Pustejovsky, James, und Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. Beijing: O’Reilly.

Schramme, Thomas (Hrsg.). 2011. *Krankheitstheorien*. Suhrkamp-Taschenbuch Wissenschaft. Berlin: Suhrkamp.

Strauss, A.L., und J. Corbin. 1996. *Grounded theory: Grundlagen qualitativer Sozialforschung*. Weinheim: Beltz Psychologie Verlags Union.

Zinsmeister, Heike. 2015. „Chancen und Grenzen von automatischer Annotation“. *Zeitschrift für germanistische Linguistik* 43 (1): 84–110.

IncipitSearch - Vernetzung musikwissenschaftlicher Vorhaben

Neovesky, Anna
Anna.Neovesky@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

von Vlahovits, Frederic
Frederic.vonVlahovits@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

Die digitalen Musikwissenschaften konzentrieren sich bis dato vor allem auf die Erstellung digitaler Editionen sowie die Entwicklung von Werkzeugen und Standards für deren Erarbeitung und Publikation. Während die Daten vieler Edition mittlerweile nachhaltig und nachnutzbar zur Verfügung gestellt werden, mangelt es noch an übergreifenden und vernetzenden Datenaggregatoren, die die musikwissenschaftlichen Vorhaben vernetzen. Anders gesagt: Vor lauter Edieren, hat man sich noch zu wenig auf die Zusammenführung der bereits existenten Daten konzentriert. Gerade das wäre jedoch mithilfe der über die Summe der recht heterogenen Datenbestände hinweg normalisierbaren Metadaten gut möglich.

Die IncipitSearch [1] der Akademie der Wissenschaften und der Literatur | Mainz setzt genau hier an, indem sie musikalische Incipits durchsuchbar macht. Da Incipits einen pragmatischen Ansatz zur Katalogisierung von notierter Musik darstellen, bei dem wenige Anfangstakte einer Partitur transkribiert werden, lassen sich hierüber Editionen, Werkverzeichnisse und Quellenkataloge digitalen wie gedruckten Ursprungs zusammenführen. Ziel eines solchen Ansatzes ist nicht nur die Möglichkeit einer repositorienübergreifenden Suche, denn es lassen sich mithilfe von Incipits sehr wohl auch rudimentäre Aussagen über Spezifika der jeweils referenzierten Musik treffen. Ein erster vergleichender analytischer Blick auf eine breite Datengrundlage ist also ebenfalls möglich.

In erster Linie versteht sich die IncipitSearch jedoch als ein Pendant zu konventionellen thematisch-bibliographischen Katalogen, wie sie in gedruckter Form schon seit Jahrhunderten vorgelegt werden. Man stelle sich vor, man sucht eine Sonate in C auf einem bestimmten Anfangston oder mit einer bestimmten Anfangsmelodie, ohne den Komponisten zu kennen. Nur allein mit dem Titel „Sonate in C“ dürfte man eine unüberschaubare Anzahl an Treffern erzielen. Mittels einer simplen Eingabe der Anfangstöne auf einer Klaviatur jedoch, lässt sich das gesuchte Stück leicht eruieren. Mit RISM hat man ja bereits ein hervorragendes Beispiel für einen solchen Suchansatz, der mit Notenincipits arbeitet. [2] Gerade eine Zusammenführung verschiedener Datenrepositorien – von Digitalen Werkverzeichnisse, Editionen über Quellenlexika und -sammlungen – wäre mehr als wünschenswert.

Eine wichtige Voraussetzung für den Erfolg eines solchen Unterfangens ist sowohl die freie Verfügbarkeit der Daten auf Seite der Datenlieferanten als auch die freie Verfügbarkeit der Software auf Seite der Anbieter des Dienstes. Die IncipitSearch ist somit neben ihrer bibliographischen Funktion gleichsam auch ein an die musikwissenschaftlichen Fachcommunity gerichteter Aufruf, ihre Daten nach einem einheitlichen Schema offen zu legen, um Sie für eine Vernetzung in diesem wie auch in weiteren Diensten nachnutzbar zu machen. Dafür wird auf einen bewährten Standard für die Codierung der Incipits gesetzt, nämlich Plaine & Easie Code. Es handelt sich um eine simple, schlanke Transkriptionssprache, die von der International Association of Music Libraries, Archives and Documentation Centres kuratiert wird (Brook / Gould 1964). Da es sich bei Plaine & Easie Code um eine Buchstabennotation handelt sind die Incipits besonders einfach zu codieren.

Der Dienst IncipitSearch selbst ist als Microservice konzipiert und in Gänze sowie bezogen auf seine einzelnen kapselbaren Bestandteile selbst nachnutzbar (Haft / Neovesky / Reimers 2016). Sein Quellcode ist vollständig MIT-lizensiert offen gelegt auf Github verfügbar. [3] Das, eine gute Dokumentation und ein zukünftiges Workshop-Angebot sollen die Zugangsschwelle für potenzielle Nutzer, vor allem seitens der Datenhalter so niedrig wie möglich legen. Damit wird die Voraussetzung für eine Anbindung eigener Daten sowie die Nachnutzung sowohl des Dienstes als auch der Daten mit möglichst geringem Aufwand geschaffen.

Mit der IncipitSearch wird den best practices moderner DH folgend, das Potenzial einer Nische genutzt, indem zahlreiche große und kleine, bisher weitestgehend als Inseln dastehende Datenrepositorien der Musik mittels kompakt-gekapselter Software, einer Datenschnittstelle und eines simplen Transkriptionsstandards von Noten-Incipts zusammengeführt werden. Bei der Implementierung weiterer Repositorien setzen wir „community driven“ bewusst auf einen bottom-up-Ansatz, um die Akzeptanz des Dienstes zu stärken und auf Bedürfnisse aus der Community angemessen reagieren zu können. Das Kernpotenzial der Anwendung liegt aber auch in seinen Anforderungen an die Community: Mehr Mut zur Datentransparenz und mehr Mut zur Vernetzung.

Fußnoten

1. Seit Dezember 2017 ist eine BETA-Version der Plattform online unter https://incipitsearch.adw-mainz.net verfügbar.
2. RISM, zweifellos das etablierteste musikwissenschaftliche Datenrepositorium, bietet bereits die Möglichkeit die dort verzeichneten Manuskripte nach Incipits zu durchsuchen (http://www.rism.info/en/home/newsdetails/select/rism_online_catalog/article/2/music-incipit-searches.html). Die Datenbestände wurden bereits in anderen Projekten nachgenutzt (Van Nuss / Giezeman / Wiering 2017).
3. Zur Softwarekomponente siehe https://github.com/annaneo/incipitSearch sowie https://github.com/digicademy/incipitSearch für die Webplattform. Für die Anzeige der Incipits kommt die vom RISM-Schweiz entwickelte Programmbibliothek Verovio zum Einsatz: https://github.com/rism-ch/verovio . Um eine komfortable Noteneingabe zu ermöglichen wurde eigens ein JavaScript-PianoKeyboard entwickelt: https://github.com/annaneo/pianoKeyboard .

Bibliographie

Brook, Barry S. / Gould, Murray. (1964): "Notating Music with Ordinary Typewriter Characters (A Plaine and Easie Code System for Musicke)", in: Fontes Artis Musicae 11 (3): 142–159.

Haft, Michael / Neovesky, Anna / Reimers, Gabriel (2016): „Digitale Nachhaltigkeit von Forschungsdaten durch Microservices.", in: FORGE 2016. Forschungsdaten in den Geisteswissenschaften: Conference Abstract: 23 -24.

Van Nuss, Jelmer / Giezeman, Geert-Jan, Wiering, Frans (2017). „Melody retrieval and composer attribution using sequence alignment on RISM incipits.", in: Proceedings of the International Conference on Technologies for Music Notation and Representation. Universidade da Coruña: 79–90 http://doi.org/10.5281/zenodo.924135 [letzter Zugriff 12. Januar 2018].

Ist die DARIAH-DE Forschungsinfrastruktur fit für Daten der realen Welt? Bericht über einen Anwendungsfall mit archäologischen Daten und seine ersten Ergebnisse

Romanello, Matteo
matteo.romanello@dainst.de
Deutsches Archäologisches Institut, Deutschland

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg, Deutschland

Hintergrund

Eine wesentliche Kritik an Forschungsinfrastrukturen behauptet:

This is the central paradox for big infrastructure design: the very wish to cater to everyone pushes the designers toward generalization, and thus necessarily away from delivering data models specific enough to be useful to anyone. (van Zundert 2012: 172)

Können generische Infrastrukturen und Datenmodelle für individuelle Forschungsfragen von Bedeutung sein? Und wenn ja, wie verhalten sich spezifische Forschungsdaten gegenüber den generischen Infrastrukturen? In diesem Beitrag diskutieren wir diese Frage im Hinblick auf die von DARIAH-DE [1] entwickelte Forschungsdateninfrastruktur und insbesondere das Data Modelling Environment (DME). Als Schema und Crosswalk Registry entstanden (Gradl et al. 2015), entwickelte sich das DME zu einem umfangreichen Werkzeug für die Modellierung, Verfeinerung, Bereinigung und Anreicherung von Daten. Die Beispieldaten, die für diesen Anwendungsfall herangezogen wurden, stammen aus einer Datenbank der aus dem Deutschen Archäologischen Institut geführten Grabung in Pergamon [2] und beschreiben etwa 100 keramische Grabungsfunde.

Für den Anwendungsfall wurde ein archäologischer Kontext gewählt, da relevante For-

schungsdaten aufgrund ihrer Heterogenität eine besondere Herausforderung für Forschungsinfrastrukturen darstellen (Gradl, Henrich 2016a). Das wesentliche Ziel dieses Beitrags besteht darin, die Verwendbarkeit des DME auch im spezifischen Kontext von Pergamon-Daten anzudeuten. Eine Integration weiterer archäologischer Daten wie 2D-Bildern, 3D-Modellen und verschiedener Arten kontrollierter Vokabulare und geographischer Daten könnten für den gewählten Anwendungsfall Erkenntnisgewinne erreicht werden, die ggf. neue Fragen für die qualitative Forschung aufwerfen.

Durch eine kombinierte Visualisierung orts- und zeitbezogener Abhängigkeiten könnte man sich schnell einen ersten Überblick über die zeitliche und geographische Verteilung der Datensätze verschaffen. Wo liegt z. B. die höchste Dichte von auf die hellenistische Zeit datierten, keramischen Funden vor? Ein solcher visueller Überblick über die Grabungsdaten könnte den WissenschaftlerInnen auch erlauben, Diskrepanzen und Sonderfälle in der archäologischen Dokumentation einer Grabung zu erkennen.

Beschreibung des Anwendungsfalls

Die Verarbeitung der Daten wird unterstützt durch das DME und insbesondere dessen Fähigkeit, auf externe Ressourcen zuzugreifen. Zwei Schnittstellen zu Diensten des DAI wurden implementiert, damit zeit- und ortsbezogene Textangaben wie „Grobdatierung: hellenistisch-kaiserzeitlich“ oder „Provenienz: Pergamon“ mit den entsprechenden und in Zahlen ausgedrückten Werten kartiert werden können. Schließlich werden die angereicherten Daten mittels des DARIAH-DE Geo-Browsers visualisiert, um die zeitliche und geographische Verteilung der in den Datensätzen beschriebenen Objekte visuell abzubilden.

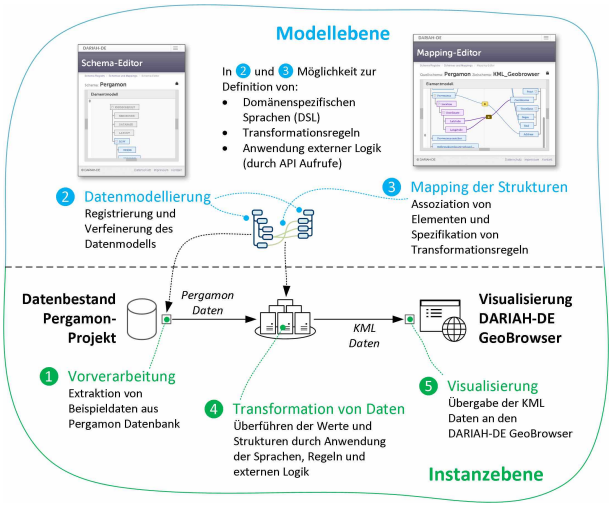


Abb. 1: Schematische Darstellung des Arbeitsablaufs.

Der Arbeitsablauf besteht aus fünf wesentlichen Schritten, die sich der *Modell-* oder *Instanzebene* zuweisen lassen.

- *Modellebene* : Datenmodelle und semantische Verbindungen zwischen diesen werden spezifiziert.
- *Instanzebene* : Während Aufgaben der Modellierung einmalig auszuführen sind, werden die Aufgaben der Instanzebene für jede Datei bzw. jede Aktualisierung der Daten durchlaufen.

Vorverarbeitung

Die archäologische Grabung in Pergamon wurde mittels der iDAI.Field Software dokumentiert. iDAI.Field ist ein modulares Dokumentationssystem für Feldforschungsprojekte, das am DAI entwickelt wurde und in ca. 50 verschiedenen Projekten eingesetzt wurde. [3] Die durch iDAI.Field gesammelten Daten werden in einer FileMaker-Datenbank gespeichert. Für eine Verarbeitung in der DARIAH-DE Infrastruktur wurde zunächst ein XML-Export aus der Datenbank ausgeführt.

Datenmodellierung

Um Pergamon-Daten in ein vom Geo-Browser unterstütztes Eingabeformat, wie die Keyhole Markup Language (KML) [4] umwandeln zu können, müssen die relevanten Datenmodelle im DME vorliegen bzw. definiert werden. Dies kann durch das Hochladen von XSD-Schemata initiiert werden. Einmal hinterlegte Modelle können nach

deren Definition in weiteren Anwendungsfällen nachgenutzt werden.

Abbildung 2 veranschaulicht neben dem Elementmodell auch die Funktionalität zur Verarbeitung von Beispieldaten, mit deren Hilfe überprüft werden kann, ob Daten korrekt prozessiert werden.

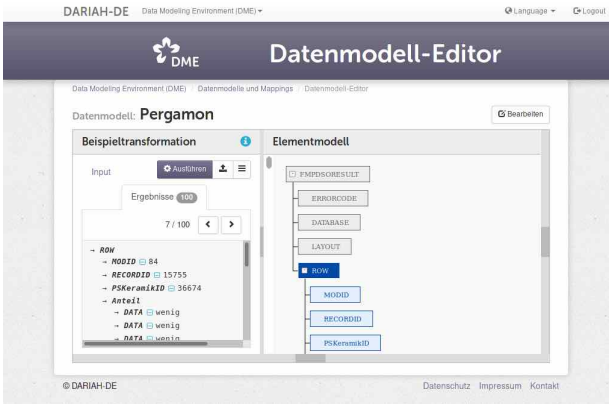


Abb. 2: Verarbeitung von Beispieldaten und Visualisierung der Datenstruktur.

Datenanreicherung

Die Funktionalität des DME stellt zwei wesentliche Methoden zur Datenmodellierung bereitgestellt (Gradl, Henrich 2016b):

- Inhaltliche Spezifikation von Daten durch die *Definition von domänenspezifischen Sprachen* (Parr 2012)
- Anwendung von Transformationsregeln. Neben bereits implementierter Funktionalität - z.B. zur automatischen Sprachverarbeitung - kann das DME flexibel durch Plugins erweitert werden, um neue Funktionen zur Verarbeitung von Daten einzubinden.

Im vorliegenden Anwendungsfall sind Daten in eine strukturierte und außerhalb der Pergamon-Datenbank interpretierbare Form zu bringen. Hierzu werden Daten u. A. durch die Nutzung des iDAI.Gazetteer [5] (Auflösung von Ortsbezeichnungen) und der iDAI.Chronontology [6] (Auflösung zeitlicher Angaben) angereichert.

iDAI.ChronOntology

Die ChronOntology API ermöglicht u. a. eine Freitextsuche. Beispielsweise ist es möglich nach Zeitangaben zu suchen, die den String „Kaiserzeitlich“ beinhalten [7] und die entsprechende Datierungen aufweisen können. So ist der Begriff „kaiserzeitlich“ mit „-27“ und „476“ als Beginn- und Enddatum verbunden.

Im Rahmen des DME wird das Modell der Pergamon-Daten dahingehend erweitert, dass unter dem in XML vorhandenen Element <Grobdatierung> zunächst Grammatik und Transformationsregel angelegt werden. Hierunter werden die zu produzierenden zusätzlichen Elemente modelliert: im konkreten Fall die strukturierte Antwort des iDAI.ChronOntology Dienstes. Abbildung 3 zeigt neben diesem erweiterten Elementmodell bereits das Ergebnis der Anwendung dieser Funktionalität auf die Beispieldaten.

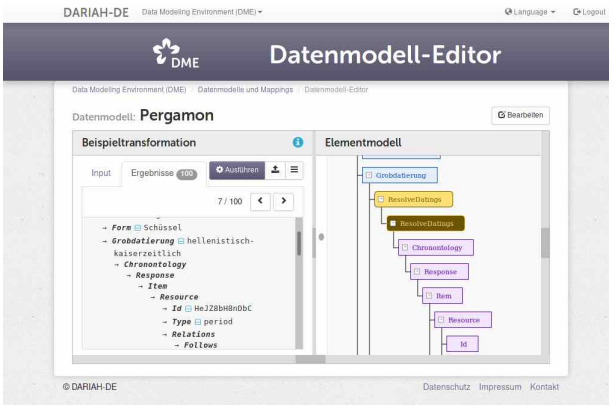


Abb. 3: Erweiterung des Datenmodells durch Zusatz der strukturierten Antwort des iDAI.ChronOntology Dienstes.

Für eine in Bezug auf die Pergamon-Daten optimierte Anfrage an den iDAI.ChronOntology Dienst wird die Semantik des Elements <Grobdatierung> expliziert. Die Grammatik in Abbildung 4 veranlasst die Zerlegung zusammengesetzter Datierungsangaben, um die vorliegende von-bis Semantik darzustellen (z. B. bei hellenistisch-kaiserzeitlich) und die einzelnen Anfrageterme zu extrahieren (Parr 2012).

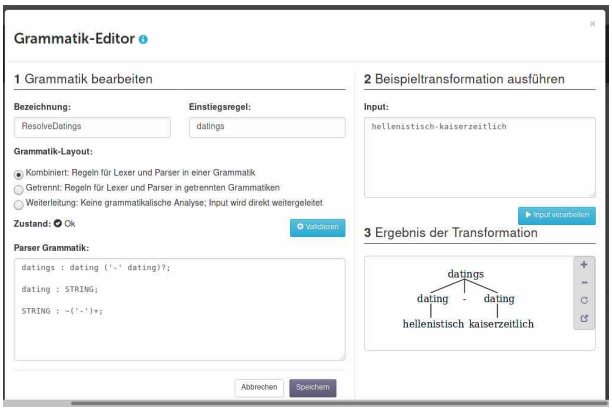


Abb. 4: Bearbeitung eines Elements des Datenmodells durch eine vom Benutzer editierbare Grammatik.

Durch die Adressierung der so gebildeten Terme in <dating> kann die anschließende Transformationsregel (vgl. Abbildung 5) auf eine verfeinerte Variante des zuvor unstrukturierten Inhalts zurückgreifen.

Die Ausführung der Chronontology API ist durch Anwendung von Funktionalität des umgesetzten DAI-Plugins möglich. Im vorliegenden Fall gestaltet sich das Kommando wie folgt:

Chronontology = dai.chronontology.query(@dating);

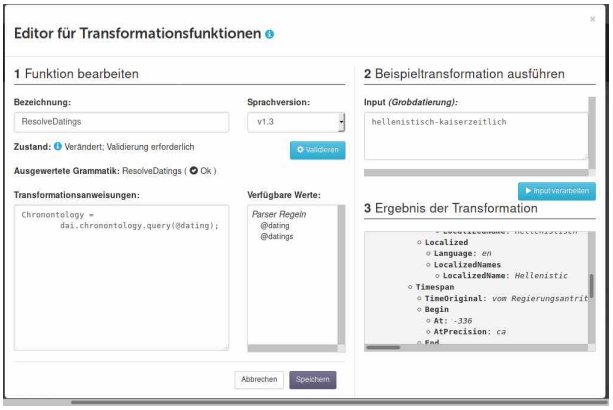


Abb. 5: Spezifikation einer Transformationsregel zur Abfrage des iDAI ChronOntology API.

Um aus der potenziellen Menge zurückgegebener Einträge ein Intervall zu berechnen, werden Kommandos aus dem math Funktionsraum verwendet:

BeginMin = math.min (@Response.Item.Resource.Timespan.Begin.At);
EndMax = math.max (@Response.Item.Resource.Timespan.End.At);

Hierdurch wird in den Daten exakt ein Zeitintervall hinterlegt, welches der gewünschten Semantik [frühester Begin, spätestes Ende] der Zeitangabe entspricht.

iDAI.Gazetteer

Vergleichbar mit der Chronontology API können auch Funktionen der Gazetteer API auf Daten angewandt werden. Im vorliegenden Beispiel wird der für eine Anfrage zurückgegebene, erste Treffer als wahrscheinlichste Koordinate verwendet und in den Daten berücksichtigt:

Location = dai.gazetteer.topcoord(@ResolveLocation);

Mapping der Datenstrukturen

Für die Transformation originärer Pergamon-Daten in das KML Format ist schließlich die Modellierung von Zusammenhängen der Datenmodelle erforderlich. Abbildung 6 zeigt die drei Mappings, die für den Anwendungsfall modelliert wurden.

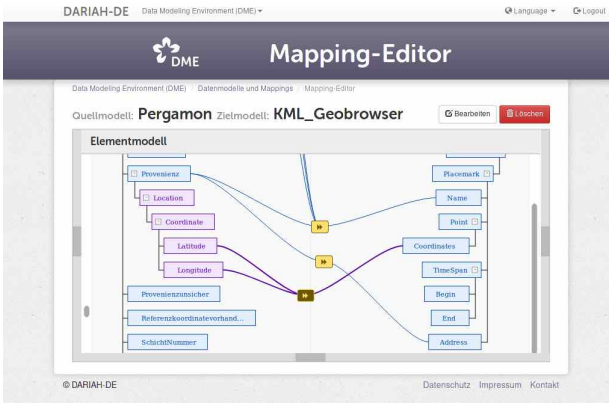


Abb. 6: Visualisierung der Mappings zwischen Quellmodell (Pergamon XML) und Zielmodell (KML) im DME.

Über einfache Wertkorrespondenzen, wie bei BeginMin (Pergamon) zu Begin (KML) hinaus können auch an dieser Stelle Transformationsregeln spezifiziert werden. Für die Übertragung der Koordinaten in das KML Schema wird so z. B. folgende Regel definiert:

[@Latitude != ""]
Coordinates = concat(@Latitude, ",", @Longitude)
[endif]

Koordinaten werden demnach nur angelegt, wenn @Latitude (für Daten im Quellschema) gesetzt ist. Zur Erzeugung eines Strings “Latitude, Longitude” wird die Konkatenationsanweisung verwendet.

Visualisierung der Mapping-Ergebnisse

Transformierte Daten können in verschiedenen Formaten heruntergeladen werden. Als KML Datei exportiert, können die 100 archäologischen Beispieldatensätze im GeoBrowser bereitgestellt und angezeigt werden (vgl. Abbildung 7).

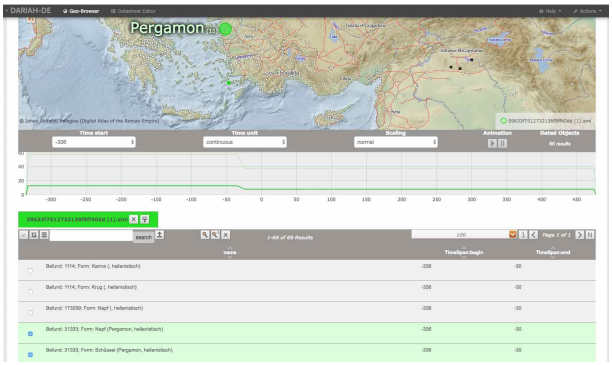


Abb. 7: Visualisierung der Mapping-Ergebnisse mittels des Geo-Browsers.

Nur 16 der 100 Datensätze haben Ortsangaben und können deshalb positioniert werden (Pergamon: 15, Knidos: 1), während fast alle eine Zeitangabe aufweisen. Die Möglichkeit, eine historische Karte (hier der *Barrington Atlas map of the Roman Empire*) auszuwählen, bietet einen zusätzlichen Nutzen, da sie eine bessere Kontextualisierung der Daten ermöglicht. Da der GeoBrowser derzeit keine XML-Namespaces unterstützt, müssen diese im Moment manuell aus den KML Daten entfernt werden.

Schlußdiskussion

Dieser Anwendungsfall basierte auf einer zu geringen Menge von Daten, als dass akute Mehrwerte erreichen werden könnten. Die Visualisierung von Daten aus mehreren Grabungsorten könnte dagegen die Einführung neuer Formen, Farben oder Keramiktypen in ort- und zeitbezogener Abhängigkeit darstellen und so zu der Generierung neuer Hypothesen führen. Das DME ist flexibel genug, um mit den heterogenen Daten der Archäologie umgehen zu können.

Indem das DME eine Modellierung von Verarbeitung von Daten spezifischer Anwendungsfälle ermöglicht, hat es das Potential das „OpenRefine für die digitalen Geisteswissenschaften“ zu werden: ein generisches Tool zur Modellierung, Verfeinerung, Bereinigung und Anreicherung von Forschungsdaten, das eine breite Vielfalt von Arbeitsabläufen unterstützen kann. [8]

Zugleich stellt sich aber auch die Frage, wer die typischen BenutzerInnen des DME sein können? Oder: ist es realistisch zu erwarten, dass GeisteswissenschaftlerInnen dieses Werkzeug ohne die Unterstützung von DH Spezialisten bedienen können? Tatsächlich scheint das DME eine gemeinsame Basis der Kollaboration und Kommunikation sein zu wollen, in der das Wissen von GeisteswissenschaftlerInnen mit der technischen Expertise von DH-Experten zusammengeführt werden. Hierdurch können Aufgaben, wie die des vorliegenden Anwendungsfalls erfüllt werden ohne sämtliche technische Problemstellungen von Grund auf neu lösen zu müssen. Durch die wachsende Zahl von bestehenden Quell-/Zielmodelle, Transformationsregeln und API-Wrappers kann Wissen und Funktionalität nachgenutzt werden.

Fußnoten

1. *http://de.dariah.eu*
2. http://www.dainst.org/projekt/-/project-display/14186
3. https://www.dainst.org/forschung/forschung-digital/idai.welt/data/-/asset_publisher/Pt831IfwO8uH/content/idai-field
4. *https://wiki.de.dariah.eu/display/publicde/Geo-Browser+Dokumentation#Geo-BrowserDokumentation-Spezifikationenf%C3%BCrdieNutzung*
5. *https://gazetteer.dainst.org* . Vgl. auch Cuy et al. 2014.
6. http://chronontology.dainst.org/
7. Z.B. http://chronontology.dainst.org/data/period/?q=kaiserzeitlich
8. Für ein Beispiel der Benutzung von OpenRefine in den digitalen Geistenwissenschaften vgl. https://programminghistorian.org/lessons/cleaning-data-with-openrefine .

Bibliographie

van Zundert, J., (2012): “If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities”. *Historical Social Research*, 37(3), pp.165–186.

Cuy, Sebastian / Gerth, Philipp / Heiden, Maximilian / Kolbmann, Wibke / Schmidle, Wolfgang (2014): iDAI.gazetteer – ein Referenzsystem für altertumswissenschaftliche Ortsinformationen als Teil einer digitalen Forschungsinfrastruktur. In *Kölner und Bonner Archaeologica* 4, S. 203-212.

Gradl, Tobias / Henrich, Andreas (2016): Die DARIAH-DE-Föderationsarchitektur: Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen, Bibliothek Forschung und Praxis. Band 40, Heft 2, Seiten 222-228, ISSN (Online) 1865-7648, ISSN (Print) 0341-4183, DOI: https://doi.org/10.1515/bfp-2016-0027 , Juli 2016

Gradl, Tobias / Henrich, Andreas (2016): „Data Integration for the Arts and Humanities: A Language Theoretical Concept“. In: Fuhr, Norbert et al. (Hg.): Research and Advanced Technology for

Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings. Cham: Springer International Publishing, S. 281–293

Gradl, Tobias / Henrich, Andreas /Plutte, Christoph (2015): „Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Förderation von Kollektionen“. In: Baum, Constanze/Stäcker, Thomas (Hg.): Grenzen und Möglichkeiten der Digital HumanitiesZeitschrift für digitale Geisteswissenschaften. 2015, H. 1. URL: http://zfdg.de/sb001_020

Parr, Terence (2012): The definitive ANTLR 4 reference. 2. Aufl. Dallas, Raleigh: Pragmatic Bookshelf (= The pragmatic programmers)

“Kann man da eben mal was eintragen und visualisieren?” Digitaler Praxistest für die DARIAH-DE-Infrastruktur

Klaffki, Lisa
klaffki@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

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

Die Ausgangslage: Diskrepanz zwischen Anspruch und Wirklichkeit

In den letzten Jahren wurden vermehrt digitale Services und Ressourcen entwickelt, die EinzelforscherInnen, aber auch Projektgruppen in der Arbeit mit digitalen Methoden, der Generierung, dem Forschungsdatenmanagement und in toto bei der Transformation von Arbeitsprozessen in das digitale Medium unterstützen sollen. Signifikante Bestandteile des Research Data Lifecycles können bereits über digitale Angebote abgedeckt werden (Puhl et. al. 2015). Ein wesentliches Ziel, welches mit dem Aufbau der digitalen Infrastrukturen verfolgt wird, ist, dass auch diejenigen Forschungsvorhaben davon profitieren sollen, in denen keine oder nur geringe technischen Kenntnisse oder eine DH-Unterstützung vorhanden sind. Aktuell besteht zwischen Anspruch und Wirklichkeit aber noch eine Diskrepanz: Auf der einen Seite scheint mittlerweile die Mehrheit der WissenschaftlerInnen den Nutzen von Normdaten, Integration ihrer Forschungsdaten in Suchmaschinen oder die Visualisierungen ihrer Daten in Forschungsumgebungen zu erkennen. Auf der anderen Seite steht aber die Kritik, dass die oftmals propagierte einfache und intuitive Nutzung der digitalen Angebote nicht vorhanden ist. Der Mehrwert, den die Services den eher auf ein analoges Projektergebnis ausgerichteten Projekten bieten, steht in keinem effizienten Verhältnis von Aufwand und Ertrag. Auch Schulungsvideos oder Workshops sind häufig zu unspezifisch oder nicht auf die vorhandenen Daten anwendbar.

Der Usecase: ein frühneuzeitlicher Auktionskatalog

Als Usecase dient das Forschungsprojekt “Erschließung frühneuzeitlicher Auktionskataloge” des Forschungsverbundes Marbach Weimar Wolfenbüttel (MWW) (Autorenbibliotheken). Die aus einem Auktionskatalog von 1670 rekonstruierte Gelehrtenbibliothek gehörte dem Chiliasten Benedikt Bahnsen (gest. 1669). Der aus Norddeutschland stammende und nach Amsterdam emigrierte Bahnsen war als Verleger, Buchhändler und Bücheragent, Mathematiker und Rechenmeister tätig. In diesem Projekt wurden in einer Excel-Tabelle alle Losnummern, d.h. Titel des Katalogs, erfasst. Die Tiefenerschließung umfasst die bibliographischen Angaben der verzeichneten Bücher, Normdaten zu Personen und Werken, Geodaten für die Druckorte sowie Links zu Digitalisaten und in Nachweissysteme. Mit der Erschließung wird dem Alleinstellungsmerkmal dieser Bibliothek Rechnung getragen, bei der es sich um das umfangreichste Einzelrepositorium für nonkonformes und heterodoxes Schrifttum handeln dürfte (Beyer et. al. 2017, 43–70).

Zwar verfügt das Projekt durch die Zusammenarbeit mit den DH-MitarbeiterInnen des Forschungsverbundes über eine eigene Datenbank und eine Präsentation im WWW (Auktionskataloge). Aber begonnen hatte das Projekt ohne deren Unterstützung und genau die in dieser Phase in einer Exceltabelle erhobenen Daten sollen für den Usecase mit DARIAH-DE verwendet werden.

Die Werkzeuge: DARIAH-DE-Dienste

Innerhalb von DARIAH-DE, einem digitalen Forschungsinfrastrukturprojekt für die Geistes- und Kulturwissenschaften, wurde eine *Data Federation Architecture* (DFA) entwickelt (Gradl / Henrich / Plutte 2015; Gradl / Henrich 2016). Unter diesem Begriff sind mehrere modulare Komponenten gebündelt, die für sich alleine oder im Zusammenspiel genutzt werden können. Davon nutzt der hier skizzierte Workflow für den Usecase das jüngst in die Produktivphase übergegangene *DARIAH-DE-Repository* zum Speichern und persistenten Adressieren der Forschungsdaten, die *Collection Registry* zum Verzeichnen der Datensammlung mitsamt Schnittstelle, um die Daten in die *Generic Search* zu integrieren und wiederauffindbar und somit nachnutzbar zu machen.

Für viele der in den verschiedenen Schritten eines Data Research Lifecycles anfallenden Aufgaben können also Komponenten der *DFA* zum Einsatz kommen (vgl. Abb. 1). Die *DFA* deckt aber nicht alle Schritte eines Forschungsprozesses und damit letztlich auch des Research Data Lifecycles ab. So bleibt der Abschnitt der Analyse oder Visualisierung offen, allerdings gibt in den digitalen Geisteswissenschaften hinreichend andere Tools, die diese Lücke zielgerichtet füllen können, zumal die häufig projektspezifischen Anforderungen hier kaum von einer generischen Lösung erfüllt werden können.

Für den Praxistest wird zur Visualisierung der räumlich-zeitlichen Daten ein ebenfalls aus dem DARIAH-DE-Projekt stammendes Tool verwendet, der *Geo-Browser* (Kollatz 2016).

Der Test: Von der grauen Theorie in die blaue DARIAH-DE-Praxis

Der Diskrepanz zwischen Wunsch und Wirklichkeit, zwischen Sinnhaftigkeit und Aufwand möchte sich diese Postereinreichung annehmen und überprüfen, wie exemplarische Forschungsdaten aus einem „analogen“, nicht von vornherein mit einer DH-Komponente geplanten Projekt in verschiedene Komponenten der Infrastruktur von DARIAH-DE integriert werden können.

Ziel ist es, dabei den Aufwand und die erforderlichen Kenntnisse zu evaluieren, die eine erfolgreiche Integration der Forschungsdaten bedingen. Die DARIAH-DE-Dienste eignen sich besonders für diesen Test, da sich die DARIAH-DE-Angebote ausdrücklich an EinzelforscherInnen richten (DARIAH-DE in Kürze).

Zwar handelt es sich nicht um eine belastbare empirische Studie, doch der mit den Projektbeteiligten durchgeführte Usecase hat dennoch aufgrund seines prototypischen Charakters richtungsweisende Bedeutung.

Bibliographie

Auktionskataloge: http://dev.hab.de/auktionskataloge [letzter Zugriff 11.09.2017].

Autorenbibliotheken: http://www.mww-forschung.de/forschungsprojekte/autorenbibliotheken [letzter Zugriff 11.09.2017].

Beyer, Hartmut et.al. (2017): “Bibliotheken im Buch: Die Erschließung von privaten Büchersammlungen der Frühneuzeit über Auktionskataloge”, in: Busch, Hannah / Fischer, Franz und Sahle, Patrick (Hrsg.): Kodikologie und Paläographie im digitalen Zeitalter 4 (Codicology and Palaeography in the Digital Age). Norderstedt: Books on Demand 43–70.

DARIAH-DE in Kürze: https://de.dariah.eu/dariah-de-in-kurze [letzter Zugriff 11.09.2017].

Gradl, Tobias / Henrich, Andreas / Plutte, Christoph (2015): “Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen”, in: Baum, Constanze / Stäcker, Thomas (Hrsg.): Grenzen und Möglichkeiten der Digital Humanities (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). text/html Format. DOI: http://dx.doi.org/10.17175/sb001_020 .

Gradl, Tobias / Henrich, Andreas (2016): “Die DARIAH-DE Föderationsarchitektur. Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen”, in: Neuroth, Heike et al. (Hrsg.): Bibliothek – Forschung und Praxis 40 (2): 222–228. DOI: https://doi.org/10.1515/bfp-2016-0027 .

Kollatz, Thomas (2016): “Raum-Zeit-Analysen mit Geo-Browser und Datasheet-Editor”, in: Neuroth, Heike et al. (Hrsg.): Bibliothek – Forschung und Praxis 40 (2): 229–233. https://doi.org/10.1515/bfp-2016-0032 .

Puhl, Johanna et. al. (2015): Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften (= DARIAH-DE Working Papers Nr. 11). Göttingen: DARIAH-DE. URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4 .

"Kinder des Buchdrucks" im Digitalen Zeitalter. Ein romanistisches Digital Humanities Modul

Burr, Elisabeth
elisabeth.burr@uni-leipzig.de
Universität Leipzig, Deutschland

Fußbahn, Ulrike
uf28bope@studserv.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

Schon länger wird die Frage nach der Lehre der Digital Humanities gestellt (cf. u. a. Hirsch (2012), Gold (2012), Warwick / Terras / Nyhan (2012)). Zudem sind Initiativen wie die EU geförderte *#dariahTeach* entstanden, die Materialien für die Lehre von Digital Humanities entwickelt und online zur Verfügung stellt. 2017 haben dann Fotis Jannidis, Hubertus Kohle und Malte Rehbein ein umfangreiches deutschsprachiges Lehrbuch zu *Digital Humanities* vorgelegt. In keiner dieser Publikationen wird aber, soweit wir sehen können, der Frage nachgegangen, wie der Auftrag, den die neue Epistemologie *Digital Humanities* als den ihren erkennt, nämlich ein umfassenderes und kritischeres Wissen von Artefakten sowie deren Relationen untereinander mittels der Nutzung von computationellen Methoden zu schaffen, in der Lehre umgesetzt werden kann. Einen Versuch in diese Richtung unternehmen wir in einem romanistischen Master-Modul, wo frühe gedruckte Grammatiken und Sprachtraktate zunächst als „Kinder des Buchdrucks“ eingeordnet und dann mittels Transkription und Auszeichnung für Forschungen im digitalen Zeitalter verfügbar gemacht werden.

Grundlagen des Versuchs

Dem Versuch zugrunde liegt ein Projekt, das es sich zum Ziel setzt, die kulturellen, politischen, gesellschaftlichen, ideologischen, genderbedingten etc. Vorstellungen, Ideen und Ziele, die sich hinter sprachbetrachtenden, sprachbeschreibenden und sprachnormierenden Werken, wie Grammatiken und Sprachtraktate, der frühen Gutenberg-Ära verbergen, vergleichend und in ihren Zusammenhängen über die einzelnen romanischen Kulturen und Sprachen hinweg zu untersuchen. Um diese Werke systematisch erforschbar zu machen, bedarf es allerdings zunächst einmal ihrer Digitalisierung und Erschließung.

Bei den Texten, die bisher berücksichtigt wurden, handelt es sich um:

- Francesco Fortunio (1516): *Regole grammaticali della volgar lingua*. Ancona: Bernardin Vercellese
- Louis Meigret (1550): *Le trętté de la GRAMMĘRE FRANCOĘZE*. Paris: Chrestien Wechel
- Speron Sperone (1542): “Dialogo delle lingue”, in: Speron Sperone: *I dialogi di Messer Speron Sperone*. Vinegia: Aldus 106-131
- Joachim Du Bellay (1549): *Deffence, et illustration de la langue francoyse*. Paris: Arnoul l’Angelier.

Als „Kinder des frühen Buchdrucks“ weisen diese Texte nicht nur eine Vielzahl von heute nicht mehr gebrauchten Lettern, von Abkürzungen und Variationen auf, sondern sie sind zugleich auch Ausdruck der durch den Buchdruck maßgeblich geförderten Hinwendung zur eigenen romanischen Volkssprache, zur eigenen Nation, zur Abgrenzung gegenüber anderen bzw. zum Vergleich oder gar zur Konkurrenz mit anderen. Zudem wird hier ein Modell der jeweiligen romanischen Sprache entwickelt und eine (soziale) sprachliche Norm fixiert, an der gerade auch die Drucker (nicht zuletzt aus ökonomischen Gründen) ein großes Interesse haben (cf. z. B. Tory (1529)). Darüber hinaus legen sie auch Zeugnis ab von der Entstehung eines Marktes, eines Lesepublikums und der Entwicklung von Layout- und Strukturierungsverfahren, die das Lesen und Verstehen von Texten fördern.

Umsetzung des Versuchs

Der Versuch wird in einem grundständigen und aus zwei Seminaren bestehenden Master-Modul zur französischen und italienischen Sprachwissenschaft unternommen. In diesem Modul sollen sich die Studierenden mit der Geschichte der Sprachbetrachtung und Normbildung bzw. Normierung und den dabei eingesetzten Werkzeugen auseinandersetzen. 2015 habe ich diesem Modul das oben genannte Projekt unterlegt und so konzipiert, dass die Studierenden selbst dazu einen Beitrag leisten können. Die beiden Seminare werden seither dazu genutzt, den Studierenden zu ermög-

lichen, alle für das Projekt benötigten Kenntnisse und Fertigkeiten zu erwerben und dabei auch ein Verständnis für die *Digital Humanities*, ihre Ziele und Methoden zu entwickeln.

Umfassendes und kritisches Wissen von diesen „Kindern des Buchdrucks“ und ihrer Bedeutung für die romanischen Sprachen sowie ein Verständnis für die Implikationen von Medienrevolutionen allgemein und der Gutenberg-Ära sowie der digitalen Revolution im Besonderen erwerben die Studierenden in den beiden Seminaren durch das gemeinsame Aufarbeiten von Medientheorien (cf. Kloock / Spahr (1997)), Darstellungen zum Buchdruck und seinen Folgen (cf. Eisenstein (1997), Giesecke (^{4}2006)) und Abhandlungen zur Grammatikographie der romanischen Sprachen und zu romanische Sprachen fokusierenden Sprachtraktaten (cf. u. a. Bierbach / Pellat (2003), Lüdtke (2001), Lubello (2003), Poggi Salani (1988), Swiggers (1990)). Die dabei zum Einsatz kommenden Technologien (Etherpad, Wiki, Datenbanken, Stylesheets bzw. Formatvorlagen) fordern sie zudem zu einem Umdenken bezüglich der Wissensproduktion im digitalen Zeitalter heraus (cf. z. B. Schöch (2017)), lange bevor sie sich – gestützt durch ein Tutorium - der Modellierung der zu digitalisierenden Texte, ihrer Transkription und Auszeichnung zuwenden. Bei letzterer spielen dann die Dokumentation des *Deutschen Textarchiv* (cf. DTA 2016), die Guidelines der *Text Encoding Initiative* (cf. TEI 2015), der *Unicode Standard* (cf. Uni 1991–2017) und die *Medieval Unicode Font Initiative* (cf. Haugen 2011) sowie der Oxygen XML-Editor eine zentrale Rolle.

Poster

Das komplexe Zusammenspiel von verschiedenen Wissensdomänen und einer Vielzahl von Technologien sowie die intendierten konzeptuellen Änderungen wollen wir in unserem Poster visualisieren. Dabei wollen wir nicht nur eine kritische Reflektion über dieses Vorgehen einbringen, sondern auch Ergebnisse, die sich in den Modulabschlussarbeiten der Studierenden zeigen, werten.

Bibliographie

Albrecht, Jörn (2001): „Sprachbewertung / Évaluation de la langue“, in: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian (eds.): *Lexikon der romanistischen Linguistik (LRL)*. Band I / 2: *Methodologie (Sprache in der Gesellschaft / Sprache und Klassifikation / Datensammlung und -verarbeitung) / Méthodologie (Langue et société / Langue et classification / Collection et traitement des données)*. Tübingen: Niemeyer 526-540.

Bierbach, Mechtild / Pellat, Jean-Christophe (2003): "Histoire de la réflexion sur les langues romanes: le français / Geschichte der Reflexion über die romanischen Sprachen: Französisch", in: Ernst, Gerhard / Gleßgen, Martin-Dietrich / Schmitt, Christian / Schweickard, Wolfgang (eds.): *Romanische Sprachgeschichte / Histoire linguistique de la Romania*. Ein internationales Handbuch zur Geschichte der romanischen Sprachen / Manuel international d'histoire linguistique de la Romania 1 (= Handbücher zur Sprach- und Kommunikationswissenschaft 23). Berlin / New York: De Gruyter 226-229.

Du Bellay, Joachim (1549): *Deffence, et illustration de la langue francoyse*. Paris: Arnoul l’Angelier (Digitalisat: PDF, BnF Gallica).

DTA (12.01.2016): „Dokumentation“, in: *Deutsches Textarchiv* <http://www.deutschestextarchiv.de/doku> [25.09.2017].

Eisenstein, Elisabeth I. (1997): *Die Druckerpresse*. Kulturrevolution im frühen modernen Europa. Wien / New York: Springer.

Fortunio, Francesco (1516): *Regole grammaticali della volgar lingua*. Ancona: Bernardin Vercellese / Bernardino Guerralda (Digitalisat: TIFF, Universitätsbibliothek Eichstätt-Ingolstadt 03.06.2015).

Giesecke, Michael (^{4}2006): *Der Buchdruck in der frühen Neuzeit*. Eine historische Fallstudie über die Durchsetzung neuer Informations- und Kommunikationstechnologien. Frankfurt am Main: Suhrkamp.

Gold, Matthew K. (ed.) (2012): *Debates in the Digital Humanities*. Minneapolis / London: University of Minnesota Press.

Haugen, Odd Einar (ed.) (2011): *Medieval Unicode Font Initative* <http://folk.uib.no/hnooh/mufi/> [25.09.2017].

Hirsch, Brett D. (ed.) (2012): *Digital Humanities Pedagogy*. Practices, Principles and Politics. Cambridge: Open Book Publishers.

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

Kloock, Daniela / Spahr, Angela (1997): *Medientheorien*. Eine Einführung (= UTB 1986). München: Wilhelm Fink.

Lubello, Sergio (2003): "Storia della riflessione sulle lingue romanze: italiano e sardo / Geschichte der Reflexion über die romanischen Sprachen: Italienisch und Sardisch", in: Ernst, Gerhard / Gleßgen, Martin-Dietrich / Schmitt, Christian / Schweickard, Wolfgang (eds.): *Romanische Sprachgeschichte / Histoire linguistique de la*

Romania. Ein internationales Handbuch zur Geschichte der romanischen Sprachen / Manuel international d'histoire linguistique de la Romania. Berlin: De Gruyter 208-225.

Lüdtke, Jens (2001): "Romanische Philologie von Dante bis Raynouard / La philologie romane de Dante à Raynouard", in: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian (eds.): *Lexikon der romanistischen Linguistik (LRL)*. Band I / 1: Geschichte des Faches Romanistik. Methodologie (Das Sprachsystem) / Histoire de la philologie romane. Méthodologie (Langue et système). Tübingen: Niemeyer 1-35.

Meigret, Louis (1550): *Le tretté de la GRAMMeRE FRANCOEZE*. Paris: Chrestien Wechel (Digitalisat: PDF, Bayrische Staatsbibliothek, Münchner Digitalisierungszentrum Digitale Bibliothek 05.09.2015, JPG 10.11.2017).

Oxygen XML Editor (2002-2017) <https://www.oxygenxml.com/> [25.09.2017].

Poggi Salani, Teresa (1988): "Italienisch: Grammatikographie / Storia delle grammatiche", in: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian (eds.): *Lexikon der Romanistischen Linguistik (LRL)* IV: Italienisch, Korsisch, Sardisch / Italiano, Corso, Sardo. Tübingen: Niemeyer 774-786.

Schöch, Christof (2017): „Digitale Wissensproduktion“, in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): *Digital Humanities*. Eine Einführung . Stuttgart: J. B. Metzler 206-212.

Schreibman, Susan / Ping Huang, Marianne / Benardou, Agiatis/ Tasovac, Toma / Scagliola, Stef / Durco, Matej / Clivaz, Claire (2015-2017): *#dariahTeach*. DH Training Materials <http://dariah.eu/teach/> [24.09.2017].

Sperone, Speron (1542): “Dialogo delle lingue”, in: Speron Sperone: *I dialogi di Messer Speron Sperone*. Vinegia: Aldus 106-131 (Digitalisat: TIFF, Universitätsbibliothek Leipzig 27.04.2017).

Swiggers, Pierre (1990): “Französisch: Grammatikographie / Grammaticographie“, in: *Lexikon der Romanistischen Linguistik* (LRL), V, 1: Französisch, Okzitanisch, Katalanisch / Le français, L'occitan, Le catalan. Tübingen: Niemeyer 843-869.

TEI (05.10.2015): “Guidelines”, in: *Text Encoding Initiative* <http://www.tei-c.org/Guidelines/> [25.09.2017].

Tory, Geoffroy (1529): *Champ Fleury*. Paris: Geoffroy Tory / Giles Gourmont.

Uni (1991–2017): “The Unicode Standard“, in: *The Unicode Consortium* <http://www.unicode.org/standard/standard.html> 25.09.2017].

Warwick, Claire / Terras, Melissa / Nyhan, Julianne (eds.) (2012): *Digital Humanities in Practice*. London: Facet Publishing.

Kleriker des Alten Reiches in der Digitalen Welt. Das Forschungsportal Germania Sacra Online

Kröger, Bärbel
bkroege@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Popp, Christian
cpopp@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

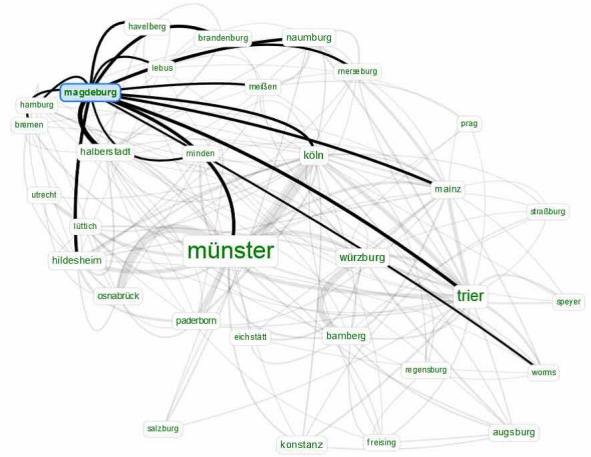
Das Poster soll die digitalen Angebote der Germania Sacra vorstellen und an einem konkreten Beispiel Möglichkeiten und Grenzen diskutieren, wie mithilfe der Netzwerkanalyse große Datenkorpora visualisiert und ausgewertet werden können, um neue Fragestellungen zu generieren.

Die Germania Sacra ist ein Forschungsprojekt an der Akademie der Wissenschaften zu Göttingen, das die Geschichte der Bistümer, Stifte und Klöster im Heiligen Römischen Reich Deutscher Nation aufarbeitet. Die geistlichen Institutionen werden von ihrer Gründung in der Spätantike bis in die Zeit der Säkularisation behandelt. Geographisch umfasst das Untersuchungsgebiet die heutige Bundesrepublik und die grenznahen Regionen ihrer Nachbarländer.

Ihre Forschungsergebnisse präsentiert die Germania Sacra in Printpublikationen und in ihrem Portal Germania Sacra Online. Die inhaltlichen Schwerpunkte des Online-Portals liegen zum einen auf der Prosopographie, zum anderen auf der überregionalen Aufarbeitung der Kloster- und Stiftslandschaft in Mittelalter und Früher Neuzeit. Für prosopographische Fragestellungen bietet das Online-Portal eine wissenschaftliche Personendatenbank, in der inzwischen mehr als 50.000 Datensätze online abrufbar sind. Das geistliche Personal der Dom- und Kollegiatstifte, das eine herausragende Rolle im mittelalterlichen Bildungs- und Universitätswesen spielt, ist ein wesentlicher Bestandteil der Datenbank und steht auch zukünftig im Zentrum der Forschungen der Germania Sacra.

Der Datenbestand wird kontinuierlich weiter anwachsen und fordert auch ein traditionelles Projekt der Grundlagenforschung heraus, Strategien und Perspektiven für die Weiterentwicklung des Online-Portals zu entwickeln. In erster Linie ist es unsere Zielsetzung, einen umfangreichen, strukturierten, soliden Datenpool zu schaffen, der aus den historischen Primärquellen erarbeitet worden ist. Darüber hinaus ist es eine unverzichtbare Aufgabe, die Daten möglichst ansprechend zu visualisieren und möglichst breit zu vernetzen und hierfür technische Lösungsansätze zu entwickeln.

Gerade für prosopographische Daten ist das Potential der historischen Netzwerkanalyse in den letzten Jahren in den Fokus gerückt. Welche Aussagen mit dieser Art von Datenvisualisierung für den Datenbestand der Germania Sacra getroffen werden können, soll an einem Beispiel veranschaulicht werden.



Grafik 1: Beziehungsgeflecht der Bistümer der Germania Sacra anhand der Ämter- und Pfründenhäufungen der Kleriker

Die Grafik zeigt das Beziehungsgeflecht der Bistümer des Alten Reiches. Datengrundlage der Grafik sind 8.000 Kleriker aus dem Bestand der Personendatenbank, die Ämter und Pfründen in unterschiedlichen Bistümern bekleidet haben. Die Visualisierung zeigt also das Beziehungsgeflecht zwischen den Diözesen, das sich aus den Ämtern des Kirchenpersonals ergibt. Hier lassen sich durch den Einsatz digitaler Methoden historische Phänomene herauslesen: So zeigt sich für das hier hervorgehobene Erzbistum Magdeburg ein dichtes Beziehungsgeflecht mit den Bistümern der eigenen Kirchenprovinz sowie mit Nachbarbistümern wie Halberstadt. Die Visualisierung macht aber zugleich sichtbar, dass Magdeburg Kontakte zur weit entfernten Trierer Kirche hatte.

Es wird in Zukunft darauf ankommen, mithilfe der Visualisierung von Forschungsdaten einen erweiterten und veränderten Blick auf historische Entwicklungen zu ermöglichen und neue, innovative Fragestellungen zu entwickeln.

Die Posterpräsentation soll Anstoß geben, alternative Strategien der Informationsauswertung durch den Einsatz digitaler Methoden vorzustellen und zu diskutieren.

Bibliographie

Serge ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck und Eveline Wandl-Vogt (Hg.): Proceedings of the First Conference on Biographical Data in a Digital World 2015, Amsterdam, The Netherlands, April 9, 2015. Aachen (CEUR workshop proceedings).

Norbert Fuhr, László Kovács, Thomas Risse und Wolfgang Nejdl (Hg.): Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings. Cham (2016).

Jörg Wettlaufer: Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern (Zeitschrift für digitale Geisteswissenschaften, 2016). Online verfügbar unter http://zfdg.de/2016_011, zuletzt geprüft am 01.09.2017.

Kollaborativ arbeiten und annotieren – Die Forschungsinfrastruktur des Spezialforschungsbereichs Deutsch in Österreich

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

Breuer, Ludwig Maximilian
ludwig.maximilian.breuer@univie.ac.at
Universität Wien

Heinisch, Barbara
barbara.heinisch@univie.ac.at
Universität Wien

Der Spezialforschungsbereich (SFB) „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (FWF F 60) beschäftigt sich mit der Vielfalt sowie dem Wandel der deutschen Sprache in Österreich. Dabei behandelt er den Gebrauch und die subjektive Wahrnehmung von deutscher Sprache in Österreich und zeigt Einflüsse durch Kontaktsprachen auf. Der SFB gliedert sich in verschiedene Teilprojekte an vier verschiedenen Institutionen (Universität Wien, Universität Salzburg, Universität Graz und Österreichische Akademie der Wissenschaften), die unterschiedliche Schwerpunkte in den Fokus nehmen.

Ein Ziel des SFB ist es, die Forschungsansätze und -ergebnisse einem möglichst breiten Publikum einfach und frei zugänglich zu machen. Hierfür ist eine gut durchdachte, funktionale und vor allem flexible Forschungsplattform sowie das Zurverfügungstellen (und die Nutzung) einiger Standards und Best Practices unumgänglich. Zum einen um die Arbeit über die verschiedenen Teilprojekte und Institutionen kollaborativ und einheitlich zu gestalten, zum anderen um den Open Science-Ansatz des Projektes verwirklichen zu können und Daten sowie Forschungsmethoden und -ergebnisse bereitstellen zu können.

Hierzu ist eine breite - auch technische - Expertise nötig, da die Forschungsplattform den gesamten Forschungsprozess unterstützt und begleitet. Die Forschungsplattform dient dabei als Virtual Research Environment (VRE) (Sarwar et al. 2013: 551, Smith et al. 2011:54). Sie bietet Tools wie eine Call-Center-Maske für die Gewährspersonenakquise und -befragung, Transkriptions-, Annotations-, Kommunikations-, Arbeitsmanagement- und Datenmanagementools – damit das ganze Spektrum des wissenschaftlichen Arbeitsprozesses von der InformantInnenakquise über die Datenauswertung bis hin zur Publikation unterstützt werden kann. Schließlich werden alle Prozesse und Daten nicht nur nachhaltig gespeichert, sondern auch weiterentwickelt und aufbereitet, damit sie WissenschafterInnen sowie der interessierten Öffentlichkeit zur Verfügung gestellt werden können. Eine Herausforderung stellen die Vielfalt der theoretischen Ansätze, praktischen Methoden, Datenformate, (u.a. Textkorpora, Audio- und Videodaten) und die (Meta-)Datenaufbereitung sowie die damit verbundene Sicherung und Nutzung dar.

Die Grundlage der einzelnen Module der Forschungsplattform basiert auf Open Source-Tools, die gegebenenfalls an die Bedürfnisse des SFB angepasst werden. Dazu gehören beispielsweise Eingabemasken für eine Personendatenbank samt „Call Center-Maske“, Tranksriptions- sowie Annotationsumgebungen, eine Literaturverwaltung sowie Analysetools. Die Tools werden dabei in Docker-Containern (Merkel 2014) gespeichert, u.a. um die Nachhaltigkeit der Tools zu unterstützen. Dies bietet den Vorteil, dass auch abseits des Projekts und sogar der Wissenschaft an den einzelnen Komponenten weiterentwickelt werden kann, insofern sich eine entsprechende Community bildet. Zudem bieten die Docker-Container die Möglichkeit, Tools sehr einfach für andere Projekte nutzen zu können. Die Docker-Container werden im SFB in Rancher (https://rancher.com/) verwaltet. Dies vereinfacht die gesamte Pipeline von der Entwicklung über das Testen bis hin zur produktiven Nutzung der Tools, da alles an einem Ort geschehen kann.

Die VRE ist zum jetzigen Zeitpunkt nur projektintern verfügbar und befindet sich in kontinuierlicher Weiterentwicklung. Eine angepasste VRE soll zum späteren Zeitpunkt auch der Öffentlichkeit zur Verfügung stehen. Andere Plattformen wie WebAnno etc. haben sich nicht als für die SFB-Bedürfnisse adaptabel erwiesen, da sie den umfassenden Aufgaben sowie der Strukturierung der Daten und Annotationen nicht gerecht werden.

Auf dem Poster soll neben der Darstellung dieser Plattform genauer auf den Bereich der Annotation eingegangen werden. Da diese auf den verschiedensten linguistischen Ebenen vorgenommen und von einer Vielzahl von WissenschafterInnen erarbeitet wird, muss es zum einen klare Vorgaben geben, wie annotiert werden soll, zum anderen müssen die Annotationen nach einem einheitlichen Schema gebildet werden. Dennoch soll die Flexibilität der Annotationen gewährleistet sein, um wissenschaftliches Arbeiten möglichst effizient gestalten zu können. Dabei werden die Tags zwar mit Hilfe einer m:n-Verknüpfung zwischen Attributen (Tags) und Werten (Antworten) gespeichert, sind aber hierarchisch projizierbar. Somit ist es möglich, dass die verschiedenen Teilprojekte unterschiedlich tief annotieren, aber dennoch auf die Annotationen der anderen Teilprojekte zurückgreifen können. Sie fügen sich zudem bestmöglich in die vorhandene Forschungslandschaft ein, z.B. durch Verwendung gängiger Taggingsysteme (wie Edisyn), um auch SFB-übergreifend leicht nutz-, adaptier- und vergleichbar zu sein.

Bibliographie

Merkel, Dirk (2014): „Docker: Lightweight Linux Containers for Consistent

Development and Deployment“, in: *Linux J.* 239 http://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment [letzter Zugriff: 25. September 2017].

Sarwar, Muhammad S. / Doherty, T. / Watt, J. / Sinnott, Richard O. (2013): „Towards a virtual research environment for language and literature researchers“, in: *Future Generation Computer Systems* 29: 549–559 https://doi.org/10.1016/j.future.2012.03.015.

Smith, Vincent / Rycroft, Simon / Brake, Irina / Scott, Ben / Baker, Ed / Livermore, Laurence / Blagoderov, Vladimir / Roberts, David (2011): „Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science“, in: *ZooKeys* 150: 53–70.

LDA Topic Modeling über ein graphisches Interface

Simmler, Severin
severin.simmler@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

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

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

LDA (Latent Dirichlet Allocation) Topic Modeling ist ein computergestütztes Verfahren zur semantischen Analyse digitaler Textsammlungen. Hierbei werden mit Hilfe eines probabilistischen Verfahrens aus Texten eine Reihe sogenannter “Topics” generiert: Gruppen semantisch ähnlicher Begriffe, die über mehrere Texte gemeinsam auftreten und im Modell als Wahrscheinlichkeitsverteilungen über die Gesamtheit des analysierten Vokabulars repräsentiert werden. Das heißt, daß zum Beispiel in einem Topic zum Thema Seefahrt nautische Begriffe besonders hohe Wahrscheinlichkeiten haben (Blei 2012, Steyvers und Griffiths 2006).

In den letzten Jahren ist das Interesse an LDA als Verfahren für die Analyse literarischer Textcorpora auf Seiten der digitalen Geisteswissenschaften stark gestiegen. Im Kontrast zu diesem gesteigerten Interesse ist die Anwendung der Methode allerdings nicht wesentlich leichter geworden. Gängige Implementierungen des LDA-Algorithmus werden entweder über ein kommandozeilenbasiertes Java-Programm (MALLET von McCallum 2002) oder über Skripte in der Programmiersprache Python (Gensim von Rehurek und Sojka 2010) angesprochen. Die Aufbereitung der Daten vor dem Topic Modeling, das sog. “Preprocessing” und die Analyse der Ergebnisse hinterher geschieht dann zumindest in Teilen häufig unter Verwendung weiterer Programme bzw. Arbeitsumgebungen. Alles in allem erfordert die Durchführung einer LDA-basierten Inhaltsanalyse damit zur Zeit relativ umfangreiche technische Kenntnisse.

Um den Zugang zu dieser Methode zu erleichtern entwickeln wir im Rahmen von DARIAH-DE (https://de.dariah.eu/) zur Zeit eine ausführlich dokumentierte Python-Programmbibliothek, die es ermöglichen soll, den gesamten Arbeitsprozess einer LDA-basierten Analyse in einer einzigen Umgebung durchzuführen (https://github.com/DARIAH-DE/Topics). Neben der Schaffung integrierter, flexibler Arbeitsabläufe, die vollständig in einer Programmiersprache und Umgebung stattfinden können, wollen wir auch Forscherinnen und Forschern ohne vorherige Programmierkenntnisse eine Möglichkeit zu bieten, Topic Modeling als Verfahren kennen zu lernen und an eigenen Daten auszuprobieren.

Um einen leichtgewichtigen Einstieg in diese Thematik zu bieten haben wir auf Basis unserer Programmbibliothek, der Python-nativen LDA-Implementierung von Allan Riddell (https://pypi.python.org/pypi/lda) und dem Python-Microframework “Flask” (http://flask.pocoo.org/) einen sogenannten GUI-Demonstrator entwickelt (Abb. 1). Dabei handelt es sich um eine browserbasierte graphische Benutzeroberfläche für die DARIAH-Topics Bibliothek, mit der sich ein basaler Topic-Modeling Analysevorgang lokal, mit eigenen Daten, aber eben ohne jegliche Programmierkenntnisse durchführen lässt.

Der GUI-Demonstrator übernimmt und erklärt hierbei exemplarisch alle Arbeitsschritte einer einfachen Analyse. Zunächst werden Textdateien über ein Auswahlmenü eingelesen und tokenisiert. Nutzerinnen und Nutzer können zur Reduktion des Vokabulars auf die Funktionswörter vorgeben, wie viele der häufigsten Wörter aus den Texten entfernt werden sollen, oder alternativ über ein weiteres Auswahlmenü eine externe Stopwortliste einbinden. Die Anzahl der zu berechnenden Topics und die Zahl der Iterationen, über die die Berechnung durchgeführt werden soll, ein Faktor, der die Qualität der Ergebnisse

entscheidend beeinflusst, können ebenfalls über das Interface gesteuert werden. In der derzeitigen Form generiert das Programm als Output eine Tabelle mit den zehn am stärksten gewichteten Wörtern in jedem Topic, sowie ein Heatmap als Übersicht über die Verteilung der Topics über die Texte.

Im Fokus der gegenwärtigen Weiterentwicklung steht die Gestaltung interaktiver Outputs mit Hilfe von Bokeh (https://bokeh.pydata.org/), die einen flexibleren Zugriff auf eine größere Zahl von Aspekten der Modellierungsergebnisse ermöglichen sollen.

Das Ziel dieser Entwicklung bleibt aber in erster Linie ein didaktisches: Der GUI-Demonstrator führt die grundsätzlichen Möglichkeiten der Methode vor und informiert gleichzeitig über die Abläufe im Hintergrund, so dass der Schritt hin zur Verwendung der gleichen Funktionalitäten in einem vorbereiteten Notebook mit interaktiven Codeblöcken, das schnell an die spezifischen Bedürfnisse eine bestimmten Forschungsfrage angepasst werden kann, nur noch klein ist.

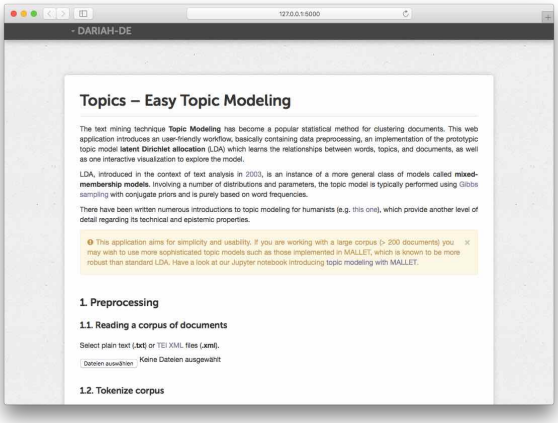


Abbildung 1.: Screenshot

Bibliographie

Blei, David M. (2012): „Probabilistic Topic Models“, in *Communication of the ACM* 55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

McCallum, Andrew K. (2002): *MALLET : A Machine Learning for Language Toolkit*. *http://mallet.cs.umass.edu* .

Rehurek, Radim/ Sojka, Petr (2010): "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Steyvers, Mark/ Griffiths, Tom (2006): „Probabilistic Topic Models“, in *Latent Semantic Analysis: A Road to Meaning*, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

MEDEA: Datenkonsistenz mittels Ontologie

Pollin, Christopher
christopher.pollin@uni-graz.at
Universität Graz, Österreich

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich

Daten werden laut der Vision der ‘High Level Expert Group on Scientific Data’ in der Zukunft einen Grad an Ausdrucksstärke und Formen der Selbstbeschreibung erhalten, dass sie in die Lage versetzt werden, ihre eigene Infrastruktur zu stellen (Neuroth, Heike / et al. 2012). Auch die Idee des Semantic Web verspricht eine Zukunft in der Maschinen selbständig mit Daten agieren können (Berners-Lee, Tim 2000). Die praktische Realität - gerade für die digitalen Geisteswissenschaften - ist noch eine Andere. Dennoch steckt in den Methoden des Semantic Webs ein Potenzial, das es mit vernünftiger Kritik zu nutzen gilt.

Das Projekt MEDEA (Modeling semantically Enriched Digital Edition of Accounts) versucht dies zu verwirklichen, indem an einem kollektiven Standard zu semantischen Anreicherung digitaler Editionen von historischen Rechnungsbüchern gearbeitet wird. Es wird der Frage nachgegangen, inwieweit Methoden des Semantic Webs bei der Erschließung, Analyse und Darstellung historischer Rechnungsbücher helfen können (MEDEA, 2017).

Ein zentrales Anliegen jedes wissenschaftlichen Projektes ist es, über qualitative und konsistente Daten zu verfügen. Dateninkonsistenz und mangelnde Datenqualität bergen die Gefahr falscher wissenschaftlicher Interpretationen und kritischer Fehlerquellen für die technische Verarbeitung der Daten. Da unterschiedliche TEI-Kodierungen von unikalen Quellen zusammengeführt werden, ist es eine besondere Herausforderung im MEDEA Projekt, Workflows zu etablieren, die dieser Herausforderung begegnen. Die Entwicklung einer domänenspezifischen Ontologie zur Formalisierung von historischen Prozessen des Rechnungswesens kann als eine solche potenzielle Lösung betrachtet werden. Die

(Bookkeeping Ontologie 2017) formalisiert in ihrem jetzigen Zeitpunkt eine grundlegende Wissensstruktur, um Einträge in Rechnungsbüchern, ihre Transaktionen von Gütern, Dienstleistungen oder Geldbeträgen von einem Akteur oder Konto zu einem anderen, standardisiert beschreiben zu können.
Aus jeder Transaktion, die mittels des Attributes @ana in einem TEI kodierten Text annotiert wurde, wird ein XML/RDF Datensatz erzeugt, der auf Konzepte der in OWL serialisierten Bookkeeping-Ontologie referenziert (Vogeler 2016). Der Ontologie Editor Protégé erlaubt es, eine Ontologie und die darin enthaltenen Daten (Individuals) einem Reasoning - dem Abarbeiten alle Vorhanden Regeln in einer Ontologie auf Basis der Description Logic - zu unterziehen (Musen 2015). Das Reasoning gilt als ein essentieller Baustein im Design, der Entwicklung, der Wartung und in der praktischen Anwendung einer Ontologie. Das Ergebnis davon sind Inferenzen. Inferenzen sind neu hergeleitete Schlussfolgerungen auf Basis des Reasoning Prozesses (Dentler / et al. 2011). Die Überprüfung strukturierter Daten mittels logischen Schlussfolgerungen kann dazu dienen, größere Datenmengen auf ihre Konsistenz und somit auch auf ihre Qualität hin zu prüfen, da logische Inkonsistenzen als Fehlermeldung angezeigt werden. Die Überprüfung und Zusammenführung der TEI-Kodierungen wird im MEDEA Projekt auf Basis dieser Ontologie durchgeführt. Use Cases für die Bookkeeping-Ontologie im Projekt umfassen:

- Formalisierung und Systematisierung von Rechnungsbüchern in einer maschinenverständlichen Wissensbasis
- Überprüfung der Datenkonsistenz und inhaltliche Zusammenführung der Daten
- Schaffung eines Definitionskonsenses
- Grundlage für semantische Retrieval und Discovery Strategien
- Wiederverwendbarkeit und Erweiterbarkeit
- Interoperables und offenes, sowie transparentes Verteilen von Daten

Dieser Zugang kann mit der Arbeit von (Steffen, Henniche / et al. 2015) verglichen werden, in der die Anwendung von Semantic Web Methoden, im Speziellen des Reasoning, auf geisteswissenschaftliche Daten angewandt wird.
So verlockend die Möglichkeiten einer Ontologie sein können, so kritisch sind diese auch zu betrachten. Ein grundlegendes Problem ist bereits durch den Widerspruch der hermeneutischen Arbeit der Historikerin und der Entscheidbarkeit von OWL gegeben. Sind Ontologien in OWL ausdrucksstark genug, um geisteswissenschaftliche Daten so beschreiben zu können, dass ein logisches Schlussfolgern Ergebnisse erzielt, das die Konsistenz und die Qualität der Daten abbildet?

Bibliographie

Berners-Lee, Tim. (2000): Weaving the Web: The Past, Present and Future of the World Wide Web by Its Inventor. London.

Bookkeeping Ontologie, http://glossa.uni-graz.at/o:medea.1951/ONTOLOGY [letzter Zugriff 21.09.2017].

Dentler, Kathrin / et al. (2011): Comparison of reasoners for large ontologies in the OWL 2 EL profile. Semantic Web 2.2, 71-87.

MEDEA, https://medea.hypotheses.org [letzter Zugriff 21.09.2017].

Musen, M.A. (2015): The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, DOI: 10.1145/2557001.25757003.

Neuroth, Heike / et al. (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Hülsbusch.

Steffen, Henniche / et al. (2015): Reasoning with Reasoning. Using Faceted Browsers to Find Meaning in Linked Data. Berlin, 1-61, https://liria-s.kuleuven.be/handle/123456789/485851

Vogeler, Georg (2016): The Content of Accounts and Registers in their Digital Edition. XML/TEI, Spreadsheets, and Semantic Web Technologies, in: SARNOWSKY, Jürgen (Hg.): Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters. Göttingen, 13-41.

Memes produzieren digitale Gefühle: Die Simpsons deuten Trump-Mania(c)

Haas, Gabriele
haasgab@fim.uni-passau.de
University of Passau, Deutschland

Koumpis, Adamantios
adamantios.koumpis@gmail.com
University of Passau, Deutschland

Handschuh, Siegfried
siegfried.handschuh@gmail.com
University of Passau, Deutschland

Einleitung

Twitter ist ein soziales Netzwerk bzw. ein Mikroblogging-Dienst, welches das Senden von Textnachrichten („Tweets“) ermöglicht. Es ist das derzeit am schnellsten wachsende soziale Netzwerk, worin die Twitter-Benutzer untereinander stark vernetzt sind und ihre eigenen Meinungen sowie Gefühle zu aktuellen Themen ausdrücken. Die textuellen und sprachlichen Inhalte der einzelnen Tweets sind nicht normativ, da die enthaltenen Informationen von umgangssprachlichen Ausdrücken, Abkürzungen, Emoticons und von Grammatikfehlern durchsetzt sind, sodass kein standardisiertes / automatisiertes Auswertungsverfahren zur Sentiment Analyse angewendet werden kann. Den Tweets können zudem Anhänge wie Bilder, Videos oder Hyperlinks beigefügt werden. Memes sind beliebte Vertreter solcher Bildanhänge in Tweets: *„I define an Internet meme as: (a) a group of digital items sharing common characteristics of content, form, and/or stance, which (b) were created with awareness of each other, and (c) were circulated, imitated, and/or transformed via the Internet by many users.” (Shifman L., 2013, S. 41)* Auf Basis dieser Definition von L. Shifman zählen somit auch Videos zur Gattung der Memes. Memes haben die Eigenschaft Bild und Text zu kombinieren, dadurch eine polysemische Nachricht zu generieren, welche die Rezipienten anspricht und Emotionen auslösen. *„Ein Internet-Meme ist die humoristische/sarkastische Reaktion der Internetgemeinde auf ein (mediales) Ereignis.“ (Marx und Weidacher, 2014, S. 143)* Zudem sind Memes in politischem Kontext aktuell wenig empirisch erforscht *(Shifman L., 2013, S. 119)*. Gerade aus diesen Gründen wurden Memes, speziell Memes in politischem Kontext, als Untersuchungsgegenstand für diese Sentiment Analyse gewählt.

Die vorliegende Analyse basiert auf zwei Memes: Das erste Meme steht im Kontext zur Wahl des US-Präsidenten inkl. des Vizepräsidenten vom 8. November 2016, das Zweite im Zusammenhang der Ersten 100 Amtstage von Präsident Donald John Trump. Parallel zur Präsidentschaftswahl verbreitete sich bereits zwei Tage nach dem Wahlstichtag ein statisches Bild der Simpsons, worin schon im Jahr 2000 der Wahlausgang mit der geografischen Karte der Wahlergebnisse prophezeit wurde *(siehe Abbildung 1)*. Als Reaktion auf die Ersten 100 Tage von D. Trump im Amt wurde am 26. April 2017 ein Video von den Simpsons *(siehe Abbildung 2)* auf YouTube veröffentlicht, welches genau diesen Zeitraum seiner Amtshandlungen parodiert. Bereits einen Tag danach setzte die Diskussion zum Video auf Twitter ein. Die vorliegende Arbeit analysiert die aus diesen beiden Memes entstandenen emotionalen Diskussionen und Reaktionen auf die beiden politischen Ereignisse via Twitter.

Abbildung 1: Simpson Prediction[1]

Abbildung 2: Donald Trump's First 100 Days In Office | Season 28 | THE SIMPSONS[2]

Forschungsstand

Dass die Meinungen anderer Mitmenschen unsere eigenen Entscheidungen beeinflussen, ist schon lange aus der Psychologie bekannt *(Friedkin, 1990)*. Auch die Sammlung und Auswertung von Meinungen wird schon lange betrieben. Mit dem Aufkommen des Web 2.0 ergeben sich viel bessere Möglichkeiten, große Datenmengen gezielt auf Meinungen hin zu analysieren. Die Sentiment Analyse zielt auf die Ergründung der Haltung, Stimmung, Meinung und die generelle Einstellung von Personen in Bezug auf ein speziell ausgewähltes Produkt, andere Personen,

Dienstleistungen oder aktuellen Themen ab. Dieses Forschungsgebiet fällt in den Bereich der *Computerlinguistik* bzw. der *linguistischen Datenverarbeitung (Natural Language Processing)*, welche Untergebiete des *Text Minings* sind. Die Sentiment Analyse ist als Klassifikationsproblem von Texten und dessen Polaritätserkennung zu verstehen. Für die Polaritätserkennung müssen Indikatoren im Text identifiziert werden, welche Rückschlüsse auf das sogenannte Sentiment zulassen. Dabei handelt es sich um sprachspezifische Ausdrücke, die aufgrund ihrer Wortbedeutung bereits positiv, neutral oder negativ vorbelegt sind. Diese Stimmungsinformation lässt sich aus sogenannten Sentimentlexika (bzw. Sentiment-Wörterbüchern) der jeweiligen Sprachen entnehmen. Hier werden stimmungstragende Ausdrücke - häufig Adjektive - als solche gekennzeichnet. Meist wird von deren kontextunabhängigen Polaritätsausprägung ausgegangen, die binär (positiv/negativ bzw. +/-) oder verhältnisskaliert (wie beispielsweise beim Wörterbuch *„SentiWordNet“*) kodiert wird. Die Vorgehensweise bei einer Sentiment Analyse kann entweder *Lexikon basiert*, anhand eines Wörterbuchs erfolgen oder *lernbasiert*, wo man sich auf die Algorithmen aus dem Fachgebiet des maschinellen Lernens stützt. Bei der Durchführung der Sentiment Analyse untergliedert man die drei Ebenen (*Level*): *Dokumenten-Ebene*, *Satz-Ebene* und die *Aspekt-Ebene*. Bei der *Dokumenten-Ebene* wird der gesamte Inhalt eines Dokuments in die Analyse mit einbezogen um eine generelle Stimmung zu deuten. Genauere Ergebnisse liefert die Betrachtung der *Satz-Ebene*, da hier die ausgedrückte Meinung für jeden einzelnen Satz berechnet wird. Die genausten Ergebnisse erhält man bei der Analyse der *Aspekt-Ebene*, da hier zu jedem Aspekt einer bestimmten Entität, die Stimmung des Meinungsvertreters zu einem bestimmten Zeitpunkt betrachtet wird.

Nach der Definition von *Liu Bing* besteht eine Meinung aus einem 5-Tupel:

opinion = (e,a,s,h,t)

Wobei *e eine Entität (Objekt)*, *a einen Aspekt (Feature)* der Entität, *s die subjektiv positive, negative oder neutrale Stimmung (Sentiment)* des Aspekts der Entität, *h den Meinungsvertreter (Opinion holder)* und *t den Zeitpunkt der Meinungsäußerung* darstellt.

Forschungsdesign und Methode

Aufbauend auf dem Forschungsstand untersucht diese Arbeit die inhaltlichen Diskussionen zu beiden eingangs erwähnten Memes auf Twitter hinsichtlich deren Emotionen und stellt diese in einem direkten Vergleich gegenüber. *„Communication messages such as tweets, emails, and digital images are by definition memes, because they are replicable transmitters of cultural meanings.” (Spitzberg B. H., 2014, S. 312)* In dieser Arbeit findet ein *Lexikon basierter Ansatz* für die Sentiment Analyse der Tweets Anwendung, da im Gegensatz zu den *lernbasierten Methoden des maschinellen Lernens*, die Lexikon basierten Verfahren für Bereiche eingesetzt werden können, für die keine Trainingsdaten existieren (Kennedy und Inkpen, 2006). Da für die vorliegende Sentiment Analyse lediglich ein relativ kleiner Untersuchungskorpus von 167 Tweets zur Verfügung steht, kommt die Anwendung der lernbasierten Methoden nicht in Frage. Des Weiteren können bei den Lexikon basierten Methoden kontextbedingte Ambivalenzen und andere sprachliche Konstrukte leichter berücksichtigt werden, da linguistische Aspekte eines Textes in Betracht gezogen werden können (Brooke et al., 2009). Dies ist vor allem zur Analyse von Mikroblogging-Einträgen geeignet, weil die Texte sehr kurz gehalten sind (Twitter erlaubt maximal 140 Zeichen pro Tweet). Grundsätzlich sind allerdings die Methoden des maschinellen Lernens bei Sentiment Analysen im Hinblick auf die Genauigkeit und Präzision der Klassifizierung meist effektiver als die Lexikon basierten Ansätze (Kennedy und Inkpen, 2006).

Es existieren zahlreiche Wörterbücher, wie beispielsweise *„MPQA Subjectivity Lexicon“*, *„Bing Liu and Minqing Hu Sentiment Lexicon“*, *„SentiWordNet“*, *„VADER Sentiment Lexicon“*, *„SenticNet“*, *„LIWC“*, *„Harvard General Inquirer“*, *„ANEW“* usw. um nur einige Beispiele zu nennen. Die meisten dieser beispielhaft genannten Wörterbücher liefern allerdings nur die tendenziellen Stimmungen *positiv*, *negativ* oder *neutral.* Das hier verwendete Wörterbuch *„SentiWordNet 3.0“* liefert vertiefend die einzelnen Gewichte der Stimmungen („Score“) zu einzelnen Wörtern.

Das Ziel dieser Sentiment Analyse ist die *Bestimmung der emotionalen Färbungen der einzelnen Tweets* und in weiterer Folge die *Ermittlung der generellen Stimmungshaltung der Diskussionen* bezüglich des Statischen im Vergleich zum bewegten Bild. Die Datenerhebung der einzelnen Tweet-Ströme (Tweets + Retweets) erfolgt über die Twitter API mit Hilfe des Web-Tools *„FollowTheHashtag“*[3]. Die Datensätze werden direkt in Microsoft Excel exportiert und dort weiterverarbeitet. Zur Grundgesamtheit gehören alle Tweets (exkl. Retweets) die im Zuge der US-Präsidentschaftswahl oder im Zuge der Ersten 100 Tage

nach Amtsantritt von D. Trump via Twitter weltweit abgesetzt wurden und beide Hashtags *„#thesimpsons“* und *„#trump“* beinhalten. Retweets werden bei der vorliegenden Sentiment Analyse nicht berücksichtigt, da diese keine neuen Aussagen, Meinungen oder Emotionen enthalten, sondern lediglich eine intendierte Wiederholung eines vorangegangenen Tweets darstellen.

Die Analyseeinheit ist der jeweils im Tweet enthaltene Text *(„Tweet Content“)*. Der textuelle Inhalt eines Tweets kann Hashtags (#), Taggings (@) oder (Medien-) Links enthalten. Alle Texte, deren Aussagen nicht in Relation mit den beiden Memes stehen, werden als Spam klassifiziert und aus dem Datensatz bereinigt. Die beiden Memes an sich werden nicht untersucht, sondern stellen nur den Auslöser der Diskussion dar.

Die Untersuchungszeiträume betragen jeweils sieben Tage ab dem Stichtag des Absetzens des ersten Tweets zu einem der beiden definierten Memes. Der Datensatz des statischen Bildes beläuft sich somit auf den Untersuchungszeitraum vom *10.11.2016* bis zum *16.11.2016* und beinhaltet *N = 94* Tweets (exkl. Retweets). Der Datensatz des Videos beläuft sich auf den Untersuchungszeitraum vom *27.04.2017* bis zum *04.05.2017* und beinhaltet *N = 73* Tweets (exkl. Retweets). Von der Kombination des Lexikon basierten Ansatzes mit lernbasierten Methoden wird aufgrund der geringen Datenmenge *(N = 167 relevante Tweets)* abgesehen.

Nach der Datenerhebung und –bereinigung folgt manuell der Prozess der

Textnormalisierung nach dem Konzept von *Tajinder Singh and Madhu Kumari*, gefolgt von der manuellen Vorverarbeitung inkl. Satztypen Erkennung der textuellen Einheiten nach dem Konzept von *Lei Zhang et al.* Die einzelnen Tweet Contents werden hinsichtlich ihrer Satz- und Wortebene unterteilt, wobei für jede textuelle Einheit die *positive*, *negative* oder *neutrale* Stimmung aus einem Wörterbuch entnommen wird. Die einzelnen Tweets werden nach den Satztypen *deklarativ*, *imperativ* und *interrogativ* kategorisiert. Interrogativsätze fließen nicht in die Auswertung ein, da dieser Satztyp keine informativen Meinungen, sondern lediglich Fragestellungen zum Thema oder zu vorausgehenden Tweets ausdrückt.

Danach folgt die Betrachtung der Wortebene, wobei nun alle Wörter mit emotionaler Stimmung aus den textuellen Einheiten extrahiert werden. Die Gewichtung der Stimmung jedes dieser Wörter wird mit Hilfe des Wörterbuchs *„SentiWordNet 3.0“* bestimmt. Für die einzelnen Abfragen aus dem Wörterbuch wird der freie Programmcode[4] von *Petter Törnberg* adaptiert. Nach der Abfrage aller Gewichte erfolgt die Auswertung der Daten. Hierzu wird der *„Score“* je Tweet (bestehend aus einzelnen bzw. mehreren Sätzen) durch die Summe der einzelnen Gewichte der Stimmungen der Wörter der textuellen Einheiten eines Tweets ermittelt.

score = score(pos(Wort)) + score(neg(Wort))

Wobei *score(pos(Wort))* die Summe der Scores aller positiven Gewichte der relevanten Wörter eines Tweets enthält.

Score(neg(Wort)) stellt analog die Summe aller negativen Gewichte dar. Durch obige Formel wird der Score des Tweets berechnet.

Punktationen, wie beispielsweise ;-), Smileys, Emoticons oder ähnliche nicht textuelle Ausdrucksformen von Stimmungen werden in der Sentiment Analyse nicht berücksichtigt.

Ergebnisse

Die Ähnlichkeiten und Unterschiede der emotionalen Diskussionen beider Memes konnten ermittelt und beschrieben werden. Dabei stellte sich insbesondere heraus, dass sich die emotionalen Richtungen der Diskussionen bezüglich des statischen Bilds im Vergleich zum Video erheblich voneinander unterscheiden. Die Auswertung der Häufigkeiten der emotionalen Färbungen der Tweets zum statischen Bild *(N = 86)* unterteilt sich in *19* positive, *51* neutrale und *16* negative Äußerungen. Die Ergebnisse für das Video *(N = 70)* beinhalten *18* positiv, *36* neutral und *16* negativ gestimmte Tweets.

Vergleicht man rein die Häufigkeiten des Auftretens der einzelnen Stimmungen, sieht man, dass die Verteilung fast ähnlich ist. Betrachtet man die Scores für das statische Bild *(score = 1,339)* bzw. das Video *(score = -.153)*, sprich die Summe aller positiven und negativen Gewichte der Wörter aller Tweets je Datensatz, so zeigt sich, dass trotz ähnlicher Häufigkeitsverteilungen die finale Stimmung für das statische *Bild in Summe positiv* gehalten ist. Beim *Video fällt die Stimmung negativ aus*. Ob bzw. in wie weit beispielsweise das Unterhaltungserlebnis oder die (ironischen) Inhalte des Videos im Gegensatz zum statischen Bild einen Einfluss auf die Emotionalität der Twitter-Nutzer beim Verfassen der einzelnen Texte der Tweets hat, wird im Rahmen dieser Forschungsarbeit nicht betrachtet.

Die Forschungsarbeit dient als Anwendungsbeispiel für eine Social Media Sentiment Analyse auf Twitter Daten und bildet einen thematisch übergreifenden Forschungsansatz für die Disziplinen der Informatik, der traditionellen Geisteswissenschaften und der Digital Humanities aufgrund der Kombination von *„User-Generated-Content“* in so-

zialen Netzwerken, über eine Programmierung bis hin zum Untersuchungsgegenstand der Memes, welcher wiederum typisch für die Geistes- und Sprachwissenschaften ist.

Schlussfolgerungen

1. Die Plattform Twitter ist ein stark genutztes soziales Netzwerk bzw. ein Mikroblogging-Dienst, wodurch dessen Nutzer ihre eigenen Meinungen sowie Gefühle zu Themen aller Art ausdrücken. Die Twitter-Benutzer kommentierten das Thema der US-Wahl und die ersten 100 Amtstage von D. Trump. Die Auslöser dieser Diskussionen stellten die beiden eingangs gezeigten Memes dar inkl. ihrer transportierten Nachrichten, welche durch das Zusammenspiel von Text und Bild vermittelt werden. Memes können durchaus Emotionen in der Internetgemeinde erzeugen - unter anderem auch in politischen Kontexten – ansonsten hätten sich diese beiden Memes nicht binnen kürzester Zeit via Twitter verbreiten können.
2. Bezüglich der inhaltlichen Ausgestaltung der einzelnen Tweet Contents kann man sagen, dass nahezu alle Tweets Abkürzungen und/ oder Emoticons beinhalteten. Die Polaritäten der Emoticons wurden allerdings nicht berücksichtigt, da hier auf keinen entsprechenden wissenschaftlichen Ansatz zurückgegriffen werden konnte.
3. Für manuell durchgeführte Sentiment Analysen auf kleinem Untersuchungskorpus eignen sich Lexikon basierte Ansätze hervorragend. Ob eine Kombination mit Methoden des maschinellen Lernens herangezogen wird, hängt vom Untersuchungsgegenstand und von der Größe des Untersuchungskorpus ab.
4. Grundsätzlich existieren zahlreiche Wörterbücher für Sentiment Analysen. Allerdings liefern die meisten Wörterbücher lediglich die tendenziellen Wortbedeutungen *positiv*, *negativ* oder *neutral*. Das Wörterbuch „*SentiWordNet*“ hingegen skaliert die stimmungstragenden Wörter auf dem Intervall [-1, 1] und verleiht den Wörtern spezifische Gewichte, wobei „-1“ negativ und „1“ positiv bedeuten. Die verhältnisskalierten Stimmungsinformationen zu den einzelnen Wörtern konnten größtenteils über dieses Wörterbuch bestimmt werden, allerdings gab es einige Wörter die selbst dieses Wörterbuch nicht beinhaltete. Das Wörterbuch „ *SentiWordNet*“ ist somit kein vollständiges Sentimentlexikon.
5. Bei der reinen Auswertung der Häufigkeiten der emotionalen Färbungen der Tweets zeigt sich, dass die Verteilung fast ident ist zwischen dem statischen Bild *(N = 86,* davon *19* positiv, *51* neutral und *16* negativ) und dem Video *(N = 70, davon 18* positiv, *36* neutral und *16* negativ). Deshalb wurden zusätzlich die verhältnisskalierten Stimmungsinformationen erhoben und einzelnen Gewichte betrachtet. Hierdurch lässt sich zeigen, dass die finale Stimmung für das statische *Bild in Summe positiv (score = 1,339)* gehalten ist. Beim *Video fällt die Stimmung negativ (score = -.153) aus*. Die Deutung/Interpretation der Ergebnisse einer Sentiment Analyse ist somit stark von der verwendeten Methode und dessen Ansatz abhängig.

Fußnoten

1. Quelle Abbildung 1: *https://img.buzzfeed.com/buzzfeed-static/static/201611/9/16/asset/buzzfeed-prod-fastlane02/sub-buzz18441-1478727536-5.png?downsize=715:*&outputformat=auto&output-quality=auto* (zuletzt aufgerufen: 30.06.2017)
2. Quelle Abbildung 2: *https://www.youtube.com/watch?v=Qo3fT0xPeHs* (zuletzt aufgerufen: 30.06.2017)
3. *http://www.followthehashtag.com/* (zuletzt aufgerufen: 26.07.2017)
4. Programmcode verfügbar unter GNU General Public License. Quelle: *https://github.com/mserrate/twitter-streamingapp/blob/master/twitter-stormtopology/src/main/java/analysis/SentiWordNet.java* (zuletzt aufgerufen: 30.06.2017)

Bibliographie

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* LREC, Vol. 10, S. 2200-2204.

Brooke, J.; Tofiloski, M.; Taboada, M.: Cross-linguistic sentiment analysis: From english to spanish. In: Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 2009, S. 50–54.

Friedkin, N.E. & Johnsen, E.C. (1990) *Social influence and opinions*. J. Math. Soc. 15. pp. 193 – 206.

Kennedy, A.; Inkpen, D.: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. In: Computational Intelligence 22 (2006), Nr. 2, S. 110–125.

Liu, B. (2010): *„Sentiment Analysis: A Multifaceted Problem“. IEEE Intelligent Systems,* S. 76-80.

Marx, K. & Weidacher, G. (2014). *Internetlinguistik. Ein Lehr- und Arbeitsbuch.* Tübingen: Narr.

Shifman, L. (2013). *Memes in Digital Culture.* MA: MIT Press.

Singh, T. und Kumari, M. (2016). *Role of Text Pre-processing in Twitter Sentiment Analysis.* Procedia Computer Science, 89, 549-554.

Spitzberg, B. H. (2014). *Toward A Model of Meme Diffusion (M3D).* Communication Theory 24. S. 311–339.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* In: Proceedings of the conference on human language technology and empirical methods in natural language processing. S. 347–354.

Zhang, L., et al. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HP Laboratories, Technical Report HPL-2011-89.

Menschen gendern? Einige Gedanken über Datenmodellierung zur Erhebung von Geschlechterverteilung anhand der TEI2016 Abstracts App

Hannesschläger, Vanessa
vanessa.hannesschlaeger@oeaw.ac.at
OEAW Österreichische Akademie der Wissenschaften, Österreich

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

Ausgangspunkt: Die TEI2016 Abstacts App

Die Abstracts zur TEI Konferenz 2016 wurden via ConfTool eingereicht und anschließend in Microsoft Word und InDesign ediert, um ein gedrucktes Book of Abstracts herzustellen. [1] Die InDesign-Datei wurde dann als PDF exportiert und online verfügbar gemacht. Diese beiden Fassungen des Book of Abstracts wurden vor der Konferenz fertiggestellt. Nach der Konferenz wurde die InDesign Datei erneut exportiert, diesmal in XML Format. Diese Datei wurde dann teils manuell, teils automatisiert in Einzeldateien (jedes Abstract als eine Datei) zerlegt, bearbeitet und mit umfangreichen Metadaten versehen, um TEI-P5-Dateien herzustellen. Diese wurden auf GitHub veröffentlicht. Die jeweiligen <teiHeader> wurden mit umfassenden Informationen ausgestattet und detailliert codiert; etwa wurden innerhalb der mit Normdaten (VIAF) angereicherten <editor>- und <author>-Tags konsequent <persName>-, <forename>- und <surname>-Tags eingesetzt. Das stellte sich für die Weiterverarbeitung, Analyse und Visualisierung der Daten später als vorteilhaft heraus.

Dieses Corpus bildete die Basis für die TEI2016 Abstracts App, in der die Abstracts in der Erstversion über ein Inhaltsverzeichnis, das nach Abstract-Titel oder nach Verfassenden sortiert werden konnte, ansteuerbar waren. Schon in dieser rudimentären Form zeigte sich das Potential der Daten und der Applikation für Auswertungen, die über Struktur und Zusammensetzung der Konferenz und ihrer Beitragenden Auskunft geben würden. Die App wurde daher sowohl im Bereich der Funktionalität als auch auf der Ebene der enthaltenen Daten substantiell weiterentwickelt.

Fragestellung und gender-theoretische Überlegungen

Wir trafen die Entscheidung, der weiteren Entwicklung des Prototyps der App eine Forschungsfrage zugrundezulegen. Wir entschieden uns für die Frage nach der Geschlechterverteilung unter den Beitragenden.

Aus Perspektive der Gendertheorie (in Butler'scher Tradition) bedarf die Entscheidung, diese Frage zu stellen, an sich eine Rechtfertigung, denn "a performative utterance (or practice) brings into being that of which it speaks." (Butler 2010, 150f) Das bedeutet, dass die Frage nach der Verteilung der Geschlechter die Unterscheidung der Geschlechter erst hervorbringt, und so die historische Situation, die unser Körper bedeutet, fortsetzt: "As an intentionally organized materiality, the body is always an embodying of possibilities both conditioned and circumscribed by historical convention. In other words, the body is a historical situation, as Beauvoir has claimed, and is a manner of doing, dramatizing, and reproducing a historical situation." (Butler 1988, 521) Berücksichtigt man außerdem die Konzepte des

Doing bzw. *Undoing Gender* (West & Zimmerman 1987; Hirschauer 1994 & 2001), so wird deutlich, dass die oben formulierte Frage deutlich präzisiert werden muss, um die zu erhebenden Daten in einer Weise strukturieren zu können, die eine zeitgenössische Auffassung von Geschlechteridentitäten wiedergibt.

Die Frage nach dem Geschlecht der Konferenzbeitragenden wurde daher wie folgt präzisiert: Da sie aus einem Interesse an der Stellung bzw. Repräsentation der Geschlechter im gegenwärtigen Wissenschaftsbetrieb heraus gestellt wurde, richtet sie sich nicht auf die biologische Beschaffenheit der Körper der Konferenzbeitragenden, sondern auf ihre gesellschaftliche (Selbst-)Wahrnehmung (also auf *gender* und nicht *sex*). Folgt man Butlers Auffassung, nach der die performative Äußerung das, wovon sie spricht, hervorbringt, wären daher eigene Angaben der Betroffenen zu ihrem Geschlecht die ideale Datenquelle für diese Fragestellung gewesen. Solche Angaben standen allerdings nicht zur Verfügung.

Datenmodellierung und -anreicherung

Nachdem eine Personenliste aus dem abstracts-Corpus extrahiert worden war, kamen wir zu dem Schluss, dass das Zuweisen von Geschlechtern an Personen unter Berücksichtigung der oben skizzierten geschlechtertheoretischen Überlegungen kein gangbarer Weg sein konnte. Um die soziodeterministische Neugier, aus der heraus die Ausgangsfrage gestellt worden war, dennoch befriedigen zu können, entwickelten wir ein gendersensibles Workaround für die Modellierung der Daten, das uns dennoch eine Erhebung der Geschlechterverteilung erlaubte: Anstatt den <persName>s ein <sex> zuzuweisen, wie es die TEI-Richtlinien vorsehen, entschieden wir uns dafür, dem jeweiligen <forename> einen @type zuzuweisen, der eine Angabe zum (sozialen) Geschlecht enthalten sollte. Diese Vorgehensweise erlaubte es, die Daten nicht den eigentlichen Personen nachzubauen und damit annahmen über deren Selbstwahrnehmung zu treffen, sondern zu erheben, welchem Geschlecht sie von einer allgemeinen Gesellschaft oder Öffentlichkeit aufgrund ihrer Namen aller Wahrscheinlichkeit nach zugeordnet würden.

Um das Vorgehen weiter zu objektivieren, entschieden wir uns gegen die Zuordnung der Geschlechter zu den Vornamen auf Basis unseres eigenen Weltwissens. Stattdessen machten wir uns auf die Suche nach Namensdatenbanken, die Angaben zum Geschlecht enthalten. Im Natural Language Toolkit (NLTK) etwa ist eine solche Liste enthalten (Kantrowitz 1997), jedoch werden keine Angaben gemacht, wie diese Liste zustande kam und auf welcher Grundlage die Geschlechter zugeordnet wurden. Wir entschieden uns daher für genderize.io, eine Datenbank, die nach eigenen Angaben auf Daten von Social-Media-Plattformen basiert, wo Personen die Informationen zu ihrem Geschlecht selbst bestimmen können. Nach dem Abgleich unserer Personenliste mit genderize.io via deren API konnten wir 102 von 124 Namen ein Geschlecht zuordnen. Da genderize.io nur zwei Geschlechter kennt, setzten wir “male”, “female” und “no-match” als @type-Werte ein. Die Namen, die nach diesem ersten Abgleich “no-match” waren, verglichen wir anschließend manuell mit der genderchecker.com-Datenbank, die ihre Geschlechtsangaben aus den "2001 and 2011 UK Census Data" bezieht, "together with multiple online sources and contributions from our 2m website visitors". Diese kommerzielle Datenbank konnte nicht für einen Gesamtvergleich genutzt werden, da nur die manuelle Verwendung kostenfrei ist. Ein interessanter Aspekt des Datenmodells von genderchecker.com ist die Berücksichtigung der Möglichkeit eines dritten Geschlechts: "If we see just one instance of a name appearing as both male and female, we categorise it as unisex." Aus gendertheoretischer Perspektive ist dieses Modell ein Schritt in die richtige Richtung, wenngleich die Benennung “unisex” diskutabel scheint. In unserer Namensliste kam kein solcher Fall vor - nach dem Abgleich der verbleibenden 22 Namen mit genderchecker.com blieben drei Namen “no-match”.

Da Daten und Code, aus denen die TEI2016 Abstracts App gebaut ist, unter einer offenen Lizenz auf GitHub frei zugänglich sind, steht es den Beitragenden frei, ihr Geschlecht und damit unsere statistische Auswertung der Geschechterverteilung zu verändern - das Ergebnis bleibt also ein vorläufiges. Im geplanten Vortrag wird es im Rahmen einer live-Demonstration der App präsentiert werden.

Fußnoten

1. Dieser Workflow ist, zugegebenermaßen, einer Konferenz im Bereich der Digital Humanities, und der TEI Konferenz im Besonderen, nicht ganz angemessen; eine Integration etwa des DH-Convalidators hätte den Prozess wesentlich gestrafft. Eingespielte Prozesse und Zeitdruck waren die Hauptgründe für diese Vorgehensweise.

Bibliographie

Peter Andorfer, Vanessa Hannesschläger (2017): TEI Abstracts 2016. A little application to publish the abstracts of the TEI conference 2016. http://tei2016app.acdh.oeaw.ac.at/

Judith Butler (1988): Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. In: Theatre Journal 40:4, 519–531.

Judith Butler (2010): Performative Agency. In: Journal of Cultural Economy 3:2, 147–161.

Genderchecker: http://genderchecker.com/

genderize.io: Determine the gender of a first name. https://genderize.io/

Vanessa Hannesschläger, Daniel Schopper (Hg.) (2017): TEI Conference and Members' Meeting 2016. Book of Abstracts [XML/TEI files]. https://github.com/acdh-oeaw/TEI2016abstracts

Stefan Hirschauer (1994): Die soziale Fortpflanzung der Zwei-Geschlechtlichkeit. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 46:4, 668–692.

Stefan Hirschauer (2001): Das Vergessen des Geschlechts: Zur Praxeologie einer Kategorie sozialer Ordnung. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie (Sonderheft 41), 208–235.

Mark Kantrowitz (1997): Name Corpus: List of Male, Female, and Pet names. CMU Artificial Intelligence Repository. http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/

Natural Language Toolkit (NLTK). http://www.nltk.org/

Susanne Oelkers (2003): Naming gender. Empirische Untersuchungen zur phonologischen Struktur von Vornamen im Deutschen. Frankfurt a. M.: Lang.

TEI Consortium (2017): TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Virtual International Authority File (VIAF). https://viaf.org/

Candance West, Don Zimmerman (1987): Doing Gender. In: Gender and Society 1:2, 125–151.

MeuchelmörderInnen, KindsmörderInnen, DiebInnen und die dazugehörigen Tatbestände: Erstellung eines Thesaurus für das österreichische Strafrecht des 18. Jahrhunderts zur Erschließung einer Flugblattsammlung

Wissik, Tanja
Tanja.Wissik@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Resch, Claudia
Claudia.Resch@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

In diesem Posterbeitrag beschreiben wir die Erstellung eines Thesaurus für das österreichische Strafrecht des 18. Jahrhunderts und gehen auf die damit verbundenen Herausforderungen ein, sowohl seitens des Themengebietes aber auch seitens der Quellen.

Als Quellen dienen kaum erforschte Flugblätter zur Bekanntmachung von Hinrichtungen im 18. Jahrhundert (vgl. Ammerer/Adomeit 2010), die derzeit mit einem modernen Methodeninventar als XML-Dokumente nach den Richtlinien der TEI annotiert und auf ihre digitale Verfügbarkeit im Netz vorbereitet werden. Die Darstellung der Sachverhalte, die zur Hinrichtung führen, sind zwar medial überformt (vgl. Peil 2002, Kosenina 2005, Wiltenburg 2009, Dainat 2009, Härter 2010), beruhen allerdings auf den Entscheidungen des damaligen Stadtgerichts.

Bei der Erschließung zu berücksichtigen ist, dass im Laufe des 18. Jahrhunderts unterschiedliche Strafrechtsbestimmungen in Kraft waren. Vor 1768 gab es in den Ländern Österreichs und Böhmens kein einheitliches Straf- und Strafpro-

zessrecht, jedoch galten die “Constitutio Criminalis Carolina Peinliche Gerichts- oder Peinliche Halsgerichtsordnung Kaiser Karls V.” und daneben unterschiedliche Landgerichtsordnungen, etwa die “Land-Gerichts-Ordnung. Deß Ertz-Hertzogthumbs Oesterreich unter der Ennß”. Erst im Jahre 1768 wurde durch die Constitutio Criminalis Theresiana, auch “Theresiana” genannt, ein einheitliches Straf- und Strafprozessgesetz eingeführt, welches aber bereits 1787 vom Allgemeines Gesetzbuch über Verbrechen und derselben Bestrafung, dem sogenannten “Strafgesetzbuch Josephs II” oder “Josephina” abgelöst wurde.

All diesen Strafrechtsbestimmungen ist gemein, dass sie unterschiedliche Tatbestände definieren, die mit der Todesstrafe belegt sind. Im Laufe der Zeit wurden aber zum Teil neue Tatbestände ergänzt und mit neuen Definitionen versehen und stattdessen andere Delikte abgeschafft. Diese Delikte werden auch in den zu erschließenden Quellen beschrieben: Wie erwähnt, handelt es sich dabei nicht eigentlich um Rechtstexte, sondern um Flugblätter, die über öffentliche Hinrichtungen im 18. Jahrhundert berichten. Die Quellen werden zwar als “Todesurteile” bezeichnet, aber da sich die Flugblätter an ein breites Publikum wenden, werden die Delikte allgemeinverständlich beschrieben, kaum aber unter Verwendung der damals zeitgenössischen Rechtsterminologie. DelinquentInnen werden als *UrphedsbrecherInnen* oder *DiebInnen* bezeichnet, ohne aber den eigentlichen Tatestand zu nennen - eine Referenz auf die jeweilige Gesetzesstelle fehlt oft gänzlich. Im gesamten Quellenmaterial konnte nur ein einziger Beleg gefunden werden, bei dem direkt auf den Gesetzestext referenziert wurde.

Ein weitere Herausforderung für die Erschließung der Quellen stellen die Schreibvarianten dar, vgl. etwa *Diebstahl* vs. *Diebstall* oder *Urfehde* vs. *Urphed*, *Vrphed* oder *Vrphedt*. Aus diesen Gründen ist die toolgestützte Analyse und die Zuordnung der Flugblätter zu den einzelnen Straftatbeständen erschwert. In ähnlich gelagerten Projekten, wie z.B. den “Proceedings of the old Bailey online”, die eine Zeitspanne von ca. 240 Jahren abdecken, wurden die Delikte für die Erschließung des Materials bewusst nicht nach den gesetzlichen Bestimmungen definiert (vgl. Hitchcock, Tim et al.), sondern allgemeine Definitionen erarbeitet. In dem vorliegenden Projekt haben wir uns für eine quellennahe Definition entschieden. Aus diesem Grund ist es für den Thesaurus auch essentiell, nicht nur die Definitionen bereitzustellen, sondern auch die Angabe der Quelle, aus der die Definition stammt, sowie die Erfassung von Varianten. Als Framework für die Erstellung und mögliche Darstellung des Thesaurus im Semantic Web wird als formale Sprache auf SKOS (Simple Knowledge Organisation System) mit der Erweiterung SKOS-XL (SKOS Simple Knowledge Organization System eXtension for Labels) zurückgegriffen.

Um eine verallgemeinernde Typologisierung aller in den Quellen vorkommenden Delikte zu vermeiden und eine quellennahe und differenzierte Zuordnung zu ermöglichen, möchten wir den von uns erstellten Thesaurus präsentieren, der auf der Web-Applikation eine von mehreren Möglichkeiten des Zugriffs darstellen wird (andere Zugriffsmöglichkeiten ergeben sich aus biographischen Angaben zu den DelinquentInnen wie Alter, Geschlecht, Familienstand, Religionszugehörigkeit, Herkunft bzw. der Volltextsuche im Text).

Anhand des Posterbeitrags wollen wir erörtern, wie die einzelnen Tatbestände aus den unterschiedlichen Rechtsnormen miteinander in Relation gesetzt werden können, wie weiters eine Verbindung zu den heutigen Tatbeständen hergestellt werden wird und worin letztlich der Mehrwert für zukünftige UserInnen innerhalb und außerhalb der Academia besteht. Schließlich sollen auch mögliche Nachnutzungsszenarien für den Thesaurus thematisiert werden.

Bibliographie

Adebayo, Kolawole John / Di Caro, Luigi / Boella Guido (2016): “Annotating Legal Documents with Ontology Concepts, Conference” in: Jusletter IT 25. Februar 2016, Proceedings IRIS 2016. https://jusletter-it.weblaw.ch/services/login.html?targetPage=http://jusletter-it.weblaw.ch/issues/2016/IRIS/annotating-legal-doc_77f5c9f8b8.html__ONCE&handle=http://jusletter-it.weblaw.ch/issues/2016/IRIS/annotating-legal-doc_77f5c9f8b8.html__ONCE [letzterZugriff 22. September 2017]

Ammerer, Gerhard / Adomeit, Friedrich (2010): „Armesünderblätter“ in: Härter, Karl / Sälter, Gerhard / Wiebel, Eva (eds.): Repräsentation von Kriminalität und öffentlicher Sicherheit. Bilder, Vorstellungen und Diskurse vom 16. bis zum 20. Jahrhundert. Frankfurt am Main: Klostermann 271-308.

Dainat, Holger (2009): „Räuber im Oktavformat: Über die printmediale Aufbereitung von Kriminalität im 18. Jahrhundert“ in: Habermas, Rebekka / Schwerhoff, Gerd (Hrsg.): Verbrechen im Blick. Perspektiven der neuzeitlichen Kriminalitätsgeschichte. Frankfurt/New York: Campus Verlag 339-366.

Dirschl, Christian (2016): „Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH“ in: Jusletter

IT 25. Februar 2016, Proceedings IRIS 2016. https://jusletter-it.weblaw.ch/services/log-in.html?targetPage=http://jusletter-it.weblaw.ch/issues/2016/IRIS/thesaurus-generation_da052418b5.html__ONCE&handle=http://jusletter-it.weblaw.ch/issues/2016/IRIS/thesaurus-generation_da052418b5.html__ONCE [letzterZugriff 22. September 2017]

Härter, Karl (2010): „Criminalbilder: Verbrechen, Justiz und Strafe in illustrierten Einblattdrucken der Frühen Neuzeit“ in: Härter, Karl / Sälter, Gerhard / Wiebel, Eva (Hrsg.): Repräsentation von Kriminalität und öffentlicher Sicherheit. Bilder, Vorstellungen und Diskurse vom 16. bis zum 20. Jahrhundert. Frankfurt am Main: Klostermann 25-88.

Hitchcock, Tim / Shoemaker, Robert / Emsley, Clive / Howard, Sharon / McLaughin, Jamie et al. (2012) (eds.) The Old Bailey Proceedings Online, 1674-1913. http:// www.oldbaileyonline.org/ [letzter Zugriff: 22. September 2017]

Košenina, Alexander (2015): „Recht – gefällig. Frühneuzeitliche Verbrechensdarstellung zwischen Dokumentation und Unterhaltung“ in: Zeitschrift für Germanistik. Neue Folge Band 15, Nummer 1 (2005) 28-47.

Peil, Dietmar (2002): „Strafe und Ritual. Zur Darstellung von Straftaten und Bestrafungen im illustrierten Flugblatt.“ in: Harms, Wolfgang / Peil, Dietmar (Hrsg.): Wahrnehmungsgeschichte und Wissenschaftsdiskurs im illustrierten Flugblatt der Frühen Neuzeit (1450-1700). Basel 265-486.

Wiltenburg, Joy (2009) „Formen des Sensationalismus in frühneuzeitlichen Kriminalberichten“ in: Habermas, Rebekka / Schwerhoff, Gerd (Hrsg.): Verbrechen im Blick. Perspektiven der neuzeitlichen Kriminalitätsgeschichte. Frankfurt/New York: Campus Verlag 323-338.

Wersig, Gernot (2016): Thesaurus-Leitfaden: Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis. Berlin / Boston: de Gruyter Saur Reprint.

W3C: SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009 https://www.w3.org/TR/2009/REC-skos-reference-20090818/) [letzter Zugriff: 22.09.2017]

Netzwerkanalytischer Blick auf die Dramen Anton Tschechows

Faynberg, Veronika
berenis0102@gmail.com
Higher School of Economics, Moskau, Russland

Fischer, Frank
frafis@gmail.com
Higher School of Economics, Moskau, Russland

Lashchuk, Svetlana
svetalashch@gmail.com
Higher School of Economics, Moskau, Russland

Orlova, Tatyana
taorkon.tootta@gmail.com
Higher School of Economics, Moskau, Russland

Palchikov, German
rebel368@gmail.com
Higher School of Economics, Moskau, Russland

Shlosman, Evgenia
zhenya96@gmail.com
Higher School of Economics, Moskau, Russland

Die literarische Netzwerkanalyse hat sich in den letzten Jahren zu einer gefragten Methode der digitalen Literaturwissenschaft entwickelt. Dabei rangiert die Größe der Arbeitskorpora im Sinne des »scalable reading« (Martin Mueller) von der Betrachtung von Einzeltexten (Schweizer/Schnegg 1998, Moretti 2011) über kleinere Korpora bis hin zur Untersuchung hunderter oder gar tausender Dramen (Fischer u. a. 2016, Trilcke u. a. 2016, Algee-Hewitt 2017). Dabei zeigt sich auch immer wieder Interesse an bestimmten, etwa autorzentrierten Subkorpora (Wade 2017).

In diesem Kontext siedelt sich auch unser Posterprojekt an, in dessen Mittelpunkt die extrahierten Netzwerkdaten zu den Stücken des russischen Dramatikers Anton Tschechow (1860–1904) stehen. Die Datengrundlage bildet das von uns aufgebaute und betriebene Russian Drama Corpus (RusDraCor), das es sich zur Aufgabe gestellt hat, russischsprachige Stücke in der Zeitspanne zwischen den 1740er-Jahren (Sumarokow, Lomonossow u. a.) und den 1930er-Jahren (mit Texten von Autoren wie Majakowski oder Gorki) im

TEI-Format zur Verfügung zu stellen (Fischer u. a. 2017). Neben Large-Scale-Analysen zur strukturellen Evolution des russischen Dramas ergibt sich so auch die Möglichkeit zur Betrachtung von nach verschiedenen Kriterien portionierten Teilkorpora, etwa der Stücke einzelner Autoren.

Anton Tschechow gehört zu den meistgespielten russischen Dramatikern, dessen Werke bis heute inszeniert werden, gerade auch an deutschsprachigen Bühnen, vor allem seine vier letzten Stücke, »Die Möwe«, »Onkel Wanja«, »Drei Schwestern« und »Der Kirschgarten«. Von der Figurenkonstellation her haben diese Werke einen hohen Wiedererkennungswert: Es gibt keine wirklichen Protagonisten; die Redeanteile und Gesprächssituationen sind relativ gleichmäßig über eine Gruppe von Figuren verteilt. Dies zeigt sich sofort auch in den Netzwerkgraphen: Die Knoten (von denen jeder für eine Figur des jeweiligen Dramas steht) bilden einen*character space* , der bei der Visualisierung einem Polyeder gleicht. Die einzigen Figuren, die nicht am gemeinsamen Gesprächskreis teilhaben, sind die Diener und sonstige Gehilfen, deren Redeanteile sich auf Dialoge mit ihren direkten Weisungsbefugten beschränken. Dieses Sichtbarwerden der sozialen Zweiteilung des Dramenpersonals ist eine der Leistungen der Netzwerkvisualisierung. Zieht man die Werkchronologie als Größe hinzu, wird außerdem deutlich, wie sich die für Tschechow typischen Personenkonstellationen allmählich herausbilden, ab den frühen Stücken »An der Landstraße« (1884), »Iwanow« (1887) und »Der Waldteufel« (1889), über mehrere Kurzdramen oder Etüden wie »Der Bär« (1888), »Tragödie wider Willen« (1889) oder »Das Jubiläum« (1891), bis 1895 mit der »Möwe« die typische Tschechow'sche Charakterkonstellation gefunden ist.

Die Beschaffenheit des Russian Drama Corpus erlaubt es, quantitative Analysen auch zugeschnitten auf bestimmte Figurengruppen zu beschränken, etwa gesondert nach Geschlecht oder sozialem Status. Bereits eine simple Worthäufigkeitsanalyse kann so etwa zeigen, dass weibliche und männliche Rollen in Tschechow-Stücken von den Redeanteilen und dem Vernetzungsgrad her vergleichbar sind (anders als etwa bei allen anderen Autoren im Korpus). Diese Verteilungsdiagramme sowie netzwerktheoretische Werte wie Dichte, Diameter, Clustering-Koeffizient und Average Path Length ergänzen die chronologisch sortierten Netzwerkvisualisierungen.

Die im Poster geschaffene Übersicht über alle Tschechow-Dramen hat auch enzyklopädischen Charakter, enthält sie doch etwa alle Figuren im Kontext ihres Auftretens im Tschechow'schen Dramenkosmos. Der netzwerkanalytische Blick ist somit durchaus geeignet, als Brücke zur inhaltlichen Auseinandersetzung mit den Werken Tschechows zu dienen.

Bibliographie

Algee-Hewitt, Mark (2017): Distributed Character: Quantitative Models of the English Stage, 1500–1920. DH2017, Montréal. URL: < https://dh2017.adho.org/abstracts/103/103.pdf >.

Fischer, Frank; Göbel, Mathias; Kampkaspar, Dario; Kittel, Christopher; Meiners, Hanna-Lena; Trilcke, Peer; Vogel, Andreas (2016): Distant-Reading Showcase. 200 Years of Literary Network Data at a Glance. DHd2016, Leipzig. DOI: < https://dx.doi.org/10.6084/m9.figshare.3101203.v1 >.

Fischer, Frank; Orlova, Tatyana; Skorinkin, Danil; Palchikov, German; Tyshkevich, Natasha (2017): Introducing RusDraCor – A TEI-Encoded Russian Drama Corpus for the Digital Literary Studies. CORPORA2017, St. Petersburg. Abstractband, S. 28–31.

Schweizer, Thomas; Schnegg, Michael (1998): Die soziale Struktur der »Simple Storys«. Eine Netzwerkanalyse. URL: < https://www.ethnologie.uni-hamburg.de/pdfs-de/michael-schnegg/simple-stories-publikation-michael-schnegg.pdf >.

Trilcke, Peer; Fischer, Frank; Göbel, Mathias; Kampkaspar, Dario; Kittel, Christopher (2016): Dramen als ›small worlds‹? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730–1930. DHd2016, Leipzig. URL: < http://www.dhd2016.de/abstracts/vorträge-060.html >.

Wade, Karen (2017): Jane Austen's Social Networks. In: The Sea of Books, 4. Juli 2017. URL: < https://theseaofbooks.com/2017/07/04/jane-austens-social-networks/ >.

NLP meets RegNLP meets Regesta Imperii

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

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

Dieser Posterbeitrag veranschaulicht die Interaktion zwischen computerlinguistischen Methoden und Regestenforschung. Es wird eine Anwendung vorgestellt, die bereits in einem graphbasierten Format vorliegendene Regesten webbasiert anzeigt und es erlaubt, Registereinträge im Text zu verorten. Die daraus entstandene Datenbasis hilft dabei neues Wissen zu generieren, so können z.B. Verwandtschaftsbeziehungen automatisch erkannt und in den Regesten-Graph integriert werden.

Das im Rahmen des Bund-Länder-geförderten Akademienprogramms angesiedelte Grundlagenforschungsprojekt Regesta Imperii erstellt deutschsprachige Inhaltsangaben (sog. Regesten) von Kaiser-, Königs und Papsturkunden, begonnen von Karl dem Großen bis hin zu Kaiser Maximilian. Seit Projektbeginn 2001 wurden 1829 Regesten erstellt und digitalisiert und stehen inzwischen als Volltext im Internet zur Verfügung. Die Publikation der digitalisierten Register befindet sich gerade in Vorbereitung.

Neben den Regesta Imperii sind immer mehr Editionen und Regestenwerke als Volltext im Internet zugänglich und können über Suchmasken abgefragt und genutzt werden. Die Nutzungsart unterscheidet sich zumeist aber nicht grundlegend von einer analogen Nutzung des Buches: Das Register wird aufgeschlagen und man kann anschließend die einem Registereintrag zugeordneten Urkunden oder Regesten aufrufen und lesen.

Mit der Nutzung von Graphentechnologien in den digitalen Geisteswissenschaften werden neue Nutzungs- und Analyseformen der bereits vorhanden digitalen Editions- und Regestenwerke möglich. Die Digitale Akademie, Mainz (*www.digitale-akademie.de*) hat auf ihrer Seite *www.graphentechnologien.de* einige beispielhafte Anwendungsszenarien für die Nutzung von Graphdatenbanken zur Erschließung von Onlineregesten vorgestellt. Für dieses Beispielprojekt wurden die Regesten Kaiser Friedrichs III. in eine Graphdatenbank konvertiert, anschließend das zugehörige Register digitalisiert und in die Graphdatenbank integriert. Im Graphenmodell ist es über die Abfrage nun möglich herauszufinden, in welchem Regest eine Person genannt wird und eine Analyse der gemeinsam mit ihr im Regest genannten Personen zu veranlassen (Abbildung 1). Das Graphenmodell erlaubt zudem die weitere Ergänzung von Kanten zwischen den Registerknoten. So ist es beispielsweise möglich, dass zwei Personenknoten, die Vater und Sohn darstellen, durch eine KIND-Kante ergänzt werden, um so deren Verwandtschaftsbeziehung explizit im Graphen zu repräsentieren. Mit solchen Zusatzinformationen kann das Register als Erschließungswerkzeug immer weiter wachsen.

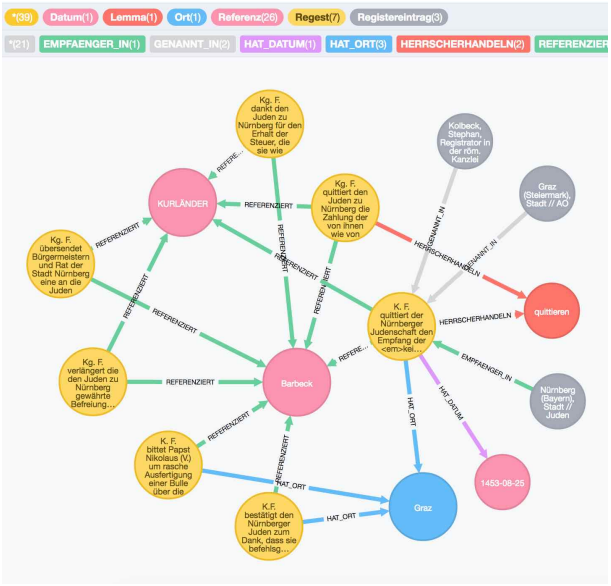


Abbildung 1 Graphbasierte Repräsentation von Regesten (gelb) mit zugehörigen Registereinträgen (rot,blau,grau).

Methoden aus der Computerlinguistik helfen diesen manuell sehr aufwendigen Grapherweiterungsschritt semi-automatisch durchzuführen. Dazu werden im ersten Schritt alle Regesten mit einer auf der Clarin-D Infrastruktur basierenden Sprachverarbeitungs-Pipeline (Malow 2012) maschinell verarbeitet: Auflösung von Abkürzungen, Tokenisierung, Part-of-Speech Tagging, Parsing und Entitätenerkennung. Im zweiten Schritt werden die Texte der einzelnen Regesten in einer interaktiven Webanwendung ausgewertet. Für jedes Regest werden alle Registereinträge (Personen, Orte, usw.), die während der Digitalisierung manuell mit Regesten verbunden wurden, angezeigt. Abbildung 2 stellt rechts den Regestentext und links alle verbundenen Registereinträge dar. In den Ausgangsdaten ist nicht gespeichert wie die Regeistereinträge im Text erwähnt werden, z.B. dass sich „der Stadt“ im Text auf den Eintrag „Aachen“ bezieht. Unsere Anwendung ermöglicht es hingegen jede im Text erkannte Entität mit den Registereinträgen per einfachem Mausklick zu verbinden. Mittels dieses Verfahrens konnten bereits ca. 4000 Registereinträge mit passenden Textstellen verbunden werden.

Abbildung 2: links: Regesttext, erkannte Entitäten sind hervorgehoben (bereits neu verortete Textstellen zum Eintrag sind grün); rechts: verknüpfte Registereinträge zum Regest

Die neu geschaffene annotierte Datenmenge bildet somit einerseits eine wichtige Grundlage, für die Regestenforschung, da nun sehr einfach Wissen aus Texten mit den verknüpften Registereinträgen automatisch abgeleitet werden kann: z.B. die Vater-Sohn-Relation zwischen Colyn Beisse und Johann Beissel (Abbildung 3).

Andererseits bietet dieser Datensatz für die computerlinguistische Forschung weitere Anknüpfungspunkte. Mittels Distant Supervision (Blessing 2012) können aus den verknüpften und im Text verankerten Relationen zu den Registereinträgen neue Modelle trainiert werden, die wiederum auf nicht manuell annotierten Textstellen der Regesten Anwendung finden. Durch diesen iterativen Ansatz können sukzessive große Regestensammlungen mit immer neuem Wissen angereichert werden und somit eine Grundlage für neue Analyseformen bieten.

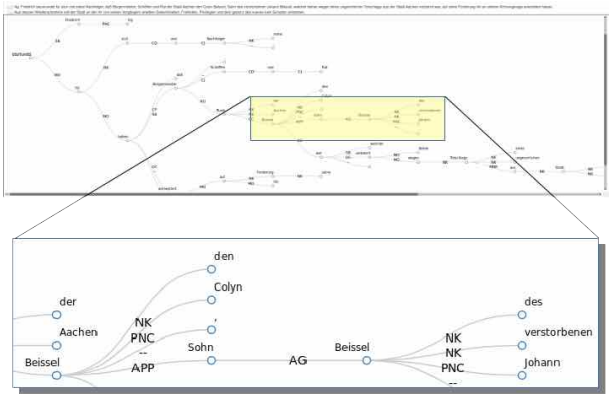


Abbildung 3: Depedenzanlayse, die zeigt wie Verwandtschaftsrelationen direkt aus der Analyse extrahiert werden können. In diesem Beispiel ist Colyn Beissel der Sohn von Johann Beissel.

Bibliographie

Blessing, Andre / Schütze, Hinrich (2012) Crosslingual Distant Supervision for Extracting Relations of Different Complexity. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*

Kuczera, Andreas (2017) Herrscherhandeln in den Regesta Imperii. Beispielprojekt an den Regesten Kaiser Friedrichs III. URL: http://www.digitale-akademie.de/forschung/graphentechnologien/beispielprojekte/ (abgerufen am 14.09.2017)

Kuczera, Andreas (2016): Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi, in: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte, 05.05.2015. URL: *http://mittelalter.hypotheses.org/5995* .

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

Nutzertests an kritischen Editionen - Print oder digital?

Caria, Federico
federico.caria@live.it
Universität zu Köln, Deutschland

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Deutschland

Zusammenfassung

In diesem Papier beschreiben wir zwei Nutzertests, die wir mit Digitalen Editionen durchgeführt haben. Im ersten Nutzertest vergleichen wir drei Digital Editionen miteinander. Dabei wird klar, dass digitale Editionen nicht immer einfach zu bedienen sind und viele der Nutzer Schwierigkeiten haben sich zurecht zu finden. Erstaunlich ist dabei, dass Nutzer Bedienbarkeit bei ihrer Wahl für relevanter halten als ein Mehr an Daten und Informationen. Im zweiten Nutzertest vergleichen kritische Editionen desselben Werks in drei verschiedenen Medien: als Buch, als PDF/eBook und als digital entstandene Online-Edition. Auch hier können wir feststellen, dass Bedienbarkeit ein relevanter Faktor bei der Wahl der Medien ist, den ein Mehr an Informationen nicht auszugleichen vermag. Das eBook gilt dabei als dasjenige, welchs am einfachsten zu bedienen ist. Es bietet die meisten Zusatzfeatures, die von den Nutzern am Häufigsten benutzt werden. Das Buch wird hingegen gerne für längere Lesepassagen genutzt. Die digitale Edition wurde allerdings von

den Nutzern größtenteils nicht richtig verstanden.

Im Folgenden werden wir zunächst die Hintergründe beleuchten, die uns dazu ermutigt haben diese Untersuchung durchzuführen. Dann werden wir dezidierter auf die Methode und die Ergebnisse eingehen und dabei insbesondere die Features behandeln, die von den Nutzern als besonders wichtig eingeschätzt wurden.

Hintergrund

In mehreren Umfragen konnte (Porter 2013, 2016) feststellen, dass zwar die Benutzung von digitalen Werkzeugen in den Geisteswissenschaften stark gestiegen ist, die Benutzung von digitalen Editionen allerdings über die Jahre konstant niedrig geblieben, oder sogar gesunken ist. Dies ist ein erstaunliches Ergebnis, wenn man bedenkt, dass immer mehr digitale Editionen verfügbar sind und diese zum Teil erhebliche Mehrwerte gegenüber den aus Printeditionen generierten PDF Versionen generieren.

Unsere Hypothese ist daher, dass digitale Editionen schwieriger zu bedienen sind und weniger auf die Bedürfnisse der Endnutzer eingehen als dies bei Printeditionen der Fall ist.

Methodik

Wir haben beide Nutzerstudien mit einer ähnlichen Methodik durchgeführt. Die Nutzergruppe bestand in beiden Fällen aus fortgeschrittenen Studenten der Geisteswissenschaften, vornehmlich aus der Philosophie, Philologie und weiteren verwandten Gebieten. Auf diese Weise haben wir sichergestellt, dass alle Teilnehmer Vorerfahrung in der wissenschaftlichen Textarbeit hatten. Keiner der Teilnehmer war ausgewiesener Experte auf dem Gebiet der Editionen, die wir betrachtet haben. Dies stellt allerdings keinen systematischen Nachteil dar. Schließlich sind die behandelten Editionen nicht nur für Experten gedacht. Tatsächlich erleichtert diese Gegebenheit sogar es den Vergleich, da alle Teilnehmer auf demselben Stand sind.

Die Teilnehmer werden für die Nutzerstudie in kleinen Gruppen in einen Raum gebeten und füllen dort zunächst einen Fragebogen zu ihren Vorerfahrungen und Präferenzen aus. Dann bekommen sie eine Liste von Aufgaben, die sie mit der vorgegebenen Edition lösen sollen. Ihre Vorgehensweise wird dabei mit Hilfe einer Screencapture Software aufgezeichnet. Nach Ablauf der Bearbeitungszeit bitten wir sie einen weiteren Fragebogen auszufüllen, in dem wir sie um pauschale Urteile zu der Edition bitten. Im Anschluss setzen sich die Teilnehmer zu einer Gruppendiskussion zusammen. Diese wird von uns an Hand eines Leitfadens geleitet und aufgezeichnet. Die Aufzeichnung wird anschließend transkribiert, kodiert und ausgewertet.

Verwandte Literatur

Nutzerstudien gelten als Standardwerkzeug um die Benutzbarkeit zu evaluieren. Sie werden in den Digitalen Geisteswissenschaften häufig eingesetzt, beispielsweise um digitale Ressourcen (Warwick 2006) oder digitale Werkzeuge für Historiker (Rücker et.al 2011) zu evaluieren. Nutzertests an digitalen Editionen (Kelly 2015, Santos 2015, Visconti 2010) konzentrieren sich typischerweise auf eine Edition und werten diese quantitativ aus. Bei der Entwicklung unserer Methodik haben wir uns vor Allem an der Untersuchung von Informationsbedürfnissen und Informationsverhalten orientiert (Barrett 2005, Belkin 1993, Buchanan 2005, Ellis 2003). Das Design der Aufgaben erfolgte in Anlehnung an (Unsworth 2000) Konzept der Primitive wissenschaftlicher Arbeit (Palmer et al 2009). Der Ansatz selbst basiert auf den Arbeiten von (Drucker 2011), allerdings erweitert, um digitale Editionen als Wissenswerkzeuge besser würdigen zu können (Bevan 1995, Porter 2016). Die transversale Analyse von (Rimmer 2008) zur Qualität von Forschung zwischen Print und Digital war dabei sehr hilfreich als Vergleich.

Editionen im Vergleich

In unserer ersten Nutzerstudie haben wir drei digitale Editionen miteinander verglichen. Die Studie selbst ist ausführlich in (Caria/Mathiak 2017) beschrieben. Daher folgt hier nur eine Zusammenfassung der Ergebnisse, um die zweite Nutzerstudie besser im Kontext verstehen zu können.

Die drei Editionen, die wir miteinander verglichen haben, hatten wir vor der Studie aus etwa 40 möglichen Editionen ausgewählt. Dabei lag unsere Wahl bei den Editionen, von denen wir dachten, dass diese am besten zu bedienen sind. Trotzdem hatten viele der Teilnehmer Schwierigkeiten die Aufgaben korrekt zu bearbeiten, da sie zum Teil die dazu notwendigen, aber vorhandenen Funktionalitäten auf der Webseite nicht gefunden haben. Dementsprechend negativ waren die Urteile der Probanden. Auffällig war, dass es jedoch einen Widerspruch zwischen Wünschen und Präferenzen der Teilnehmer gab. Gewünscht wurde

vor allem mehr Inhalt. Präferiert wurde jedoch die einzige Edition, die keine Faksimiles hatte, obwohl dies oft gewünscht wurde, weil sie die höchste Benutzbarkeit aufwies. Dies steht im Kontrast zu der allgemeinen Erfahrung, dass Wissenschaftler Inhalt vor Form bevorzugen (Kern/Mathiak 2015).

Wir schließen daraus, dass es so eine Art unsichtbare Schwelle der Bedienbarkeit gibt, die erreicht werden muss, damit der Inhalt einer digitale Edition überhaupt zum Tragen kommen kann.

Eine Frage des Mediums

Die Frage, die sich daran unmittelbar anschließt ist, ob dies eine Erklärung für die Ergebnisse von (Porter 2016) ist. Mit anderen Worten: ob die digitalen Editionen deshalb nicht benutzt werden, weil ihnen die Bedienbarkeit fehlt, die das entsprechende Buch bzw. die digitalisierte Version eines Buches von Natur aus mitbringen.

Wir haben dazu das Werk “Also sprach Zarathustra” von Friedrich Nietzsche ausgewählt. Als Print und PDF wurde die Ausgabe von (Nietzsche/Colli/Montinari 1997) repräsentiert. Die Onlineversion stammt von nietzschesource.org.

Für den Eingangsfragebogen haben wir Fragen aus dem Fragebogen von Porter übersetzt. Die Antworten sind mit den Ergebnissen von Porter vergleichbar. In den Abbildungen 1 und 2 sind die am häufigsten genutzten digitalen und analogen Ressourcen unserer Nutzer angegeben. Abbildung 3 zeigt den präferierten Zugriffsweg zu verschiedenen Arten von Ressourcen.

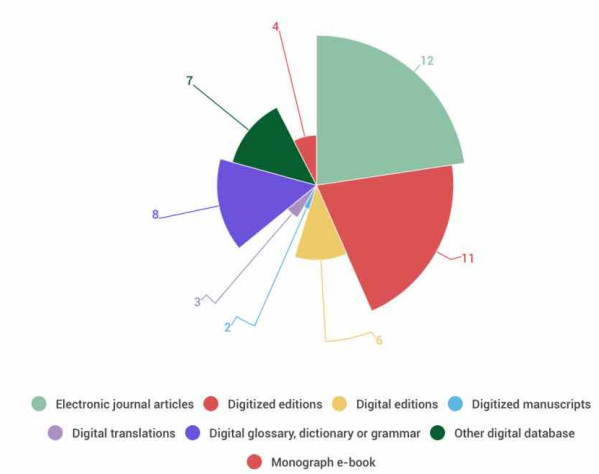


Abbildung 1: Die am häufigsten benutzten digitalen Ressourcen der letzten drei Monate.

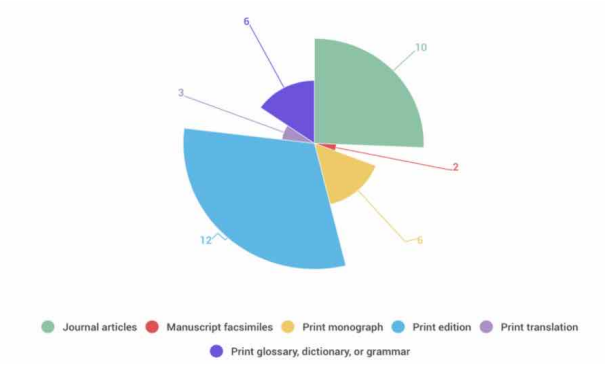


Abbildung 2: Die am häufigsten benutzten analogen Ressourcen der letzten drei Monate.

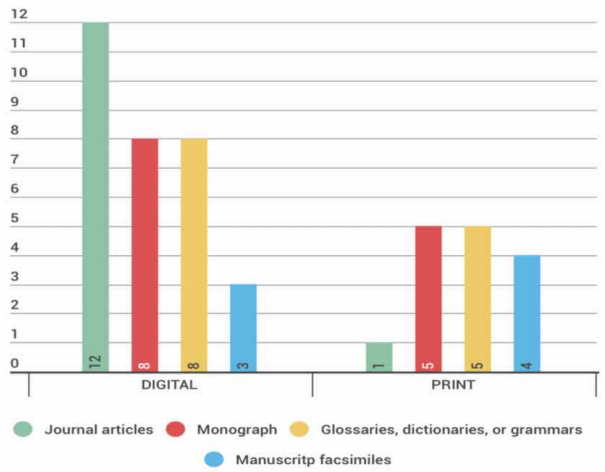


Abbildung 3: Wie wurde auf diese Ressourcen vorwiegend zugegriffen (Print oder Digital)?

Für den Nutzertest haben wir Aufgaben entwickelt, die den Nutzer auf eine natürliche Art an die Medien heranführen. Die gestellten Aufgaben waren vor allem interpretativ und verlangten von den Nutzern sich in den verschiedenen Primär- und Sekundärquellen zu orientieren und Änderungen mittels der Kommentare nachzuvollziehen. Im Anschluss an die Tests haben wir die klassischen Usability Metriken zu Effectiveness (Können die Ziele erreicht werden?), Efficiency (Wie viel Aufwand ist es die Ziele zu erreichen?) und Satisfaction (Wie zufrieden sind sie mit der Benutzbarkeit?) erhoben (vgl. Abb. 4).

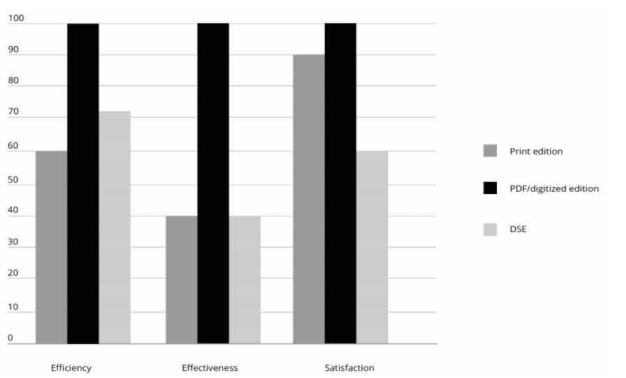


Abbildung 4: Usability Metriken der drei Medien im Vergleich.

In der abschließenden Gruppendiskussion wurden schließlich die Vor- und Nachteile der verschiedenen Medien von den Nutzern diskutiert. Neben der Frage nach dem Medium, welches sie bevorzugen, waren auch die Kernfunktionalitäten der Editionen, also die Suche, Annotationen, Kompatibilität und Portabilität ein Thema.

Die Nutzer waren sich einig, dass sie am Liebsten mit der PDF-Version arbeiten. Genannte Gründe waren: “die Suchfunktion”, “klare Navigationsstruktur”, “die Option Anmerkungen zu machen” und “bereits vertraut mit dem Interface”.

Nietzsche.org wurde dabei durchaus aufgrund der Inhalte gewürdigt: „Die [Digitale Edition] sollte eigentlich die beste Wahl sein, dadurch, dass sie viele verschiedene Texte und Daten beinhaltet, allerdings habe ich lieber 5 Tabs bzw. PDF's offen. Somit kann ich fliegend hin und her wechseln, ohne dass ich wieder und wieder nach Texten in der digitalen Edition suchen muss.“

Am Buch wurde hingegen vor Allem die Bequemlichkeit geschätzt. Es sei “besser für die Augen” und “lenkt nicht so ab”.

Gewünschte Funktionalität

Aus den Diskussionen konnten wir relativ klare und konkrete Kritikpunkte an der digitalen Edition identifizieren, die so oder so ähnlich auf die meisten der von uns untersuchten digitalen Editionen zutreffen (vgl. Abb. 5). Besonders häufig kam die Kritik an der Struktur bzw. Navigation innerhalb der Edition auf. Das Inhaltsverzeichnis gibt zwar eine relativ einfache Struktur vor, diese wurde allerdings von den Nutzern besser verstanden als die Struktur der Webseite. Ebenfalls sehr viele Nutzer arbeiten mit Annotationen. Dass dies auf der Webseite nicht unterstützt wurde, fiel unangenehm auf.

Weitere häufige Beschwerden wurden über die Suchfunktion geäußert. Diese funktionierte nicht wie von den Nutzern erwartet. Hinzu kam die Komplexität der Webseite, welche durch die Suchfunktion und Navigation nicht ausreichend kompensiert wird.

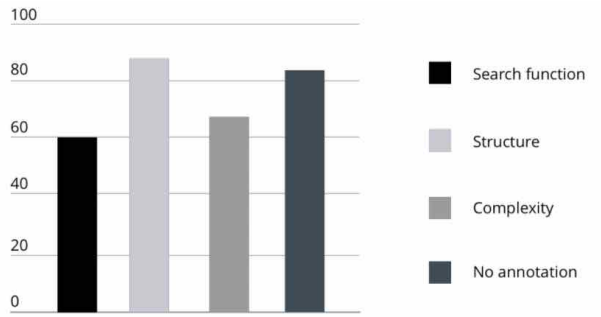


Abbildung 5: Häufig geäußerte Kritik an der digitalen Edition.

Fazit

Kritische Editionen sind nicht nur Datenquellen, sondern dienen auch als Arbeitswerkzeuge. Dabei darf die Usability nicht als nettes Feature abgetan werden, denn sie sorgt dafür, dass Nutzer eher eine digitale Edition benutzen. Das Interface sollte daher nicht nur als letzter Schritt eines editorialen Prozesses gedacht werden, sondern von Anfang an berücksichtigt und systematisch erprobt werden.

Als Kernfunktionalität sollten dabei mindestens eine Suchfunktion, eine Möglichkeit zum Kommentieren/Annotieren und zusätzlich noch eine Vergleichsfunktion in Betracht gezogen werden. Texte zu vergleichen gehört zu den häufigsten Tätigkeiten und aus unserer ersten Studie wissen wir, dass die Nutzer technische Unterstützung dazu sehr schätzen.

Bibliographie

Barrett, A. (2005): “The Information-Seeking Habits of Graduate Student Researchers in the Humanities” in: *The Journal of Academic Librarianship* 31(4), 324–331.

Belkin, N. (1993): “Interaction with Texts: Information Retrieval as Information Seeking Behavior” in: Knorz, G., Krause, J., Womser-Hacker, C. (Eds.) Information Retrieval '93: Von der Modellierung zur Anwendung, Regensburg, vol. 12, pp. 55-66 . Schriften zur Informationswissenschaft.

Bevan, N. (1995): “Measuring Usability as Quality of Use” in: *Software Quality Journal* 4: 115.

Buchanan, G./Cunningham, S.J./ Blandford, A./ Rimmer, J. / Warwick, C. (2005): "Information seeking by humanities scholars" in: *International Conference on Theory and Practice of Digital Libraries* (pp. 218-229). Springer, Berlin, Heidelberg.

Caria, F./ Mathiak. B. (2017): "Hybrid Focus Group for the Evaluation of Digital Scholarly Editions of Literary Authors", IDE. *In Druck*

Chu, C.(1999): "Literary Critics at Work and their Information Needs: A Research-Phases Model" in: *Library and Information Science Research*, 21(2): 247-73.

Cooper, A./ Reimann, R. / Cronin, D. (2007): "About face 3: the essentials of interaction design". John Wiley & Sons.

D'Iorio, P.(2015): "On the Scholarly Use of the Internet. A Conceptual Model" in A. Bozzi (éd.), *Digital Texts, Translations, Lexicons in a Multi-Modular Web Application : the method and samples*, Firenze, Olschki, 1-25.

Drucker, J. (2011): "Humanities Approaches to Interface Theory" in: *Culture Machine*, vol. pp. 1-20.

Duff, W. M., Johnson, C. (2002): "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives" in: *The Library Quarterly*, Vol. 72, No. 4, pp. 472-496.

Ellis, D. (2003): "A Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences." Journal of Documentation, vol. 49, no. 4, 356 - 369.

Gibbs, F./ Owens, T. (2012): "Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs". *Digital Humanities Quarterly*, 6:2.

Leblanc, E. (2016): "Thinking About Users and Their Interfaces: The Case of Fonte Gaia Bib" in: Book of Abstracts for *Digital Scholarly Editions as Interfaces*. International Symposium. Austrian Centre for Digital Humanities at the University of Graz 49-50

Kelly, A. (2015): "Tablet computers for the dissemination of digital scholarly editions" in: *Manuscritica*, no. 28, pp. 123-140.

Kern, D. / Mathiak, B., (2015): "Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?" in: *International Conference on Theory and Practice of Digital Libraries* 197-208. Springer, Cham.

McCarthy, J./ Wright, P. (2007): "Technology as Experience" MIT Press.

Nielsen, J. (1994): "Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier", 1994. http://www.nngroup.com/articles/guerrilla-hci/. [letzter Zugriff 23. September 2017]

Mueller, P., Oppenheimer, D. (2014): "The pen is mightier than the keyboard" in *Psychological Science* 6(25) 1159-1168.

Nietzsche, F.W. / Colli, G. / Montinari, M. (1997): "Werke: Kritische Gesamtausgabe (Vol. 5, No. 1)" de Gruyter.

Palmer, C.L./ Teffeau, L.C. / Pirmann, C.M. (2009): "Scholarly information practices in the online environment". Report commissioned by OCLC Research. Published online at: www. oclc.org/programs/publications/reports/2009-02.pdf. [letzter Zugriff 23. September 2017]

Porter, Dorothy (2013): "Medievalists and the Scholarly Digital Edition." in: *Scholarly Editing* 34, 1-26.

Porter, Dorothy (2016): "What is an Edition anyway? A critical examination of Digital Editions since 2002". Keynote lecture. *Digital Scholarly Editions as Interfaces.* International Symposium. Austrian Centre for Digital Humanities at the University of Graz. https://static.uni-graz.at/fileadmin/gewi-zentren/Informationsmodellierung/PDF/ Porter_what-is-an-edition-anyway-min.pdf [letzter Zugriff 23. September 2017]

Rimmer, J. (2008): "An examination of the physical and digital qualities of humanities research." in: *Information Processing and Management*, vol. 44, no. 3, pp. 1374-1392.

Ruecker, S./ Radzikowska, M./ Sinclair, S. (2011): "Visual Interface Design for Digital Cultural Heritage. A Guide to Rich Prospect Browsing", Routledge : New York, 2011.

Sahle, P. (2013): "Digitale Editionsformen. Zum Umgang mit der Uberlieferung unter den Bedingungen des Medienwandels. Teil 1: Textbegriffe und Recodierung". Dissertation, Universitaẗ zu Koln, 2013.

Santos, T. (2015): "LdoD Archive: User Experience and Identity Design Processes" in: *Digital Literary Studies*, Coimbra. https://www.researchgate.net/publication/280621302_LdoD_Archive_User_Experience _and_Identity_Design_Processes [letzter Zugriff 23. September 2017]

Unsworth, J. (2000): "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?" in *Humanities Computing: formal methods, experimental practice* sponsored by King's College, London.

Visconti, A. (2010): "Songs of Innocence and of Experience: Amateur Users and Digital Texts" Dissertation, School of Information, University of Michigan. https:// deepblue.lib.umich.edu/handle/2027.42/71380 [letzter Zugriff 23. September 2017]

Warwick, C./ Terras, M. / Nyhan, J. eds. (2012): “Digital humanities in practice”. Facet Publishing.

Warwick, Claire (2006) “The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities”. *Final Report to the Arts and Humanities Research Council.* Arts and Humanities Research Council. http://discovery.ucl.ac.uk/189677/ [letzter Zugriff 23. September 2017]

Peer-to-Peer statt Client-Server: Der Mehrwert kollegialer Beratung und agiler DH-Treffen

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

Dogunke, Swantje
swantje.dogunke@klassik-stiftung.de
Forschungsverbund Marbach Weimar Wolfenbüttel; Klassik Stiftung Weimar

Mayer, Corinna
corinna.mayer@dla-marbach.de
Forschungsverbund Marbach Weimar Wolfenbüttel; Deutsches Literaturarchiv Marbach

Neumann, Katrin
Neumann@MaxWeberStiftung.de
Max Weber Stiftung

Cremer, Fabian
Cremer@MaxWeberStiftung.de
Max Weber Stiftung

Wübbena, Thorsten
twuebbena@dfk-paris.org
Max Weber Stiftung

Ein wesentliches Kennzeichen der Digital Humanities ist ihr hoher Grad an Interdisziplinarität, Vernetzung und Kommunikation. Angesichts ihrer Schnittstellenfunktion und des schnellen technologischen Wandels ist der regelmäßige Austausch zwischen DH-Vorhaben ebenso unerlässlich wie impulsgebend. Aus diesen Gründen finden seit drei Jahren regelmäßige Treffen zwischen den DH-MitarbeiterInnen des Forschungsverbundes Marbach Weimar Wolfenbüttel und der Max Weber Stiftung statt. Beide Institutionen vereinen strukturelle Gemeinsamkeiten: Als Zusammenschlüsse geografisch verteilter Forschungs- und Infrastruktureinrichtungen, die innerhalb der Geisteswissenschaften unterschiedliche wissenschaftliche Schwerpunkte verfolgen, streben beide Verbünde im Bereich der digital gestützten Forschung nach Synergien und Vernetzung.

Die Treffen werden als Peergroup-Treffen auf operativer Ebene nach dem Ansatz der kollegialen Beratung mit wechselnden Rollen und Formaten, wie z.B. Partnerinterviews, Impulsreferate, Buzzgroups oder Think-Pair-Share, durchgeführt (Tietze 2010). Ursprünglich begonnen als offener Austausch über Fragen des Aufbaus digitaler Infrastrukturen, des Wissenstransfers von fachlicher Expertise und zu aktuellen DH-Einwicklungen rückten durch den systemischen Ansatz zunehmend andere, unerwartete Themenfelder in den Fokus: Bei den DH-Vorhaben in den Verbünden werden vor allem Projektvorhaben initiiert, bei denen eine Konvergenz zwischen digitalen und analogen Forschungsmethoden angestrebt wird. Hieraus entsteht ein Spannungsverhältnis zu den etablierten Organisationstrukturen der Institutionen, die in ihren Abläufen und Strukturen nicht auf multidisziplinäre Projekte mit einem digitalen Anteil ausgerichtet sind. Bei der Übertragung des traditionellen Organisations- und Projektmanagements auf die DH-Projekte ergeben sich folgende Herausforderungen:

1. Dissonanz zwischen analogen und digitalen Projektworkflows: Der DH-Anteil in Forschungsprojekten ist nicht selten durch Förderrichtlinien motiviert und viele GeisteswissenschaftlerInnen kommen in diesen Projekten erstmalig in Kontakt zu DH (Pitti 2004).Technische Möglichkeiten werden dadurch sowohl unter- als auch überschätzt. Zu Projektbeginn existieren daher in den DH-Anteilen häufig nur abstrakte Vorstellungen über die Umsetzung der Projektinhalte in technische Verfahren oder Werkzeuge. Bei DH-Projekten werden die notwendigen konzeptionellen Arbeiten zu Projektbeginn – etwa die Evaluierung von Software oder die Identifizierung von Best-Practice-Modellen zu Kernaufgaben – nur unzureichend in den Projektplänen berücksichtigt (Tabak 2017).

2. Strukturelle Isolation: In vielen Forschungsprojekten arbeitet häufig nur eine einzelne Person als DHler/in; diese fungiert als singuläre Wissensressource, die schließlich zu den übrigen ProjektmitarbeiterInnen entweder in einem dienstleistungsähnlichen Verhältnis steht oder

für die Vermittlung und Legitimation neuartiger digitaler Methoden gegenüber den technisch weniger versierten Kollegen viel Engagement aufbringen muss. Da in Gedächtniseinrichtungen und -organisationen nur wenige bis keine Planstellen für DH vorhanden sind, können die Projektstellen nur bedingt auf institutionelles Wissen und Kompetenzen, z. B. in Bereichen wie Softwareentwicklung oder Datenmodellierung, zurückgreifen.

3. Diversität der Aufgaben: Alle technischen Fragen landen in den Forschungsprojekten in der Regel auf dem Tisch der DH. Förderanträge enthalten selten Anforderungsanalysen oder validierte Zielsetzungen bezüglich der DH-Anteile, die sich auch in den überlangen Kompetenzanforderungen widerspiegeln. Die DH-MitarbeiterInnen müssen dadurch Mehrfachfunktionen erfüllen: Einerseits sollen sie Dienstleistungen für Forschungsprojekte erbringen, andererseits eigene Forschungsprojekte durchführen und (implizit) Projektmanagement übernehmen (Reed 2014).

Durch diese Situation gehen einerseits die Kernkompetenzen der DH, wie die Gestaltung, Moderation und Begleitung von digitalen Prozessen verloren. Zeitgleich steigt der Rechtfertigungsdruck, wenn Anforderungen nicht kurzfristig realisiert werden können. Basierend auf den ersten Ergebnissen hat die Gruppe konkrete Ziele und methodische Grundlagen der Treffen ausgearbeitet. Diese umfassen neben dem eingangs anvisierten Austausch und gegenseitiger Beratung auch Qualitätssicherung, Innovationsleistung und Fortbildung. Neben diesen Funktionen zählen außerdem die Stärkung eigener Entscheidungen durch Verifikation des jeweiligen Partners, durch Impulse beschleunigte Entwicklungspfade und direkter Technologietransfer zu den Resultaten des Kooperationsmodells. Durch den systemischen Ansatz werden die projektbedingten Dissonanzen der Arbeitspläne offengelegt, die strukturelle Isolation zeitweise aufgelöst und die Diversität der Aufgaben durch Wissens-, Erfahrungs- und Technologietransfer erleichtert.

Das Poster soll zum einen die Funktionen und Impulse präsentieren, die sich durch das Peer-to-Peer Format ergeben haben. Ziel ist es, mit anderen KollegInnen ins Gespräch zu kommen und auch unangenehme Fragen zu diskutieren, etwa warum die DH in bestimmten Projektstrukturen primär mit Sensibilisierung und Legitimation beschäftigt sind. Dabei soll im Sinne der Tagung auch kritisch hinterfragt werden, inwieweit die DH nicht auch selber zu einer Wahrnehmung beitragen, die sie in die Rolle eines technischen Dienstleisters oder der digitalen Wollmilchsau drängt.

Bibliographie

Tietze, Kim-Oliver (2010): Wirkprozesse und personenbezogene Wirkungen von kollegialer Beratung – Theoretische Entwürfe und empirische Forschung. Wiesbaden.

Pitti, Daniel V. (2004): "Designing Sustainable Projects and Publications", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): A Companion to Digital Humanities. Oxford 2004, http://www.digitalhumanities.org/companion/ [letzter Zugriff 11.09.2017].

Reed, Ashley (2014): Managing an established digital humanities project: Principles and practices from the twentieth year of the William Blake archive, in: DHQ, 8.1, http://www.digitalhumanities.org/dhq/vol/8/1/000174/000174.html [letzter Zugriff 11.09.2017].

Tabak, Edin (2017): A Hybrid Model for Managing DH Projects, in: DHQ, 11.1, http://www.digitalhumanities.org/dhq/vol/11/1/000284/000284.html [letzter Zugriff 11.09.2017].

Personen- und Figurennetzwerke in Fernando Pessoas Publikationsplänen

Bigalke, Ben
bbigalke@smail.uni-koeln.de
Universität zu Köln, Deutschland

Drach, Sviatoslav
sdrach@smail.uni-koeln.de
Universität zu Köln, Deutschland

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

Sepúlveda, Pedro
pmpsepulveda@gmail.com
Neue Universität Lissabon, Portugal

Theisen, Christian
ctheise2@uni-koeln.de
Universität zu Köln, Deutschland

Im Nachlass des portugiesischen Dichters Fernando Pessoa (1888-1935) finden sich zahlreiche Listen geplanter Publikationen. Diesen stehen nur wenige zu Lebzeiten tatsächlich realisierte Veröffentlichungen gegenüber. Daraus ergibt sich ein Kontrast zwischen der Ebene des Möglichen und des Verwirklichten in der Literatur und Literaturproduktion. Vor diesem Hintergrund entsteht die digitale Edition "Fernando Pessoa. Projekte und Publikationen" und mit ihr die im Folgenden vorgestellte Netzwerkvisualisierung. Mit ihr wird ein Werkzeug zur Verfügung gestellt, das die Exploration des Personen- und Figurenkosmos in Pessoas Publikationsplänen über die Zeit ermöglicht.

Pessoas Werk zwischen Planung und Publikation

Die Dynamik der Schriften Pessoas ist im Spannungsverhältnis zwischen dem projizierten Werk, das der Dichter im Sinne eines vollkommen Ganzen konzipiert hat, und dem tatsächlich Geschriebenen und nur in geringem Maße Publizierten zu verstehen. Pessoa war bei der Veröffentlichung seiner Werke extrem selektiv und die Dynamik seines Schreibens weist auf eine ständige Bedeutungsverschiebung hin, die sowohl an fragmentarischen Schriften seines Nachlasses als auch an seinen Publikationsplänen zu erkennen ist (vgl. dazu Cunha 1987, Martins 2003, Gusmão 2003 und Sepúlveda 2013). Diese Bedeutungsverschiebung hängt stark mit dem Wahl der Autorennamen zusammen, die ebenfalls einem ständigen Wechsel unterlag und in den Plänen zur Edition und Publikation des Werkes eine hohe Bedeutung gewann. Zur Definition eines Publikationsvorhabens gehörte für Pessoa die Zuordnung einer bestimmten Autorfigur, deren fiktionale Persönlichkeit sowohl durch ein bestimmtes Werk konstruiert werden sollte als auch dieses Werk in seiner Besonderheit definieren würde.

Die Spannung zwischen Planung und Publikation des Werkes wird daher noch dadurch verstärkt, dass Pessoa unter verschiedenen, insgesamt etwa 120 Autorennamen geschrieben hat - oder geplant hat zu schreiben (vgl. dazu Pessoa 2012). Eine besonders wichtige Rolle in Pessoas Werk und seinen Werkplänen spielen dabei die Namen Alberto Caeiro, Álvaro de Campos und Ricardo Reis, die er (in Abgrenzung zu den weiteren Pseudonymen) Heteronyme genannt hat.

Digitale Edition "Projekte und Publikationen"

Die Notizzettel, Seiten aus Notizbüchern und andere Papiere aus Pessoas Nachlass, auf denen er die Pläne für seine Werke handschriftlich oder mit Schreibmaschine geschrieben festgehalten hat, werden in einer Kooperation zwischen dem Institut für Literatur und Tradition (IELT) der Neuen Universität Lissabon und dem Cologne Center for eHumanities (CCeH) der Universität zu Köln digital ediert (Sepúlveda und Henny-Krahmer 2017). [1]

Neben den Dokumenten aus dem Nachlass umfasst die digitale Edition auch die zu Lebzeiten von Pessoa publizierten Gedichte. Gegenstand des hier vorgestellten Netzwerkes sind jedoch ausschließlich die Dokumente, auf denen die Figuren und Personen genannt sind und deren Beziehungen ausgehend von ihrer gemeinsamen Erwähnung über die Zeit untersucht werden. Die Dokumente sind in der Edition in TEI codiert, wobei die Namensvorkommen erfasst werden und eine Identifikation der hinter den Namen stehenden Personen und Figuren in einem zentralen Index erfolgt. Transkriptionen und Index bilden zusammen mit den für jedes Dokument festgehaltenen Metadaten die Datengrundlage für das Figuren- und Personennetzwerk, wobei insbesondere die für die Chronologie relevante Datierung zu nennen ist.

Netzwerkvisualisierung zu Personen und Figuren

Die Vorkommen von Namen in Pessoas Publikationsplänen werden hier mit Hilfe einer interaktiven Netzwerkvisualisierung analysiert, um zu untersuchen, wie sich der von ihm in den Dokumenten entworfene Personen- und Figurenkosmos über die Zeit entwickelt. Für dynamische Netzwerkvisualisierungen gibt es in den DH bereits verschiedene Ansätze (vgl. u. a. Rigal et al. 2016, Xanthos et al. 2016). Für das Pessoa-Netzwerk ergibt sich die Dynamik aus der Möglichkeit, das Gesamtnetzwerk der Personen und Figuren über alle Dokumente hinweg auf Dokumente aus bestimmten Zeiträumen oder Jahren einzugrenzen. Es ist als heuristisches Instrument gedacht, um Hypothesen zur Chronologie von Personen- und Figurenkonstellationen zu generieren und im Ansatz überprüfen zu können.

Für das Netzwerk, das unter http://www.pessoa-digital.pt/de/network verfügbar ist, sind 249 Do-

kumente ausgewertet worden, die insgesamt 369 Namen enthalten. Es werden sowohl historische Personen als auch fiktive Figuren aus Pessoas Werkwelt gezeigt. Analysiert wird das Vorkommen der Namen auf Pessoas Publikationsplänen von 1913 bis zu seinem Tod im Jahr 1935. Die frühen Dokumente (vor 1913) werden hier nicht berücksichtigt, da sie noch in Bearbeitung sind.

Die Netzwerkdaten sind mit Hilfe von XSLT aus den TEI-Dokumenten generiert worden und liegen im JSON-Format vor. [2] Die interaktive Visualisierung ist mit der Bibliothek D3 erstellt worden, wobei ein Netzwerklayout von Mike Bostock adaptiert und um weitere Funktionalitäten ergänzt wurde. [3] Erweiterungen, die für die vorliegende Anwendung vorgenommen wurden, sind u. a. die Möglichkeit, Teile des Netzwerks ein- und auszublenden (nach Chronologie; Teilnetzwerke für die Verbindungen, die von einzelnen Personen ausgehen; nur Knoten mit oder auch Knoten ohne Verbindungen) sowie Optionen für die Darstellung (Einblenden von Labels; Dichte bzw. Weite der Anzeige des Netzwerks).

Abbildung 1: Optionen für die Anzeige des Netzwerks

Die Größe der Knoten im Netzwerk zeigt an, wie häufig einzelne Namen auf den Dokumenten erwähnt werden. Zur Ermittlung der Knotengröße wurde die Formel 2 + log2(size) angewandt, wobei size für die tatsächliche Häufigkeit des Vorkommens steht. Um zwischen fiktiven Figuren und historischen Personen unterscheiden zu können, sind die Knoten unterschiedlich eingefärbt (türkis = fiktiv; dunkelbau = historisch). Bei den Netzwerkkanten verdeutlicht die Dicke, wie häufig Namen gemeinsam auf Dokumenten vorkommen: je häufiger das gemeinsame Auftreten, umso dicker die Linien in der Visualisierung. Dabei ist die minimale Kantendicke 1 Pixel (bei einer gemeinsamen Erwähnung). Pro weiterer gemeinsamer Erwähnung nimmt die Kantendicke um 1 Pixel zu.

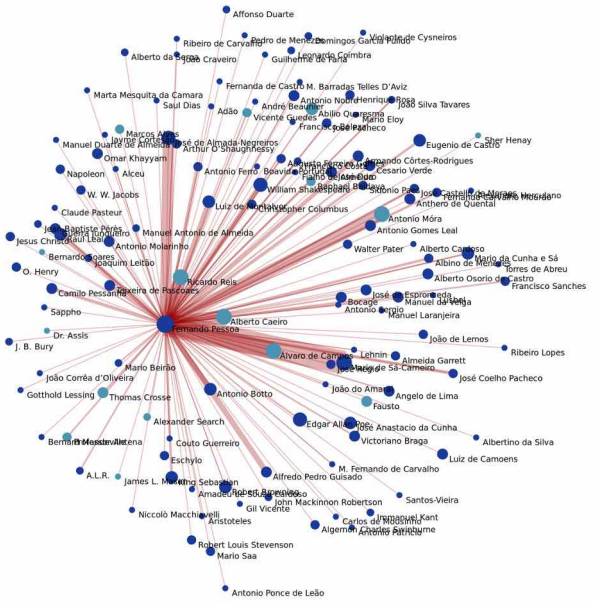


Abbildung 2: Teilnetzwerk mit Fernando Pessoa als zentralem Knoten

Zentrale Funktionalitäten in der Netzwerkanwendung sind Auswahloptionen, welche die Chronologie der Namenserwähnungen betreffen. So kann das Netzwerk neben einer Gesamtdarstellung auch für Vorkommen in einzelnen Jahren angezeigt werden. Darüber hinaus ist eine Anzeige nach Perioden möglich (z. B. 1919-1927). Da es in der Pessoa-Forschung konkurrierende Vorschläge für eine Periodisierung des Werkes gibt, werden zwei verschiedene Einteilungen in Perioden zur Auswahl angeboten. Auf diese Weise wird es möglich, die Entwicklung von Pessoas Personen- und Figurenkosmos, wie er sich in den Publikationsplänen darstellt, kritisch zu untersuchen.

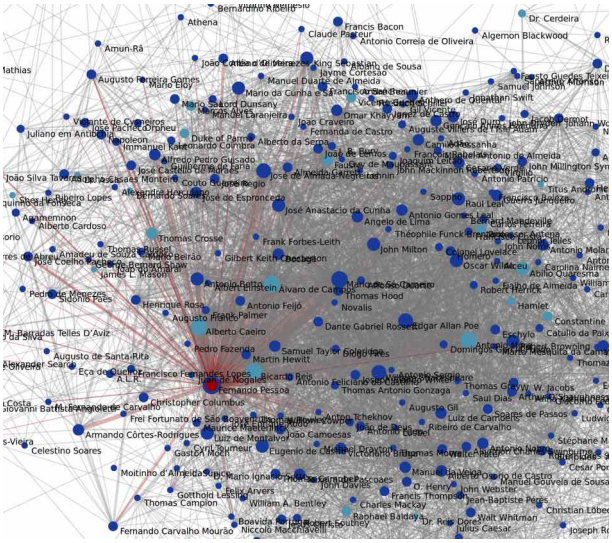


Abbildung 3: Gesamtnetzwerk mit Pessoa als häufigstem Namen

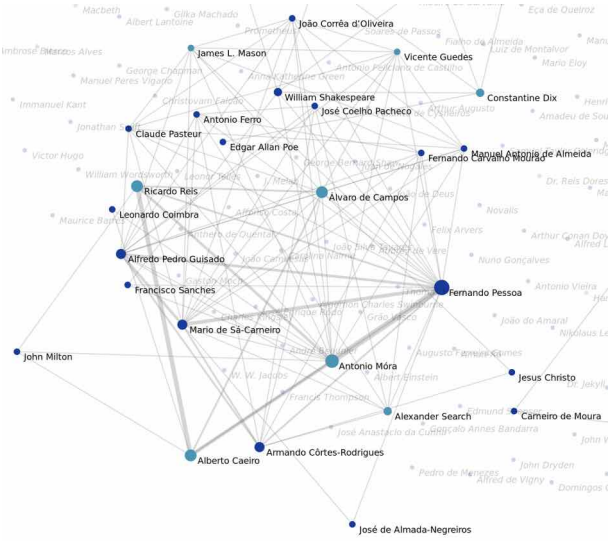


Abbildung 4: Teilnetzwerk 1915

Wenn man das Vorkommen aller Namen im gesamten Netzwerk betrachtet, so ist zu erkennen, dass Fernando Pessoa als Name am häufigsten vorkommt, direkt gefolgt von den Namen der Heteronyme Alberto Caeiro, Ricardo Reis und Álvaro de Campos, was etabliertes Wissen zu der Bedeutung der Heteronyme in Pessoas Werk bestätigt (vgl. Abb. 2 und 3). An zweiter Stelle stehen dann William Shakespeare, José Almada-Negreiros, Edgar Allan Poe, Mário de Sá-Carneiro und António Mora. Dabei ist beispielsweise interessant zu sehen, dass die Heteronyme (und Mora, der zeitweise als starker Kandidat für die Rolle eines Heteronyms galt) auch zusammen mit Fernando Pessoa selbst, aber besonders unter sich verbunden sind, was auf eine gewisse Autonomie des heteronymischen Universums hindeutet. Die Namen von Shakespeare und Poe weisen auf die zwei wichtigsten Referenzen Pessoas aus der englischsprachigen Literatur hin, während Almada und Sá-Carneiro die zwei für ihn bedeutendsten zeitgenössischen portugiesischen Schriftsteller waren.

Betrachtet man das Netzwerk chronologisch anhand ausgewählter Jahre und Perioden, innerhalb derer die Publikationspläne verfasst wurden, so sind Tendenzen zu erkennen, die historisch, editorisch und auch poetisch für das Werk von maßgeblicher Bedeutung sind. So kann man beispielsweise sehen, wie in den Jahren von 1913 bis 1919, und besonders zwischen 1914 und 1915, die Namen der Heteronyme zusammen mit dem von Fernando Pessoa am häufigsten vorkommen und vor allem untereinander verbunden sind (vgl. Abb. 4).

Namen anderer Schriftsteller und historischer Figuren kommen tendenziell in späteren Perioden, etwa ab den 20er Jahren, häufiger vor. Sie zeigen ein zunehmendes Interesse Pessoas an der Veröffentlichung eigener Übersetzungen von einigen für ihn entscheidenden Werke der Weltliteratur. Auch die wichtigsten Beziehungen Pessoas zu zeitgenössischen portugiesischen Schriftstellern und Kritikern sind im Netzwerk deutlich zu erkennen, besonders zwischen 1913 und 1918 (vgl. Abb. 5). Es handelt sich dabei vor allem um die modernistische Generation, die sich um die Zeitschrift Orpheu versammelt hat, und ab 1928 und bis zu Pessoas Tod 1935 dann die sogenannte zweite modernistische Generation um die Zeitschrift Presença.

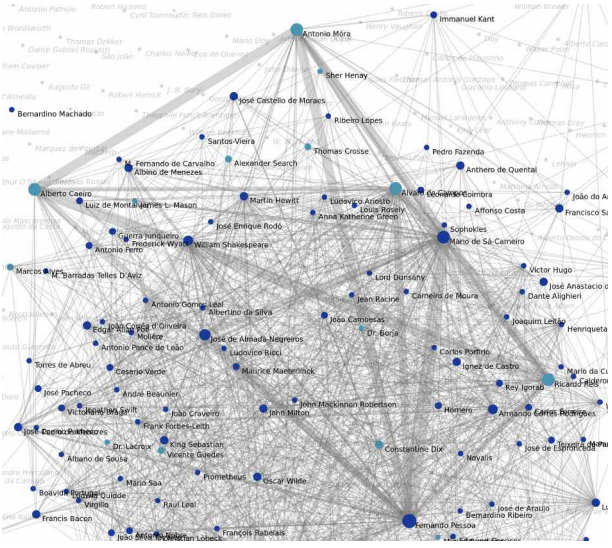


Abbildung 5: Teilnetzwerk 1913-1918

Fazit

Die aus der Netzwerkinterpretation gewonnenen literaturhistorischen Erkenntnisse bestätigen, dass für Pessoa die Edition und Publikation des Werkes und dessen Planung nicht von der Bedeutungsebene des Werkes selbst zu unterscheiden sind. Dass bestimmte Namen insgesamt oder zu bestimmten Zeiten besonders häufig auf den Publikationsplänen auftauchen, ist gleichbedeutend mit deren Wichtigkeit für das Werk in der entsprechenden Zeitperiode. Das gilt neben dem Vorkommen einzelner Namen auch für die gemeinsamen Vorkommen mehrerer Namen, wodurch die Bedeutung bestimmter Konstellationen zu bestimmten Zeiten sowohl in Pessoas Publikationsplänen als auch in seinem Werk deutlich wird. Diese grundlegende Erkenntnis für die Interpretation von Pessoas Werk bestätigt einige Intuitionen der Kritiker über den größeren Zusammenhang mehrerer Ebenen von Pessoas Werk (z. B. der Ebene der Edition des Werkes, vgl. dazu Sepúlveda und Uribe 2016), sowie die Vorstellung von Pessoas Werkganzem, die einer materiellen Fragmentarität seiner Schriften gegenübersteht (vgl. dazu Martins 2003, Gusmão 2003, Sepúlveda 2013, Feijó 2015).

Methodisch eröffnet die Visualisierung durch das interaktive Element und den höheren Grad der Abstraktion gegenüber den edierten Dokumenten, die in der digitalen Edition studiert werden können, neue Interpretationsspielräume. Dabei ist es jedoch sehr wichtig, die Dokument-, Text- und Datengrundlage sowie die methodischen Wege zum Netzwerk und zur visuellen Darstellung stets im Blick zu behalten. Für diesen Beitrag bieten wir dafür eine Dokumentation an, die direkt mit dem interaktiven Netzwerk verbunden ist. [4] Der Zusammenhang zwischen Quellen und Analyse wird auch dadurch hergestellt, dass die interaktive Visualisierung an die digitale Edition angebunden ist. [5]

Fußnoten

1. Die digitale Edition ist in einer Beta-Version unter http://www.pessoadigital.pt [letzter Zugriff 14. Januar 2018] verfügbar; die Entwicklung wird fortlaufend über GitHub organisiert: https://github.com/cceh/pessoa [letzter Zugriff 14. Januar 2018]. Zur editorischen Herangehensweise aus digitaler Perspektive vgl. Henny-Krahmer und Sepúlveda 2017.
2. Die dem Netzwerk zugrunde liegenden Daten sind unter https://github.com/cceh/pessoa/tree/master/app/data/network [letzter Zugriff 14. Januar 2018] einsehbar.
3. Das Ausgangs-Layout von Bostock trägt den Titel “Force Layout with Mousover” (Bostock 2017).
4. http://www.pessoadigital.pt/de/network/documentation [letzter Zugriff 14. Januar 2018].
5. Derzeit über den Menüpunkt “Chronologie”, vgl. http://www.pessoadigital.pt/de/index.html [letzter Zugriff 14. Januar 2018].

Bibliographie

Bostock, Mike (2017): “Force Layout with Mouseover Labels”, in: Mike Bostock’s Blocks. https://bl.ocks.org/mbostock/1212215 [letzter Zugriff 14. Januar 2018].

Cunha, Teresa Sobral (1987): "Planos e projectos editoriais de Fernando Pessoa: uma velha questão", in: Revista da Biblioteca Nacional, Série 2, Vol. 2, N. 1: 92-107.

Feijó, António M. (2015): “Uma admiração pastoril pelo diabo (Pessoa e Pascoaes)”, in: Pessoana. Ensaios. Lissabon: Imprensa Nacional-Casa da Moeda.

Gusmão, Manuel (2003): "O Fausto — um teatro em ruínas", in: Românica 12: 67-86.

Henny-Krahmer, Ulrike / Sepúlveda, Pedro (2017): “Pessoa’s editorial projects and publications: the digital edition as a multiple form of textual criticism”, in: Boot, Peter / Cappellotto, Anna / Dillen, Wout / Fischer, Franz / Kelly, Aodhán / Mertgens, Andreas / Sichani, Anna-Maria / Spadini, Elena / van Hulle, Dirk (eds.): Advances in Scholarly Editing. Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp. Leiden: Sidestone Press, 125-133 https://www.sidestone.com/books/advances-in-digital-scholarly-editing [letzter Zugriff 14. Januar 2018].

Martins, Fernando Cabral (2003): "Breves notas sobre a alta definição", in: Românica. N.º 12. 157-164.

Pessoa, Fernando (2012): Teoria da Heteronímia. Hsg. von Fernando Cabral Martins und Richard Zenith. Lissabon: Assírio & Alvim.

Rigal, Alexandre / Rodighiero, Dario / Cellard, Loup (2016): “The Trajectories Tool: Amplifying Network Visualization Complexity”, in: Digital Humanities 2016. Conference Abstracts. Kraków: Jagiellonian University & Pedagogical University, 328-330 http://dh2016.adho.org/abstracts/340 [letzter Zugriff 14. Januar].

Sepúlveda, Pedro (2013): Os livros de Fernando Pessoa. Lissabon: Ática.

Sepúlveda, Pedro / Henny-Krahmer, Ulrike (eds., 2017): Fernando Pessoa – Digitale Edition. Projekte und Publikationen. Editorische Lei-

tung Pedro Sepúlveda, technische Leitung Ulrike Henny-Krahmer. Lissabon und Köln: IELT, Neue Universität Lissabon und CCeH, Universität zu Köln http://www.pessoadigital.pt [letzter Zugriff 14. Januar 2018].

Sepúlveda, Pedro / Uribe, Jorge (2016): O Planeamento editorial de Fernando Pessoa. Lissabon: Imprensa Nacional-Casa da Moeda.

Xanthos, Aris / Pante, Isaak / Rochat, Yannick / Grandjean, Martin (2016): "Visualizing the Dynamics of Character Networks", in: Digital Humanities 2016. Conference Abstracts. Kraków: Jagiellonian University & Pedagogical University, 417-419 http://dh2016.adho.org/abstracts/407 [letzter Zugriff 14. Januar 2018].

Perspektiven auf ein Korpus. Kombinationen quantitativ-qualitativer Analysemethoden zur Ermittlung von Textgliederungsprinzipien

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

Einführung

Im Bereich digital basierter Untersuchungen wird zunehmend eine Verzahnung quantitativen und qualitativen Arbeitens gefordert. In der konkreten Arbeit der Korpusanalyse wird aus dieser scheinbaren Dichotomie jedoch schnell eine Methodenvielfalt, denn gerade durch Kombinationen verschiedener Perspektiven auf die Daten werden unterschiedliche Phänomene greifbar und entfaltet sich das volle Potential quantitativ-qualitativen Arbeitens. Das hier präsentierte Poster soll dies an einem konkreten Beispiel veranschaulichen.

Fragestellung

Inhaltlicher Ausgangspunkt ist die Frage nach Textgliederungsprinzipien, welche für bestimmte erbauliche Textsorten kennzeichnend sind. Mittel der Textgliederung können als Mittel der Markierung von Teiltexten innerhalb eines Gesamttextes beschrieben werden (Hausendorf/Kesselheim 2008: 41) und können wiederum für Textsorten charakteristisch sein. Sie finden sich auf verschiedenen Ebenen des Textes, u.a. im Bereich der Typographie (Stein 2003: 422).

Gerade diese typographischen Gliederungsmerkmale stellen einen guten Ausgangspunkt für eine quantitative Analyse von Textgliederungsmerkmalen dar, da sie in TEI-annotierten Korpora z.B. durch die XML-Strukturierung automatisch greifbar werden. Anders als bei dem Verfahren, textbezogene Phänomene reduziert auf bestimmte TEI-Strukturen zu untersuchen (z.B. Schöch 2016: 351ff., Haaf 2016), gelangen hier die im Korpusvergleich möglicherweise signifikanten Häufigkeiten der TEI-Strukturierungen selbst in den Blick.

Die inhaltliche Frage nach Textgliederungsprinzipien erbaulicher Textsorten wird ausführlich behandelt in Haaf (in Vorber.). Im vorliegenden Beitrag stehen – der thematischen Ausrichtung der Konferenz entgegenkommend – Überlegungen zur adäquaten Methodik einer solchen Untersuchung im Vordergrund.

Korpus- und Analysegrundlage

Der hier präsentierten Studie liegen drei Teilkorpora des 17. Jahrhunderts aus dem Deutschen Textarchiv (2017) zugrunde:

- Prosaische Erbauungsliteratur: 25 Bände (10 Autoren, 10.501 Seiten)
- Funeralschriften: 334 Schriften (14.316 Seiten)
- Referenzkorpus: 187 Bände verschiedener Textsorten (60.798 Seiten)

Die Texte des DTA-Korpus wurden nach einheitlichen Richtlinien und mittels eines TEI-Subsets, das Ambiguitäten der Auszeichnung möglichst reduziert, ausgezeichnet (Haaf et al. 2014/15).

Für die vorliegende Untersuchung wurden einzelne TEI-Strukturen hinsichtlich der Häufigkeit ihres Auftretens (relativ zur Token-Anzahl) und ihrer Verteilung im jeweiligen Korpus verglichen, um speziell die Unterschiede in der Textgliederung zwischen den untersuchten Textsorten herauszuarbeiten. Dabei wurden solche Tags einbezogen, die voraussichtlich Textgliederungsmerkmale repräsentieren. So kann z.B. tei:div die Kapitelstruktur eines Textes anzeigen, durch tei:l wird der Wechsel zwischen Prosatext und Lyrik greifbar, tei:note zeigt Metatexte in Form von Anmerkungen, z.B. Marginalien, an, tei:hi repräsentiert Hervorhebungen von Textpassagen ge-

genüber dem Grundtext. Die Ergebnisse wurden einer qualitativen Beurteilung unterzogen.

Ergebnisse

Zur Evaluation eines Merkmals war hier nicht allein der Blick auf seine relativen Häufigkeiten in und deren signifikante Unterschiede zwischen den untersuchten Korpora relevant. Die signifikant erhöhte Häufigkeit eines Merkmals kann vielmehr unterschiedliche Gründe haben. So kann sie einerseits zwar durchaus (1) auf die höhere Relevanz des Merkmals im Korpus hindeuten, wie sich am Merkmal der Marginalie zeigt, das signifikant häufig und regelmäßig verteilt im Korpus der Funeralschriften auftritt (Abb. 1). Sie kann andererseits aber auch (2) aufgrund der unausgewogenen Verteilung des Merkmals im Korpus gar nicht aussagekräftig sein, entweder weil (2a) sich das Korpus selbst als in sich unausgewogen und nicht repräsentativ für den zu beschreibenden Gegenstand erweist oder weil (2b) das Merkmal im gegebenen Kontext nicht relevant ist, wie etwa die horizontale Trennlinie zwischen Textteilen, die in allen drei Vergleichskorpora unregelmäßig verteilt war (Abb. 2). Andererseits kann es (3) auch vorkommen, dass die bestehende Ausgewogenheit der Verteilung eines Merkmals in einem Korpus letztlich nicht aussagekräftig für dessen Relevanz ist.

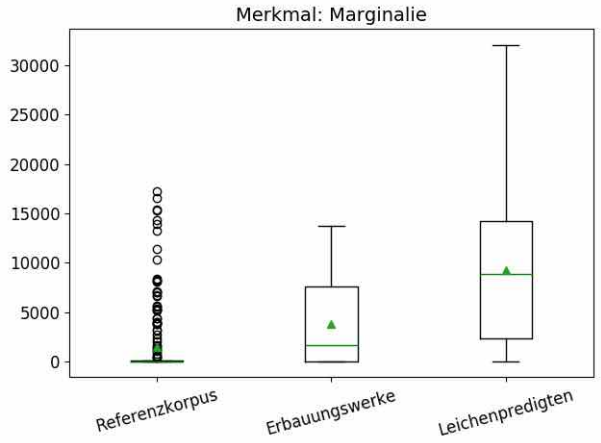


Abb. 1: Relative Häufigkeiten je 1 Mio. Token und deren Verteilung in den drei untersuchten Korpora für das Merkmal „Marginalie“

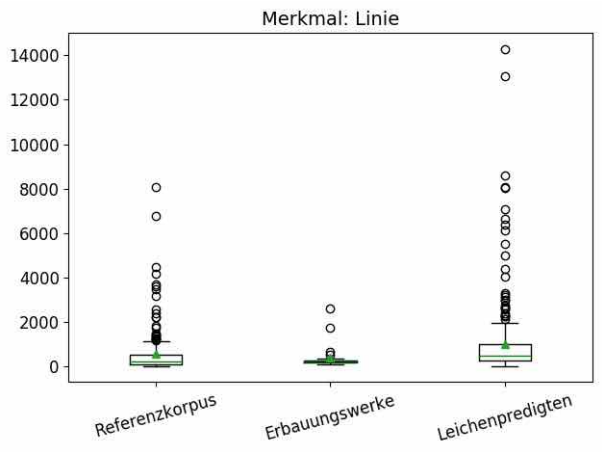


Abb. 2: Relative Häufigkeiten je 1 Mio. Token und deren Verteilung in den drei untersuchten Korpora für das Merkmal „Horizontale Trennlinie“

Weiterhin konnten im gegebenen Kontext (4) auch Merkmale mit geringeren Häufigkeiten relevant sein, und schließlich ist nicht zuletzt (5) auch der Ort, an dem ein Merkmal im Dokument auftritt, zu berücksichtigen. Beide Aspekte (4 und 5) zeigen sich z.B. am Merkmal der Liste, die in den erbaulichen Prosawerken erwartungsgemäß selten, aber relativ regelmäßig auftritt, und zwar in Form von Registern am Buchbeginn oder Buchende (in tei:front oder tei:back).

Methodisch zeigte sich also, dass für eine adäquate Beurteilung der untersuchten Merkmale verschiedene Blickwinkel notwendig sind. Das Poster veranschaulicht anhand der erwähnten und weiterer Beispiele diese genannten methodischen Aspekte.

Inhaltlich führte die Untersuchung zutage, dass z.B. Merkmale, die den Zugang zum Text erleichtern und Hilfe zur Orientierung im Text geben, für die erbaulichen Textsorten relevant sind (Näheres vgl. Haaf (in Vorber.)).

Bibliographie

Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin 2017. http://www.deutschestextarchiv.de [letzter Zugriff: 24.09.2017]

Haaf, Susanne (i. Vorb.): „Art und Funktion von typographischen Mitteln zur Textgliederung in erbaulichen Textsorten des 17. Jahrhunderts. Automatische Analyse im Korpusvergleich und qualitative Einordnung“, in: Simmler, Franz / Baeva, Galina (Hrsg.): *Textgliederungsprinzipien.*

Ihre Kennzeichnungsformen und Funktionen vom 8. bis 18. Jahrhundert. Akten zum Internationalen Kongress vom 22. bis 24. Juni 2017 an der Universität St. Petersburg. Berlin: Weidler [2018].

Haaf, Susanne (2016): "Corpus Analysis based on Structural Phenomena in Texts. Exploiting TEI Encoding for Linguistic Research", in: Nicoletta Calzolari et al.: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 23.–28. Mai 2016, Portorož (Slovenia)*. Paris.

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2014/15): "The DTA 'Base Format'. A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources", in: *Journal of the Text Encoding Initiative* 8.

Hausendorf, Heiko / Kesselheim, Wolfgang (2008): *Textlinguistik fürs Examen*. Göttingen: Vandenhoeck & Ruprecht.

Schöch, Christoph (2016): „Ein digitales Textformat für die Literaturwissenschaften. Die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse", in: *Romanische Studien* 4.

Stein, Stephan (2003): *Textgliederung. Einheitenbildung im geschriebenen und gesprochenen Deutsch. Theorie und Empirie*, Berlin.

Professionalisierung der Ausbildung von Geisteswissenschaftlern in der Digitalisierung von Texten

Dahnke, Michael
fedja_anatevka@web.de
Universität Würzburg, Deutschland

Motivation

»Angesichts der steigenden Sichtbarkeit der Digital Humanities, auch und gerade bei universitären Schwerpunktsetzungen, ist die Frage, wie sie am sinnvollsten gelehrt werden sollen, von steigender Bedeutung« […] »Um diese Bemühungen längerfristig in der Community zu verankern, wurde auf der ersten Jahreskonferenz der Digital Humanities der deutschsprachigen Länder im März 2014« […] »eine Arbeitsgruppe der DHd gegründet. Die Proponenten der Arbeitsgruppe schlugen vor,« […] »die bisher losen Diskussionen stärker auf ein »Referenzcurriculum« zu fokussieren.« (https://dig-hum.de/ag-referenzcurriculum-digital-humanities)

Diesem Anliegen fühlt sich der als Dozent für die Vermittlung von Digitalisierungskompetenz an der Universitätsbibliothek Würzburg arbeitende Autor verpflichtet. Er engagiert sich in der *AG Referenzcurriculum Digital Humanities,* ist Mitglied des Würzburger Arbeitskreis Digitale Editionen und unterstützt das Editionsprojekt *Narragonien digital.* Er plädiert mit seinem Beitrag dafür, die DH-Ausbildung bezüglich der Bilddigitalisierung und des OCR (Optical Character Recognition) stärker zu kanonisieren. Mit seiner eigenen Lehrveranstaltung *Bilddigitalisierung und OCR für Geisteswissenschaftler* , deren Schwerpunkt auf der Erstellung der Digitalisate und dem OCR liegt, bietet er eine Referenz, die er hiermit zur Diskussion in der Community stellt.

Das Ziel der Lehrveranstaltung ist die Vermittlung von Kenntnissen des gesamten Digitalisierungsprozesses für MA-, BA- und LA-Studentinnen und Studenten aller geisteswissenschaftlichen Fachrichtungen. Diese nachfolgend als Zielgruppe Bezeichneten sollen in die Lage versetzt werden, selbständig strukturiert ausgezeichnete, digitale Volltexte zu erzeugen und die Ergebnisse der Erstellung derselben beurteilen zu können. Diese Kenntnisse sind wichtig, weil Projekte häufig auch dadurch gefährdet sind, dass Vertretern der Zielgruppe die mangelnde Qualität der ihnen vorliegenden Digitalisate zu spät bewusst wird. Darum müssen sie von Beginn an Digitalisate auf ihre Brauchbarkeit für die automatische Texterkennung beurteilen können. Strukturiert ausgezeichnete, digitale Volltexte sind unabdingbar beispielsweise für Topic Modeling oder Sentiment Analysis auf größeren Textcorpora und als Zwischenstufe für die Erstellung digitaler Editionen.

Ablauf

Kurz gefaßt sind die sechs Arbeitschritte dafür

1. die Sensibilisierung für juristische Aspekte der Bilddigitalisierung,
2. die Suche nach vorhandenen oder die Erstellung von eigenen Digitalisaten,
3. deren Vorverarbeitung,
4. das OCR,
5. die Anreicherung der Digitalisate und des generierten Rohtextes – im Idealfall einer diplomatische Transkription – mit Metadaten, sowie

6. die inhaltliche Auszeichnung des generierten Rohtextes.

2.1. Rechtliche Grundlagen der Bilddigitalisierung

Um den Teilnehmern den typischen Arbeitsablauf der Textdigitalisierung möglichst stringent und ohne thematische Abschweifungen vorzuführen, werden sie zuerst intensiv mit den juristischen Grundlagen der Bilddigitalisierung vertraut gemacht. Dazu gehören

1. die Vorstellung des UrhG beziehungsweise speziell der § 60d und 60g UrhG in der novellierten Fassung des UrhWissG 2017,
2. die Unterscheidung zwischen Immaterial- und Materialgüterrecht und was daraus für die Digitalisierung zweidimensionaler Objekte folgt,
3. die Persönlichkeitsrechte des Urhebers und weiterer Betroffener sowie
4. der Umgang mit Werken, die unter der Creative Commons Lizenz stehen und die Möglichkeit, diese selbst zu benutzen.

2.2. Suche nach vorhandenen Digitalisaten beziehungsweise deren Erstellung

Am Anfang der Transformation vom gedruckten zum digitalen Corpus steht die Suche nach möglicherweise bereits vorhandenen Digitalisaten. Diese Suche setzt neben nicht DH-spezifischen Kenntnissen der Erschließung das Wissen um Metadaten zu digitalen Bildformaten voraus. Die Vertreter der Zielgruppe müssen in dieser Situation wissen, dass beispielsweise die Chancen eines erfolgreichen OCR mit einem JPEG mit 72 dpi deutlich geringer sind als mit einem unkomprimierten TIFF, True Color und 300 dpi. Sollte er schließlich feststellen, dass ihm Digitalisate in der gewünschten Form nicht zugänglich sind, bedarf er der Kenntnisse zu digitalen Bildformaten genauso, um im nächsten Schritt erfolgreich den Scan selbst durchzuführen oder nach seinen Vorgaben durchführen zu lassen.

Ausgehend von den skizzierten Anforderungen werden den Vertretern der Zielgruppe in der Veranstaltung die Grundlagen der Bilddigitalisierung nahe gebracht. Vertieft wird hier auf das menschliche Sehen und die Farbreproduktion, die Entstehung digitaler Bilder (Rastergraphik, optische und interpolierte Auflösung, Farbtiefe), Farbräume, Color-Management-Systeme, verschiedene Graphikspeicherformate, Speichermedien und verschiedene Scannertypen eingegangen. Auch für die Berücksichtigung konservatorischer Aspekte werden die Teilnehmer sensibilisiert.

2.3. Aufbereitung der Digitalisate für das OCR

Die Forschung im Bereich OCR, insbesondere auf Inkunabeln und Wiegendrucken, sowie die eigene Praxis des Autors belegen die Bedeutung einer vorherigen Aufbereitung der Digitalisate für das OCR, der darum entsprechend Platz in der Veranstaltung eingeräumt wird (Springmann 2015: 9). Als Tätigkeiten sind hier in der Reihenfolge ihrer Ausführung die Bereinigung und anschließende Binarisierung der Digitalisate, deren Segmentierung in einzelne Textabschnitte und schließlich einzelne Textzeilen sowie die Transkription (›ground truth‹) einer Anzahl der Textzeilen zu nennen. Die gesamte Aufbereitung der Digitalisate für das OCR führen die Vertreter der Zielgruppe in der Veranstaltung selbst an ausgewähltem Trainingsmaterial durch. Nach der Teilnahme an der Veranstaltung sollen sie auch diesen Arbeitsschritt selbständig erledigen und Arbeitsergebnisse anderer in diesem Bereich beurteilen können.

2.4. OCR

Entsprechend der zunehmenden Spezialisierung des Digitalisierungszentrums der UB Würzburg ist es erstens wünschenswert, in der Lehrveranstaltung besonders auf das Training eigener Modelle beispielsweise mit *OCRopus* einzugehen. Zweitens soll ein Arbeitsablauf für die Digitalisierung eigener Texte vorgestellt werden, der von den Vertretern der Zielgruppe selbständig mit möglichst geringem technischen Aufwand realisierbar ist. Diese Form der Digitalisierung von Texten soll als handhabbares Mittel zum Zweck wahrgenommen werden.

Schließlich wird in diesem Zusammenhang auf die Frage nach dem Zeitpunkt der Normalisierung des mit dem OCR erstellten Textes eingegangen. Soll bereits mit dem OCR ein normalisierter Text erstellt werden und wenn ja, nach welchen Regeln? Ist also beispielsweise von der verwendeten OCR-Software das Schaft-s bereits automatisch als Rund-s zu lernen und anschließend zu transkribieren? Oder soll das Ergebnis des OCR graphisch so dicht wie möglich am Original bleiben und normalisierende Eingriffe erst hinterher erfolgen?

2.5. Auszeichnung/Anreicherung

Nach dem OCR wird erst die Notwendigkeit der Auszeichnung sowohl der Digitalisate mit Metadaten als auch des Rohtextes erläutert, die die Teilnehmer dann auch selbst vornehmen sollen. Für den extrahierten Rohtext gilt das in zweifacher Hinsicht: Erstens sind ihm Metadaten hinzuzufügen, welche die spätere, eindeutige Identifikation des Werkes und dessen Auffindbarkeit ermöglichen. Zweitens muss der Text strukturiert mit inhaltsbezogenen Elementen angereichert werden.

Konkret sind bei der digitalen Repräsentation eines Romans beispielsweise die Figuren, Orte, Zeitpunkte und gegebenenfalls weitere signifikante Entitäten im Text für das spätere, automatisierte Retrieval zu kodieren. Andere Anforderungen stellen digitalierte Transkriptionen gesprochener Sprache und wiederum andere die Erstellung einer Urkundenedition (Vogeler 2015). Für die visuelle Präsentation, beispielsweise auf einem Webportal, sind textstrukturierende Merkmale wie die Einteilung nach Kapiteln, Abschnitten, Fußnoten etc. zu kennzeichnen. Dem unterschiedlichen Kenntnisstand der Teilnehmer geschuldet muss hier vor der Vorstellung der TEI Guidelines zuvor zweifelsohne XML dargestellt werden. Wie ausführlich daneben Dublin Core und bibliotheksspezifische Formate (MARC21) thematisiert werden, ist noch nicht entschieden. Wieviel Zeit für weiterführende Themen wie Named Entity Recognition , PID und Normdaten wie GND bleibt, muß ebenfalls die Praxis weisen.

Forschungsbezug und Weiterentwicklung

Die eine Stärke der Würzburger Lehrveranstaltung ist die Praxisorientierung, der nach der zweitägigen Einführung noch stärker mit einer drei Tage dauernden Übung Rechnung getragen wird. Bei dieser werden die Vertreter der Zielgruppe mit vorbereiteten Scans selbständig die genannten Arbeitsschritte von der Bereinigung und anschließenden Binarisierung über die Segmentierung und Transkription, dem OCR bis zum anschließenden Auszeichnen beziehungsweise der Anreicherung des digitalen Corpus mit den nötigen Metadaten vornehmen.

Unverzichtbar für die gesamte Ausbildung im DH-Bereich ist neben dem Praxisbezug die Orientierung am neuesten Stand der Forschung in allen Teilbereichen. Dem wird bei der skizzierten Lehrveranstaltung erstens durch die Forschung einzelner Mitglieder des Digitalisierungszentrums als die Veranstaltung verantwortende Abteilung Rechnung getragen (1. Reul/Wick/Springmann/Puppe: 2017. 2. Springmann: 2016. 459–462). Zweitens ist die enge Zusammenarbeit des Digitalisierungszentrums mit dem Lehrstuhl für Informatik VI der Universität Würzburg zu nennen.

Denkbare Erweiterungen für eine zukünftig breiter angelegte Lehrveranstaltung der beschriebenen Art sind die Vorstellung a) des OCR von Handschriften, beispielsweise in der Kooperation mit einer Transkribus anwendenden Institution, idealerweise der *Digitalisierung und elektronische Archivierung – DEA* der Universität Innsbruck, und b) des Einsatzes virtueller Forschungsumgebungen zur Herstellung digitaler Ressourcen.

Aus Sicht des Autors ist nach der erfolgreichen Durchführung die kritische Diskussion mit den Autors ähnlicher oder gleicher Veranstaltungen von anderen Institutionen unverzichtbar. Ausgehend von http://cceh.uni-koeln.de/digitale-geisteswissenschaften-studiengange-2011/ befragt er aktuell die Mitarbeiter einschlägiger Institutionen nach deren Angeboten im Bereich der Digitalisierung von Texten. Er hofft mit seinem Beitrag wie dem AG Treffen an der DHd2018 auf einen fruchtbaren Meinungsaustausch.

Bibliographie

Corbach, Almuth: *Bestandsschonendes Digitalisieren von schriftlichem Kulturgut*. In: Digital und analog. Die beiden Archivwelten. 46. Rheinischer Archivtag. Ratingen 21.-22. Juni 2012.

Jannidis, Fotis / Hubertus Kohle / Malte Rehbein [Hrsg.]: *Digital Humanities. Eine Einführung*. Springer-Verlag GmbH Deutschland, 2017.

Kneißl, Michael: *Scannen wie die Profis : Text- und Bildvorlagen perfekt digitalisieren*. München: DTV. (2)2002.

Loewenheim, Ulrich / Adolf Dietz / Gerhard Schricker: *Urheberrecht*. Kommentar. München: Beck. (4)2010.

Reul, Christian / Christoph Wick / Uwe Springmann / Frank Puppe: *Transfer Learning for OCRopus Model Training on Early Printed Books*. In: *Zeitschrift für Bibliothekskultur*. Bd. 5, Nr. 1 (2017).

Springmann, Uwe: *A high accuracy OCR method to convert early printings into digital text. A Tutorial. Center for Information and Language Processing (CIS)*. LMU. München. 2015. S. 9.

Springmann, Uwe: *OCR für alte Drucke. Informatik-Spektrum*. 39(6):459–462. 2016.

Vogeler, Georg: *Die Text Encoding Initiative (TEI) als Werkzeug des Urkundeneditors – Erfahrungen und Desiderate*. In: Fees, Irmgard Prof. Dr.; Hotz, Benedikt; Schönfeld, Benjamin (Hrsg.): *Papsturkundenforschung zwischen internationaler Vernetzung und Digitalisierung. Neue Zugangs-*

weisen zur europäischen Schriftgeschichte. Göttingen. 2015.

Weitzmann, John H. / Paul Klimpel: Rechtliche Rahmenbedingungen für Digitalisierungsprojekte von Gedächtnisinstitutionen. Berlin: Zuse Institute Berlin. digiS – Servicestelle Digitalisierung Berlin. (3)2016.

Projektvorstellung – Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse

Brunner, Annelen
brunner@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Engelberg, Stefan
engelberg@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

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

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einführung

Das laufende DFG-Projekt „Redewiedergabe“ stellt einen Anwendungsfall quantitativer Sprach- und Literaturwissenschaft dar und beschäftigt sich mit dem Phänomen „Redewiedergabe“ auf der Grundlage großer Datenmengen. Zu diesem Zweck wird zum einen ein Korpus manuell mit Redewiedergabeformen annotiert, zum anderen werden Verfahren zur automatischen Erkennung des Phänomens entwickelt. Ziel ist es, Forschungsfragen nach der Entwicklung von Redewiedergabe vor allem im 19. Jahrhundert zu beantworten.

Das Poster präsentiert einen Überblick über das Gesamtprojekt sowie erste Projektergebnisse.

Stand der Forschung

Sowohl aus linguistischer als auch aus literaturwissenschaftlicher Perspektive ist Redewiedergabe ein interessantes Phänomen. Die Art und Weise, wie die Figurenstimme in die Erzählung eingebunden ist, steht in engem Zusammenhang mit Erzählweise und -haltung, sowie der Konstruktion der erzählten Welt. Folglich wird dem Phänomen in der Erzählforschung viel Aufmerksamkeit geschenkt und es liegen zahlreiche systematische Analysen vor (vgl. z.B. Genette 1998; Martínez / Scheffel 2007). Zu Phänomenen wie der erlebten Rede, dem Bewusstseinsstrom usw. gibt es eine umfangreiche Spezialforschung (Überblick bei McHale 2014). Aus linguistischer Perspektive ist Redewiedergabe vor allem in Bezug auf den Funktionswandel des Konjunktivs im Zusammenhang mit seinem Auftreten in indirekter Rede untersucht worden (vgl. z.B. Übersicht in Ágel 2000). In geringem Umfang sind auch Redewiedergabeverben und ihr Verhältnis zur wiedergegebenen Rede in das Blickfeld der Forschung gerückt (eine kurze Synopse bei Fritz 2005).

Ein Vorbild für die ausführliche, manuelle Annotation von Redewiedergabe ist v.a. Semino / Short 2004. Implementierungen der automatischen Erkennung stammen vor allem aus dem Bereich der Computerlinguistik und werden oft als Vorverarbeitungsschritt für andere Anwendungen durchgeführt (z.B. Wissensextraktion, Sprechererkennung oder dem Aufbau von sozialen Netzwerken literarischer Figuren, vgl. z.B. Krestel / Bergler / Witte 2008; Elson / Dames / McKeown 2010; Iosif / Mishra 2014). Eine literaturwissenschaftlich motivierte Anwendung ist die Untersuchung von Schöch et al. 2016 zur Erkennung von direkter Wiedergabe in französischen Romantexten. Die wichtigste Vorarbeit für das vorgestellte Projekt ist die Studie Brunner 2015, auf deren Ergebnissen es aufbaut. In dieser Studie wurde ein Korpus von 13 Erzähltexten manuell annotiert und Prototypen für die automatische Erkennung (sowohl regelbasiert als auch mit Hilfe von maschinellem Lernen) wurden entwickelt und ausgewertet.

Datengrundlage, Methodik und Ziele

Das Untersuchungskorpus umfasst die Jahre 1840-1920 und enthält sowohl fiktionale als

auch nicht-fiktionale Texte. Der nicht-fiktionale Teil setzt sich zusammen aus Texten des „Mannheimer Korpus Historischer Zeitungen und Zeitschriften“ und der Zeitschrift „Die Grenzboten“ (digitalisiert durch die Staats- und Universitätsbibliothek Bremen), der fiktionale Teil aus Erzählungen der Sammlung der Digitalen Bibliothek (textgrid). So sind sowohl Beobachtungen von Entwicklungen über die Zeit hinweg als auch Vergleiche zwischen Textsorten möglich.

Auszüge aus den Texten werden manuell annotiert. Das in Brunner 2015 vorgestellte und an Kategoriensystemen der Literaturwissenschaft orientierte Annotationssystem wurde für das Projekt erweitert und präzisiert. Es unterscheidet zwischen Wiedergabe von gesprochener Sprache, von Schrift und von Gedanken sowie den Typen direkte Wiedergabe (*Er sagte: "Ich bin hungrig."*), indirekte Wiedergabe (*Er sagte, er sei hungrig.*), erzählte Wiedergabe (*Er sprach über das Mittagessen.*) und freie indirekte Wiedergabe ('erlebte Rede') (*Wo sollte er jetzt nur etwas zu essen bekommen?*). Attribute spezifizieren die Annotation (z.B. Verschachtelungstiefe) und markieren Sonderfälle (z.B. nicht-faktische Wiedergabe). Zusätzlich werden Sprecher, Rahmenformeln und die redeeinleitenden Verben bzw. Nomen markiert.

Die Annotatoren arbeiten mit dem im Projekt Kallimachos (www.kallimachos.de) von Markus Krug entwickelten Eclipse-basierten Annotationswerkzeug ATHEN (https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen), für welches eine spezielle Annotationsoberfläche für Redewiedergabeformen implementiert wurde. Es werden, zumindest in Teilen, Mehrfachannotationen durchgeführt und Annotatorenvergleiche angestellt.

Die zweite Projektphase, welche zum Einreichungszeitpunkt dieses Posters gerade beginnt, umfasst die Entwicklung eines automatischen Erkenners für Redewiedergabeformen. Hierbei dient das manuell annotierte Material als Test- und Trainingsmaterial. Die in Brunner 2015 implementierten Prototypen dienen als Ausgangspunkt, die Implementierung erfolgt unter Nutzung des UIMA-Frameworks sowie in Python. Geplant ist eine Verbesserung des maschinellen Lernens durch Optimierung der Attributauswahl sowie Tests mit verschiedenen Lernalgorithmen (RandomForest, SVM, eventuell Conditional Random Fields und Deep Learning) und verschiedenen Parametereinstellungen. Auch regelbasierte Ansätze sollen weiter verfolgt werden, eventuell auf Grundlage einer aufwendigeren Vorverarbeitung (z.B. Parsing). Zudem ist eine Ergänzung und Verfeinerung einer Liste von Wörtern geplant, die auf Redewiedergabe hinweisen, welche sich bereits in Brunner 2015 als wertvolles Werkzeug bei der automatischen Erkennung erwiesen hat.

Der Redewiedergabe-Erkenner wird dann auf weitere Texte in unserem Untersuchungszeitraum angewendet, um größere Entwicklungslinien beobachten zu können und verschiedene offene narratologische und linguistische Forschungsfragen auf quantitativer Basis zu untersuchen, z.B.: Welche Entwicklungen in der Verwendung und Form von Redewiedergabe lassen sich im Untersuchungszeitraum beobachten? Welche Rolle spielen Textsortenunterschiede bei der Entwicklung von Redewiedergabeformen? Wie kommt die Dynamik im Bestand an Verben zustande, die als Redeeinleiter gebraucht werden?

Sowohl das manuell annotierte Korpus als auch der automatische Erkenner werden am Ende des Projekts der Forschungsgemeinschaft zur Verfügung gestellt. Es werden dafür sowohl das CLARIN-D-Forschungsdatenrepositorium des Instituts für Deutsche Sprache als auch das DARIAH-DE-Repository genutzt.

Bibliographie

Ágel, Vilmos (2000): "Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts", in: Besch, Werner / Betten, Anne / Reichmann, Oskar (eds.): *Sprachgeschichte.* Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Berlin / Boston: de Gruyter 1855-1903.

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe.* Ein Beitrag zur quantitativen Narratologie (= Narratologia 47). Berlin / Boston: de Gruyter.

Elson, David K. / Dames, Nicholas J. / McKeown, Kathleen (2010): "Extracting Social Networks from Literary Fiction", in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* 138-147.

Fritz, Gerd (2005): *Einführung in die historische Semantik.* Tübingen: M. Niemeyer.

Genette, Gérard (1998): *Die Erzählung.* München: Wilhelm Fink.

Iosif, Elias / Mishra, Taniya (2014): "From Speaker Identification to Affective Analysis: A Multi-Step System from Analyzing Children's Stories", in: *Proceedings of the Third Workshop on Computational Linguistics for Literature* 40-49.

Krestel, Ralf / Bergler, Sabine / Witte, René (2008): "Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles", in: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)* 2823-2828.

Martínez, Matías / Scheffel, Michael (2007): *Einführung in die Erzähltheorie.* 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 18. September 2017].

Schöch, Christof / Schlör, Daniel / Popp, Stefanie / Brunner, Annelen / Henny, Ulrike / Calvo Tello, José (2016): "Straight talk! Automatic Recognition of Direct Speech in Nineteenth-century French Novels", in: *Conference Abstracts. Jagiellonian University & Pedagogical University* 346-353.

Semino, Elena / Short, Mick (2004): *Corpus stylistics.* Speech, writing and thought presentation in a corpus of English writing. London / New York: Routledge.

Schlüsseldokumente zur deutsch-jüdischen Geschichte: Eine digitale Edition des Instituts für die Geschichte der deutschen Juden

Burckhardt, Daniel

burckhardtd@geschichte.hu-berlin.de
Institut für die Geschichte der deutschen Juden (IGdJ), Deutschland

Menny, Anna

anna.menny@igdj-hh.de
Institut für die Geschichte der deutschen Juden (IGdJ), Deutschland

Die vom Institut für die Geschichte der deutschen Juden (IGdJ) erstellte, seit Juli 2015 von der DFG geförderte und seit September 2016 frei zugängliche zweisprachige (deutsch/englisch) Online-Quellenedition „Hamburger Schlüsseldokumente zur deutsch-jüdischen Geschichte“ (http://juedische-geschichte-online.net/) wirft am Beispiel von derzeit etwa 75 ausgewählten Quellen thematische Schlaglichter auf zentrale Aspekte der deutsch-jüdischen Geschichte Hamburgs.

Mit der Auswahl und Digitalisierung von Text-, Bild-, Ton- und Sachquellen, die exemplarisch Einblick in historische Zusammenhänge und Ereignisse von der frühen Neuzeit bis in die Gegenwart bieten – den sog. Schlüsseldokumenten – führt sie das aufgrund von Vertreibung und Migration verstreute jüdische Erbe der Stadt digital wieder zusammen und trägt zu seiner langfristigen Sicherung für zukünftige Generationen bei. Ziel ist dabei, das Digitale nicht nur als ein weiteres Medium zu begreifen, sondern als einen Werkzeugkasten, mit dem das Material auf unterschiedlichen Ebenen bearbeitet werden kann. Zum einen führt die Digitalisierung selbst zur besseren Zugänglichkeit und nachhaltigen Sicherung, zum anderen erlauben die technische Auszeichnung nach TEI und Verknüpfung der bereitgestellten Materialien die Auswertung bislang nicht systematisch erfasster Informationen. Und schließlich bietet eine digitale Publikationsumgebung die Möglichkeit, neben Textquellen Bild-, Ton- und Videodokumente (sowie zukünftig 3D-Repräsentationen von Objekten) einzubinden und damit in den Geschichtswissenschaften bislang eher stiefmütterlich behandelte Quellengattungen verstärkt in den Blick zu nehmen.

Von diesen Überlegungen ausgehend, bilden die digitalisierten und technisch aufbereiteten Quellen konsequenterweise den Dreh- und Angelpunkt der Edition, die zugleich so strukturiert ist, dass sie hypertextuell angelegt und modular aufgebaut ist. Dass die Auseinandersetzung über konkrete Deutungen und Einordnungen am Beispiel konkreter Dokumente erfolgt und diese zugleich neuartig aufbereitet präsentiert werden, erlaubt ihre Fruchtbarmachung für neue Fragestellungen und kann Impulse für die deutsch-jüdische Geschichte geben. Alle Quellen werden als Transkript und digitales Faksimile bereitgestellt. Sowohl Quellen als auch Interpretationstexte werden nach TEI P5 gemäß dem Basisformat des Deutschen Textarchivs ausgezeichnet (Haaf et al 2015). Dies erlaubt neben der Auszeichnung der grundlegende Textstruktur die Verknüpfung mit Normdaten (Personen, Institutionen, Orte) sowie eine interne Verlinkungen mit weiteren Quellen, um so den Texten eine zweite Informationsebene einzuschreiben. Zugleich werden alle bereitgestellten Materialien mit einer DOI versehen, die die langfristige Referenzierbarkeit sicherstellt und die Bearbeitungshistorie transparent werden lässt.

Da die Digitalisierung und Online-Stellung von Quellen jedoch auch immer ein Herauslösen aus dem Überlieferungszusammenhang bedeutet und damit mit einer Entkontextualisierung und Entmaterialiserung verbunden ist, wird bei dieser Edition Wert darauf gelegt, neben der Bereitstellung der digitalisierten Quelle, diese durch begleitende Interpretations- und Hintergrundtexte

verstärkt in ihre historischen Kontexte einzubetten und zusätzliche Informationen zur Überlieferung, Rezeptionsgeschichte und zu wissenschaftlichen Kontroversen bereitzustellen.

Indem für die Digitalisierung, Textauszeichnung und Metadatenerschließung auf existierende Standardformate digitaler Editionen und der Langfristarchivierung wie MODS (Katalogdaten), METS (Digitalisate), TEI (Textauszeichnung der Transkriptionen und Übersetzungen), DOI (persistente Adressierung) sowie GND-Beacon-Dateien zurückgegriffen wird und bestehende Werkzeuge (Oxygen XML Editor) und technische Infrastrukturen (MyCoRe, Zotero) nachgenutzt werden, zugleich aber die Nutzerfreundlichkeit und Bedienbarkeit im Vordergrund steht, wurde eine innovative digitale Quellenedition zur jüdischen Geschichte Hamburgs geschaffen, die das Digitale als eine Möglichkeit ansieht, analoge Quelle neuartig zu präsentieren und mit weiteren (Informations-)schichten anzureichern und damit neue Impulse für die Forschung zu geben. Der Quellcode der Webanwendung ist für andere Projekte frei nachnutzbar (https://github.com/burki/jewish-history-online).

Neben der Auswahl und Aufbereitung des ausgewählten heterogenen Quellenmaterials zeichnet sich das Angebot durch umfassende Recherchemöglichkeiten (Karte, Zeitstrahl, Themen) sowie eine attraktive Präsentationsform aus. Auf diese Weise werden digitale Edierungstechniken für Quellen zur jüdischen Geschichte erprobt und einer breiten Öffentlichkeit vorgestellt. Die systematische Auszeichnung von Personen, Organisationen und Orte mit Normdaten ermöglicht die bidirektionale Verknüpfung der Edition mit externen Angeboten. So können ergänzende Informationen aus Linked Data Services wie denen der DNB und von Getty automatisiert ergänzt werden. Umgekehrt ermöglicht die Generierung von eigenen GND-Beacon-Listen externen Anbietern eine einfache Verknüpfung ihrer Angebote mit den entsprechenden Inhalten im Quellenportal.

Unser Poster zeigt die zentralen Eigenschaften der Online-Edition und hilft, den konzeptionellen Rahmen sowie die technische Umsetzung zu verstehen. Es illustriert die Verknüpfung zwischen TEI-Kodierung der Dokumente sowie ihrer Präsentation und Navigation. Es soll damit den konzeptionellen Grundgedanken des Projektes veranschaulichen, das Digitale mit seinen Möglichkeiten ernst zu nehmen, jedoch nicht in Konkurrenz, sondern in Ergänzung zum Analogen.

Bibliographie

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2015): "The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources", in: Journal of the Text Encoding Initiative 8, December 2014 - December 2015, http://journals.openedition.org/jtei/1114 [letzter Zugriff 10. Januar 2018].

Science as a Service? Chancen und Limits von serviceorientierten Softwarearchitekturen für die Digital Humanities

Hoffmann, Christoph
christoph.hoffmann@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Einer der am weitesten verbreiteten Ansprüche von Digital Humanities Projekten und Forschungsvorhaben ist jener, nachhaltig nutzbare Daten und Services zu produzieren bzw. zu hinterlassen. Neben archivierbaren Datenformaten und quelloffener Software ist die Einrichtung von REST – APIs eine vielgenutzte Möglichkeit, erstellte Services in anderen Projekten nutzbar zu machen.[1] Der Vorteil solcher Schnittstellen liegt darin, dass die ihnen zugrunde liegenden Technologie (im wesentlichen HTTP) in so gut wie allen Plattformen und Programmiersprachen bereits implementiert ist. So sind sowohl Services als auch Daten welche über eine sauber definierte REST – Schnittstelle zugänglich sind, relativ einfach in einer Vielzahl anderer Projekte zu integrieren. In vielen Bereichen der Software – Entwicklung hat dieser Vorteil dazu geführt dass Service orientierte Softwarearchitektur und ein „API first approach“ immer populärer geworden sind.

Auch das Austrian Center for Digital Humanities an der Österreichischen Akademie der Wissenschaften hat es sich für die kommenden zwei Jahre zu einer Aufgabe gemacht, die in den vergangenen drei Jahren in einer Vielzahl von Projekten entstandenen Daten und Services in stan-

dardisierten REST – Schnittstellen verfügbar zu machen bzw. bereits entstandene Schnittstellen entsprechend aufzubereiten und zu dokumentieren. Hierzu sollen zunächst wiederkehrende, generisch abbildbare Aufgaben in den Workflows identifiziert werden, sodann die Ihnen korrespondierenden, bereits bestehenden, Services für eine Verwendung außerhalb des Projektkontextes abgeändert werden.

Ziel dieses Vorhabens ist es, einen Katalog an REST-Services zu erstellen, welcher in einer Sandbox zum einen die direkte Verwendung erlaubt, zum anderen die Funktionen der verschiedenen Endpunkte unmittelbar kritisch dokumentiert. Des Weiteren sollen bereits vorhandene Standards[2] und semantische Erweiterungen [3] auf Ihre Tauglichkeit und Sinnhaftigkeit in Digital Humanities Kontexten hin geprüft, und gegebenenfalls in den Schnittstellen implementiert werden werden. Exemplarisch soll auch modular wiederverwendbare Front – End Komponenten angeboten werden (bspw. Autocompletes u.a.)

Parallel dazu gilt es, anhand der verwendeten Projekte kritisch zu reflektieren, welche Schritte, Funktionalitäten und Daten sich aus einem vorhandenen Projekt sinnvoll zur Nachnutzung herauslösen lassen. Wie kleinteilig lassen sich die Schritte eines geisteswissenschaftlichen Projektes modularisieren und an externe Services auslagern, wenn ein zur Kritik der Resultate fähiger Gesamtblick gewahrt bleiben soll? Wie exakt muss, umgekehrt gefragt, eine Schnittstelle dokumentiert sein, um einer kritischen Prüfung im Rahmen des verwendenden Projektes standhalten zu können? Diese und andere Betrachtungen sollen schlussendlich in einem White Paper zu REST – APIs für wiederkehrende Szenarien in Digital Humanities Projekten münden.

Das Poster soll zum einen den Katalog und die Sandbox als Tools präsentieren, zum anderen erste Ergebnisse der Standardisierung von APIs zur Diskussion stellen. Eine Debatte zu den oben genannten Fragen soll dem Vorhaben eine noch breitere Grundlage bei der Erstellung des White Papers geben.

Fußnoten

1. http://digitalhumanities.berkeley.edu/blog/15/05/01/project-sustainability-dh-collaboration-and-community - 14.1.2018
2. bspw. https://github.com/OAI/OpenAPI-Specification oder http://jsonapi.org/ - 14.1.2018
3. http://www.hydra-cg.com/spec/latest/core/ und http://microformats.org/ - 14.1.2018

Sechs Wege der FRBRisierung von Textverknüpfungen

Helling, Patrick
patrick.helling@uni-koeln.de
Universität zu Köln, Deutschland

Mathiak, Brigitte
mathiak@gmail.com
Universität zu Köln, Deutschland

Einleitung

Linked Open Data wird auch in den Geisteswissenschaften immer wichtiger (Barbera 2013: 91 – 105). Dabei geht es sehr oft um die Verknüpfung von Text mit weiterem Text, z.B. mit Kommentaren, Referenzen auf andere Werke oder anderen Entitäten wie Personen, Orte, etc. Mit Ontologien wie den Functional Requirements for Bibliographic Records object-oriented (FRBRoo) (Bekiari et al. 2015) können zwar bibliographische Gegenstände im Rahmen einer digitalen Datenbank auf eine Art und Weise modelliert und bereitgestellt werden, dass diese allein durch ihre Organisation und objektübergreifende Strukturierung und Verknüpfung bibliographisch-informationellen Mehrwert im Rahmen der Recherche erzeugen können (Förster / Becker 2010: 15 - 25). Allerdings ist nicht immer klar, wie Verknüpfungen aus den Texten heraus mit anderen Texten oder aber zu komplett anderen Entitäten zu modellieren sind. Die Verknüpfung von verschiedensten Dokumenten und Entitäten ist allerdings eine der Hauptideen bei der Benutzung von Ontologien und Linked Open Data.

Einige Forschungsprojekte haben sich dieses Problems in Spezialfällen bereits angenommen. So wird bei HuCit (Romanello / Pasin 2011: 216 - 218) betrachtet, wie kanonische Zitation, z.B. der klassischen Literatur, modelliert werden kann. In (Bartalesi / Meghini 2016: 385 - 394) wird eine Ontologie speziell für die Texte von Dante Alighieri entwickelt. In (Mathiak / Boland 2015) wird der Spezialfall einer Verknüpfung zwischen Texten und Datensätzen, auf denen diese Texte beruhen, betrachtet.

Wir nehmen die zuvor genannten Ansätze als Ausgangspunkt, systematisieren diese und arbeiten Vor- und Nachteile heraus. Ziel ist es ein Modell für textbasierte Materialien zu konzipieren,

mit dem Text(-stellen) mit anderen Textstellen oder Entitäten verknüpft und modelliert werden können.

Sechs Wege

1.) Die direkte Verbindung
In diesem einfachen Fall gibt es einfach eine Verbindung zwischen dem Dokument und der Entität, die mit diesem verknüpft werden soll. Die Property wird dazu idealerweise aus einer bereits etablierten Ontologie gewählt (siehe Abb. 1). Das klassische Beispiel hierfür ist eine Verschlagwortung oder die Zuordnung zu Autoren. Obwohl diese Art der Modellierung sehr einfach ist, hat sie doch starke Einschränkungen. Es bleibt unklar, welcher Teil des Dokuments für die Verbindung verantwortlich ist und es ist schwierig die Art der Verbindung über die Property hinaus zu beschreiben.

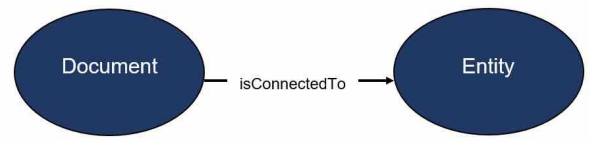


Abb. 1: Die direkte Verbindung.

2.) Die Verbindung über die Textstelle
Bei dieser Art der Modellierung wird zunächst die Textstelle innerhalb des Dokuments identifiziert und dann im nächsten Schritt mit der Entität verbunden, wodurch im Gegensatz zu Variante 1 eindeutig definiert wird, welcher Teil des Dokuments für die Verbindung verantwortlich ist (siehe Abb. 2).

Abb. 2: Die Verbindung über die Textstelle.

3.) Die Verknüpfung als eigene Entität
Dabei wird das Dokument bzw. die Textstelle zunächst mit einem Verknüpfungsobjekt verbunden und dann dieses mit der Zielentität (siehe Abb. 3), wobei im Vergleich zu Variante 1 die Erfassung relationsbeschreibender Informationen möglich wird. Diese Art der Modellierung ist vor allem nützlich, wenn es sehr viele Zusatzinformationen zu der Verknüpfung selbst gibt.

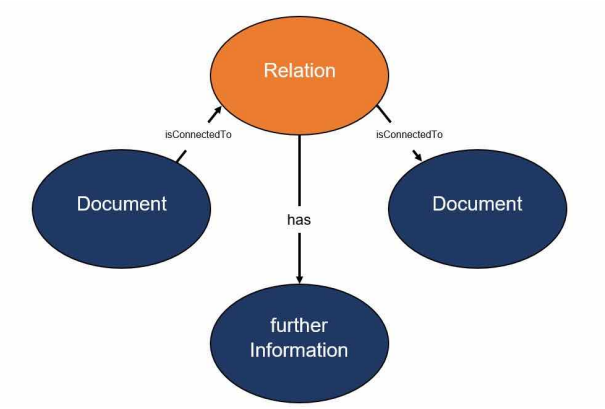


Abb. 3: Die Verknüpfung als eigene Entität.

4.) Die Verknüpfung mit einer Stellvertreterentität
Manchmal ist es nicht eindeutig möglich das Ziel der Verknüpfung als Entität zu identifizieren. In diesem Fall kann es sinnvoll sein, zunächst ein fiktives Ziel zu definieren und dann dessen Beziehungen zu bereits bekannten Entitäten zu etablieren (siehe Abb. 4). In gewissem Sinne ist dies invers zu Methode 2 zu sehen, bei der ja auch zunächst das Objekt der Textstelle neu erschaffen wird, indem es zu seinem Quelldokument in Beziehung gesetzt wird. In (Mathiak / Boland 2015) wird dies benutzt um fehlende Informationen zur Zielentität zu modellieren.

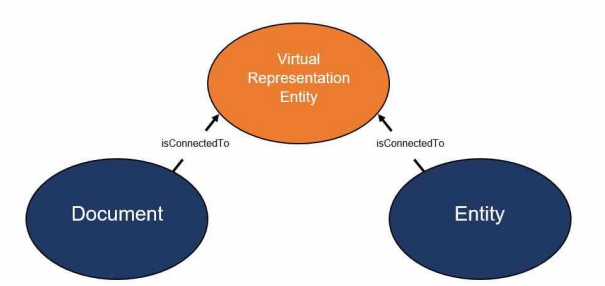


Abb. 4: Die Verknüpfung mit einer Stellvertreterentität.

5.) Das unabhängige Netzwerk
In vielen Fällen ist der Textbezug für die Ergebnisdarstellung nur von sekundärer Bedeutung, stattdessen werden die referenzierten Entitäten direkt miteinander in Verbindung gesetzt (siehe Abb. 5). Als Beispiel seien Beziehungsnetzwerke von Protagonisten in literarischen Werken genannt. Die Entitäten sind zwar auch mit dem Dokument verknüpft in dem sie auftauchen, aber die Kerninformation liegt in den Beziehungen, die diese untereinander haben.

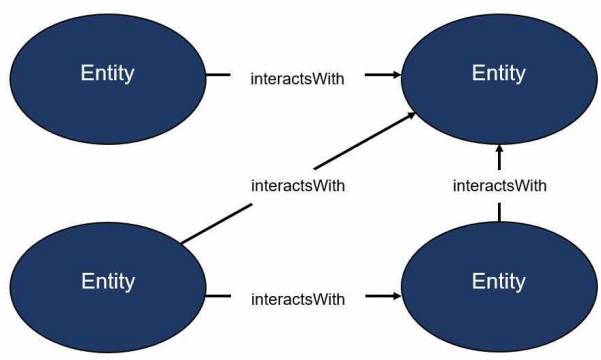


Abb. 5: Das unabhängige Netzwerk.

6.) Multiple Dokumente
Verschiedene Versionen desselben Dokuments werden in typischen Ontologien bisher wenig betrachtet, obwohl sie in den Digital Humanities eine häufige Modellierungsherausforderung sind.
Poster-Präsentation
Auf dem Poster präsentieren wir die grundlegende Ontologie-Struktur und stellen die verschiedenen modularen Modellierungsstrategien genauer vor. Zur Veranschaulichung werden wir eine Umsetzung unserer ontologischen Konzepte an einem Beispiel aus den Geisteswissenschaften demonstrieren.

Bibliographie

Barbera, Michele (2013): "Linked (open) data at web scale: research, social and engineering challenges in the digital humanities" *in: Global Interoperability and Linked Data in Libraries: Special issue. JLIS.it, Vol. 4, No. 1*: 91 – 105 http://dx.doi.org/10.4403/jlis.it-6333 .

Bartalesi, Valentina / Meghini, Carlo (2016): „Using an Ontology for Representing Knowledge on Literary Texts: the Dante Alighieri Case Study " *in: Semantic Web, Volume 8, Number 3, 6. Dezember 2016*: 385 – 394 http://www.semantic-web-journal.net/content/using-ontology-representing-knowledge-literary-texts-dante-alighieri-case-study-0 [letzter Zugriff 22. September 2017].

Bekiari, Chryssoula / Doerr, Martin / Le Boeuf, Patrick / Riva, Pat (2015): „Definition of FRBRoo. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism " Den Haag: https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf [letzter Zugriff: 25. September 2017].

Förster, Frank / Becker, Hans-Georg (2010): „Vernetztes Wissen – Ereignisse in der bibliographischen Dokumentation" *in: Zeitschrift für Bibliothekswesen und Bibliographie.* 57. Jahrgang/Heft Nr. 1: 15 – 25 http://dx.doi.org/10.3196/186429501057133 .

Mathiak, Brigitte / Boland, Katarina (2015): „Challenges in Matching Dataset Citation Strings to Datasets in Social Science " *in: D-Lib Magazine, Volume 21, Number 1/2, Januar/Februar 2015*, https://doi.org/10.1045/january2015-mathiak .

Romanello, Matteo / Pasin, Michele (2011): „An Ontological View of Canonical Citations " *in: DH 2011 Book of Abstracs. Stanford*: Stanford University Library 216 - 218 http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-143.xml [letzter Zugriff 22. September 2017].

Semantische Extraktion auf antiken Schriften am Beispiel von Keilschriftsprachen mithilfe semantischer Wörterbücher

Homburg, Timo
timo.homburg@gmx.de
Hochschule Mainz, Deutschland

Einleitung und Motivation

Semantische Extraktionsmechanismen (z.B. Topic Modelling) werden seit vielen Jahren im Bereich des Semantic Web und Natural Language Processings sowie in den Digital Humanities als Verfahren zur Visualisierung und automatischen Kategorisierung von Dokumenten eingesetzt. Oft ergeben sich durch den Einsatz neue Aspekte der Interpretation von Dokumentensammlungen die vorher noch nicht ersichtlich waren. Als Beispiele solcher Verfahren kommen häufig Machine Learning Algorithmen zum Einsatz, welche eine Grobeinordnung von Texten vornehmen können. Gepaart mit Metadaten von Texten können anschließend beispielsweise thematische Übersichten von Dokumenten mit geographischem Bezug auf Kartenmaterialien in GIS Systemen oder mittels historischer Gazetteers zeitliche Zusammenhänge automatisiert dargestellt werden. In dieser Publikation möchten wir die Möglichkeiten der semantischen Extraktion nutzen und diese auf ei-

ner Sammlung von Texten in Keilschriftsprachen anwenden.

Keilschriftsprachen

Keilschriftsprachen haben in den letzten Jahren ein größeres Interesse in der Digital Humanities und Linguistik Community erfahren. (Inglese 2015, Homburg et. al. 2016, Homburg 2017, Sukhareva et. al. 2017). Neben der andauernden Standardisierung in Unicode werden unter anderem Part Of Speech Tagger und Mechanismen der automatisierten Übersetzung erprobt um Keilschrifttexte besser mit dem Computer zu erfassen und zu interpretieren. Desweiteren wurde die Erlernbarkeit der Keilschriftsprachen durch digitale Tools wie Eingabemethoden oder Karteikartenlernprogramme verbessert. (Homburg 2015) Trotz all der erreichten Fortschritte verbleiben jedoch zahlreiche Probleme bei der maschinellen Verarbeitung von Keilschriftsprachen, die unter anderem mit der geringen Verfügbarkeit annotierter Ressourcen und der fehlenden Verfügbarkeit maschinenlesbarer und semantisch sowie linguistisch annotierter Wörterbücher zusammenhängt. Diese Limitierungen hindern viele Natural Language Processing und semantische Extraktionsalgorithmen daran ein besseres Ergebnis zu erzielen. Wir möchten mit dieser Publikation einen Beitrag leisten diese Situation zu verbessern und stellen das "Semantic Dictionary for Ancient Languages" vor, welches ein Versuch ist durch Annotierung vorhandener in der Forschungscommunity anerkannter Wörterbuchressourcen mit Unicode Characters, Semantic Web Konzepten, etymologischen Daten, gemeinsamen Vokabularen und POSTags eine semantische Ressource in RDF für die Optimierung solcher Algorithmen auf Basis der Sprachen Hethitisch, Sumerisch und Akkadisch zu schaffen.Das Wörterbuch basiert auf dem Lemon-Standard, ein W3C Standard der es erlaubt ebenfalls multilinguale Resourcen abzubilden. So können Entwicklungen der Sprache und gemeinsame Vokabulare wie zum Beispiel Akkadogramme und Sumerogramme in Hethitisch mit erfasst werden.

Semantisches Wörterbuch und Semantische Extraktion

Wir testen die Performance des Wörterbuchs auf einer der größten Sammlungen von digitalen Keilschrifttexten, der CDLI, aus der wir repräsentative Texte in hethitischer, sumerischer und akkadischer Keilschrift aus verschiedenen Epochen extrahieren und mittels Machine Learning klassifizieren, sowie verschlagworten. Das Ergebnis der semantischen Extraktion ist eine Sammlung von Themen pro Keilschrifttafel, die sich wiederum in Überkategorien gruppieren lassen und in einen zeitlichen, sprachlichen, dialektischen, sowie örtlichen Kontext gestellt werden können. Anhand der verschiedenen Metadaten der CDLI war es uns möglich eine thematische Karte der Fundorte der Keilschrifttafeln sowie deren Inhalt pro Epoche darzustellen aus der das relevante Fachpublikum schließen kann welche Themen zu welcher Zeit an welchem Fundort relevant für die Schreiber der jeweiligen Epoche waren. Im Zuge einer Weiterentwicklung möchten wir diese Informationen mit weiteren Metadaten wie beispielsweise der Jurisdiktion, den Daten der jeweiligen Herrscher sowie rekonstruierten Orten aus der antiken Zeit vervollständigen um Rückschlüsse auf interessante historische Ereignisse zu ziehen.

Aufbau des Posters

Auf unserem Poster möchten wir gerne den Prozess des Aufbaus, sowie die Struktur des semantischen Wörterbuchs sowie die Karte die durch unsere semantische Extraktion entstanden ist präsentieren um die jeweiligen Fachwissenschaftler zur Diskussion über die Entwicklung eines Semantic Web von Keilschriftsprachen und Keilschriftartefakten einzuladen. Desweiteren soll unser Poster eine Reihe von Anwendungen demonstrieren die sich in Zukunft mit unserer semantischen Ressource entwickeln lassen können um einen Beitrag zu einem hoffentlich zukünftig existierenden LinkedData Datensatz für Keilschriftartefakte zur Dokumentation von Keilschrift zu leisten.

Bibliographie

Inglese, G. (2015): “Towards a hittite treebank. basic challenges and methodological remarks.” In: Corpus-Based Research in the Humanities (CRH) p. 59 1.1

Homburg, T. (2017): “Postagging and semantic dictionary creation for Hittite cuneiform.” In:DH2017 (2017)

Homburg, T., Chiarcos, C. (2016): “Word segmentation for Akkadian cuneiform.” In: LREC2016 1.1

Homburg, T., Chiarcos, C., Richter T., Wicke, D. (2015): “Learning Cuneiform the Modern Way.” In: DhD2015

Sukhareva, M., Fuscagni, F., Daxenberger, J., Görke, S., Prechel, D., Gurevych,I. (Aug 2017): “Distantly supervised pos tagging of low-resource languages under extreme data sparsity: The case of Hittite.” In: LaTeCH-CLfL ’17 Proceedings of the 11th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp.95–104 1.1

Stadtgeschichtliche Forschung und Vermittlung anhand historischer Fotos als Forschungsgegenstand – Ein Zwischenbericht der Nachwuchsgruppe HistStadt4D

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

Barthel, Kristina
kristina.barthel@tu-dresden.de
TU Dresden, Deutschland

Bruschke, Jonas
jonas.bruschke@uni-wuerzburg.de
JMU Würzburg, Deutschland

Friedrichs, Kristina
kristina.friedrichs@tu-dresden.de
JMU Würzburg, Deutschland

Kröber, Cindy
cindy.kroeber@tu-dresden.de
TU Dresden, Deutschland

Maywald, Ferdinand
ferdinand.maiwald@tu-dresden.de
TU Dresden, Deutschland

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

1. Einleitung

Historische Fotografien sowie Pläne sind eine wesentliche Quellengrundlage baugeschichtlicher Forschung (Burke, 2003, Paul, 2006, Wohlfeil, 1986, Pérez-Gómez and Pelletier, 1997) und ebenso wie diese zentrale Gegenstände der Digital Humanities (Kwastek, 2014). Im Zuge von Digitalisierungsvorhaben wurden eine Reihe digitaler Bildarchive errichtet und umfangreiche Fotografie- und Planquellen in derartige Repositorien überführt. Dabei stellt sich jedoch nicht nur die Schwierigkeit, für eine Erforschung relevante und aussagekräftige Quellen zu finden und zu identifizieren, sondern auch, diese zu untersuchen, zu kontextualisieren und zu vergleichen sowie die darin beschriebenen historischen Objekte vorstellbar zu machen. Die durch das BMBF geförderte eHumanities-Nachwuchsgruppe HistStadt4D adressiert anhand stadt- und baugeschichtlicher Forschungsfragen und Vermittlungsanliegen zur Historie der Stadt Dresden die Untersuchung und Entwicklung von methodischen und technologischen Ansätzen, umfangreiche Repositorien historischer Medien und Kontextinformationen räumlich dreidimensional sowie zeitlich zusammenzuführen, zu strukturieren und zu annotieren sowie diese für Wissenschaftler und Öffentlichkeit mittels eines 4D-Browsers sowie einer ortsabhängigen Augmented-Reality-Darstellung als Informationsbasis, Forschungswerkzeug und zur Vermittlung geschichtlichen Wissens nutzbar zu machen. Prototypische Datenbasis stellen dabei ca. 30.000 digitalisierte historische Fotografien und Pläne des historischen Dresdens dar.

2. Ergebnisse

Photogrammetrische Methoden zur Erschließung historischer Bilddaten

Um einen visuellen Zugang zu großen Bildrepositorien zu schaffen, können verschiedene photogrammetrische Methoden hilfreich sein. Die Spannweite der Verfahren reicht von der automatischen Bildsortierung in einer Datenbank mittels kontextbasierter Ansätze über die temporale und spatiale Verortung von Bildern in virtuellen Umgebungen bis zur Generierung komplexer historischer dreidimensionaler Modelle. Das Potential historischer Fotografien liegt hierbei vor allem in der Dokumentation aber auch in der Wiederherstellung von zerstörten Objekten (Grün et al., 2004, Falkingham et al., 2014). Seit ca. 20 Jahren wird die klassische analytische Photogrammetrie hierbei durch digitale Bildverarbei-

tungsverfahren unterstützt. Es ist außerdem ein zunehmender Automatisierungsgrad in diesen Technologien zu sehen. Anwendung finden diese Verfahren heute hauptsächlich in großen Bildrepositorien mit aktuellen Fotografien (Agarwal et al., 2009). Aber auch einzelne Gebäude/Städte werden mithilfe von historischen Materialien modelliert (Schindler & Dellaert, 2012, Bitelli et al., 2017). Schwierigkeiten, die in der Arbeit mit historischen Bildern entstehen, sind fehlende Informationen zu Kamera, Aufnahmeort und Aufnahmezeit. Auch Bildfehler, starke Bildunterschiede und geringe Scanauflösung wirken sich auf eine automatische photogrammetrische Mehrbildauswertung negativ aus. Vor diesem Hintergrund erwiesen sich die in der Nachwuchsgruppe erprobten Verfahren des Structure-from-Motion (SfM) und der Direct-Linear-Transformation (DLT) als hinsichtlich der Ergebnisse lückenhaft bzw. schlecht automatisierbar (vgl. Abb. 1). Während SfM als automatisches Verfahren prinzipiell sehr gut funktioniert, gelangt der Algorithmus beim Umgang mit heterogenen Bildmaterial wie eben in historischen Aufnahmen an Grenzen. Dann können zwischen den verschiedenen Bildern nur wenige oder gar keine Korrespondenzen gefunden werden und die Orientierung und die anschließende Erstellung einer dreidimensionalen Punktwolke schlägt fehl. DLT berechnet die Kameraorientierung aus homologen Punkten in Gebäudemodell und Fotografie. Eine automatische Bestimmung der homologen Punkte auf stark generalisierten Modellen und verrauschten Bildern ist eine komplexe Aufgabe, die bisher noch nicht algorithmisch gelöst wurde. Die Punkte werden deshalb noch manuell durch einen Operateur gesucht, wobei mit 6 Punktpaaren eine große Anzahl homologer Punkte benötigt wird. Einen demgegenüber vielversprechenden, aktuell in Prüfung befindlichen Ansatz bietet die Kongruenzprüfung von im Bild dargestellten Formen.

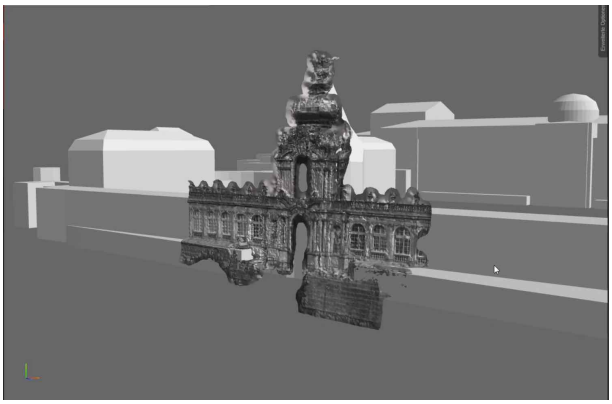


Abb. 1: Verortetes 3D-Modell generiert aus historischen Bildern (Prototyp)

Nutzung und Zugänge zu Bildrepositorien in der architekturgeschichtlichen Forschung

Digitale Bildrepositorien erfüllen ganz verschiedene Zwecke, sowohl innerhalb der geisteswissenschaftlichen Forschung als auch im Rahmen der musealen oder touristischen Vermittlung, bis hin zu informationswissenschaftlichen Aspekten (Münster, 2011). Die technischen Möglichkeiten solcher Repositorien erlauben Kunst- und Architekturhistorikern, eine deutlich größere Menge an Material in ihre Forschungen einzubeziehen. Wichtige Ergebnisse eines ersten Jahres Forschungstätigkeit ist ein Verzeichnis existierender Bildrepositorien und deren Bewertung mit Blick auf Anforderungen architekturhistorischer Arbeit: Wissenschaftler müssen zum Beispiel die Möglichkeit haben, ihre genutzten Quellen miteinander zu vergleichen und sie im Kontext zu verorten, (Münster et al., 2015, Brandt, 2012, Wohlfeil, 1986), aber auch, das Verhältnis von Bildquelle zu Abbild nachvollziehen zu können (Favro, 2006, Niccolucci and Hermon, 2006).

Mit Blick auf eine Nutzungspraxis von Bildrepositorien hängt der Erfolg bei der Suche und dem Zugang zu den Informationen stark von den Fähigkeiten der Nutzer sowie der Bedienbarkeit der Webplattformen ab (Kemman et al., 2014). Nutzer legen z.B. Wert auf effiziente Such- und Filterfunktionen sowie eine intuitiv bedienbare Nutzeroberfläche und Navigation (Barreau et al., 2014). Je nach Nutzerkreis ist eine Dokumentation mittels Metadaten (Bentkowska-Kafel et al., 2012, Maina and Suleman, 2015) oder eine grundlegende Einführung in die Themen und Informationen gewünscht (Maina and Suleman, 2015). Empirisch überprüft werden konnte dies mit 15 Studenten der Kunst- und Architekturgeschichte, welche im September 2016 im Rahmen einer Fokusgruppen-Diskussion zur Suche und Nutzung von Bildern befragt und inhaltsanalytisch ausgewertet werden. Wesentliche Erkenntnis stellten beispielsweise dar, dass ein Suchprozess neben den dargebotenen Bildinhalten stark durch zugeordnete textuelle Beschreibungen sowie Vorschläge und Links zu weiteren Ressourcen beeinflusst wird. Eine Suchstrategie verändert sich dabei im Laufe des Suchprozesses von (a) einem ungerichteten Stöbern über ein (b) Sichten von Beständen zur Gewinnung eines Überblicks hin zur (c) zielgerichteten Suche nach spezifischen Inhalten.

Die dabei gewonnenen Erkenntnisse bildeten die Grundlage für die Konzeption und prototypische Entwicklung eines 4D-Browserinterfaces für eine räumlich und zeitlich verortete Suche in Medienrepositorien (Abb. 2). Eine besondere Herausforderung einer ersten Entwicklungsphase stellten die semantische Verknüpfung der Daten und

die Visualisierung zeitlich und räumlich verorteter Informationen (Gouveia et al., 2015) dar. So muss die angestrebte 4D-Browseranwendung ein ganzes Stadtmodell handhaben können, das sich zudem über die Zeit stetig verändert. Mit Blick auf eine weitere Entwicklung stehen aktuell Strategien zur Präsentation einer Vielzahl von Datensätzen (Samuel et al., 2016), Schnittstellen zum automatischen Datenbezug aus Repositorien sowie Möglichkeiten zur nutzergesteuerten, interaktiven Informationsverknüpfung auf einer Entwicklungsagenda.

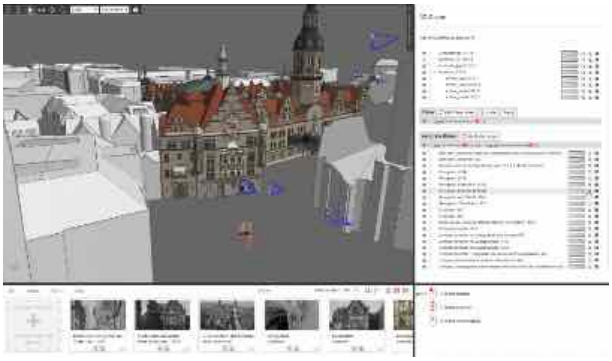


Abb. 2: 4D-Browser (Prototyp)

Vermittlung von Stadtgeschichte mittels Augmented-Reality (AR)

Eine räumliche und zeitliche Einordung in einem dreidimensionalen Stadtmodell (Abb. 3) sowie in einer Augmented Reality – Anwendung sollen hierbei im Vordergrund stehen und eine Beziehung zur heutigen stadträumlichen Situation vermitteln. Ein diesbezüglicher Forschungsgegenstand ist der Einsatz im Rahmen von Bildungsszenarien für die Vermittlung von Stadtgeschichte. In den letzten Jahren wurden weltweit zahlreiche touristische AR-Anwendungen für die stadträumliche Nutzung sowie Softwarelösungen zur Erstellung dieser Anwendungen entwickelt (Kounavis et al., 2012, Smirnov et al., 2014, Yovcheva et al., 2012). In wissenschaftlichen Studien wurden Nutzungsanforderungen touristischer AR-Anwendungen (Zaibon et al., Han et al., 2014), deren Akzeptanz beim Nutzer (Haugstvedt and Krogstie, 2012, tom Dieck and Jung, 2015), sowie deren Potenziale für das Lernen (Kysela and Štorková, 2015) oder für die Sensibilisierung der Nutzer für kulturhistorische Aspekte beschrieben (Chang et al., 2015b). Nur wenige wissenschaftliche Arbeiten beschäftigen sich demgegenüber mit Lerneffekten bei Nutzern (tom Dieck et al., 2016). Zudem fehlen Untersuchungen zur Integration von pädagogischen und motivationalen Strategien in wissensvermittelnden AR-Anwendungen und daraus resultierende Gestaltungsempfehlungen. Diese Lücke soll mit der weiteren Forschungsarbeit geschlossen werden.

Abb. 3: Stadträumliche Augmented-Reality-Darstellung (Mockup)

Aus technologischer Sicht wurden bisher in der Nachwuchsgruppe vor allem Modi der Interaktion mit geschichtswissenschaftlichen Daten erprobt und untersucht (vgl. Livingston et al., 2008, Zöllner et al., 2010, Walczak et al., 2011, Chang et al., 2015a, Chung et al., 2015). Dabei wesentliche Fragen waren: Wie können Interaktionsmöglichkeiten mit virtuellen Gebäuden und mit ihnen verknüpften Informationen gestaltet werden? Können aus dem Umgang mit Mobilgeräten bekannte Interaktionsmetaphern in der Augmented Reality weiterverwendet werden? Des Weiteren wurden mobile AR Anwendungen für die Vermittlung stadtgeschichtlichen Wissens anhand der eingesetzten Darstellungskonzepte analysiert und kategorisiert. Davon ausgehend soll in einem nächsten Schritt nun ein Anwendungsprototyp entwickelt und validiert werden.

3. Resümee

Welches Resümee lässt sich nach einem Jahr Forschungsarbeit in der Nachwuchsgruppe ziehen? Neben der vorgestellten Forschung in den jeweiligen Untersuchungsgebieten stellt insbesondere die Schaffung von Schnittstellen zwischen diesen Bereichen eine wichtige und kontinuierliche Aufgabe dar. Praktisch bewährt haben sich in unserer Arbeit beispielsweise Elemente wie eine gemeinsame Ergebnisvision, eine engmaschige Abstimmung und ein agiles Arbeitsvorgehen. Vor diesem Hintergrund soll der hier vorgeschlagene Beitrag nicht nur Forschungsergebnisse aufzeigen sondern auch Strategien zur interdisziplinären Gruppenarbeit vorstellen und diskutieren.

Bibliographie

Deutsche Fotothek [Online]. http://www.deutschefotothek.de/ [9.5.2014].

Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg [Online]. http://www.fotomarburg.de/ [9.5.2014].

Barreau, J.-B./ Gaugne, R./ Bernard, Y./ Le Cloirec, G. / Gouranton, V. (2014). Virtual reality tools for the West Digital Conservatory of Archaeological Heritage. *Proceedings of the 2014 Virtual Reality International Conference.*

Bentkowska-Kafel, A./ Denard, H. / Baker, D. (2012). *Paradata and Transparency in Virtual Heritage,* Burlington, Ashgate.

Brandt, A. V. (2012). *Werkzeug des Historikers,* Stuttgart [u. a.], Kohlhammer.

Burke, P. (2003). *Augenzeugenschaft. Bilder als historische Quellen,* Berlin.

Chang, Y.-L./ Hou, H.-T./ Pan, C.-Y./ Sung, Y.-T. / Chang, K.-E. (2015a). Apply an Augmented Reality in a Mobile Guidance to Increase Sense of Place for Heritage Places. *Educational Technology & Society,* 18, 166-178.

Chang, Y.-L./ Hou, H./ Pan, C./ Sung, Y. / Chang, K. (2015b). Apply an Augmented Reality in a Mobile Guidance to Increase Sense of Place for Heritage Places. *J. Educ. Technol. Soc.,* 18, 166–178.

Chung, N./ Han, H. / Joun, Y. (2015). Tourists' intention to visit a destination: The role of augmented reality (AR) application for a heritage site. *Computers in Human Behavior,* 50, 588-599.

Favro, D. (2006). In the eyes of the beholder. Virtual Reality re-creations and academia. *In:* HASELBERGER, L., HUMPHREY, J. & ABERNATHY, D. (eds.) *Imaging ancient Rome: Documentation, visualization, imagination: Proceedings of the 3rd Williams Symposium on Classical Architecture, Rome, 20.- 23. 5. 2004.* Portsmouth: Journal of Roman Archaeology.

Gouveia, J./ Branco, F./ Rodrigues, A. / Correia, N. (2015). Travelling Through Space and Time in Lisbon's Religious Buildings. *In:* GUIDI, G., SCOPIGNO, R., TORRES, J. C. & GRAF, H. (eds.) *2nd International Congress on Digital Heritage 2015.* Granada.

Han, D./ Jung, T. / Gibson, A. (2014). Dublin AR: Implementing Augmented Reality in Tourism. *In:* Z., X. & I., T. (eds.) *Information and Communication Technologies in Tourism.* Cham: Springer.

Haugstvedt, A.-C. / Krogstie, J. (Year) Published. Mobile augmented reality for cultural heritage: A technology acceptance study. 2012 IEEE Int. Symp. Mix. Augment. Real., 2012.

Kemman, M./ Kleppe, M. / Scagliola, S. (2014). Just Google It. Digital Research Practices of Humanities Scholars. *Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities. Sheffield: HRI Online*.

Kounavis, C. D./ Kasimati, A. E. / Zamani, E. D. (2012). Enhancing the tourism experience through mobile augmented reality: Challenges and prospects. *Int. J. Eng. Bus. Manag.,* 4, 1–6.

Kwastek, K. (2014). Vom Bild zum Bild. Digital Humanities jenseits des Texts (Keynote). *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014).* Passau.

Kysela, J. / Štorková, P. (2015). Using Augmented Reality as a Medium for Teaching History and Tourism. *Procedia - Soc. Behav. Sci.,* 175, 926–931.

Livingston, M. A./ Bimber, O. / Saito, H. (2008). *Proceedings of the 7th IEEE International Symposium on Mixed and Augmented Reality. Cambridge, UK,* Piscataway, IEEE Xplore.

Maina, J. K. / Suleman, H. (2015). Enhancing Digital Heritage Archives Using Gamified Annotations. *In:* ALLEN, B. R., HUNTER, J. & ZENG, L. M. (eds.) *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings.* Cham: Springer International Publishing.

Münster, S. (2011). Entstehungs- und Verwendungskontexte von 3D-CAD-Modellen in den Geschichtswissenschaften. *In:* MEISSNER, K. & ENGELIEN, M. (eds.) *Virtual Enterprises, Communities & Social Networks.* Dresden: TUDpress.

Münster, S./ Jahn, P.-H. / Wacker, M. (2015). Von Plan- und Bildquellen zum virtuellen Gebäudemodell. Zur Bedeutung der Bildlichkeit für die digitale 3D-Rekonstruktion historischer Architektur. *In:* AMMON, S. & HINTERWALDNER, I. (eds.) *Bildlichkeit im Zeitalter der Modellierung. Operative Artefakte in Entwurfsprozessen der Architektur und des Ingenieurwesens.* München: Wilhelm Fink Verlag.

Niccolucci, F. / Hermon, S. (2006). A Fuzzy Logic Approach to Reliability in Archaeological Virtual Reconstruction. *In:* NICCOLUCCI, F. & HERMON, S. (eds.) *Beyond the Artifact. Digital Interpretation of the Past.* Budapest.

Paul, G. (2006). Von der Historischen Bildkunde zur Visual History. *Visual History. Ein Studienbuch.* Göttingen.

Pérez-Gómez, A. / Pelletier, L. (1997). *Architectural Representation and the Perspective Hinge,* Cambridge, London, University Press.

Samuel, J./ Périnaud, C./ Servigne, S./ Georges, G. / Gesquirère, G. (2016). Representation and Visualization of Urban Fabric through Historical Documents. *14th Eurographics Workshop on Graphics and Cultural Heritage 2016.* Genua.

Smirnov, A./ Kashevnik, A./ Shilov, N./ Teslya, N. / Shabaev, A. (2014). Mobile application for guiding tourist activities: Tourist assistant -

TAIS. *Conference of Open Innovation Association, FRUCT, 2014*, 95–100.

Tom Dieck, M. C. / Jung, T. (2015). A theoretical model of mobile augmented reality acceptance in urban heritage tourism. *Curr. Issues Tour.,* 3500, 1-21.

Tom Dieck, M. C./ Jung, T. H. / Tom Dieck, D. (2016). Enhancing art gallery visitors' learning experience using wearable augmented reality: generic learning outcomes perspective. *Curr. Issues Tour.* 1-21.

Walczak, K./ Cellary, W. / Prinke, A. (2011). Interactive Presentation of Archaeological Objects Using Virtual and Augmented Reality. *In:* JEREM, E., REDÖ, F. & SZEVERÉNYI, V. (eds.) *On the Road to Reconstructing the Past. Proceedings of the 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Budapest: Archaeolingua.

Wohlfeil, R. (1986). Das Bild als Geschichtsquelle. *Historische Zeitschrift,* 243, 91–100.

Yovcheva, Z./ Buhalis, D. / Gatzidis, C. (2012). Overview of smartphone augmented reality applications for tourism. *e-Review of Tourism Research,* 10, 63–66.

Zaibon, S. B./ Pendit, U. C./ Aida, J. / Bakar, A. (Year) Published. User Requirements on Mobile AR for Cultural Heritage Site towards Enjoyable Informal Learning. Multimedia and Broadcasting (APMediaCast), 2015 Asia Pacific Conference, 2015. 23–25.

Zöllner, M./ Becker, M. / Keil, J. (2010). Snapshot Augmented Reality - Augmented Photography. *In:* ARTUSI, A., JOLY-PARVEX, M., LUCET, G., RIBES, A. & PITZALIS, D. (eds.) *11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2010).* Paris: Eurographics Association.

Strings&Structures

Rolshoven, Jürgen
rols@spinfo.uni-koeln.de
Universität zu Köln, Deutschland

Etimi, Valmir
vetemi@smail.uni-koeln.de
Universität zu Köln, Deutschland

Seipel, Peter
pseipel1@uni-koeln.de
Universität zu Köln, Deutschland

Wiehe, Thomas
twiehe@uni-koeln.de
Universität zu Köln, Deutschland

Der vorliegende Beitrag befasst sich mit dem Projekt "Strings&Structures. Codes of Sense and Function in Genomics and Linguistics". Dieses Projekt wird von der Sprachlichen Informationsverarbeitung und der Bioinformatik im Rahmen der Exzellenzinitiative der Universität zu Köln durchgeführt. Beide Bereiche befassen sich intensiv mit der Prozessierung von Texten. Bei der Sprachlichen Informationsverarbeitung handelt es sich um natürlichsprachliche Texte und Textkorpora, bei der Bioinformatik um genomische Texte. Das Projekt zielt auf die Aufdeckung von Mustern in Texten und die Analyse der Beziehung der Muster untereinander. Vor dem Hintergrund dieser Fragestellungen werden gemeinsam nutzbare Algorithmen entwickelt. Jedoch sollten dabei wesentliche Unterschiede der zugrundeliegenden Textarten nicht übersehen werden. Natürlichsprachliche Texte sind das Resultat grammatischer Produktionssysteme, genomische Texte sind Produktionssysteme. Das linguistische Vorhaben zielt auf die Rekonstruktion der erzeugenden Produktionssysteme aus zugrundeliegenden Textkorpora. Weitere Unterschiede zwischen natürlichsprachlichen Texten und genomischen Texten liegen in der Größe der zugrunde liegenden Alphabete und der zweigliedrigen Kombinatorikebenen natürlicher Sprachen. Wenngleich die Interaktion und Dynamik der Einheiten in genomischen Texten hochkomplex ist, so kann die Funktion einer einzelnen Einheit gut bestimmt werden. In natürlichen Sprachen dagegen ist die Bedeutung einzelner Einheiten oftmals nur schwierig zu bestimmen. Sie ist hochgradig kontext- und situationsabhängig. Dies hängt auch damit zusammen, dass sprachliche Einheiten weitgehend polysem sind.

Die automatische Aufdeckung der Bedeutung und Funktion sprachlicher Zeichen vollzieht sich in einem vierstufigen Prozess:

1. Ermittlung minimaler bedeutungs- oder funktionstragender Einheiten.
2. Kombinatorik dieser Einheiten durch Aufdeckung morphologischer Prozesse.
3. Syntaktische Kombinatorik der morphologisch erkannten Einheiten.
4. Auswertung syntaktischer Strukturen für die Bestimmung der Bedeutung sprachlicher Einheiten.

Dieses vierstufigen Verfahren wird in schrittweiser Verfeinerung in weitere Komponenten zerlegt, die algorithmisch als Module in einem

Prozesskettensystem frei verschaltet werden. Ein solches graphisch orientiertes System ermöglicht auch Laien, Prozessketten für die Lösung eigener Fragestellungen zu schaffen.

Ad 1. Ermittlung minimaler bedeutungs- oder funktionstragender Einheiten.

Bei der Ermittlung minimaler Bedeutung oder funktionstragende Einheiten wird von dem strukturalistischen Grundgedanken der Zeichenkonstitution durch Opposition ausgegangen. Dieser Gedanke wird mit Hilfe von Suffixbäumen umgesetzt. In Suffixbäumen verweisen Verzweigungen auf potenziell in Opposition stehende Zeichenketten hin. Allerdings führt eine direkte Auswertung von Verzweigungen Suffixbäumen zu einer viel zu mächtigen Menge potentieller Morpheme. Daher müssen Filtermechanismen für deren Reduktion konstruiert werden. Ein Filtermechanismus beruht darin, nur identische Zeichenketten aus vorwärts und rückwärts aufgebauten Suffixbäumen zu verwenden.

Ad 2. Kombinatorik dieser Einheiten durch Aufdeckung morphologischer Prozesse.

Ein weiterer Filtermechanismus liegt in der Begrenzung der Kombinatorik von kleinsten funktions- oder bedeutungstragenden Einheiten. Formal kann morphologische Kombinatorik als Typ-2-Sprache im Sinne der Chomsky-Hierarchie formaler Sprachen betrachtet werden. Mit zusätzlichen Kriterien zur Unterscheidung bedeutungs- oder funktionstragender Einheiten kann die Übermenge, die der Suffixbaumgenerator liefert, drastisch reduziert werden. Die verbleibenden Einheiten sind in den nachfolgenden Schritten syntaktisch und semantisch zu analysieren.

Ad 3. Syntaktische Kombinatorik der morphologisch erkannten Einheiten

Eines der Probleme maschineller syntaktischer Sprachverarbeitung liegt in der Kontextsensitivität natürlicher Sprachen. Dies hat unter anderem zur Folge, dass Einheiten, die bedeutungsmäßig zusammengehören, in Sätzen oftmals weit voneinander getrennt sind. Für die Erkennung semantischer Zusammengehörigkeit und semantischer Abhängigkeit werden in dem vorliegenden Projekt Kookurrenzmatrizes ausgewertet. Die Kookurrenzmatrizes speichern semantische Vektoren, die semantische Abhängigkeit ausdrücken. Starke semantischer Abhängigkeit -etwa eines Verbs zu seinem Objekt -können werden in einer Baumstruktur direkt durch benachbarte Knoten ausgedrückt, selbst dann, wenn es Vorkommen des Objekts gibt, die gar nicht unmittelbar neben dem Verb im Textkorpus stehen. Letztlich könnten auf diese Weise kontextabhängige Phänomene aufgedeckt werden.

Ad 4. 4. Auswertung syntaktischer Strukturen für die Bestimmung der Bedeutung sprachlicher Einheiten.

Syntaktische Strukturen in natürlichen Sprachen haben die Funktion, die Prozessierung sprachlichen Inputs zu erleichtern und zu beschleunigen. Syntaktische Strukturbäume ermöglichen es, korrekte Beziehungen zwischen sprachlichen Elementen herzustellen. Für die Bestimmung von Bedeutungspotenzial sind syntaktische Strukturen daher von grundlegender Bedeutung. Wird das Bedeutungspotenzial wiederum durch vektorielle Kookurrenzmatrizes erfasst, dann tragen syntaktischer Strukturbäume dazu bei, die Zahl der Komponenten der Matrizes stark zu reduzieren und folglich die vektorielle Semantik zu schärfen.

Eine Besonderheit des hier gewählten Vorgehens liegt in der Interaktion von subsymbolischen, vektoriellen und symbolischen, baumstrukturorientierten Verfahren. Die Stärke symbolischer Verfahren liegt in ihrer Kompaktheit und der Möglichkeit der Falsifikation. Subsymbolischer Verfahren sind nicht oder nur schwierig falsifizierbar. Sie machen semantische Unschärfe und semantische Ähnlichkeit fassbar

SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften

Barzen, Johanna
johanna.barzen@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Deutschland

Breitenbücher, Uwe
uwe.breitenbuecher@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kronenwett, Simone
simonekronenwett@gmail.com
Universität zu Köln, Deutschland

Leymann, Frank
frank.leymann@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Deutschland

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland

Der digitale Wandel verändert die Wissenschaft grundlegend (Kramp 2013). Das exponentielle Wachstum, die steigende Komplexität sowie der zunehmende Gebrauch von digitalen Forschungsdaten beeinflussen den Forschungsprozess signifikant. Um das Potential der fortschreitenden Digitalisierung optimal nutzen zu können, müssen entsprechende Infrastrukturen geschaffen werden, die das Management von Forschungsdaten, die Möglichkeit ihrer Vernetzung, ihre dauerhafte Verfügbarkeit und einen freien Zugang gewährleisten.

Die Vielzahl von wissenschaftspolitischen Empfehlungen, Bestandsaufnahmen, Umfragen und institutionellen Richtlinien rund um Forschungsdaten, die in den vergangen Jahren veröffentlicht wurden, sind ein Zeichen zunehmenden Problembewusstseins, politischen Handlungswillens, aber eben auch anhaltenden Handlungsbedarfs (Pampel / Bertelmann 2011; Wissenschaftsrat 2012; DV-ISA 2016; Rat für Informationsinfrastrukturen 2016). Durch die großen europäischen und nationalen Infrastrukturprojekte in den Geisteswissenschaften (CLARIN, DARIAH), die Einrichtung von fachspezifischen Datenzentren und spezifisch geisteswissenschaftlichen Datenzentren, wie dem Data Center for the Humanities (DCH, s. http://dch.uni-koeln.de), hat sich die Versorgungslage für Forschungsdaten stetig verbessert. Doch längst werden nicht alle produzierten Daten und digitalen Forschungsergebnisse tatsächlich für die Nachnutzung verfügbar gemacht bzw. sind für eine dauerhafte Verfügbarkeit in einer hochdynamischen digitalen Welt gerüstet (Razum / Neumann 2014; Wissenschaftsrat 2012).

Die Diskussion fokussiert bisher stark auf Forschungsinformationssysteme zur standardisierten Vorhaltung von Forschungsprimärdaten. Dabei bleibt weitgehend unberücksichtigt, dass ein Großteil der digitalen Produkte in den Geisteswissenschaften in einer Form vorliegen, die sich einer normierten Versorgung bislang entzieht. Gemeint sind Forschungsanwendungen, oder "lebende Systeme" (Sahle/Kronenwett 2013), die einen wesentlichen Bestandteil digitaler Ergebnissicherung darstellen: Präsentationssysteme, interaktive Visualisierungen, Recherche-Datenbanken, digitale Editionen und digitale Arbeitsumgebungen – um nur einige Formen zu nennen – sind Arbeitswerkzeuge, Plattformen der Dissemination und Aggregation von Forschungsergebnissen und sind aus dem Arbeitsalltag in Geisteswissenschaften nicht mehr wegzudenken. Ihre dauerhafte Erhaltung, Betreuung und Bereitstellung über den Zeitraum der Projektförderung hinaus stellt eine große organisatorische und letztlich finanzielle Herausforderung dar.

Ein grundlegendes Problem ist die Sorge um softwaregestützte Präsentationen und Funktionalitäten, die in Forschungsprojekten entstehen (Sahle/Kronenwett 2013). Nicht selten sind diese die eigentlichen Träger von „Informationsgehalt beziehungsweise wissenschaftliche[m] Mehrwert" der im Projekt erbrachten Forschungsleistung (Wuttke et al. 2016). Eine Nachhaltigkeitsstrategie, die auf eine Abtrennung und Archivierung allein der Datenbasis zurückfällt, führt unweigerlich zum Verlust von Information und reduziert im Extremfall den wissenschaftlichen Nutzen auf null (Blumtritt/Mathiak 2016).

Forschungsanwendungen sind keineswegs statische Objekte, sondern unterliegen einem kontinuierlichen Veränderungszyklus. Viele Anwendungen fungieren als Plattformen, die User-Input entgegennehmen und damit ihre Datenbasis laufend verändern. Browser-Updates und Veränderungen der Nutzungsgewohnheiten können bestimmte Komponenten unbrauchbar oder obsolet machen und damit eine Überarbeitung des Codes anstoßen. Notwendige Sicherheitsupdates erfordern regelmäßiges Eingreifen und können Kaskaden von weiteren Updates und Softwareanpassungen nach sich ziehen. Der Verzicht auf kontinuierliche Wartung spart kurzfristig Kosten, verschärft mittelfristig aber das Problem. Erfahrungen aus dem "LAZARUS-Projekt" zeigen, dass Anwendungen, die über längere Zeit brach liegen, nur unter großem Ressourceneinsatz wieder revitalisiert werden können (Bingert et al. 2016). Ein erster Ansatz zur Archivierung von Forschungssoftware auf Basis des TOSCA-Standards wurde vorgestellt, löst jedoch nur das Problem der initialen automatisierten Bereitstellung, nicht der Revitalisierung terminierter Systeme (Breitenbücher et al. 2017).

Darüber hinaus sehen sich Forschungseinrichtungen mit dem Dilemma konfrontiert, dass befristete Projektlaufzeiten einen dauerhaften Betrieb organisatorisch deutlich erschweren. Dies ist insbesondere in den Geisteswissenschaften fatal, da hier oftmals ein anderer Maßstab bezüglich der Nachhaltigkeit angelegt wird, als dies in

den vergleichsweise schnelllebigen Naturwissenschaften üblich ist. Die institutionelle Vorhaltung von Forschungsdaten für einige Jahrzehnte erfüllt den Zweck der Überprüfbarkeit und Reproduzierbarkeit im Sinne der guten wissenschaftlichen Praxis, für die Vorhaltung von Gegenständen des kulturellen Erbes ist eine zeitliche Beschränkung nicht sinnvoll. Selbst dann, wenn sich eine Abschaltung aus Sicherheits- oder Kostengründen nicht vermeiden lässt, so ist zumindest zu gewährleisten, dass Anwendung und Datenbasis auf eine Art und Weise archiviert werden, dass sie sich jederzeit verlustfrei und benutzbar wiederherstellen lassen.

Das Projekt „SustainLife“, das in Zusammenarbeit zwischen des Data Center for the Humanities (DCH) aus Köln und des Instituts für Architektur von Anwendungssystemen (IAAS) der Universität Stuttgart durchgeführt wird, adressiert diese Probleme. In diesem Poster stellen wir das Projekt vor und präsentieren den ersten Teil einer Anforderungsanalyse an lebende Systeme der Digital Humanities. Als zweiten Teil soll vor Ort eine Umfrage unter den Teilnehmern der Konferenz durchgeführt werden, um die Anforderungen an lebende Systeme aus unterschiedlichsten Domänen möglichst zielgenau zu analysieren. Diese Anforderungsanalyse stellt die Basis für im Projekt SustainLife geplante Erweiterungen der open-source Software „OpenTOSCA“ (Binz et al. 2013) dar, welche auf dem OASIS-Standard TOSCA (OASIS 2013) basieren. Dieser Standard ermöglicht die Automatisierung des Deployments und Managements von Anwendungen und stellt damit eine vielversprechende Grundlage zur kostengünstigen Sicherung der Nachhaltigkeit lebender Systeme dar.

Bibliographie

Bingert, Sven / **Blumtritt, Jonathan** / **Buddenbohm, Stefan** / **Engelhardt, Claudia** / **Kronenwett, Simone** / **Kurzawe, Daniel** (2016): "Anwendungskonservierung und die Nachhaltigkeit von Forschungsanwendungen", in: Forschungsdaten in den Geisteswissenschaften (FORGE 2016) 13-16 https://www.fdm.uni-hamburg.de/ueber-uns/a-nachrichten/aktivitaeten/forge16/praesentationen/programmheft.pdf [letzter Zugriff 20. September 2017].

Binz, Tobias / **Breitenbücher, Uwe** / **Haupt, Florian** / **Kopp, Oliver** / **Leymann, Frank** / **Nowak, Alexander** / **Wagner, Sebastian** (2013): "OpenTOSCA - A Runtime for TOSCA-based Cloud Applications", in: Proceedings of the 11th International Conference on Service-Oriented Computing (ICSOC 2013). Springer.

Blumtritt, Jonathan / **Mathiak, Brigitte** (2016): “Consulting Workflow for Humanities Research Data”, in: Forschungsdaten in den Geisteswissenschaften (FORGE 2016) 21-22 https://www.fdm.uni-hamburg.de/ueber-uns/a-nachrichten/aktivitaeten/forge16/praesentationen/programmheft.pdf [letzter Zugriff 20. September 2017].

Breitenbücher, Uwe / **Barzen, Johanna** / **Falkenthal, Michael** / **Leymann, Frank** (2017): "Digitale Nachhaltigkeit in den Geisteswissenschaften durch TOSCA: Nutzung eines standardbasierten Open-Source Ökosystems", in: DHd 2017: Digitale Nachhaltigkeit 235-238 http://www.dhd2017.ch/wp-content/uploads/2017/01/Abstractband_def_24.1.17-1.pdf [letzter Zugriff 20. September 2017].

DV-ISA (2016): "Umgang mit digitalen Daten in der Wissenschaft: Forschungsdatenmanagement in NRW. Eine erste Bestandsaufnahme", Version 0.7, https://www.dvisa-nrw.de/veroeffentlichungen/veroeffentlichungen-container-oeffentlich/dv-isa-vorstudie-bestandsaufnahme-forschungsdatenmanagement [letzter Zugriff 20. September 2017].

Kramp, Leif (2013): “Wie die Digitalisierung die Wissenschaft verändert”, in: sueddeutsche.de, 20.11.2013, http://www.sueddeutsche.de/wissen/digitales-morgen-debatte-zur-digitalisierung-wie-die-digitalisierung-die-wissenschaft-veraendert-1.1823133[letzter Zugriff 20. September 2017].

OASIS (2013): Topology and Orchestration Specification for Cloud Applications Version 1.0.

Pampel, Heinz / **Bertelmann, Roland** (2011): "Data Policies im Spannungsfeld zwischen Empfehlung und Verpflichtung", in: Handbuch Forschungsdatenmanagement 49–61, Bad Honnef: Bock + Herchen.

Rat für Informationsinfrastrukturen (2016): "Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland", http://www.rfii.de/download/rfii-empfehlungen-2016/ [letzter Zugriff 20. September 2017].

Razum, Matthias / **Neumann, Janna** (2014): "Das RADAR Projekt: Datenarchivierung und -publikation als Dienstleistung - disziplinübergreifend, nachhaltig, kostendeckend", in: o|bib Das offene Bibliotheksjournal 1/1: 30–44 https://www.o-bib.de/article/view/2014H1S30-44/117 [letzter Zugriff 20. September 2017].

Sahle, Patrick / **Kronenwett, Simone** (2013): "Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities'", in: LIBREAS. Library Ideas 23: 76–96.

Wissenschaftsrat (2012): "Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020", http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf [letzter Zugriff 20. September 2017].

Wuttke, Ulrike / Engelhardt, Claudia / Buddenbohm, Stefan (2016): "Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum", in: Zeitschrift für digitale Geisteswissenschaften 2017, text/html Format, doi:10.17175/2016_003.

Syndred - A Syntax-Driven Editor for Lexical Resources

Mondaca, Francisco
f.mondaca@uni-koeln.de
Universität zu Köln, Deutschland

Rolshoven, Jürgen
rols@spinfo.uni-koeln.de
Universität zu Köln, Deutschland

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Universität zu Köln, Deutschland

Vogt, Andreas
vogt.andreas@uni-koeln.de
Universität zu Köln, Deutschland

Kontinuierlich steigt die Menge der erzeugten und prozessierten Information und damit der Bedarf an technologisch assistierter bishin zu autonomer Datenverarbeitung. Textuell repräsentierte Inhalte mittels strukturgebender Methoden aufzubereiten ist meist aufwendiger als deren Erzeugung, insofern gewinnt automatische Textprozessierung zunehmend an Bedeutung (Bernstein et al.: 2016).

Mithilfe der formalen Grammatiken RBNF (Rich Backus-Naur Form) [1] und ABNF (Augmented Backus-Naus Form) können Produktionssysteme im Sinne domänenspezifischer, formaler Sprachen (engl. Domain-Specific Languages) modelliert und auf textuelle Daten angewandt werden: Syndred [2] nutzt als Parser konkretisierte Grammatiken zur Abbildung von Texten auf Strukturbäume (engl. Parse Trees), die in vielfältiger Form (z. B. als TEI-Dokumente) persistiert, mittels Treewalkern (Parr 2013: 17ff) für weitere Prozessierung transformiert oder beispielhaft zur Überprüfung formaler Korrektheit unter Berücksichtigung von Merkmalen wie bspw. Schriftgröße, -stil und -farbe eingesetzt werden.

Wegweisend für die Entwicklung des hier beschriebenen Systems sind die Arbeiten von Wirth und Gutknecht (Gutknecht 1985 bzw. Wirth / Gutknecht 1992: 78ff). Entgegen der Funktionsweise eines Compilers jedoch, der die Analyse des zu übersetzenden Programms meist direkt in Programmcode vornimmt, definiert Syndred textuelle Strukturen anhand domänenspezifischer Grammatiken und bildet sie auf Syntaxgraphen ab.

Der syntaxkontrollierte Editor (engl. Syntax-Driven Editor, kurz: Syndred) soll als flexibles Werkzeug in unterschiedlichsten Projekten (z. B. Lexikoneditoren und denzentraler Korrektur) dienen, kollaboratives Arbeiten durch effiziente formale Kontrolle unterstützen und damit die allgemeine Produktivität durch die Erstellung domänenspezifischer Sprachen erhöhen (Fowler 2011: xxi).

Syndred wartet mit einem Split-View Design auf, welches einerseits das Erstellen und Bearbeiten von Grammatiken und andererseits syntaxkontrolliertes Editieren textueller Inhalte in direkten, visueller Bezug zueinander bringt. Kernkompetenzen der Benutzerschnittstelle sind die Herleitungen von Grammatiken aus textuellen Bestandsdaten oder Rückmeldungen formaler Produktionsregeln zu textuellen Inhalten.

Dezentrales Arbeiten erfordert eine netzwerkbasierte Architektur, idealerweise eine Web-Applikation mit clientseitigem Editor und serverseitigem Parser; Syndred nutzt das auf ReactJS [3] basierende Editor-Framework DraftJS [4] und kommuniziert im JSON-Format über WebSockets mit einer zentralen, auf Basis des Spring-Frameworks [5] implementierten Java-Applikation. Neben dem Vorteil der Plattformunabhängigkeit ermöglichen die eingesetzten Web-Technologien die zentralisierung der Hardwareleistung wie auch des Speicherbedarfs und sichern gegen Datenverlust ab. Auch garantiert die Bidirektionalität der Client-Server-Verbindung die Kohärenz und Persistenz der kollaborativen Instanzen. Zeitnahe Verfügbarkeit aller für die kollaborative Arbeit mit Syndred notwendigen Ressourcen wird durch einen serverseitigen Cached Thread Pool sichergestellt; aktiven Instanzen wird bei Bedarf ein

Parser-Threads zur Verfügung gestellt, der nach Verwendung beendet wird.

Syndred ist somit ein Werkzeug zur intuitiven Entwicklung domänenspezifische Sprachen und Überprüfung textueller Inhalte anhand dieser formalen Grammatiken, realisiert als kollaborative Web- Applikation; zusammengenommen ein Meilenstein in der Entwicklung dieser Art von Programmiersprachen.

Fußnoten

1. RBNF ist eine EBNF (Extended Backus-Naus Form) Erweiterung um Rich-Text Auszeichnungen.
2. *https://github.com/spinfo/syndred*
3. *https://facebook.github.io/react*
4. *https://draftjs.org*
5. *https://projects.spring.io/spring-framework*

Bibliographie

Bernstein, Abraham / Hendler, James / Noy, Natalya (2016): "A New Look at the Semantic Web." in: *Communications of the ACM*. 59:9. 35-37.

Fowler, Martin (2011): *Domain-Specific Languages*. Addison-Wesley: Upper Saddle River, New Jersey.

Gutknecht, Jürg (1985): "Concepts of the Text Editor Lara." in: *Communications of the ACM*. 28:9. 942-960.

Parr, Terence (2013): *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf: Dallas Texas.

Wirth, Niklaus (1986[4]): *Compilerbau*. Teubner: Stuttgart.

Wirth, Niklaus (2000[5]): *Algorithmen und Datenstrukturen*. Teubner: Stuttgart.

Wirth, Niklaus / Gutknecht, Jürg (1992): *Project Oberon: Design of an Operating System and Compiler*. Addison-Wesley Longman: Amsterdam.

TEASys: Kollaboratives digitales Annotieren als Lehr- und Lernprozess

Zirker, Angelika
angelika.zirker@uni-tuebingen.de
Humboldt Universität zu Berlin; Eberhard Karls Universität Tübingen

Bauer, Matthias
m.bauer@uni-tuebingen.de
Eberhard Karls Universität Tübingen

Kirchhoff, Leonie
leonie.kirchhoff@uni-tuebingen.de
Eberhard Karls Universität Tübingen

Lahrsow, Miriam
miriam.lahrsow@uni-tuebingen.de
Eberhard Karls Universität Tübingen

Das Poster präsentiert TEASys (Tübingen Explanatory Annotations System; Bauer/Zirker 2015) und die Möglichkeiten, die es als heuristisches *tool* in Lehr- und Lernprozessen bietet. Im Projekt annotieren Studierende kollaborativ Texte der englischsprachigen Literatur. Nachdem TEASys und seine Struktur wie auch Funktionen bereits bei der DHd 2016 und 2017 vorgestellt wurden, liegt der Schwerpunkt nun auf dem *Prozess* der kollaborativen Annotation im digitalen Medium.

Die Praxis des erklärenden Annotierens, d.h. der Anreicherung eines Textes mit Zusatzinformationen um ihn verständlicher zu machen, wird durch die Digitalisierung grundlegend beeinflusst. Während die Erstellung von annotierten Buch-Editionen vor allem Einzelforschern vorbehalten war (und ist), eröffnen sich durch Digitalität neue Möglichkeiten der Generierung von *social knowledge*: Viele digitale Editionen bzw. Plattformen (z.B. *The Readers' Thoreau*, *Infinite Ulysses*, *PynchonWiki*, *Genius.com*) und Tools (z.B. *Annotation Studio*, *A.nnotate*, *hypothes.is*) ermöglichen es Lesern, zur Erläuterung von Wörtern und Passagen beizutragen. Dadurch löst sich die Grenze zwischen Leser und Annotierendem auf (Sahle 2013: 177, 258). Gleichzeitig gibt es in einer digitalen Edition keine zeitlichen oder räumlichen Einschränkungen, so dass Annotatoren kontinuierlich neue Informationen zu einer Annotation hinzufügen können. Damit ist Digitalität im Sinne des Tagungsthemas auch der Kritik zu unterziehen. Denn sie kann dazu führen, dass endlose Abhandlungen zu jedem einzelnen Wort eines Textes entstehen, was zu Informationsflut, irrelevanten oder unstrukturierten Informationen und infolgedessen zu einem Qualitätsverlust der digitalen Annotation führen kann. Damit laufen digitale erklärende Annotationen Gefahr, den Nutzer bei seinem Anliegen, einen Text zu verstehen, zu verwirren oder in die Irre zu führen (s. Bauer/Zirker 2017b).

TEASys steuert dem entgegen, indem sowohl der Prozess des Annotierens wie auch der Nutzen der entstandenen Annotationen kritisch re-

flektiert werden. Dies geschieht in studentischen *peer- learning*-Gruppen und in Lehrveranstaltungen, die häufig den Anstoß für die selbständige Weiterarbeit in den Gruppen geben. In den Lehrveranstaltungen dient das Annotieren als Methode zur Erarbeitung historisch und/oder kulturell distanter Texte ebenso wie zum Erwerb textanalytischer Fähigkeiten. Dabei lernen Studierende, ihre eigene Vorgehensweise zu reflektieren und sie überprüfbar zu machen. TEASys wird als Lehrmethode eingesetzt, indem es Studierende dazu anregt, ihr eigenes Unverständnis eines Textes zu reflektieren: welche Elemente eines Textes tragen dazu bei, dass er ‚schwierig' ist (oder so empfunden wird), und welche (Art von) Informationen werden benötigt, um diese Schwierigkeiten zu überwinden? Hier kommt die Heuristik von TEASys ins Spiel: Annotationen sind strukturiert nach Kategorien der Information (z.B. Sprache, Form, Intertextualität etc.; s. Bauer/Zirker 2015) sowie nach Ebenen der Komplexität (insgesamt drei) und zielen damit auf konkrete Leserbedürfnisse (Bauer/Zirker 2017b).

Die Strukturierung der Annotationen hilft Studierenden somit auch, Fragen zu formulieren, auf die sie andernfalls vielleicht nicht gestoßen wären (z.B. „Konnte das Wort „travel" Wort im 16. Jahrhundert nicht noch etwas anderes bedeuten?"). Häufig wird dabei deutlich, dass die Erläuterungsbedürftigkeit eines Textes oder von Textteilen oft nur aufgrund eines gewissen Expertenwissens erkannt wird. Dieses wird oftmals von den Lehrenden beigesteuert; aufgrund der Strukturierung können aber auch Studierende leichter zu Experten werden. Die Kombination von Expertise und Kritik führt zu Generierung von *social knowledge* durch den Austausch innerhalb der *peer*-Gruppen, wodurch in effizienter Weise qualitativ verifizierte Arbeitsergebnisse erzeugt werden (vgl. Jannidis, Kohle, Rehbein 2017: 211). Die Publikation der Annotationen online stellt eine zusätzliche Motivation für die Studenten dar, Output auf hohem fachlichem Niveau zu produzieren (Stroud 2006: 215), sowie ihre Annotation zu verbessern. Die kontinuierliche Weiterarbeit (vgl. Ralle 1016: 147) birgt aber auch Risiken in Hinsicht auf die Verwend- und Zitierbarkeit des kollaborativ entstandenen Wissens.

Das Poster stellt die einzelnen Lehr- und Lernprozesse des kollaborativen Annotierens dar und soll die Nachhaltigkeit des Projektes und sowie den Mehrwert der Heuristik von TEASys im Hinblick auf digitale Methoden im Literaturunterricht veranschaulichen.

Bibliographie

Bauer, Matthias / Zirker, Angelika (2015): "Whipping Boys Explained: Literary Annotation and Digital Humanities." in: *Literary Studies in the Digital Age: An Evolving Anthology*. https://dlsanthology.mla.hcommons.org/whipping-boys-explained-literary-annotation-and-digital-humanities/ [letzter Zugriff 23. September 2017].

Bauer, Matthias / Zirker, Angelika (2017a): "TEASys (Tübingen Explanatory Annotations System): Die erklärende Annotation literarischer Texte in den Digital Humanities." in: *DHd2017: Digitale Nachhaltigkeit. Konferenzabstracts* 274-276.

Bauer, Matthias / Zirker, Angelika (2017b): "Explanatory Annotation of Literary Texts and the Reader: Seven Types of Problems." in: *IJHAC* 11.2: 212-232.

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

Ralle, Inga H. (2016): "Maschinenlesbar – menschenlesbar." in: *Editio: internationales Jahrbuch für Editionswissenschaft* 30.1: 144-156.

Sahle, Patrick (2013): *Digitale Editionsformen: Teil 2: Befunde, Theorien und Methodik*. Norderstedt: Book on Demand.

Stroud, Matthew D. (2006): "The Closest Reading: Creating Annotated Online Editions." *in: Bass, Laura R. / Greer, Margaret R. (eds.): Approaches to Teaching Early Modern Spanish Drama. New York: Modern Language Association of America* 214-219 .

TEI-Editionswerkstatt Urkunden@UPB.

Schwengelbeck, Isabel
schwengelbeck.isabel@gmail.com
Universität Paderborn, Deutschland

Wahl, Dominik
dwahl@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Foester, Karl
foesterkarl@gmail.com
Universität Paderborn, Deutschland

Friedl, Dennis
dennisfriedl@paderborn.com
Universität Paderborn, Deutschland

Fluss, Fabian
fabian.fluss92@gmx.de
Universität Paderborn, Deutschland

Mersch, Isabelle
imersch@campus.uni-paderborn.de
Universität Paderborn, Deutschland

Voss, Fabian
vossf@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Dröge, Martin
martin.droege@upb.de
Universität Paderborn, Deutschland

Stadler, Peter
peter.stadler@upb.de
Universität Paderborn, Deutschland

Voges, Ramon
ramon.voges@upb.de
Universität Paderborn, Deutschland

Ein DH-Lehr-Lernprojekt zu den Gründungsurkunden der Jesuitenuniversität Paderborn

Die „digitale Revolution“ greift zunehmend in komplexer Weise auf alle Lebensbereiche über. Im „digitalen Zeitalter“ ist auch ein Wandel in der Geschichtsschreibung zu beobachten, der schon bei der Unterscheidung analoger von digitalen Quellen sichtbar wird (Pfanzelter 2016: 85). Die Arbeitsweise von HistorikerInnen wird durch die Verfügbarkeit von Quellen in digitaler und digitalisierter Art verändert (Bernsen 2017: 295, Kelly 2013). Daraus ergibt sich nicht nur das Desiderat nach entsprechenden Kompetenzen im Umgang mit digitalen sowie digitalisierten Quellen, sondern auch nach einer „der digitalen Welt angepassten, technikgestützten Quellenkritik“ (Pfanzelter 2016: 93). Herausforderungen hinsichtlich des Umgangs mit Quellen lassen sich für alle Ebenen der Gesellschaft ableiten – von einer Schülerschaft, die zur Partizipation in der Gesellschaft befähigt werden soll, über (Lehramts-) Studierende, die dieses ermöglichen sollen, bis zu den FachwissenschaftlerInnen und Lehrenden an den Universitäten. Dementsprechend hoch ist die Relevanz, das im folgenden vorgestellte Projekt im Kontext der „Kritik der digitalen Vernunft“ zu diskutieren und einen besonderen Fokus darauf zu legen, welche und wie stark ausgeprägte digitale Kompetenzen HistorikerInnen benötigen, um dem „digitalen Zeitalter“ gerecht zu werden und den kritischen Anforderungen der Geisteswissenschaften zu genügen. Dies gilt in besonderem Maße mit Blick auf zukünftige GeschichtslehrerInnen, die in ihrer Position als gesellschaftliche Multiplikatoren einer adäquaten Ausbildung bedürfen.

Das Lehrkonzept des Forschenden Lernens eröffnet die Möglichkeit, Studierende mit Projekten, Methoden und Werkzeugen der Digital Humanities kritisch und reflektiert vertraut zu machen. Die aus (Lehramts-)StudentInnen und DozentInnen des Historischen Instituts der Universität Paderborn zusammengesetzte Projektgruppe ‚TEI-Editionswerkstatt‘ will zugleich wichtige Kompetenzen der Digital Humanities wie auch Fachwissen vermitteln. Dies geschieht außerhalb des Curriculums auf freiwilliger Basis, sodass die Motivation aller Beteiligten sehr hoch ist. Das gemeinsame Ziel ist es, vier Urkunden über die Gründung der Jesuitenuniversität Paderborn um 1600 (Meyer zu Schlochtern 2014) in einer digitalen Quellenedition der Forschung online zur Verfügung zu stellen (Sahle 2013).

Organisiert ist die Arbeitsgruppe in vier Teilgruppen, die jeweils eine Urkunde bearbeiten, woraus einerseits eine zeitökonomische Arbeitsweise für die heterogene Projektgruppe resultiert und andererseits sichergestellt ist, dass jedes Gruppenmitglied sämtliche Arbeitsschritte auf dem Weg zur digitalen Quellenedition selbstständig durchführt. Diese Vorgehensweise ist einer effektiveren Arbeitsweise dienlich und fördert im Sinne eines sekundären Erkenntnisinteresses die Kompetenzen der Arbeitsgruppe hinsichtlich des Umgangs mit den entsprechenden Tools und Methoden sowie einen elaborierten Erkenntnisgewinn hinsichtlich der erarbeiteten Inhalte. Die Ausdifferenzierung der Vorgehensweise und die gruppeninterne Organisation der Arbeitsschritte erfolgen bei regelmäßigen Treffen, bei denen die Ziele immer wieder dem aktuellen Stand des Projekts angepasst sowie Fragen und Ideen diskutiert werden.

Die Arbeitsschritte im Detail:

- Transkription: Es wurde eine eigenständige Transkription der Originaldokumente in ein digitales Format angefertigt.
- TEI-Encoding: Die Quelle wurde in XML nach den Richtlinien der TEI ausgezeichnet. Hier

liegt eine der Hauptaufgaben der Arbeitsgruppe.
- TEI-Schema: Um sicher zu stellen, dass die gleichen Standards eingehalten werden, wurde ein für das Projekt maßgeschneidertes TEI-Schema erarbeitet.
- Übersetzung: Da die Urkunden in lateinischer Sprache vorliegen, soll eine deutsche Übersetzung angeboten werden, die eigens angefertigt werden muss.
- Historische Einleitung: Der Edition soll eine Einleitung vorangestellt sein, in welcher Informationen (u.a. zur Überlieferung) enthalten sind, die dem Leser eine Quellenkritik erleichtern und den historischen Kontext präsentieren.
- HTML-Darstellung: Die Urkunden-Edition soll online zugänglich gemacht werden. Der Benutzer soll die Möglichkeit erhalten, bestimmte Versionen (Original, TEI-Edition, Übersetzung) der Quelle zu vergleichen.

Es wird in unserem Lehr-Lernkontext bewusst TEI – und nicht das textsortenspezifischere CEI – als Auszeichnungssprache eingesetzt, um den Projektmitgliedern möglichst generische Methoden der Textauszeichnung zu vermitteln und ein größtmögliches Spektrum an Anwendungsbereichen hinsichtlich der erlernten Fähigkeiten zu ermöglichen. Für die kollaborativen Arbeiten an den Dokumenten, am TEI-Schema und den Transformationsskripten ist bei GitHub eine entsprechende Gruppe inkl. eines öffentlichen Repositories (https://github.com/gedigiupb/urkunden_upb) eingerichtet worden, wodurch sichergestellt wird, dass alle Beteiligten mit der aktuellen Dateiversion arbeiten. Die TEI-Auszeichnung findet mithilfe des XML-Editors Oxygen statt.

Im Kontext von „Kritik der digitalen Vernunft“ ist das Projekt als praktisches Beispiel im Rahmen der digitalen Angebote, Projekte und Werkzeuge zu verorten. Dabei liegt nicht nur die Erstellung einer „zeitgemäßen“ Quellenedition im Fokus der Arbeitsgruppe. Darüberhinausgehend wird im Rahmen des Projektes der digitale Horizont der (angehenden) HistorikerInnen erweitert sowie bereits vorhandene Kompetenzen im Sinne historischen Denkens und wissenschaftlichen Arbeitens gefördert. Dem kritischen Anspruch der Geisteswissenschaften wird insofern Rechnung getragen, als dass die permanente Reflexion und Ausdifferenzierung der Vorgehensweise und das kritische Hinterfragen der Sinnhaftigkeit des Einsatzes angewandter Tools eine reflektierte Vereinbarkeit der „daten-, algorithmen- und werkzeuggetriebenen“ Wissenschaft mit geisteswissenschaftlichen Ansprüchen generieren. Das praktische Beispiel der „Editionswerkstatt“ ermöglicht die Diskussion gesellschaftlicher Dimensionen der konkret in diesem Kontext wirksamen Digitalisierungsprozesse besonders unter Berücksichtigung heterogener Begrifflichkeiten wie Interaktionsformen, Partizipation und Bildung – es postuliert geradezu die Diskussion ihrer Konsequenzen in Wissenschaft und Gesellschaft.

Bibliographie

Bernsen, Daniel (2017): *“Arbeiten mit digitalen Quellen im Geschichtsunterricht“, in: Bernsen, Daniel / Kerber, Ulf (eds.): Praxishandbuch Historisches Lernen und Medienbildung im digitalen Zeitalter, Opladen/ Berlin/ Toronto: Verlag Barbara Budrich 295-303.*

Kelly, T. Mills (2013): *“Teaching History in the Digital Age“, Ann Arbor: MI: University of Michigan Press, http://dx.doi.org/10.39 98/ dh.12146032.0001.001 [letzter Zugriff: 08. September 2017].*

Meyer zu Schlochtern, Josef (2014): *“Die Academia Theodoriana. Von der Jesuitenuniversität zur Theologischen Fakultät Paderborn 1614-2014“, Paderborn: Schöningh.*

Pfanzelter, Eva (2017): *“Analoge vs. digitale Quellen: eine Standortbestimmung“, in: Bernsen, Daniel / Kerber, Ulf (eds.): Praxishandbuch Historisches Lernen und Medienbildung im digitalen Zeitalter, Opladen/ Berlin/ Toronto: Verlag Barbara Budrich 85-94.*

Sahle, Patrick (2013): *“Digitale Editionsformen. Teil I: Das typografische Erbe, Teil II: Befunde, Theorie und Methodik, Teil III: Textbegriffe und Recodierung.“ Norderstedt: Schriften des Instituts für Dokumentologie und Editorik 7-9, urn:nbn:de:hbz:38-53510 , urn:nbn:de:hbz:38-53523 , urn:nbn:de:hbz:38-53534 [letzter Zugriff: 08. September 2017].*

TEIHencer - Enhance your TEI-Documents

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

Karner, Stefan
stefan.karner@onb.ac.at
Österreichische Nationalbibliothek

Die Forschungstätigkeiten Georeferencing und Entity Linking sind wichtiger Bestandteil vieler DH-Projekte. Webservices/APIs und Tools versuchen diese Tätigkeiten zu vereinfachen und zu beschleunigen. Eines der bekannteren Tools, wenigstens im deutschsprachigen Raum ist dabei vermutlich der 'DARIAH-DE Datasheet Editor'[1]. Dieser zeichnet sich durch seine einfache Benutzung aus, sei es was den Datenimport (ausfüllen einer Tabelle oder Hochladen einer CSV-Tabelle) betrifft oder die anschließende Disambiguierung/Verifizierung der vom 'Getty Thesaurus of Geographic Names'[2] zurückgelieferten Treffer über ein Graphical User Interface.

Das Projekt TEIHencer greift diese Vorzüge des 'DARIAH-DE Datasheet Editors' auf und versucht diese einerseits mit der 'TEI-Welt' zu verknüpfen sowie mit GeoNames und der GND zwei alternative Normdaten Ressourcen einzubinden.

Konkret handelt es sich bei TEIHencer[3] um ein Plug-In zu dem Python/Django basierten prosopographisch-geographischen Informationssystem APIS[4]. Mit Hilfe des TEIHencers ist es möglich, XML/TEI kodierte Texte in denen Lokalitäten, Orte ausgezeichnet sind, über eine Webformular in APIS zu importieren. Während des Imports werden die Orts-Entitäten entsprechend eines vom Benutzer wählbaren X-Path Ausdruckes geparst, gegen GeoNames und GND abgeglichen und im Falle von Übereinstimmung angereichert und in einer relationalen Datenbank gespeichert. Die gespeicherten Entitäten können anschließend über das APIS-Web-Interface im Falle mehrerer Treffer disambiguiert werden. Dies erfolgt über eine Kartendarstellung, in welcher die verschiedenen Treffer zu einer Entität aufscheinen. Darüber hinaus können über das APIS-Web-Interface noch weitere Informationen zu den Entitäten ergänzt (z.B. alternative Schreibweisen, Datierungen) sowie die einzelnen Entitäten miteinander in typisierte Beziehungen gesetzt werden (z.B. Ort A ist Nachfolger von Ort B.; Ort A ist Teil von Ort B).

Die mit Hilfe von TEIHencer angereicherten Daten können dann wieder als XML/TEI Dokument (kodiert als <tei:listPlace> Element) exportiert bzw. über HTTP GET request abgerufen und so etwa in andere Applikationen eingebunden werden.

Im Zuge der Posterpräsentation soll der TEIHencer der einschlägigen DH-Comunity vorgestellt werden und zwar an dem konkreten Fallbeispiel der "Andreas Okopenko: Tagebücher aus dem Nachlass (Hybridedition)"[5]. Dabei handelt es sich um ein digitales Editionsprojekt, das eine Auswahl der Tagebücher Andreas Okopenkos im Zeitraum von 1949 bis 1955 inhaltlich erschließen und einem breiteren Publikum zugänglich machen möchte. Einer der Schwerpunkte des Projekts liegt hierbei auf der inhaltlichen Erschließung des örtlichen Wirkungs- und Schaffensraums des Nachkriegsavantgardisten, indem nicht nur erwähnte Orte (<tei:placeName>), sondern nach Maßgabe auch Werke und Organisationen (<tei:title> und <tei:orgName>) mit geographischen Normdaten verknüpft werden, um so ein umfassenderes Bild von Okopenkos kulturellem Kontext vermitteln zu können.

Neben der eigentlich Applikation und des konkreten Use-Cases wird am Poster auch das Konferenzthema "Kritik der Digitalen Vernunft" bzw. das Subthema "Kritik digitaler Angebote, Projekte und Werkzeuge" in Form der Frage nach der Nachhaltigkeit des vorgestellten Tools reflektiert. Eine solche glauben wir nämlich insofern gewährleisten zu können, als das Tool a) in ein konkretes Projekt (Okopenko) eingebettet ist, b) einen weit verbreiteten Standard (TEI) unterstützt, c) auf bestehende Eigenentwicklungen (APIS) aufbaut und d) Teile des Codes als selbstständige Module (TEI-Modul als python-package) konzipiert sind, die auch jenseits der konkreten Applikation Anwendung finden können. Darüber hinaus, e) ist der gesamte Code auf GitHub publiziert [3].

Fußnoten

1. http://www.getty.edu/research/tools/vocabularies/tgn/index.html
2. https://geobrowser.de.dariah.eu/edit/index.html
3. https://github.com/acdh-oeaw/teihencer
4. https://github.com/acdh-oeaw/apis-core
5. https://www.onb.ac.at/bibliothek/sammlungen/literatur/forschung/projekte/andreas-okopenko-tagebuecher-aus-dem-nachlass-hybridedition/

Text Mining und Computersimulation zur Analyse autobiographischer Texte: Einflüsse auf das literarische Schaffen Klaus Manns

Hess, Jan
hess@uni-trier.de
Trier Center for Digital Humanities (TCDH); Universität Trier

Lebherz, Daniel
lebherz@uni-trier.de
Center for Informatics Research and Technology (CIRT); Universität Trier

Zeyen, Christian
zeyen@uni-trier.de
Center for Informatics Research and Technology (CIRT); Universität Trier

„Ist das alles?“ (Mann 1995, S. 151) – Die vorwurfsvoll anmutende (rhetorische) Frage, die Klaus Mann Ende Juni 1933 unter einer Auflistung seiner Werke des zurückliegenden Halbjahrs in seinem Tagebuch notiert, gibt einen Hinweis darauf, wie selbstkritisch sich der Schriftsteller mit dem eigenen Schaffen auseinandersetzt. Angesichts der persönlichen und politischen Umstände sowie der Vielzahl an Kontakten und Aktivitäten, die er in seinem Journal verzeichnet, erscheint es fast etwas überraschend, dass die gut drei Monate nach seiner Emigration erstellte Werkliste immerhin 21 – wenn auch größtenteils kürzere – Texte umfasst. In Anbetracht seiner ständigen Rastlosigkeit stellt sich nicht nur die Frage, wann Klaus Mann überhaupt seiner eigentlichen schriftstellerischen Arbeit nachgeht, sondern auch, welche Faktoren zum Ge- oder Misslingen seines Schaffens beitragen. Wie wirken sich beispielsweise die beinahe allabendlichen Theater-, Kino- und Barbesuche oder die mehr oder weniger regelmäßig eingenommenen Rauschmittel auf seine literarische Produktivität aus? Welche Rolle spielt das tägliche Lektürepensum? Haben die zahlreichen Treffen, Telefonate und Korrespondenzen Einfluss auf seine schriftstellerische Arbeit? Oder erweisen sich die politischen Umstände, Angst, Verzweiflung und Hass gegenüber dem Nationalsozialismus als entscheidendes Vehikel literarischer Produktivität?

Aufgrund des Umfangs und der Detailgenauigkeit seiner alltäglichen Schilderungen von Gedanken und Aktivitäten – insbesondere auch der teilweise bis auf die Tageszeit genauen Protokollierung seiner Arbeitsprozesse – erscheinen Klaus Manns Tagebücher als idealer Untersuchungsgegenstand zur Beantwortung der aufgeworfenen Fragen. Gerade die Vielzahl und Komplexität der darin enthaltenen Informationen machen es jedoch schwierig, sich potentiellen Faktoren literarischer Produktivität tiefergehend analytisch zu nähern. Um diese große Anzahl an Informationen und deren Zusammenhänge besser bzw. überhaupt untersuchen zu können, sollen im Rahmen von *eXplore!* [1] am Fallbeispiel der Klaus Mann-Tagebücher verschiedene Methoden des Text Mining angewendet werden. Zur Unterstützung und Strukturierung der Analyseprozesse werden Scientific Workflows erstellt und ausgeführt. Auf Basis der mit Text Mining erzielten Ergebnisse sollen Methoden der Computersimulation eingesetzt werden, um die eingangs aufgeworfenen Fragen beantworten zu können (Abb. 1).

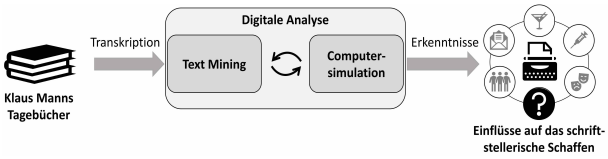


Abbildung 1: Forschungsprozess zur Analyse der Tagebücher Klaus Manns

Methoden der Computersimulation haben sich bereits seit längerer Zeit in verschiedenen Wissenschaftsdisziplinen etabliert. Seit den 1990er Jahren halten diese auch vermehrt Einzug in sozialwissenschaftliche Forschungsbereiche und bieten dort verschiedene Möglichkeiten zur Darstellung und Analyse von gesellschaftlichen Systemzusammenhängen (Gilbert 2007). Eine ähnliche Entwicklung lässt sich auch in den Digital Humanities nachvollziehen, wenngleich diese Methoden – insbesondere im literaturwissenschaftlichen Bereich – eine eher untergeordnete Rolle spielen (Kohle 2017). Die jedoch insgesamt zunehmende Anwendung von Computersimulation als Forschungsmethode liegt darin begründet, dass sie die Möglichkeit bietet, komplexe, unzugängliche Systeme zu untersuchen und experimentell zu analysieren. In solchen Experimenten lassen sich u.a. durch die Betrachtung verschiedener Szenarien Resultate erzielen, die

Rückschlüsse auf Plausibilitäten von Handlungen, Strukturen und Zusammenhängen im zugrundeliegenden System erlauben. Agentenbasierte Computersimulation bietet darüber hinaus insbesondere bei der Analyse von individuellem Entscheidungsverhalten die Möglichkeit, sogenannte emergente Effekte aufzudecken. Diese ergeben sich aus dem Zusammenspiel individueller Handlungen und Interaktionen von einzelnen Akteuren, lassen sich jedoch nicht ausschließlich durch diese erklären (Bonabeau 2002). Ein direkter Zusammenhang zwischen Systeminput und -output muss bei agentenbasierter Simulation also nicht bestehen. Dies steht im Gegensatz zu anderen statistischen Analyseverfahren, bei denen i.d.R. eine (lineare) Abhängigkeit zwischen verschiedenen Objekten vorausgesetzt wird. Ferner ermöglicht agentenbasierte Computersimulation detaillierte Analysen von Komponenten der Makroebene. Der Weg zu verlässlichen Ergebnissen einer Simulationsstudie führt in einem ersten, wesentlichen Schritt über die Modellbildung und Sammlung dafür relevanter Daten (Law 2015). Zu diesem Zweck soll das zugrundeliegende System, also konkret Klaus Mann, seine sozialen Kontakte und sein Umfeld möglichst realitätsnah mit allen wesentlichen Eigenschaften, Aktivitäten und Zusammenhängen in einer agentenbasierten Sozialsimulation (Davidsson 2002) abgebildet werden (Abb. 2).

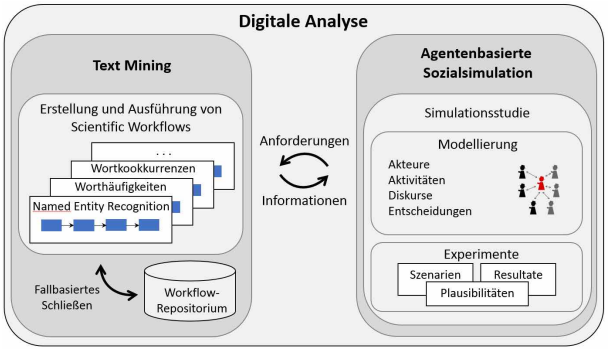


Abbildung 2: Digitale Analyse im Detail

Die dafür notwendige Generierung einer geeigneten Datengrundlage ist ein anspruchsvoller Prozess, in dem anhand von Methoden aus dem Bereich des Text Mining zunächst die wesentlichen, auf den Untersuchungsgegenstand einwirkenden Faktoren identifiziert und analysiert werden müssen. Konkret sollen dabei u.a. Methoden wie Named Entity Recognition, Worthäufigkeits- oder Kookkurrenzanalysen angewendet werden, um vorab Kontakte, Aktivitäten, Aufenthalte, Tätigkeiten, persönliche oder politische Ereignisse zu identifizieren, welche mit Klaus Manns literarischer Produktivität in Zusammenhang stehen könnten. Anders als im Falle der Computersimulation kommen Text Mining-Methoden, -Werkzeuge und Programmbibliotheken im literaturwissenschaftlichen Kontext bereits ungleich häufiger zum Einsatz. Zur Komplexitätsreduktion werden die zugrundeliegenden Text Mining-Prozesse dabei jedoch oft als Black Box betrachtet, was den Nachteil birgt, dass das Zustandekommen der erzielten Ergebnisse nur schwer nachvollzogen werden kann. Um dem entgegenzuwirken sollen die im Rahmen von *eXplore!* benötigten Text Mining-Prozesse explizit in Form von Scientific Workflows (Taylor et al. 2014) definiert und ausgeführt werden. Die in den Digital Humanities vergleichsweise wenig verbreiteten Scientific Workflows dienen der systematischen und transparenten Automatisierung wiederkehrender Datenverarbeitungsprozesse und haben sich in den Naturwissenschaften als geeignetes Mittel erwiesen, wissenschaftliche Prozesse zu strukturieren, zu dokumentieren und damit die Reproduktion und Validierung von Ergebnissen erheblich zu erleichtern.

Im Kontext von *eXplore!* sollen daher erfolgreich eingesetzte Workflows in einem Repositorium gespeichert und mithilfe von Fallbasiertem Schließen (Richter / Weber 2016), einer Methode zum erfahrungsbasierten Lösen von Problemen, wiederverwendbar gemacht werden. Ein darauf aufbauendes Tool soll die Erstellung von Workflows in der Praxis vereinfachen. Letztlich sollen somit nicht nur die literaturwissenschaftlichen Forschungsprozesse im Fallbeispiel Klaus Mann, sondern zukünftig auch ähnliche Text Mining-Problemstellungen sowie die Modellierungsarbeit in Simulationsstudien zur Erforschung weiterer autobiographischer Texte sinnvoll unterstützt werden (Abb. 2).

Fußnoten

1. *eXplore!*: *Computergestützte Modellierung, Analyse und Exploration als Grundlage für eScience in den eHumanities* (Fz: 01UG1606) wird vom Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Bibliographie

Bonabeau, Eric (2002): “Agent-based modeling: Methods and techniques for simulating human systems” in: *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280– 7287

Davidsson, Paul (2002): "Agent based social simulation" in: *Journal of Artificial Societies and Social Simulation* 5(1): http://jasss.soc.sur-

rey.ac.uk/5/1/7.html [letzter Zugriff 11.Januar 2017]

Gilbert, Nigel (2007): *"Computational social science: Agent-based social simulation"* in: Amblard, Frédéric / Phan, Dennis (eds.): Agent-based Modelling and Simulation in the Social and Human Sciences. Oxford: The Bardwell Press 115-133.

Kohle, Hubertus (2017): *"Digitale Rekonstruktion und Simulation"* in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): Digital Humanities: Eine Einführung. Stuttgart: J. B.Metzler 315-327.

Law, Averill M. (2015): Simulation Modeling and Analysis. 5.Auflage. Tucson: McGraw Hill Education.

Mann, Klaus (1995): Tagebücher 1931 – 1933. Reinbek: Rowohlt.

Richter, Michael M. / Weber, Rosina O. (2013): Case-Based Reasoning: A Textbook. Berlin: Springer.

Taylor, Ian J. / Deelman, Ewa / Gannon, Dennis B. / Shields, Matthew (2014): Workflows for e-Science: Scientific Workflows for Grids.London: Springer

Universal Morphology zwischen Sprachtechnologie und Sprachwissenschaft: Sprachressourcen für Kaukasussprachen

Chiarcos, Christian
chiarcos@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Donandt, Kathrin
donandt@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Ionov, Maxim
ionov@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Rind-Pawlowski, Monika
pawlowski@lingua.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Sargsian, Hasmik
Sargsyan@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Wichers Schreur, Jesse
wichersschreur@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Hintergrund

Das Projekt `Linked Open Dictionaries' (LiODi, 2015-2020) ist eine vom Bundesministerium für Bildung und Forschung (BMBF) finanzierte eHumanities-Forschungsgruppe, die daran arbeitet, einen nutzerorientierten Zugang zu LOD-Technologien in den Sprachwissenschaften zu entwickeln und diesen in Einzelstudien zum Sprachkontakt im Kaukasus zu demonstrieren. Ein wichtiges Element dafür sind Lehnwortuntersuchungen, und in morphologisch reichen Sprachen ist es möglich, dass eine flektierte Form Gegenstand der Entlehnung war. Diese automatisch gestützt generieren und identifizieren zu können, erfordert einen morphologischen Generator, der auch über unvollständige Daten hinweg generalisieren, und beispielsweise Paradigmen komplettieren kann. Hierfür stellt die Universal Morphology derzeit Standardressourcen bereit, auf die hin Softwareimplementierungen optimiert werden, etwa im Rahmen aktueller SIGMORPHON Shared Tasks (Cotterell et al., 2016).

Universal Morphology (*Unimorph*, http://unimorph.github.io/) ist ein aktuelles Communityproject zur Erfassung und automatischen Generierung der Flexionsmorphologie unterschiedlichster Sprachen. Ziel ist sowohl die Entwicklung von Vollformenwörterbüchern, deren Einsatz zur Annotation, weshalb Unimorph auf Kompatibilität mit den Universal Dependencies (http://universaldependencies.org/) angelegt ist, aber auch der Aufbau einer Referenzressource zur Entwicklung morphologischer Analyse- und Generierungskomponenten innerhalb der Sprach *technologie*. Die sprach *wissenschaftliche* Nutzung jedoch steht bislang aus und bestimmt daher den Fokus unseres Beitrages. Unimorph-Ressourcen und -Technologien sind dabei eine höchst willkommene Ergänzung unserer Arbeit, ihre praktische Anwendung des Schemas auf das Kaukasusgebiet erweist sich jedoch als problembehaftet.

Der **Kaukasus** ist für die Diversität seiner Sprachen bekannt, die oftmals eurozentristische Ansichten traditioneller Sprachwissenschaften infrage gestellt haben, und sich daher sehr gut zur Prüfung von linguistischen Modellen mit universellem Anspruch eignen. Viele Kaukasussprachen

sind bedroht, die meisten (mit Ausnahme von Georgisch, Armenisch und Albanisch/Udi) wurden erst in jüngerer Vergangenheit verschriftlicht. Allen gemeinsam ist ein großer Lehnwortschatz (u.a. aus dem Iranischen, Türkischen, Russischen) und Lehnbeziehungen untereinander. Alle Kaukasussprachen sind sprachtechnologisch schlecht erschlossen, hier betrachten wir daher aktuelle Ansätze zur Schaffung von sprachübergreifenden (`universellen') morphologischen Annotationen im Rahmen des *Unimorph*-Projektes. Wir berichten Ergebnisse zu unserer Arbeit zu Batsbi (nakh-dagestanisch), Mingrelisch (kartvelisch), Khinalug (nakh-dagestanisch) und Armenisch (indoeuropäisch). Auf dieser Basis diskutieren wir behutsame Erweiterungen von Unimorph, um dessen Anwendbarkeit für die kaukasischen Sprachen im Besonderen und für Sprachdokumentationsdaten im Allgemeinen zu ermöglichen.

Unimorph für Sprachdokumentation im Kaukasus?

Schema und Schemaerweiterungen

Unimorph verwendet ein **TSV-Format**, d.h., eine Liste von tab-separierten Einträgen für jeweils Wortform, Lemma und Unimorph-Tags. Letztere sind *nicht qualifizierte* Merkmale, durch Semikolon getrennt und *unsortiert*. Der Eintrag für deutsch „(ich) treffe (dich)“ wäre beispielsweise

```
treffen  treffe  V;IND;PRS;1;SG
```

Der Eintrag für mingrelisch *kešerxvaduk* (`Ich werde dich treffen') besitzt folgende Glosse:

```
kešerxvaduk
ke-   še-   r-    xvad  -u    -k
AFF   PV    O2SG  meet  TM    S1SG
```

In Unimorph wird diese Analyse wie folgt repräsentiert:

```
xvad  kešerxvaduk  AFF;LGSPEC4;ARGDA2S;V;LGSPEC6;ARGNO1S
```

Das mingrelische Verb kongruiert mit beiden syntaktischen Argumenten: dem Subjekt (1S, Nominativ) und dem Objekt (2S, Dativ), für die **zusammengesetzte Merkmale** gebildet werden, die ein Argument mit dessen Person, Numerus usw. zusammenstellt; z.B. werden hier ArgNo1S für „Nominativargument=1. Person Singular“ und ArgDa2S für „Dativargument=2. Person Singular“ aufgeführt. Ein methodisches Problem ist, dass das Unimorph-Schema dieselbe Information hierbei in unterschiedlicher Weise ausdrückt, wie im Vergleich deutlich wird: mingrelisch ArgNo1S entspricht deutsch 1;SG. Da der Zusammenhang zwischen Kasus und grammatischen Rollen für das Deutsche nicht explizit definiert ist, gibt es keine Möglichkeit, diese automatisch als äquivalent zu interpretieren. Leider erlaubt es Unimorph zudem nicht, herkömmliche Terminologie zu verwenden, in der beide Argumente bzgl. ihrer syntaktischen Rollen (`Subjekt' und `Objekt') beschrieben werden, sondern zieht stattdessen die Kasusmorphologie heran. Dies ist insofern problematisch, als der Kasus im Verb nicht morphologisch realisiert ist, und Argumente im Satz nicht (pro)nominal realisiert werden müssen. Konventionell werden statt dessen grammatische Rollen verwendet.

Ähnlich zu (mehreren Argumenten von) Verben gibt es Nomina mit **mehrfacher Kodierung derselben Merkmalskategorie** (v.a. Kasus), die mit Unimorph nicht behandelt werden können. In der sog. *Suffixaufnahme* spezifizieren adnominale Elemente *neben* ihrem inhärenten Kasus auch Agreement-Merkmale ihres Kopfnomens, z.B. durch Wiederholung von dessen Kasusmorphologie. Dies wurde ursprünglich für Georgisch beschrieben, gilt aber als verbreitet im Kaukasus. Bedauerlicherweise kann diese Information in Unimorph nicht positional kodiert werden, sondern erfordert eine Erweiterung des Label-Inventars. Deshalb schlagen wir die Einführng numerischer Indizes in der nominalen Morphologie vor, wobei der inhärente Kasus nicht bezeichnet wird, der Kasus des direkten Kopfes durch Anhängen von -1 an das Feature-Label, der Kasus von dessen Kopf durch -2, usw.

Diese **numerischen Indizes** sind auch auf die verbale Domäne übertragbar. Gegeben etablierte Hierarchien grammatischer Rollen bzw. der zugeordneten Kasus, kann die bisherige Verbundmarkierung multipler Argumente durch eine Indizierungsstrategie ersetzt werden, die sich auf diese bezieht, und bei der das höchstrangige Element (z.B. das Subjekt) unbezeichnet bleibt, während andere Argumente nach ihrer Stellung im Ranking gekennzeichnet werden. Eine alternative Repräsentation des mingrelischen *kešerxvaduk* wäre also

```
V;... 1;SG; 2-1;SG-1
```

Dies korrigiert auch die Asymmetrie zwischen zusammengesetzten und individuellen Merkmalen. Auch die Zuschreibung mehrerer Merkmale einer Kategorie kann nominal und verbal einheitlich gehandhabt werden, und die Vergleichbarkeit über Sprachen hinweg wird vereinfacht.

Datenformat und Alternativen

Ein zweites Problem ist das in Unimorph verwendete, nicht erweiterbare TSV- **Format**, das gegenüber den in der Sprachdokumentation üblichen Softwarelösungen (FLEx, Toolbox, ELAN) aber nur *stark eingeschränkte* Informationen bereitstellt: Im Vergleich zur wörterbuchgestützten Interlinearglossierung stellen Morphem-Inventorien in unvollständigen und weniger gut interpretierbaren Repräsentationen ein Akzeptanzproblem dar. Daher sollte Unimorph nicht als eigenständiger Formalismus verstanden werden, sondern als Austauschformat zwischen reichen und hochwertigen Sprachressourcen auf der einen Seite und morphologischen Generatoren auf der anderen.

Allerdings sind *Format und Speicherort* festgeschrieben, so dass zugrundeliegende Ressourcen an anderen Orten gespeichert und gepflegt werden müssen, und Ergänzungen aus der Sprachdokumentationsarbeit womöglich nicht eingepflegt werden. Wir schlagen daher vor, das jetzige Format nur bei Bedarf zu generieren. Der Schlüssel hierzu liegt darin, die Quellformate gemäß W3C-Standards zur Ressourcentransformation auf einheitliche RDF-Datenstrukturen nach lemon (https://www.w3.org/2016/05/ontolex/) zu mappen und mit Hilfe der Anfragesprache SPARQL das derzeitige Tabellen-Format zu erzeugen. Das Repository enthält dann für jede Sprache die (a) *vollständigen* Daten, und (b) ein standardisiertes Mapping auf lemon. Die TSV-Generierung ist nicht ressourcenspezifisch. Der Gebrauch von RDF-Technologien für Datenkonversion und Abfrage kann so die Entwicklung einer technischen Infrastruktur für Unimorph ermöglichen, die es erlaubt, über die Grenzen des TSV-Formats hinauszuwachsen, wovon SprachwissenschaftlerInnen, ForscherInnen und NLP-IngenieurInnen, die mit *low-resource*-Sprachen arbeiten, profitieren könnten.

Die Integration mit gängigen Annotationswerkzeugen kann hierbei auf von uns entwickelten RDF-Konvertern für FLEx, Toolbox und weitere Formate aufsetzen (Chiarcos et al., 2017, https://github.com/acoli-repo/LLODifier).

Zusammenfassend plädieren wir für die Einführung numerischer Indizes für verschiedene Argumente polyvalenter Verben und rekursive Merkmale in der Nominalflexion in Unimorph. Für die bessere Integration von existierenden Ressource aus der Sprachdokumentation insgesamt schlagen wir zudem eine Erweiterung der unterstützten Formate und einen einheitlichen Zugriff auf diese auf Basis von RDF-Technologien vor, so dass das jetzige Unimorph-Format nicht mehr von den zugrundeliegenden, reicheren Quelldaten separiert wird, sondern bedarfsabhängig daraus generiert wird. Sind beide Mängel behoben, steht einer sprachwissenschaftlichen Nutzung von Unimorph hinsichtlich der Sprachkontaktforschung im Kaukasus nichts mehr entgegen.

Bibliographie

Chiarcos, Christian / **Ionov, Maxim** / **Rind-Pawlowski, Monika** / **Fäth, Christian** / **Wichers Schreur, Jesse** / **Nevskaya, Irina** (2017): “LLODifying linguistic glosses” in: *Proceedings of the First International Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 2017*. Springer (Lecture Notes in Artificial Intelligence (LNAI)), 89-103 https://doi.org/10.1007/978-3-319-59888-8_7 [letzter Zugriff 14. Januar 2018]

Cotterell, Ryan / **Kirov, Christo** / **Sylak-Glassman, John** / **Yarowsky, David** / **Eisner, Jason** / **Hulden, Mans** (2016). “The SIGMORPHON 2016 Shared Task Morphological Reinflection” in: *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, Berlin, Germany, August, 2016*. Association for Computational Linguistics, 10-22 http://anthology.aclweb.org/W16-2002.pdf [letzter Zugriff 14. Januar 2018]

Sylak-Glassman / **John** (2016). “The composition and use of the universal morphological feature schema (UniMorph schema)”. Technical report, Department of Computer Science, Johns Hopkins University, working draft, v.2, https://unimorph.github.io/doc/unimorph-schema.pdf [letzter Zugriff 14. Januar 2018]

VedaWeb – eine webbasierte Plattform für die Erforschung altindischer Texte

Reinöhl, Uta
uta.reinoehl@uni-koeln.de
Universität zu Köln

Kölligan, Daniel
d.koelligan@uni-koeln.de
Universität zu Köln

Kiss, Börge
kiss@mailbox.org
Universität zu Köln

Mondaca, Francisco
f.mondaca@uni-koeln.de
Universität zu Köln

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln

Sahle, Patrick
sahle@uni-koeln.de
Universität zu Köln

Einleitung

Das hier beschriebene Poster thematisiert den technischen Entwurf und den funktionalen Aufbau von *VedaWeb*, einer webbasierten Plattform für die sprachwissenschaftliche Erforschung altindischer Texte (siehe http://vedaweb.uni-koeln.de). Das 2017 begonnene Vorhaben wird als Kooperationsprojekt an der Universität zu Köln durchgeführt, in enger Zusammenarbeit zwischen Fachwissenschaftlern des Instituts für Linguistik (Allgemeine Sprachwissenschaft, Historisch-Vergleichende Sprachwissenschaft und Sprachliche Informationsverarbeitung) sowie des Cologne Center for eHumanities (CCeH). Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) in der LIS-Förderlinie „eResearch-Technologien“ gefördert.

Über die *VedaWeb*-Plattform werden altindische, in Sanskrit verfasste Texte morphologisch und metrisch annotiert, sowie nach lexikographischen und korpuslinguistischen Kriterien durchsuchbar zur Verfügung gestellt. Als Pilottext dient zunächst der *Rigveda*, einer der ältesten und wichtigsten Texte der indogermanischen Sprachfamilie. Der *Rigveda* ist in der ältesten Sprachform des Altindischen, dem Vedischen, verfasst. Seine Entstehung kann auf das späte zweite Jahrtausend v. Chr. datiert werden. Mit einem Umfang größer als Homers *Ilias* und *Odyssee* zusammen stellt er eine überaus reiche Datengrundlage dar. Perspektivisch sollen auch weitere Texte wie etwa der *Atharvaveda*, *Yajurveda* und vedische Prosatexte in die *VedaWeb*-Plattform integriert werden. Das Projekt wird Forschungen in allen Bereichen des Vedischen erleichtern und voranbringen, beispielsweise in Bezug auf Fragestellungen der Syntax (siehe z.B. zu referentiellen Null-Objekten Keydana/Luraghi 2012, zu Nicht-Konfigurationalität Reinöhl 2016), der Morphologie (siehe z.B. zum vedischen vṛkī-Typus Widmer 2007, zu ya-Präsentien Kulikov, 2012) oder der Wortbildung (siehe z.B. zu Komposita Scarlata/Widmer 2015). Es wird angestrebt, die *VedaWeb*-Plattform längerfristig zur zentralen Anlaufstelle für die internationale Fachgemeinschaft, die mit altindischen Primärtexten arbeitet, auszubauen, um den in Köln bestehenden Schwerpunkt auf südasiatische Sprachen weiter zu stärken.

Projektziele

Ausgangspunkt und Grundlage des Projekts ist eine vollständige morphologische (d.h. wortstrukturelle) Annotation des *Rigveda*, die im Vorfeld an der Universität Zürich durchgeführt und dem Projekt zur Verfügung gestellt wurde. Hinzu kommen metrische Informationen (Kevin Ryan, Harvard University, siehe http://www.meluhha.com/rv/) sowie perspektivisch auch syntaktische Informationen aus verschiedenen abgeschlossenen und andauernden Forschungsprojekten als weitere Annotationsebenen. Anhand dieser Annotationen werden im Projekt verschiedene Recherche- und Analysewerkzeuge entwickelt und sukzessive in die *VedaWeb*-Plattform integriert. Hierzu gehören eine kombinierte Suchfunktion nach linguistischen Parametern (u.a. Lemmata, Wortformen, morphologische und metrische Informationen), die Verknüpfung mit dem Standardwörterbuch zum *Rigveda* von Hermann Grassmann (1873), die Anzeige von Übersetzungen (u.a. Grassmann 1876, Geldner 2003, Griffith 1896) und von Kommentaren (Oldenberg 1909/1912), sowie die Möglichkeit des Exports von annotierten Textabschnitten nach vom Nutzer gewählten Kriterien.

Von zentraler Bedeutung ist die Verknüpfung des *Rigveda* mit der am CCeH angesiedelten Portalseite für Sanskritwörterbücher (*Cologne Digital Sanskrit Dictionaries*, siehe http://www.sanskrit-lexicon.uni-koeln.de), einer zentralen Anlaufstelle für die internationale Sanskritforschung. Auf Basis einer Modellierung der Daten in TEI werden die Wortformen über die jeweiligen Lemmata mit dem digital erfassten Wörterbuch von Grassmann verknüpft, so dass sowohl vom Text auf das Wörterbuch verwiesen wird als auch umgekehrt vom Wörterbuch aus Textstellen aufgesucht werden können. Auf diese Weise kann über das Lemma gleichzeitig auch eine Verknüpfung zu den weiteren Sanskrit-Wörterbüchern hergestellt werden, etwa um vergleichende, wörterbuchübergreifende Recherchen zu ermöglichen. In der direkten Verknüpfung von Text und Wörterbuch und der damit verbundenen Erweiterung der Analysemöglichkeiten besteht das wesentliche Alleinstellungsmerkmal des Projekts gegenüber bestehenden Ressourcen des Altindischen wie z.B. dem *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS, siehe http://titus.uni-frankfurt.de).

Technische Aspekte

Der Schwerpunkt des Posters liegt auf der Präsentation des Systementwurfs der *VedaWeb*-Plattform sowie der dort eingesetzten Technologien. Dies umfasst zum einen eine Beschreibung der funktionalen Elemente der Nutzerschnittstelle, die dem Anwender als Forschungsumgebung dient, indem sie verschiedene Werkzeuge bereitstellt (z.B. Suche, Verlinkung, Export in TEI-Format). Zum anderen werden die Systemarchitektur und die für deren Umsetzung verwendeten Technologien thematisiert. Die *VedaWeb*-Plattform wird als Webanwendung auf Basis des *Spring-Frameworks* umgesetzt (siehe https://spring.io). Durch dessen weite Verbreitung und den großen Funktionsumfang sind sowohl die langfristige Wartbarkeit als auch zahlreiche Möglichkeiten zur späteren Erweiterung der Plattform gewährleistet. Für die Umsetzung der Suchlogik wurde zunächst der Einsatz von etablierten, auf *Lucene* (siehe http://lucene.apache.org) aufbauenden Such-Servern wie *Solr* oder *Elasticsearch* geprüft, jedoch wieder verworfen, da ihre Stärken v.a. im Durchsuchen sehr umfangreicher, dabei aber verhältnismäßig einfach strukturierter Textsammlungen liegen. Die *VedaWeb*-Suche hingegen soll mittels komplexer, kombinierbarer Suchkriterien über die reine Volltextsuche hinaus auch linguistisch motivierte Suchanfragen auf Grundlage der verschiedenen Annotationsebenen zulassen. Anstelle der genannten Such-Server wird deshalb eine eigene, auf den gegebenen Anwendungsfall zugeschnittene Suchlogik direkt mit *Lucene* implementiert. Das Frontend wird unter Verwendung einer Template Engine (*Thymeleaf,* siehe http://www.thymeleaf.org) und Client-seitigen JavaScript-Technologien umgesetzt, um eine möglichst flexible und effiziente Arbeitsweise zu ermöglichen. Mit der Fokussierung auf den Systementwurf möchten wir mit dem Poster vor allem einen kompakten Überblick über die Nutzungsmöglichkeiten sowie über die technische Funktionsweise der VedaWeb-Plattform geben.

Bibliographie

Geldner, Karl Friedrich (2003) [1951-57]: *Der Rig-Veda. Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen von Karl Friedrich Geldner*. Cambridge (Mass.): Harvard University Press.

Grassmann, Hermann (1873): *Wörterbuch zum Rig-veda*. Wiesbaden, O. Harrassowitz.

Grassmann, Hermann (1876): *Rig-veda. Übersetzt und mit kritischen und erläuternden Anmerkungen versehen von Hermann Grassmann*. Leipzig: F.A. Brockhaus.

Griffith, Ralph T. H. (1896): *The Hymns of the Rigveda*. Benares: Lazarus.

Keydana, Götz / Luraghi, Silvia (2012): *Definite referential null objects in Vedic Sanskrit and Ancient Greek*. Acta Linguistica Hafniensia 44 (2):116–28. https://doi.org/10.1080/03740463.2013.776245 .

Kulikov, Leonid (2012): *The Vedic -ya-Presents: Passives and Intransitivity in Old Indo-Aryan*. Amsterdam, Netherlands: Rodopi.

Oldenberg, Hermann (1909/1912): *Ṛgveda. Textkritische und exegetische Noten*. Berlin: Weidmann.

Reinöhl, Uta (2016): *Grammaticalization and the Rise of Configurationality in Indo-Aryan.* Oxford: Oxford University Press

Scarlata, Salvatore / Widmer, Paul (2015): *Vedische exozentrische Komposita mit drei Relationen*. Indo-Iranian Journal, 58(1):26-47.

Widmer, Paul (2007): *Der altindische vrkî-Typus und hethitisch nakki-: Der indogermanische Instrumental zwischen Syntax und Morphologie.* Zeitschrift für Sprachwissenschaft, (1-2):190-208.

Verhaltensmuster in Massendiskursen: Ein Opinion Dynamics - Modell

Heckelen, Malte
malte.heckelen@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Seit der letzten Präsidentschaftswahl in den Vereinigten Staaten und der Verbreitung rechten Gedankenguts in Europa stellt sich für Forscher die interessante und wichtige Frage nach dem "Warum?". Phänomene wie die Verbreitung radikaler, eigentlich unbeliebter Meinungen werden in der Presse mit Social Media und ihren technisch bedingten Dynamiken in Verbindung gebracht. Daneben sind es aber auch die auf das Diskursverhalten bezogenen Normen und strategisches Verhalten, die die Entwicklung des Meinungsbildes in Massendiskussionen prägen.

Die sozialpsychologische Persuasionsforschung untersucht die Determinanten der Überzeugung auf der Mikroebene von Kleingruppen. Opinion Dynamics - Simulationsmodelle hingegen befassen sich mit der Verbindung der Mikroebene zur Makroebene: mit bewusst einfach konzeptualisierten Agenten als "Black Box", deren simple Botschaftenübertragungen interagieren und zu emergenten Meinungsverteilungen und -dynamiken führen. Um theoretische Konzepte der Überzeugung von Individuen aggregierend auf Massendynamiken der Meinungslandschaft beziehen zu können, werden in diesem Projekt entscheidungstheoretische Modelle der Soziologie und Persuasionsforschung mit Opinion Dynamics - Simulationen verbunden (Agentenbasiertes Modell).

Das agentenbasierte Modell stattet Agenten mit einer kognitiven Architektur aus, die auf Dual Process - Modellen der Meinungsbildung und der strategischen Handlungswahl basiert. Dual Process - Modelle repräsentieren Entscheidungen in zwei Modi: einem langsamen, elaborierten Modus und einem schnellen, heuristischen Modus. Im Modell "wollen" alle Agenten eine homogene Meinungslandschaft erzeugen und verfolgen dabei abhängig von ihrer Energie und Involvierung innerhalb der zwei Modi verschiedene Sender- und Empfängerstrategien. Die Agenten interagieren auf typischen Netzwerktypen mit verschiedener Parametrisierung. Eine geplante Erweiterung ist die eines Parameters für Modi der Botschaftenübertragung, die an typische Social Media - Mechanismen (etwa Filtering) angelehnt sind. Das Modell kann somit potenziell typische Dynamiken des Meinungsaustauschs nachbilden und auf problematische Konstellationen hinweisen, etwa im Sinne einer Verbreitung radikaler Meinungen oder der Bildung nicht mehr interagierender Fraktionen. Dem Diskurs externes Wissen, dass unter Umständen im Sinne einer "fake news" bewusst falsch dargestellt wird, kann in diesem Modell nicht abgebildet werden.

Dual Process - Framework für das Opinion Dynamics - Modell

Das entwickelte Modell baut auf Dual Process - Modellen der Persuasionsforschung und der handlungstheoretisch orientierten Soziologie auf. Das soziologische Modell der Frame-Selektion (Kroneberg 2010, Esser 1996) konzeptualisiert "bewusstes" Entscheiden und "unbewusstes" Verhalten in einem Rational Choice - Metaframework: Ob Informationen und Handlungsmöglichkeiten rational verarbeitet oder anhand sozial geteilter Frames und Skripte automatisch-spontan ausgewählt werden, hängt von der vorhandenen Energie und der Einordenbarkeit der Situation in durch Erfahrung gebildete Klassen ab. Das Elaboration-Likelihood-Modell (Petty und Cacioppo 1986) und das Systematic-Heuristic-Model (Chaiken und Eagly 1989) aus der Sozialpsychologie sehen die Verarbeitung persuasiver Botschaften ebenfalls auf zwei Wegen, bedingt durch Energie und Involviertheit: Verarbeitung der argumentativen Struktur (systematisch) und Orientierung an Sender- und Botschaftenattributen (heuristisch).

Das vorliegende Modell integriert diese Sichtweisen zu einem handlungstheoretischen Blick auf Massendiskurse: Individuen maximieren die Meinungsähnlichkeit in ihrer Umgebung und minimieren den aus ihrem Handeln resultierenden Energieverlust. Agenten wählen entweder kalkulierend-strategisch oder normbezogen-automatisch Aktionen der Botschaftenrezeption und -produktion, um die Meinungsähnlichkeit in ihrer Umgebung zu maximieren. Lokale Diskussionsnormen bilden sich situationsabhängig als automatisch-spontan gewählte Handlungen durch wiederholte rationale Wahl im Sinne des Reinforcement Learning.

Computermodell

Das Computermodell basiert auf Bounded Confidence – Modellen (Deffuant et al. 2000, Hegselmann/Krause 2002): Agenten verfügen über eine Ziffer als "Meinung" und passen ihre Meinungen um einen parameterabhängigen Grad an und zwar abhängig davon, dass die Differenz der Meinungen den Wert eines Confidence-Parameters unterschreitet. Diese Agenteninteraktion wird im vorliegenden Modell erweitert: Agenten rezipieren Botschaften entweder im systematischen Modus oder im heuristischen Modus. Der systematische Modus ist dem klassischen Bounded Confidence - Modell im Ablauf gleich. Im heuristischen Modus hängt die Änderung der Agentenmeinung von Senderattributen wie Degree, der Ähnlichkeit von diskreten Agenteneigenschaften und der Ähnlichkeit der gesendeten Meinung zu Nachbarschaftsmeinungen ab. Agenten können Meinungen unverändert senden oder sie an die Botschaften des Empfängers geringfügig anpassen, wobei beide Varianten mit zusätzlichen Referenzen auf Botschaftenqualitäten (Popularität, Meinungsähnlichkeit zu Rezipientenpeers u.a.) versehen werden können.

Alle Entscheidungen werden über die Maximierung von Nutzenfunktionen im rc-Modus oder die automatische Wahl im as-Modus getroffen. Der as-Modus der Handlungswahl wird mittels eines mit einem adaptiven k-means-Clusteralgorithmus gekoppelten Q-Learning-Algorithmus implementiert (Karimpanal und Wilhelm 2017, Wen u.a. 2006). Kann eine neue "Situation" (Vektor aus Umgebungsparametern) nicht eindeutig einem bereits bekannten Cluster zugeordnet werden, wird sie im rc-Modus verarbeitet. Momentane Energie und Involviertheit (Energieinvestition über Zeit) sind bei der Moduswahl jedes Entscheidungsschritts weitere Parameter, was zu realistischen Mustern aktiven und passiven Verhaltens führen sollte.

Projektfortschritt

Das Modell ist zurzeit im Java-Framework Repast Symphony mit voller Funktionalität des rc-Modus implementiert. Das geplante Poster wird die Ziele des Projekts, die Funktionsweise des Computermodells sowie erste Analysergebnisse vorstellen.

Bibliographie

Chaiken, Shelly / Eagly, Alice H. Eagly (1989): "Heuristic and Systematic Information Processing Within and Beyond the Persuasion Context." In: Uleman, J.S. / Bargh, J.A.: *Unintended Thought.* New York: Gilford, 212–52.

Deffuant, Guillaume / Neau, David / Amblard, Frederic / Weisbuch, Gérard (2000): „Mixing Beliefs among interacting agents." *Advances in Complex Systems 3* (1-4): 87-98.

Esser, Hartmut. (1996): "Die Definition der Situation." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 48 (2): 1–34.

Hegselmann, Rainer / Krause, Ulrich (2002): "Opinion dynamics and bounded confidence. Models, Analysis and Simulation." *Journal of Artificial Societies and Social Simulation 5 (3). URL: http://jasss.soc.surrey.ac.uk/5/3/2.html.*

Karimpanal, Thommen George, / Wilhelm, Erik (2017): "Identification and Off-Policy Learning of Multiple Objectives Using Adaptive Clustering." *Neurocomputing* 263: 39-47.

Kohle, Hubertus (2017): "Digitale Rekonstruktion und Simulation". In: Jannidis, Fotis / Kohle, Hubertus (eds.): *Digital Humanities. Eine Einführung*. Stuttgart: J.B. Metzler, 315-327.

Kroneberg, Clemens. (2010): "Das Modell Der Frame-Selektion: Grundlagen Und Soziologische Anwendung Einer Integrativen Handlungstheorie." PhD, Universität Mannheim.

Petty, Richard E., und John T. Cacioppo. (1986): "The Elaboration Likelihood Model of Persuasion." In: *Advances in Experimental Social Psychology* 19: 123–205.

Wen, Feng / Chen, Zonghai / Zhuo, Rui / Zhou, Guangming (2006): "Reinforcement Learning Method of Continuous State Adaptively Discretized Based on K-Means Clustering." *Control and Decision* 21 (2): 143-146.

Virtuelle Ausstellungen und Rundgänge: digitalisiertes Kulturerbe vermitteln und präsentieren

Steiner, Elisabeth
elisabeth.steiner@uni-graz.at
Universität Graz, Österreich

Ausgangslage

Virtuelle Rundgänge, virtuelle Ausstellungen oder virtuelle Touren: Die Begriffe bezeichnen oft ähnliche Vorgehensweisen. Zusätzlich zu struktu-

rierten Suchzugängen werden sie auf den unterschiedlichsten Portalen für digitalisiertes Kulturerbe in der Regel als Ergänzung angeboten. Dazu gehören prominente Beispiele wie die Europeana [1], die Deutsche Digitale Bibliothek [2] oder zahlreiche Regionalportale im deutschsprachigen Raum (stellvertretend sei Bavarikon [3] oder Kulturerbe Niedersachsen [4] genannt). Doch auch kommerziellen Anbietern ist das Konzept nicht fremd, wie Google zeigt. [5] Am häufigsten wird die Bezeichnung (Virtuelle/Digitale) Ausstellung ((*virtual/digital) exhibition*) verwendet.

Unabhängig von der Begrifflichkeit teilen sich die Umsetzungen meist folgende Merkmale (vgl. auch INDICATE 2012, 17):

- es handelt sich um Sammlungen von Informationen und Quellen/Objekten zu einem bestimmten Thema, einer Periode oder einer Person (oft auch anlassbezogen zu Jubiläen, z.B. Reformation [6], Weltkrieg [7], etc.)
- die Quellen werden von Fachleuten kontextualisiert und beschrieben
- verschiedene Inhalte (Text, Bild, Audio, Video) werden miteinander verbunden
- die Objekte können aus einem oder mehreren digitalisierten Sammlungen oder Beständen kommen
- meist gibt es einen vorgegebenen (zumindest bevorzugten) Ablauf der Informationen

So vielfältig wie die möglichen Themen der Rundgänge sind die technischen Umsetzungen und Visualisierungen. Zahlreiche Werkzeuge sind für die Erstellung und Darstellung von Virtuellen Ausstellungen verfügbar. [8] Bei der Auswahl des Werkzeuges sind die gewünschte Funktionalität, Kompatibilität mit bestehender Infrastruktur, Unabhängigkeit von äußeren Quellen wie auch eine möglichst geringe Einarbeitungszeit wichtige Kriterien. Eine Wahrung der Prinzipien der Digital Humanities (vor allem mit Hinblick auf open source Software) ist ebenso wünschenswert.

Ein Vorbildprojekt ist AthenaPlus, das nicht nur theoretische und methodische Vorarbeit im Bereich Virtuelle Ausstellungen geleistet hat, sondern mit Movio auch ein open source Tool entwickelt hat.

Umsetzung im Portal „Kultur- und Wissenschaftserbe“ Steiermark

Auch im Webportal „Kultur- und Wissenschaftserbe Steiermark“, einem Ergebnis des Projektes „Repositorium Steirisches Wissenschaftserbe“, ergänzen Virtuelle Rundgänge den strukturierten Suchzugang. Das Konzept zu den Touren wurde gemeinsam mit den Datenlieferanten erarbeitet. Die kuratierten Rundgänge erlauben eine spielerische Auseinandersetzung mit den Objekten und ein „Virtuelles Schlendern“ durch Bestände.

Für die Umsetzung wurde die JavaScript-Bibliothek StoryMapJS verwendet, die in Form eines eigenen Objektmodells in die dem Webportal zu Grunde liegende Infrastruktur GAMS integriert wurde. Die Wahl fiel deswegen auf StoryMapJS, weil es sich aus den zahlreich verfügbaren Werkzeugen am besten in das bestehende System implementieren ließ. Viele Werkzeuge bieten ein Rundumpaket von Content Management System bis Präsentation, für die Integration in GAMS war ein reines Disseminationswerkzeug zu bevorzugen. Das Objektmodell kapselt den Inhalt der Ausstellung (Text, Medienreferenzen, Ablauf) und übergibt die Daten an die Javascript-Bibliothek, die die Darstellung übernimmt. So sind alle verwendeten Daten unabhängig von der Anwendung Teil des digitalen Archivs und können somit dauerhaft adressiert sowie gemeinsam verwaltet und (langzeit)archiviert werden. Eine persistente Identifikation der Ausstellung ist auf Basis von Handle möglich.

Derzeit wird der für StoryMapJS verlangte JSON Input eins zu eins aus einer XML Datei konvertiert. Das bringt eine teilweise Vermischung von Form und Inhalt mit sich. Um dem entgegen zu wirken, wäre ein spezielles Set zur Beschreibung der Meta(Daten) der Virtuellen Ausstellung sinnvoll. Teilweise wird das durch DEMES (*Digital Exhibition Metadata Element Set*) ermöglicht, allerdings werden hier stets die Metadaten der gesamten Ausstellung beschrieben, eine Metabeschreibung der einzelnen Elemente, des Inhalts wie auch von Ablauf und Zusammenhang ist hier nicht möglich.

Fazit

Die gewählte Bibliothek StoryMapJS erfüllt den Zweck der Visualisierung der Rundgänge für dieses Projekt gut. Ein Nachteil liegt in der relativ eingeschränkten Reihenfolge der einzelnen Stationen; NutzerInnen können nur begrenzt eigene Wege durch das Material finden. Die Einarbeitungszeit ist gering, die Integration in die eigene Infrastruktur funktionierte problemlos.

Ein Desiderat ist jedoch die bessere Strukturierung und Standardisierung der Daten und Metadaten. Hier wäre eine einheitliche Beschreibung nicht nur der gesamten Ausstellung (wie durch DEMES ermöglicht, oder vielleicht begrenzt auch

mit Dublin Core oder TEI umsetzbar), sondern der einzelnen Stationen und deren Zusammenhang sinnvoll. Zusätzlich könnten durch Anreicherung Anknüpfungspunkte für die Vernetzung der erzeugten Daten entstehen, beispielsweise GeoNames-Referenzen, wenn moderne Karten als Hintergrund verwendet werden. Die vollständige Trennung von Darstellung und Inhalt ist auch für die Weiternutzung und Archivierung zentral.

Insgesamt wird die wissenschaftliche Auseinandersetzung mit Virtuellen/Digitalen Ausstellungen/Rundgängen oft vernachlässigt, weil sie in erster Linie als Verschönerung bzw. nachrangiges Angebot von Suchportalen verstanden werden. Ganz im Gegenteil sind diese Einstiegspunkte aber für das Publikum oft leichter verständlich und können einen besseren Überblick über die zu vermittelnde Thematik geben als hochstrukturierte Interfaces und rein textuelle Beschreibungen.

Fußnoten

1. http://www.europeana.eu/portal/en/exhibitions/foyer (2017-08-09)
2. https://www.deutsche-digitale-bibliothek.de/content/ausstellungen (2017-08-09)
3. https://www.bavarikon.de/topics (2017-08-09)
4. http://kulturerbe.niedersachsen.de/viewer/kontexte/ (2017-08-09)
5. https://www.google.com/culturalinstitute (2017-08-17)
6. https://bavarikon.de/object/bav:BSB-CMS-0000000000001151 (2017-08-09)
7. http://wk1.staatsarchiv.at/ (2017-08-09)
8. Siehe bspw. http://oedb.org/ilibrarian/5-free-and-open-source-tools-for-creating-digital-exhibitions/ (2017-08-09)

Bibliographie

AthenaPlus (2015): Digital storytelling and cultural heritage: stakes and opportunities. http://www.athenaplus.eu/index.php?en/207/digital-storytelling [letzter Zugriff 22. August 2017]

AthenaPlus (2016): Metadata for the description of digital exhibitions: the DEMES element set. Version 1.0. http://www.athenaplus.eu/index.php?en/206/demes [letzter Zugriff 22. August 2017]

AthenaPlus: Movio. http://www.athenaplus.eu/index.php?en/211/movio)

Geisteswissenschaftliches Asset Management System – GAMS. http://gams.uni-graz.at [letzter Zugriff 22. August 2017]

INDICATE (2012): Handbook on virtual exhibitions and virtual performances. Version 1.0. http://www.indicate-project.org/getFile.php?id=412 [letzter Zugriff 22. August 2017]

Kultur- und Wissenschaftserbe Steiermark. http://www.kulturerbe-stmk.at [letzter Zugriff 22. August 2017]

Repositorium Steirisches Wissenschaftserbe. https://wissenschaftserbe.uni-graz.at/ [letzter Zugriff 22. August 2017]

StoryMapJS. https://storymap.knightlab.com/ [letzter Zugriff 22. August 2017]

Vom geschützt zugänglichen Datenbankverbund zur offenen Editions- und Forschungsplattform: kritischer Rückblick auf halber Strecke

Forney, Christian
christian.forney@hist.unibe.ch
Universität Bern, Historisches Institut

Rojas Castro, Antonio
arojasca@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Dängeli, Peter
p.daengeli@uni-koeln.de
Universität Bern, Historisches Institut; Universität zu Köln, Cologne Center for eHumanities

Ausgangslage

Die Vielschichtigkeit der Aufklärungszeit, die Funktionsweise der Gelehrtenrepublik und ihre Wechselwirkungen mit Wirtschaft, Politik und Gesellschaft, aber auch die Ausdifferenzierung der Naturgeschichte werden an der Universität Bern seit den frühen 1990er-Jahren intensiv beforscht. [1] Zentrale Figur des Forschungsinteresses ist dabei Albrecht von Haller (1708-1777), Schweizer Universalgelehrter und Schöpfer eines viel-

fältigen Oeuvres, das unter anderem in einem außerordentlich reichhaltigen handschriftlichen Nachlass überliefert ist.

Wichtigstes Hilfsmittel ist seit 1991 eine Forschungsdatenbank, die schrittweise durch den Zusammenschluss mit mehreren Forschungsprojekten (etwa zur Oekonomischen Gesellschaft Bern) zur Verbunddatenbank erweitert und bis heute laufend ausgebaut wurde. [2] Sie enthält detaillierte und untereinander stark verlinkte Daten zu rund 40 '000 Publikationen (Hallers Bibliothek, Forschungsliteratur), 22 '000 Personen, 20 '000 Briefen, 3 '000 Pflanzenarten, 2 '500 Orten und 800 Institutionen, jeweils in projektspezifisch gewachsenen Datenstrukturen abgelegt.

Um diese wertvolle Forschungsquelle umfassender verfügbar und sie zugleich als Grundlage für die anstehende Edition von Primärquellen nutzbar zu machen, wird die Datenbank gegenwärtig in einer Forschungskooperation zwischen dem Historischen Institut der Universität Bern und dem Cologne Center for eHumanities (CCeH) bereinigt, umstrukturiert und in eine zugleich entstehende Forschungs- und Editionsplattform integriert. [3] Zielformat sind XML-Daten, die sich eng an die TEI-Guidelines anlehnen. [4]

Charakteristika der Datenbank und der entstehenden Editionen

Das Potenzial der Datenbank lässt sich gut anhand der Personenobjekte veranschaulichen, die für rund 24 '000 historische Persönlichkeiten (dazu zählen Hallers rund 1 '050 Korrespondenten und 156 Korrespondentinnen) verschiedene Aspekte sehr detailliert und mit Quellennachweisen versehen beschreiben. Je nach Objekttyp enthalten die Personendatensätze bis zu 385 Felder: Neben den gängigen Lebensdaten wurden zu diesen Personen alle Verwandtschaftsbeziehungen, Ausbildungsgang, sowie Arbeits-, Forschungs- und Reisetätigkeiten erhoben und wo möglich mit den entsprechenden Datenbankobjekten verlinkt. [5] Für die historische Forschung stellt dieses dichte und qualitativ gesicherte Informationsnetz eine herausragende Forschungsressource dar.

Die Datenbank wird im Rahmen der Forschungskooperation ergänzt durch eine prototypische Korrespondenzedition des Briefverkehrs zwischen dem von Hannover aus wirkenden Universitätskurator Gerlach von Münchhausen und Albrecht von Haller. [6] Die Edition, die sowohl innerhalb der Forschungs- und Editionsplattform als auch mit externen Ressourcen wie correspSearch.net verlinkt wird, dient als Modell für die *Open Access*-Edition von zunächst 8 '000 Briefen und 9 '000 Rezensionen Hallers, die ab 2018 im Rahmen des SNF-Projekts *Online-Edition der Rezensionen und Briefe Albrecht von Hallers: Expertise und Kommunikation in der entstehenden Scientific community* ediert werden.

Hinsichtlich digitaler Forschungspraktiken sind die laufenden Modellierungs-, Konversions- und Entwicklungsarbeiten in zweierlei Hinsicht von Interesse: einerseits werfen sie die Frage auf, inwiefern Datenstrukturen durch die zugrunde liegenden Werkzeuge beeinflusst werden und bieten zugleich einen Rahmen konkreter Reflexion, andererseits erfordert die Konzeption eines übergreifenden Portals für unterschiedliche, aber zusammenhängende Inhalte eine geeignete Präsentation und Kommunikation.

Datenmodellierung und -transformation als Reflexionsprozess

Die *ad hoc* gewachsenen Datenstrukturen werden in ein neu definiertes Textformat überführt, das den TEI-Guidelines inhärente bestehende semantische Konzepte nutzt und sich mithin (zu einem gewissen Grad) selbst beschreibt. Angestrebt wird Verständlich- und Nachvollziehbarkeit nicht nur in der Präsentation, sondern auch auf der Datenebene. Da die TEI-Guidelines für viele vorliegende Forschungsdaten (z.B. Eigenschaften von Personen) keine entsprechenden Elemente kennen, wurden generische Lösungen durch Abstraktion gesucht. Damit bedingt die Überführung aus einer Feld-, Tabellen- und Maskenbasierten Datenbank in ein TEI-basiertes Format eine analytische Durchdringung der Datenstruktur, die auch den mit den Daten vertrauten Forschenden einen neuen Blick auf ebendiese gibt. Aus dieser Perspektive betrachtet erscheinen die Forschungsdaten deutlicher als Objekte der realen Welt mit semantisch klar definierten Eigenschaften. Die Reflexion verstärkt auf diese Weise die gegenseitige Wechselwirkung zwischen Datenmodellierung und wissenschaftlicher Analyse. Die Generalisierung spezifischer Datenbeschreibungsformen ist wesentlicher Bestandteil, um die neue Datenbank nicht nur im bestehenden, um Albrecht von Haller und die Oekonomische Gesellschaft Bern zentrierenden Kontext nutzen zu können, sondern z.B. auch für die Naturforschende Gesellschaft Zürich. Sie ermöglicht

inskünftig auch Anschluss an Schnittstellen Dritter wie *correspSearch* [7] im Bereich der Korrespondenzdaten oder die im Entstehen begriffene *prosopogrAPhi* [8] im Bereich der Personendaten.

Spezifizierung eines übergreifenden Editionsportals

Das entstehende Editionsportal will eine einheitliche Grundlage schaffen für verschiedene Inhalte, seien es beschreibende und Metadaten, digitale Neu-Editionen bestehender gedruckter Briefeditionen oder neu erarbeitete digitale Brief- und Rezensionseditionen mit unterschiedlichen thematischen, geographischen und personellen Schwerpunkten. Dabei sollen bestimmte Inhalte der Plattform den Nutzern in verschiedenen Kontexten (z.B. ein Brief als atomares Datum, als Bestandteil einer “Sammlung”, als edierter Brief) jeweils so präsentiert werden, dass die Situierung innerhalb des Portals jederzeit nachvollziehbar bleibt. Der Anspruch des Portals geht dadurch konzeptionell beträchtlich über den Standardfall einer digitalen Edition (im Sinne tiefer editorischer Erschließung) eines einzelnen Textes oder verschiedener Texte eines Autors hinaus. Zugleich greift durch die Tiefe der Erschließung aber auch der Vergleich mit digitaler Datenbereitstellung, wie sie in Bibliotheken und Archiven oftmals betrieben wird, zu kurz. Eine einfache und verständliche visuelle Kommunikation und Umsetzung dieser Ansprüche erweist sich in der konzeptuellen Arbeit als Herausforderung.

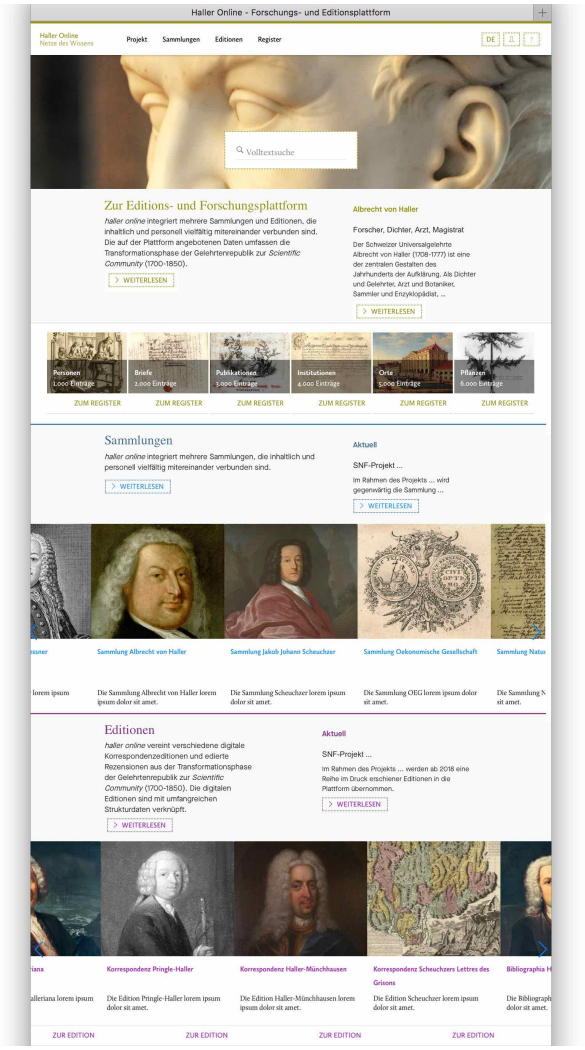


Abbildung 1: Segmentierte Landingpage der in Entwicklung begriffenen Webanwendung: Datenbank, Sammlungen, Editionen

Fußnoten

1. Wegweisend war das SNF-Projekt *Albrecht von Haller und die Gelehrtenrepublik der Aufklärung* (Laufzeit 1991-2003, Leitung: Urs Boschung, medizinhistorisches Institut der Universität Bern, in Kooperation mit der Burgerbibliothek Bern).
2. Realisiert wurde die Datenbank mit dem Dokumentations- und Retrievalsystem *FAUST Professional* der Firma Land Software: ursprünglich in der Version 1.0, zuletzt wurde das Programm in der Version 6.0 genutzt.
3. Der Umbau ist das Ziel des Projekts Haller Online (Laufzeit 2016-2019), das von der Burgergemeinde Bern mit der Albrecht von Haller-Stiftung sowie der Universität Bern finanziert wird. Über den Abschluss des Projekts hinaus ist die Zugänglichkeit und der Unterhalt der Plattform

institutionell durch die Haller-Stiftung und die Burgerbibliothek Bern abgesichert.
4. Die Datentransformation erfolgt in XProc-orchestrierten XSL-Transformationen und die künftige Datenpflege im oXygen-Autorenmodus; die Webanwendung vereinigt ein XML-Backend (XSLT 3.0, Solr) und ein Vue.js-Frontend, die Digitalisate sollen über eine IIIF-Schnittstelle eingebunden und verfügbar gemacht werden.
5. Verlinkungen können sowohl auf andere Personen wie auch auf andere Objekte (wie Orte, Pflanzen oder Institutionen) referenzieren.
6. Otto Sonntag (ed.): The Albrecht von Haller-Gerlach Adolph von Münchhausen Correspondence, Digitale Edition, Bern 2018 (in Vorbereitung).
7. Stefan Dumont et al.: correspSearch, URL: http://correspsearch.net (14.1.2018).
8. Georg Vogeler et al.: prosopogrAPhI, URL: https://github.com/GVogeler/prosopogrAPhI (14.1.2018).

Bibliographie

Boschung, Urs et al. (eds.): Repertorium zu Albrecht von Hallers Korrespondenz 1724-1777 (Studia Halleriana VII), Basel: Schwabe 2006.

Dumont, Stefan et al.: correspSearch, URL: http://correspsearch.net (14.1.2018).

Flückiger, Daniel / Stuber, Martin: Vom System zum Akteur. Personenorientierte Datenbanken für Archiv und Forschung, in: Kirchhofer, André et al. (eds.), Nachhaltige Geschichte. Festschrift für Christian Pfister, Zürich: Chronos 2009, S. 253-269.

Sonntag, Otto (ed.): The Albrecht von Haller-Gerlach Adolph von Münchhausen Correspondence, Digitale Edition, Bern 2018 (in Vorbereitung).

Steinke, Hubert: Archive databases as advanced research tools: the Haller Project, in: Monti, Maria Teresa (ed.), Antonio Vallisneri. L'edizione del testo scientifico d'età moderna, Firenze: Olschki 2003, S. 191–204.

Steinke, Hubert / Profos, Claudia: Bibliographia Halleriana: Verzeichnis der Schriften von und über Albrecht von Haller (Studia Halleriana VIII), Basel: Schwabe 2004.

Stuber, Martin: Findmittel und Forschungsinstrument zugleich. Die Datenbank des Berner Haller-Projekts, in: Arbido 14, 1999, S. 5-10.

Stuber, Martin / Hächler, Stefan / Lienhard, Luc (eds.): Hallers Netz. Ein europäischer Gelehrtenbriefwechsel zur Zeit der Aufklärung, Schwabe: Basel 2005.

Vogeler, Georg et al.: prosopogrAPhI, URL: https://github.com/GVogeler/prosopogrAPhI (14.1.2018).

Von Drupal 8 zur virtuellen Forschungsumgebung - Der WissKI-Ansatz

Fichtner, Mark
m.fichtner@wiss-ki.eu
Germanisches Nationalmuseum, Deutschland

Im Rahmen des von der DFG finanzierten Projekts “WissKI” entstand in zwei Projektphasen eine digitale Forschungsumgebung für die Anwendung im Bereich der Digital Humanities. Mit dem Ende der zweiten Projektphase 2017 wurde die Forschungsumgebung grundlegend aktualisiert und setzt nun auf das Open Source Content Management System Drupal 8 auf. Damit ging eine Aktualisierung der gesamten zugrundeliegenden Frameworks und Technologien (php 7, SPARQL 1.1) einher. Die aktuelle Version der Forschungsumgebung steht nun der wissenschaftlichen Öffentlichkeit als Open Source zur freien Verfügung.

Auch die aktuelle Fassung der Software setzt auf die bewährten Kernaspekte: Die Datenerfassung und -haltung in WissKI wird zentral bestimmt durch die semantischen Zusammenhänge zwischen einzelnen Fakten und Datensätzen. Dies wird durch umfassende Unterstützung aktueller Semantic Web Technologien erreicht. Die Einordnung und Speicherung der erhobenen Daten erfolgt auf Grundlage einer Domänenontologie, deren Konzepte und Relationen - zu sogenannten Pfaden verbunden - als Vorlage für die Masken und Felder im System dienen. Auf Basis dieser Technologie werden solitär erscheinende Daten zu einem gemeinsamen, semantischen Netzwerk verbunden und damit die unmittelbare Sichtbarkeit weiterer, tiefergehender Zusammenhänge ermöglicht. Hierdurch werden intuitiv Zusammenhänge in den Daten sichtbar, die sich für den Nutzer als Mehrwert anbieten. Das Webbasierte Systemdesign und der dadurch ermöglichte Zugriff über das Internet, die Anbindung von externen kuratierten Datenquellen (sog. Authority Files) und die Möglichkeit zur Bereitstellung ausgewählter Daten über gängige Online-Schnittstellen (Web-Frontend, SPARQL-Endpoint, ...) betonen den Semantic-Web-Gedanken hinter der

Infrastruktur. Die Speicherung der Daten erfolgt in einem TripleStore, der die eingegebenen Fakten in einer Subjekt-Prädikat-Objekt-Satzform ablegt. Die Aneinanderreihung der hier verwendeten Prädikate zu Pfaden erfolgt im Kern des Systems, dem sogenannten Pathbuilder, mit dem die semantische Bedeutung der einzelnen Einträge in Bezug auf das beschriebene Objekt (auch Person, Ort o. Ä.) anhand der Ontologie festgelegt wird. Die Eingabe der Daten erfolgt über eine, mit den gängigen Datenbankoberflächen vergleichbare, Editier-Oberfläche. Sie ist aus Feldern aufgebaut, die wiederum je einem bestimmten Feldtyp zugeordnet sind. Feldtypen bestimmen die Ein- und Ausgabemodalitäten der Daten.

Dabei verzichtet die Software nicht auf die aus dem Bereich der Content Management Systeme bekannten Funktionalitäten wie z. B. die Generierung von Websites, Foren, Wikis oder auch die detaillierte Verwaltung der Nutzer und ihrer Zugriffsrechte. Inzwischen ist die Software in verschiedenen Forschungsprojekten an unterschiedlichen, namhaften Institutionen im kunst- und kulturhistorischen, sowie biologischen und technischen Bereich erfolgreich im Einsatz. Als Domänenontologie im Museums- und Sammlungsbetrieb kommen individuelle Erweiterungen des “Conceptual Reference Model” des Comité international pour la documentation zum Einsatz (CIDOC-CRM: ISO 21127), dessen Umsetzung in der Web-Ontology-Language OWL ebenfalls vom Projekt besorgt wurde und über die Website http://erlangen-crm.org frei zur Verfügung steht.

Das Poster stellt den aktuellen Stand der WissKI-Software nach Vollendung der beiden Projektphasen dar. Neben dem bewährten Modell der Anpassung der Software durch die beiden am DFG-Projekt beteiligten Museen und der Friedrich-Alexander-Universität Erlangen-Nürnberg unterstützt die Interessengemeinschaft für semantische Datenverarbeitung e.V. (http://www.igsd-ev.de/) die gemeinnützigen Aspekte der Software weiter. Darüber hinaus werden bewusst auch Dritte zum Einsatz und zur Anpassung von WissKI eingeladen. Daraus resultierte im vergangenen Jahr der zahlreiche Einsatz der Software in Forschungsprojekten z.B. in Kooperation mit der Landesstelle der Nichtstaatlichen Museen in Bayern oder dem Zentralinstitut für Kunstgeschichte. Das System stellte v.a. durch die Nutzung aller Drupal-Basis-Funktionalitäten wie z.B. „Views“ seine Stärken unter Beweis. So können neben den altbewährten Textfeldern und -bereichen und Bildern (incl. Zoomviewer für sehr hochauflösende Bilder) auch interaktive Landkarten, 3D-Animationen, Zeitstrahlen und alle denkbaren Medientypen, sowohl zur direkten Ansicht als auch zum Download als Funktionalität genutzt werden. Zusätzlich zu diesen gängigen Formaten ermöglicht die Standardkonformität von WissKI-D8 auch die Einbindung anderer, gängiger Feldtypmodule, die für Drupal 8 zur Verfügung stehen. Zu den erwähnten Erleichterungen zählt ebenso ein Update des System-Kerns, dem Pathbuilder, mit dem die Pfadschablonen durch die Domänenontologie auf einer graphischen Oberfläche ausgewählt bzw. erzeugt werden können. Daneben wird eine umfassende Bibliothek mit Musterontologien, -masken und -pfaden bereitgestellt, die die Einstiegshürde für Erstbenutzer minimal zu halten.

Index der Autorinnen und Autoren