DHd2015

Von Daten zu Erkenntnissen

23. bis 27. Februar 2015, Graz

Book of Abstracts

UNI GRAZ
DHd
ZIM
CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

DHd2015

Von Daten zu Erkenntnissen

23. bis 27. Februar 2015, Graz

Workshops

CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

DHd-Tagung 2015

»Von Daten zu Erkenntnissen:
Digitale Geisteswissenschaften als Mittler
zwischen Information und Interpretation«

Workshop Proposal

Frosch oder Prinz?
TUSTEP als Werkzeug der Digitalen Geisteswissenschaften – ein Workshop der International TUSTEP User Group

Thomas Kollatz · Ute Recker-Hamm · Matthias Schneider

Frosch oder Prinz?
TUSTEP als Werkzeug der Digitalen Geisteswissenschaften – ein Workshop der International TUSTEP User Group

Das Tübinger System von Textverarbeitungsprogrammen (TUSTEP)[1] ist eines der ältesten, seit über 40 Jahren bis heute beständig weiterentwickelten und in zahlreichen geisteswissenschaftlichen Projekten[2] erfolgreich eingesetzten Programme der Digitalen Geisteswissenschaften. Es bietet einen Werkzeugkasten von speziell auf philologisch-sprachwissenschaftliche Anforderungen zugeschnittenen, auf einander abgestimmten Werkzeugen, die in hohem Maß an projektspezifische Erfordernisse angepasst werden können. Die Bandbreite der Programmmodule reicht dabei vom Vergleichen und Kollationieren, dem Zerlegen, Annotieren und Indexieren von Texten in XML oder anderen Formaten über das Skripting bis hin zum professionellen Satz und kann – beispielsweise für dynamische Webpublikationen – auch direkt auf Servern eingesetzt werden. Seit Kurzem verfügt TUSTEP über eine XML-basierte, alternative Kommandosyntax namens TXSTEP, die eine Steuerung des Programms von außerhalb, z. B. aus der XML-Entwicklungsumgebung Oxygen erlaubt.

Obwohl die Leistungsfähigkeit und hohe Qualität des Programmpakets national wie international unbestritten ist,[3] eilt ihm auf Grund der reduzierten Benutzeroberfläche und seiner vermeintlich besonderen Arbeitsweise der Ruf voraus, schwer erlernbar zu sein. So kommt beispielsweise das TEI-Wiki – nachdem die Vorzüge von TUSTEP gelobt worden sind – zum Schluss »learning TUSTEP is not a walk in the park […] a scholar with no or minor technical background will need support to get things work.«[4]

Diese Aufgabe übernimmt die International TUSTEP User Group (ITUG e. V.)[5], die seit über 20 Jahren TUSTEP-Anwender und -Lerner durch Kurse, das TUSTEP-Wiki[6] und

---

[1] ⟨tustep.uni-tuebingen.de⟩

[2] Für Projekte in Auswahl, die TUSTEP einsetzen, siehe: ⟨itug.de/projekte.html⟩

[3] Siehe z. B. John Bradley: Text Tools, in: A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004 – ⟨digitalhumanities.org/companion/⟩; Susan Hockey: The History of Humanities Computing, Ebenda; Michael Sperberg-McQueen: ⟨tustep.uni-tuebingen.de/sperberg2007.html⟩; John Unsworth: ⟨new.livestream.com/accounts/2263400/events/3512861⟩

[4] ⟨wiki.tei-c.org/index.php/Publishing_printed_critical_editions_from_TEI⟩

[5] ⟨itug.de⟩

[6] ⟨tustep.wikispaces.com/TUSTEP-Wiki⟩

Fachkolloquien unterstützt sowie die Weiterentwicklung von TUSTEP fördert, das seit mehreren Jahren als Open Source Software kostenlos verfügbar ist.

Der Workshop »Frosch oder Prinz?« der ITUG im Rahmen der DHd 2015 soll zeigen, dass es vielleicht nicht gerade ein Spaziergang im Park ist, TUSTEP zu erlernen, aber gewiß auch keine Hochgebirgstour. Dafür werden drei Programmmodule (Vergleich / Kollationierung, Skripting, Satz) exemplarisch heraus gegriffen und mit den Workshop-Teilnehmenden anhand eines durchgehenden Beispiels (»Froschkönig« der Brüder Grimm) praktisch erprobt. Der Workshop ist als Hands-on Workshop konzipiert.

Vergleichen

Texte miteinander zu vergleichen, kann in verschiedenen Situationen nützlich sein: Neben dem üblichen Vergleichen und Kollationieren von Textvarianten spielen auch Qualitätssicherungsaspekte eine Rolle, wie z. B. das Feststellen von unerlaubten oder versehentlichen Eingriffen in Texte oder der Vergleich von mehrfach erfassten Eingabedaten zur Herstellung eines korrigierten Texts.

Für den Vergleich und die Weiterverarbeitung der Vergleichsergebnisse stellt TUSTEP mehrere Programmbausteine zur Verfügung. Das Vergleichsmodul erlaubt die Ausführung in verschiedenen Modi, die Spezifikation von als gleichwertig zu betrachtenden Zeichenfolgen, die Definition von Aufsatzpunkten u. v. a. m. Das Ergebnis des Vergleichs wird in Form einer Protokoll- bzw. Korrekturdatei ausgegeben, die manuell oder programmgesteuert weiter verarbeitet und / oder in die Ausgangsdaten rücküberführt werden kann. Der Vergleich von mehr als zwei Textzeugen ist dadurch möglich, dass jede Textvariante separat mit dem Grundtext verglichen wird und die Vergleichsprotokolle anschließend automatisch kumuliert werden.[7]

Im Workshop werden die Arbeitsweise des Vergleichsmoduls und die Möglichkeiten der Weiterverarbeitung der daraus resultierenden Daten anhand von zwei Auflagen des »Froschkönig« erprobt.

---

[7] Dies Vorgehen impliziert keine vorweggenommene editorische Entscheidung für einen Archetyp, sondern stellt lediglich ein technisches Vorgehen dar, das in beliebiger Reihenfolge und beliebig oft in verschiedenen Abfolgen ausgeführt werden kann.

Skripting mit TUSCRIPT

Mit TUSCRIPT stellt TUSTEP eine vollwertige Skriptsprache bereit, die insbesondere zur philologischen Bearbeitung und Analyse von Textdaten geeignet ist.[8] Im Workshop werden den Teilnehmenden einige Skripting-Funktionen sowie exemplarische Dateizugriffe mit TUSCRIPT am »Froschkönig« näher gebracht. So wird in einem prototypischen Arbeitsablauf ein Import des Textes aus einem Büroformat (MS-Word), sowie dessen Aufbereitung in eine TEI–XML konforme und valide Datei mit den Teilnehmenden durchgeführt. Anschliessend soll gezeigt werden, wie Texte mit einfachen Mitteln (semi-)automatisch angereichert und ausgezeichnet werden können. Die Skriptsprache eignet sich selbstverständlich auch zur systematischen Textanalyse und Recherche. Im Workshop werden zudem komplexe Suchfunktionen, Indexerstellung und dergleichen mehr vorgestellt.

Satz

TUSTEP stellt zwei Programmbausteine für die Aufbereitung von Textdaten mit dem Ziel der Ausgabe als Postscript- respektive PDF-Datei bereit. Mit dem Modul #SATZ können einfache Texte bis hin zu anspruchsvollsten Editionen inklusive zeilensynoptischer Ausgabe sowie mehreren Apparaten und Registern hergestellt werden. Ein sogenannter »Roundtrip«, also das Zurückfließenlassen von Satzdaten (z. B. Seiten / Zeileninformationen) in die (XML-)Ausgangsdaten (z. B. zur nachträglichen Registererstellung) ist möglich. Das Modul *SATZ hingegen stellt einen einsteigerfreundlich komprimierten Ausschnitt dieses Funktionsumfangs zur Verfügung, welcher für die Produktion von Hausarbeiten, Dissertationen und Monographien oder Sammelbänden ausreichend ist. Hierfür stehen ein XML-Tagset, eine spezielle GUI, verschiedene Dokument- und Formatvorlagen sowie Schriftdefinitionen zur Verfügung. Die Lernkurve ist entsprechend flach. Das Modul *SATZ überzeugt nicht nur durch Benutzerfreundlichkeit, sondern darüber hinaus auch durch die vielfältigen Verwendungsmöglichkeiten out of the box (u. a. klassische Druckvorstufe für die Druckerei, on the fly-PDF-Erstellung aus Online-Datenbanken sowie die plattform- und medienunabhängige PDF-Erstellung).

---

[8] Für Beispiele s. ⟨rosettacode.org/wiki/TUSCRIPT⟩

Im Rahmen des Workshops soll das Werkzeug *SATZ benutzt werden, um die Ausgabe der zuvor mit den Vergleichs- und TUSCRIPT-Programmbausteinen aufbereiteten Textdaten als PS-/PDF-Datei zu exemplifizieren. Hierbei wird großer Wert auf die schnelle und komfortable Erschließung des Werkzeugs sowie die sehr guten Ergebnisse der Druckvorstufe gelegt.

Die Teilnehmenden erhalten die Möglichkeit, den in den vorhergehenden Abschnitten verarbeiteten Beispieltext selbst in unterschiedlichen Formaten zu erzeugen sowie mit der Druckausgabe zu experimentieren (Spaltensatz, Schriftarten, Bearbeitung des Textes in der *SATZ-GUI).

TUSTEP-Café

Während der Pause sowie unmittelbar vor und nach dem Workshop erhalten Projekte, die TUSTEP anwenden, Gelegenheit ihre Arbeit in informeller Atmosphäre durch Poster vorzustellen. Einige Zusagen liegen bereits vor. Falls der Workshop wie geplant durchgeführt werden kann, wird zudem ein »call for posters« über die ITUG Mailing Liste ausgerufen.

Wissenschaftsblogs und wissenschaftliches Bloggen bei de.hypotheses.org

Was bedeutet der digitale Wandel für die Wissenschaft? Neue Herausforderungen. Aber auch neue Möglichkeiten. Nicht nur können Publikationen online zur Verfügung gestellt werden, auch ändert sich zunehmend die wissenschaftliche Umgebung, mit all den Facetten, die das Web 2.0 zu bieten hat. Universitäten, Wissenschaftliche Institutionen sowie Wissenschaftlerinnen und Wissenschaftler sind auf Twitter, Facebook und Co. präsent. Und dann sind da noch all die Plattformen, die explizit der Vernetzung von Forschenden dienen. Dass jedoch das Internet nicht mit Qualitätsverlust gleichzusetzen ist, zeigen die regen wissenschaftlichen Austausche und Diskussionen im Internet: wissenschaftliche Kommunikationspraktiken in Ergänzung zu den wissenschaftlichen Fachzeitschriften und Tagungsbesuchen.

Mit Wissenschaftsblogs entwickelt sich rasant ein neues Genre, das bislang nicht im Methoden-Kanon und den überkommenen Reputationsmechanismen geistes- und sozialwissenschaftlicher Disziplinen vorgesehen war. Was genau bedeutet Bloggen für das akademische Schreiben und Publizieren? Wie verändert diese Kommunikationsform den wissenschaftlichen Alltag? Als neue Form der fachwissenschaftlichen Kommunikation nutzen Blogs die Möglichkeiten des Internets und des Web 2.0 für eine direkte und interaktive Publikation. Angesprochen wird neben der akademischen Community immer auch die breite Öffentlichkeit, denn jedes Blog ist ein Fenster zum Elfenbeinturm Wissenschaft. Als öffentlich geführte wissenschaftliche Notizbücher eignen sich Blogs zur selbstkritischen Reflektion des eigenen Forschungsprozesses wie auch zur Dokumentation desselben. Nicht nur Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern bietet Bloggen die Möglichkeit, bereits in einem frühen Stadium auf ihr Projekt aufmerksam zu machen, mit erfahrenen Wissenschaftlerinnen und Wissenschaftlern in Austausch zu treten und sich zu vernetzen. Denn Wissenschaftsblogs haben ein hohes Potential für die schnelle Verbreitung und Diskussion aktueller Forschungsinhalte. Mit de.hypotheses.org wurde Anfang 2012 eine Plattform für geistes- und sozialwissenschaftliche Blogs geschaffen, in deren Umfeld seither eine stetig wachsende deutschsprachige Community als Teil eines europäischen Netzwerks entstanden ist.

Im Rahmen des Workshops soll zum einen die theoretische Seite des wissenschaftlichen Bloggens angesprochen, zum anderen ein praktischer Teil angeboten werden. Zunächst sollen unter anderem verschiedene Arten des Bloggens, der besondere Schreibstil und die Interaktion mit der Leserschaft thematisiert werden. Hier steht vor allem die Frage nach erfolgreichem wissenschaftlichen Bloggen im Vordergrund. Im Anschluss daran werden Schritt für Schritt die einzelnen Aspekte der Blogpraxis vorgestellt und vorgeführt. Die Teilnehmerinnen und Teilnehmer erhalten eigene Schulungsblogs auf der Plattform de.hypotheses.org (Wordpress) und üben die einzelnen Schritte, vom Anlegen eines Artikels über die Formulierung einer guten Überschrift bis hin zum Einbetten von Videos. Während des Workshops werden außerdem Tipps für die Anfangsphase eines wissenschaftlichen Blogs gegeben sowie rechtliche Belange erörtert.

Der Workshop richtet sich vor allem an Teilnehmerinnen und Teilnehmer, die keine oder wenige Vorkenntnisse im wissenschaftlichen Bloggen haben. Ein eigenes Blog ist nicht Voraussetzung zur Teilnahme.

Proposal für einen Pre-Conference Workshop zur
2. Jahrestagung *Digital Humanities im deutschsprachigen Raum (DHd)*

Computerlinguistische Methoden der Inhaltsanalyse in den Sozialwissenschaften: Forschungspraktische Herausforderungen, Werkzeuge und Technologien

Organisiert von den Kooperationspartnern des BMBF-Verbundprojekts *e-Identity*

Überblick

Zeit (Vorschlag):	Montag, 23. Februar 2015, 14:00 bis 19:00 Uhr Dienstag, 24. Februar 2015, 09:00 bis 13:00 Uhr
Organisatoren:	Prof. Dr. Manfred Stede, Jonathan Sonntag (beide Universität Potsdam), Prof. Dr. Cathleen Kantner, Maximilian Overbeck (beide IfS Stuttgart), Prof. Dr. Jonas Kuhn, Dr. André Blessing (beide IMS Stuttgart), Prof. Dr. Ulrich Heid, Fritz Kliche (beide Universität Hildesheim)
Teilnehmerzahl:	ca. 15 – 20
Technische Ausstattung:	Beamer (VGA)
Kontakt:	Projekt-Webpage: http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html

Inhalt des Workshops

Aufgrund der dramatisch angestiegenen Verfügbarkeit großer Korpora sozialwissenschaftlich relevanter Textdaten erlebt die Forschungslandschaft der Sozialwissenschaften aktuell einen regelrechten Boom der Methoden für die computerlinguistische Inhaltsanalyse. Dabei werden die Akzente mal stärker auf quantitative Auswertungen von Texten, mal stärker auf qualitative Interpretation und Annotation von Textdaten gesetzt. Unser Workshop möchte die Perspektiven beider Seiten zusammenbringen und dabei insbesondere auch die Möglichkeiten der sinnvollen Ergänzung quantitativer und qualitativer Analyse in den Blick nehmen.

Die Organisatoren des Workshops sind in dem interdisziplinären Forschungsprojekt *e-Identity* vernetzt und untersuchen aus politikwissenschaftlicher und computerlinguistischer Perspektive die internationale Diskussion über Kriege und humanitäre Interventionen seit dem Ende des Kalten Krieges. Dabei stehen folgende Fragestellungen im Vordergrund: Wie mobilisieren internationale Akteure in Krisensituationen kollektive Identitäten? Spielen sie ethnische, religiöse, nationale, europäische, u.a. Bindungen gegeneinander aus? Welche Ursachen und Effekte hat diese Identitätspolitik? Das Projekt untersucht internationale Diskussionen über Kriege und humanitäre Interventionen seit dem Ende des Kalten Krieges. Das Forscherteam greift auf ein bereinigtes mehrsprachiges Korpus von mehreren hunderttausend Zeitungsartikeln aus der Qualitätstagespresse mehrerer europäischer Länder (Österreich, Deutschland, Irland, Frankreich, Vereinigtes Königreich) und den USA zurück (kontinuierlich erhobener Untersuchungszeitraum: Januar 1990 - Dezember 2011).

Um die Analyse dieser komplexen, theoretischen Konzepte auf großen Textkorpora von Zeitungsartikeln zu bewältigen, verwendet das *e-Identity*-Projekt diverse sprachtechnologische Werkzeuge. Das *e-Identity*-Projekt befindet sich aktuell in seinem letzten Projektjahr und hat bereits Tools und Verfahren entwickelt, die nun im Rahmen des Workshops der breiteren Forschungslandschaft der Digital Humanities im deutschsprachigen Raum präsentiert werden sollen. Neben der Präsentation unserer Forschungsergebnisse soll ein weiterer Schwerpunkt auf der Präsentation externer Forschungsprojekte liegen, die aktuell an der Schnittstelle von Computerlinguistik und Sozialwissenschaften durchgeführt werden.

Format der Workshops

Der zweitägige Workshop soll im Vorfeld zur Digital Humanities Jahrestagung in Graz an den Tagen 23. und 24. Februar 2015 stattfinden. Als Format für den Workshop schlagen wir zwei Phasen vor:

1) In einer ersten Phase erhalten die Workshop-Teilnehmer die Möglichkeit, über ihren Einsatz von Software-Werkzeugen oder anderen quantitativen (z.B. korpuslinguistischen) Methoden der Inhaltsanalyse großer Textmengen zu berichten. Sie sollen dabei konkret von ihrer Forschungspraxis im Rahmen ihrer sozialwissenschaftlichen Forschungsprojekte berichten. Ziel der Vorträge ist es, möglichst viele Einblicke in methodische und technische Einzelheiten der empirischen Analysen zu gewinnen, was auch die Demonstration von Software-Werkzeugen einschließt. Innerhalb dieses Blocks sollen auch die im *e-Identity*

Verbund entstandenen Werkzeuge präsentiert werden: Eine Explorations-Werkbank für die Konstruktion und manuelle Annotation von Korpora aus heterogenen Textquellen, und der *Complex Concept Builder* – eine mehrschichtige Analyse-Pipeline für die automatische Annotation der Texte mit Linguistik-naher Information. Insbesondere von Relevanz sind hier die Endprodukte der Pipeline, die darauf ausgelegt sind, von Sozialwissenschaftlern verwendet zu werden.

Das Format der ersten Workshop-Phase besteht aus jeweils 15 bis max. 20-minütigen Vorträgen (inklusive Demos) und anschließender 10-minütiger Diskussion. Insgesamt sollen ca. 5 –7 externe Forschungsgruppen die Möglichkeit erhalten, ihre Forschungsprojekte vorzustellen. Die Beitragenden werden von den Workshop-Organisatoren unmittelbar angesprochen – es wird also kein offizieller Call for Papers veröffentlicht. Nichtsdestotrotz wird der Workshop für weitere Forscherinnen und Forscher der e-Humanities geöffnet und über unterschiedlichste Kanäle, wie die Newsletter der FAG 8 der Clarin-D-Community publik gemacht.

2) Die Vorträge dienen dann in der zweiten Phase als Grundlage für eine breitere Reflexion und vergleichende Analyse der vorgestellten Werkzeuge und Resultate. Die TeilnehmerInnen werden aufgefordert, möglichst selbstkritisch und transparent über Schwierigkeiten und Herausforderungen ihrer methodischen Ansätze zu berichten. Dabei stehen u.a. der Vergleich der gesetzten Forschungsziele und die erreichte Funktionalität sowie eine Diskussion hinsichtlich der Übertragbarkeit auf unterschiedliche Anwendungsszenarien im Vordergrund. Ziel des zweiten Themenblocks besteht darüber hinaus in einer gemeinsamen Bestandsaufnahme von Erfahrungswerten der konkreten Zusammenarbeit von Sozialwissenschaften und Informatik-nahen Disziplinen:

- Was wurde bisher erreicht – was „funktioniert“ nunmehr?
- Was bedeutet es wenn etwas „funktioniert“? Während Informatik-nahe Disziplinen sich beispielsweise über eine 80%ige Trefferrate auf natürlichsprachlichem Text durchaus freuen, gibt es in den Sozialwissenschaften meistens ein anderes Verständnis von „funktionieren“. Wie kann hier eine Brücke geschlagen werden?
- Welche ursprünglichen Ziele oder Pläne haben sich als noch nicht umsetzbar erwiesen?
- Welche neuen Pläne oder Ziele ergeben sich aus Anstößen der bisherigen Zusammenarbeit?

Das Format des zweiten Themenblocks unterscheidet sich methodisch vom ersten Themenblock. Hier sollen keine Präsentationen stattfinden, sondern vielmehr moderierte Diskussionen, sowie kürzere Gruppenarbeits-Phasen. Die Beitragenden werden von den Workshop-Organisatoren unmittelbar angesprochen, es wird also keinen öffentlichen Call for Papers geben. Das bedeutet freilich nicht, dass es sich um einen "geschlossenen" Workshop handeln soll; im Gegenteil sind weitere Teilnehmer sehr willkommen. Die Vorträge werden so konzipiert, dass sie als Grundlage für die Diskussion in Phase 2 dienen können.

Je nach Lage der Interessen und Zusammensetzung der Teilnehmergruppe sind auch kurze Phasen der Gruppenarbeit denkbar.

Für den Erfolg dieses Szenarios erscheint es uns wichtig, dass beide Phasen nicht unmittelbar aufeinander folgen, sondern die Teilnehmer nach Abschluss der Präsentationen eine „Bedenkzeit“ haben, bevor die Diskussionsphase einsetzt. Im Unterschied zum eigentlich vorgegebenen System halbtägiger Pre-Conference Workshops schlagen wir daher vor, unseren Workshop an zwei halben Tagen stattfinden zu lassen: Phase 1 am Montagnachmittag, Phase 2 am Dienstagvormittag.

Zielgruppe des Workshops

Dieser Workshop richtet sich an andere Forschungsgruppen, die sich bereits im fortgeschrittenem Stadium ihres Projektes befinden. Sowohl Teilnehmer aus den Informatik-nahen Bereichen als auch aus den Sozialwissenschaften sind angesprochen.

Workshop Computational Narratology

In den letzten fünf Jahren hat die computergestützte erzähltechnische Analyse von Einzeltexten wie von großen Textsammlungen solche Fortschritte gemacht, dass sich inzwischen eine ganze Reihe von Methoden und Werkzeugen etabliert haben. Durch eine Reihe von Einzeluntersuchungen [z.B. Brunner 2014, Mani 2013] ist dadurch ein neues Forschungsfeld sichtbar geworden, das man als *Computational Narratology* bezeichnen kann.[1] Dieser Workshop soll Interessierte an diesem innovativen Forschungsfeld, bei dem sich Erzählforschung, formale Modellierung und quantitative Textanalyse begegnen, zusammenbringen, um möglichst frühzeitig Felder zu identifizieren, in denen Kooperationen die unterschiedlichen Forschungsinteressen stärken und gegenseitig befruchten können.

Im Zentrum des Workshops stehen dabei folgende sechs Themenkomplexe:

I. Entwicklung von Standards und Datenstrukturen für gemeinsam nutzbare Textkorpora

Ein erstes Aufgabenfeld für eine Computational Narratology, das eine Grundlage für die koordinierte, verteilte Arbeit an einzelnen narratologischen Problemfeldern bildet, ist die Verabredung von Standards und Datenstrukturen für gemeinsam nutzbare Textkorpora. Die Aufgabenstellung umfasst dabei folgende Einzelaspekte:

- Entwicklung eines gemeinsamen Formats für Metadaten, die für narratologische Analysen relevante, die Texte beschreibende Informationen enthalten (einheitliche Feldnamen und kontrolliertes Vokabular, um den Anpassungsaufwand an neue Korpora zu minimieren; siehe auch unten: "Narratologische Basiskategorien").
- Ein gemeinsames Format für annotierte Trainingskorpora, sodass händisch bezüglich linguistischer und/oder narratologischer Phänomene annotierte Trainingskorpora eines Teams von anderen Teams für die Optimierung von automatischen Analyseverfahren nachgenutzt werden können.
- Ein gemeinsames Repository für solche Trainingskorpora, das die sichere Speicherung, eindeutige Identifizierung und Rechteverwaltung zulässt, sodass neue Ressourcen publiziert, ausgetauscht, gefunden und nachgenutzt werden können.
- Eine gemeinsame Strategien zum Einsatz von Crowdsourcing beim Erstellen von Ressourcen durch händische Annotationen sollte entwickelt werden, auch unter Berücksichtigung verschiedener institutioneller Akteure wie Forschungsgruppen oder Anbietern von Textarchiven.

II. Narratologische Basiskategorien für die Analyse von Textkorpora

[1] Wir fokussieren im Folgenden auf den dritten der in Mani 2013 angeführten Gegenstandsbereiche, nämlich die computergestützte narratologisch-literaturwissenschaftliche Textanalyse. Das umfassender Forschungsfeld umfasst nach Mani „ (…) approaches to storytelling in artificial intelligence systems and computer (and video) games, the automatic interpretation and generation of stories, and the exploration and testing of literary hypotheses through mining of narrative structure from corpora."

Als zweites Aufgabenfeld soll auf dem Workshop die Definition eines generischen, projektübergreifend relevanten Sets von narratologischen Beschreibungskategorien in Form einer sog. Tag-Library diskutiert werden. Dabei stellen sich zwei Kernfragen:

- Theoretische Fundierung: Bereits seit den 1990er Jahren wird eine intensive fachwissenschaftliche Debatte zwischen Vertretern der "klassischen" (formalistisch-strukturalistisch orientierten) Narratologie und Vertretern der sog. "new narratologies" (d.h. der kognitivistische, thematische und ideologiekritische Forschungsfragen verfolgenden) geführt. Vor diesem Hintergrund ist zu erörtern, ob es möglich und forschungsmethodisch überhaupt sinnvoll ist, diesen narratologischen Methodenpluralismus in Form einer gemeinsamen Tag-Library kategorial abbilden zu wollen. Im Kontext der Computational Narratology wäre prinzipiell auch die Alternative denkbar: d.h. die Modellierung mehrerer alternativer, dafür jedoch in sich homogener Klassifikationssysteme.

- Wie kann man ausgehend von den 'klassischen' Beschreibungskategorien insbesondere Gennette'scher Provenienz, die konzeptionell überwiegend auf eine 'dichte' narratologische Beschreibung von Einzeltexten etwa bis zur Satzebene hinauf abzielen, zu neuen synthetischen Kategorien fortschreiten, die auf einer deutlich höheren Komplexitäts- und Abstraktionsebene angesiedelt sind, wie sie für synchrone wie diachrone Korpusstudien erforderlich ist?

III. Automatisierung und Optimierung von Search&Retrieval-Operationen

Eine dritte Herausforderung an eine Computational Narratology stellt - im Anschluss an die formale Modellierung - das Retrieval von entsprechend beschreibbaren Textinstanzen dar. Hierzu sollen in dem Workshop zwei Operationalisierungsstrategien diskutiert werden, die auch komplementär eingesetzt werden können:

- die manuelle Auszeichnung aller Texte auf der Grundlage der formalen Modellierung durch idealerweise mindestens zwei Annotatoren;
- die manuelle Annotation eines definierten Teilkorpus, um auf dieser Basis entweder mit maschinellem Lernen oder regelbasiert den Rest der Textsammlung automatisch zu annotieren.

Beide Verfahren, die für ein Retrieval narratologischer Konzepte notwendig sind, werden aktuell bereits für Teilbereiche des narratologischen Kategoriensystems angewendet – dazu gehören beispielsweise Redewiedergabe (vgl. Brunner 2013) und Zeitgestaltung (vgl. Bögel et al. im Erscheinen). Dabei zeigt sich, dass einige „weiche" Konzepte, wie beispielsweise die Kategorie *erlebte Rede*, deutlich schlechtere Ergebnisse beim Retrieval ergeben als die klarer definierbaren Begriffe. Ein Desiderat für die Computational Narratology ist entsprechend, *best practices* für das Retrieval von Konzepten unterscheidlichen Explikationsgrades zu etablieren. Auch dabei sind unterschiedliche Strategien möglich: so kann z.B. die Übereinstimmungsquote, die menschliche Annotatoren in der Anwendung der fraglichen Konzepte maximal erreichen (*interannotator agreement*), als Ziel des Retrievals festgelegt werden. Denkbar ist aber auch ein Ansatz, der weniger auf Normierung und stärker auf die Exploration und Abbildung von Deutungsvielfalt setzt, wie dies z.B. im Ansatz des sog. kollaborativen taggings (vgl. Meister 2012) versucht wird.

IV. Gattungsklassifikation auf Grundlage narratologischer Merkmale

Auf der Grundlage von detaillierten Informationen über die Präsenz bzw. Verteilung von kategorial definierten narrativen Phänome in zahlreichen Einzeltexten ist heute erstmals eine narratologisch begründete, korpusbasierte und damit auch empirisch validierbare Textklassifikation möglich. Unter diesem Gesichtspunkt soll der Workshop zwei Teilfragen diskutieren:

- Inwieweit stehen die Mechanismen und Ergebnisse dieses Ansatzes in einem produktiven Spannungsverhältnis zu bestehenden (historischen oder fachwissenschaftlichen) Gattungskategorien?
- Wie ist es um die Anschlussfähigkeit der Computational Narratology an literaturwissenschaftliche Teilbereiche wie Literaturgeschichte oder Gattungstheorie bestellt?

V. Werkzeug-Entwicklung

Die Entwicklung und nachhaltige Bereitstellung von Werkzeugen und Plattformen für die Arbeit mit gemeinsam nutzbaren Textkorpora ist für die Computational Narratology von besonderer Relevanz. Aufgrund der Vielfältigkeit narratologischer Forschung stellen sich dabei besondere Herausforderungen an das Beschreibungsinventar, wie z. B. die Berücksichtigung dynamisch entwickelbarer Taxonomien, von Hierarchieüberlappungen und widersprüchlichen Annotationen. Analysewerkzeuge sollten immer auch Anpassungsmöglichkeiten für konkrete Forschungsfragen vorsehen. Vor dem Hintergrund dieser Problemstellung soll der Workshop zwei Fragen thematisieren:

- Textanalyse unter Einbeziehung von Textannotationen: Werkzeuge zur einfachen Analyse reiner Textdaten sind nicht zuletzt durch die Entwicklung im Rahmen von Suchmaschinen gut erforscht. Zur tiefergehenden Analyse von Textdaten ist allerdings die Einbeziehung von Kontextinformationen in Form von Annotationen nötig. Die Werkzeugentwicklung einer Computational Narratology sollte daher ein besonderes Augenmerk auf die Erstellung und Verarbeitung von Annotationen richten.
- Verknüpfung von Crowd-Sourcing und Machine Learning-Verfahren: Aufgrund der benötigten großen Menge von qualitativ hochwertigen manuellen Annotationen werden Crowd Sourcing fähige Werkzeuge benötigt. Angestrebt wird jedoch zugleich auch die automatische Generation von Annotationen. Gefordert sind daher Werkzeuge, die aus den manuell angefertigten Annotationen lernen und auf diesen aufbauen können.

VI. Tutorien zur Computational Narratology

Um die Nachnutzbarkeit der im Rahmen einer Computational Narratology gewonnenen Erkenntnisse und Ergebnisse zu gewährleisten und einer möglichst breiten Nutzergruppe zugänglich zu machen, sollen frei im Netz verfügbare Tutorien entstehen. Diese sollen die Verwendung neu entwickelter Methoden und Tools veranschaulichen, prototypische Arbeitsabläufe darlegen und die Einhaltung von best practices beschreiben. Dadurch soll im Bereich der computergestützten Textanalyse unerfahrenen Personen, weniger technisch versierten Narratologen oder auch Studenten der Digital Humanities oder Literaturwissenschaft der Einstieg in das Feld der Computational Narratology erleichtert werden.

Vor dem Hintergrund dieser Zielsetzung soll zum Abschluss des Workshops ein Gesamtkonzept für den Aufbau einer umfassenden Sammlung “Tutorien zur Computational Narratology” entworfen werden.

Der Workshop wird geleitet von:

Prof. Dr. Fotis Jannidis

Universität Würzburg, Institut für Deutsche Philologie, Philosophische Fakultät I, Am Hubland, D - 97074 Würzburg, fotis.jannidis@uni-wuerzburg.de

Prof. Dr. Jan Christoph Meister

Universität Hamburg, Institut für Germanistik, Fakultät für Sprache, Literatur und Medien, Von-Melle-Park 1, D - 20146 Hamburg, jan-c-meister@uni-hamburg.de

Dr. Christof Schöch

Universität Würzburg, Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie, Am Hubland, D - 97074 Würzburg, christof.schoech@uni-wuerzburg.de

Organisatorisches

Der Workshop ist für bis zu 15 Teilnehmer ausgelegt. Als technische Grundausstattung wird ein Beamer und ausreichend Stromanschlüsse für die Teilnehmer sowie Zugang zum Internet benötigt. Ein workshopspezifischer Call wird nicht veröffentlicht.

Die oben umrissenen sechs Themenkomplexe sollen in einer Vormittags- und einer Nachmittagssession bearbeitet und die Zwischenergebnisse live in einem begleitenden Workshop-Blog dokumentiert werden.

Literatur

Bögel, Thomas/Gertz, Michael/Gius, Evelyn/Jacke, Janina/Meister, Jan Christoph/Petris, Marco/Strötgen, Jannik: Collaborative Text Annotation Meets Machine Learning. heureCLÉA, a Digital Heuristics of Narrative, in: DHCommons Journal (im Erscheinen).

Brunner, Annelen: Automatic recognition of speech, thought, and writing representation in German narrative texts. In Literary & Linguistic Computing 28: 4 (2013), S. 563-575.

Mani, Inderjeet: "Computational Narratology". In: Hühn, Peter et al. (eds.): *the living handbook of narratology*. Hamburg: Hamburg University. URL = http://www.lhn.uni-hamburg.de/article/computational-narratology [view date:10 Nov 2014]

Meister, Jan Christoph: „Crowd sourcing “true meaning”. A collaborative markup approach to textual interpretation.“ In: Willard McCarty, Marylin Deegan (eds.), Collaborative Research in the Digital Humanities. Festschrift for Harold Short (Ashgate Publishers) 2012, 105-122.

ediarum – Eine digitale Arbeitsumgebung für Editionsvorhaben

Stefan Dumont (dumont@bbaw.de), Martin Fechner (fechner@bbaw.de)

An der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) sind zahlreiche geisteswissenschaftliche Forschungsvorhaben unterschiedlichster Fachrichtungen angesiedelt. Die TELOTA-Arbeitsgruppe (»The Electronic Life of the Academy«) unterstützt diese Vorhaben in allen digitalen Belangen und entwickelt Softwarelösungen für die tägliche Forschungsarbeit der Wissenschaftler/-innen.

Die Erfahrung hat gezeigt, dass die Bereitschaft, TEI-Kodierung in Editionsvorhaben zu verwenden, von der Benutzerfreundlichkeit der Eingabeoberfläche abhängt. Aus der Perspektive der Wissenschaftler/-innen erscheint es als ein Rückschritt, direkt im XML-Code zu arbeiten, wenn man vorher in Programmen wie MS Word gearbeitet hat. Eine neue Softwarelösung muss daher mindestens den gleichen Komfort bieten wie das zuvor benutzte Programm. Idealerweise würde sie sogar den gesamten Lebenszyklus einer Edition abdecken: von der ersten Phase der Transkription bis hin zur Publikation in Web und Druck.

TELOTA hat mit »ediarum« eine solche digitale Arbeitsumgebung entwickelt. Diese Lösung besteht aus mehreren Softwarekomponenten, die es den Wissenschaftler(inne)n erlauben, Transkriptionen von Manuskripten in TEI-XML anzufertigen, zu bearbeiten und zu veröffentlichen.

Als zentrale Softwarekomponente der neuen Arbeitsumgebung wird »oXygen XML Author« eingesetzt. Die Bearbeiter arbeiten in oXygen XML Author nicht in einer Codeansicht, sondern in der benutzerfreundlichen »Autorenansicht«, die über Cascading Stylesheets (CSS) gestaltet wird. Außerdem kann der Endanwender über eine eigene Werkzeugleiste per Knopfdruck Auszeichnungen vornehmen. So können z.B. in Manuskripten Streichungen markiert oder Sachanmerkungen eingegeben werden. Auch können Textstellen ausgezeichnet und gleichzeitig über eine komfortable Auswahlliste mit dem jeweiligen Eintrag eines zentralen Registers (Personen-, Ortsregister etc.) verknüpft werden. Der gesamte Text kann dadurch einfach und schnell mit TEI-konformen XML ausgezeichnet werden.

Die digitale Arbeitsumgebung nutzt die native XML-Datenbank »exist-db« als zentrales Repositorium für die XML-Dokumente. Die Datenbank ist auf einem Server installiert und online zugänglich. Dadurch können alle Projektmitarbeiter auf ein und denselben Datenbestand zugreifen und zusammenarbeiten.

Neben der eigentlichen Arbeitsumgebung in oXygen XML Author, wird für die Forschungsvorhaben auch jeweils eine Website auf Basis von eXist, XQuery und XSLT erstellt. In ihr kann von den Wissenschaftler(inne)n der aktuelle Datenbestand leicht durchblättert bzw. durchsucht werden. Die Website kann - je nach Bedarf - nur den Bearbeitern oder der gesamten Öffentlichkeit gemacht werden.

Als weitere Ausgabemöglichkeit wird mit Hilfe von ConTeXt eine Druckausgabe implementiert, die automatisch aus den aktuellen TEI-XML-Dokumenten ein PDF erstellt. Die Gestaltung und Formatierung kann - nach entsprechender Konfiguration - dabei gedruckten Bänden der jeweiligen Edition entsprechen. Jedem TEI-Element wird über eine Konfigurationsdatei eine entsprechende Formatierungsanweisung für den Druck übergeben. So können z.B. Text- und Sachapparat als Fußnoten dargestellt werden, die mit Hilfe von

Zeilennummerierung und Lemmata auf den Fließtext verweisen. Die Druckausgabe erstellt bei Bedarf auch das passende Register zu den jeweiligen Transkriptionen und löst Querverweise zwischen Texten auf.

Die Arbeitsumgebung wird seit 2012 von Wissenschaftler(inne)n verschiedener Forschungsvorhaben bei ihrer täglichen Arbeit benutzt. Nach ihrer Meinung befragt, waren sich die Nutzer darin einig, dass durch die neue Arbeitsumgebung die Editionsarbeit erleichtert und viel Zeit gespart wird. Auch die Möglichkeit, die Ergebnisse der Arbeit direkt in einer Webpräsentation oder Druckausgabe zu kontrollieren, wurde positiv gesehen. Sehr erleichtert äußerten sich die Mitarbeiter/-innen darüber, dass ihnen keine Arbeit im XML-Code selbst zugemutet wird, sondern alle Texte in einer grafischen und einfach zu bedienenden Programmoberfläche mit XML ausgezeichnet werden können.

Nach der erfolgreichen Pilotumsetzung im Akademievorhaben »Friedrich Schleiermacher in Berlin 1808-1834. Briefe, Vorlesungen, Tageskalender« wurde »ediarum« in zwei weiteren Akademienvorhaben eingesetzt: »Commentaria in Aristotelem Graeca et Byzantina« und »Regesta Imperii - Friedrich III.« (letzteres in Kooperation mit der AdW Mainz) Für jedes Projekt wurden die TEI-XML-Schemata sowie die Funktionen an die verschiedenen Manuskripttypen und Forschungsanforderungen angepasst. Derzeit wird »ediarum« für die Historisch-kritische Edition der Schriften Jeremias Gotthelf zur Verfügung gestellt (in Kooperation mit der Universität Bern). Weitere Implementierungen befinden sich in Planung.

Inhalt des Workshops
Im Workshop wird Interessenten das technische Konzept von ediarum vorgestellt. Nach einem kurzen Überblick werden die einzelnen Komponenten – oXygen XML Author, eXistdb und ConTeXt – und deren Zusammenspiel erklärt und demonstriert. Anhand praktischer Beispiele soll dabei jede Softwarekomponente, deren Funktionsweise und Implementierung eingehend dargestellt werden. Dabei können die Teilnehmer/-innen diese Beispielfunktionen über eine Testinstanz auch selbst nachvollziehen. Ein wesentlicher Schwerpunkt des Workshops wird es sein, auf Fragen und Problemstellungen der Teilnehmer/-innen einzugehen. Gerne können die Teilnehmer/-innen diese schon vorher den Referenten übermitteln.

Zielgruppe
Der Workshop richtet sich besonders an Wissenschaftler/-innen und Entwickler/-innen, die mit der konkreten technischen Umsetzung eines digitalen Editionsvorhaben konfrontiert sind. Selbstverständlich können auch Interessierte ohne Vorkenntnisse am Workshop teilnehmen, um sich einen Eindruck von ediarum zu machen. Aufgrund des zeitlich beschränkten Umfangs des Workshops können die Grundlagen der verwendeten Technologien leider nicht vermittelt werden. Um in möglichst hohem Maße vom Workshop zu profitieren, sollten die Teilnehmer/-innen daher nach Möglichkeit Basiskenntnisse in XML, XSLT und XQuery haben. Auch grundlegende TeX-Kenntnisse wären vorteilhaft.

Teilnehmerzahl
Max. 20 Teilnehmer/-innen

Benötigte Ausstattung seitens der Veranstalter
Beamer und Internetzugang. Möglichkeit für die Teilnehmer/-innen eigene Laptops mitzubringen, anzuschließen und mit dem Internet zu verbinden

Empfohlene Ausstattung seitens der Teilnehmer/-innen
Eigener Laptop mit bereits installiertem oXygen XML Editor (nach Möglichkeit in der aktuellen Version, siehe http://www.oxygenxml.com/download_oxygenxml_editor.html).

Referenten

Stefan Dumont
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23
10117 Berlin
030 / 20 370 -684
dumont@bbaw.de

Stefan Dumont arbeitet als wissenschaftlicher Mitarbeiter bei der TELOTA-Initiative der Berlin-Brandenburgischen Akademie der Wissenschaften. Er forscht dort über digitale Editionen und entwickelt Werkzeuge zu deren Erstellung. Darüber hinaus beschäftigt er sich mit der Vernetzung von Briefeditionen mit Hilfe digitaler Methoden.

Martin Fechner
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23
10117 Berlin
030 / 20 370 -684
fechner@bbaw.de

Martin Fechner forscht zur Wissenschaftskommunikation physikalischer Themen im Bereich der Wissenschaftsgeschichte, außerdem forscht er zu neuen, digitalen Methoden für die Geisteswissenschaften und zu digitalem Publizieren. Als wissenschaftlicher Mitarbeiter von TELOTA ist er Entwickler mehrere digitaler Werkzeuge aus dem Umfeld der Editionswissenschaften.

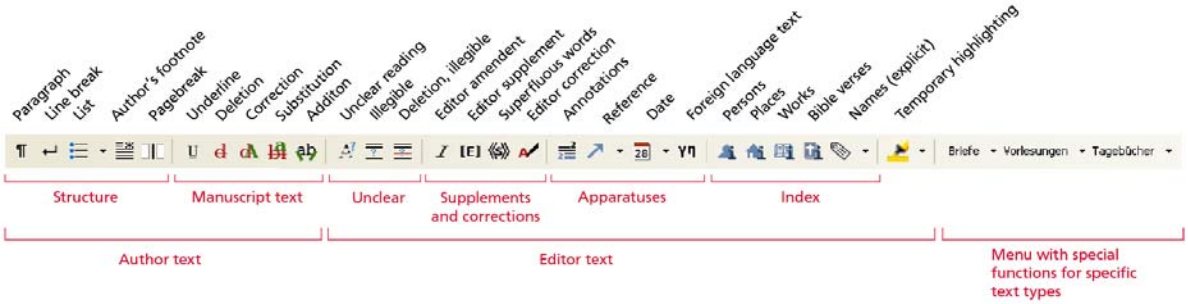
Weitere Informationen

- Projektwebsite: http://www.bbaw.de/telota/software/ediarum

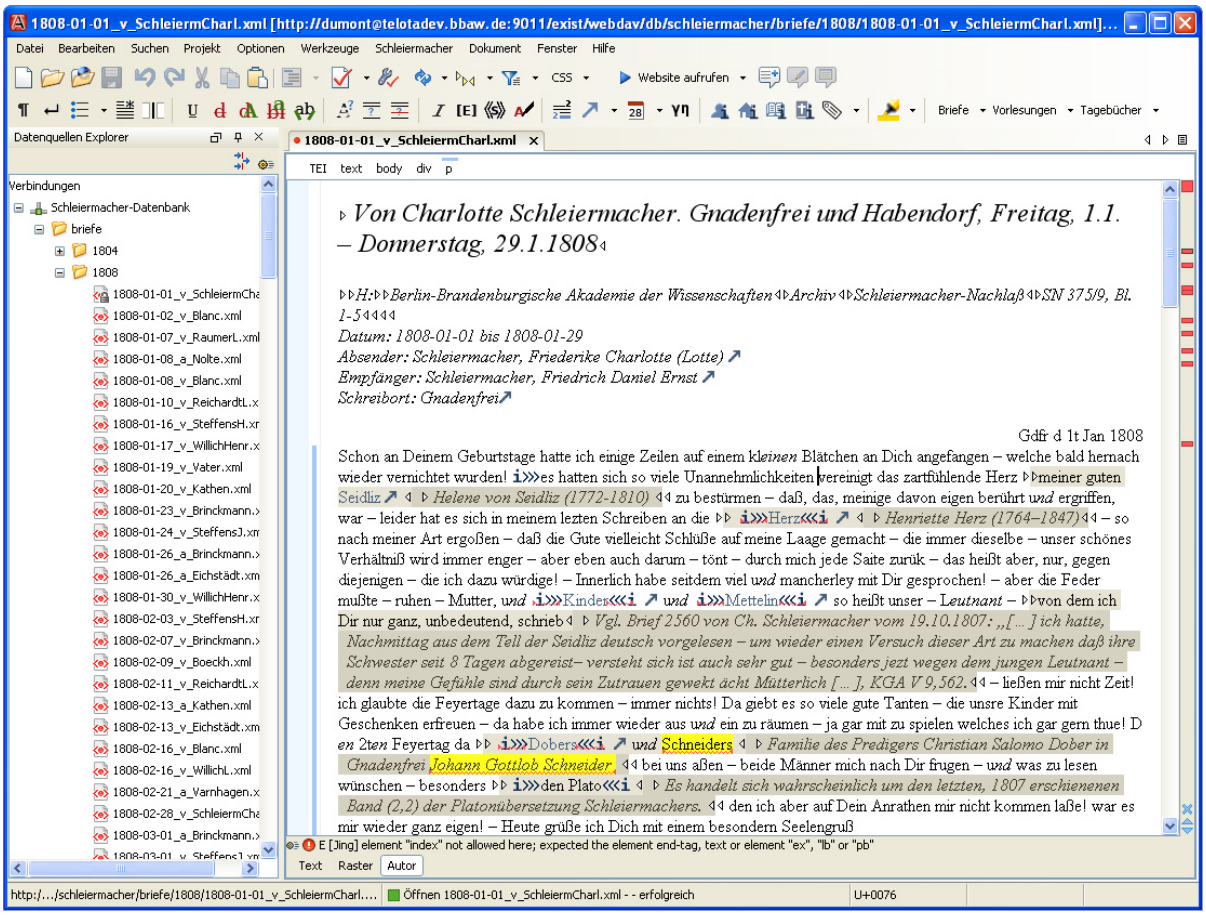
Literatur

- Dumont, Stefan; Fechner, Martin: Digitale Arbeitsumgebung für das Editionsvorhaben »Schleiermacher in Berlin 1808—1834« In: digiversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften. URL: <http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834/>
- Burnard, Lou; Bauman, Syd (Hg.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Charlottesville, Virginia, USA 2014. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>
- User Manual oXygen XML Author 14. URL: <http://www.oxygenxml.com/doc/ug-editor/>
- eXist Main Documentation. URL: <http://www.exist-db.org/exist/documentation.xml>
- ConTeXt Dokumentation. URL: <http://wiki.contextgarden.net/Main_Page>

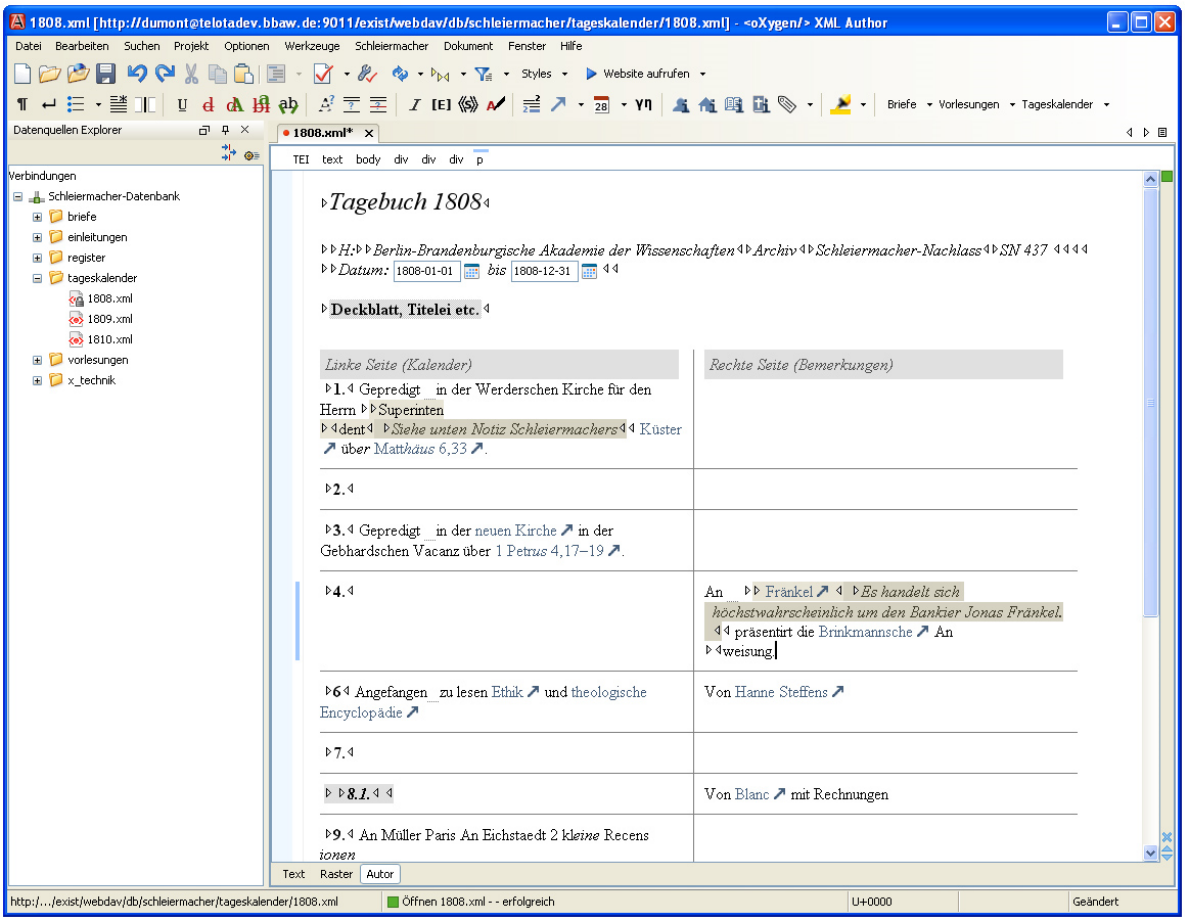
Screenshots



Eigene Werkzeugleiste für die Schleiermacher-Edition in oXygen XML Author



Transkription eines Briefes in oXygen XML Author



Transkription eines Tageskalenders in oXygen XML Author

Workshop: Automatisierte Handschriftenerkennung mit der Transcription & Recognition Platform (TRP)

Hindergrund

Im Rahmen des Projekts tranScriptorium, das sich mit der automatisierten Erkennung historischer Handschriften beschäftigt, entwickelt die Projektgruppe „Digitalisierung und Elektronische Archivierung“ (DEA) am Institut für Germanistik der Universität Innsbruck eine Plattform (TRP) mit deren Hilfe handschriftliche Dokumente in neuartiger Weise erschlossen werden können. (Sánchez et al. 2013)

Wie die Vorträge bei der International Conference on Frontiers in Handwriting Recognition (2014) gezeigt haben, handelt es sich dabei um eine Technologie, die am Sprung zum Praxiseinsatz steht. Das größte Hindernis stellt jedoch die ungenügende Menge an Referenzdaten dar, die daher bei den eingesetzten statistischen Verfahren oftmals zu unbefriedigenden Ergebnissen führen. (Plötz & Fink 2009)

Die von uns im Rahmen es Workshops vorgestellte Plattform will diesem Mangel nun dadurch begegnen, dass sie Geisteswissenschaftlern die Möglichkeit bietet, die Transkription eines handschriftlichen Textes in einer besonders komfortablen und teilweise mit automatisierten Methoden unterstützten Art und Weise durchzuführen. Also besonderer Anreiz ist hierbei die enge Verbindung zwischen Text und Bild auf Block, Zeilen und Wortebene zu nennen, zum anderen die standardisierten Exportformate: TEI (Text Encoding Initiative) sowie PDF (Portable Document Format) zur lokalen Benutzung, aber auch METS (Metadata Encoding and Transmission Standard) für die Integration in Repositorien wie etwa FEDORA. (Mühlberger 2014)

Gleichzeitig können nun die von Geisteswissenschaftlern produzierten Transkriptionen auch für das Training von Handwritten Text Recognition (HTR) Maschinen genützt werden. Mithilfe der automatisierten Erkennung kann nicht nur die Transkription selbst unterstützt werden, sondern können auch noch nicht transkribierte, größere Mengen von Dokumenten automatisiert erkannt werden.

Die zu erzielenden Genauigkeiten hängen von vielerlei Faktoren ab. Erste Experimente zeigen jedoch, dass bei einem nicht zu komplexen Layout und gerader Linienführung durchaus Ergebnisse von unter 30% Word Error Rate zu erzielen sind. (Romero et al. 2013)

Ziele des Workshops

Die Teilnehmer des Workshops erhalten die Möglichkeit mit einer Betaversion der Plattform zu arbeiten und die automatisierte Erkennung an Beispieldokumenten durchzuführen. Die Software wird Anfang 2015 öffentlich von der Webseite des Projekts zum Download angeboten. Ab diesem Zeitpunkt haben die Teilnehmer folgende Möglichkeiten.

- Sich als Benutzer in der Plattform zu registrieren
- Eigene Bilddateien auf die Plattform hochzuladen
- Eine manuelle Transkription durchzuführen
- Eine automatisierte Block- und Zeilenerkennung durchzuführen
- Blöcke, Zeilen, Wortsegmente edieren

- Ausgewählte Beispieldokumente (Schriften des Jeremy Bentham aus dem Transcribe Bentham Projekt) automatisiert zu erkennen
- Die entsprechenden Dokumente in den angebotenen Standardformaten (TEI, METS, PAGE, PDF) zu exportieren

Ablauf des Workshops

Der Workshop wird aus drei Teilen bestehen:

1. Einführung in das Thema Handwritten Text Recognition (Vortragender: Joan Andreu Sanchez) – ca. 30‘
   Vorgestellt werden die grundlegenden Technologien und Tools, die der automatisierten Handschriftenerkennung zugrunde liegen. Joan Andreu Sanchez ist wissenschaftlicher Koordinator des EU Projekts tranScriptorium und Professor für Comuter Science an der Technischen Universität Valencia.

2. Vorstellung der Transcription & Recogniton Platform (Vortragender: Günter Mühlberger) – ca. 30‘
   Hier wird auf das grundlegende Konzept der Plattform eingegangen und die Idee einer digitalen Infrastruktur zur Erkennung von Handschriften im Detail erläutert. Günter Mühlberger leitet die Gruppe „Digitalisierung und elektronische Archivierung“ (DEA) am Institut für Germanistik der Universität Innsbruck und ist für das Arbeitspaket „Datenmanagement“ im oben genannten Projekt verantwortlich.

3. Einführung in das Tool „Transcribus“ (Vortragende: Sebastian Colutto und Philip Kahle) – ca. 30‘
   Das für den Geisteswissenschaftler wichtigste Interface zur HTR Technologie im Rahmen der Transcription & Recognition Platform ist das Tool „Transcribus“. Es ist mit JAVA und SWT programmiert und muss lokal installiert werden. Allerdings werden die Bilder und Daten mittels einer Remoteverbindung zum TRP Server geladen und gespeichert. Auf diese Weise kann ein sehr flüssiges Arbeiten mit einem „Rich Client“ erzielt werden. Sebastian Colutto und Philip Kahle sind seit mehreren Jahren Projektmitarbeiter in der DEA Gruppe und arbeiten seit knapp zwei Jahren intensiv an dem vorliegenden Prototypen.

4. Selbständiges Arbeiten mit der Plattform bzw. Transcribus - ca. 2,5h
   Die Teilnehmer sollen die Möglichkeiten und Grenzen der Technologie in allen Einzelheiten an ihrem PC ausprobieren können und werden dabei von den vier Vortragenden unterstützt.

Zielgruppen für den Workshops

1. Geisteswissenschaftler, die mit der wissenschaftlichen Edition handschriftlicher Texte befasst sind. Sie erhalten mit Transcribus bzw. TRP ein mächtiges Werkzeug, das sie in ihrer täglichen Arbeit unterstützen soll.
2. Archive und Bibliotheken die interessiert sind, ihre digitalisierten handschriftlichen Bestände für die Öffentlichkeit zu öffnen. In diesem Fall kann TRP benutzt werden, um ein Team von ehrenamtlichen Mitarbeitern koordinieren zu können.

Literatur

Mühlberger, G., 2014. Die automatisierte Volltexterkennung historischer Handschriften als gemeinsame Aufgabe von Archiven, Geistes- und Computerwissenschaftlern. Das Modell einer

zentralen Transkriptionsplattform als virtuelle Forschungsumgebung. In *Digitalisierung im Archiv. Neue Wege der Bereitstellung des Archivguts. Beiträge des 18. Archivwissenschaftlichen Kolloquiums am 26. und 27. November 2013*.

Plötz, T. & Fink, G. a., 2009. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(4), pp.269–298. Available at: http://link.springer.com/10.1007/s10032-009-0098-4 [Accessed August 20, 2013].

Romero, V.V. et al., 2013. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6), pp.1658–1669. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0031320312005080.

Sánchez, J.A. et al., 2013. tranScriptorium: an European Project on Handwritten Text Recognition. *DocEng'13, September 2013, Florence, Italy*, pp.227–228.

Workshop

Es geht auch ohne Formeln

Der Einsatz von TEX in den Digital Humanities am Beispiel kritischer Editionen

Martin Sievers

20. Januar 2015

1 Einleitung

Die Diskussion rund um digitale Editionen als Ergänzung oder sogar Ersatz für die klassische Buchausgabe ist in vollem Gange. Grund dafür ist auch der Siegeszug der plattform- und implementationsunabhängigen Metasprache XML in den „Digital Humanities“. Insbesondere der TEI-Standard[1] hat dafür gesorgt, dass Informationen aller Art in einem Sammelformat vorliegen, das von vielen neu entwickelten Werkzeugen als Ausgabe- und Austauschformat verwendet wird.

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

Projekte wie XML-Print[2] oder Apache FOP[3] setzen hier an und wollen das Textsatzproblem innerhalb der „X-Technologien“[4] lösen. Es ist in den vergangenen Jahren jedoch deutlich geworden, dass der wissenschaftliche Textsatz von diesen Werkzeugen (noch) nicht in all seinen Facetten erfasst werden kann. Daher greifen aktuelle Editionsprojekte in der Regel auf etablierte

1 Vgl. Burnard und Bauman 2007.

2 Siehe dazu *XML-Print: typesetting arbitrary XML documents in high quality* 2014.

3 Siehe dazu *Apache(tm) FOP – a print formatter driven by XSL formatting objects (XSL-FO) and an output independent formatter* 2014.

4 Unter den X-Technologien versteht man die W3C-Standards XML, XSL und XPath sowie je nach Kontext weitere im XML-Umfeld entstandene Sprachen und Formate wie XQuery oder XLink.

Werkzeuge wie TUSTEP oder andere nicht notwendigerweise XML-basierte Ansätze zurück, indem der XML-Eingabetext mittels XSLT in die entsprechenden Zwischenformate transformiert wird.

Genau hier möchte der Workshop ansetzen und die Möglichkeiten des Textsatzsystems TEX[5] im Bezug auf die Erstellung einer historisch-kritischen Ausgabe (kritische Edition) vorstellen. Leider wird für dieser Weg aus Unwissenheit bzw. wegen falscher oder schlicht veralteter Informationen bzgl. des Funktionsumfangs viel zu selten beschritten. Umso erfreulicher sind Forschungs- und Arbeitsumgebungen wie FuD[6] oder ediarum[7], die am Ende des editorischen XML-basierten Workflows eine mit TEX erzeugte PDF-Datei zur Kontrolle bzw. als Vorstufe des Druckergebnisses ausgeben.

2 Funktionsweise von LATEX

2.1 Allgemein

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

Die so entstandenen Algorithmen, z. B. derjenige für den Zeilenumbruch[8] waren bahnbrechend und sind bis heute „State of the Art“. Entsprechend werden sie auch in jüngerer Software wie Adobe InDesign oder Apache FOP nahezu unverändert verwendet. Für den Autor eines Texts bedeutet dies, dass er sich vollständig auf die inhaltliche bzw. strukturelle Gestaltung konzentrieren kann. Dies entspricht der Arbeit mit XML-Quelldaten, die in der Regel keinerlei typographische Anweisungen enthalten.

Somit lebt die klassische Trennung zwischen Autor und Setzer wieder auf, die durch Text*verarbeitungs*programme stückweise aufgeweicht worden ist – mit negativen Folgen für die Druckqualität. In heutigen digitalen Arbeitsumgebungen entspricht der Setzer einem Satzprogramm, das eine Druckvorlage auf die Quelldokumente eines Autors anwendet.

5 TEX wird seit langem schon nur noch selten direkt angewendet, sondern in der Regel über eine Makrosprache wie LATEX (siehe z. B. http://www.latex-project.org/) oder ConTEXt (siehe z. B. http://wiki.contextgarden.net/What_is_ConTeXt) angesprochen. Der Workshop konzentriert sich auf LATEX.

6 Siehe dazu *FuD – Eine virtuelle Forschungsumgebung für die Geistenwissenschaften* 2014.

7 Siehe dazu *ediarum – eine digitale Arbeitsumgebung für Editionsvorhaben* 2014.

8 Vgl. dazu Knuth und Plass 1981.

2.2 Anwendungsfall Kritische Edition

Eine kritische Edition stellt besondere Anforderungen an ein Textsatzwerkzeug. Daher eignet sich dieser Dokumenttyp besonders gut, um die Qualität von LaTeX zu demonstrieren. Plachta definiert für eine kritische Edition zehn elementare Bestandteile.[9] Diese lassen sich zu den folgenden drei Themenkomplexen zusammenfassen:

- Edierter Text
- Apparate
- Verzeichnisse

Zu all diesen Bereichen liefert LaTeX entweder direkt oder über Erweiterungen Möglichkeiten, hochwertige Ergebnisse zu erzielen. Sie stehen über das zentrale Paketarchiv CTAN[10] allen Nutzern kostenfrei zur Verfügung.

3 Inhalte des Workshops

Entlang der in Abschnitt 2.2 genannten Bestandteile einer kritischen Edition wird der Workshop den Teilnehmern die Gelegenheit geben, LaTeX als Satzprogramm kennenzulernen. Im Detail werden die folgenden Inhalte vermittelt:

Edierter Text: Neben den grundlegenden Algorithmen werden mikrotypographische Fragen thematisiert. Dazu gehören neben der Nutzung typographisch korrekter Zeichen (z. B. bei Anführungszeichen, Gedankenstrich oder Ellipse) auch der automatische optische Randausgleich oder die Laufweitenänderung, die beide durch das Paket `microtype`[11] bereitgestellt werden.

Apparate: Die Anforderungen an den Satz von Apparaten gehen weit über diejenigen „normaler“ Fußnoten hinaus. Es wird das Paket `eledmac`[12] vorgestellt und neben verschiedenen Anpassungen auch Lösungen für Probleme wie z. B. überlappende Lemmata erarbeitet.

Verzeichnisse: Eine kritische Edition enthält neben einem Quellenverzeichnis verschiedene Register. Diese sollen im Idealfall direkt aus den Druckdaten dynamisch erzeugt werden. Dazu werden die Erweiterungen `biblatex`[13] sowie `splitindex`[14] vorgestellt, die genau dies bewerkstelligen.

9 Plachta 2006, S. 14 f.
10 Das Comprehensive TeX Archive Network stellt über zwei zentrale Server und mehr als hundert Spiegelserver (Mirrors) weltweit unter http://www.ctan.org insgesamt über 4500 Erweiterungen bereit.
11 Vgl. dazu http://www.ctan.org/pkg/microtype.
12 Vgl. dazu http://www.ctan.org/pkg/eledmac.
13 Vgl. dazu http://www.ctan.org/pkg/biblatex.
14 Vgl. dazu http://www.ctan.org/pkg/splitindex.

4 Teilnehmerkreis / Technische Ausstattung

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

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

Literatur

LaTeX – A document preparation system (2014). http://www.latex-project.org/ (besucht am 10. 11. 2014).

Apache(tm) FOP – a print formatter driven by XSL formatting objects (XSL-FO) and an output independent formatter (2014). http://xmlgraphics.apache.org/fop/ (besucht am 10. 11. 2014).

Burnard, Lou und Syd Bauman (2007). *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative.

ediarum – eine digitale Arbeitsumgebung für Editionsvorhaben (2014). http://www.bbaw.de/telota/software/ediarum (besucht am 10. 11. 2014).

FuD – Eine virtuelle Forschungsumgebung für die Geistenwissenschaften (2014). http://fud.uni-trier.de/de/ (besucht am 10. 11. 2014).

Knuth, Donald E. und Michael F. Plass (1981). "Breaking paragraphs into lines". In: *Journal Software: Practice and Experience* 11.11, S. 1119–1184. DOI: 10.1002/spe.4380111102.

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

What is ConTeXt (2014). http://wiki.contextgarden.net/What_is_ConTeXt (besucht am 10. 11. 2014).

XML-Print: typesetting arbitrary XML documents in high quality (2014). https://sourceforge.net/projects/xml-print/files/.

DHd-Jahrestagung 2015, Graz – Pre-Conference Workshop

Einführung in die Nutzung der Edirom Tools im Kontext digitaler Musikeditionen

Übersicht

Unter dem Namen Edirom Tools hat sich beginnend mit deren Entwicklung im Jahr 2004 eine Sammlung von Softwarewerkzeugen für die Bedürfnisse digital arbeitender Musikeditionsprojekte etabliert. Die Edirom Tools unterstützen solche Projekte bei der Sammlung, Erschließung und Organisation digitalisierter und digitaler Materialen. Sie werden damit aber auch der im Bereich der Musikwissenschaft respektive der Musikedition zunehmenden Hinwendung zu digitalen Publikationsmodellen gerecht. Die Entwicklung dieser Werkzeuge fand von Beginn an in enger Kooperation mit Editionsprojekten statt, um einen praxisnahen Einsatz in unterschiedlichen Werk- und Komponistenkontexten zu gewährleisten. So wird beispielsweise Mitte 2015 mit dem Abschluss der siebenbändigen ersten Abteilung „Orgelwerke“ der „Max Reger-Werkausgabe“ (RWA)[1] erstmals ein vollständiger Schaffensbereich eines Komponisten mit Hilfe der Edirom Tools digital ediert und publiziert vorliegen.

Weitere bisher mit Edirom arbeitende Projekte sind die „Carl-Maria-von-Weber-Gesamtausgabe“[2], „OPERA – Spektrum des Europäischen Musiktheaters in Einzeleditionen“[3], „Freischütz Digital“[4] und „A Cosmopolitan Composer in Pre-Revolutionary Europe – Giuseppe Sarti“[5]. Im Rahmen dieser Editionsprojekte werden die Edirom Tools in unterschiedlichen Werk- und Komponistenkontexten sowie Arbeits- und Publikationsumgebungen eingesetzt.

Ziel des Workshops ist die Erstellung einer digitalen Musikedition unter Anleitung auf Basis der Edirom Tools durch die Workshopteilnehmer. Er richtet sich damit an interessierte Musikwissenschaftler und Editoren, die sich grundlegend über den Einsatz und die Arbeit mit den Edirom Tools informieren möchten.

Ablauf des Workshops

I – Einführung und Kurzüberblick

Zunächst werden die verschiedenen Werkzeuge der Sammlung vorgestellt und ihr jeweiliger Einsatzzweck anhand von Beispielen aus unterschiedlichen Editionsprojekten skizziert.

II – Digitales Edieren

Der zweite Teil des Workshops widmet sich den grundlegenden Arbeitsschritten des digitalen Edierens mit dem Edirom-Editor. Es wird gezeigt, wie Quellendigitalisate importiert, strukturiert

1 Siehe: http://www.max-reger-institut.de/de/rwa.php
2 Siehe: http://www.weber-gesamtausgabe.de
3 Siehe: http://www.opera.adwmainz.de
4 Siehe: http://www.freischuetz-digital.de
5 Siehe: http://www.udk-berlin.de/sites/musikwissenschaft/content/forschung/a_cosmopolitan_composer_in_pre_revolutionary_europe__giuseppe_sarti/index_ger.html

und in Werkkontexten organisiert werden können. Dazu zählt vor allem die Definition von Quellen-, Satz- und Taktbeziehungen, die im Datenformat MEI (Music Encoding Initiative)[6] transparent hinterlegt werden. Darauf aufbauend sollen die Möglichkeiten der digitalen Quellenautopsie, des Kollationierens „am Bildschirm“ und der Erfassung von Annotationen mit dem Edirom-Editor behandelt werden.

III – Digitales Publizieren

Im dritten Teil des Workshops werden die Möglichkeiten der Publikation und der inhaltlichen Anreicherung im Rahmen der Edirom-Online behandelt. Dazu wird die Übernahme der Editionsdaten aus dem Edirom-Editor gezeigt und wie diese bei Bedarf um weitere Inhalte wie Texte (z.B. TEI) oder Abbildungen erweitert und – bei Bedarf – projektspezifisch angepasst werden können.

Voraussetzungen

Für die Teilnahme an diesem Workshop ist ein eigener Rechner (Windows oder OS X) mit einer aktuellen Java-Umgebung sowie eine Testversion des oXygen XML Editors notwendig. Diese ist kostenlos auf der Herstellerseite[7] erhältlich. Weitere technische Vorkenntnisse werden nicht vorausgesetzt.

6 Siehe: http://music-encoding.org
7 Siehe: http://www.oxygenxml.com

DHd2015

Von Daten zu Erkenntnissen

23. bis 27. Februar 2015, Graz

Vorträge

CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

Distant Waching.
Ein quantitativer Zugang zu YouTube-Videos

Mag. Dipl.-Ing. Gernot Howanitz

Stipendiat der Österreichischen Akademie der Wissenschaften (DOC)
am Lehrstuhl für Slavische Literaturen und Kulturen, Universität Passau

Zusammenfassung

Um virale Videos auf *YouTube* untersuchen zu können, ist ein quantitativer Zugang unumgänglich. Die Literatur- und Kulturwissenschaften – insbesondere die Slavistik – beschränken sich allerdings größtenteils auf textbasierte Verfahren. Das hier vorgestellte ‚distant watching' kann die Schranken des Textes überwinden, weil es erlaubt, eine große Anzahl an Videos mit einem Blick zu erfassen und Rückschlüsse auf deren grundlegende kreative Mechanismen zu ziehen.

1. Überblick

Ein nicht unbedeutender Teil der Internetkultur besteht aus Bildern und viralen Videos, die kopiert, verändert und weiterverbreitet werden und damit zum „Mem" werden (Dawkins 2013; Shifman 2013). Die schiere Masse an Material, die dabei entsteht, widersetzt sich den ‚klassischen' Mitteln der Literatur- und Kulturwissenschaft – vor allem dem „close reading". Vor ein vergleichbares Problem gestellt – nämlich angesichts der bis zu 60.000 Romane, die im England des 19. Jahrhunderts gedruckt worden sind, hat der Literaturwissenschaftler Franco Moretti vorgeschlagen, Texte ‚aus der Ferne' zu lesen. Statistiken wie etwa Worthäufigkeiten verknüpft mit mathematischen Visualisierungsstrategien werden im Rahmen des sogenannten „distant reading" dazu genutzt, grobe Zusammenhänge zwischen den Texten ‚herauszulesen', ohne sie im Detail studieren zu müssen. Franco Moretti plädiert dabei für eine Kombination von „distant reading" und „close reading". Mittels ‚Fernlesens' werden die Texte ausgewählt, mit denen man sich anschließend eingehender beschäftigen will, die man also ganz ‚klassisch' liest:

> [W]e know how to read texts, now let's learn how *not* to read them. Distant reading: where distance [...] *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, [...] one can justifiably say, Less is more. (Moretti 2000:57, Hervorh. i. O.)

Allgemein sind quantitative Methoden mittlerweile in der Literatur- und Kulturwissenschaft angekommen. Texte stehen dabei eindeutig im Vordergrund, was nicht weiter verwundert, da Wörter zu zählen technisch sehr einfach zu bewerkstelligen ist. Doch auch für Bilder ist ein ähnlicher Zugang möglich (Kwastek 2013). Nachfolgend soll deshalb eine neue Methode vorgeschlagen werden, die es erlaubt, *YouTube*-Videos nach quantitativen Gesichtspunkten zu untersuchen: ‚distant watching' (‚Fern-Sehen'). Als konkretes Anwendungsbeispiel soll das tschechoslowakische Musikvideo „Jožin z bažin" („Seppl aus dem Sumpf") aus dem Jahr 1978 dienen, das 2008 im Internet seinen Durchbruch gefeiert hat und zu einem Mem geworden ist.

2. Vorarbeiten

Entstanden ist die Idee zu ‚distant watching' im Rahmen eines Projekts zu Online-Erinnerungen an den umstrittenen ukrainischen Nationalhelden Stepan Bandera in Polen, Russland und der Ukraine (Fredheim/Howanitz/Makhortykh 2014). Dabei wurden zunächst die Aktivitäten von Nutzerinnen und Nutzern auf *Wikipedia*, *YouTube* und *Twitter* verglichen. Konkret wurden Zugriffszahlen, Geoinformationen, Artikel, Tweets und Kommentare quantitativ erfasst und verarbeitet. Um auch die Videoinformationen auf YouTube miteinbeziehen zu können, wurde von mir ein erster, rudimentärer Prototyp des ‚distant watching'-Verfahrens entwickelt. Diesen Prototypen habe ich nun grundlegend überarbeitet und weiterentwickelt, um ihn auf eine große Bandbreite von Problemen der Internetkultur anwenden zu können.

3. Methodologie

Die Grundidee von ‚distant watching' ist, die ersten 200 Resultate für einen Suchbegriff – beispielsweise „Jožin z bažin" – auf *YouTube* als Korpus zu verwenden. Die Einzelbilder (Frames) dieser 200 Videos dienen als statistische Grundlage. Zunächst werden die Unterschiede zwischen aufeinanderfolgenden Frames berechnet. Dabei entstehen charakteristische ‚Schnittkurven', die zeigen, dass sich alle Videos in fünf verschiedene (technische) Genres einreihen: geschnittene und ungeschnittene Videos, Videocollagen, Standbilder und Diashows. In den Schnittkurven fungiert die x-Achse als Zeitachse, die allerdings nicht Sekunden, sondern Einzelbilder erfasst. Auf der y-Achse werden die Unterschiede zwischen benachbarten Frames eingetragen, wobei 1,0 als 100% Übereinstimmung zu lesen ist, also kein Unterschied feststellbar ist, während 0,5 für 50% Übereinstimmung steht, etc. In Abbildung 1 sind Beispiele für alle fünf technischen Genres abgebildet. Sie stammen aus den *YouTube*-Suchresultaten für den Begriff „Jožin z bažin" und zeigen, auf welch verschiedene Weisen Nutzerinnen und Nutzer das Originalmaterial aufgreifen, bearbeiten und weiterverbreiten.

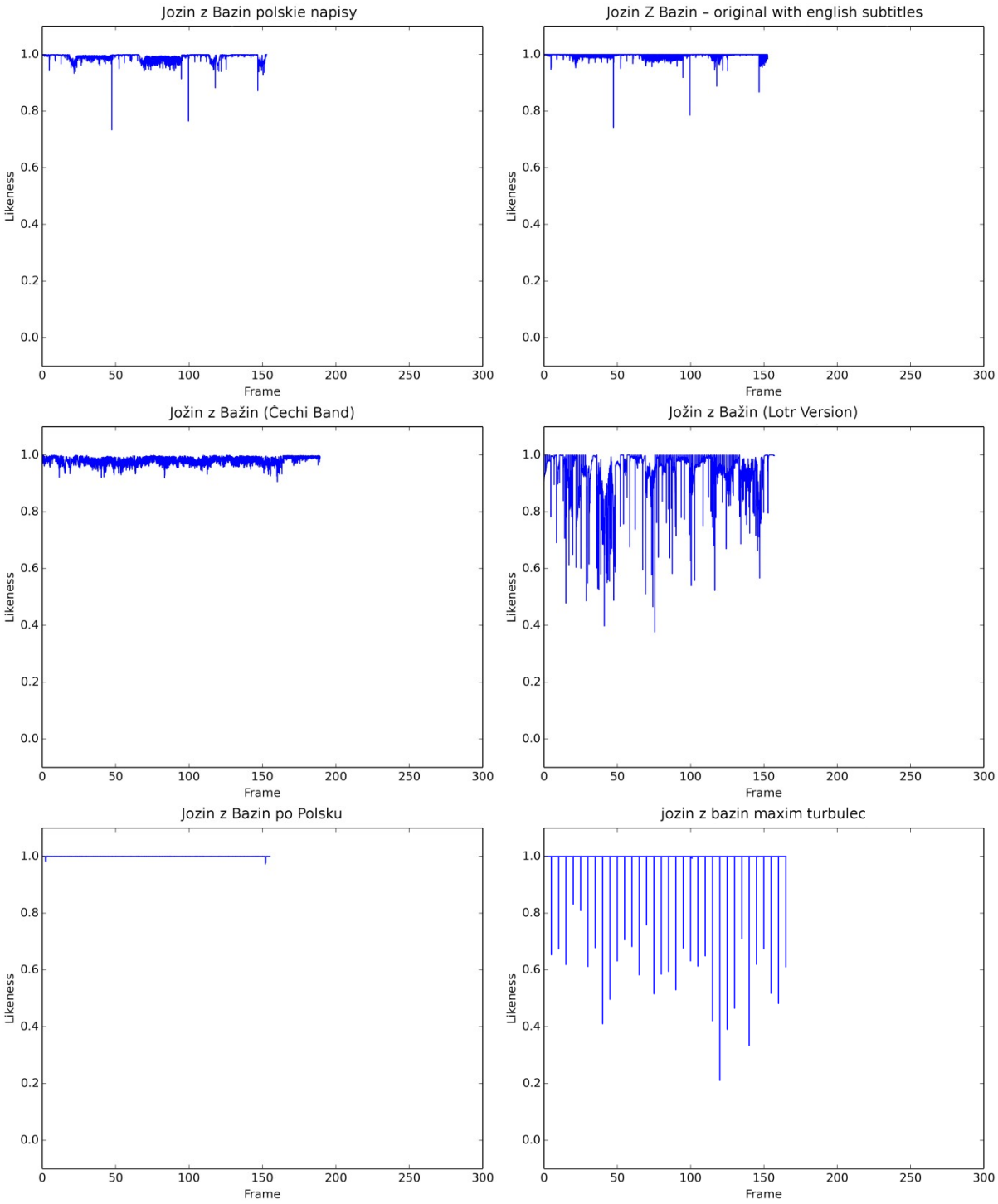


Abbildung 1. Schnittkurven zweier Versionen des (geschnittenen) Originalvideos (oben rechts und links), eines ungeschnittenen Amateurvideos (Mitte links), einer Collage aus unterschiedlichen Videos (Mitte rechts), eines Standbildes (unten links) und einer Diashow (unten rechts).

Die Schnittkurven beschleunigen die (manuelle) Analyse der 200 Videos im Korpus, unter Zuhilfenahme der Hauptkomponentenanalyse (*Principal Component Analysis*, PCA) kann man darüber hinaus einzelne Videos automatisiert zueinander in Bezug zu setzen. Die PCA erlaubt es, hochkomplexe Probleme kompakt in Form einer zweidimensionalen Grafik darzustellen. Dazu muss man allerdings die Daten in eine einheitliche Form bringen. Im Falle der Videos ist dies aufgrund ihrer unterschiedlichen Länge notwendig. Um die Videos zu vereinheitlichen, werden sie über die 70 größten Framesunterschiede in absteigender Reihenfolge codiert, wobei jeweils die ersten und letzten zehn Sekunden eines Videos nicht berücksichtigt werden. Diese Herangehensweise wurde experimentell ermittelt. Weiters muss berücksichtigt werden, dass der größte Framesunterschied in manchen Videos (Abb. 1 links unten) keine Bedeutung hat, weil *alle* Unterschiede 1,0 betragen. Deshalb wurde 0,9 als Schwellenwert eingeführt, ist der Framesunterschied kleiner, so wird stattdessen ‚-1' eingetragen. Damit sind die Daten hinreichend aufbereitet und man kann die PCA durchführen. Diese reduziert die ursprünglich 70 Dimensionen (die 70 größten Framesunterschiede) des Problems auf zwei, wodurch sich die 200 Videos ganz einfach darstellen lassen, siehe Abbildung 2.

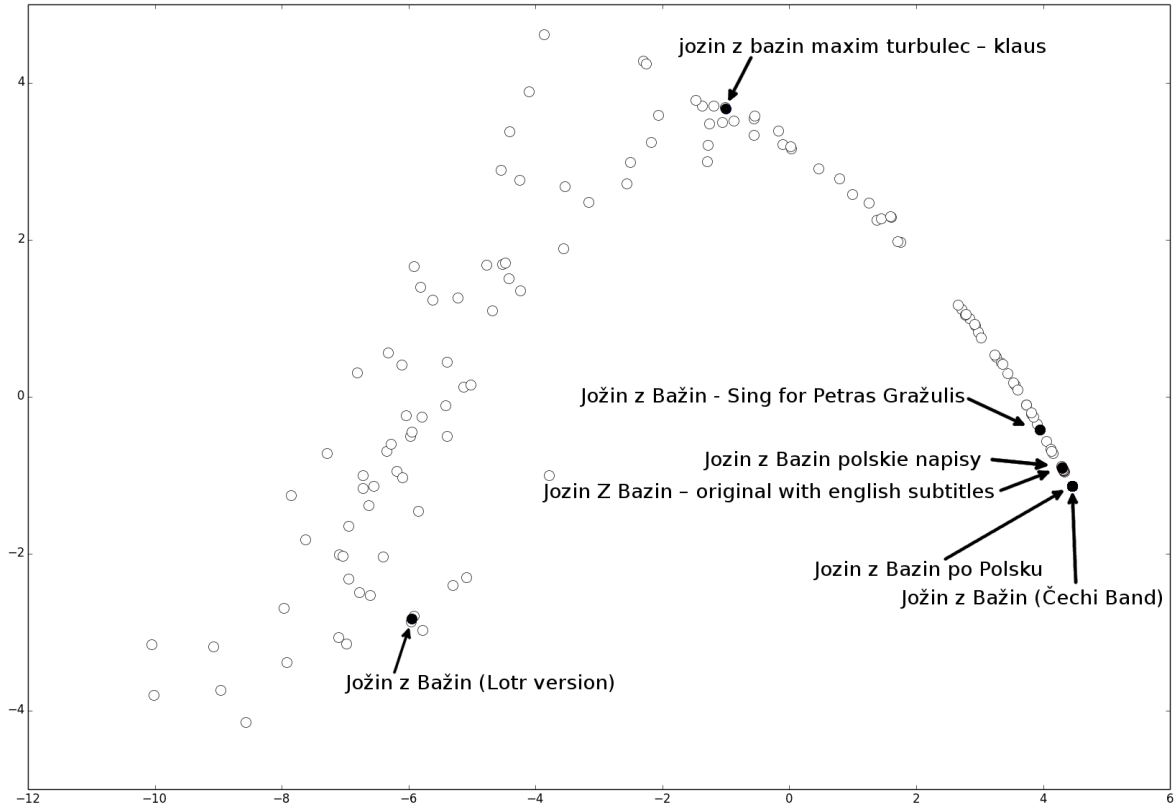


Abbildung 2. PCA der Einzelbilderunterschiede mit anschließender Projektion der Daten auf die von den ersten beiden Hauptkomponenten aufgespannte Ebene.

4. Ergebnisse

Die Schnittkurven sind ein gutes Hilfsmittel um das Genre eines einzelnen Videos auf einen Blick zu erfassen. So ist es offensichtlich, dass es sich bei den Videos oben links und rechts in Abbildung 1 um zwei unabhängig voneinander hochgeladene identische Versionen des originalen *Jožin z Bažin*-Videos handelt. Die Unterschiede ergeben sich aus der unterschiedlichen Videoqualität und den verwendeten Kompressionsverfahren. Die Schnittkurve in der Mitte links zeigt hingegen ein Video, das keine Schnitte enthält. Meist sind dies Amateurvideos. Rechts daneben sehen wir eine Videocollage, die passend zur *Jožin z Bažin*-Musik aus vielen sehr kurzen Videoschnipseln zusammengesetzt worden ist – hier konkret aus Peter Jacksons *Lord of the Ring: The Two Towers* (2002). Unten links ist ein Video, das nur aus einem einzelnen Standbild besteht, und unten rechts eine Diashow, die aus rhythmisch wechselnden Einzelbildern besteht. Da sich die technischen Genres hinsichtlich ihrer Aufwändigkeit unterscheiden, lassen sie erste Rückschlüsse bezüglich des Videoinhalts zu. Bei Standbildern ist es wahrscheinlich, dass die Musik im Vordergrund steht, weil die Videos optisch wenig zu bieten haben. Standbilder und Diashows sind einfach zu erstellen, geschnittene Videos hingegen anspruchsvoller. Dementsprechend selten sind selbst erstellte Videos unter den geschnittenen Videos; hier überwiegen Inhalte, die ursprünglich für das Fernsehen produziert worden sind.

Die Überblicksdarstellung der PCA erlaubt es, das Gesamtkorpus in den Blick zu nehmen. Wie man in Abbildung 2 sieht, fächert die PCA die Videos je nach Anzahl der Schnitte auf. Rechts unten finden sich sowohl das Amateurvideo der „Čechi Band“ als auch das Standbild *Jozin z Bazin po Polsku*. Nicht weit davon entfernt sind die zwei Versionen des Originalvideos – sie weisen relativ wenig Schnitte auf. Nach einer größeren Lücke in der ‚Perlenkette‘ kommt dann die Region der Videos mit häufigen Schnitten. Dementsprechend findet sich oben in der Mitte die Diashow *maxim turbulec*, und links unten die Videocollage *Lotr version*.

5. Diskussion

Der Masse an Material, welche die Netzkultur kreiert, kann man sinnvoll nur über quantitative Verfahren erfassen – beispielsweise mit ‚distant watching‘. Zwar gibt es hier noch Verbesserungspotential, etwa hinsichtlich der Erkennung ähnlicher Schnittfolgen, vor allem, weil es vielversprechend erscheint, diese Methode auf andere Probleme – etwa in der Filmanalyse – anzuwenden. Auch könnte eine quantitative Analyse der Tonspuren zusätzliche Einblicke bringen – etwa, ob das Originallied verwendet wird oder nicht. Statistische Informationen wie die Aufrufzahlen oder Anzahl der Kommentare auf *YouTube* können als zusätzliche Unterstützung verwendet werden. Auch können

die Frames mit Bilderkennungsalgorithmen weiter bearbeitet werden, um etwa das Auftreten bestimmter Symbolkombinationen quantitativ zu erfassen. All diese zusätzlichen Verfahren müssen natürlich auf die jeweilige konkrete Fragestellung abgestimmt werden.

Trotzdem liefert ‚distant watching' wesentliche Einblicke in das Videokorpus. Es legt die grundlegenden technischen Differenzen frei, aufgrund derer sich die fünf technischen Genres bestimmen lassen. Die graphische Darstellung der PCA kann dazu benutzt werden, sich einen Überblick über das Korpus in seiner Gesamtheit zu verschaffen und Videos eines bestimmten Genres schnell und unkompliziert zu finden. Sie zeigt Gruppen von Videos auf, die insofern interessant sind, als sie eigene thematische Akzente setzen. Zu betonen ist, dass erst eine qualitative Untersuchung einiger repräsentativer Videos bzw. Video-Gruppierungen genaueren Aufschluss über die innere Beschaffenheit des Korpus zulässt.

Quellen

Dawkins, R. 2013. „Just for Hits – Richard Dawkins", The Saatchi & Saatchi New Directors' Showcase. http://www.youtube.com/watch?v=GFn-ixX9edg (aufgerufen am 10. 11. 2014).

Fredheim, R. / Howanitz, G. / Makhortykh, M. 2014. „Stepan Bandera through the Lens of Quantitative Memory Studies", in *Digital Icons* 12 [im Druck].

Kwastek, K. 2013. „Vom Bild zum Bild: Digital Humanities jenseits des Texts". Keynote DHd 2013. http://video.uni-passau.de/video/Vom-Bild-zum-Bild%253A-Digital-Humanities-jenseits-des-Texts/92b59154b625a12070cc9d6947d6a611 (aufgerufen am 10. 11. 2014).

Moretti, F. 2000. „Conjectures on World Literature", in *New Left Review* 1, 54–68.

Shifman, L. *Memes in Digital Culture*, Cambridge (MA) 2013.

Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation, 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum, Graz, 23.-27.02.2015
Abstract zum Vortrag
Waltraud v. Pippich, Ludwig-Maximilians-Universität München, Institut für Kunstgeschichte

Farbe und Maß

Die Fibonacci-Zahlen in der Kunstgeschichte

Zahlenverhältnisse in Bildern

Farbe ordnet sich spektral im Farbkreis und ist als Licht eine der faszinierendsten Erscheinungen der Natur. Ist sie messbar? Als Anteile im Farbspektrum werden Farbwerte durch ein Computerprogramm erfasst, das Björn Ommer von der *Computer Vision Group* der Ruprecht-Karls-Universität Heidelberg entwickelt hat. Farbwerte werden durch das Programm quantifizierbar, indem feste Stellen der Farbskala als Schwellen gewertet werden und alle Pixel eines Bildes, die diese Schwellen erreichen, bestimmt werden können. Die Grundlagen der Studien zu Zahlenverhältnissen in Bildern liefern die durch das Computerprogramm bereit gestellten Daten zu den Farbfrequenzen. Durch die Arbeit mit dem Computerprogramm stellte sich heraus, dass bestimmte Zahlenverhältnisse in den Farbrelationen der Werkcluster häufiger auftauchen als andere. In zahlreichen Werken klingt im Miteinander der Farben die Relation der Fibonacci-Zahlen an und erzeugt die Wirkung der irrationalen Zahl Phi und des goldenen Schnittes.

Jede Zahl der Fibonacci-Folge (1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...) ist die Summe der zwei vorhergehenden Fibonacci-Zahlen, ihr Verhältnis zur vorangehenden Fibonacci-Zahl nähert die goldene Relation (1,6180...) an. Der Vortrag wird diese Zahlenverhältnisse für bestimmte Werkgruppen vorstellen und erklären. Unter den Werken sind berühmte Gemälde wie Raffaels Madonnendarstellungen, Menzels *Flötenkonzert in Sanssauci* und Barnett Newmans Gemälde *Vir heroicus sublimus*. Das irrationale Verhältnis der Zahl Phi ist nicht in ganzen Zahlen darzustellen, wohl aber visuell in Größenverhältnissen wahrzunehmen.

Es ist zu fragen, welche Rolle die Fibonacci-Zahlen in den Bildkompositionen einnehmen können. Einige Theorien um die Fibonacci-Folge befassen sich mit natürlichem Wachstum und besonders ergiebigen Zahlenverhältnissen in der Natur (D´Arcy Wentworth Thompson: On Growth and Form, 1992 (1917), S. 923-924). Ist der Malende als vor der Leinwand stetig wachsende Farbsummen produzierend zu imaginieren? Warum findet sich so häufig die Relation der Fibonacci-Zahlen in den Werken?

Relationen von Farbsummen

Die Forschung behandelt die Frage, welche Farbgruppen insbesondere in welcher Zeit zueinander in das Verhältnis der Fibonacci-Relation treten. Als sogenannter Modalwert einiger Werke zu den Nach-

barwerten auf der Farbskala ist die Fibonacci-Relation beispielsweise in zahlreichen beim Publikum erfolgreichen Gemälden des 19. Jahrhunderts aufweisbar. Einige dieser Werke sind dem sogenannten Protoimpressionismus zuzuordnen. Aber auch einige den kunsthistorischen Kanon lange bestimmende offizielle Herrscherporträts finden sich unter diesen Werken. In der Relation besonders dunkler und besonders lichter Bildbestandteile taucht die Fibonacci-Relation ebenso auf, auch in den Kleinstsummen unterschiedlicher Farbbestandteile eines Bildes. Ebenso wurden in Gemälden Zahlenverhältnisse ausgerechnet, die fast eine Hälfte der vom Bild erreichten Farb-*Range* in den Größen der auf der Skala aufeinander folgenden Farbwerte in die Fibonacci-Relation stellen. Welche Wirkung erzeugen diese Werke? Gibt es historische Entwicklungslinien der Kompositionsschemata? Welche *implicit patterns* beruhen in den Werkclustern, die zum Teil eine beträchtliche Größe aufweisen, auf Farbrelationen und Farbsummenrelationen (Beispiele einer Farbspektrenanalyse, Berechnungen zu Relationen des goldenen Schnittes, sowie vergleichende Farbwertanalysen in den Abb. 1, 2)?

Farbe und Musik

Bei der stilometrischen Farbforschung stellte sich, auch beim Ringen um Namen für die Muster und Kookkurrenzen in den Werken, heraus, dass sich Vergleiche zu musikalischen Kompositionen eher anbieten, als zu Clustern, die den *data mining* Verfahren der Linguistik zugrunde liegen (Textcorpora bestehend aus Sprachwerken). Die Farbbestandteile der Bilder verhalten sich in bestimmten Mengen zueinander und zum Gesamt des Bildes. Ihr Maß ist an das Maß der anderen Farbbestandteile gebunden. Wie Akkorde klingen die Farben zusammen, wie eine Konsonanz lässt sich ein Gleichklang der Farbsummen und ihr Abgestimmtsein zueinander, zum Modalwert und zu den Kleinstsummen beschreiben (Beispiele zu Versuchen stilometrischer *pattern detection* in den Darstellungen in Abb. 3). Das Abgestimmtsein von Farbgruppen exakt gleicher Größe und ein Entstehen von Bezügen durch Proportionsverhältnisse ähnelt der Bildung harmonischer Akkorde. Wie Tonhöhen, die sich innerhalb der Oktavordnung bestimmen lassen, erhalten die verschiedenen Farbwertsummen im Bild durch ihre nun mögliche numerische Ordnung eine Artikulation.

Schwierig zu beantworten ist die Frage, inwieweit innerhalb einer Ko-Partizipation von Farben in unterschiedlichen Farbsummenverhältnissen diese Farben in einem Bild in verschiedenen Fibonacci-Relationen zugleich wirken können, in denen sie auftauchen, sich selbst in verschiedene Bezüge einbinden und an den Farbrelationen teilhaben. Zu fragen ist auch, wie Zahlen periodischer Ordnung, beispielsweise die 5, 10, 15, 20, als Prozentanteile von Farbwerten in einem Bild gemeinsam mit den sich der Periodik entziehenden Fibonacci-Zahlen wirken können. Diese Zahlenverhältnisse wurden in Werken des französischen Klassizismus und des deutschen Realismus gemessen. Dies sind Fragen des Simultanen und die stilometrische Bildforschung liegt abermals, nun auch aus rezeptionsästhetischer Sicht, eher im Bereich zugleich wahrzunehmender Akkorde der Musik als im Bereich des sich sukzessive entwickelnden Verhaltens beim Lesen eines Textes oder des Hörens von Sprache.

Farbfrequenzen und Erkenntnis

Die durch das Software-Tool der *Computer Vision Group* des *Heidelberg Collaboratory for Image Processing* ermöglichten Farbfrequenzanalysen gewähren Einblicke in der Bildforschung auf diese Weise bislang nicht zugänglichen Relations- und Proportionseigenschaften der Farbgestaltungen. Für die stilometrische, aber auch die historisch ausgerichtete Forschung liegen in diesem in immer weiteren Kreisen zu erarbeitendem Wissen weitreichende Potentiale. Durch die Verlautbarung der Farben in numerischen Mengen haben diese vormals lautlosen Elemente der Bilder für die Wissenschaft nun eine Stimme gewonnen. Natürlich sind die Schwierigkeiten von *messy data* auch in der Bildforschung zu thematisieren. Innerhalb einer Demokratie der Pixel tragen alle Bildelemente zum Ergebnis der Farbsummenberechnungen bei. So tragen fehlende Pixelmengen ebenso wie fälschlicherweise mitgezählte Pixelmengen (etwa durch schmalste weiße Streifen an den Rändern der Abbildungsvorlagen) zu *messy data* als Grundlage der Berechnungen und damit zu unpräzisen Ergebnissen bei.

Der Punkt, an dem diese quantitativ ausgerichteten Studien an die Ebene der Semantik zurückzubinden sind, ist während stilometrisch und historisch ausgerichteter Arbeit mit dem Datenmaterial immer wieder zu befragen und bietet zahlreiche Perspektiven für die Einbindung weiterer, etwa komparativ ausgerichteter Forschungsfragen. An dieser Schwelle zwischen von einem Computerprogramm bereit gestellten Daten hin zu durch den Menschen zu gewinnenden Interpretationen und Versuche der Einordnung des Datenmaterials in stilgeschichtliche, kunsthistorische und geisteswissenschaftliche Zusammenhänge erneuern sich Fragen der bildwissenschaftlichen Disziplin nach dem Zusammenhang von Gestaltung und Wirkung, von Form und Sinn. So gelangt diese Forschung von Zahlen und auf der Grundlage von diesen errechneten Proportionsverhältnissen hin zu Fragen der Rezeptionsästhetik, Stilanalyse und Einbindung in kunsthistorische Traditionen. Nicht die suggestive Kraft von Farben, wohl aber Farbfrequenzen können nun durch das Computerprogramm erfasst und für die weitere Forschung gewonnen werden. Für die bildwissenschaftliche Forschung aufschlussreich sind dabei auch Bezüge zwischen der Linien- und Flächenkomposition eines Werkes, den Farbrelationen und dem Inhalt, dem dargestellten Thema eines Werkes.

Schönheit und Farbmaß

Auch Fragen zu Schönheit und Maß sind durch bildwissenschaftliche Beschäftigung mit Farbsummenverhältnissen zu stellen. Im Vortrag wird das sogenannte Pringsheim-Rätsel vorgestellt. Dies löst ein eindrückliches Porträt von Katharina Pringsheim, das Franz Lenbach anfertigte, in seiner eindringlichen Farbästhetik auf anhand der in den Farbproportionen aufzuweisenden irrationalen Relation der Fibonacci-Zahlen. Dieses Zahlenverhältnis trägt zur Schönheit der Darstellung, dem Wohlklang der Komposition bei.

Der Vortrag fasst die Ergebnisse der Farbwertmessungen zusammen und stellt diese an Beispielen dar, er thematisiert die Schwierigkeit des Erfassens der nicht distinkt, sondern graduell vorliegenden Farbeigenschaften durch ein Computerprogramm. Er beschreibt ein heuristisches Vergleichen der Ordnun-

gen musikalischer Harmonik mit Farbproportionen in Bildern und handelt von den Herausforderungen stilometrischer *pattern detection* auf der Grundlage von Farbwertanalysen. Der Vortrag gibt ein Beispiel der Perspektiven computergestützter Bildanalysen – für die Stilometrie der Bilder, ihre historische Ordnung und den Blick in die Disziplin der Musikwissenschaft.

Abb. 1

Abb. 2

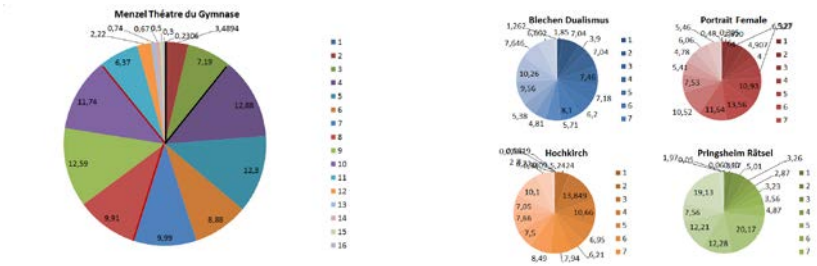
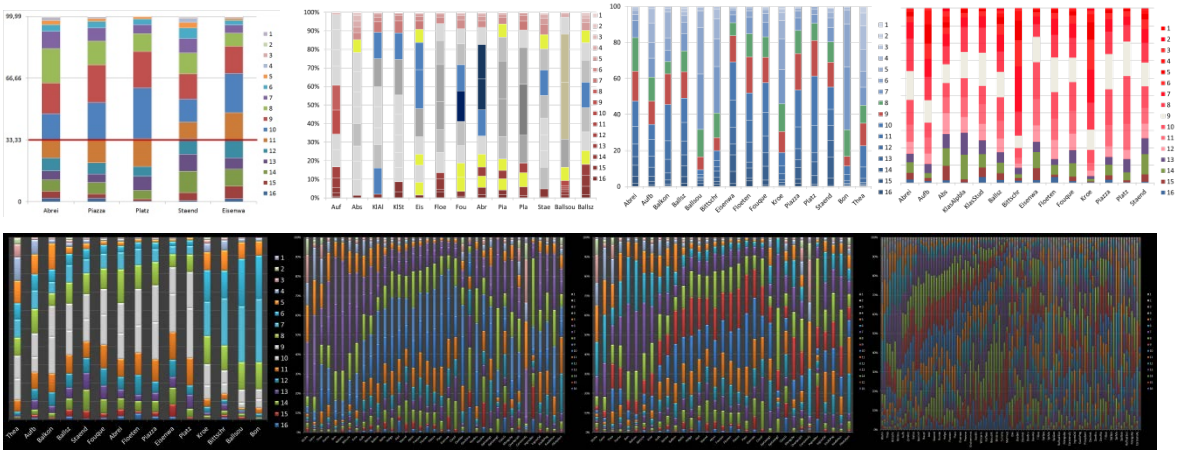


Abb. 3



Prof. Dr. Christian Kassung
Institut für Kulturwissenschaft
Humboldt-Universität zu Berlin

10.11.2014

Making Things Chatter. Historische Debatten und ihre Objekte im Museum

Am 10. Juni 1908 wurde das Märkische Museum feierlich eröffnet, nachdem sich der Kaiser eine Woche zuvor höchstpersönlich von der angemessenen Zurschaustellung der Berliner Stadtgeschichte überzeugt hatte. In der Morgenausgabe der Berliner Börsen-Zeitung wird der vom Geheimen Baurat Dr. Ludwig Hoffmann errichtete »Dicke Wilhelm« für seine »äußere und innere, durch ihre Einfachheit besonders wirkungsvolle Ausschmückung gelobt«, in der Abendausgabe des Berliner Tageblatts wird von einer schlichten Feier berichtet, »bei der jeder äußere Prunk vermieden war«, was dann auch den Bau charakterisieren sollte. Was beide Zeitungen indessen verschweigen, ist das nicht unprekäre Faktum, dass die Eröffnungsfeierlichkeiten bei hellichtem Tage stattfinden mussten, weil Hoffmann eine künstliche Beleuchtung seines Museums schlichtweg abgelehnt hatte. Obwohl bereits seit gut zwanzig Jahren mit der elektrischen Glühbirne eine ebenso effiziente wie ungefährliche Alternative zum Gaslicht zur Verfügung stand, setzte Hoffmann auf eine »raffinierte natürliche Lichtführung«, wie die Vossische Zeitung schreibt, und zwang das Museum damit zum Toresschluss, sobald die Dämmerung nahte.

Was auf den ersten Blick wie eine echte Berliner Schote anmutet, erweist sich bei näherem Hinsehen als ein, wenn auch prominenter, aber eben nur als ein Ausläufer einer erbitterten Debatte zwischen den Anhängern des alten Gaslichts und den Protagonisten der neuen Glühbirne, von der Allgemeine Elektricitäts-Gesellschaft zur Ikone einer zweiten industriellen Revolution erhoben. Noch zwanzig Jahre später, als der Museumsdirektor Walter Stengel das Märkische Museum elektrifizieren ließ, bettete er diese Maßnahme in eine groß angelegte Kampagne ein, die unter dem Titel »Berlin im Licht« um Akzeptanz der Glühbirne warb. Dokumentiert ist diese Debatte u.a. in dem von Johann Gottfried Dingler begründeten »Polytechnischen Journal« (1820–1931), das vom Institut für Kulturwissenschaft der Humboldt-Universität zu Berlin mit Förderung der DFG digitalisiert wurde. In einem Folgeprojekt, das die Humboldt-Universität gemeinsam mit der Stiftung Stadtmuseum Berlin und der Hochschule für Technik und Wirtschaft durchführt, werden diese und andere zeitgenössische Quellen so mit dem Wissensraum des Museums und den dort präsentierten Objekten verbunden, daß der Besucher die Geschichte der Elektrifizierung in ihrer ganzen Breitenwirkung auf das alltägliche Leben erfahren kann. Als Endgeräte kommen dabei sowohl Tablets über eine entsprechende App zum Einsatz, wie speziell für das Projekt entwickelte »Licht-Würfel«, die als Token fungieren, um unterschiedlichste multimediale Inhalte zu steuern.

Elektrizität ist aus einem sehr spezifischen Grunde besonders geeignet, den dinghaften Wissensraum eines Museums mit dem symbolischen Wissensraum einer Datenbank zu verknüpfen. Einerseits lässt sich die Sozial-, Kultur- und Technikgeschichte einer Stadt, v. a. einer »Elektropolis« wie Berlin nicht ohne die vielfältigen Einflüsse dieser neuen Energieform verstehen. Andererseits aber ist die Elektrizität selbst schlichtweg unsichtbar, d. h. sie wird nur erfahr- und begreifbar in ihren jeweiligen Manifestationen und Wirkungen im Objekt. Dies eröffnet einen sehr breiten Raum für Spekulationen, Ängste oder Euphorien, die sich in Debatten dieser Zeit nicht nur niederschlagen, sondern zu einer öffentlichen Frage ersten Ranges werden. Im Zentrum dieser Debatten stehen das Märkische Museum mit ihrem Protagonisten Ludwig Hoffmann und die Glühbirne mit ihrem Protagonisten Emil Rathenau. Im »Polytechnischen Journal« haben sich diese Debatten in Form von gut zweihundert

Artikeln niedergeschlagen, die explizit das Problem der Einführung des elektrischen Lichts in Berlin diskutieren. So lernen wir aus einem Text von 1889 beispielsweise, dass die Berliner Elektricitätswerke Strom folgendermaßen verkaufen:

> Für Elektromotoren ist eine monatliche Grundtaxe von 1 M. für 1 Ampère der Maximalleistung zu zahlen. Diese Taxe wird nicht erhoben, wenn der Verbraucher sich bereit erklärt, auf die Lieferung des elektrischen Stromes während der Wintermonate von Sonnenuntergang bis 11 Uhr Abends zu verzichten, im Falle die Beanspruchung der Centralstationen für die elektrische Beleuchtung dies erfordern sollte. (http://dingler.culture.hu-berlin.de/article/pj272/ar272015, 23.10.2014)

Anders formuliert: Im Gegensatz zu den Städtischen Gaswerken waren die Elektrizitätswerke also nicht in der Lage, die erforderlichen Energiemengen kontinuierlich zur Verfügung zu stellen, was zu entsprechend komplizierten und undurchsichtigen Geschäftsmodellen führte. So schreibt noch 1914 der Karlsruher Ingenieur A. Sander im »Polytechnischen Journal«, dass » die Gasindustrie den Kampf mit dem elektrischen Licht sehr wohl aufnehmen kann«. (http://dingler.culture.hu-berlin.de/article/pj329/ar329054, 23.10.2014)

Die gesamte Debatte um die Elektrifizierung Berlins lässt sich also über die Quellen im »Polytechnischen Journal« rekonstruieren, verstehen und anschaulich machen. Ziel unseres Projektes ist es, die Schnittstelle zwischen den Objekten im Museum und den Texten in der Datenbank technisch so zu realisieren, dass daraus ein »Blended Museum« (Harald Reiterer) entsteht (Klinkhammer, Daniel; Reiterer, Harald: Blended Museum – Perspektiven für eine vielfältige Besuchererfahrung. In: i-com 7/2 (2008), pp. 4–10.).

Eine der Strategien, die vom Vortrag exemplarisch vorgestellt werden, ist der Einsatz von »Licht-Würfel«. Dabei wird einem Objekt im Museum ein solcher Würfel zugeordnet, wobei im Museum an zentralen Stellen dem Besucher mehrere Würfel beispielsweise auf Stelen angeboten werden. Zusätzlich befindet sich im jeweiligen Raum ein multimediales Display. Aktiviert der Besucher nun einen der Würfel durch Veränderung seiner räumlichen Lage, werden die zugehörigen Informationen auf dem Display dargestellt. Der Würfel fungiert also als ein Token oder als Schnittstelle zwischen Objekt und Information. Entscheidend ist nun, dass im Display weder der Besucher noch das Objekt verschwinden, sondern über eine Kamera selbst Bestandteil der dargestellten Informationen werden. Beispielsweise kann der Besucher ein Objekt, das als Mixed Reality im Display erscheint, von allen Seiten betrachten, indem er den Würfel dreht. Den fünf Würfelseiten (eine dient der Stromversorgung) sind unterschiedliche Informationsebenen zugeordnet, so dass aus der Manipulation des Token eine Geschichte des Objekts wird.

Während im hier skizzierten Vortrag eine von der Projektgruppe verwendete Technologie mit Fokus auf dem Museum ausführlich präsentiert werden soll, zielt die Postereinreichung »Dingler Dissemination« (M. Hug et al.) v.a. darauf, einen Einblick in die zu Grunde liegenden Daten zu geben.

Filmbild, Filmschnitt, Filmstil – die Quantifizierung und Visualisierung von filmischen Strukturen

Adelheid Heftberger

Die Digital Humanities sehen sich als ein „Big Tent", unter dem die unterschiedlichsten Disziplinen Platz haben sollen. Bei näherer Betrachtung gewinnt man allerdings den Eindruck, dass sich die Projekte in einem engen, textlastigen Fokus bewegen und unter anderem eine Kanon-Debatte befeuert (siehe die Diskussion rund um die polemischen Publikationen von Franco Moretti). Zur Filmwissenschaft und Filmgeschichte gibt es, abgesehen von einzelnen Initiativen in den USA, noch kaum Forschung im Rahmen der Digital Humanities. Die Gründe dafür sind nicht klar ersichtlich. Hängt es damit zusammen, dass Filmwissenschaft prinzipiell eine Randposition innerhalb der Geisteswissenschaften einnimmt? Oder gibt es Vorbehalte innerhalb der Disziplin gegenüber quantitativen Verfahren und ihrer Verwendung für die Interpretation? Ist schlicht die audiovisuelle Datenmenge zu dicht? In meinem Vortrag möchte ich unter anderem auf solche Fragen eingehen und zur Diskussion einladen.

Grundsätzlich soll mein Beitrag das Potential von interdisziplinärer Zusammenarbeit, z.B. der Informationsvisualisierung, für die filmwissenschaftliche und filmhistorische Forschung ausloten. Konkret soll dies am Beispiel des sowjetischen Avantgarde- und Dokumentarfilmemachers Dziga Vertov (1896 bis 1954) exemplifiziert werden. Für eine formale Untersuchung eignet sich Vertovs Werk deshalb besonders gut, weil der Regisseur seine politischen Botschaften über formale film-inhärente Verfahren, wie Einstellungslänge, Einstellungsgröße, Bildkomposition oder Bewegungsintensität, konzipierte und vermitteln wollte. Seine Filme entstanden in der noch jungen Sowjetunion der 1920er und 30er Jahre, einer Zeit des kreativen Aufbruchs, in der Künstler und Wissenschaftler keine Berührungsängste hatten und die formale Analyse von Kunstwerken als wesentliches Mittel zum Verständnis von Literatur und Film gesehen wurde. Der damals entwickelte russische Formalismus (auch Russische Formale Schule) hat mit den Digital Humanities allerdings viele Fragestellungen und vielleicht auch Methoden gemeinsam.

Die empirische Grundlage für meine weiterführenden Untersuchungen bildet ein Datenkorpus, das im Zuge eines interdisziplinären Projekts mit dem Titel „Digital Formalism" (Laufzeit von 2007 bis 2010) sowohl in manueller als auch computergestützter Annotation erarbeitet wurde. Jetzt würde man es ohne Zweifel im Rahmen der Digital Humanities einreichen, damals standen der interdisziplinäre Ansatz, vor allem die Zusammenarbeit von Geistes- und Naturwissenschaften im Vordergrund. Über einen Zeitraum von drei Jahren befasste sich eine Gruppe von Filmwissenschaftler/innen, Archivar/innen und Informatiker/innen gemeinsam mit der Datengewinnung und Interpretation von acht Langfilmen. Die jeweils unterschiedlichen, fachspezifischen Zielsetzungen und Interessen machten das Projekt zu einer Herausforderung für alle Beteiligten. Auch deshalb blieben 2010, am Ende des Projekts, vielleicht mehr Fragen als Erkenntnisse. Wie gelangt man denn nun von den quantitativen Daten tatsächlich zu Erkenntnissen und wie sollten diese formuliert werden? Welchen Beitrag kann die quantitative Analyse zur Filmgeschichte leisten? In welcher Weise können die Daten dargestellt werden, damit unterschiedliche Disziplinen damit arbeiten können?

Mein Vortrag soll mögliche Antworten und Herangehensweisen aufzeigen, wenn auch noch keine disziplinübergreifenden neuen Methoden und Techniken entwickelt wurden. Eine davon ist die Datenvisualisierung, die sich als ein hilfreiches Werkzeug herausstellte. Aufbauend auf der langen Tradition der visuellen Darstellungen von formalen Filmeigenschaften im Lauf der Filmproduktion, können nun mit Hilfe des technologischen Fortschritts neue Wege beschritten werden. Aufgrund der sprunghaften Entwicklung von Computerleistungen ist es möglich sogenannte reduktionslose Visualisierungen zu erstellen. Die Autorin hat in Zusammenarbeit mit dem Medientheoretiker Lev Manovich einige Beispiele davon anhand von Vertovs Filmen entwickelt und durchgeführt. Nur so steht die semantische Information unmittelbar für die Analyse zur Verfügung und kann direkte Anhalts- und Orientierungspunkte für relevante Teile im Film liefern. Um am Ende jedoch zu

sinnvollen Erkenntnissen über Vertov und seine Filme zu gelangen, muss die manuelle oder computergestützte Datenanalyse und Visualisierung mit filmhistorischem Wissen und Quellenstudium gepaart sein.

Nicht nur die Filmwissenschaft ist ein möglicher, wenn auch logischer geisteswissenschaftlicher Partner für die interdisziplinäre Zusammenarbeit. In diesem speziellen Fall bietet sich auch die Slawistik an und kann einen wertvollen Beitrag zum Verständnis des geschichtlichen und kulturellen Hintergrunds liefern. Sinnvolle Kooperationen scheinen sich aktuell auch mit der Kognitionswissenschaft oder der Statistik ergeben. Die amerikanischen Filmwissenschaftler David Bordwell und Kristin Thompson sehen Forschungsbedarf und Potential für die Filmwissenschaft vor allem in Hinsicht auf die Untersuchung des Filmstils. Darunter versteht man, laut Bordwell, im engsten Sinn den systematischen und signifikanten Einsatz von filmischen Verfahren. Eine quantitative Analyse von Filmen kann dafür die Grundlage erst liefern, wie es z.B. der Filmwissenschaftler Barry Salt bereits begonnen hat. Auch die Psychologie kann interessante Beiträge liefern. So führte zum Beispiel James Cutting interessante Untersuchungen an Hollywoodfilmen durch, die unter anderem zeigen, wie die formale Bauart eines Filmes die Filmrezeption lenkt und umgekehrt – wie also Filmproduktion und das Publikum im Wechselspiel stehen. Die Erkenntnisse der australischen Tänzerin und Cutterin Karen Pearlman zum filmischen Rhythmus ergänzen dessen ausschließlich formalen Zugang um wesentliche kinästhetische Überlegungen, was wiederum den Kreis zum Regisseur Vertov schließt.

Gerade Synergien zwischen den erwähnten Disziplinen können dazu beitragen, die quantitativen und visuellen Daten aus Filmwerken im Hinblick auf filmhistorische und filmwissenschaftliche Erkenntnisse zu interpretieren. In meinem Vortrag möchte ich daher anhand von Beispielen aus dem Werk Dziga Vertovs zeigen, wie quantitative Filmanalyse, Visualisierung und kulturhistorische Interpretation zusammenwirken können und in welche Richtung die Forschung dabei gehen könnte.

Dr. **Adelheid Heftberger** ist wissenschaftliche Mitarbeiterin und Archivarin im Österreichischen Filmmuseum in Wien. Studium der Slawistik und Vergleichenden Literaturwissenschaft, von 2007 bis 2010 Mitarbeiterin im interdisziplinären Forschungsprojekt „Digital Formalism“. Dissertation zum Thema Visualisierung in der Filmwissenschaft (am Beispiel des russischen Regisseurs Dziga Vertov).

Abstract

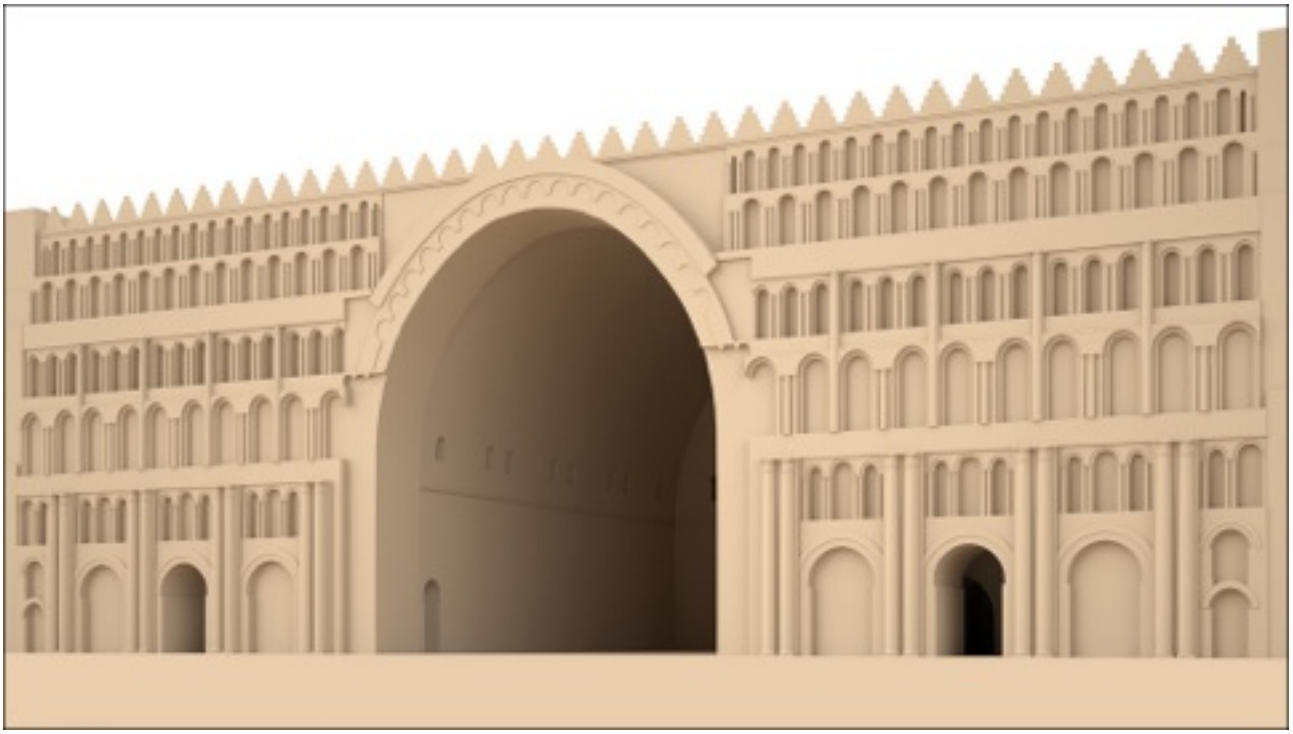
Die Bedeutung architektonischer Gestaltung in der visuellen Vermittlung wissenschaftlicher Unschärfe am Beispiel von Ktesiphon und weiteren archäologischen Stätten.

(Dominik Lengyel, Catherine Toulouse)

Forschungskontext

Das interdisziplinär angelegte, von der Deutschen Forschungsgemeinschaft (DFG) geförderte Excellence Cluster TOPOI (The Formation and Transformation of Space and Knowledge in Ancient Civilizations) beinhaltet im zweiten Fünfjahreszeitraum ein Forschungsprojekt, das eine Best Practice Methode für die Kooperation zwischen Wissenschaft und Museen anhand einer exemplarischen archäologischen Stätte erarbeiten soll. Ruinen im außermusealen Kontext, aber auch Funde im musealen Kontext sowie Ergänzungen in Form von Zeichnungen, Modellen in unterschiedlichen Maßstäben vom städtebaulichen Überblick bis zum Nachbau im Originalmaßstab, aber auch virtuelle 3D-Rekonstruktionen, prägen die Perzeption und damit das Wissen über die historischen Orte. Während TOPOI die Formation und Transformation von Raum und Wissen in der Antike behandelt, untersucht dieses Forschungsprojekt insbesondere die Perzeption und Repräsentation im Museum.

Untersuchungsgegenstand ist die spätantike sasanidisch Residenz Ktesiphon im heutigen Irak, von dem bis auf den aufragenden Palast mit dem bekannten Bogen des Kisra die Grabungen heute durch landwirtschaftliche Nutzung wieder verschüttet sind. Von den großen Stadtbereichen sind daher nur die Grabungsdokumentation vorhanden. Das bedeutendste Zeugnis der Stadt, die bogenförmige Audienzhalle (Iwan) des Palastes gilt als Prototyp dieser Bauform, ein ikonisches Zeichen, das im späteren Verlauf ein Merkmal vor allem islamischer Architektur wurde. Größe und Bedeutung dieses Ur-Iwan sollen im Museum vemittelt werden.



Das im Projekt federführende Museum für Islamische Kunst im Pegamon Museum Berlin der Staatlichen Museen zu Berlin (SMB) steuert sein umfangreiches Grabungskonvolut aus seinen Archiven bei, das aus Fragmenten der Palastanlage, aber auch zahlreicher über das Areal verstreuter Gebäude besteht. Von Seiten der Wissenschaft sind die Bereiche Restaurierung, Museumsforschung und Visualisierung beteiligt. Die Restaurierung untersucht nicht nur die Funde selbst und bereitet sie für die museale Präsentation auf, sondern vermittelt auch die einhundertjährige Restaurierungsgeschichte. Die Museumsforschung begleitet und evaluiert den Prozess, um die Ziele der Projektbeteiligten laufend aufeinander abzustimmen und zu optimieren. Das Ergebnis wird eine Ausstellung im Museum für Islamische Kunst in den Jahren 2015/2016 sein, bevor es im Jahr 2020 in die dann neu eröffnete Dauerausstellung einfließen wird.

Visualisierung

Die Aufgabe der Visualisierung ist die visuelle Vermittlung des historischen Topos auch durch Rekontextualisierung der Funde in Bildern, Filmen und dreidimensional gedruckten, haptischen Modellen im Rapid Prototyping Verfahren mithilfe einer virtuellen 3D-Rekonstruktion. Die besondere Herausforderung ist hierbei die äußerst weit verstreute Fundsituation. Das als städtisches Areal bezeichnete, in seinen Konturen äußerst unscharfe Gebiet umschließt große nicht ausgegrabene Bereiche, disjunkte Zeitphasen aus einer großen Zeitspanne und unterschiedliche Gebäudetypen, von denen einige in ihrer Deutung unbestimmt geblieben sind. Im Wissen über Ktesiphon, dessen Einheit als Stadt mehr aus ihrer räumlichen Gegenüberstellung mit der Stadt Seleukia, die auf der anderen Seite des Tigris liegt, denn aus ihrer inneren Struktur erwachsen ist, überwiegen deutlich die Unschärfen. Eine visuelle Vermittlung hat daher die Aufgabe, erstens trotz dieser Unschärfen ein Bild von Ktesiphon, das dem aktuellen Stand der Forschung entspricht, und zweitens gerade diese Unschärfe im Wissen als integralen Bestandteil wissenschaftlicher Forschung zu vermitteln. Dieser Aspekt der ausdrücklichen Vermittlung eines Wissens, das nicht nur unvollständig ist, sondern dessen Unschärfe auch in gleichwertigen Widersprüchen liegen kann, ist in einem so ausgeprägt fragmentarischen Topos wie Ktesiphon entscheidend.

Darstellung von Unschärfe

Die Darstellung von Unschärfe, also die visuelle Vermittlung unscharfen Wissens, hier in Bauforschung und Archäologie, besteht in Hinblick auf ihre Vermittlerrolle aus den beiden gleich bedeutenden Aspekten der Treue der Visualisierung zum Stand des Wissens unter besonderer Beachtung der Unschärfe, also vor allem der Hypothesen mitsamt ihren Unsicherheiten und Widersprüchen, sowie der expliziten Ablesbarkeit ebenjener Unschärfe. Die visuelle Übersetzung dieser komplexen Inhalte ist damit keine 3D-Rekonstruktion im herkömmlichen Sinne, also definierter historischer Zustände, sondern vielmehr eine Erschaffung virtueller visueller Repräsentationen abstrakter Inhalte (Lengyel, 2011a).

Durch den räumlichen Charakter der bauforscherischen und archäologischen Hypothesen entsteht zwar ein dreidimensionales CAD-Modell als Rohling, ähnlichem einem Werkstück zur weiteren Bearbeitung. Dieses wird aber, sofern es nicht als dreidimensional gedrucktes Modell materialisiert wird, erst durch die Projektion, die Visualisierung zu einem Medium der Vermittlung. Das heißt, erst in dem Zusammenwirken zwischen dem Modell und der sich mit diesem auseinander setzenden virtuellen Fotografie entsteht eine

Visualisierung, die in der Lage ist, dem Anspruch der Forschungstreue einschließlich sämtlicher Unschärfen gerecht zu werden.

Virtuelles Modell und virtuelle Fotografie verstehen sich, auch wenn sie sich aufeinander beziehen und voneinander abhängen, als digitalen Analogien unterschiedlicher tradierter Darstellungsmethoden, dem Modellbau im architektonischen Entwurfsprozess und der Architekturfotografie (Lengyel, 2011b).

Gestaltung

Traditioneller Modellbau im architektonischen Entwurfsprozess belässt vieles bewusst in einem wenig definierten Stadium, um spätere Entscheidungen nicht vorweg zu nehmen. Trotz dieser Unbestimmtheit im Detail werden konkrete Aussagen zu Volumen, Präsenz, Raumbeziehung oder auch Bautypologie getroffen. Mit CAD modelliert werden dabei also keine Gebäude im herkömmlichen Sinn, sondern Volumina, deren abstrakte Form erst im Kontext eine Bedeutung erhält. Die erzeugten Formen ergeben sich dabei nicht von allein aus den Vorgaben, sie können ebenso die geometrische Vereinfachung bestehender Gebäude wie geplanter Gebäude sein. Gerade die bewusste Gestaltung der formalen Ähnlichkeit existierender und geplanter Gebäude versetzt diese in die Lage, zueinander in Dialog zu treten. Diesem Verhältnis zwischen der Gewissheit des Bestandes und der Offentheit des Entwurfs entspricht die Unschärfe in einer archäologischen Hypothese. Es wird daher in beiden Fällen keine reale Architektur (re-) konstruiert, sondern es werden völlig neue räumliche Repräsentationen entworfen. Die Unbestimmheit der formalen Ausprägung, durch die sich dabei die Aussage weg vom Individuum hin zum Gebäudetypus verschiebt, erzeugt eine diagrammatische Architektur, die ähnlich einem Diagramm in einer übergeordneten Bedeutungsebene Prinzipien vermittelt. Hierdurch wird deutlich, dass erst durch Gestaltung eine Form des Diagramms gefunden wird, die in der Lage ist, die abstrakten Inhalte der Hypothesen zutreffend visuell wiederzugeben (Lengyel, 2013)

Architekturfotografie ist vor allem durch ihren hohen Anteil an Dokumentarischem charakterisiert. Ihr Anspruch an die Wiedergabe natürlicher Rezeption schließt allerdings nicht den unvermeidbaren interpretatorischen Gehalt einer Fotografie aus. In unterschiedlicher Gewichtung folgen eine Reihe von Aspekten fotografischer Interpretation, sei es perspektivischer Standpunkt (in der virtuellen Fotografie auch die bloße Blickrichtung der im realen Raum nicht möglichen Parallelprojektion), Perspektivität als die Wahrnehmung stark beeinflussende Aufweitung oder Stauchung des Raumes, Bildausschnitt als Selektion sowie Beleuchtung und Belichtung als Mittel der Plastizität, der Kompensation des räumlichen Sehens und der nicht zu unterschätzenden emotionalen Wirkung des Bildes. Alle genannten Aspekte der Fotografie beeinflussen in ebenjenem Maße, wie Architekturfotografie und Fotografie überhaupt als eigenständige Gestaltungsdisziplinen die Wahrnehmung beeinflussen, die Rezeption des Dargestellten explizit aber auch subtil. Im Kontext der Vermittlung unscharfen Wissens gilt es daher, so weit wie möglich dokumentarisch vorzugehen, bevor die interpretatorischen Mittel dazu eingesetzt werden, die Vermittlung der Hypothesen zu unterstützen, dass heißt ein trotz zum Teil hohem Abstraktionsgrad plausibles, immersives und emotional überzeugendes Bild zu erschaffen, das eine sachgerechte und historisch relevante Interpretation des Dargestellten erlaubt (Lengyel, 2011c)

Fazit

Der Vortrag soll die herausgehobene Bedeutung der Gestaltung bei der visuellen Vermittlung des unscharfen Wissens anhand des aktuellen Forschungsstandes von und der spezifischen Problematik im Wissen um Ktesiphon beleuchten, und anhand weiterer Vergleichsprojekte grundsätzliche Aspekte bei der Gestaltung der Visualisierung unscharfen Wissens mit virtuellen 3D-Rekonstruktionen erläutern, insbesondere der virtuellen 3D-Rekonstruktion von Pergamon, das im Rahmen des ersten Fünfjahreszeitraums von TOPOI, aber auch des Berliner Skulpturennetzwerks (Lengyel, 2011d) gemeinsam mit der Abteilung Istanbul des Deutschen Archäologischen Instituts entstanden ist und laufend weiterentwickelt wird (Laufer, 2011). Exemplarisch sollen Referenzen zu tradierten und aktuellen Mitteln der bildenden Kunst gezeigt werden, derer sich der virtuelle Modellbau und die virtuelle Fotografie bedienen, und von denen einige Aspekte unmittelbar in die Gestaltung eingeflossen sind, wodurch die Perzeption der Vermittlung sich auch auf kulturspezifische Seherfahrungen stützt.

Literatur

Laufer, Eric / Lengyel, Dominik / Pirson, Felix / Stappmanns, Verena / Toulouse, Catherine (2011). Die Wiederentstehung Pergamons als virtuelles Stadtmodell. In: Scholl, A., Kästner, V., & Grüssinger, R. (Hrsg.): Antikensammlung Staatliche Museen Berlin. Pergamon. Panorama der antiken Metropole. Petersberg: Verlag Imhof. S. 82–86.

Lengyel, Dominik / Toulouse, Catherine (2011a). Die Gestaltung der Vision Naga - Designing Naga's Vision. In: K. Kröper, S. Schoske, & D. Wildung (Hrsg.): Königsstadt Naga - Naga, Royal City. Grabungen in der Wüste des Sudan - Excavations in the Desert of the Sudan. München - Berlin, S. 163-175, Abb. 210-212, Naga-Projekt Berlin - Staatliches Museum Ägyptischer Kunst München.

Lengyel, Dominik / Toulouse, Catherine (2011b). Darstellung von unscharfem Wissen in der Rekonstruktion historischer Bauten. In K. Heine, K. Rheidt, F. Henze & A. Riedel (Hrsg.), Von Handaufmaß bis High Tech III. 3D in der historischen Bauforschung (S. 182-186). Darmstadt/Mainz: Philipp von Zabern.

Lengyel, Dominik / Schock-Werner, Barbara / Toulouse, Catherine (2011c). Die Bauphasen des Kölner Doms und seiner Vorgängerbauten. Cologne Cathedral and Preceding Buildings. Köln: Verlag Kölner Dom e.V.

Lengyel, Dominik / Toulouse, Catherine (2011d). Ein Stadtmodell von Pergamon - Unschärfe als Methode für Darstellung und Rekonstruktion antiker Architektur. in: Lars Petersen, Ralf von den Hoff (Hg.), Skulpturen in Pergamon – Gymnasion, Heiligtum, Palast, Freiburg, S. 22–26, Abb. 22, 25, 29, 32. Archäologische Sammlung der Albert-Ludwigs-Universität Freiburg, ISBN 978-3-86206-088-7

Lengyel, Dominik / Toulouse, Catherine (2013). Die Bauphasen des Kölner Domes und seiner Vorgängerbauten: Gestaltung zwischen Architektur und Diagrammatik. In: D. Boschung & J. Jachman (Hrsg.): Diagrammatik der Architektur, Tagungsband Internationales Kolleg Morphomata der Universität zu Köln. Paderborn: Verlag Wilhelm Fink. S. 327-352

„Befund“ und „Deutung“ in Beethovens Schreibprozessen

Aus philologischer Sicht stellen genetische Ausgaben eine besondere Herausforderung dar, da sie mit dem sogenannten *avant-texte*[1] Unfertiges in den Blick nehmen. Dies führt gerade im Bereich der Edition von Musik zu einem „terminologischen“ Problem, das zugleich die Deutung betrifft: Eine skizzierte Melodielinie ohne verbindlich notierte Tondauern ist eben keine Aneinanderreihung z.B. von Viertelnoten, sondern schlicht graphisch unvollständig niedergeschrieben. Dabei ist ein solches Notat keineswegs lücken- oder gar fehlerhaft – alle aus Sicht des Komponisten zu diesem Zeitpunkt relevanten Parameter sind ja tatsächlich bereits festgelegt. Dennoch erweist es sich als schwierig, derartige Notate angemessen zu beschreiben. Eine echte Brücke von Befund zu Deutung schlagen diese Ausgaben jedoch oft nicht, da auch bei Vorliegen einer Transkription die Zuordnung des vom Editor Gelesenen bzw. Interpretierten zu den zugrunde gelegten Graphemen durch mehrfache Überschreibungen, Löschungen, Restitutionen und Querbezüge oft kaum möglich ist. Selbst wenn es (scheinbar) gelingt, diese Zuordnung zwischen Transkription und originaler Handschrift zu klären, fehlt es dem Leser der Ausgabe in der Regel an Leseerfahrung im Umgang mit Beethovens Schreibgewohnheiten, so dass manch eindeutiger Sachverhalt schwer nachvollziehbar, Unsicheres hingegen durch die vermeintliche Bestimmtheit der Transkription unhinterfragt bleibt. Gleichwohl können verbale Erläuterungen seitens des Editors dieses Problem nur bedingt lösen, da die Umständlichkeit der Adressierung einzelner Aktionen des Schreibers mit steigender graphischer Komplexität ebenfalls zunimmt.

Im Jahr 2011 stellte die Text Encoding Initiative (TEI) ein Modell zur Codierung von textgenetischen Prozessen zur Diskussion, welches seit 2007 von einer gesonderten Arbeitsgruppe entwickelt worden war. Das Grundkonzept dieses Modells ist es, neben der Codierung des Textes (mit den bisherigen Mitteln der TEI) eine weitere, parallele Codierung anzubieten, die sich ausschließlich am zu erfassenden Dokument orientiert und statt semantischer Einheiten wie Absätzen oder Verszeilen ausschließlich mit graphischen Einheiten wie Seiten und Textzeilen operiert. Diese beiden Codierungen, die sich nebeneinander im gleichen Dokument ablegen lassen, können dann über Querverweise in Bezug gesetzt werden, um etwa die zu einer Textschicht gehörenden Grapheme zu identifizieren. Damit wird es theoretisch möglich, die Einheiten „Dokument“ und „Text“ in einer digitalen textgenetischen Edition sauber zu differenzieren.

Allerdings gibt es bis heute teils sehr kritische Äußerungen und Diskussionen, in deren Kern immer wieder die Frage steht, ob statt einer parallelen Codierung nicht ein „eingebettetes“ Modell, bei dem also Text und Dokument zeitgleich adressiert werden, geeigneter sei, auch wenn dabei besondere Herangehensweisen zur Vermeidung von überlappenden Hierarchien nötig seien. Der wesentliche Grund für diese Diskussionen dürfte die mitunter schwierige Abgrenzung zwischen Dokument und Text sein, zumal einerseits das Modell der TEI in beiden Bereichen teils auf die gleichen Elemente zurückgreift, andererseits das Verhältnis zu den prozessbeschreibenden Elementen innerhalb der Textcodierung (<tei:del> für Streichungen etc.) nicht abschließend geklärt zu sein scheint. Entsprechend finden sich Positionen, die je nach Hintergrund des jeweiligen Wissenschaftlers eine beide Aspekte einschließende Codierung entweder aus der Perspektive des Textes (also basierend auf dem bisherigen TEI-Modell) oder des Dokuments (und damit des neuen textgenetischen Modells) fordern[2].

Das Projekt „Beethovens Werkstatt“ hat sich zum Ziel gesetzt, die hochkomplexe Dynamik kompositorischer Prozesse im Œuvre Beethovens zu erforschen und in exemplarischen digitalen Editionen zu dokumentieren. Dazu sollen nach Möglichkeit die kompositorischen Schreibprozesse sowohl innerhalb einzelner Autographen als auch in der Abfolge aufeinander beziehbarer Werkstattmanuskripte rekonstruiert werden, um so Aufschluss über Beethovens kompositorisches Denken, Handeln und Entscheiden zu erlangen. Dabei steht das Projekt in der

[1] Zum Begriff vgl. Almuth Grésillon: *Literarische Handschriften. Einführung in die „critique génétique“*, Bern 1999, S. 139f.

[2] Ein gutes Beispiel einer solchen Diskussion mit allen geschilderten Positionen findet sich im Februar 2014 auf der Mailing-Liste der TEI unter dem Titel „Embedded transcription and text structure“ (archiviert unter http://listserv.brown.edu/archives/cgi-bin/wa?A2=tei-l;29c06075.1402).

Tradition der gedruckten Editionen von Beethovens Skizzen, sieht sich aber gleichzeitig innerhalb der Digital Humanities als möglicher Nutzer einer abstrahierten Umsetzung des TEI-Modells für genetische Editionen. Beide Bezugspunkte erweisen sich allerdings als brüchig und nur bedingt anschlussfähig. Während die gedruckten Ausgaben aufgrund ihrer abweichenden medialen Rahmenbedingungen keine unmittelbar ins Digitale übertragbaren Konzepte bereitstellen, gelten ähnliche Einschränkungen für das auf literarische Texte ausgelegte Modell der TEI: Durch das im Falle von „Beethovens Werkstatt“ andere Schriftmedium – Notenschrift vs. Texte, die man in der Informatik als mehr oder weniger „ASCII-based“ bezeichnen würde[3] – verschieben sich einige der zu behandelnden Probleme vom Sonder- zum Regelfall. Dazu zählt in besonderer Weise die sinnentstellende Unvollständigkeit des graphischen Einzelzeichens – selbst in Skizzen werden in der Literaturwissenschaft i.d.R. vollständige Buchstaben (als kleinste sinnvolle Einheit) notiert. In der Musik gilt diese Voraussetzung (wie beschrieben) so nicht: Ein Melodieverlauf kann durch Notenköpfe skizziert werden, ohne dass diese bereits eine rhythmische Abfolge festlegen. Dabei wird die mangelnde Festlegung der rhythmischen Komponente am einzelnen Notenzeichen nicht zwingend ersichtlich. Eine Skizze ist also ggf. nicht als solche zu erkennen und verleitet damit zu Deutungen, die so vom Befund nicht gedeckt werden.

Eine Codierung von Musik, etwa mit den Mitteln der Music Encoding Initiative (MEI), operiert ohnehin in anderer Weise als etwa eine Codierung mit TEI. Während in TEI ein Text, dessen grundlegender Zeichenvorrat mit ASCII-Zeichen erfasst, durch XML-Elemente angereichert und im vom englischen „Markup“ kommenden Sinne deutend ausgezeichnet wird, gibt es für die Musik keine „direkte“ Beziehung zwischen einer Note und ihrer digitalen Entsprechung, wie es im Textbereich eine Beziehung zwischen Buchstabe und ASCII-Codepoint gibt. Schon für diese kleinste Einheit muss also ein „beschreibendes Markup“ gewählt werden, bei dem sämtliche relevanten Parameter der Note explizit erfasst werden[4]. Es dürfte offensichtlich sein, dass diese Art der Codierung der beschriebenen graphischen Offenheit gerade skizzenhafter Musiknotation sehr deutlich im Wege steht und alle bestehenden Möglichkeiten zum Umgang mit Mehrdeutigkeit zwingend zu sehr umfangreichen und komplexen Codierungen führen müssen. Eine parallele Codierung von Text und Dokument, wie sie das Modell der TEI vorsieht, erscheint daher aus musikwissenschaftlicher Perspektive als wenig praktikabel.

Stattdessen wurde für das Projekt ein Codierungsmodell entworfen, das eine sehr viel klarere Unterscheidung zwischen Dokument und Text erlaubt und auf diese Weise eine Diskussion, wie sie auf Seiten der TEI zu beobachten ist, vermeiden hilft. Da „Beethovens Werkstatt“ in Teilen am Beethoven-Haus Bonn angesiedelt ist, stehen dem Projekt für einen Großteil der behandelten Werke hochauflösende Faksimiles zur Verfügung. In diesen werden nun sämtliche Eintragungen auf der Ebene der einzelnen Strichführungen nachgezeichnet und als SVG-Dateien[5] vorgehalten. Die einzelnen Textschichten (soweit erkennbar) werden dann in einer sauberen Transkription, d.h. in heutiger Notenschrift ohne den Versuch einer diplomatischen Wiedergabe, codiert, wobei editorische Zusätze ebenso wie Mehrdeutigkeiten und offene Interpretationen selbstverständlich also solche gekennzeichnet werden. Jedes einzelne Zeichen wird dann mit den ihm zugehörigen Eintragungen Beethovens im Faksimile verknüpft, so dass der Benutzer der Ausgabe unmittelbar durch einen Klick auf eine transkribierte Note den dieser Deutung zugrunde liegenden Befund im Faksimile durch eine (ausblendbare) farbliche Hervorhebung nachvollziehen kann. Sämtliche Prozess-Informationen zu Streichungen, Ersetzungen, Restitutionen etc. werden dabei als Funktion des Textes bzw. der Textentwicklung erfasst, wobei die einzelnen Eingriffe wie Streichungen durchaus auch den graphischen Manifestationen im Faksimile zugeordnet werden.

Dieses Modell zeichnet sich dadurch aus, dass es auf Seiten des Dokuments (also im Faksimile) weitgehend ohne

[3] ASCII (American Standard Code for Information Interchange) ist eine Auflistung von 128 Zeichen, die hier als Grundlage der Belegung einer Computertastatur verstanden und gleichzeitig als stellvertretend für umfangreichere Zeichencodierungen wie etwa UTF-8 als technische Festlegung der Beziehung von (Einzel-)Zeichen und Bedeutung interpretiert wird.

[4] Die Codierung einer Viertelnote „C“ in der vierten Oktave wäre in MEI etwa: <note pname=“c“ oct=“4“ dur=“4“/>.

[5] Scalable Vector Graphics (SVG) ist ein Bildformat, welches beliebig skaliert werden kann und es erlaubt, die so markierten Teile jederzeit nachträglich beliebig einzufärben oder hervorzuheben.

Interpretation auskommt[6], während alle editorischen Deutungen auf Seiten des Textes gebündelt werden. Durch den Einsatz des gegenüber den Inhalten agnostischen Datenformats SVG bietet es dabei eine klarere Trennung von „Befund“ und „Deutung“ bzw. „Dokument“ und „Text“ als die parallele Codierung im Modell der TEI, die letztlich auf beiden Seiten auf die gleichen Grundbegriffe zurückgreift. Gleichzeitig werden auf diese Weise (und in Abgrenzung zu den herkömmlichen gedruckten Ausgaben) jedoch Dokument und Text(schichten) so eng verzahnt, dass der Benutzer der Ausgabe jederzeit nachvollziehen kann, worauf sich die Deutungen des Editors beziehen.

Im Vortrag soll das Modell anhand ausgewählter Beispiele ausführlicher vorgestellt und diskutiert werden.

[6] Die Identifizierung bzw. Differenzierung einzelner Schreibvorgänge kann bei mehrfacher Überschreibung bereits Interpretationen erforderlich machen. Dies ist aber eher als grundsätzliches editorisches Problem zu verstehen und ist kein Charakteristikum des Modells.

Modellierung von Annotationen in der digitalen Musik- und Medienedition

BMBF-Projekt “Zentrum Musik – Edition – Medien”
Anna Maria Komprecht (komprecht@edirom.de), Musikwissenschaftliches Seminar Detmold/Paderborn
Andreas Oberhoff (oberhoff@upb.de), Heinz Nixdorf Institut, Universität Paderborn

Das vom Bundesministerium für Bildung und Forschung geförderte *Zentrum Musik – Edition – Medien* vereint die Kompetenzen von Forscherinnen und Forschern musikwissenschaftlicher, medienwissenschaftlicher und verschiedener informatischer Disziplinen und wird in den kommenden Jahren die Forschungsarbeit rund um die digitale Musik- und Medienedition sowie den Umgang mit musikalischen und nicht-textuellen Objekten in einer digitalen Forschungsumgebung vorantreiben. Aufbauend auf den Vorarbeiten mehrerer Projektgruppen (beispielsweise Software zur Umsetzung digitaler Musikeditionen, virtuelle Wissensräume und musikalische Mensch-Computer-Interaktion) ist das Ziel, neue Konzepte, Modelle und softwaretechnische Umsetzungen in Begleitung von qualitativen sowie quantitaven Nutzerstudien zu erarbeiten. Hierbei spielen die Modellierung von mehrschichtigen Annotationsmodellen und die interaktive Eingabe, Darstellung und Gestaltung von Anmerkungen und somit der Umgang mit einer feingliedrigen Objektstruktur in der digitalen Musik- und Medienedition eine große Rolle. Das vorliegende Abstract beschreibt erste Überlegungen zur konzeptionellen und technischen Umsetzung dieser Modelle.

Ausgehend von dem von Frans Wiering entworfenen “mehrdimensionalen Modell” in digitalen Musikeditionen werden Annotationen innerhalb diesem als eigenständige Ebene gesehen und im digitalen Archiv neben allen anderen Objektarten der Edition abgespeichert, um im zweidimensionalen Raum dargestellt werden zu können (vgl. Wiering, 2009, S. 28). Die Öffnung des Editionsprozesses durch Anreicherung der Texte und Dokumente mit editorischen Annotationen und das Einbeziehen des Nutzers, dem man im digitalen Raum die Möglichkeit geben kann, eigene Anmerkungen und Kommentare zu verfassen, stellt vor allem innerhalb der Objektstruktur von Musiknotationen eine neue Herausforderung dar. Da die Arbeit an einer Edition in der digitalen Umgebung nicht mehr durch abgeschlossene Prozesse charakterisiert wird, sondern vielmehr eine permanente Aneinanderreihung unterschiedlicher Aktionen unter Einbezug verschiedener Editoren- bzw. Nutzersichten ist, muss innerhalb der verschiedenen Medienobjekte in der Musik- und Medienedition ein spezifisches Augenmerk auf den Gegenstand der Annotation geworfen werden. Aufbauend auf den Arbeiten verschiedener digitaler Musikeditionen (vgl. Opera[1] (Verbindung Libretto – musikalischer Text), Reger-Werkausgabe[2] (umfangreiche Kontextinformationen) und Freischütz Digital[3] (Variantenmodell sowie erstmalige Verbindung von Text, Musik und Audio)), sollen im Kompetenzzentrum innovative Wege gefunden werden, Annotationen in einer digitalen Umgebung zu verfassen.

[1] vgl. http://www.opera.adwmainz.de/das-projekt.html, Abruf am 23. Oktober 2014
[2] vgl. http://www.max-reger-institut.de/de/rwa.php, Abruf am 23. Oktober 2014
[3] vgl. http://freischuetz-digital.de/, Abruf am 23. Oktober 2014

Aber was bedeutet das Annotieren von musikalischem Text und verschiedener Medien wie Bild, Audio und Video in einer derartigen Umgebung überhaupt und wie kann dieser Vorgang innerhalb der digitalen Musik- und Medienedition softwaretechnisch unterstützt werden? Die Erstellung von Kommentaren und Anmerkungen in der klassisch gedruckten Buchausgabe ist ein intuitiver Vorgang, den es auch im digitalen Medium zu unterstützen gilt. Annotationen können situations- und personenbedingt ganz unterschiedlich verfasst und gestaltet werden und somit in verschiedene Schichten und auch Sichten aufgeteilt sein. Vor allem muss hier aber die Objekstruktur innerhalb der Musiknotation untersucht und nach ihrer Sinnhaftigkeit sowie der Verortung der jeweiligen Referenzen hingehend zu den vielschichtigen Dimensionen erforscht werden. Dabei gilt es, die Hinzufügung von kritischen oder kontextuellen Informationen oder Folgerungen ihrer Bedeutung entsprechend im Medium bzw. im Wissensraum am Objekt selbst zu verankern und die Dateneingabe und -darstellung nutzerfreundlich zu visualisieren und technisch intuitiv umzusetzen. Verschiedene Konzepte der Oberflächendarstellung sowie des technischen Unterbaus von Annotationen im digitalen Raum sind in unterschiedlichen Projekten bereits erforscht worden (vgl. Annotation Studio[4], annotator[5] bzw. AnnotateIt[6] oder co-ment[7]). Grundlage für verschiedene Annotationsmodelle in der digitalen Musik- und Medienedition sollen daher auch Methoden der Social bzw. Web Annotations sein und somit den Anspruch an Interoperabilität und die Orientierung am traditionellen Annotationsverhalten erfüllen.

Aufbauend auf Studien zu Materialität und Schriftlichkeit in der digitalen Musik- und Medienedition wird im Projekt versucht, das Verhältnis von materiellen und gedanklichen Einheiten von Annotationen zueinander zu durchleuchten. Ebenso ist angestrebt, die Prozesse des Annotationsvorgangs eines Editors, aber auch eines Benutzers der digitalen Edition zu erforschen, um diese optimal durch verschiedene Anpassungen und Werkzeuge zu unterstützen. Die Wahrnehmung von Objekten und deren semantische und strukturelle Beziehungen sowie der Einfluss digitaler (Re-)Präsentation spielen hierbei eine tragende Rolle. Daher muss ein möglichst kongruenter Handlungs- und Wahrnehmungsraum geschaffen werden, damit Objekte der Wahrnehmung gleichzeitig Objekte der Handlung werden. So können manipulierende Operationen, wie z. B. das Annotieren, direkt an den Objekten selbst durchgeführt werden, um Medienbrüche zu verhindern. Entscheidend ist dabei der Granularitätsgrad der Objekte für eine kleinstmögliche, aber dennoch semantisch sinnvolle Differenzierung, wobei keinesfalls die mögliche Aggregation von Objekten zu größeren Einheiten vernachlässigt werden darf.

Teil eines Forschungsdiskurses sind immer auch unterschiedliche Kontexte, in denen gearbeitet wird, um verschiedene Hypothesen zu untersuchen oder Forschungsgegenstände aus unterschiedlichen Blickwinkeln zu betrachten. Aus diesem Grund ist es neben einer angestrebten Verschmelzung von Handlungs- und Wahrnehmungsraum ebenfalls notwendig,

[4] vgl. http://www.annotationstudio.org/project/, Abruf am 23. Oktober 2014
[5] vgl. http://annotatorjs.org/, Abruf am 23. Oktober 2014
[6] vgl. http://annotateit.org/#about, Abruf am 23. Oktober 2014
[7] vgl. http://www.co-ment.com/, Abruf am 23. Oktober 2014

dass unterschiedliche Sichten auf die einzelnen Objekte und deren zum Teil komplexe Arrangements möglich sind. Diese Sichten müssen sowohl mit Blick auf die Darstellung von Objekten als auch der Funktionalität, mit denen sich diese bearbeiten lassen, anpassbar sein, um den entsprechenden Forschungskontext bzw. -prozess bestmöglich zu unterstützen.
Die spezielle Anpassung an die Medialität der Forschungsgegenstände innerhalb der Musik- und Medienedition erfordert die Modellierung eines Konzepts, welches einen ersten Ausgangspunkt darstellt, aber genug Flexibilität für weitere Entwicklungen und Konzepte bietet. Das Annotationsschema, die Codierungsformate bzw. -modelle, die die logische und technische Verbindung auf verschiedenen Ebenen darstellen und somit die Datenarchitektur vorgeben sowie Werkzeuge und Applikationen, aber auch das Design und die grafische Oberfläche bauen auf dieser ersten Konzeption auf. Die Einordnung der Annotationsreferenzen auf unterschiedlichen Ebenen sowie die Kontextualisierung innerhalb der verschiedenen Medien und der Detailgrad verschiedener Anmerkungen sind hierbei besonders zu beachten.

Die softwaretechnische Umsetzung setzt sich mit der Zusammenführung verschiedener, vorhandener Infrastruktur- und Softwarekomponenten auseinander und versucht unterschiedliche Vorarbeiten zu modularisieren und zu adaptieren. Ebenso sollen Neuentwicklungen, die durch Benutzerstudien in den jeweiligen Gruppen begleitet werden, realisiert werden. Um die stark benutzergenerierten und -interpretierten Metainformationen der Annotationsdaten besser maschinell prozessierbar zu machen, wird die Verwendung des Resource Description Framework (RDF) als Baustein des Semantic Web untersucht. Daten des Annotationsmodells müssen dazu in Form von Elementaraussagen in Triple Strores persistiert werden, um mit Hilfe der RDF Query Language (RDQL) für bestimmte, kontextabhängige Fragestellungen im Forschungsprozess abgefragt und aufbereitet werden zu können.

Ein weiteres Ziel bei der Modellierung wird es sein, vielfältige Typen der Eingabe von Annotationen, wie zum Beispiel am Bildschirm, aber auch Stift- und touch-basiert an verschiedenen mobilen Endgeräten sowie Multi-Touch-Tischen, zu ermöglichen. Die digital-physische Brücke spielt also ebenfalls eine große Rolle, die bereits im Prozess der Modellierung beachtet werden muss. Bearbeitungswerkzeuge unterschiedlicher Art und die Entwicklung einer Architektur, die eine flexible Anpassung an die sich ständig ändernden Anforderungen der Wissensarbeit garantiert, müssen entwickelt werden. Da das Projekt hauptsächlich auf die Software zur Arbeit an digitalen Musik- und Medieneditionen aufbaut und den Umgang mit diesen neu erforscht, wird die kontextspezifische Annotation mannigfaltiger Medienobjekte ebenso Gegenstand und Voraussetzung der Forschung sein wie die Gestaltung privater Datenräume für (mehrere) Nutzer der digitalen Edition und die Realisierung flexibler Interaktionsmöglichkeiten innerhalb der Anwendungen. Ebenso müssen die Annotationsmodelle auf die Harmonisierung der Repräsentationen auf unterschiedlichen Gerätebildschirmen untersucht werden.

Kern des Vortrags ist es, erste Konzepte für verschiedene Annotationsmodelle in der digitalen Musik- und Medienedition vorzustellen und zur Diskussion zu stellen. Es soll gezeigt werden, wie auf verschiedenen bereits entwickelten Konzepten und Methoden von Annotationen im

traditionellen Sinne, aber insbesondere auch im digitalen Umfeld aufgebaut werden kann und wie sich diese Prozesse durch digitale Techniken adäquat unterstützen lassen. Gerade bei nicht-textuellen Objekten im Allgemeinen und der komplexen Musiknotation im Speziellen besteht an dieser Stelle noch erheblicher Forschungsbedarf. Die hier vorgestellte Bestrebung des Projekts ist die Entwicklung innovativer Konzepte, die sowohl bei der Modellierung als auch bei der softwaretechnischen Realisierung von Annotationsprozessen im Bereich der digitalen Musik- und Medieneditionen die zweifellos vorhandenen Potenziale für Editoren und Benutzer ausschöpft.

Literatur

Wiering, Frans (2009): Digital Critical Editions of Music: A Multidimensional Model. In: Crawford, Tim/Gibson, Lorna (Hrsg.): Modern Methods for Musicology - Prospects, Proposals, and Realities. London, S. 23 - 46.

Referenzen

Annotation Studio. http://www.annotationstudio.org/project/, Abruf am 23. Oktober 2014.
annotator. http://annotatorjs.org/, Abruf am 23. Oktober 2014.
AnnotateIt. http://annotateit.org/#about, Abruf am 23. Oktober 2014.
co-ment. http://www.co-ment.com/, Abruf am 23. Oktober 2014.
Freischütz Digital. http://freischuetz-digital.de/, Abruf am 23. Oktober 2014
OPERA - Spektrum des europäischen Musiktheaters in Einzeleditionen. http://www.opera.adwmainz.de/das-projekt.html, Abruf am 23. Oktober 2014.
Reger-Werkausgabe. http://www.max-reger-institut.de/de/rwa.php, Abruf am 23. Oktober 2014.

Repräsentationen von Migration in musikwissenschaftlichen Datenbanken

Abstract zum Vortrag auf der DHd 2015 in Graz

von Torsten Roeder (Würzburg)

1. Musik und Migration in den Digital Humanities

Die Untersuchung von Migration und Mobilität hat sich in den letzten Jahren als ergiebiger Forschungsbereich in den Geisteswissenschaften etabliert. Dabei trat zutage, dass gerade Musiker aufgrund ihrer hochspezialisierten, teils akademisierten und zudem repräsentativen Tätigkeiten vielfach eine hohe Mobilität an den Tag legten und dass diese Migrationsbewegungen spätestens seit der frühen Neuzeit maßgeblich die Herausbildung einer gesamteuropäischen Musikkultur stimulierten.[1]

Längst erschließen auch die Digital Humanities dieses Forschungsfeld. Schon vor einigen Jahren erschienen die ersten musikwissenschaftlichen Personendatenbanken. Diese waren zunächst nicht mehr als digitale Abbilder klassischer Musiklexika, wie z. B. das *Österreichische Musiklexikon* (*ÖML*)[2] und *Eitner digital*[3]. Spätere Projekte setzten auf eine semantische Strukturierung und Vernetzung der Daten mithilfe von Verschlagwortung und Normdateien, wie z. B. das *Bayerische Musiker-Lexikon Online* (*BMLO*)[4]. Eine beachtliche Anzahl von digital orientierten Musikforschungsprojekten setzte den Schwerpunkt schließlich explizit auf das Thema der Musikermigration. Zu diesen zählen bislang *Musica Migrans*[5], *MUSICI*[6] und dessen Nachfolgeprojekt *Music Migrations in the Early Modern Age* (*MusMig*)[7], sowie das

1 Vgl. Ehrmann-Herfort 2013.
2 Österreichisches Musik-Lexikon. Online-Ausgabe, Wien: Österreichische Akademie der Wissenschaften, 2002–2013, <http://www.musiklexikon.ac.at/> (31.10.2014) = Oesterreichisches Musiklexikon, Wien: Verlag der Österreichischen Akademie der Wissenschaften, 2002–2006.
3 Eitner digital, Universität Zürich, <http://www.musik.uzh.ch/research/eitner-digital.html> (31.10.2014) = Robert Eitners Quellen-Lexikon, Leipzig: Breitkopf & Härtel, 1900–1904.
4 Bayerisches Musiker-Lexikon Online, hrsg. von Josef Focht, München: Ludwig-Maximilians-Universität, 2004–2014, <http://bmlo.de/> (31.10.2014).
5 Musica Migrans, Universität Leipzig, <http://www.musicamigrans.de/> (31.10.2014).
6 MUSICI. Musicisti europei a Venezia, Roma e Napoli (1650–1750): musica, identità delle nazioni e scambi culturali, <http://www.musici.eu/> (31.10.2014).
7 Music Migrations in the Early Modern Age. The Meeting of the European East, West and South, <http://musmig.eu/> (31.10.2014).

zeitgeschichtliche *Lexikon verfolgter Musiker und Musikerinnen der NS-Zeit* (*LexM*)[8]. Diese sollen im folgenden näher betrachtet werden.

Während musikwissenschaftlich-biographische Lexika zwangsläufig auf einzelne Personen fokussieren, bieten Datenbanken prinzipiell die Möglichkeit, Fragestellungen entlang alternativer Sichtachsen zu verfolgen. Der Blickwinkel des Forschungsfeldes muss in Datenbanken zudem nicht auf berühmte oder einflussreiche Persönlichkeiten beschränkt bleiben, sondern kann auch die „hinteren Reihen" der Musikschaffenden mit all ihren unterschiedlichen Schwerpunkten mit einbeziehen. Dies kann beispielsweise der Erforschung bestimmter Berufsgruppen dienlich sein. So ergibt eine kleine Recherche im *BMLO* (siehe Abbildung), dass ein guter Teil bayerischer Lautenbauer ihr Handwerk in Italien ausübte: Von den insgesamt 173 erfassten bayerischen Lautenmachern wirkten 72 in Italien. Dieses Ergebnis kann durch weitere Nachforschungen außerhalb der Datenbank geprüft und zu einer allgemeingültigen Aussage entwickelt werden.[9]

Thesenimpulse durch Datenbanken müssen sich jedoch nicht auf rein quantitative Aussagen beschränken, sondern können auch strukturelle Sachverhalte abbilden und sichtbar machen, wie es sich z. B. die MusMig-Datenbank zum Ziel gesetzt hat: Dort soll unter anderem gezeigt werden, inwieweit die Verlegung von Herrschersitzen Einfluss auf die Lebenswege von Musikern ausübte, indem diese mit- oder auch abwanderten.[10] Datenbanken dienen folglich nicht nur dazu, Migrationsmasse zu orten, sondern können möglicherweise zukünftig auch dazu herangezogen werden, Daten anhand bestimmter Kontexte und Strukturen aufzuzeigen und auslegbar zu machen. Die Musikmigrationsforschung eignet sich zur Erprobung der Möglichkeiten als anschauliches Beispiel.

2. Grundlinien der Datenmodellierung

Das Grundbedürfnis der Migrationsforschung lässt sich in folgenden grundlegenden Fragen zusammenfassen: *Wann* und *warum* gingen *wer wohin* um *was* zu tun? Als Hauptparameter der

8 Lexikon verfolgter Musiker und Musikerinnen der NS-Zeit, <http://www.lexm.uni-hamburg.de/> (31.10.2014).
9 Siehe z. B. Schulz 2010.
10 Vgl. Over/Roeder 2015.

Migration dienen somit die fünf Entitäten Individuum, Ort, Zeit, Motiv und Tätigkeit. Die jeweiligen Schnittpunkte zwischen diesen Entitäten liefern weiteren Fragen Ansatzpunkte für weitere Fragen: Wie verhielten sich beispielsweise die Kollegen eines jungen Leipziger Sängers, der nach Venedig emigrierte? Auf welches Umfeld stieß der Sänger im Ausland? Wie integrierte er sich in das dortige kulturelle Leben? Was brachte er dort ein, was nahm er von dort mit zu seiner nächsten Station? Diese Parameter zu modellieren stellt für die Entwicklung von Datenbanken eine Herausforderung dar, einerseits aufgrund der komplexen Fragen, auf die anhand der Daten eine Antwort gefunden werden soll, andererseits aber auch aufgrund der vielschichtigen Informationen, die in eine handhabbare Struktur gebracht werden müssen und die zudem ein spezielles Fachvokabular benötigen.

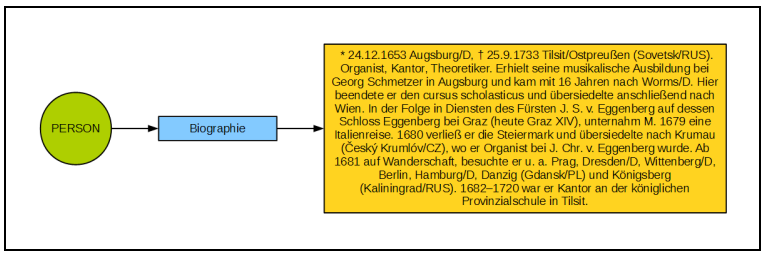
An die aktuell existierenden Datenbanken stellen sich demzufolge mehrere Fragen: Welche Möglichkeiten stellen die jeweiligen Datenbanken für Recherchen von sich aus bereit und welches Potenzial bieten sie für die Entwicklung von weiterführenden Fragestellungen? Welchen Einfluss haben die Datenstrukturen auf die möglichen Abfragen und Erkenntnisse? Zu erwarten ist, dass jede Datenbank-Oberfläche nur einen geringen Teil der tatsächlich geforderten Anwendungsmöglichkeiten antizipieren kann. Somit schließen sich weitere Fragen an: Stellen die Datenbanken ihre Ressourcen in einem semantisierten Format offen und wissenschaftlich nutzbar zur Verfügung? Kann dann aus den verschiedenen Datenbanken eine Datenmenge generiert werden, die trotz der Heterogenität der Daten in puncto Herkunft, Auswahl und semantischer Erschließung einen Mehrwert liefert? Ist es z. B. möglich, anhand einer Gegenüberstellung der verschiedenen Modellierungsansätze eine allgemeinere, synthetisierte Datenstruktur für Musiker-Migrationen zu entwerfen? Welche Erkenntnisse, die über die ursprünglichen Angebote der Datenbanken hinausgehen, wären dann zu erwarten? Und welche visuellen Methoden müssen angewendet werden, um diese lesbar zu machen?

3. Datenbanken im Vergleich

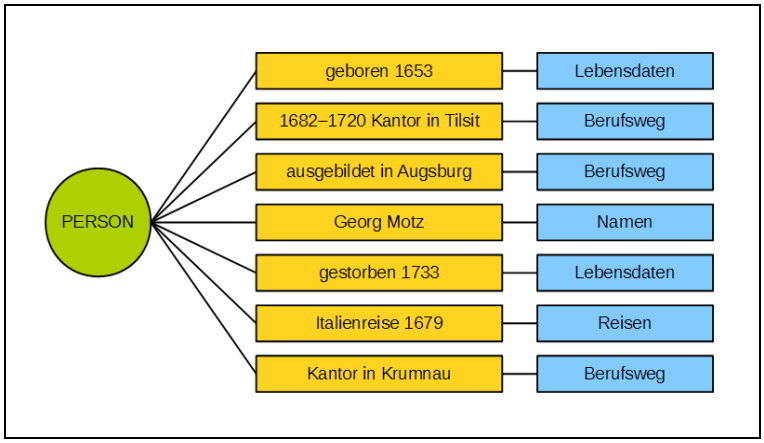
Biographische Datenmodelle sind zahlreich vorhanden: Ob man *TEI*[11], *BioDes*[12], *EAC-CPF*[13] oder ein eigenes Modell verwendet, obliegt der Entscheidung des jeweiligen Projektes, die stets auf der Grundlage von Vorbedingungen und Zielsetzungen gefällt wird. Wie aber sind Migrationen in den konkreten Datenbanken repräsentiert? Drei unterschiedliche Prinzipien seien hier kurz am Beispiel des weit bereisten Kirchenmusikers Georg Motz (1653–1733) vorgestellt.

11 Text Encoding Initiative (TEI): 13.3 Biographical and Prosopographical Data, <http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ND.html#NDPERS> (31.10.2014).
12 Biografisch Portaal van Nederland: BioDes, <http://www.biografischportaal.nl/about/biodes> (31.10.2014).
13 Staatsbibliothek zu Berlin: Encoded Archival Context for Corporate Bodies, Persons, and Families (EAC-CPF), <http://eac.staatsbibliothek-berlin.de/> (31.10.2014).

(1) Ein rein textuell ausgerichteter Ansatz, wie z. B. im *ÖML*, erlaubt eine Auswertung der Informationen im ganz klassischen Sinne. Migrationen können hier über das Lesen des Volltextes erschlossen werden. Die Informationen sind unterhalb der Textebene jedoch nicht weiter ausdifferenziert, so dass der systematischen, digital gestützten Auswertung wenig geholfen ist (siehe Abbildung).[14]



(2) Ein derzeit recht verbreiteter Ansatz ist es, alle biographischen Informationen zu einer Person in unabhängige Aussagen zu zerlegen (siehe Abbildung). Diese können dann wieder systematisch gruppiert werden, behalten dabei aber ihren Status als Einzelaussage. Derartige Ansätze wurden z. B. von *Topic Maps Lab*[15] und dem *Personendaten-Repositorium*[16] entwickelt, welche die Basis der Datenbanken *Musica Migrans* und *MUSICI* bilden. Dieses Modell verlässt den zwingenden Fokus auf die Person. Der Vorteil dieses Verfahrens besteht darin, dass hier spezifische Tätigkeiten an einem Ort zu einer bestimmten Zeit in einer einzelnen Aussage festgehalten werden können. Aus der Abfolge mehrerer Aussagen lässt sich folglich eine Migrationsbewegung erschließen. Parallel dazu kann z. B. verglichen werden, ob in anderen Biographien Migrationsbewegungen mit ähnlichen Zeit- und Ortsparametern vorliegen.



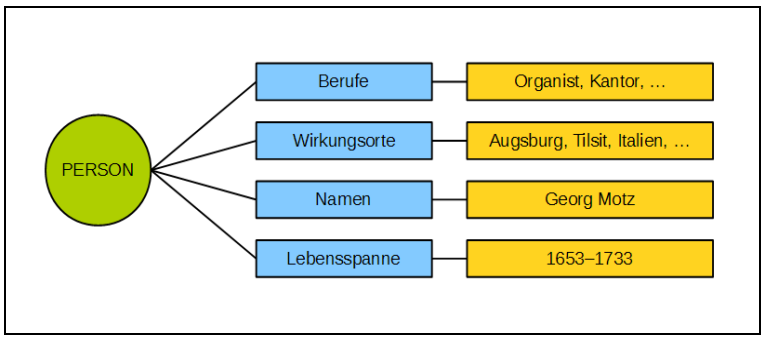
(3) Ein dritter Ansatz besteht darin, Personen eine beliebige Menge an Schlagwörtern zuzuordnen, die jeweils einer bestimmten Kategorie angehören (Namen, Lebensdaten, Wirkungsorte, Tätigkeiten etc.). Hier „zersplittert“ die Person in entitätische Facetten ihrer Biographie (siehe Abbildung). Dies

14 Man könnte hier zunächst eine semantische Auszeichnung von Ortsnamen versuchen und dann anhand dieser Namen – die im Text jedoch auch ganz anders gemeint sein könnten – Vermutungen über die Mobilität der jeweiligen Person aufstellen. Webservices für diese Zwecke existieren z. B. im Webservice-Angebot des Personendaten-Repositoriums, siehe <http://pdr.bbaw.de/> (31.10.2014).

15 „In einer Topic Map ist ein Topic […] der Repräsentant eines Aussagegegenstandes […] der realen Welt im Modell. An diesem Repräsentanten werden alle Informationen, die in dem Modell über den zugehörigen Aussagegegenstand repräsentiert werden sollen, geheftet.“ – Topic Maps Lab, Universität Leipzig, <http://www.topicmapslab.de/> (31.10.2014).

16 „Eine Person wird […] als die Menge aller Aussagen definiert, die zu ihr getroffen werden. […] Die kleinste Dateneinheit im Personendaten-Repositorium ist daher eine einzelne Aussage zu einer Person – im Datenmodell ‚Aspekt‘ genannt.“ – Personendaten-Repositorium, Berlin-Brandenburgische Akademie der Wissenschaften, <http://pdr.bbaw.de/> (31.10.2014). Vgl. auch Walkowski 2009.

reduziert den Sinnzusammenhang in der Biographie auf ein Minimum, wodurch gleichzeitig eine – informationstechnisch gesprochen – effiziente, da stark abstrahierte Datenform entsteht. Diese Herangehensweise verfolgt das *BMLO*. Migrationen können in diesem Ansatz weder qualitativ noch in ihrer Abfolge festgehalten werden, da der Zusammenhang zwischen Ort, Tätigkeit und Zeit aufgehoben ist. Jedoch ergeben sich bereits durch die Verschlagwortung geographische und professionelle Profile der jeweiligen Musiker. Mit diesem Verfahren können z. B. Herkunft und Mobilität von verschiedenen Berufsgruppen leicht gegenübergestellt oder Listen von an einem Ort wirkenden Personen zusammengestellt werden.



In dem besten Falle für den Nutzer gelingt es einer Ressource, die Vorteile mehrerer Ansätze miteinander zu verbinden. Dies ist der Fall im *LexM*, das sowohl eine strukturierte Daten als auch Volltextbiographien von ausgewählten Personen anbietet.[17] Dies gibt dem Forscher die Möglichkeit, sowohl systematisch als auch hermeneutisch zu arbeiten.

4. Zusammenführung

Dem Großteil der hier kurz vorgestellten Datenbanken ist gemein, dass die Such- und Ansichtsmodi ausschließlich auf Personen zentriert sind. Das oftmals vorhandene große Potenzial für andere Sichtachsen auf die Daten – vorstellbar sind statistische Übersichten nach Berufsgruppen und/oder Aufenthaltsorten, geographische Darstellungen, Zeitleisten mit Parallelvergleich von Musikeraufenthalten an verschiedenen Orten und vieles mehr – wird als Alternative nicht ausgeschöpft. Allein *MUSICI* unternimmt in dieser Richtung einen Versuch, indem ein Suchergebnis sowohl als Liste als auch als Balken- oder Kreisdiagramm, Karte oder Zeitleiste visualisiert werden kann.[18] Insofern zwingen die Datenbanken den Nutzer fast ausschließlich immer noch in die herkömmliche biographische Perspektive.

Der Versuch, aus den vorgestellten Datenbanken Information zu extrahieren und in ein einheitliches Format zu bringen, wird durch die Tatsache extrem erschwert, dass fast durchgehend die Unterstützung von allgemein anerkannten Austauschformaten, etwa RDF, und Ontologien wie CIDOC-CRM, nicht gegeben ist. Zur Zusammenführung der Informationen sollten dennoch solche allgemeine Formate und Referenzsysteme genutzt werden. Aufgrund der Heterogenität der Daten

17 Auch die *Deutsche Biographie* verfolgt ein solches zweigleisiges Verfahren, wobei hier methodisch von bestehenden Volltexten ausgegangen wird, die im Nachgang semantisch strukturiert werden. Siehe: Deutsche Biographie (ADB/NDB), Bayerische Staatsbibliothek, <http://www.deutsche-biographie.de/> (31.10.2014).

18 Vgl. Berti/Roeder 2014.

wird dabei ein minimalistischer Ansatz verfolgt: Die Informationen werden zu „Events“ mit den fünf oben genannten Basiseigenschaften (Individuum, Ort, Zeit, Motiv und Tätigkeit) zusammengefasst, welche dann mithilfe eines Geo- oder Netzwerkbrowsers visualisiert und analysiert werden können, um Erkenntnispotenziale für die Migrationsforschung auszuloten.

Literatur

<Berti/Roeder 2014> Michela Berti; Torsten Roeder: The “Musici” Database. An Interdisciplinary Cooperation, in: Musicisti europei a Venezia, Roma e Napoli (1650–1750). Europäische Musiker in Venedig, Rom und Neapel. Les musiciens européens à Venise, à Rome et à Naples (= Analecta musicologica 51). Hg. von Anne-Madeleine Goulet und Gesa zur Nieden, 2 Bde., Kassel u. a. 2014, im Erscheinen.

<Ehrmann-Herfort 2013> Migration und Identität. Wanderbewegungen und Kulturkontakte in der Musikgeschichte (= Analecta musicologica 49). Hg. von Sabine Ehrmann-Herfort. Kassel u. a. 2013.

<Over/Roeder 2015> Berthold Over; Torsten Roeder: MUSICI and MusMig. Continuities and Discontinuities, in: Zeitschrift für Digital Humanities 1, in Vorbereitung.

<Schulz 2010> Knut Schulz: Brot und Lautenspiel. Bayerische Handwerker in Italien vom 15. bis zum Beginn des 17. Jahrhunderts, in: Von Bayern nach Italien. Transalpiner Transfer in der Frühen Neuzeit, hrsg. von Alois Schmidt, München: Beck, 2010. S. 97–113.

<Walkowski 2009> Niels-Oliver Walkowski: Zur Problematik der Strukturierung und Abbildung von Personendaten in digitalen Systemen. <urn:nbn:de:kobv:b4-opus-9221>

WissensSpielRäume – Digital Humanities in der Theaterforschung

(Frau) Dr. Nic Leonhardt, LMU München, Theaterwissenschaft
e-mail: n.leonhardt@lmu.de

Theater zu erforschen, bedeutet Wissen über ein ephemeres Phänomen zu schaffen, das im Moment seiner Ausführung auch schon Geschichte ist. Was vom Theater bleibt, sind Spuren in Materialien unterschiedlicher Textur – Kritiken, Skripte, Zeichnungen, Bilder, Fotografien, Videos, Memoiren, Programmhefte etcetera –, aber nie Theater selbst. Diese methodologische Herausforderung treibt die Theaterwissenschaft seit ihrer Gründung um und macht sie besonders offen für immer neue Methoden und Perspektivwechsel.

Theater ist, mehr vielleicht als andere Künste und kulturelle Praktiken, inter-/transmedial, inter-/transkulturell, ist ein Produkt gemeinschaftlichen Erarbeitens und vereint verschiedene Ebenen von Zeitlichkeit in sich. Damit scheint seine Erforschung, so ließe es sich formulieren, „prädestiniert" zu sein für die Anwendung und kritische Prüfung von Methoden der digitalen Geisteswissenschaft, die nicht gedruckt, sondern digital, die nicht rein textlich, sondern multimedial sind, in der Anlage multi-referentiell und -relational sowie trans-disziplinär, und die ohne eine rege Nutzerschar – ein interaktives Publikum – nicht auskommen.
Die „zweite Welle" der Digital Humanities fußt auf einer Fülle an Materialien – also Texten, Zeitungen, Bildern, Archivbestände, die für die öffentliche oder auch teilöffentliche Nutzung zugänglich gemacht wurden. Auf die erste große Phase quantitativen Datenproduktion und -analyse folgt gegenwärtig eine mehr qualitative Annäherung, die von einer gezielten Adressierung von geisteswissenschaftlich relevanten Fragestellungen begleitet wird. In den vergangenen Jahren wurden auch innerhalb der Theaterwissenschaft digitale Projekte entworfen, die sich historisch oder zeitgenössisch mit der Herausforderung der Transitorik ihres Gegenstandes auseinandersetzen.

In meinem Vortrag „WissensSpielRäume – Digital Humanities in der Theaterforschung" möchte ich eine Zwischenbilanz über DH in der Erforschung von Theater ziehen: Auf der Grundlage des historisch angelegten DH-Projektes *Theatrescapes. Mapping Theatre Histories* an der LMU München (www.theatrescapes.theaterwissenschaft.uni-muenchen.de) und verwandter Projekte, möchte ich diskutieren, welches Surplus die Anwendung digitaler Technologien, relationaler Datenbanken und das digitale Verknüpfen und Teilen von Wissen zwischen Akteuren wie Wissenschaftlern, Praktikern und interessierten Laien, auf die Erforschung von Theater haben kann. *Theatrescapes* zielt mit Hilfe einer relationalen Datenbank auf die Untersuchung der Etablierung von Theaterhäusern seit 1850 (und in einem späteren Schritt der Mobilität von Theater-Professionals) auf globaler Ebene ab. Um dieser anspruchsvollen

Aufgabe gerecht zu werden, wird momentan eine Single-Page-Web-Application (Theatrescapes Research Tool) entwickelt, die einerseits zur Erfassung notwendiger Daten dient und andererseits deren Erforschung unter Zuhilfenahme geographischer sowie statistischer Auswertungsmethoden ermöglichen soll. Gleichzeitig ist die Arbeit an diesem Projekt nur kollaborativ und im Austausch mit anderen DH-Theaterprojekten, aber auch disziplinär verschiedener Projekte und Datenbestände (Kunstgeschichte, Literatur, Geographie etc..) möglich. Ausgehend von *Theatrescapes* möchte ich die Möglichkeiten der Wissensgenerierung und -erweiterung durch DH kritisch diskutieren und dabei einerseits das innovative Potential von DH markieren, wie andererseits die Probleme, Fallstricke und wissenschaftshistorischen Redundanzen aufzeigen, die ich in meiner Arbeit an dem DH-Theaterprojekt zu verhandeln habe. In meinen Ausführungen folge ich strukturell den vier Wissenspraktiken, wie sie Peter Burke in seiner jüngsten Studie *Die Explosion des Wissens. Von der Encyclopédie bis Wikipedia* (2014) formuliert: Sammeln, Analysieren, Verbreiten oder auch Teilen und Anwenden.

Durch Werkzeuge und Methoden digitaler Geisteswissenschaften lässt sich zwar der ‚Zauber' des Theatermomentes nicht wiederherstellen, aber, so meine These, es werden Praktiken der Wissensproduktion ermöglicht und/ oder vereinfacht, die für die Theaterwissenschaft wie für die Theaterpraxis äußerst Gewinn bringend sind und klug verbundene interdisziplinäre Spielräume entfalten.

Nic Leonhardt studierte Theaterwissenschaft und audiovisuelle Medien, Deutsche Philologie, Kunstgeschichte und Musikwissenschaft in Erlangen und Mainz, wo sie 2006 mit einer Arbeit über *Piktoral-Dramaturgie. Visuelle Kultur und Theater im 19. Jahrhundert* promovierte. Nach Tätigkeiten in Forschung und Lehre in Mainz, Köln, Leipzig, New York, Heidelberg ist sie seit 2010 Associate Director des DFG- Projektes "Global Theatre Histories" an der LMU München und seit 2013 Leiterin des DH-Projektes „Theatrescapes. Mapping Theatre Histories". Ihre Arbeitsschwerpunkte sind: Theater- und Mediengeschichte, Globalgeschichte, Urban Studies, Visual Culture und Digital Humanities.

Digitale Netzwerkanalyse dramatischer Texte

Peer Trilcke[1], Frank Fischer[2], Dario Kampkaspar[3]

[1] Georg-August-Universität Göttingen
[2] Göttingen Centre for Digital Humanities
[3] Herzog August Bibliothek Wolfenbüttel

1 Einleitung

Das Projekt 'Digitale Netzwerkanalyse dramatischer Texte' steht in der Tradition strukturanalytischer Ansätze in der Literaturwissenschaft (allgemein Titzmann 1977), die es einerseits im Sinne eines konsequent netzwerkanalytischen Relationismus (mit Rekurs auf die Social Network Analysis, siehe u. a. Wasserman/Faust 1998), andererseits unterstützt durch Verfahren der automatisierten Datenerhebung und -auswertung weiterentwickelt, um sie auf größere Textkorpora anzuwenden und so umfassende relationale Daten über Prozesse des literaturgeschichtlichen Strukturwandels gewinnen zu können.

Als theoretisches Fundament dient dabei eine netzwerkanalytische Konzeptualisierung dramatischer Interaktion (erste Ideen dazu prominent bei Moretti 2011; Kritik und literaturtheoretisch begründete Rekonzeptualisierung bei Trilcke 2013 – dort auch ein ausführlicher Forschungsüberblick), die – in Fortführung von Konzepten der dramatischen Konfiguration (Marcus 1973, Pfister 1977; problematisch hingegen, weil mit diffuser Konzeptualisierung: Pohlheim 1997) – zunächst bei einer rudimentären Operationalisierung ansetzt, nach der eine 'Interaktion' dann vorliegt, wenn zwei Figuren innerhalb einer durch die überlieferte Struktur des Textes vorgegebenen Subsegmentierungseinheit (in der Regel 'Szene' oder 'Auftritt') als Sprecher aufgeführt werden.

'Interaktion' wird in diesem Sinne – und zu Zwecken einer ersten Automatisierung – verstanden als 'szenische Kopräsenz zweier Sprecher'. Auf Grundlage der so definierten Relation werden im Rahmen des Projekts automatisiert netzwerkanalytische Daten erhoben, die sowohl global die 'Interaktions'-Netzwerke der Dramen (Density, Average Degree, Connectedness u. dergl.) als auch fokussiert einzelne Akteure charakterisieren (Degree sowie diverse weitere Centrality-Indices). Der erstellte Workflow ermöglicht auch die Datenerhebung auf Mesoebene (u. a. Identifizierung von Clustern) und beinhaltet darüber hinaus Visualisierungen der Netzwerkdaten, die wiederum zur Analyse des literaturgeschichtlichen Strukturwandels beitragen.

2 Wahl des Dramenkorpus

Für die automatisierte Analyse von Dramen war ein verlässliches und genügend großes Dramenkorpus vonnöten. Infrage kamen hier:

- Deutsches Textarchiv (DTA): 49 Dramen[1]
- Wikisource: 50 Dramen[2]
- Projekt Gutenberg-DE: 641 Dramen[3]
- Textgrid Repository: 690 Dramen[4]

[1] `http://www.deutschestextarchiv.de`
[2] `http://de.wikisource.org/wiki/Kategorie:Drama`
[3] `http://projekt.gutenberg.de`
[4] `http://www.textgridrep.de`

Das DTA hat zwar das qualifizierteste (TEI-)Markup, besteht aber bisher nur aus vergleichsweise wenigen Texten. Letzteres gilt auch für die Dramen im deutschsprachigen Zweig von Wikisource. Das Projekt Gutenberg-DE wiederum, das seit 2002 bei Spiegel Online gehostet wird, hat das Problem, dass es nicht mit brauchbarem Markup versehen ist, nur rudimentärem XHTML. Deshalb kam eigentlich nur das TextGrid Repository infrage, das sich aus den alten Zeno.org-Volltexten speist und basale TEI-Auszeichnungen aufweist.

Aus dem TextGrid-Gesamtkorpus wurden zunächst die in den Metadaten mit dem Genre 'drama' versehenen Texte extrahiert, insgesamt 690. Dazu gehören vor allem deutschsprachige Dramen von ca. 1500 bis 1930 sowie ferner u. a. Übersetzungen von einem Dutzend griechischer Tragödien und einiger Shakespeare-Dramen. Aus der Gesamtmenge lassen sich prinzipiell auch recht einfach zeitlich gestaffelte Teilkorpora erstellen, denn im TEI-Header stehen innerhalb von `<creation></creation>` rudimentäre Entstehungsdaten (Beispiele: `<date when="1802"/>`, aber auch weitläufige Eingrenzungen wie `<date notBefore="1738" notAfter="1758"/>`).

3 Erhebung der Netzwerkdaten

Als Zwischenschritt wurde für jede der 690 TEI-Dateien eine Relationsliste (CSV-Datei) erzeugt, die den gängigen Formaten der Speicherung netzwerkanalytischer Daten entspricht. Zur Extraktion der Sprecherdaten sind in der Regel zwei getrennte Schritte nötig: Das Erkennen der einzelnen Teile eines Theaterstückes und danach das Erkennen der einzelnen Sprecher.

Zur Erleichterung der nachstehenden Arbeiten teilt das Skript die vorliegenden Dateien auf: Für jede erkannte Ebene (die Datei selbst ist dabei auch eine) wird ein Unterverzeichnis angelegt, in dem wiederum TEI-Dateien mit den einzelnen Teilen stehen und in das auch die jeweiligen Registerdateien geschrieben werden. Anhand der aufgeteilten Dateien werden verschiedene Arten von Ausgaben erstellt. Zum einen ist dies ein kleinteiliges Register aller `<speaker>`-Tags, aber auch aller Auszeichnungen `<rs>` und `<person>`. Um eindeutige Referenzziele zu erhalten, werden ggf. ID-Nummern vergeben (dies erleichtert insbesondere auch spätere Eingriffe, wenn problematische Namen manuell korrigiert werden müssen). Zum andern werden die Kookkurrenzlisten erstellt. Im untersten Verzeichnis werden die Vorkommen aller Sprecherpaare in allen Dateien gezählt. In den darüberliegenden werden die Werte aller Unterverzeichnisse addiert.

Neben dem Erkennen der Struktur ist die korrekte Zuordnung von Namen die größte Herausforderung. Im Idealfall sind alle `<speaker>` mit einem Attribut `@who` versehen, über das eine normierte Form des Namens erreicht werden kann. Ist dies nicht der Fall (oder sind stattdessen die Tags `<rs>` oder `<person>` verwendet worden), muss das Skript den Textinhalt des Tags auswerten, wobei neben möglichen Verschreibungen (bei der Transkription oder in der Vorlage) auch syntaktische Änderungen auftreten können. So findet sich bei Lessing, *Nathan der Weise* V/1, neben Saladin "Ein Mameluck", der nach seinem ersten Auftreten als "Der Mameluck" geführt wird. Es folgt ein weiterer: "Ein zweiter Mameluck", danach "zweiter Mameluck". Ist es hier noch möglich, durch ein Berücksichtigen der Artikel mit einfachen Mitteln gute Ergebnisse zu erzielen, stellt sich dies bei Emilia Galotti etwas schwieriger dar, da teils nur "Odoardo", teils aber "Odoardo Galotti" erscheint. Auch Fälle mit mehreren Sprechern (z. B. "Alle") sind nicht ganz trivial zu bearbeiten.

Neben dem Versuch, diese Fälle automatisch zu klären, besteht in diesen Zweifelsfällen aber immer noch die Möglichkeit des manuellen Eingriffs, wozu die erstellten Indexdateien mit eindeutiger ID beitragen können. In einer weiteren Überarbeitung des Skriptes ist es vorgesehen, eine einfache graphische Oberfläche anzubieten, über die solche Zweifelsfälle bearbeitet werden können.

4 Datenauswertung und Visualisierung

Die Datenauswertung erfolgt über Python (3.4.x) mit dem igraph-Paket, das sowohl zum Visualisieren der Graphen als auch zum Berechnen der netzwerkanalytischen Daten genutzt wird.

Für eine erste Visualisierung des Datenbestands wurden die Graphdaten an eine spring-embedding-Methode übergeben (Fruchterman-Reingold), die versucht, affine Knoten näher beieinander anzuordnen und dadurch deutlich sichtbar zu clustern. Einen Eindruck des gesamten Korpus vermittelt

Abbildung 1, die 671 Dramen aus 2500 Jahren Dramengeschichte enthält, chronologisch links oben mit den Griechen beginnend und bis rechts unten ins zweite Viertel des 20. Jahrhunderts reichend:

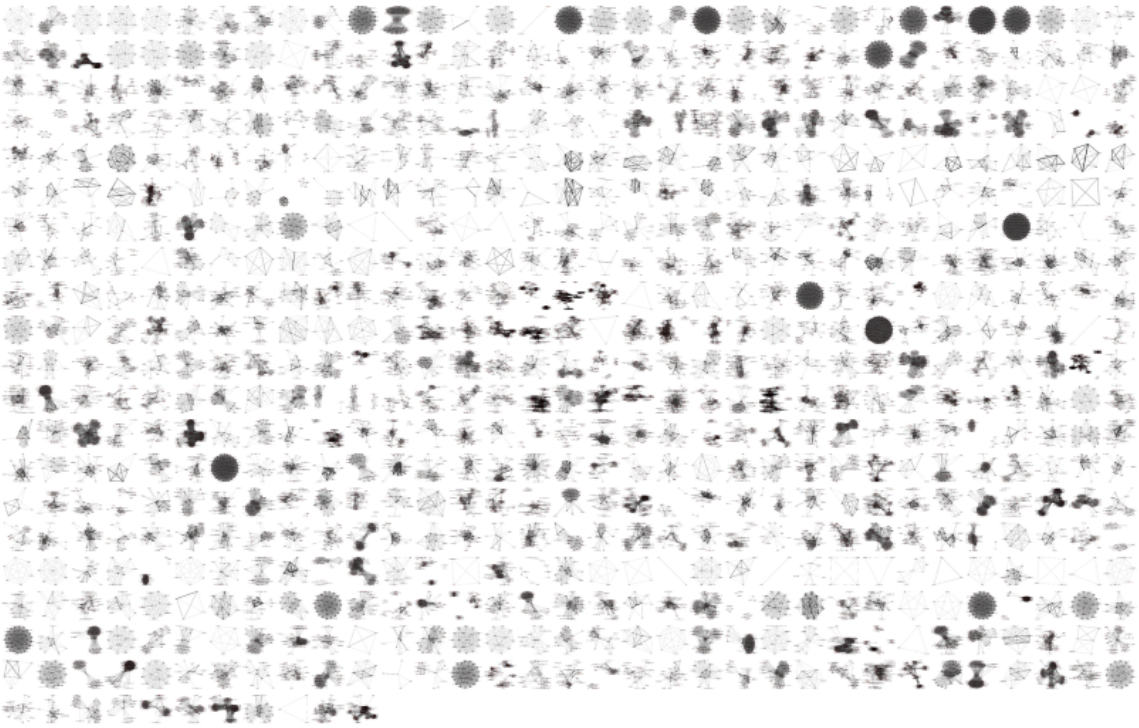


Abbildung 1: Netzwerkgraphen von 671 Dramen aus dem TextGrid Repository

Die visualisierten Graphen haben auch deutlich gemacht, dass die meisten berechneten CSV-Dateien wegen des teils nicht eindeutigen Markups zumindest kleine Fehler aufwiesen. Diese Erkenntnisse konnten zur Fehlerbehandlung an den vorhergehenden Schritt (Erhebung der Netzwerkdaten) zurückgegeben werden.

Erste strukturanalytische Berechnungen erfolgten auf Basis der 12 (vollendeten) Lessing-Dramen. Entsprechende Diagramme sind in Abbildung 2 zu finden.

5 Ausblick

Die erhobenen und bereinigten Netzwerkdaten sollen als Grundlage für alle statistischen Berechnungen dienen und auch öffentlich zur Verfügung gestellt werden. Im Mittelpunkt der Forschung steht nun die Implementierung zusätzlicher netzwerkanalytischer Berechnungstools (etwa zur Bestimmung der Betweenness Centrality, mit der die Wichtigkeit einzelner Figuren für das Netzwerk bestimmt werden kann). Darüber hinaus wird an der Qualifizierung der Netzwerkdaten gearbeitet (außer dem reinen Fakt, dass Figuren miteinander sprechen: Redeanteile quantifizieren, Bühnenpräsenz nicht sprechender Personen mit einbeziehen usw.) sowie an der Erstellung multiplexer Netzwerke, die nicht nur die oben definierten ‘Interaktions’-Relationen erfassen, sondern auch u. a. Verwandschafts- oder instrumentelle Beziehungen berücksichtigen.

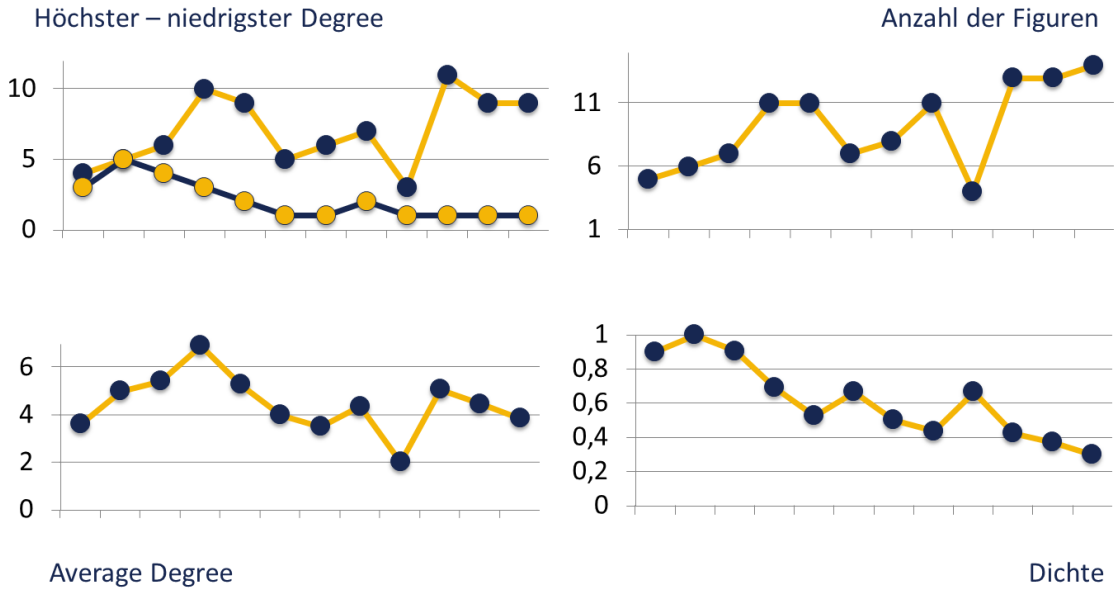


Abbildung 2: Beispielberechnungen anhand der Lessing-Dramen (x-Achse): Damon, 1747 – Der junge Gelehrte, 1747 – Der Misogyn, 1748 – Die alte Jungfer, 1748 – Der Freigeist, 1749 – Die Juden, 1749 – Der Schatz, 1750 – Miß Sara Sampson, 1755 – Philotas, 1759 – Minna von Barnhelm, 1767 – Emilia Galotti, 1772 – Nathan der Weise, 1779

Literatur

Marcus, Solomon. *Mathematische Poetik.* Frankfurt/M. 1973.

Moretti, Franco. *Network Theory, Plot Analysis.* Stanford Literary Lab Pamphlets 2 (1. 5. 2011). `http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf`.

Pfister, Manfred. *Das Drama. Theorie und Analyse.* München 1977 u. ö.

Pohlheim, Karl Konrad (Hg.). *Die dramatische Konfiguration.* Paderborn u. a. 1997.

Titzmann, Michael. *Strukturale Textanalyse. Theorie und Praxis der Interpretation.* München 1977 u. ö.

Trilcke, Peer. *Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft.* In: Philip Ajouri, Katja Mellmann u. Christoph Rauen (Hg.): Empirie in der Literaturwissenschaft. Münster 2013. S. 201–247.

Wasserman, Stanley; Katherine Faust. *Social Network Analysis. Methods and Applications.* Cambridge u. a. 1998.

Comedia - Comédie: Topic Modeling als Perspektive auf das spanische und französische Theater des 17. Jahrhunderts

Abstract für die zweite Jahrestagung des Verbandes der *Digital Humanities im deutsprachigen Raum* zum Thema “Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation”, 23.-27. Februar 2015 an der Universität Graz

Christof Schöch
Universität Würzburg

Nanette Rißler-Pipka
Universität Siegen

1. Hintergrund

Im Europa des 17. Jahrhunderts entwickelten sich zeitgleich verschiedene Formen des Theaters. Trotz unterschiedlicher sozialer und poetologischer Kontexte weisen das spanische und französische Theater viele (stoffliche / stilistische) Verbindungen auf, die auf eine gemeinsame europäische Theatergeschichte hindeuten (Couderc 2013). Die Frage, ob man dazu nicht Methoden der Digital Humanities nutzen sollte, stellte sich bisher in der Romanistik nicht (anders als in weiteren Philologien, vgl. Rybicki 2012, Jockers 2013, Eder 2014). Allgemein gilt, dass sprachübergreifende, quantitative Textanalysen eine Herausforderung bleiben (vgl. Steinberger 2009, Eder/Rybicki 2011).

In romanistischer Tradition über Sprachgrenzen hinweg quantitative Verfahren anzuwenden, scheint mit Topic Modelling möglich zu sein: Die Topics mehrerer einheitlich strukturierter Textsammlungen können zunächst unabhängig voneinander modelliert werden, um dann auf der Grundlage von Topic-Labels und strukturellen Merkmalen Ähnlichkeiten und Unterschiede zu ermitteln.

2. Fragestellungen

Anstelle von Einzelhypothesen und konkreter Passagenvergleiche sind hier zwei spanische und französische Textsammlungen durch Topic Modeling verglichen worden. Welche Arten von Topics liegen vor, und wie verhalten sie sich zueinander? Welche Relation besteht zwischen den Topics und Kategorien wie Untergattungen (Komödie / Tragödie)? Wie gestaltet sich dies im Vergleich des spanischen und französischen Theaters?

Über diese Fragen hinaus soll die Eignung der Methode für Theaterstücke geprüft werden. Wie verhalten sich die "Topics" zu theaterwissenschaftlich relevanten "Themen"? Welche Perspektivenverschiebung ergibt sich durch ein quantitatives Verfahren wie Topic Modeling?

3. Textsammlungen

Die spanische Textsammlung enthält 145 Theaterstücke von sechs Autoren. Die Stücke sind zwischen 1585 und 1688 erschienen. Die Untergattungen sind "drama", "comedia" und "auto sacramental". Die Texte stammen von www.comedias.org, Wikisource und der Biblioteca Cervantes.

Die französische Textsammlung enthält 143 Theaterstücke von neun Autoren. Die Stücke sind zwischen 1630 und 1708 erschienen, und stammen von www.theatre-classique.fr. Die Untergattungen sind "comédie", "tragédie" "tragi-comédie" und "pastorale".

4. Methode: Topic Modeling

Topic Modeling ist ein quantitativer Ansatz, um in größeren Textsammlungen thematische Muster zu entdecken (Blei 2003; Anwendungen in den DH: Blevins 2010, Rhody 2012, Jockers 2013). Mathematisch gesehen sind „Topics“ Verteilungen von Auftretenswahrscheinlichkeiten von Wörtern. Die Wörter eines Topic mit der höchsten Auftretenswahrscheinlichkeit sind sich semantisch (oder anderweitig) ähnlich (vgl. Blei 2011 und Steyvers & Griffiths 2007). Durch Verknüpfung mit Metadaten können thematische Trends über einen Zeitverlauf oder thematische Differenzen zwischen Textgattungen entdeckt werden.

Wichtige Parameter sind das Präprozessieren der Texte (bspw. Lemmatisierung), die Auswahl der zu berücksichtigenden Wörter (nach Wortarten, Wortfrequenzen oder Stoplist), die Textsegmentierung sowie die Anzahl der Topics, die gefunden werden sollen. Für diese Studie wurden die Texte mit TreeTagger (Schmidt 1994) lemmatisiert und nach Wortarten annotiert. Es wurden verschiedene Textfassungen generiert, die bspw. nur Substantive und Verben enthalten, die Texte in Segmente von 40 Lemmata zerlegt und Topic Modeling mit MALLET (McCallum 2009) durchgeführt. Die Anzahl der Topics wurde auf 50 bzw. 200 festgelegt.

5. Ergebnisse und Diskussion

5.1 Die ermittelten Topics

Die ermittelten 50 Topics lassen sich meist mit einem Begriff zusammenfassen, der die inhaltliche Gemeinsamkeit der wichtigsten Worte im Topic fasst. Es gibt allgemeinere und spezifische Topics mit unterschiedlichem Gewicht in der Textsammlung, was hier am Beispiel der französischen Topics gezeigt wird (Abb. 1).

Topic „Label“	Topic-Score	Topic-Worte mit höchstem Score
Topic 14: „Liebe“	0.154	aimer amour cœur amant haïr œil âme flamme feu
Topic 34: „Komödie“	0.014	comédie pièce jouer monsieur trouver monde auteur rôle comédien
Topic 33: „Suchen-Finden“	0.180	trouver attendre sortir lieu chercher heure temps quitter ami
Topic 1: „Vergnügen“	0.100	homme esprit trouver gens femme monde rire plaire discours

Abb. 1: Auswahl von Topics aus der französischen Textsammlung.

Einige Topics betreffen allgemein gefasste, erwartbare Themen, wie bspw. Liebe / Intrigen (5 der 6 wichtigsten Topics gehören in diesen Themenbereich). Das Liebestopic enthält oft ein Element des Schmerzes und Hasses, das auf die Tragödie hindeutet (Topic 14). Dagegen lässt sich ein inhaltlich typisches Komödientopic nur durch selbstreferentiellen Begriffe erkennen (Topic 34). Faktisch am distinktivsten für die Komödie sind dagegen ein relativ unbestimmtes Topic (33, "Suchen-Finden") sowie Topic 01 ("Vergnügen"; vgl. 5.3).

Andere Topics sind spezifischer (bspw. Topic 11, "Gefahr" oder Topic 24 "Geheimnis") und könnten vermuten lassen, dass sie mit bestimmten Untergattungen des Theaters verknüpft sind (bspw. Topics 30 und 38, "Verbrechen"). Allerdings zeigt ein Vergleich von Topics und Textklassen (vgl. 5.3.), dass Topic 38 zwar der Tragödie zugeordnet werden kann, es in Topic 30 aber offenbar um ein “Verbrechen” geht, das sich in Komödien abspielt.

5.2 Die Topics im Vergleich

Vergleicht man die Topics der französischen und der spanischen Textsammlung miteinander, stellt man einige Übereinstimmungen und Unterschiede fest (Abb. 2).

Französische Sammlung		Spanische Sammlung	
Topic 41: „Liebe-Hoffnung“	cœur **amour aimer** oser **espoir** gloire **âme** vœu souffrir	**amor** celo **amar alma** amantar olvidar ver **esperanza** favor	Topic 35 „Liebe-Hoffnung“
Topic 43: „Krieg“	**guerre soldat** armée bataille chef ennemi camp champ muraille	**soldado guerra** arma valor gente tocar caja vencer armar	Topic 13 „Krieg“
Topic 7 „Arzt“	**médecin** mal **remède guérir** monsieur **maladie** fille demander homme	fingir pinzón **médico** doctor **curar** salud **enfermedad cura remedio**	Topic 17 „Arzt“
(keine Entsprechung)	-/-	justicia rey juez castigo delito mandar muerte sentencia prisión	Topic 16 „Gericht-König“
Topic 38 „Verbrechen“	crime venger mort punir sang haine vengeance perdre fureur	-/-	(keine Entsprechung)

Abb. 2: Topics im Sprachvergleich

In beiden Textsammlungen präsent sind allgemeine Topics, wie diejenigen um das Thema “Liebe” (bspw. Topic 41fr vs. Topic 35sp). Zwar kann man, abgesehen von einer leichten Tendenz in Richtung Lust (“celo, ver”) im spanischen und Leid (“souffrir”) im französischen Topic, kaum von einer semantischen Differenz sprechen. Dennoch kann Topic 35sp in der Topicverteilung nach Gattungen (vgl. 5.3) der (spanischen) "Comedia" zugeordnet werden, während Topic 41fr der (französischen) Tragödie zugeordnet wird. Die ähnlich große Wichtigkeit beider Topics in den jeweiligen Korpora belegt die stoffgeschichtliche Verwandtschaft des Theaters beider Länder.

Auch bei noch spezifischeren Topics gibt es zahlreiche Übereinstimmung, bspw. Topic 4fr und Topic 2sp, die beide mit dem Titel “Gnade-Gottes” versehen werden könnten, oder sehr konkrete Topics wie "Krieg" (Topic 46fr und 13sp) oder "Arzt" (Topic 7fr und 17sp), die mit fast identischen Wörtern vorkommen.

Topics in der spanischen Sammlung ohne Übereinstimmung in der französischen Sammlung sind bspw. Topic 45 (“Schuld-Unschuld”) oder 16 (“Gericht-König”). Umgekehrt sind Topics in der französischen Sammlung ohne Übereinstimmung in der spanischen Sammlung bspw. Topic 18 "Gehorsam" oder 38 "Verbrechen"). Diese Ergebnisse bieten Ausgangspunkte für einen Abgleich mit Erkenntnissen der Literaturgeschichte.

5.3 Topics und Textklassen

Mit unterschiedlicher Ausprägung zeigt sich in beiden Textsammlungen, dass die Untergattungen jeweils mindestens einen charakteristischen Topics besitzen (Abb. 3 und 4).

topics	AutoS	Comedia	Drama	(sd)
tp45	0.083	0.002	0.003	0.047
tp35	0.025	0.095	0.064	0.035
tp09	0.083	0.023	0.034	0.032
tp11	0.065	0.007	0.023	0.030
tp21	0.062	0.005	0.021	0.030
tp44	0.064	0.016	0.018	0.027
tp00	0.032	0.060	0.061	0.017
tp12	0.068	0.098	0.084	0.015
tp46	0.007	0.020	0.035	0.014
tp36	0.032	0.008	0.011	0.013
tp38	0.010	0.032	0.017	0.011
tp48	0.017	0.039	0.030	0.011
tp49	0.007	0.027	0.009	0.011
tp37	0.019	0.023	0.038	0.010
tp41	0.040	0.045	0.057	0.008
tp07	0.039	0.042	0.054	0.008
tp34	0.010	0.021	0.009	0.007
tp03	0.016	0.028	0.017	0.007
tp27	0.029	0.017	0.029	0.007
tp19	0.017	0.029	0.022	0.006

Abb. 3: Heatmap für Topic-Scores in Genres (Spanisch)
(20 Topics mit größter Varianz, gemessen als Standardabweichung)

Der wesentliche Kontrast bei den spanischen Stücken (Abb. 3) liegt zwischen "Auto sacramentales" einerseits, “Comedias” und “Dramas”, andererseits. Letztere haben schwächer kontrastive Topics, bspw. Topic 35 ("Liebe-Hoffnung") oder, auf niedrigerem Niveau, Topic 49 (unklar).

Bei den französischen Stücken hat jede Untergattung zumindest einen charakteristischen Topic (Abb. 4): Topic 33 ("Suchen-Finden") für die Komödie, Topic 41 ("Liebe-Hoffnung") für die Tragödie, Topic 08 ("Liebe-Schönheit") für die Pastorale. Ausnahme ist die Tragikomödie.

topic	Comedie	Tragedie	Tracom.	Pastorale	(sd)
tp08	0.052	0.029	0.087	0.129	0.044
tp41	0.059	0.120	0.061	0.050	0.032
tp19	0.027	0.076	0.057	0.015	0.028
tp33	0.088	0.029	0.051	0.051	0.024
tp38	0.026	0.078	0.043	0.029	0.024
tp01	0.061	0.012	0.016	0.019	0.023
tp03	0.002	0.002	0.004	0.044	0.021
tp28	0.041	0.004	0.006	0.007	0.018
tp26	0.028	0.044	0.069	0.049	0.017
tp14	0.042	0.051	0.049	0.078	0.016
tp11	0.026	0.058	0.033	0.029	0.015
tp20	0.045	0.068	0.044	0.036	0.014
tp00	0.040	0.011	0.015	0.015	0.013
tp48	0.013	0.031	0.033	0.011	0.012
tp49	0.025	0.002	0.003	0.003	0.011
tp02	0.038	0.055	0.063	0.059	0.011
tp22	0.008	0.010	0.010	0.031	0.011
tp09	0.010	0.028	0.016	0.006	0.010
tp05	0.030	0.012	0.015	0.028	0.009
tp24	0.038	0.047	0.029	0.030	0.008

Abb. 4: Heatmap für Topic-Scores in Genres (Französisch)
(20 Topics mit größter Varianz, gemessen als Standardabweichung)

Insgesamt scheint die gattungsbezogene Trennschärfe in den französischen Texten deutlicher als in den spanischen Texten. Dieser Befund entspricht den unterschiedlichen französischen und spanischen Poetiken der Zeit.

5.4 Gruppierung auf Grundlage von "topic scores"

Es ist nicht auszuschließen, dass die verwendeten Gattungsbezeichnungen tatsächlich vorhandene Differenzierungen verdecken. Ohne vorgängige Kategorien, nur auf Grundlage der Ähnlichkeit von Stücken nach der Verteilungen von 200 Topics sollten daher mit Principal Component Analysis Strukturen in den Textsammlungen gefunden werden.

Die räumliche Verteilung der Stücke zeigt für die spanischen Texte kaum Struktur und bildet eine recht einheitliche Wolke (Abb. 6). Die französischen Texten (Abb. 5) zeigen mehr Struktur: ein kompakterer, leicht separierter Bereich rechts oben sowie ein weiterer, besonders dichter Bereich links oben. Die in den ersten beiden Komponenten enthaltene Varianz der Daten ist mit zusammen 17,3% (französisch) und 9,2% (spanisch) verhältnismäßig gering.

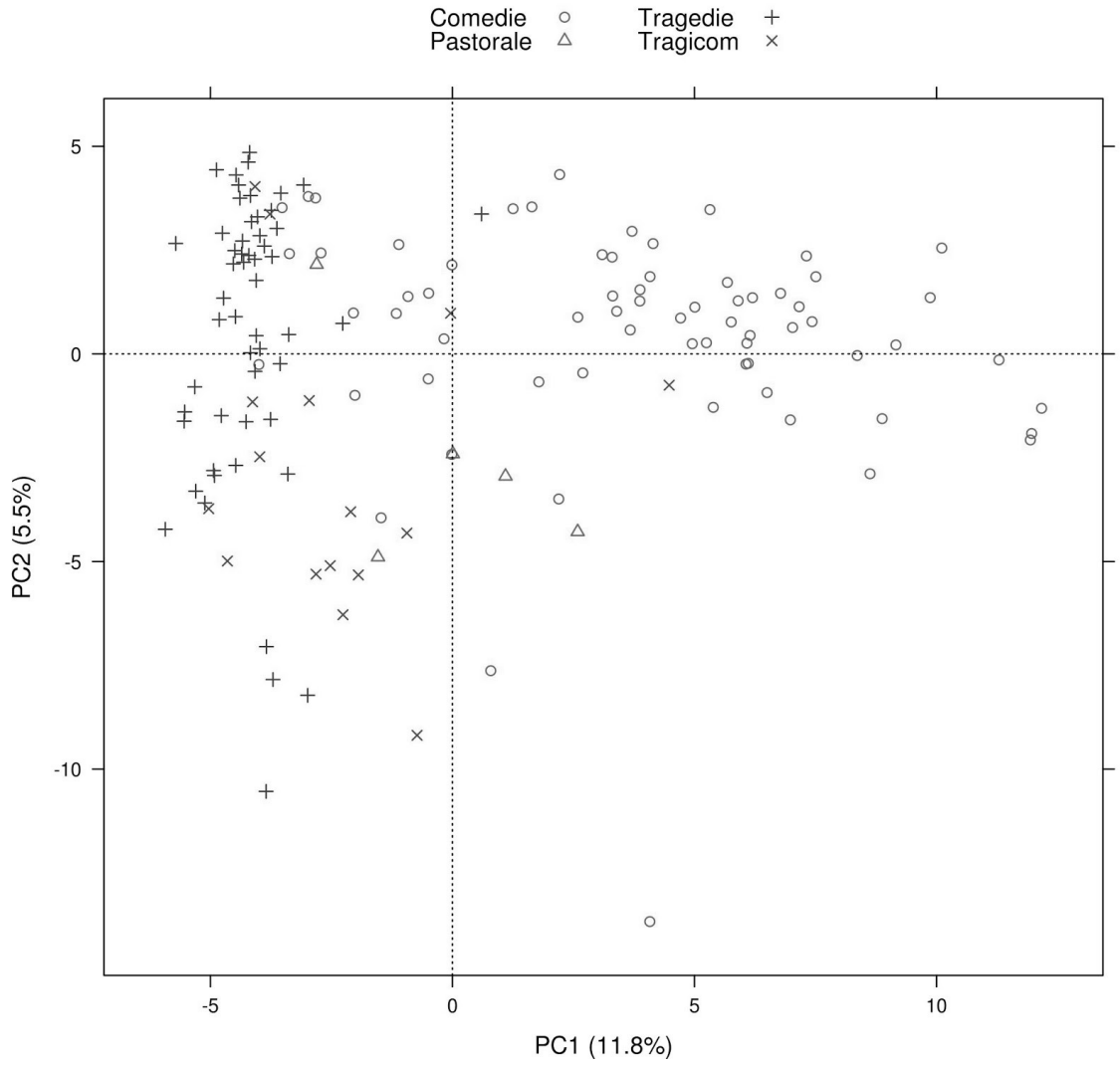


Abb. 5: PCA-Plot auf Grundlage von 200 topic scores (französische Sammlung, Genre-Labels)

Die Verteilung der Gattungssymbole zeigt, dass die französischen Texte nach Gattungen gruppiert sind: rechts oben die Komödien, links oben die Tragödien; die stärker verteilten Tragikomödien überlappen vor allem mit den Tragödien.

Bei den spanischen Texten gibt es ebenfalls Gruppen: die “Auto Sacramentales” im linken unteren Quadranten, die “Comedias” eher in der rechten Hälfte, die Dramen breit gestreut in der Mitte.

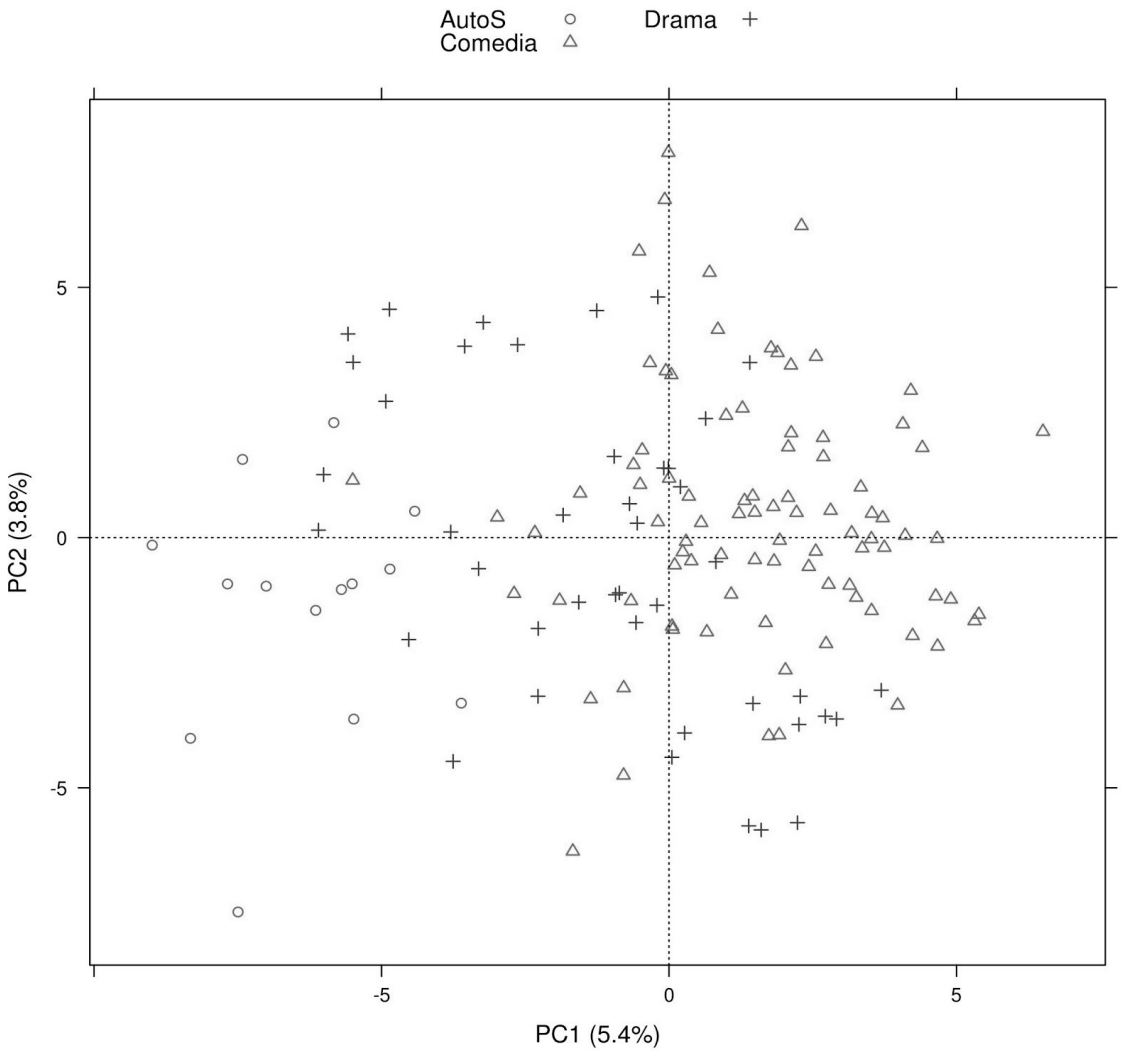


Abb. 6: PCA-Plot auf Grundlage von 200 topic scores (spanische Sammlung, Genre-Labels)

Einfache Korrelationstests bestätigen den Gesamteindruck (Abb. 7). In den französischen Stücken korreliert Genre sehr deutlich nur mit PC1, Autorschaft dagegen vor allem mit PC2. Bei den spanischen Stücken ist nur die Korrelation zwischen Autorschaft und PC1 stark.

	Französische Stücke			**Spanische Stücke**		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Korrelation mit Autorschaft	0.33 ***	-0.56 ***	-0.05 ns	0.72 ***	-0.28 ***	-0.21 *
Korrelation mit Gattung	0.72 ***	0.18 *	-0.18 *	0.14 *	0.22 ns	0.04 ns
Varianz (sd)	4.86	3.31	2.89	3.28	2.75	2.61

Abb. 7: Korrelationtests zwischen Principal Components und Autorschaft bzw. Gattungszugehörigkeit

Die thematische Differenzierung der Stücke ist also in der französischen Textsammlung stärker ausgeprägt und korreliert auch stärker mit den vorhandenen Gattungs-Kategorien als in der spanischen Textsammlung.

Bilanz und nächste Schritte

Zahlreiche Einzelergebnissen zum Verhältnis der inhaltlichen Bestimmung einzelner Topics und ihrer eventuellen Zuordnung zu Untergattungen des Theaters zeigen, dass sich spanisches und französisches Theater auf Grundlage der Topic-Verteilungen auf eine Weise unterscheiden, die gattungspoetischen Positionen der Zeit entspricht und an vorhandene literaturwissenschaftliche Erkenntnisse anschlussfähig ist.

Außerdem zeigen die Ergebnisse den Unterschied zwischen “Topics” und “Themen” im literaturwissenschaftlichen Sinn. Der semantische Gehalt des Topics, der in einem Begriff wie “Liebe-Leidenschaft” (Topic 14) gebündelt werden kann, beschreibt *nicht* unbedingt das zentrale Thema der sich dahinter verbergenden Theaterstücke (vgl. die Diskussion der Tragödien-, Komödien und Pastoralentopics). Diese vermeintliche Kluft zwischen Topics und literaturwissenschaftlichen Themen ist aber eher eine Chance als ein Dilemma: so lassen sich vorschnelle Interpretationsansätze überprüfen und neue Erkenntnisse gewinnen.

Methodisch wird deutlich, dass Topic Modeling selbst nur ein Schritt in der Analyse- und Interpretationskette sein kann, der durch linguistische Annotation und Metadaten vorbereitet werden muss, und dessen Ergebnisse durch weitere Verarbeitung und Kontextualisierung erst bedeutungsvoll werden.

Als nächste Schritte könnte die Textsammlung erweitert und um weitere Metadaten ergänzt werden, um die Vergleichbarkeit der Textsammlungen zu erhöhen. Es könnte mit "Multilingual Topic Modeling" (Boyd-Graber & Blei 2009) operiert werden, das unmittelbar thematische Bezüge zwischen Dokumenten in unterschiedlichen Sprachen ermittelt. Alternativ wäre ein algorithmisches Verfahren zur Ähnlichkeitsbestimmung verschiedensprachiger Topics zu entwickeln (vgl. Pouliquen 2006).

Bibliographie

Blei, David M. 2011. “Introduction to Probabilistic Topic Models.” *Communication of the ACM*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3,March: 993–1022.

Blevins, Cameron. 2010. “Topic Modeling Martha Ballard’s Diary.” *Historying*. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Boyd-Graber, Jordan, and David M. Blei. 2009. “Multilingual Topic Models for Unaligned Text.” In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 75–82. UAI ’09. Arlington, Virginia, United States: AUAI Press. http://dl.acm.org/citation.cfm?id=1795114.1795124.

Couderc, Christophe 2013: *La tragédie Espagnole et son contexte Européen : XVIe - XVIIe siècles*, Paris : Presses Sorbonne Nouvelle.

Eder, M. 2014. Stylometry, network analysis and Latin literature. In: *Digital Humanities 2014: Book of Abstracts*, EPFL-UNIL, Lausanne, pp. 457-58. http://dharchive.org/paper/DH2014/Poster-324.xml

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

McCallum, Andrew K. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Pouliquen, Bruno, Ralf Steinberger, and Camelia Ignat. 2006. “Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus.” *arXiv:cs/0609059*, September. http://arxiv.org/abs/cs/0609059.

Rhody, Lisa M. 2012. “Topic Modeling and Figurative Language.” *Journal of Digital Humanities* 2,1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Rybicki, Jan. 2012. The great mystery of the (almost) invisible translator: stylometry in translation. In M. Oakley and M. Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, pp. 231-248.

Rybicki, Jan, and Maciej Eder. 2011. Deeper Delta across genres and languages: do we really need the most frequent words?Literary and Linguistic Computing 26(3), 315-21.

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

Steyvers, Mark, and Tom Griffiths. 2006. “Probabilistic Topic Models.” In *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum.

Automatische Erkennung von Figuren in deutschsprachigen Romanen

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

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

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

Toepfer, Martin
toepfer@informatik.uni-wuerzburg.de, Universität Würzburg

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

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

Eine wichtige Grundlage für die quantitative Analyse von Erzähltexten, etwa eine Netzwerkanalyse der Figurenkonstellation, ist die automatische Erkennung von Referenzen auf Figuren in Erzähltexten, ein Sonderfall des generischen NLP-Problems der Named Entity Recognition [Sharnagat 2014]. Mit dem Stanford Parser [Finkel 2005] unter Verwendung eines Modells für deutsche Sprache [Faruqui and Pado 2010] liegen inzwischen auch freie Werkzeuge für Texte in deutscher Sprache vor. Allerdings ist die Erkennungsrate des Modells, das an einem Korpus von Zeitungstexten trainiert wurde, für literarische Texte nur eingeschränkt brauchbar (Abb. 1). Eine Auswertung anhand unseres Testkorpus (265 000 Tokens) hat einen F1-Score von nur 31% ergeben, was vor allem am sehr niedrigen Recall lag. Dieser Befund deckt sich mit vergleichbaren Erfahrungen aus der Computerlinguistik: Viele NLP-Werkzeuge müssen erst für einen neuen Anwendungsbereich angepasst werden, um brauchbare Resultate zu erbringen. Im Fall des Romankorpus führt die Einbeziehung von Appellativen in die Named Entity-Definition und deren häufige Verwendung in Romantexten zu dem schlechten Ergebnis. Da die Figurenreferenzen allerdings für fast alle nachfolgenden Verarbeitungsschritte eine hohe Relevanz haben, sind wir *nicht* den Weg einer automatischen Domänenadaption [Qi Li 2012] gegangen, sondern haben ein umfangreiches Trainingskorpus aufgebaut, um auf diese Weise möglichst hohe Erkennungsraten zu erhalten. Im Folgenden berichten wir über unser Vorgehen, diese Aufgabe möglichst effizient zu gestalten. Zusammenfassend können wir feststellen, dass wir die Erstellung des notwendigen Trainingskorpus durch ein Werkzeug erheblich beschleunigen konnten, das den Annotatoren bereits gute Vorschläge machte. Außerdem konnten die Resultate des verwendeten Lernverfahrens dadurch deutlich verbessert werden, dass über die üblichen Standardfeatures hinaus word2vec-Informationen (s.u.) als Feature verwendet wurden.

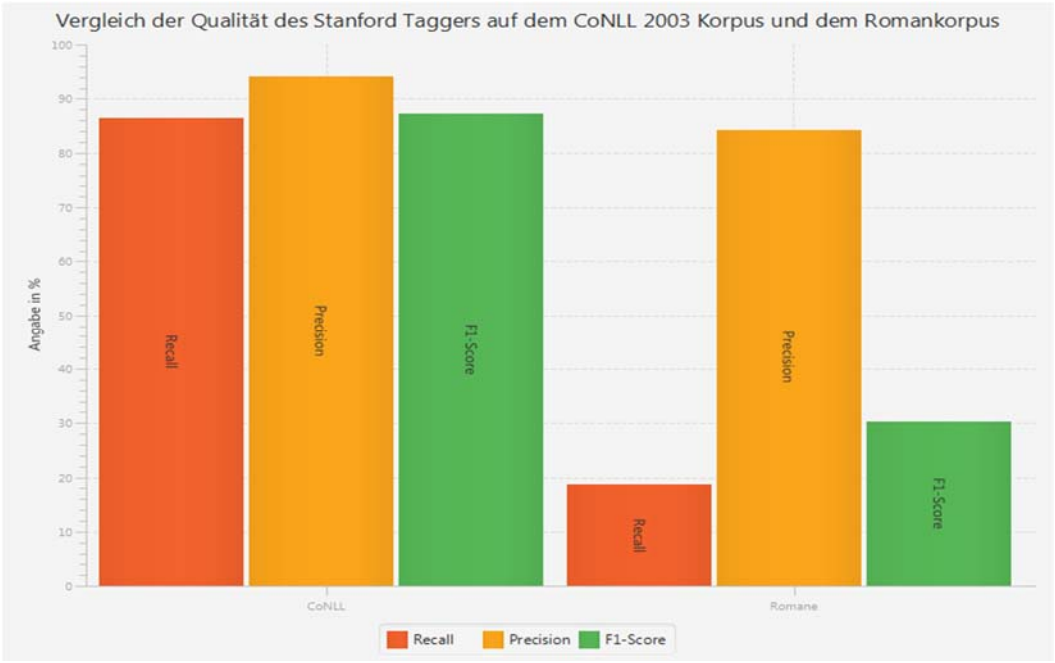


Abb. 1: Ergebnisse des Stanford-Parsers mit deutschem Modell (Faruqui and Pado 2010) angewandt auf ein Zeitungskorpus (CoNLL 2003) und ein Korpus deutschsprachiger Romane.

Material und Methoden

Als annotierte Trainings- und Testdaten dienten das Zeitungskorpus der CoNLL 2003 [Sang 2003] (ca. 220 000 Tokens) und ein von uns aufbereitetes Romankorpus mit je 130 zusammenhängenden Sätzen aus 50 Romanen mit 140 000 Tokens für das erste Experiment, und 85 Romanen mit 265 000 Tokens für das zweite Experiment. Die Annotation geschah mittels einem eigens für diesen Zweck entwickelten Werkzeug, das über eine komfortable grafische Benutzeroberfläche dem Annotator die mit einfachen Regeln ermittelten Vorschläge zur Bearbeitung anbietet, wodurch sich die Annotation erheblich beschleunigen ließ (die vorher direkt in XML-Dateien und dann in einem Annotationswerkzeug durchgeführt wurde, das nicht spezifisch für die Aufgabe angepasst wurde). Notiert wurden folgende Eigenschaften:

a) Handelt es sich um einen wirklichen Namen, z.B. „Effi Briest“, oder um einen Appellativ, z.B. der „Lehrer“.
b) Handelt es sich um eine einzelne Person oder um eine Personengruppe bzw. um mehr als eine Person, z.B. die „Gäste“.
c) Koreferenz per Identität (ID), d.h. alle Referenzen auf die gleiche Figur erhalten die gleiche grafisch angezeigte ID.

Für die Anwendung unüberwachter Lernverfahren verwendeten wir Texte aus der FAZ (ca. 15 Millionen Tokens) und unser Erweiterungskorpus deutschsprachiger Romane (ca. 60 Millionen Tokens), beide Textsammlungen nicht annotiert.
In der ersten Serie von Experimenten wurde die Frage untersucht, mit welchen Features das maschinelle Lernverfahren Conditional Random Fields (CRF), das auch im Stanford Parser eingesetzt wird, die besten Ergebnisse erbringt. Folgende sechs Features, die vom Stanford-Tagger [Finkel 2005] verwendet werden, wurden als Basis betrachtet:

1) Current Word: das Wort an Position i
2) Previous Word: das Wort an Position i-1
3) Next Word: das Wort an Position i+1
4) Word Shape: für Groß/Kleinschreibung oder Zahlen
5) Part-Of-Speech Tags (POS-Tags) an den Positionen i, i-1 und i+1, die mit Hilfe des TreeTaggers [Schmid 1995] bestimmt wurden.
6) Präfix bzw. Suffix, das aus den ersten oder letzten 2 Zeichen besteht.

Außerdem getestete Features:

7) Gazeteers: Listen bestehend aus rd. 5200 männlichen, 3400 weiblichen Vornamen, 160 Adelstiteln, Anreden und 8700 Berufen.
8) Semantische Felder, je nach Wortart 15-23, auf der Grundlage von GermaNet
9) Satzsubjekt ermittelt mit dem Mate-Dependecy Parsers [Bohnet 2010].
10) Compound-Words: alle von SFST [Fitschen 2004] erkannten Teilworte des Eingabewortes inkl. Prä- und Suffixe.
11) Head-Lemma: Grundform des zum Subjekt gehörenden Verbes.
12) LDA-Cluster: Es wird die Zugehörigkeit aller Nicht-Stop-Wörter zu dem wahrscheinlichsten von 250 Clustern mit der Latent-Dirichlet-Allocation (LDA) [Blei 2003] in Anlehnung an [Chrupala 2011] auf der Basis der oben erwähnten nicht annotierten Korpora mit 15 Millionen bzw. 60 Millionen Token ermittelt. Das LDA wurde mit dem Framework MALLET [MALLET 2002] implementiert.
13) Word2Vec-Cluster: Es wird ebenfalls die Zugehörigkeit aller Nicht-Stop-Wörter zu einem semantischen Cluster ermittelt. Dabei wurde eine effiziente Implementierung des “Continuous Bag-of-Words” Modells nach [Mikolov 2013] genutzt und die resultierenden Vektoren mit einem k-means Verfahren geclustert.

Ergebnisse

Zum Testen der gelernten CRFs wurde eine 10-fache Kreuzvalidierung auf der Trainingsmenge des Romankorpus (120.000 Tokens) durchgeführt. Die Baseline mit den Features 1-6 erbrachte einen F1-Score von 86,66%. Die Kombination der besten Features (letzte Zeile) erzielte einen F1-Score von 89,98, d.h. eine Steigerung um 3,32 Prozentpunkte. Der mit Abstand größte Anteil an dieser Steigerung ging auf das semantische Feature "Word2Vec-Cluster" zurück. Dagegen erbrachte das semantische Clustering mit LDAs einen eher negativen Effekt. In [Tkachenko 2012] wird der gleiche Effekt berichtet und die Vermutung geäußert, dass die LDA-Cluster redundant zu den POS-Tagging-Features sind. Beim Trainingskorpus mit den Zeitungsartikeln war die Baseline mit 87,9% etwas besser, aber die Steigerung durch Hinzunahme des Word2Vec-Cluster mit 1,6 Prozentpunkten (auf 89,5%) etwas schlechter.

Verfahren	**Precision in %**	**Recall in %**	**F1-Score in %**	**Unterschied zur Baseline (F1-Score) in %**
Baseline (Features 1-6)	95.12	79.60	86.66	+0
Baseline + (Feature 7)	95.73	79.28	86.70	+0.04
Baseline +(8)	94.53	81.74	87.65	+0.99
Baseline + (9)	94.96	79.74	86.67	+0.01
Baseline + (10)	95.07	81.00	87.45	+0.79
Baseline + (11)	95.03	79.63	86.63	-0.03
Baseline + (12)	**96.47**	77.83	86.13	-0.53
Baseline + (13)	94.97	85.28	89.84	+3.18
Baseline + (7),(8),(10),(13)	94.86	**85.60**	**89.98**	**+3.32**

Tab. 1. Einfluss verschiedener Features auf die NER mit CRFs; Trainingsset ca. 120 000 Tokens.

Wir haben beim Feature 13 "Word2Vec-Cluster" untersucht, welchen Einfluss die Anzahl der vorgegeben Cluster im k-means Verfahren zwischen 100 und 1000 auf die Qualität der NER hat. Dabei stellte sich heraus, dass bei einer Clusteranzahl ab 250 (relativ konstant bis 1000) das beste Ergebnis erzielt wird, so dass in weiteren Experimenten die Clusteranzahl von 250 gewählt wurde.

In unserem zweiten Experiment beschäftigten wir uns mit den Fragen, wie groß unser annotiertes Korpus für das Training eines praktisch nutzbaren NER-Modells sein muss, bzw. ab welcher Größe eine Erweiterung des Trainingsmaterials keine nennenswerte Verbesserung der Erkennungsleistung mehr bringt. Als zweiten Aspekt gilt es das für unseren Task beste Lernverfahren zu ermitteln. Für diesen Zweck haben wir die Erkennungsgenauigkeit mit immer größeren Mengen von Trainingsdaten gemessen: Für beide Domänen wurde zunächst nur eine Trainingsmenge von 30 000 Tokens genutzt, die dann in Schritten von 10 000 Tokens auf die Maximalzahl von 230 000 Tokens bei den Romanen bzw. 170 000 Tokens bei den Zeitungsartikeln gesteigert wurde. Als Features haben wir die jeweils beste Feature-Menge für das CRF verwendet. Neben dem CRF-Klassifikator wurden auch Maximum-Entropy, Naive Bayes und Decision-Trees mit der gleichen Menge an Features getestet. Abb. 2 zeigt, dass die beiden besten, von uns getesteten Klassifikationsverfahren MaxEnt, sowie CRFs sind. Auf dem Zeitungskorpus sind CRFs ca. 3-5% besser als MaxEnt, die Evaluation auf dem Romankorpus zeigt genau entgegengesetzte Ergebnisse. Eine Ausnutzung der Zustandsübergangsinformation, die CRFs zusätzlich zu MaxEnt nutzen, scheint im Fall der Romane keine nützlichen Informationen zu liefern, sondern das Ergebnis zu verschlechtern. Dies könnte in einer deutlich höheren durchschnittlichen Satzlänge (24,2 Tokens vs. 16,3 Tokens) in unserer Domäne begründet liegen. Ab einer Trainingsmenge von etwa 150 000

Tokens zeigt sich keine signifikante Verbesserung der Ergebnisse mehr. Wenn statt dieser 10-Fold Cross-Validation eine Leave-One-Out-Evaluation verwendet wird, bei der der zu testende Roman nicht in der Trainingsmenge enthalten ist, verringert sich der durchschnittliche F1-Score um ca. fünf Prozentpunkte von 88% auf 83.4%. Entgegen unserer Erwartung führte die Hinzunahme von 35 Romanen in dem Trainingskorpus zu keiner Verbesserung der Erkennungsrate, sondern sogar zu einer Verschlechterung um ca. 2%. Eine genauere Analyse zeigte, dass unter diesen zufällig ausgewählten Romanen auch solche mit Dialekten und anderen Besonderheiten waren, was die Verschlechterung erklären könnte.[1]

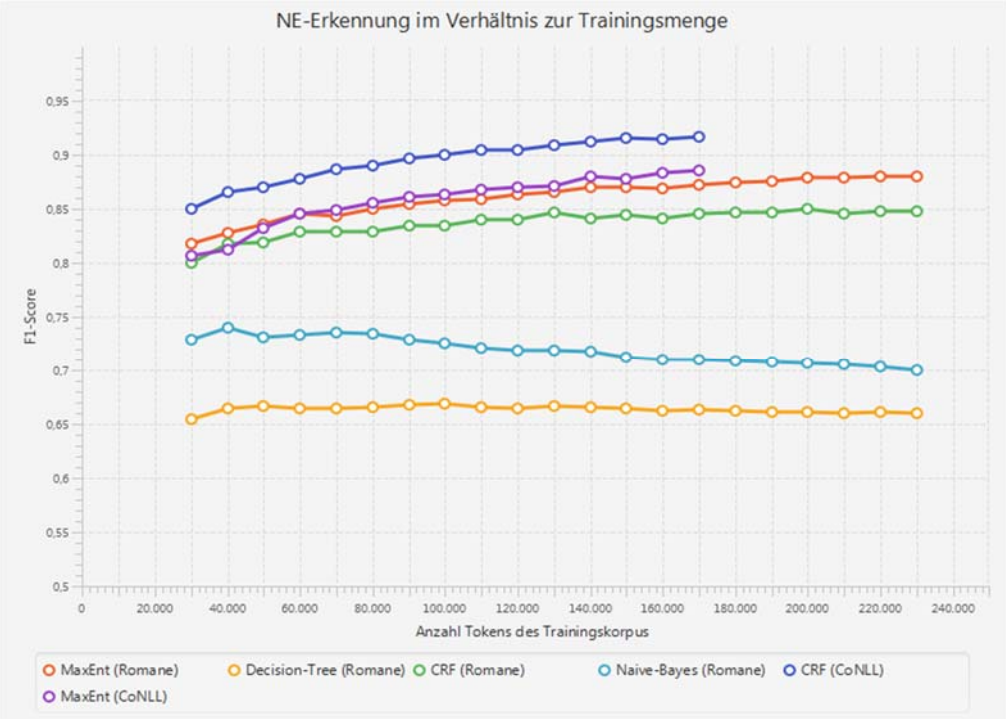


Abb. 2. Einfluss verschiedener Größen von Trainingsdaten von 30 000 bis 230 000 bzw. 170 000 Tokens auf den F1-Wert der NER mit CRFs in zwei verschiedenen Domänen (Romane und Zeitungsartikel) und verschiedenen maschinellen Lernverfahren

Ausblick

Es gibt eine Reihe von weiteren Optimierungsverfahren, die im Anschluss an die berichteten Experimente exploriert werden sollen. Wir haben bisher nur Lernverfahren für die NER in Romanen auf der Basis annotierter Textkorpora untersucht. Wir versprechen uns sowohl beim Erstellen eines Goldstandards, als auch bei dem erzielbaren F1-Wert der NER Verbesserungen durch die Integration von komplexeren regelbasierten Verfahren [Klügl et al. 2014] zur Information Extraction. Außerdem soll der Vermutung nachgegangen werden, dass die Erkennungsleistung durch Verwendung von Strategien der Domänenanpassung noch verbessert werden kann, wenn diese auf das vorhandene umfangreiche Korpus mit nicht-annotierten Daten angewandt werden [Qi Li 2012]. Außerdem sollen Alternativen zum

1 Unsere Implementierung des MaxEnt-Modells ist unter https://github.com/MarkusKrug/NERDetection/ zu finden. Sie ist so aufbereitet, dass sie mit dem DkPro-Framework kompatibel ist. Die Eingliederung dort soll demnächst folgen.

word2vec-Feature erprobt werden, die in NLP-Tasks gleichwertige Ergebnisse erbracht haben [Pennington 2014].

Literatur

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3, 993–1022.

Bohnet, B. (2010). *Very High Accuracy and Fast Dependency Parsing is not a Contradiction*. The 23rd Int. Conference on Computational Linguistics (COLING 2010), Beijing, China.

Chrupala, G. (2011). Efficient induction of probabilistic word classes with LDA. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 363-372.

Faruqui, M. and Pado. S. (2010) Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. Proceedings of Konvens 2010, Saarbrücken, Germany.

Finkel, F., Grenager, T. and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Fitschen, A., Schmid, H. and Heid, U. (2004) SMOR: A German computational morphology covering derivation, composition, and inflection. *Proceedings of the IVth International Conference on Language Resources and Evaluation* (LREC 2004), 1263–1266.

Klügl, P., Toepfer, M., Beck, P.D., Fette, G., Puppe, F. (2014) UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering* First View, 1–40 (2014). DOI 10.1017/S1351324914000114.

McCallum, A. MALLET: A Machine Learning for Language Toolkit. 2002.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Nadeau, D. and Sekine, S. (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes 30 (1), 3-26

Pennington, J, Socher, R. and Manning, C. (2014) Glove: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Qi Li (2012): Literature Survey. *Domain Adaption Algorithms for Natural Language Processing*. nlp.cs.rpi.edu/paper/qisurvey.pdf

Sang E. and Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (*4), 142-147.

Schmid, H. (1995) Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Sharnagat, R. (2014) *Named Entity Recognition: A Literature Survey*. Surveys of the Center for Indian Language Technology. http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf

Tkachenko, M. and Simanovsky, A. (2012) Named entity recognition: Exploring features. *Proceedings of KONVENS 2012*, 118-127.

Die explorative Visualisierung von Texten

Von den Herausforderungen der Darstellung geisteswissenschaftlicher Primär- und Annotationsdaten

Evelyn Gius und Marco Petris, Universität Hamburg

1. Zur Komplexität von Textdaten

Die Visualisierung von Textdaten ist innerhalb des Bereichs der Datenvisualisierung eine besondere Herausforderung, da es sich bei ihnen um unstrukturierte Daten handelt: Bevor man Textdaten visualisieren kann, muss aus ihnen eine Struktur abgeleitet werden. Hinzu kommt, dass Textdaten eine Vielzahl an Betrachtungsmöglichkeiten eröffnen, die durch die zahlreichen Bedeutungsdimensionen von Texten bedingt werden. Die einzelnen Dimensionen von Texten können durch Annotationen herausgearbeitet werden, wobei jede Annotationsschicht eine oder mehrere Dimensionen des Textes offenlegen kann. In diesem Sinne sind Textdaten also multidimensional. Insbesondere im Bereich der geisteswissenschaftlichen Textanalyse ist aufgrund des hermeneutischen Zugangs zu Texten auch in spezifischen Analysen nicht von vornherein klar, auf welche Weise Analyse und Interpretation zusammenhängen. Entsprechend muss eine sinnvolle Visualisierung von Textdaten im geisteswissenschaftlichen Kontext als exploratives Werkzeug zum Herausarbeiten möglicher Zusammenhänge fungieren können.[1]

Ein Blick in die einschlägige Literatur zur Datenvisualisierung zeigt, dass der Besonderheit von Textdaten häufig nicht Rechnung getragen wird. Zumindest scheint im traditionell mit Datenvisualisierung befassten informationswissenschaftlichen Bereich die Komplexität von annotierten Textdaten nicht immer im vollen Umfang wahrgenommen zu werden. So verweisen etwa Ward et al. (2010) in ihrer umfassenden Einführung zu Datenvisualisierung im Kapitel zu Textdaten auf drei mit Texten zusammenhängende Sucharten, die für die Anforderungen an die Visualisierung von Texten ausschlaggebend sind.[2] Die auf die Sucharten folgende Darstellung von Möglichkeiten der Textvisualisierung beschränkt sich allerdings auf die Darstellung von Texten und Korpora mit Metadaten (Erscheinungsjahr, Publikation o.ä.). Das Problem wird in der Zusammenfassung des Textvisualisierungskapitels offensichtlich: Die diskutierten Ansätze beträfen das "transforming unstructured text into structured data suitable for visualization and analysis" (Ward et al. 2010: 311). Die Option, dass der Text bereits mit Analysedaten in Form von

[1] Vorgehen, die darauf basieren, die Komplexität der Daten automatisiert zu reduzieren, erscheinen uns deshalb auch nicht geeignet für das beschriebene Problem (vgl. zu solchen Ansätzen z.B. Yang et al. 2003; Tatu et al. 2011).

[2] Typischerweise würden Zeichenketten in Form von Wörtern, Phrasen oder Themen gesucht, im Falle von partiell strukturierten Daten könnte außerdem nach Beziehungen zwischen Wörtern, Phrasen, Themen oder Dokumenten gesucht werden und schließlich ginge es in strukturierten Texten oder Textkorpora meistens um das Identifizieren von Mustern oder Auffälligkeiten innerhalb von Texten bzw. Dokumenten (vgl. Ward et al. 2010:291). Annotierte Textdaten fallen also potentiell unter die letzten beiden Fälle.

Annotationen angereichert sein könnte, wird nicht in Betracht gezogen. Das, obwohl in der Einleitung auf die drei Ebenen von Texten verwiesen – die lexikalische, die syntaktische und die semantische – und im Fall der syntaktischen Ebenen sogar explizit die Möglichkeit von Annotationen im Rahmen von *named entity recognition* (NER)-Prozessen erwähnt wurde (Ward et al. 2010:294).

2. Geisteswissenschaftliche Textdaten

In der stark geisteswissenschaftlich orientierten Position von Drucker (2014) werden hingegen die vielfältigen Interpretationsmöglichkeiten in den Fokus gerückt. Sie schreibt über die Visualisierung geisteswissenschaftlicher Interpretation: "The challenge is enormous, but essential, if the humanistic worldview, grounded in the recognition of the interpretive nature of knowledge, is to be part of the graphical expressions that come into play in the digital environment" (Drucker 2014: 136). Drucker geht es v.a. darum, die mit geisteswissenschaftlichen Analysen einhergehende Unsicherheit in der Darstellung des Wissens zu verdeutlichen, wobei sie sich nicht nur auf Texte beschränkt.

Was bedeutet das im Falle von Texten? Betrachten wir die Problematik an mit CATMA[3] annotierten Texten, die durch die flexiblen Annotationsmöglichkeiten des Werkzeugs exemplarisch für die große Bandbreite und gleichzeitig eingeschränkte Vorhersagbarkeit geisteswissenschaftlicher Analysen sind.[4] Für die Visualisierung von in CATMA erzeugten Text- und Annotationsdaten ist die von Drucker angesprochene Unsicherheit geringer, da es um die Analyse von Texten geht: Sie beschränkt sich auf (Text-)Interpretationen und liegt zudem nur in Form von Annotationen vor, die diese Unsicherheit konzeptionell durch entsprechende Tags fassen. Die Tags selbst beinhalten aber keine Unsicherheit, die für die weitere Analyse berücksichtigt werden muss.[5] Trotzdem ist Druckers Beobachtung zur Besonderheit geisteswissenschaftlicher Aussagen auch für unseren Zweck gültig und muss für die Visualisierung der Text- und Annotationsdaten berücksichtigt werden: "[…] we need to conceive of every metric 'as a factor of X', where X is a point of view, agenda, assumption, presumption, or simply a convention. By qualifying any metric as a factor of some condition, the character of the 'information' shifts from self-evident 'fact' to constructed interpretation motivated by a human agenda." (Drucker 2014:131). Aufgrund des freien Annotationsschemas, das CATMA zur Verfügung stellt, ist die Art der "Information", die die Annotationen enthalten, nämlich nicht über die vorliegenden Daten zugänglich: Man kann in CATMA genauso gut strukturelle Textmerkmale wie inhaltliche Aspekte annotieren und dafür eine eigene Annotationshierarchie entwickeln, deren Struktur zwar von der Anlage her hierarchisch ist, die aber prinzipiell überlappendes und widersprüchliches Markup zulässt.

[3] CATMA = Computer Aided Text Markup and Analysis, vgl. www.catma.de (gesehen am 10.11.2014).
[4] In CATMA können Texte anhand von frei gewählten Tags annotiert werden, die zu so genannten Tagsets zusammengefasst werden. Die so entstehende Taxonomie oder Systematik kann wiederverwendet werden. Die Texte und die Annotationen können außerdem mit einer umfangreichen Suchfunktionalität durchsucht und analysiert – und ggf. weiter annotiert werden. Zum damit außerdem verbundenen Konzept des hermeneutischen Markups vgl. Bögel et al. (im Erscheinen).
[5] vgl. dazu Jacke & Meister (2014).

3. Anforderungen an Visualisierung als Exploration

Aufgrund der nicht a priori eingrenzbaren Zwecke der Annotation und der Analyse muss die Visualisierung von Textdaten so generisch wie möglich gestalten werden. Nur so kann sie ohne ein tieferes Verständnis über die jeweils vorliegenden Text- und Annotationsdaten eingesetzt werden und einen Mehrwert bei der Analyse der Daten erzeugen.[6] Grundsätzlich konzipieren wir Visualisierungen deshalb ausgehend von der Frage, wie viele und welche Dimensionen der Daten dargestellt werden sollen.[7]

Für die Auswahl der Dimensionen stellt CATMA über die Struktur der Ergebnismenge der Abfragen folgende Kategorien zur Verfügung:

- Metadaten der Dokumente (z.B. Titel, Autor, etc.),
- Tag bzw. Typ der Annotation,
- Properties der Annotation und die für den annotierten Text vergebenen Werte,
- annotierter Text,
- Position im Text (via Zeichen-Offset),
- Textkontext des annotierten Textes (variable Anzahl von Token),
- Vorkommenshäufigkeit des annotierten Textes,
- Vorkommenshäufigkeit der Annotation,
- weitere berechnete Kategorien, wie der z-Faktor oder der TF-IDF

Neben dem generischen Zugang über die Dimensionen der Daten muss auch ein Mechanismus zur Verfügung gestellt werden, mit dem der Zweck einer spezifischen Analyse in der Visualisierung der Daten herausgearbeitet werden kann – und der die erzeugten Visualisierungen als explorative Heuristik nutzbar macht. Für die damit zusammenhängenden spezifischen Erkenntnisinteressen werden deshalb zusätzliche Anpassungsmöglichkeiten in Form von wählbaren Parametern eingeführt. Diese sollen typische Varianten abfangen, wie etwa die Frage, ob die Häufigkeit einer Annotation oder aber der annotierte Textumfang dargestellt werden soll, wie mit überlappenden Annotationen verfahren werden soll (soll etwa eine zweifach annotierte Stelle zweimal oder nur einmal gezählt bzw. dargestellt werden?) oder ob die Struktur der Tagsets so ist, dass sich Tags auf derselben Hierarchieebene gegenseitig ausschließen oder ob sie sich ergänzen können.[8]

4. Beispiele

Die oben angestellten Überlegungen sollen an folgenden Beispielen demonstriert werden. Datengrundlage ist das in Gius (2013) beschriebene und analysierte umfangreich

[6] Dies gilt nicht für die Visualisierung zu Demonstrations- bzw. Kommunikationszwecken – also von Daten, die bereits analysiert und interpretiert wurden.

[7] Mit “Dimensionen” sind also nicht räumliche Dimensionen gemeint. Dies wird allerdings von einigen gängigen Ansätzen zur Visualisierung mehrdimensionaler Daten angenommen, die Textdaten nur als einen – eindimensionalen – Datentyp betrachten und allgemeine Modelle entwickeln (vgl. etwa Shneiderman 1996).

[8] Auch hier unterscheidet sich der vorgestellte Ansatz durch seinen Fokus auf die Spezifik von Texten wieder deutlich von Ward et al. (2010) oder Shneiderman (1996), die so genannte *tasks* als Basis für zusätzliche explorative Funktionalitäten betrachten.

annotierte Korpus.[9] Die hier nur kurz beschriebenen Visualisierungen werden ebenso wie eine Reihe weiterer Visualisierungen in CATMA zur Verfügung gestellt. Ihre Funktionen und der damit verbundene explorative Gewinn werden im Rahmen des Vortrags näher vorgestellt werden.

Interaktive TreeMap

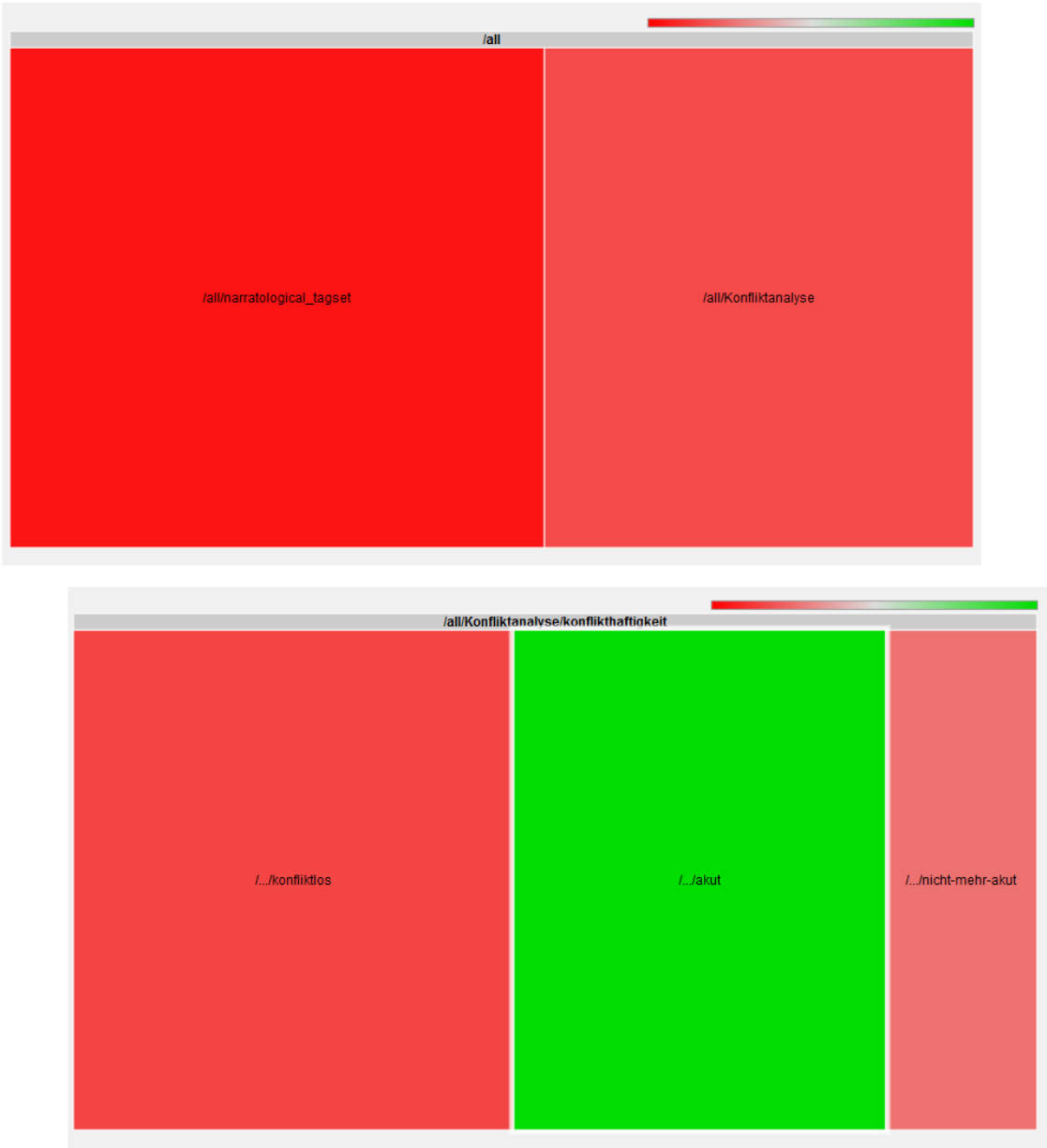


Abbildung 1: Interaktive TreeMap[10]

[9] Das Korpus besteht aus 24 Texten mit insgesamt 86.246 Wörter, die auf etwa 150 als Tags eingeführte narratologische Konzepte untersucht und mit insgesamt 24.347 Annotationen versehen wurden.

[10] Erstellt auf Basis von Google Charts: https://developers.google.com/chart/interactive/docs/gallery/treemap?hl=de (gesehen am 10.11.2014).

Das erste Beispiel ist eine interaktive TreeMap. Jede einzelne Sicht zeigt zwei Dimensionen: (1) Die Vorkommenshäufigkeit der Abfrageergebnisse (Tags, Wörter, o.ä.) als Größe des zugehörigen Rechtecks und (2) die durchschnittliche annotierte Textmenge als Farbintensität auf einer Skala von rot (weniger) bis grün (mehr). Die interaktive Komponente ermöglicht das Ergründen einer dritten Dimension: Durch Klicken der einzelnen Rechtecke kann man durch (3) die Hierarchie des Tagsets navigieren.
Abbildung 1 zeigt zwei Ebenen: Rechts die höchste Ebene mit den beiden Top-Level Tags "narratological_tagset" und "Konfliktanalyse" und links die Ebene 1 den Zweig entlang dem Tag "Konfliktanalyse" mit den Tags der darunter liegenden Ebene. Gezeigt wird also die Verteilung von Vorkommenshäufigkeit und annotierter Textmenge für die Hierarchieebene. Für die dargestellte Datenbasis ist das insofern interessant, als hier die Konflikthaftigkeit von Erzählungen bzw. die als konflikthaft oder konfliktlos annotierten Passagen dargestellt werden. Für die Analyse des Korpus ist sowohl die Frage nach der Häufigkeit, in der konflikthafte Passagen auftauchen (sie unterbrechen nämlich von den Erzählerinnen eigentlich als konfliktlos deklarierte Erzählabschnitte und deshalb ist ihre Anzahl relevant), als auch die reine Textmenge, die sie umfassen (wird ausgiebiger über konfliktlose oder über konflikthafte Situationen erzählt?), interessant. Die Visualisierung als dreidimensionale TreeMap ermöglicht es, die beiden Betrachtungsweisen – Anzahl vs. Textmenge – überblickshaft in Beziehung zu setzen und dabei durch die hierarchisch angeordneten Tags zu navigieren, also zusammengefasste und detailliertere Perspektiven zu wählen.

Small Multiples

Das zweite Beispiel (vgl. Abbildung 2) zeigt die Vorkommenshäufigkeit von zwei Annotationen (Wiedergabe von mentalen Prozessen und Wiedergabe von Rede) im Textverlauf bei neun Texten des Korpus. Die Vorkommenshäufigkeit wird auf der y-Achse und der Textverlauf in 10%-Schritten auf der x-Achse dargestellt. Für jeden ausgewählten Text wird jeweils ein Koordinatensystem als dritte Dimension erstellt, in dem die Annotationen als farbige Linien abgebildet werden.[11]

Diese Darstellung ermöglicht eine explorative Betrachtung der Verteilung der beiden annotierten Phänomene in den Einzeltexten und einen ersten Überblick über mögliche Muster im gesamten Korpus. Für eine weitere Analyse können auffällige Stellen – wie etwa besondere Häufigkeiten in einem Textabschnitt oder der Wechsel von dominierender Redewiedergabe zu dominierender Wiedergabe von mentalen Prozessen – genauer betrachtet werden: Das Anklicken der entsprechenden Punkte im Graphen erzeugt eine KWIC(=*KeyWord In Context*)-Anzeige der Annotationen im betreffenden Textabschnitt, von denen aus wiederum durch Klicken in den Volltext gesprungen werden kann.

[11] Die Darstellung als Linie wurde aus Gründen der Übersichtlichkeit gewählt, mathematisch gesehen handelt es sich natürlich um diskrete Werte.

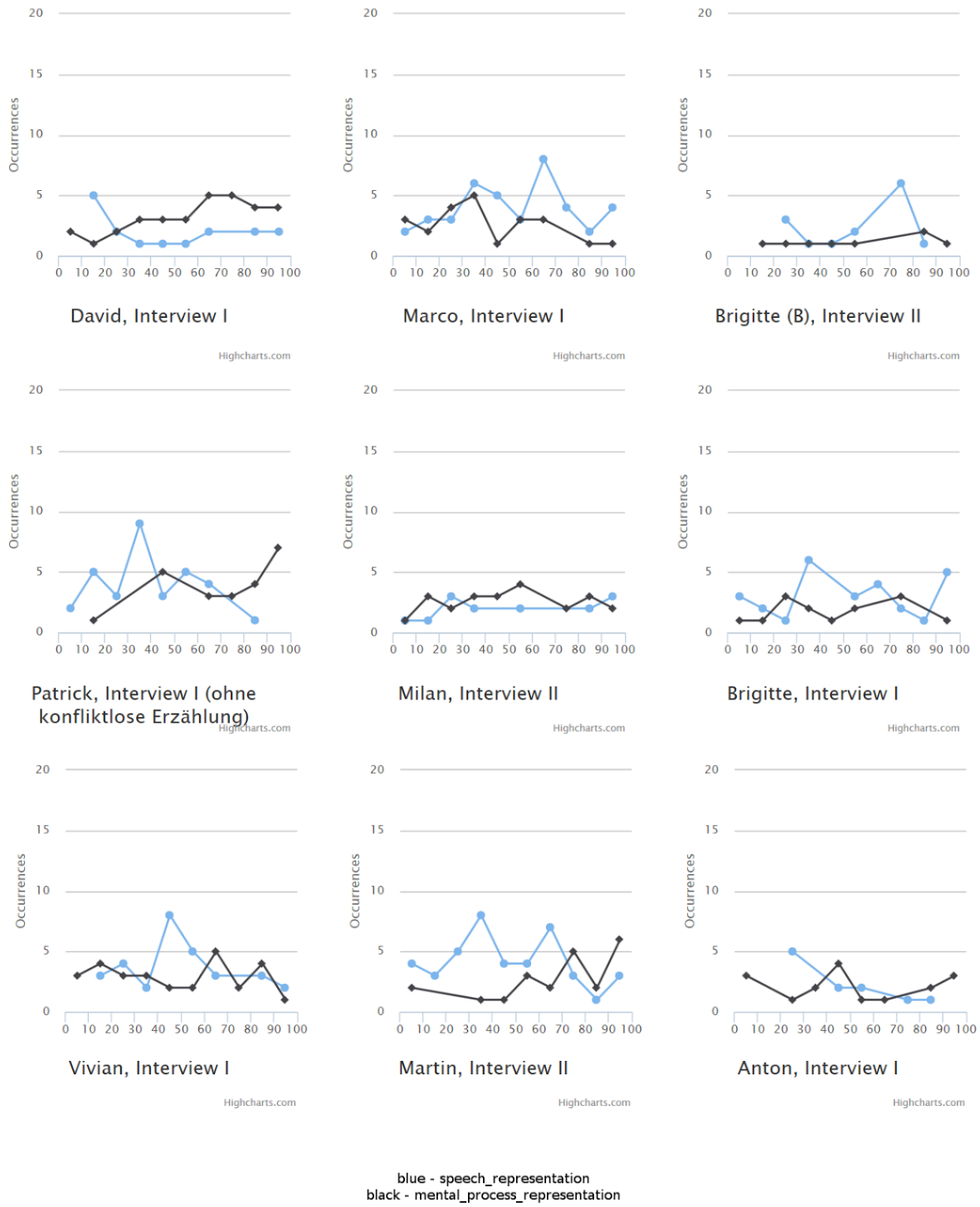


Abbildung 2: Small Multiples: Distributionsgraphen[12]

[12] Erstellt auf Basis von Highcharts: http://www.highcharts.com/ (gesehen am 10.11.2014).

5. Ausblick

Für das dargestellte Korpus sind oben beschriebene Visualisierungen von großem Gewinn. Sie ermöglichen einen Überblick über die Daten, für die nicht bereits bei der Annotation festgelegt wurde, welche Dimensionen genauer betrachtet und in Zusammenhang gebracht werden müssen. Dadurch sind die für die Analyse der Daten von großem Nutzen. Inwiefern sich nach in diesem Beitrag vorgestellten Überlegungen entwickelte Visualisierungen auch systematisch als heuristisches Werkzeug eignen und ob sie sich als Alternative gegen generelle Datenvisualisierungen durchsetzen können, ist zum momentanen Zeitpunkt allerdings noch nicht abschätzbar. Neben der Vieldimensionalität der Daten ist dafür insbesondere die Frage der explorativen Funktion der Visualisierungen zentral: Reichen (1) die dargelegte Aufschlüsselung der Analysen nach ihrer Dimensionalität und (2) die Explorationsmöglichkeit durch einstellbare Parameter aus, um Visualisierungen zu erzeugen, die systematische Rückschlüsse auf die dargestellten Daten und Strukturen zulassen – und nicht nur assoziative Denkanstöße zu liefern?

Dies wird in breit angelegten Nutzerstudien zu ergründen sein, ebenso wie untersucht werden muss, ob und auf welche Weise die in den Visualisierungen zum Einsatz kommenden visuellen Metaphern den Verstehensprozess beeinflussen.

Referenzen

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, and Jannik Strötgen. “Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristics of Narrative.” *DHCommons Journal*, im Erscheinen.

Drucker, Johanna. *Graphesis: Visual Forms of Knowledge Production*. MetaLABprojects. Cambridge, Massachusetts: Harvard University Press, 2014.

Gius, Evelyn. “Erzählen Über Konflikte. Eine Computergestützte Narratologische Untersuchung von Narrativen Interviews Zu Arbeitskonflikten.” Dissertation, Universität Hamburg, 2013.

Jacke, Janina, und Jan Christoph Meister. „Pushing Back the Boundary of Interpretation: Concept, Practice and Relevance of a Digital Heuristic“. In *Digital Humanities 2014 – Book of Abstracts*, 264–66. Lausanne, 2014.

Shneiderman, Ben. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” In *In IEEE Symposium on Visual Languages*, 336–43, 1996.

Tatu, Andrada, Georgia Albuquerque, Martin Eisemann, Peter Bak, Holger Theisel, Marcus Magnor, and Daniel Keim. “Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data.” *IEEE Transactions on Visualization and Computer Graphics* 17, no. 5 (May 2011): 584–97.

Ward, Matthew, Georges G. Grinstein, and Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, Mass: A K Peters, 2010.

Yang, J., M. O. Ward, E. A. Rundensteiner, und S. Huang. „Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets“. In *Proceedings of the*

Symposium on Data Visualisation 2003, 19–28. VISSYM ’03. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2003.

Abstract
DHd2015 „Von Daten zu Erkenntnissen“
Dr. Angelika Zirker (Eberhard Karls Universität Tübingen) angelika.zirker@uni-tuebingen.de
Fabian Schwabe (Eberhard Karls Universität Tübingen) fabian.schwabe@uni-tuebingen.de

Vortrag

Theorie und Praxis der erklärenden Annotation im Kontext der Digital Humanities

Der Vortrag geht aus einem aktuellen Forschungsprojekt an der Eberhard Karls Universität Tübingen hervor, das von Prof. Dr. Matthias Bauer und Dr. Angelika Zirker (beide Literaturwissenschaft Anglistik) initiiert wurde (www.annotating-literature.org) und mit dem eScience Center der Universität kooperiert. Es befasst sich mit der erklärenden, interpretatorischen Annotation literarischer Texte unterschiedlicher Gattungen und verfolgt drei Ziele: (1) Die erläuternde Annotation vorwiegend literarischer Texte auf eine theoretische Basis zu stellen, um daraus Praxismodelle abzuleiten, die auch für nichtliterarische Texte von Belang sind; (2) die bislang eher geringe Verankerung der erläuternden Annotation in den Digital Humanities konzeptuell und durch Beispiele voranzutreiben und damit sowohl eines der Anwendungsfelder der Digital Humanities zu erweitern als auch neue Anwendungsmethoden zu generieren; (3) bildungswissenschaftliche Fragestellungen in die literaturwissenschaftliche Arbeit einzubeziehen, um den Nutzen von Annotationen besser erforschen zu können und um zu klären, ob und wie das Textverständnis des Lesers verbessert werden kann.

Der Schwerpunkt des Vortrags wird im zweiten skizzierten Bereich des Projekts liegen, weil sich hier eine Anbindung an das Tagungsthema „Von Daten zu Erkenntnissen“ in besonderer Weise anbietet: die Verankerung der erläuternden Annotation literarischer Texte in den Digital Humanities setzt bei genau der Frage an, die sich mit dem Mehrwert digitaler Methoden bzgl. der Erkenntnisprozesse in den Geisteswissenschaften befasst. Die Fragestellung schließt zunächst an den Bereich der Hermeneutik an, vor allem an die Definition von Annotation im texterläuternden Sinn und ihrem Verhältnis zur Interpretation. Die hermeneutische Theorie ist (ebenso wie die Textlinguistik) auch dann relevant, wenn es um Fragen der Auswahl zu annotierender Aspekte eines Textes geht und um das Verhältnis von Teil und Ganzem. Annotationen sind in der Regel auf einen bestimmten Textteil oder ein Textelement bezogen; die theoretische Bestimmung ihrer Funktion für die Erklärung bzw. das Verständnis des Gesamttextes impliziert Fragen nach der Hierarchisierung und Systematisierung von Information. So gewinnt das fundamentale Problem des hermeneutischen Zirkels in der Annotation besondere Virulenz, weil Annotationen von Teilaspekten ein Gesamtverständnis des Textes voraussetzen, welches sie erst ermöglichen. Ein weiterer wichtiger Aspekt der Theoriebildung im Bereich erläuternder Annotation ist die systematische Beschreibung der Notwendigkeit von Annotationen. Welche Elemente eines Textes bedürfen der Erläuterung und warum? Ein weiterer Aspekt der Reflexion von Annotation betrifft die Funktion der Zusammenarbeit bei der Erstellung von erläuternden Annotationen. Das digitale Medium erlaubt und fördert die Kollaboration; zugleich sind Erläuterungen größerer Textkorpora und komplexe Annotationen nicht von einzelnen zu leisten. Was bedeutet es aber, wenn Texte gemeinschaftlich erschlossen werden? Hat diese Vorgehensweise Auswirkungen auf unser Verständnis der Bedeutung von Texten und wie wird Erläuterungsautorität verhandelt?

Diese theoretischen und inhaltlichen Erwägungen werden im Vortrag anhand aktueller Ergebnisse aus dem laufenden Projekt diskutiert, indem Annotationen zu englischsprachigen Gedichten aus dem Ersten Weltkrieg exemplarisch präsentiert und analysiert werden. Dabei wird ein mehrschichtiges Annotationskonzept verwendet, das gemeinsam mit Spezialisten aus den Digital Humanities entwickelt wurde und das von basalen Definitionen (etwa Worterklärungen) innerhalb der literarischen Texte bis hin zu Interpretationsvorschlägen reicht, d.h. das verschiedenen Ebenen der Annotation im Sinne einer Komplexitätssteigerung unterscheidet. An dieser Stelle kommen neben Fragen der Hermeneutik und Textinterpretation nun auch die Techniken und Werkzeuge der Digital Humanities zum Tragen, die ein solch vielschichtiges Annotationskonzept visualisierbar, nutzbar und sinnvoll erfahrbar machen. Um dies zu realisieren, kommt mit der TEI-XML-Kodierung ein etabliertes und standardisiertes Werkzeug zum Einsatz. Ziel der Anwendung dieser technischen Lösung muss der höchstmögliche Gewinn für den Rezipienten sein. Dabei ist zu bedenken, dass die digitalen Medien in dieser Hinsicht nicht per se eine

Abstract
DHd2015 „Von Daten zu Erkenntnissen“
Dr. Angelika Zirker (Eberhard Karls Universität Tübingen) angelika.zirker@uni-tuebingen.de
Fabian Schwabe (Eberhard Karls Universität Tübingen) fabian.schwabe@uni-tuebingen.de

Verbesserung der Annotationspraxis mit sich bringen, ebenso wenig wie Kollaboration nicht automatisch einen Qualitätsgewinn bedeutet. Die Vorteile der digitalen Annotation (insbesondere gegenüber dem gedruckten Buch) und der mit ihr einhergehenden Möglichkeiten der Zusammenarbeit kommen nur bei einer durchdachten Konzeption und der Einbeziehung aller Risiken des Informationsverlusts durch Informationsfülle zur Geltung. Gerade weil auf der Grundlage digitaler Medien eine prinzipiell offene und endlose Annotierung möglich ist, geht es also auch um grundsätzliche Elemente der Qualitätssicherung, die ebenfalls in die Theorien und Modelle der erläuternden Annotation einzubeziehen sind.

Der Aufwand der digitalen Kodierung sowie die Dokumentation des verwendeten Schemas übersteigen den einer analogen Annotation um ein Vielfaches. Aus diesem Grund kommt einer nachhaltigen Verfügbarmachung und Sicherung der entstandenen Daten eine immense Bedeutung zu. Wird die inhaltliche Annotation als Prozess verstanden, der sich immer wieder, abhängig von Personen, Forschungsfrage oder Zeitgeist, erneuern und erweitern lässt, verstärkt sich nochmals die Bedeutung der Nachhaltigkeit. Um diesem Umstand zu begegnen werden gemeinsam mit dem eScience-Center der Universität Strategien entwickelt, die diesen Anforderungen Rechnung tragen. Wir möchten in unserem Vortrag die Perspektiven von Theorie und Praxis sowohl aus literaturwissenschaftlicher Sicht wie aus Perspektive der Digital Humanities verbinden und anhand erster Projektergebnisse illustrieren, welche neuen Möglichkeiten sich dadurch erschließen, uns aber gleichzeitig auch der Diskussion inhärenter Probleme der Verbindung von Theorie und Praxis stellen.

Das Zusammenspiel interpretativer und automatisierter Verfahren bei der Aufbereitung und Auswertung mündlicher Daten

Ein Fallbeispiel aus der angewandten Wissenschaftssprachforschung

Cordula Meißner (Universität Leipzig)
Franziska Wallner (Universität Leipzig)

Obwohl die Erforschung der Wissenschaftssprache auch im Bezug auf das Deutsche in den letzten Jahren verstärkt Beachtung gefunden hat, stehen selbst für die geschriebene Modalität der Wissenschaftssprache nur in sehr begrenzten Umfang elektronisch verfügbare Datensammlungen als empirische Grundlage für diesbezügliche Untersuchungen zur Verfügung. Für die gesprochene Wissenschaftssprache fehlten sie lange Zeit vollkommen. Sowohl für den Bereich der Diskurs- und Variationsforschung als auch für den Bereich der Sprachlehr- und lernforschung stellt die Untersuchung der gesprochenen Wissenschaftssprache auf breiterer empirischer Basis ein noch weitgehend unbearbeitetes Desiderat dar.

Mit dem an der Universität Leipzig (Herder-Institut), der Aston University, Birmingham und der Universität Wrocław erarbeiteten Korpus GeWiss („Gesprochene Wissenschaftssprache kontrastiv – Deutsch im Vergleich zum Englischen und Polnischen") wurde 2013 für die gesprochene Wissenschaftssprache exemplarisch eine flexibel nutzbare Korpusressource der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt (vgl. Fandrych/Meißner/Slavcheva 2012, 2014). Sie umfasst ca. 120 Aufnahmestunden von gesprochener deutscher, englischer und polnischer Wissenschaftssprache (mehr als 1 Mio Token, die als Transkripte vorliegen und mit Audiofiles synchronisiert analysierbar sind). Es handelt sich um ein Vergleichskorpus, welches zwei zentrale Genres der mündlichen Wissenschaftskommunikation umfasst – Vorträge/Referate und Prüfungsgespräche. Datengrundlage sind dabei zum einen Aufnahmen von L1-Sprecher/inne/n der drei Vergleichssprachen, zum anderen deutschsprachige Realisierungen dieser Genres von L2-Sprecher/inne/n in Deutschland, Großbritannien und Polen.

Ziel des aktuellen Folgeprojekts „Gesprochene Wissenschaftssprache digital" ist die Optimierung der Nutzungsmöglichkeiten des Gewiss-Korpus und die Erprobung weiterer methodischer Möglichkeiten zur Auswertung und Analyse der Daten. Im Fokus stehen dabei neben den Realisierungsmöglichkeiten wissenschaftlicher Sprachhandlungen wie beispielsweise Diskurskommentierungen, Verweisen und Zitaten auch der Wortschatz der allgemeinen gesprochenen Wissenschaftssprache sowie lexikalische, morphologische und

syntaktische Besonderheiten der gesprochenen Wissenschaftssprache. Der Vortrag stellt anhand eines Beispiels für die aktuelle Arbeit mit den Korpusdaten vor, wie hierbei interpretative und formbasiert automatisierbare Ansätze zusammenwirken können.

Der erste Teil des Vortrags gibt einen Überblick über den Aufbau und das Design des GeWiss-Korpus und dokumentiert die aktuell erarbeiteten Aufbereitungsschritte, welche die orthografische Normalisierung, das Wortarten-Tagging, die Lemmatisierung sowie die pragmatische Annotation umfassen. Sowohl die Verfügbarkeit einer solchen Datenbasis zur gesprochenen Wissenschaftssprache als auch die genannten Formen der Aufbereitung eröffnen neue methodische Zugänge zum Untersuchungsgegenstand. In seinem zweiten Teil wird der Vortrag zeigen, wie diese Zugänge im Forschungsprozess genutzt werden können. Er will damit auch einem Beitrag zur Methodenreflexion leisten im Hinblick auf das Zusammenspiel von interpretativ-manuellen und formbasiert-automatisierbaren Verfahren bei der Aufbereitung und Auswertung von Korpusdaten.

Exemplarisch sollen hierfür die Möglichkeiten zur Beschreibung der wissenschaftssprachlichen Handlung des Diskurskommentierens durch manuelle Annotation und darauf aufbauende korpusmethodische Ermittlung „guter Kandidaten" betrachtet werden.

Zur korpusbasierten Untersuchung pragmatischer Phänomene wurde bislang entweder der Weg der manuellen Annotation (vgl. Baur et al. 2013, Maynard/Leicher 2007, Alsop/Nesi 2012) oder der Weg eines automatischen Zugriffs auf der Formebene über konkrete Lexeme oder datengeleitet ermittelte N-Gramme (vgl. z.B. Scharloth/Bubenhofer 2012, Rühlemann 2010) verfolgt. Der erste Weg ermöglicht eine funktional orientierte, erschöpfende Erfassung des Phänomens. Eine derartige Korpusaufbereitung ist jedoch mit einem hohen Maß an Zeitaufwand verbunden und erfordert mehrstufige Korrekturdurchgänge sowie eine intensive Abstimmung unter den Annotierenden. Der zweite Weg ermöglicht zwar eine schnelle Datengewinnung, beschränkt die Ergebnisse jedoch auf angenommene, zuvor ausgewählte Formmerkmale oder datengeleitet ermittelte rekurrierende Wortfolgen, die anschließend der funktionalen Interpretation bedürfen.

Im Vortrag wird exemplarisch anhand einer Analyse zur sprachlichen Handlung der Diskurskommentierung ein Ansatz vorgestellt, der beide Wege verbindet: Ausgehend von den im GeWiss-Korpus für ein Teilkorpus von Konferenzvorträgen annotiert vorliegenden Diskurskommentierungen (vgl. Fandrych 2014) werden über korpuslinguistische Analysen (u.a. Keyword-, N-Gramm- und Kookkurrenzanalysen) typische Formmerkmale dieser Sprachhandlung ermittelt. Daraus wird eine Suchabfrage gebildet, mit der in nicht-annotierten Korpora nach „guten Kandidaten" für die betrachtete sprachliche Handlung gesucht werden

kann. Es wird somit durch ein ursprünglich interpretatives Herangehen an die Daten die Grundlage für einen formbasiert-automatisierbaren Zugriff geschaffen, der wiederum Datensamples für weiterführende interpretative Analysen erschließt. Es findet dabei ein Zusammenspiel zweier Ansätze statt: Das interpretative Vorgehen, nähert sich den Daten über die Bedeutungs- bzw. Funktionsebene und ermittelt deutend-verstehend Vorkommen der sprachlichen Handlung im Äußerungskontext. Beim formbasiert-automatisierten Herangehen hingegen erfolgt der Zugriff über die Form, d.h. über Oberflächenmerkmale (z.B. Schlüsselwörter). Die sprachliche Handlung wird dann durch die automatische Abfrage dieser Merkmale im Korpus ermittelt.

Wie wirken diese Herangehensweisen bei der dargestellten Untersuchung von Diskurskommentierungen zusammen? Es erfolgt zunächst die Ermittlung von Belegen für die Sprachhandlung durch Interpretation sprachlicher Äußerungen im Kontext, wobei alle interpretativ zugänglichen Belegstellen im Bezugskorpus identifiziert und manuell annotiert werden. Die anschließende korpuslinguistische Analyse dieser Belege identifiziert wiederkehrende Oberflächenmerkmale, die typisch für die Realisierung der Sprachhandlung sind. Diese werden in einen Suchausdruck zusammengeschlossen und erlauben so den formbasiert-automatisierten Zugriff auf Sprachdatensequenzen, die diese Merkmale tragen. Dadurch können aus nicht-annotierten Vergleichskorpora „gute Kandidaten" für die sprachliche Handlung ermittelt werden, d.h. sprachliche Äußerungen, die über die für diese Handlung typischen Merkmale verfügen. Durch die Evaluation und quantitative Auswertung der so zugänglichen Belege lassen sich Hypothesen über die Realisierung der Sprachhandlung im untersuchten Korpus bilden.

Welchen Beitrag leisten hierbei interpretativer und formbasiert-automatisierter Zugriff? Ohne eine ursprüngliche interpretative Ermittlung von Belegstellen wäre die sprachliche Handlung in den Korpusdaten nicht zugänglich. Vorhandene Belegstellen können durch diese Herangehensweise zudem über unterschiedliche Realisierungsformen hinweg vollständig erfasst werden. Durch den interpretativen Zugriff wird somit die Vielfalt aller möglichen Realisierungen erschlossen. Der formbasiert-automatisierbare Zugriff hingegen macht formale Muster explizit, die beim interpretativen Zugriff aufgrund der gewonnenen Datenvielfalt implizit bleiben. Er erlaubt somit auch eine Bündelung der Beschreibung der Sprachhandlung im Hinblick auf typische Formmerkmale. Darüber hinaus macht er einen Teil der Realisierungen formal erfassbar und abfragbar. Formbasiert-automatisiert können aus den annotierten Daten *beste Beispiele*, aus nicht-annotierten Vergleichskorpora *gute Kandidaten* ermittelt werden, anhand derer sich Hypothesen über die Realisierung der Sprachhandlung in der Sprachverwendungskonstellation dieser Korpora aufstellen lassen. Die Ergebnisse des formbasiert-automatisierten Zugriffs erfordern eine interpretative Weiterbearbeitung in Form

der Evaluation der Belege und ihrer Einordnung im Hinblick auf die durch die ursprüngliche interpretative Erfassung bekannte Varianz in der Realisierung der Sprachhandlung. Interpretativer und formbasiert-automatisierbarer Zugang bedingen sich somit gegenseitig: ohne ursprünglichen interpretativen Zugriff wäre eine Formalisierung nicht möglich, ohne diese nicht ein Zugriff auf Belege mit handlungstypischen Merkmalen für eine interpretative Auswertung höherer Stufe.

Im Vortrag wird der Ansatz anhand einer Analyse zur Realisierung von Diskurskommentierungen in Konferenzvorträgen von polnischen Deutsch-L2-Sprecher/inne/n exemplarisch illustriert. Das Vorgehen und der ermittelte Suchausdruck sind jedoch auch auf weitere verwandte Sprachverwendungskonstellationen (z.B. studentische Seminarreferate) übertragbar. Die ermittelten Belege geben Aufschluss über das Realisierungsspektrum der Handlung in anderen, sich sprachlich, kompetenz- und diskursartbezogen unterscheidenden Konstellationen. Die Ergebnisse eröffnen dadurch weitere Möglichkeiten für kontrastive und kompetenzorientierte Untersuchungen: Aus den Treffern, die der Suchausdruck (als Muster einer fachkompetenten L1-Realisierungsform) in Daten von fachkompetenten L2-Sprecher/inne/n und in Daten von fachlichen Novizen erzielt, lassen sich Schlüsse über Variation und Kernmerkmale der Handlung in diesen sprachlichen Konstellationen ziehen. Die Verbindung aus interpretativem und datenorientiertem Zugriff ermöglicht es zudem, aus den manuell annotierten Daten typische Beispiele zu extrahieren, die über häufige Formeigenschaften verfügen. Dies bietet eine weitere Anwendungsmöglichkeit für die Fremdsprachenvermittlung.

Das vorgestellte Fallbeispiel zeigt somit, dass sich durch das Zusammenwirken interpretativer und formbasiert-automatisierbarer Ansätze eine umfassendere Beschreibung der betrachteten Sprachhandlung gewinnen lässt, die formale Muster explizit macht, diese zur Gewinnung von besten Beispielen und Hypothesen über die Realisierung in verwandten Sprachverwendungskonstellationen nutzt und sie gleichzeitig in Bezug auf Varianz und die Grenzen der formalen Erfassbarkeit einordnen kann.

Literatur:

Alsop, Sian/Nesi, Hilary (2012): Annotating a corpus of spoken English: the Engineering Lecture Corpus (ELC). In: Mello, Heliana/Pettorino, Massimo/Raso, Tomaso (Hgg.): *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*. Firenze: Firenze University Press, 58 – 62.

Baur, Benedikt/Gräfe, Karen/Lange, Daisy/Schmidt, Julia (2014). *Dokumentation zur Annotation der Diskurskommentierungen*. Abrufbar unter: https://gewiss.uni-leipzig.de [Stand: 06.11.2014].

Fandrych, Christian (2014): Metakommentierungen in wissenschaftlichen Vorträgen. In: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hgg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation), 95 – 111.

Fandrych, Christian/Meißner, Cordula/Sadowski, Sabrina/Wallner, Franziska (in Vorb.): *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora*. Tübingen: Stauffenburg.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hgg.) (2014): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012): The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In: Schmidt, Thomas/Wörner, Kai (Hgg.): *Multilingual corpora and multilingual corpus analysis*. Amsterdam: Benjamins, 319 – 337. (=Hamburg Studies in Multilingualism 14).

Maynard, Carson/Leicher, Sheryl (2007): Pragmatic Annotation of an Academic Spoken Corpus for Pedagogical Purposes. In: Fitzpatrick, Eileen (Hg.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodophi, 107 – 116.

Meißner, Cordula (in Vorb.): Die Realisierung mündlicher wissenschaftssprachlicher Handlungen im Deutschen als L1 und L2. Eine gebrauchsbasierte Analyse. In: Kontutytė, Eglė / Žeimantienė, Vaiva (Hgg.): *Sprache in der Wissenschaft. Germanistische Einblicke*. Duisburger Arbeiten zur Sprach- und Kulturwissenschaft. Frankfurt: Peter Lang.

Rühlemann, Christoph (2010): What can the corpus tell us about pragmatics? In: O'Keeffe, Anne / McCarthy, Michael (Hgg.): *The Routledge handbook of corpus linguistics*. London/New York: Routledge, 288 – 301.

Scharloth, Joachim/Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik. Korpusvergleiche als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hgg.): *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analyse*. Berlin / N.Y.: de Gruyter, 195 – 230.

Erfahrungsberichte aus zweiter Hand:
Erkenntnisse über die Autorschaft von Arztbewertungen in Online-Portalen

Michaela Geierhos & Frederik S. Bäumer, Universität Paderborn

Für den Internetnutzer[1] entstehen immer mehr Möglichkeiten, Bewertungen über eine Vielzahl an Produkten (z. B. Amazon Reviews), Leistungen (z. B. jameda) und Erlebnissen (z. B. TripAdvisor) abzugeben. Diese suchen Bewertungsplattformen auf, um aktiv ihre Erfahrungen mit Dienstleistungen wie z. B. im Versandhandel, mit Hotelurlauben oder von Arztbesuchen mit anderen zu teilen. Häufig bestehen diese Bewertungskommentare aus Freitexten (sog. User Generated Content), die in Struktur und inhaltlicher Fokussierung deutlich voneinander abweichen. Vor der wissenschaftlichen Interpretation dieser Erfahrungsberichte steht jedoch der Prozess der Datenerhebung – die Entscheidung darüber, was die eigentlichen *Daten* für bestimmte Forschungszwecke sind. Was wir als Computerlinguisten unter empirischen Daten verstehen, und wie wir und andere Disziplinen den Prozess der Datenbeschaffung bezeichnen (z. B. Datenakquise vs. Datenkonstruktion), unterscheidet sich stark davon, wie Daten z. B. in der Soziologie erhoben und wie dieser Prozess dort verstanden wird.

Daten – Informationen – Erkenntnisse

Daten sind für uns nur Angaben über Sachverhalte und Vorgänge, die in einer Datenbank gespeichert werden. So liegen dieser Studie zufällig ausgewählte Datensätze der Portale jameda und DocInsider aus den Jahren 2009 bis 2015 zugrunde. Das Korpus umfasst 860.000 individuelle Erfahrungsberichte, die allesamt nach einem Arztbesuch in Deutschland verfasst und online gestellt wurden. Jede Arztbewertung besteht dabei aus einer Überschrift, dem eigentlichen Text und bis zu 17 verschiedenen numerischen Bewertungskategorien (u.a. Behandlung, Vertrauensverhältnis, Wartezeit, Barrierefreiheit). Bei der computergestützten Datenakquise wurden diese authentischen, online-verfügbaren Datensätze (sog. empirische Daten) originalgetreu zur weiteren Auswertung lokal bereitgestellt.

Erst durch die kognitive Verarbeitung des Lesers können aus diesen Daten *Informationen* werden. Hierfür setzt die Computerlinguistik Methoden zur Informationsextraktion ein, um sowohl Informationen aus den Meta-Daten (z. B. Erstellungsdatum, Name und Alter der Autoren) als auch aus dem Inhalt von Datenformaten (z. B. HTML) zu gewinnen. Anfangs wird definiert, welche Arten von Informationen extrahiert werden sollen. Auf diese Weise ist es möglich, aus Online-Arztbewertungen beispielsweise auszulesen, wer behandelt wurde, was die Symptome waren oder die Krankheitsgeschichte ist, wie (un)zufrieden derjenige war (z. B. „lange Wartezeit“; „kompetente Behandlung“), und ob er den Arzt weiterempfehlen würde. Hierzu müssen einerseits Zuordnungsregeln formuliert werden, die darüber entscheiden, welche „Daten“ welche *Informationen* repräsentieren, wie beispielsweise „Kassenpatient“ → *Gesetzliche Krankenversicherung*. Andererseits ist zu beachten, dass der quantitative Teil der Daten in Abhängigkeit der Ursprungsplattform (jameda oder DocInsider) anders zu interpretieren ist. Während jameda mit Schulnoten arbeitet, setzt DocInsider Sterne zur Bewertung der unterschiedlichen Aspekte beim Arztbesuch ein. Aufgrund dessen ist eine portalspezifische Abbildungsfunktion von Zahlen auf Denotationen zu definieren, so dass sichergestellt wird, dass eine 1,0 bei jameda der Bedeutung von „sehr gut“ entspricht, wohingegen derselbe numerische Wert bei DocInsider ein „ungenügend“ repräsentiert.

Anschließend können die Bewertungstexte erst maschinell analysiert werden, um *Erkenntnisse* über deren Inhalte zu gewinnen. Dies birgt zwar einige Herausforderungen hinsichtlich der Subjektivität (Wiebe, 2004; Bruce & Wiebe, 1999), Polarität (Bakliwal et al., 2012; Kim & Hovy, 2006; Qin et al.,

[1] Aus Gründen der leichteren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Entsprechende Begriffe gelten im Sinne der Gleichbehandlung für beide Geschlechter.

2008; Pang et al., 2002) und der Semantik (Hu & Liu, 2004b; Kim & Hovy, 2006; Turney, 2002) der Aussage(n), führt aber nach Anwendung von automatischen Textanalyseverfahren zu qualitativ hochwertigen, interdisziplinär verwertbaren Ergebnissen. Zwar untersucht die computerlinguistische Forschung bereits automatisiert die Zufriedenheit von Konsumenten hinsichtlich bestimmter Produkte oder Dienstleistungen (Gamon, 2004; Hu & Liu, 2004a), unterstützt aber bisher nicht die Identifikation von thematischen Rollen in Bewertungstexten. Die Forschung im Bereich der Bewertungskultur im Web 2.0 deckt bereits eine Vielzahl unterschiedlicher Disziplinen (u.a. Soziologie, Marketing, Psychologie, Linguistik, Informatik) und verschiedene Aspekte (u.a. Bewertungsgegenstand, demographische Einflüsse, Emotionserkennung) ab. Selbst im Bereich des Textmining in Sozialen Medien werden Bewertungen und Kommentare vorwiegend auf Polarität (He et al., 2011; Jiang et al., 2011) und/oder Inhalt (Kushal et al., 2003) automatisch analysiert. Hierbei werden in der Regel die offensichtlichen sprachlichen Indikatoren (z. B. Emoticons, positive und negative Wörter) zur Identifizierung der Grundstimmung in den einzelnen Texten berücksichtigt (sog. Sentiment Analysis). Vereinzelt gibt es jetzt auch Arbeiten, welche sich mit der Stilistik von Bewertungstexten (Ludwig et al., 2013) auseinandersetzen. Während Autoreninformationen meist zur Erkennung von gefälschten Bewertungen (Liu, 2012:127; O'Connor, 2010; Hu et al., 2011) eingesetzt werden, konzentrieren wir uns in diesem Beitrag auf die Identifikation von Akteuren (sog. Rollen) im Online-Ärztebewertungsprozess.

Rezensenten – Patienten – Ärzte

Vermutlich sind es Patienten, die ihre Ärzte bewerten und die eigenen Erfahrungen nach dem Arztbesuch weitergeben möchten. Allerdings finden sich auch einige Belege im Korpus, dass Kinder über den Arztbesuch ihrer alternden Eltern berichten oder Eltern für ihre minderjährigen Kinder Bewertungen vergeben. Damit wird nicht nur die Behandlung, sondern ebenfalls der Behandelte Gegenstand des Erfahrungsberichtes, so dass Autor und Patient nicht mehr in persona auftreten.

Der damit einhergehende Perspektivwechsel spiegelt sich in den morpho-syntaktischen Strukturen der Texte wider. So entwickelte Geierhos (2007) mit der Grammatik der Menschenbezeichner eine Methode, um Prädikat-Argument-Strukturen mittels lokaler Grammatiken (Gross, 1997) zur Identifikation von Personennennungen und ihrer Funktion (semantische Rolle) im unmittelbaren Kontext zu modellieren. Als Menschenbezeichner sind diejenigen Lexeme oder Mehrwortlexeme zu verstehen, welche auf Personen referieren. Hierbei handelt es sich aus syntaktischer Sicht um Eigennamen (z. B. Dr. med. Dieter Rempe), Personalpronomen (z. B. er, ihm, sein) oder Nomen (sog. „allgemeine Menschenbezeichnungen“ nach Geierhos, 2007:69), die u.a. Berufe, Nationalitäten oder Verwandtschaftsverhältnisse denotieren (Geierhos, 2007:74).

In Erfahrungsberichten über Arztbesuche konnten wir folgende *semantische Rollen* identifizieren:

(1) Wenn der Rezensent selbst der Patient ist, so nehmen beide in persona die Rolle des *Experiencers* (Wahrnehmenden) ein.
(2) Ist der Patient nicht der Rezensent, wird er zum *Instrument* der Handlung. Er ist damit der Anlass bzw. das Mittel zum Zweck, um eine Arztbewertung zu verfassen. Der Rezensent nimmt dabei die Rolle des *Agens* ein, der willentlich handelt, indem er die Rezension schreibt.
(3) Unabhängig davon, ob Rezensent und Patient in gleicher oder unterschiedlicher Person auftreten, ist der Arzt und gegebenenfalls sein Praxisteam der Bewertungsgestand, der sogenannte *Patiens* des Erfahrungsberichts.
(4) Zum *Thema* eines Erfahrungsberichts werden sämtliche Qualitätskriterien (u.a. Behandlung, Freundlichkeit, Zufriedenheit und Weiterempfehlung, Aufklärung, Genommene Zeit, Patientenumgang) bei der Benotung niedergelassener Ärzte durch den Rezensenten.

Insbesondere konnte mithilfe der automatischen Erkennung von Menschenbezeichnern im Korpus identifiziert werden, wer Rezensent und Patient im Fall (2) sind. So handelt es sich in ungefähr 17.800

Bewertungen um ein Kind, das den Arztbesuch mit einem oder beiden Elternteilen schildert. Bei weiterer Textanalyse konnte ebenfalls das Alter der Patienten aufgedeckt werden, die in der Regel weit über 70 sind, wenn ihre Kinder für sie stellvertretend den Arztbesuch bewerten. Meist sind es Angehörige, die in der Rolle der Rezensenten schlüpfen. Ähnliches gilt für Eltern, die für ihre Kinder sprechen (29.990 Datensätze). In diesen Fällen sprechen wir von Fremdwahrnehmung, da die Erlebnisse der Patienten – der ursprünglichen *Experiencer* – durch Dritte wiedergegeben werden und damit nicht mehr sichergestellt wird, inwiefern diese Meinungsbilder nicht durch die Rezensenten beeinflusst wurden.

Die Mutter ist nicht immer die Patientin

Ausgehend von der These, dass die Nennung von Verwandtschaftsverhältnissen ein Indiz für eine fremde Autorschaft ist, kann vermutet werden, dass eine Grammatik zur Modellierung der Prädikat-Argument-Beziehungen von Menschenbezeichnern überflüssig ist, um die semantischen Rollen zu bestimmen. Das Korpus liefert jedoch Belege, die ihre Anwendung rechtfertigen:

(a) Meine Ärztin sollte nicht wie meine Mutter sein.
(b) […], da ich mittlerweile die gleiche Erkrankung wie meine Mutter habe.
(c) Ich bin seit Jahren, genau wie meine Mutter, bei Dr. Esser in Behandlung.

Stellvertretend für weitere allgemeine Menschenbezeichnungen (Geierhos, 2007) wird anhand der „Mutter"-Beispiele in (a)-(c) deutlich, dass ihre syntaktische Einbettung im Kontext von bestimmten prädikativen Ausdrücken ihnen andere semantische Rollen zuweist. Eine Muster-basierte Textsuche nach „meine/unsere Mutter" würde jegliches Vorkommen im Korpus unabhängig von seiner semantischen Rolle im Erfahrungsbericht berücksichtigen und eine Differenzierung über den Gebrauch im Satz wäre nicht mehr möglich.

Informiertes Einverständnis im Web 2.0?

Im Korpus sind zwar die Rezensenten anonym (ohne Namen), aber nicht die Ärzte. Die Veröffentlichung dieser personenbezogenen Daten unterliegt damit der komplexen Abwägung zwischen dem allgemeinen Persönlichkeitsrecht der Ärzte und dem Grundrecht auf Meinungs- und Informationsfreiheit. Gerichtlich wurde für diesen Anwendungsfall entschieden, dass das Erheben, Speichern, Verändern oder Nutzen im Sinne des Bundesdatenschutzgesetzes grundsätzlich zulässig ist. Weiterhin werden im Fall (1), wenn Rezensent und Patient ein und dieselbe Person sind, keine Persönlichkeitsrechte verletzt, da dieser seinen Erfahrungsbericht freiwillig veröffentlicht hat. Aber wie verhält es sich nun im Fall (2) bei denjenigen Datensätzen, die wir bereits erhoben haben und nun neue Erkenntnisse über deren Autorschaft gewonnen haben? Wissen die betroffenen Patienten überhaupt, dass ihre Erfahrungen beim Arztbesuch online – für jeden lesbar – zugänglich sind? Kann man bei der Datenakquise das informierte Einverständnis der Behandelten voraussetzen? Erheben wir Daten, sehen wir nicht die Person(en) dahinter, sondern nur eine Fülle an Informationen, aus denen wir weitere Schlussfolgerungen ziehen können. Doch werfen diese Erkenntnisse unter Umständen weitere (interdisziplinäre) Fragestellungen auf, die Konsequenzen haben, ob und wie wir forschungsethische Diskussionen führen.

Literatur

Bakliwal, A.; Arora, P.; Madhappan, S.; Kapre, N., Singh, M. & Varma. V. (2012): *Mining sentiments from tweets.* In Proceedings of the 3rd WASSA-Workshop 2012, Jeju, Island, Republic of Korea.

Bruce, RF. & Wiebe, JM. (1999): *Recognizing subjectivity: a case study in manual tagging.* In Natural Language Engineering, 1(1), pp. 187-205.

Gamon, M. (2004): *Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis*. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004, pp. 611-617.

Geierhos, M. (2007): *Grammatik der Menschenbezeichner in biographischen Kontexten.* Arbeiten zur Informations- und Sprachverarbeitung. Band 2. Centrum für Informations- und Sprachverarbeitung. LMU München.

Gross, M. (1997): *The Construction of Local Grammars*. In E. Roche und Y. Schabès (Hrsg.): Finite-State Language Processing, pp. 329–354. Language, Speech und Communication, Cambridge, Mass.: MIT Press.

He, Y.; Lin, C. & Alani, H. (2011): *Automatically extracting polarity-bearing topics for cross-domain sentiment analysis*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oreg., USA, 19–24 June 2011, pp. 123-131.

Hu, M. & Liu, B. (2004a): *Mining opinion features in customer reviews*. In Proc. of the 19th National Conference on Artificial Intelligence, San Jose, Kalifornien, 25.-29. Juli 2004, pp. 755-760.

Hu, M. & Liu, B. (2004b): *Mining and summarizing customer reviews*. In Proceedings of SIGKDD, pp. 168–177.

Hu, N.; Bose, I.; Koh, NS. & Liu, L. (2012): *Manipulation of online reviews: An analysis of ratings, readability, and sentiments*. In Decision Support Systems 52(3), pp. 674-684.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X. & Zhao, T. (2011): *Target-dependent Twitter sentiment classification*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oreg., USA, 19.-24. Juni 2011, pp. 151-160.

Kim, S.-M. & Hovy, E. (2006): *Automatic identification of pro and con reasons in online reviews*. In Proceedings of the Poster Session at the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17.-21. Juli 2006, pp. 483–490.

Kushal, D.; Lawrence, S. & Pennock, D. M. (2003): *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In Proceedings of the 12th World Wide Web Conference, Budapest, Hungary, 20.–24. Mai 2003, pp. 519–528.

Ludwig, S.; de Ruyter, K.; Friedman, M.; Brüggen, E. C.; Wetzels, M. & Pfann, G. (2013): *More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates*. Journal of Marketing 77:1, pp. 87-103.

O'Connor, P. (2010): *Managing a Hotel's Image on TripAdvisor.* In Journal of Hospitality Marketing & Management. 19(7), pp. 754-772.

Qin, B.; Zhao, Y.; Gao, L. & Liu, T. (2008): *Recommended or not? Give advice on online products*. In Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Jinan, Shandong, China, 18.-20. Oktober 2008, pp. 208-212.

Turney, P. (2002): *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn., 7.-12. Juli 2002, pp. 417-424.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004): *Learning subjective language*. In: Computational linguistics, 30(3), pp. 277-308.

WebLicht: Bombardieren bevor die Services explodieren

Daniël de Kok, Wei Qiu, Marie Hinrichs

Einleitung

Web-Analyse Software, die Informationen über das Nutzerverhalten sammelt und auswertet, kann eingesetzt werden, um Probleme zu verhindern, bevor diese entstehen. In diesem Abstrakt zeigen wir, wie wir eine derartige Software auf ein System mit verteilten Webservices, WebLicht [1], angewandt haben und wie wir die entstandene Analyse benutzt haben, um das System zu verbessern.

WebLicht ist eine Webanwendung zur automatischen Annotation von Texten und multimodalen Daten. WebLicht nutzt eine serviceorientierte Architektur. Dies ermöglicht den CLARIN-Zentren Annotationswerkzeuge hinzuzufügen durch diese als Webservice zu implementieren und die CMDI-Metadaten [2] über diesen Service an ihr Repositorium zuzufügen. WebLicht aggregiert diese Metadaten und bietet dem Nutzer die Annotationswerkzeuge an. Wenn ein Benutzer mehrere Annotationswerkzeuge ausführen will oder ein Annotationswerkzeug vom Output eines anderen Werkzeugs abhängig ist, erlaubt WebLichts Verkettungsmechanismus es den Benutzern, Annotationswerkzeuge auf eine kompatible Art und Weise zu kombinieren.

WebLicht ist eine voll entwickelte Anwendung, die mehr als hundert Annotationswerkzeuge anbietet. Dadurch weitet sich die Nutzung von WebLicht in zwei Dimensionen aus: (1) Die Anzahl der WebLicht-Benutzer steigt; (2) durchschnittlich lassen die Benutzer größere Datasets annotieren. Um diese Wachstumsstrukturen zu verstehen, sammeln wir Nutzungsstatistiken. Um WebLicht und die Annotationswerkzeuge für unsere Nutzerbasis zu optimieren, müssen wir zuerst die Nutzungsmuster der derzeitigen Benutzer kennen. Zweitens müssen wir zukünftige Kapazitätsengpässe vorhersagen können, damit wir uns mit ihnen auseinandersetzen können, bevor sie die User Experience beeinträchtigen.

In den folgenden Abschnitten werden wir zuerst beschreiben, wie wir Benutzeraktivität und Nutzungsmuster messen. Dann werden wir die Simulation verschiedener Nutzungsszenarien behandeln. Schließlich werden wir einen kurzen Überblick über die Veränderungen geben, die wir im vergangenen Jahr aufgrund der Messungen vorgenommen haben.

Die Messung von Benutzeraktivität

Das Sammeln von Nutzerstatistiken für Webseiten ist ein wohlverstandenes Feld, in dem es mehrere konkurrierende und voll entwickelte Produkte, wie etwa Google Analytics, gibt. Es sind auch viele gute Open Source Webanalyse-Tools, wie zum Beispiel Piwik[1], Webalizer[2] and AWStats[3] verfügbar. Diese Tools funktionieren gewöhnlich auf eine von zwei Arten: (1) Sie analysieren die vom Webserver protokollierten Logdateien; oder (2) sie benötigen einen Maintainer, der auf jede Seite einen Javascript-Snippet einfügt, wodurch der Browser des Benutzers mit der Analysesoftware Kontakt aufnimmt.

Bereits existierende Lösungen sind nicht direkt auf WebLicht anwendbar, da ihre Verwendung nur zeigen würde, wie oft, aus welchem Land, etc. WebLicht besucht wurde. Die nützlichste Information würden fehlen: welche Annotationswerkzeuge verwenden die Benutzer, und wie oft? Piwik bietet eine Programmierschnittstelle (Application Programming Interface) an, durch die jedes Annotationswerkzeug seine eigene Nutzung registrieren kann. Jedoch würde dies erfordern, dass knapp hundert Dienste upgedatet werden müssten, um die Programmierschnittstelle aufrufen zu können. Ein weiterer Nachteil dieser Lösung ist, dass interessante Informationen, wie etwa das Land des Benutzers, nicht verfügbar sind, da die Annotationswerkzeuge von WebLicht aufgerufen werden und nicht direkt vom Browser des Benutzers.

Wir haben die oben genannten Probleme durch das Melden von Statistiken im WebLicht-Verkettungsmechanismus gelöst. Da der Verkettungsmechanismus jedes Annotationswerkzeug aufruft, kann er gleichzeitig die Verwendung des Werkzeugs über die Piwik-Programmierschnittstelle melden. Dies macht die Statistiken sofort erhältlich für alle Annotationswerkzeuge, die in WebLicht verfügbar sind, ohne auch nur eine von ihnen zu verändern. Da der Verkettungsmechanismus in der WebLicht-Anwendung ausgeführt wird, hat es den Zusatznutzen, dass wir einige Metadaten bereitstellen können, die es Piwik erlauben das Land des Besuchers, den Webbrowser, etc. zu ermitteln.

[1] http://piwik.org/
[2] http://www.webalizer.org/
[3] http://www.awstats.org/

⊞ **ws1-clarind.esc.rzg.mpg.de**	12	13
⊞ **weblicht.sfs.uni-tuebingen.de**	372	18026
⊞ **kaskade.dwds.de**	12	12
⊞ **dspin.dwds.de:8080**	51	64
⊞ **dlexdb.de**	5	5
⊟ **clarin05.ims.uni-stuttgart.de**	139	3832
/treetagger2008	35	55
/treetagger	4	4
/RFTaggerMorph	26	3645
/rftagger	1	1
/cgi-bin/dspin/tokeniser4.perl	43	75
/cgi-bin/dspin/tei2tcf4.perl	1	1
/cgi-bin/dspin/tei2tcf3.perl	2	3
/cgi-bin/dspin/smor4.perl	1	1
/cgi-bin/dspin/bitpar4.perl	26	47
⊞ **chopin.ipipan.waw.pl:8083**	1	1

Abb. 1: Piwik Nutzerstatistiken für die Stuttgarter Werkzeuge

Die Simulation von Nutzungsmustern

Eines der Probleme, auf das wir mit der gestiegenen Nutzung von WebLicht gestoßen sind, ist, dass manche Annotationswerkzeuge nicht zur Bewältigung vieler Simultanbenutzer oder großen Inputs entwickelt wurden. Leider wurden diese Probleme oft erst entdeckt, wenn ein Benutzer eines der Annotationswerkzeuge nicht ausführen konnte.

Durch das Simulieren von WebLicht-Nutzungsmustern konnten wir solche Probleme aufdecken, bevor die Benutzer ihnen begegnen. Zu diesem Zweck haben wir ein Simulationswerkzeug namens Bombard entwickelt. Bombard ermöglicht es den Entwicklern von WebLicht-Annotationswerkzeugen, Testfälle zu spezifizieren. Jeder Testfall besteht aus: der Kette von Annotationswerkzeugen, die getestet werden soll; dem Input; einem Intervall, das anzeigt, wie oft der Testfall ausgeführt werden soll; und die maximal erlaubte Bearbeitungsdauer. Eine Bombard-Konfigurierung kann aus vielen solcher Testfälle bestehen. Wenn Bombard gestartet wird, beginnt es jeden Testfall zu einem willkürlich gewählten Zeitpunkt und wiederholt den Testfall ab da in dem festgelegten Intervall. Während des 'Bombardements' behält Bombard den Überblick über Fehlschläge und inakzeptable Bearbeitungsdauern. Danach können die Entwickler einen Bericht mit Statistiken für jedes Annotationswerkzeug abrufen.

Im CLARIN-Zentrum in Tübingen wird Bombard zum Testen neuer Dienste genutzt, indem es das folgende Szenario simuliert: Zwei Gruppen von jeweils 40 Studierenden reichen innerhalb von zwei Minuten einen Text ein. In Gruppe A bearbeitet jeder Studierende zwei Textabschnitte aus Wikipedia, in Gruppe B bearbeitet jeder Studierende den Roman *Alice im Wunderland* (oder eine Übersetzung davon). Mithilfe dieses Testszenarios konnten wir Kapazitätsengpässe

in früher entwickelten Annotationsdiensten bestimmen und sicherstellen, dass die neuen Dienste in der Lage sind, solche Szenarien zu verarbeiten.

Da die Konfigurierung der Testfälle in Bombard anpassungsfähig ist, kann sie einfach auf andere Szenarien oder Annotationswerkzeuge, für die andere Erwartungen gelten, angewendet werden.

Änderungen an WebLicht

Wir haben unser Klassenzimmer-Szenario an Annotationsketten getestet, die unseren Feststellungen nach häufig genutzt werden, nämlich: Part-of-speech Tagging, Lemmatisierung, Konstituenz-Parsen, Dependenz-Parsen und Named-entity Recognition. Während der Simulation entdeckten wir die folgenden Probleme: (1) Manche Dienste versagten, wenn viele Instanzen des längeren Textes gesandt wurden; (2) manche versagten an dem längeren Text, da er relativ verrauscht ist; (3) bei manchen Diensten, insbesondere Parsern, kamen die Anfragen schneller herein als der Dienst sie verarbeiten konnte.

Das erste Problem war am leichtesten zu lösen – diese Dienste hatten eine ältere Version der TCF interchange format library genutzt, die sich nicht linear an die Größe des Inputs anpasste. Das zweite Problem musste von Fall zu Fall einzeln gelöst werden. Diese Dienste hatten einige Programmierfehler, die sie bei unerwartetem Input versagen ließen. Das dritte Problem jedoch war schwerer zu lösen und wird im Rest dieses Abschnitts besprochen werden.

Parser für natürliche Sprachen sind oft langsam im Vergleich zu anderen Annotationsdiensten. Zum Beispiel können gebräuchliche Konstituenz-Parser normalerweise höchstens ein paar Sätze pro Sekunde parsen. Um die von uns vorhergesehenen Gebrauchsfälle bewältigen zu können, mussten wir die Fähigkeiten moderner Server und Computercluster, Prozesse parallel durchführen zu können, ausnutzen. Dafür haben wir ein Framework entwickelt [3], das auf einer verteilten Task-Warteschlange (Jesque) basiert und das folgende bietet: Parallele Prozessverarbeitung innerhalb der Anfragen, gleichzeitige Verarbeitung von Anfragen, Garantien, was die Verwendung von Ressourcen und Fairness betrifft (z.B.: Eine große Anfrage sollte keine sichtbaren Auswirkungen auf kleine Anfragen haben.)

Wir benutzen dieses Framework in den upgedateten Diensten für die Malt-, Stanford- und Berkeley-Parser. Dadurch können wir solche Szenarien leicht bewältigen. Was vielleicht noch wichtiger ist: Unser Framework erlaubt es uns, Webservices an noch größere simultane Benutzerzahlen oder größere Inputs anzupassen, indem wir weitere Prozessorkerne oder Geräte hinzufügen.

Fazit

Wir haben diesem Abstract zwei neue Vorgehensweisen zur Messung der Nutzung und Kapazität von WebLicht dargestellt. Zuerst haben wir Nutzungsmeldungen zum Verkettungswerkzeug hinzugefügt, sodass wir die Nutzungsstatistiken der Annotationswerkzeuge bekommen ohne die CLARIN-Partner darum bitten zu müssen, ihre Werkzeuge anzupassen. Zweitens haben wir das Dienstprogramm Bombard vorgestellt, das es uns ermöglicht die Auswirkung der Hochrechnung des momentanen Wachstums zu messen. Solche Messungen machen uns die Nutzung zur Bestimmung von Kapazitätsengpässen in der WebLicht-Infrastruktur und ihre frühe Behandlung möglich.

Bibliografie

[1] Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25– 29. Association for Computational Linguistics.

[2] ISO 24622-1:2014. 2014. Language resource management -- Component Metadata Infrastructure (CMDI) -- Part 1: The Component Metadata Model. Technical Report, ISO.

[3] De Kok, D., De Kok, D., and Hinrichs, M. (2014). Build your own treebank. In: *Proceedings of the CLARIN Annual Conference 2014.* Soesterberg, Netherlands.

Zeitliche Verlaufskurven in den DTA- und DWDS-Korpora: Wörter und Wortverbindungen über 400 Jahre (1600–2000)

Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish,
Christian Thomas, Frank Wiegand, Kay-Michael Würzner
(Berlin-Brandenburgische Akademie der Wissenschaften)

Einführung

In diesem Beitrag werden zwei Referenzkorpora der deutschen Sprache verwendet, um daraus frequenznormierte Verlaufskurven für Wörter und Wortverbindungen zu berechnen: das DWDS-Kernkorpus des 20. Jhs. sowie das Referenzkorpus des Deutschen Textarchivs (1600–1900). Da beide Korpora bezüglich ihrer Metadaten vereinheitlicht und auch mit denselben linguistischen Informationen annotiert wurden, können korpusübergreifende Abfragen gestellt werden. Beispiele hierfür sind schreibweisentolerante Lemmasuchen oder textsortenspezifische Suchen. Die auf dieser Grundlage generierten Verlaufskurven stehen auf der Website des Deutschen Textarchivs für die Abfrage zur Verfügung.

1 Korpusgrundlage und Annotation

Die Grundlage für die zeitlichen Verlaufskurven (Histogramme) bilden die 100 Millionen Textwörter des DWDS-Kernkorpus des 20. Jhs. sowie weitere 140 Millionen Textwörter des Deutschen Textarchivs, welches Werke des 17. bis 19. Jh. als Erstausgaben umfasst.

Beide Korpora sind hinsichtlich der repräsentierten Textsorten aus Belletristik, Gebrauchsliteratur, Wissenschaft und (im DWDS:) Journalistischer Prosa sowie hinsichtlich der enthaltenen Disziplinen ausgewogen (Geyken 2007, 2013; Geyken et al. 2011). Sie wurden beide gemäß den TEI/P5-Richtlinien annotiert (Geyken et al. 2012; Haaf et al., forthcoming) und sind bezüglich der Metadaten untereinander interoperabel, insbesondere bezüglich der in den Histogrammen verwendeten Angaben zum Datum und zu den Textsorten.

Beide Korpora wurden für die Auswertung linguistisch annotiert, insbesondere wurden alle Texte tokenisiert, lemmatisiert, nach lexikalischen Kategorien analysiert (PoS-Tagging) und mit GermaNet-Kategorien versehen. Von besonderer Bedeutung für die Generierung der Histogramme ist die CAB-Analyse der historischen Texte: Mit CAB werden historische Varianten einer Wortform auf die wahrscheinlichste Normalform reduziert und ihrem neuhochdeutschen Lemma zugeordnet (Jurish 2013). Damit ist nicht nur die Suche in den Korpora, sondern auch die Darstellung der Wortverläufe schreibweisenübergreifend möglich. Dies geschieht, indem das Suchwort extensional zu all denjenigen Wortformen expandiert wird, die eine (möglicherweise flektierte) Variante des Suchwortes darstellen. In den

folgenden Abschnitten wird gezeigt, dass diese Vorgenerierung der möglichen Formen für den Benutzer eine effektive Hilfe darstellt (s. Abschnitt 3: Beispiele).

Beide Korpora, das DWDS-Kernkorpus und das DTA-Korpus, sind mit der Suchmaschine DDC (Dialing DWDS Concordancer) indiziert (Jurish et al. 2014). DDC verfügt über reichhaltige Metadatenfilter sowie die Möglichkeit, mehrere Annotationen auf einer Wortposition zu indizieren und abfragbar zu machen. Darüber hinaus können reguläre Ausdrücke, Boolesche Verknüpfungen und Abstandsoperatoren genutzt werden. Insbesondere ist es möglich, Metadatenfilter und mehrfache linguistische Annotationen miteinander zu verbinden. Die Indizes von DDC wurden so optimiert, dass Abfragen über die zeitliche Verteilung (Histogramme) ausreichend schnell für eine dynamische Berechnung zur Laufzeit sind. Dadurch ist es auch möglich, zeitliche Verläufe für die gesamte Mächtigkeit der DDC-Abfragesprache dynamisch zu berechnen.

2 Visualisierung

Grundlage der Visualisierung sind die nach Messpunkten (Jahreszahlen bzw. Datumsintervallen) normierten relativen Häufigkeiten pro Million Textwörter. Da die beiden Korpora weder gleich verteilt noch gleich groß sind und zudem die Wortanzahl je Zeitintervall variiert, ist die Normierung notwendig, um die Histogramme beider Korpora einheitlich zu präsentieren. Wie im vorigen Abschnitt erwähnt, können normierte relative Frequenzen nicht nur für einzelne Wortformen oder Lemmata, sondern auch für Kombinationen aus linguistischen Annotationen und Phrasensuchen ausgegeben werden. Darüber hinaus können Histogramme textsortenspezifisch (z. B. nur für Belletristik oder nur für Wissenschaft) oder textsortenübergreifend gebildet werden.

Da die Häufigkeiten von Wortformen (Types) in Textkorpora zipfsch verteilt sind, ist aufgrund der Größe der beiden Referenzkorpora bereits bei Wortformen einer mittleren Häufigkeit damit zu rechnen, dass die Histogramme bei einzelnen Messpunkten Nullstellen aufweisen können. Umgekehrt kann auch die Unausgewogenheit der Textkorpora an einem Messpunkt zu „Ausschlägen“ mit zu hoher Frequenz führen. Aus diesem Grund wurden in die Visualisierungskomponente verschiedene Parameter zur Glättung implementiert.

Aus Platzgründen soll hier nur auf die beiden wichtigsten eingegangen werden: die Parameter „window“ und „pruning“. Der Parameter „window“ gibt die Fensterbreite (als natürliche Zahl) für die Glättung nach dem gleitenden Mittelwert an. Der Parameter „pruning“ implementiert ein zweistufiges Verfahren. Im ersten Schritt wird eine Fehlerverteilung für die normierten Datenpunkte berechnet. Die beobachteten „Fehler“ werden unter Annahme einer Normalverteilung in p-Werte überführt, und alle Datenpunkte mit p-Werten außerhalb des angegebenen Konfidenzbereichs (p=0.05) werden als Ausreißer behandelt. Datenpunkte, die Ausreissern entsprechen, werden durch eine lineare Interpolation der nächstliegenden Datenpunkte, die innerhalb des Konfidenzbereichs liegen, ersetzt.

Die Visualisierung selbst erfolgt mittels der Javascript-Bibliothek Highcharts[1] und ist auf der Website des DTA abfragbar.[2]

3 Beispiele, Ergebnisse und Diskussion

Im folgenden, abschließenden Abschnitt sollen einige Vorzüge der Visualisierung auf der Grundlage der oben beschriebenen Referenzkopora und ihrer linguistischen Annotationen anhand konkreter Beispiele illustriert werden. Die Abfragemöglichkeiten und Ergebnisse werden mit den entsprechenden Ergebnissen im *Google Ngram Viewer*[3] in Beziehung gesetzt.

Beispiel 1: *Tatsache*

Als erstes Beispiel dient das Lemma „Tatsache“, ein Begriff, der laut dem Etymologischen Wörterbuch (Pfeifer) erst mit einer Publikation aus dem Jahr 1756 Eingang in die deutsche Sprache fand.[4] Die Wortverlaufskurve in *Abb._1* bestätigt den Beginn dieser „Wortkarriere“ und zeichnet ihn von der Mitte des 18. Jhs. bis in das 20. Jh. nach.[5]

Die Verlaufskurve im *Google Ngram Viewer*, *Abb._2*, zeigt eine vergleichbare Tendenz. Der Vergleich beider Histogramme lässt dennoch Probleme bei der Arbeit mit dem *Ngram Viewer* sichtbar werden: Historische Schreibweisen wie „Thatsache“ – sowie ggf. etliche weitere möglich Varianten[6] und idealerweise auch alle möglichen Flexionsformen bzw. Expansionen – müssen bei der Eingabe der Anfrage explizit ergänzt werden, um die historische Entwicklung des Begriffs zu visualisieren. Auf den Punkt Flexionsformen/Expansionen wird im dritten Beispiel („billig“) noch näher eingegangen; hier sind zunächst nicht erklärbare Ausschläge der Kurve *vor* 1756 von Interesse, als Spalding das Wort in die deutsche Sprache brachte. Besonders deutlich zeigt sich darin ein Problem des *Google-Books*-Korpus. Eine Recherche in den Dokumenten daselbst zeigt, dass es sich bei sämtlichen früheren Treffern um falsch datierte Dokumente handelt, darunter so prominente Beispiele aus dem 19. Jh. wie Goethes Schriften zur Morphologie (II. Teil, datiert auf 1659) und die Fliegenden Blätter (in Google Books datiert auf 1692). Die Wortverlaufskurven werden hier also durch Metadatenfehler verfälscht.

---

[1] http://www.deutschestextarchiv.de/search/plot.
[2] http://www.highcharts.com/products/highcharts.

[3] https://books.google.com/ngrams.
[4] Vgl. z. B. das Etymologische Wörterbuch des Deutschen (nach W. Pfeifer), digitale Version via DWDS: „[...] nachgebildet (1756) von dem Theologen Spalding für engl. matter of fact [...].“ Siehe *Bestätigung der natürlichen und geoffenbarten Religion*, Leipzig, 1756, sowie Johann Joachim Spaldings Übersetzung von Joseph Butlers *The analogy of religion, natural and revealed, to the constitution and course of nature*, 1736. Die Texterfassung dieses Werks ist für das DTA in Arbeit; derzeit stammt der früheste Beleg im DTA-Korpus aus Münter 1772.

[5] Aus Platzgründen kann hier auf den Abfall der Kurve ab Mitte des 20. Jh. nicht eingegangen werden.
[6] Im DTA-Korpus sind für den in dieser Hinsicht einfach erscheinenden Begriff „Tatsache“ immerhin 15 Expansionen belegt, vgl. http://kaskade.dwds.de/dstar/dta/lizard.perl?q=Tatsache.

Beispiel 2: *merkwürdig*

Das zweite Beispiel illustriert die Möglichkeiten, die eine Textsortendifferenzierung für die Interpretation der Histogramme bietet. Der Begriff „merkwürdig“ wurde Pfeifers Etymologischem Wörterbuch zufolge ab dem 19. Jh. vorrangig in der Bedeutung „seltsam, verwunderlich“ verwendet; bis dahin dominierte die seit dem 17. Jh. gebräuchliche Verwendung im Sinne von „bemerkenswert, bedeutsam“. *Abb._3* zeigt das nach Textsorten differenzierte Histogramm. Belegt ist die relativ hochfrequente Verwendung von „merkwürdig“ in den Textsorten Wissenschaft und Gebrauchsliteratur bis über die Mitte des 19. Jhs. hinaus. Folgt man der These (wie die im DTA verfügbaren Belege bestätigen), dass die Verwendung in der Wissenschaft und der Gebrauchsliteratur vorrangig im Sinne von „bemerkenswert, bedeutsam“ geschah, legt dies einen etwas länger andauernden Gebrauch des Wortes in dieser Bedeutung nahe, als bei Pfeifer angegeben. Auch hier zeigt sich der Vorteil gegenüber der Verlaufskurve des *Google Ngram Viewers* (*Abb._4*), wo die Textsortendifferenzierung fehlt.[7]

Beispiel 3: *billig*

Das abschließende Beispiel („billig“) illustriert zwei weitere Vorzüge der auf den Referenzkorpora beruhenden Histogramme gegenüber dem *Google Ngram Viewer*: die bessere Abdeckung der DTA-Korpora bezüglich des 17. Jhs. und die bessere Handhabung der großen graphematischen Varianz v. a. in diesen historischen Texten. *Abb._5* zeigt das Histogramm für „billig“ aus DTA und DWDS; die Kurve veranschaulicht die breite Verwendung des Begriffs im 17. Jh., die dazugehörigen Belege zeigen das Bedeutungsspektrum zwischen ‘angemessen, gerechtfertigt’ und ‘mäßig, wohlfeil, günstig’. Die Kurve aus dem *Google Ngram Viewer*, *Abb._6*, zeigt dagegen keinen kontinuierlichen Verlauf im 17. Jh., was angesichts der u. a. im DTA-Korpus belegten Verbreitung des Begriffs die verhältnismäßig geringe Substanz des *Google-Books*-Korpus für diesen Zeitraum belegt. Daher erscheint auf dessen Grundlage quellenbasierte Forschung zumindest in diesem Zeitraum kaum möglich. Ein weiteres, bereits angesprochenes Problem, kommt erschwerend hinzu: Insbesondere bei Abfragen für Zeiträume vor 1700 liefert das *Google-Books*-Korpus aufgrund der sehr heterogenen Graphie schlichtweg keine befriedigende Anzahl von Belegen; hier liefert die CAB-Analyse der DTA-Korpora klare Vorteile. Allein für das in dieser Hinsicht relativ unproblematisch erscheinende Lemma „billig“ sind im DTA-Korpus die in *Abb._7* gezeigten immerhin 67 Flexions- und Expansionsformen belegt. Bei einer Abfrage via DTA/DWDS werden alle diese Formen berücksichtigt, während der Nutzer des *Google Ngram Viewers* sie manuell eingeben (und zu diesem Zweck selbstverständlich überhaupt erst einmal präsent haben) müsste.

4 Ausblick

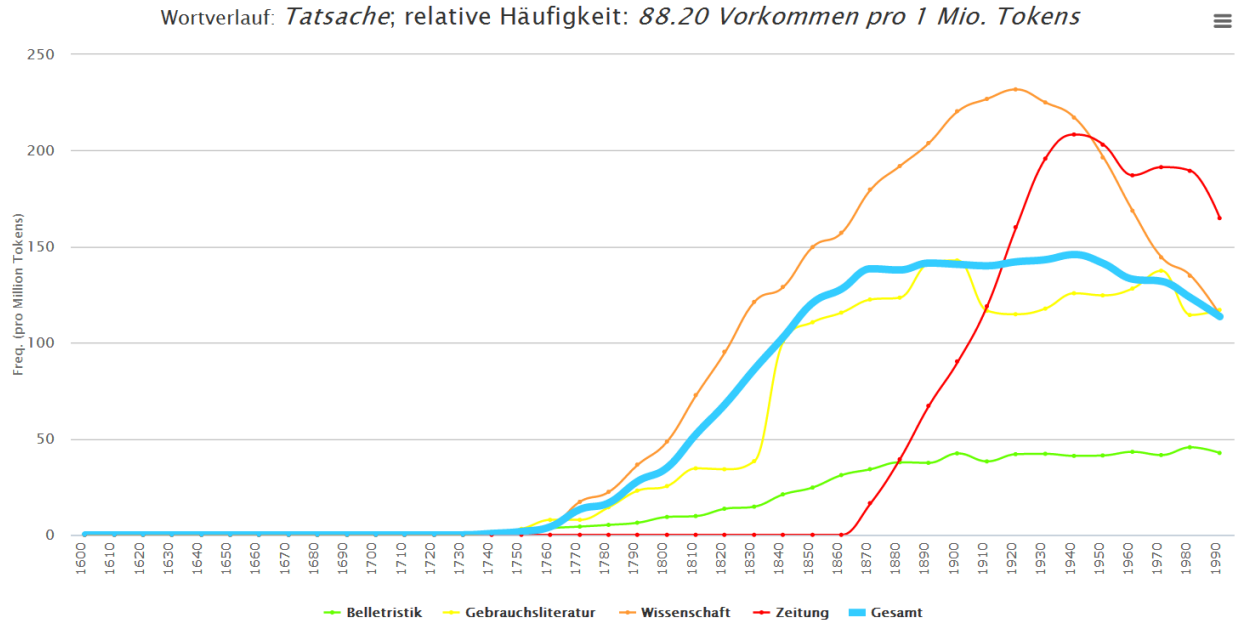
Wie im vorigen Abschnitt gezeigt werden konnte, liefert die Zeitverlaufskurve auf der Grundlage der beiden Referenzkorpora interessante Ergebnisse, die mit dem wesentlich

[7] *NB:* Überraschend ist zudem, dass die Graphien „merkwirdig“ und „merckwirdig“ in dem der Kurve zugrundeliegenden Korpus *German* (das nicht identisch mit dem abfragbaren aktuellen GB-Korpus ist) nicht belegt sind.

größeren *Google-Books*-Korpus entweder nicht oder nur mit großem Rechercheaufwand ermittelbar gewesen wären. Grund dafür sind die verlässlichen Metadaten, die Zuordnung nach Textsorten sowie die aufgrund der genauen Texterfassung möglichen Erschließungsmethoden (CAB-Software).

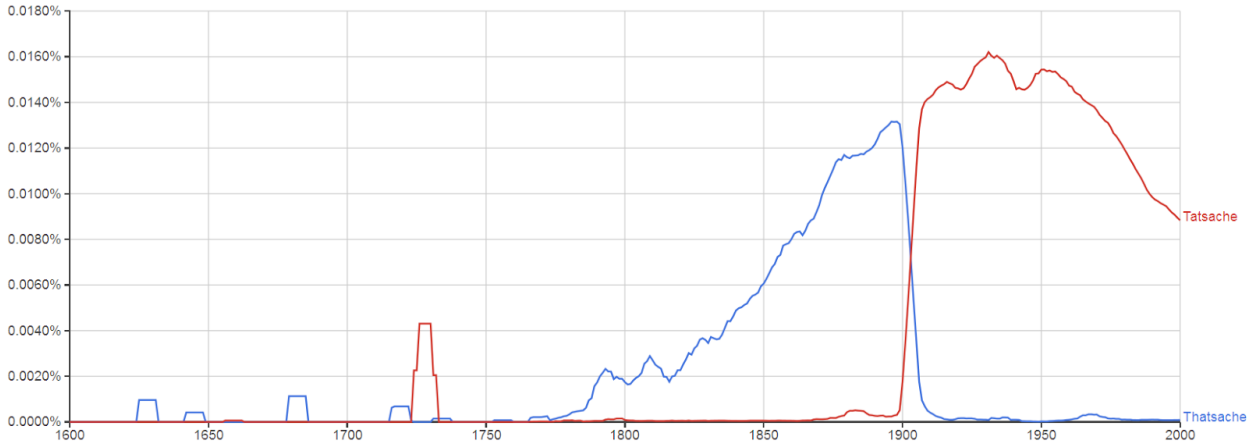
Es ist damit zu rechnen, dass sich die Lage der auf Referenzkorpora basierten zeitlichen Verlaufskurven künftig weiter verbessern wird. Zum einen liegt dies daran, dass immer mehr historische Volltexte in hoher Qualität entstehen. Diese Bemühungen werden dadurch verstärkt, dass die DFG erst unlängst eine spezifische OCR-Förderlinie aufgelegt hat. Zum anderen hat das DTA für die sich dynamisch verändernden Korpusgrundlagen bereits die geeigneten technischen Lösungen: Das Korpus des DTA wird automatisch im Wochenrhythmus indiziert.

Abbildungen[8]



http://www.deutschestextarchiv.de/search/ddc/lemmata/?lemma=Tatsache&mode=extended;norm=date%2Bclass&smooth=spline&single=0&grand=1&slice=10&prune=1&window=3&wbase=0&logavg=0&logscale=0&xrange=1740%3A2000&totals=0

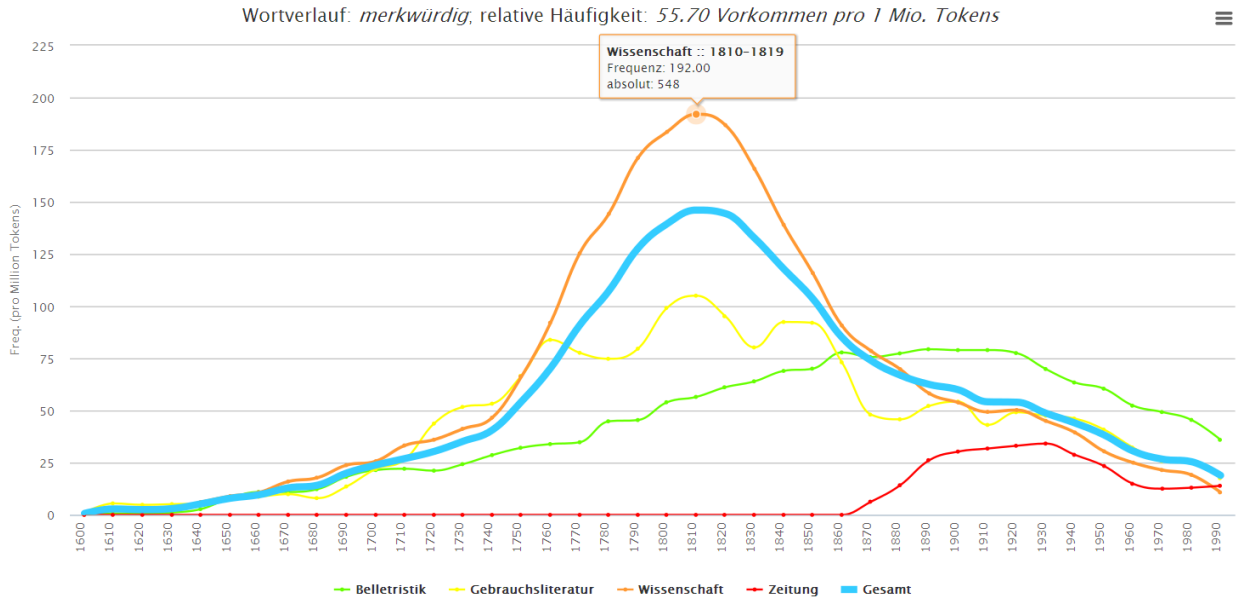
Abb._1: DTA-DWDS-Histogramm „Tatsache“



https://books.google.com/ngrams/graph?content=Thatsache%2CTatsache&year_start=1600&year_end=2000&corpus=20&smoothing=3&share=&direct_url=t1%3B%2CThatsache%3B%2Cc0%3B.t1%3B%2CTatsache%3B%2Cc0

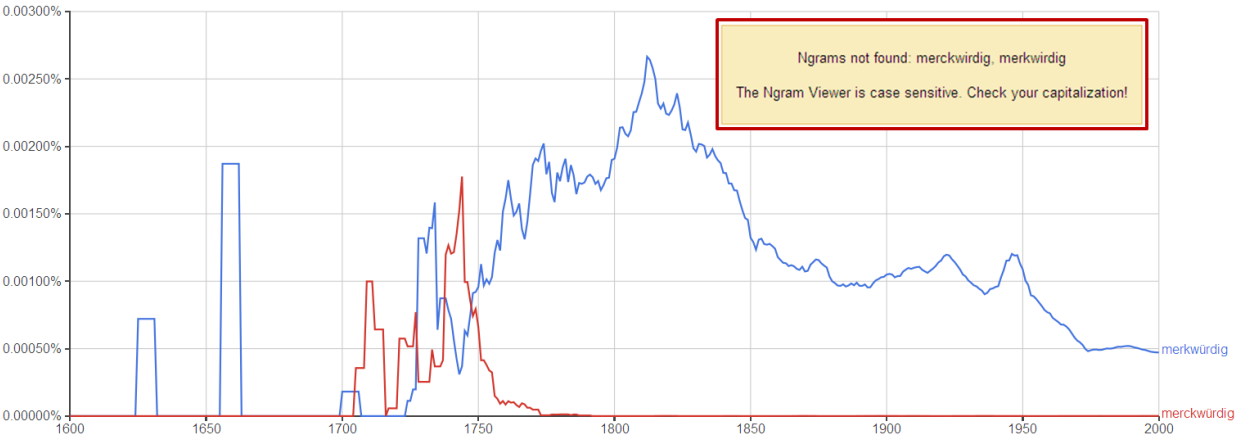
Abb._2: Google-Ngram-Histogramm „Tatsache,Thatsache“

[8] Zu den gewählten Parametern der Kurvengenerierung siehe die jeweils angegebene URL.



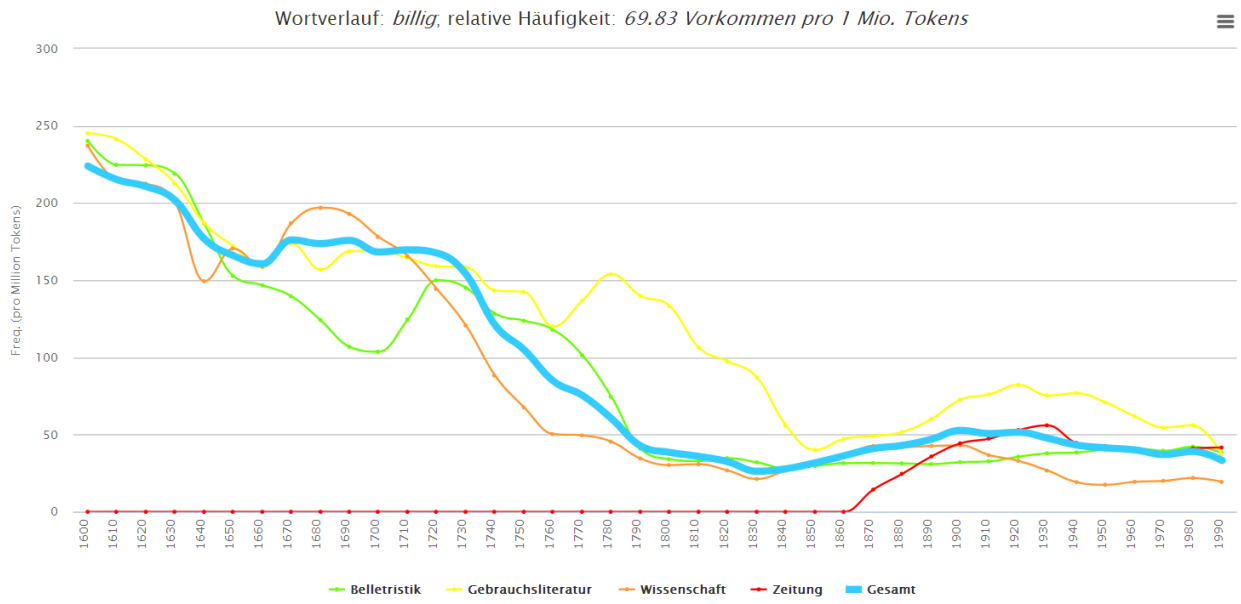
http://www.deutschestextarchiv.de/search/ddc/lemmata/?lemma=merkw%C3%BCrdig&mode=extended;norm=date%2Bclass&smooth=spline&single=0&grand=1&slice=10&prune=1&window=3&wbase=0&logavg=0&logscale=0&xrange=1600%3A2000&totals=0

Abb._3: DTA-DWDS-Histogramm „merkwürdig“



https://books.google.com/ngrams/graph?content=merkw%C3%BCrdig%2Cmerckw%C3%BCrdig%2Cmerckwirdig%2Cmerkwirdig&year_start=1600&year_end=2000&corpus=20&smoothing=3&share=&direct_url=t1%3B%2Cmerkw%C3%BCrdig%3B%2Cc0%3B.t1%3B%2Cmerckw%C3%BCrdig%3B%2Cc0

Abb._4: Google-Ngram-Histogramm „merkwürdig,merckwürdig,merckwirdig,merkwirdig“



http://www.deutschestextarchiv.de/search/ddc/lemmata/?lemma=billig&mode=extended;norm=date%2Bclass&smooth=spline&single=0&grand=1&slice=10&prune=1&window=3&wbase=0&logavg=0&logscale=0&xrange=1600%3A2000&totals=0

Abb._5: DTA-DWDS-Histogramm „billig“

https://books.google.com/ngrams/graph?content=billich%2Cbillig&year_start=1600&year_end=2000&corpus=20&smoothing=3&share=&direct_url=t1%3B%2Cbillich%3B%2Cc0%3B.t1%3B%2Cbillig%3B%2Cc0

Abb._6: Google-Ngram-Histogramm „billig“

BIllich	biilich	billicherem	billigerer
BJllich	bil-lig	billicherer	billigeres
BJllig	biliigen	billichers	billigern
Billich	bill'g	billiches	billiges
Billiche	bill'ge	billichst	billiglich
Billicher	bill'ger	billichste	billigs
Billig	billg	billichsten	billigst
Billige	billg'	billig	billigste
Billiger	billge	billig'	billigstem
Billigere	billgem	billig's	billigsten
Billigeren	billgen	billige	billigster
Billigeres	billich	billigem	billigstes
Billiges	billiche	billigen	billing
Billigste	billichem	billiger	pillichen
Billigsten	billichen	billigere	pillicher
allerbillichsten	billicher	billigerem	pillig
allerbillichster	billichere	billigeren	

http://kaskade.dwds.de/dstar/dta/dstar.perl?fmt=expand-html&q=billig&x=Token

Abb._7: im DTA belegte Expansions- und Flexionsformen des Lemma „billig“

Bibliographie

Geyken 2007: Alexander Geyken: *The DWDS corpus – A reference corpus for the German language of the 20th century*. In: Fellbaum, Christiane (Hg.): Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies. London: Continuum Press, 2007, S. 23–40.

Geyken et al. 2011: Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, Frank Wiegand: *Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv*. In: Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010. Beiträge der Tagung. Hrsg. von Silke Schomburg, Claus Leggewie, Henning Lobin und Cornelius Puschmann. 2., ergänzte Fassung. hbz, 2011, S. 157–161. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159

Geyken et al. 2012: Alexander Geyken, Susanne Haaf, Frank Wiegand: *The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora*. In: 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference. Hrsg. von Jeremy Jancsary. Wien, 2012 (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf

Geyken 2013: Alexander Geyken: *Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv*. In: Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011. Hrsg. von Ingelore Hafemann, Berlin 2013, S. 221–234. urn:nbn:de:kobv:b4-opus-24424

Haaf et al., forthcoming: Susanne Haaf, Alexander Geyken, Frank Wiegand: *The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources*. To appear in: Journal of the Text Encoding Initiative (jTEI), Issue 8. [Abstract des korrespondierenden Vortrags im Rahmen der Veranstaltung: TEI Conference and Members Meeting, 2.–5. Oktober 2013, Sapienza, Università di Roma (IT): http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C137]

Jurish 2013: Bryan Jurish: *Canonicalizing the Deutsches Textarchiv.* In: Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen

Akademie der Wissenschaften, 12.–13. Dezember 2011. Hrsg. von Ingelore Hafemann, Berlin 2013, S. 235–244. urn:nbn:de:kobv:b4-opus-24433

Jurish et al. 2014: Bryan Jurish, Christian Thomas, Frank Wiegand: *Querying the Deutsches Textarchiv*. In: Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities (co-located with iConference 2014, Berlin, Germany, 4th March, 2014). Hrsg. von Udo Kruschwitz, Frank Hopfgartner, Cathal Gurrin, Berlin 2014, S. 25–30. http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf

Münter 1772: Balthasar Münter: Bekehrungsgeschichte des vormaligen Grafen [...] Johann Friederich Struensee. Kopenhagen, 1772. In: Deutsches Textarchiv, www.deutschestextarchiv.de/muenter_bekehren_1772, abgerufen am 07.11.2014.

Citation Segmentation from Sparse & Noisy Data:

An Unsupervised Joint Inference Approach with Markov Logic Networks

Dustin Heckmann, Anette Frank
Institut für Computerlinguistik
Universität Heidelberg
{heckmann,frank}@cl.uni-heidelberg.de

Matthias Arnold, Peter Gietz, Christian Roth
Exzellenzcluster "Asien and Europa in a Global Context"
Universität Heidelberg
{arnold,croth}@asia-europe.uni-heidelberg.de, gietz@daasi.de

Bibliographische Daten sind das Rückgrat der wissenschaftlichen Forschung. Liegen sie nur in Druckform vor, können sie ihr Potential nicht voll entfalten. Erst ihre Speicherung in bibliographischen (Online-) Datenbanken öffnet effiziente Suchmöglichkeiten für den zeitgemäßen und breitgefächerten Einsatz in internationalen Forschungsgemeinschaften. Zu diesem Zweck ist es erforderlich, die Materialien zu digitalisieren und die den bibliographischen Referenzen innewohnende Struktur automatisch zu erkennen, indem einzelne Felder (z. B. Autor, Titel, Erscheinungsort) extrahiert werden.

In diesem Paper präsentieren wir ein Verfahren für die Zitationsanalyse auf spärlichen und verrauschten OCR-Daten[1]. Als Datenbasis nutzen wir den *Turkologischen Anzeiger* (TA)[2], ein wichtiges Referenzwerk für die Turkologie und die Osmanistik. Der *Turkologische Anzeiger* ist eine systematische Bibliographie in 28 Bänden, die bisher nur in gedruckter Form verfügbar war. Er umfasst Einträge in vielen verschiedenen Sprachen, einschließlich Transkriptionen aus dem

[1] Das hier vorgestellte Projekt ist eine Zusammenarbeit mehrerer Institutionen der Universität Heidelberg, dem Seminar für Sprachen und Kulturen des Vorderen Orients (Islamwissenschaft), dem Institut für Computerlinguistik und der Heidelberg Research Architecture, der Abteilung Digitale Geisteswissenschaften am Exzellenzcluster "Asia and Europe in a Global Context". Gestützt auf eine Vereinbarung mit dem Redaktionsausschuss des TA, die Datenbank als Open Access Ressource der Öffentlichkeit zur Verfügung zu stellen, formierte sich eine Arbeitsgruppe, die beim Exzellenzcluster erfolgreich Mittel zur Umsetzung einwerben konnte. Die Ergebnisse des Projektes sowie die Online Datenbank selbst können unter dieser Adresse aufgerufen werden: http://kjc-fs2.kjc.uni-heidelberg.de:8000/.

[2] Siehe Hazai & Kellner-Heinkele (1975) sowie die Online Datenbank.

Arabischen und aus Sprachen mit kyrillischem Alphabet, wobei selbst einzelne Einträge aus Abschnitten in verschiedenen Sprachen bestehen können.

Bestehende Ansätze für die Zitationsanalyse stützen sich auf sprachspezifische lexikalische Daten und Mehrfachvorkommen von Zitationen in Online-Publikationsverzeichnissen. Bei der Verarbeitung mehrsprachiger Daten wird die Nutzung sprachspezifischen Wissens jedoch erschwert. Darüber hinaus enthalten in sich abgeschlossene Datenquellen, wie gedruckte Bibliographien, naturgemäß wenige wiederkehrende Referenzen, was die Berufung auf Datenredundanz verhindert.

Mit den in hoher Auflösung digitalisierten und mit der OCR Software Abbyy FineReader Professional in Volltext umgewandelten Zitationen des *Turkologischen Anzeigers* verhielt es sich genauso. Der Mangel an Redundanz in den Zitationen, Erkennungsfehler in der OCR-Volltextumwandlung sowie Inkonsistenzen in der Zitationsstruktur der Druckausgabe erschwerten die Anwendung bestehender statistischer Ansätze für Zitationsanalyse.

Nach Beispiel von Poon & Domingos (2007) stützt sich unser Verfahren auf Markov Logic Networks (MLN), einem Framework für Statistical Relational Learning, das Prädikatenlogik mit probabilistischer Modellierung verbindet (Richardson & Domingos, 2006). Die Formulierung in Prädikatenlogik bietet hohe Ausdrucksstärke und Flexibilität. Dadurch kann die Zitationsanalyse auf die besonderen Konventionen einer Bibliographie -- in unserem Fall dem *Turkologischen Anzeiger* -- zugeschnitten werden. MLNs können auf der Basis annotierter Daten in einem überwachten Lernverfahren trainiert werden. Sie können aber mit Hilfe manuell gewichteter Regeln auch in einem nicht-überwachten Verfahren angewandt werden. Bei fehlenden Trainingsdaten und Mangel an Datenredundanz in bibliographischen Quellen bieten MLNs ein attraktives Framework für die Zitationsanalyse durch Verwendung unüberwachter Verfahren.

In unserer Arbeit präsentieren wir ein Verfahren für Zitationsanalyse mittels Markov Logic Networks und Joint Inference. Wir wenden dieses auf eine umfangreiche mehrsprachige Bibliographie an, die aus verrauschtem OCR-Output gewonnen wurde. Die zu bewältigenden Probleme beinhalten insbesondere Rauschen durch OCR-Fehler, die Mehrsprachigkeit der einzelnen Einträge, komplexe Zitationsstrukturen, Inkonsistenzen in den Zitationen, sowie Mangel an Redundanz. Unser Joint Inference Verfahren erweitert den Ansatz von Poon & Domingos (2007), indem Redundanz auf Feldebene ausgenutzt wird. Dadurch sind wir in der Lage, dem Fehlen redundanter Zitationen beizukommen.

Die Ergebnisse einer einfachen MLN Formalisierung für einzelne Datensätze übertreffen sowohl die guten Referenzdaten der traditionellen Herangehensweise mit auf regulären Ausdrücken basierendem Parsing, als auch die des überwachten statistischen Ansatzes unter Nutzung von Conditional Random Fields (CRF). Werden auch joint references auf *Entitäts*- und *Feldebene* einbezogen, steigt die

Perfomanz bei recall und precision, wobei Joint Inference auf Feldebene die besten Ergebnisse produziert. Dabei nutzt unser Verfahren manuell gewichtete Regeln und ist komplett unüberwacht.

Unsere Evaluationsergebnisse zeigen, dass unsere MLN Formalisierung angesichts besonderer Herausforderungen bei der Analyse von Zitationen aus einer digitalisierten Bibliographie sowohl regelbasierte als auch moderne statistische Verfahren übertreffen. Auf unseren Testdaten erreichen wir einen F_1-Wert von 88% für exakte Übereinstimmung von Feldern, was einen Zuwachs von 21.9% gegenüber einem CRF-basierten Vergleichssystem darstellt.

Zusammenfassend kann festgestellt werden, dass wir im Gegensatz zu früheren Arbeiten Datensätze aus dem Bereich der Digitalen Geisteswissenschaften adressieren, die verrauscht und nur gering strukturiert sind. Dabei erweitert unsere Methode den Ansatz von Poon & Domingos (2007), indem wir *Joint Inference* auf *Feldebene* einsetzen. Dadurch sind wir in der Lage, ohne wiederkehrende Referenzen und mit verrauschten Daten umzugehen. Unser Verfahren ist komplett unüberwacht und benötigt keine annotierten Trainingsdaten. Das von uns erarbeitete Regelwerk kann auch auf die Verarbeitung anderer Bibliographien oder digitale Quellen, wie historische Wörterbücher oder Enzyklopädien, angewandt werden.

Die Ergebnisse unseres Projektes stellen wir auf der Website *Turkology Annual Online*[3] der Öffentlichkeit zur Verfügung. Das Interface bietet sowohl Such- und Browse-Funktionalitäten für den TA an. Bibliographische Subfelder (wie Titel oder Autor) sind explizit gemacht und können als Suchkriterien oder zum Sortieren von Ergebnissen genutzt werden, Querverweise sind als Hyperlinks ausgegeben. Referenzen können selektiert und in verschiedene bibliographische Formate wie beispielsweise BibTeX exportiert werden.

Abbildung 1 zeigt einen Beispieldatensatz, wie er im Webinterface dargestellt wird.
Abbildung 2 zeigt ein Beispiel für einen Datensatz im Originalscan.

[3] http://kjc-fs2.kjc.uni-heidelberg.de:8000/

Hazai, G. and Kellner-Heinkele, B. eds. (1975ff). *Turkologischer Anzeiger*, Universität Wien. Institut für Orientalistik and Universität Wien. Orientalisches Institut. Available at: http://orientalistik.univie.ac.at/forschung/publikationen/turkologischer-anzeiger, [Accessed: July 19, 2013].

Poon, H. and Domingos, P. (2007). Joint Inference in Information Extraction. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press.

Richardson, M. and Domingos, P. (2006). Markov Logic Networks. *Machine Learning*, 62(1), pp.107–136.

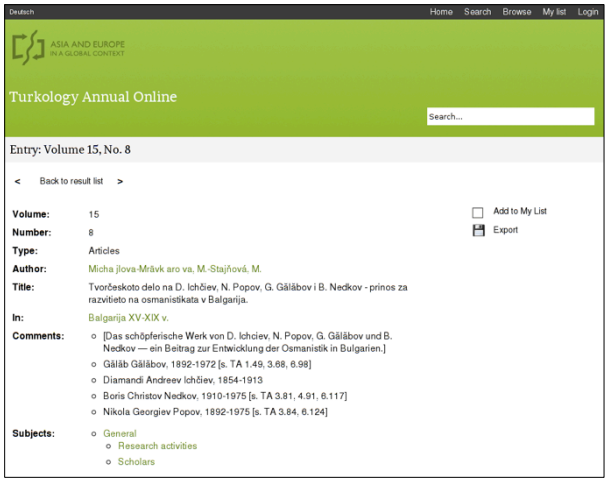


Abbildung 1: Turkologischer Anzeiger Online: Darstellung eines einzelnen Datensatzes.

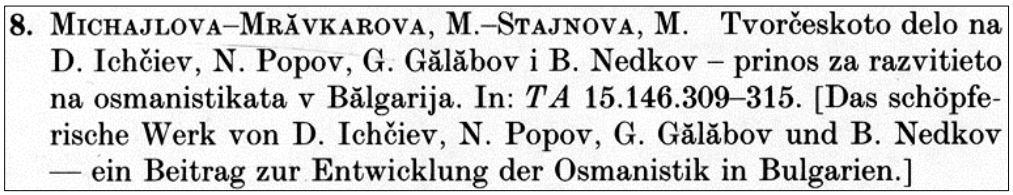

8. MICHAJLOVA–MRĂVKAROVA, M.–STAJNOVA, M. Tvorčeskoto delo na D. Ichčiev, N. Popov, G. Gălăbov i B. Nedkov – prinos za razvitieto na osmanistikata v Bălgarija. In: *TA* 15.146.309–315. [Das schöpferische Werk von D. Ichčiev, N. Popov, G. Gălăbov und B. Nedkov — ein Beitrag zur Entwicklung der Osmanistik in Bulgarien.]

Abbildung 2: Turkologischer Anzeiger: Beispiel für einen Datensatz im Originalscan.

Parameter zur Klassifizierung stilistischer Varianz bei E-Mails

Ulrike Krieg-Holz

Institut für Germanistik
Universität Leipzig

ulrike.krieg-holz@uni-leipzig.de

Udo Hahn

Institut für Germanistische Sprachwissenschaft
Friedrich Schiller-Universität Jena

udo.hahn@uni-jena.de

Zusammenfassung. Auf der empirischen Grundlage eines im Aufbau befindlichen deutschsprachigen E-Mail-Korpus werden Aspekte der stilistischen Varianz innerhalb der Kommunikationsform ‚E-Mail' untersucht.

Mit der wachsenden Bedeutung der empirischen Fundierung linguistischer Analysen ist auch für die deutsche Sprache eine große Vielfalt von Korpora entstanden.[1] Aus linguistischer Sicht stand dabei lange die Idee der Erfassung möglichst vielfältiger und großvolumiger Mengen von sprachlichen Rohdaten im Vordergrund,[2] die auch dem Anspruch genügen sollten, den Sprachgebrauch weitgehend „repräsentativ" abzubilden (exemplarisch gilt dies etwa für das DeReKo-Korpus[3] [KBKW10] sowie das DWDS-Kernkorpus[4] [Geyk07]). In diesen Korpora spielen nach wie vor Zeitungstexte die weithin dominierende Rolle, ergänzt durch überwiegend literarisch-belletristische Quellen und eher geringe Vorkommen von Gebrauchstexten (Kochrezepte, Montageanleitungen usw.). Trotz der angestrebten Vielfalt von Textsorten liegt der Schwerpunkt der in diesen Korpora auftretenden Texte eindeutig im Bereich der deutschen Standardsprache mit klarer Ausrichtung auf formelle Kommunikation.

Auf dem entgegengesetzten Pol des Kontinuums formeller vs. informeller Sprachgebrauch können aktuelle Arbeiten zur Erhebung von Korpora geschriebener Alltagssprache eingeordnet werden [Stor13]. Die Sammlung und Annotation informeller Sprache ist derzeit für das Deutsche am weitesten bei DeRiK gediehen [BEG+13], einem Korpus zur Erfassung computervermittelter Kommunikation (Blogs, Chats usw.) als Ergänzung des DWDS-Kernkorpus. In dieser Textkollektion werden jedoch explizit keine E-Mails berücksichtigt.

Wegen ihrer bedeutenden Rolle im öffentlichen wie privaten Kommunikationsumfeld moderner IT-basierter Gesellschaften ist damit ein bedeutsames korpuslinguistisches Desiderat beschrieben – reichen doch die Textsorten der Kommunikationsform ‚E-Mail' inhaltlich mittlerweile von zum Teil hochgradig formalisierten professionellen Diskursen (Geschäfts- bzw. Verwaltungspost) bis hin zu gänzlich persönlichen und somit rein informellen Interaktionen. Deshalb werden sie in Bezug auf ihre

[1] Eine aktuelle Übersicht enthält http://de.clarin.eu/de/sprachressourcen/corpora.html (letzter Aufruf: 3.11.2014)
[2] Die sprachlichen Rohdaten sind zum Teil bereits auch automatisch lemmatisiert und nach Wortarten annotiert (POS-Tagging).
[3] http://www1.ids-mannheim.de/kl/projekte/korpora/ (letzter Aufruf: 3.11.2014)
[4] http://www.dwds.de/ressourcen/korpora/ (letzter Aufruf: 3.11.2014)

stilistische Ausformung äußerst unterschiedlich gestaltet und bilden damit – so unsere leitende Hypothese – innerhalb einer Kommunikationsform eine sehr große Bandbreite der performativen Varianz auf dem gesamten Kontinuum formeller und informeller Sprache ab. E-Mails eignen sich somit ganz besonders für empirisch fundierte stilistische Untersuchungen.

Im Vergleich zu literaturwissenschaftlichen Stilanalysen, in deren Mittelpunkt die stilistische Figuriertheit von Texten oder auch Autorenstile stehen, fokussiert der sprachwissenschaftliche Stilbegriff ganz allgemein die Spezifik der sprachlichen Ausgestaltung von Textstrukturen. Diese Spezifik sprachlicher Formulierungen resultiert prinzipiell aus der Möglichkeit innerhalb von im Sprachsystem angelegten Varianten auszuwählen. Stil ist deshalb als ein Phänomen der Wahl anzusehen und ein Ergebnis von Entscheidungsprozessen, die sich einerseits an Vorgegebenem, Prototypischem und Musterhaftem orientieren, andererseits immer auch eigenständige Umsetzungen in Verbindung mit individualstilistischen Merkmalen darstellen. Derartige Wahlentscheidungen sind in sämtlichen Kommunikationsformen von größter pragmatischer Relevanz, weil sie das kommunikative Handeln ganz entscheidend prägen. Zum einen können sprachliche Handlungen desselben Typs auf verschiedene Weise durchgeführt werden, zum anderen unterscheiden sich sprachliche Handlungen verschiedenen Typs in stilistischer Hinsicht. Dadurch können sowohl Textproduktions- als auch Textrezeptionsprozesse erheblich beeinflusst werden, denn das Wissen über Stil ist Teil der Textsortenkompetenz, also der Fähigkeit, auf der Grundlage eines mehr oder weniger bewussten Wissens über Textsortenqualitäten in der Kommunikation operieren zu können.

Das Ziel der sprachwissenschaftlichen Stilistik besteht darin, innerhalb von Texten und kommunikativen Zusammenhängen diejenigen Elemente und Strukturen aufzudecken, mit denen das Spezifische der sprachlichen Gestaltung einer kommunikativen Handlung charakterisiert werden kann; es gilt also, die Träger stilistischer Information zu charakterisieren. Dazu hat sich ein terminologisches Inventar herausgebildet, das zum Teil durch verschiedene Beschreibungsansätze geprägt und entsprechend ungleich stark etabliert ist. So bezeichnen die Begriffe ‚Stilelement' und ‚Stilzug' ursprünglich grundlegende funktionalstilistische Kategorien (vgl. Fleischer et al. 1993 [FlMS93, S. 27]), wobei sich Stilelemente immer auf einzelne sprachliche Mittel innerhalb eines Relationsgefüges beziehen (z.B. markierte lexikalische Elemente, eine spezifische Wortstellung oder bestimmte Elemente auf der lautlichen und graphemisch-ikonischen Ebene). Demgegenüber werden als Stilzüge bestimmte Stilstrukturen zusammengefasst, die aus einer typischen Kombination verschiedener Stilelemente resultieren. Sie lassen sich als Bündel miteinander vorkommender, kookkuierender Merkmale auffassen, die als eine bedeutsame, sinnhafte Gestalt interpretiert werden können (vgl. Selting & Hinnenkamp 1989 [SeHi89, S.5f.]; Sandig 2006 [Sand06, S. 54f.]). Für die Beschreibung derartiger Merkmalbündel bzw. Stilzüge steht eine breite, bisher systematisch kaum erfasste Vielfalt an Kriterien zur Verfügung, aus denen hier diejenigen gefiltert werden sollen, die innerhalb der Textsorten der E-Mail-Kommunikation distinktiv wirken.

Um unsere Arbeiten auf eine solide empirische Basis zu stellen, wird derzeit am Institut für Germanistik der Universität Leipzig und am Institut für Germanistische Sprachwissenschaft der Friedrich Schiller-Universität Jena am Aufbau eines umfassenden **Ko**rpus **d**eutschsprachiger **E**mails (KodE Alltag) gearbeitet. Alle bislang verfügbaren E-Mail-Korpora (hier ist vor allem das Enron-Korpus [KlYa04] zu erwähnen) erfassen nur die englische Sprache, für das Deutsche existiert bislang kein vergleichbares Korpus. Auch diese Lücke wollen wir mit unseren Arbeiten füllen. In unserem Beitrag werden wir die leitenden Entwurfsprinzipien für und den konkreten Aufbau von KodE Alltag sowie den aktuellen

Stand der Datenerhebung im Detail beschreiben. Hierzu gehören auch Ausführungen zu Aspekten des Urheberrechts an E-Mails und zu ihrer (semi-automatischen) Anonymisierung.

Im Zentrum unseres Beitrags werden jedoch erste Befunde zur Klassifizierung der stilistischen Varianz in KodE Alltag stehen. In diesem Zusammenhang unterscheiden wir formellen vs. informellen sprachlichen Stil graduell anhand einer Fülle von Parametern, die jeweils in Verbindung mit anderen zur Ausprägung entsprechender Stilzüge bzw. Merkmalbündel beitragen können. Dazu gehören

- die lexikalische Vielfalt (in Bezug auf kanonische Lexika) und Variabilität (in Bezug auf alternative Korpora wie DeReKo/DWDS oder DeRiK),
- die Orientierung an Mündlichkeit (z.B. klitisierte oder reduzierte Wortformen, spontansprachliche Syntax),
- die Distribution von Abkürzungen und Icons,
- Abweichungen in Formenbildung, Orthographie und Interpunktion usw.

Anhand solcher Parameter wird eine Untergliederung von KodE Alltag in Subkorpora möglich sein, die unterschiedliche Grade an Formalisierung von Sprache in Form von Stilzügen ausdrücken. Diese Stilzüge werden dann als Grundlage für das Training von statistischen Klassifikatoren genutzt, um ungesehene Sprachdaten entlang unterschiedlicher Stilformen automatisch klassifizieren zu können.

Unsere Arbeiten verbinden also germanistische Stilforschung und Korpuslinguistik mit Verfahren der automatischen Textklassifikation aus dem Bereich der Computerlinguistik. Die von uns erarbeiteten Ergebnisse werden wir im Rahmen von Fragestellungen aus dem Bereich der Forensischen Linguistik (Autorenerkennung durch stilistische Write-Prints) auf Anwendbarkeit prüfen.

Literatur

[BEG+13] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531-537.

[FlMS93] Wolfgang Fleischer, Georg Michel & Günter Starke (1993). *Stilistik der deutschen Gegenwartssprache.* Tübingen: Niemeyer.

[Geyk07] Alexander Geyken (2007). The DWDS corpus: a reference corpus for the German language of the 20th century, In: Christiane Fellbaum (Ed.), *Collocations and Idioms*. London: Continuum, pp. 23–40.

[KBKW10] Marc Kupietz, Cyril Belica, Holger Keibel & Andreas Witt (2010). The German reference corpus DeReKo: a primordial sample for linguistic research. In: *LREC '10 – Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta, May 2010, pp. 1848-1854.

[KlYa04] Bryan Klimt & Yiming Yang (2004). The Enron corpus: a new dataset for email classification research. In: Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti & Dino Pedreschi (Eds.), *ECML 2004 - Proceedings of the 15th European Conference on Machine Learning*, Pisa, Italy, September 20-24, 2004. Berlin, Heidelberg: Springer, pp.217–226 (Lecture Notes in Computer Science, 3201)

[Sand06] Barbara Sandig (2006). *Textstilistik des Deutschen*. Berlin, New York: de Gruyter.

[SeHi89] Margret Selting & Volker Hinnenkamp (1989). Stil und Stilisierung in der interpretativen Soziolinguistik. In: Volker Hinnenkamp & Margret Selting (Eds.), *Stil und Stilisierung*. Tübingen: Niemeyer, pp. 1-23.

[Stor13] Angelika Storrer (2013). Sprachstil und Sprachvariation in sozialen Netzwerken, in: Barbara Frank-Job, Alexander Mehler & Tilmann Sutter (Eds.), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*, Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 329–364.

Das Wissen der Bilder. Spielarten des digitalen Annotierens

Harald Lordick, Salomon-Ludwig-Steinheim-Institut für deutsch-jüdische Geschichte, Essen
Stefan Schmunk, SUB Göttingen (DARIAH-DE)
Sibylle Söring, SUB Göttingen (TextGrid)

Das Annotieren als eine Form spontanen und/oder systematischen Kommentierens und Explizierens ist traditionell ein fester Bestandteil geistes- und kulturwissenschaftlichen Arbeitens. Nicht nur ihre prominenteste Spielart, die Fußnote, erfährt dabei im digitalen Medium einen radikalen Wandel. Annotiert, kommentiert, erläutert, verwiesen, ergänzt, vermerkt, aber auch strukturiert, geordnet, (re-)organisiert werden jenseits der Gutenberg-Galaxis nicht mehr nur Texte, sondern auch Bilder und audiovisuelle Medien, öffnen sich neue, auch kollaborative Wissensräume jenseits der linearen Struktur, die die gedruckte Annotation vorgab. Zudem entstehen - z.B. durch algorithmische Verarbeitungen wie etwa die automatisierte Mustererkennung - neue, nicht mehr ausschließlich manuell erzeugte Annotate, die wiederum neues Wissen generieren. In digitaler Form in den von John Unsworth aufgestellten “Scholarly Primitives” dokumentiert, hat das Annotieren als geisteswissenschaftliche Praxis und Methode mit der Überführung von Forschungsgegenständen in digitale Formate und dem Generieren genuin digitalen Materials (born digital) nicht zuletzt auch im Zuge der Entwicklungen im Bereich der Linked Open Data disziplinübergreifend noch an Relevanz für die Digital Humanities gewonnen. Zeitgemäßes Annotatieren verspricht Eigenschaften wie *punktgenaue Referenzierung* oder *bidirektionale Verknüpfung* (*backlink*). Initiativen wie *Open Annotation*, *annotatorjs.org* oder *Hypothes.is* sehen dies Thema im Zentrum künftiger Webentwicklungen.

Annotationen in ihrem digitalen Lebenszyklus 'einzufangen', ihnen einen Ort zu bieten, an dem sie methodisch angemessen erschaffen werden können, kontrolliert erreichbar bleiben und nachhaltig gespeichert werden, ist auch deshalb eine Herausforderung, weil die vielfältige Motivation, die Zielsetzung des annotierenden Forschers, der Grund der Annotation höchst relevant sind für den Entwurf und die konkrekte Implementation von Annotationssystemen, die im Ergebnis infrastruktureller Unterstützung bedürfen.

Die Sektion “Das Wissen der Bilder. Spielarten des digitalen Annotierens“ hat das Ziel, anhand der Virtuellen Forschungsinfrastrukturen TextGrid und DARIAH-DE aufzuzeigen, dass Werkzeuge und Services des digitalen Annotierens, die im Rahmen einer Digitalen Forschungsumgebung entwickelt bzw. eingesetzt werden und in einer digitalen Forschungsinfrastruktur implementiert sind, genutzt werden können, um vielfältigste (disziplinäre) Forschungsfragen zu beantworten. Im Rahmen der Sektion möchten wir unterschiedliche digitale Verfahren und Methoden des Annotierens vorstellen und zugleich aufzeigen, dass diese nicht nur in unterschiedlichen fachwissenschaftlichen Forschungskontexten verwendet werden, sondern zugleich auch mediale Grenzen des “Textuell-Bildlichen” überwinden können. Insbesondere soll der Frage nachgegangen werden, auf welche Weise Annotationen von Bildern, Digitalisaten und Fotografien - z.B. Zeichen, Muster, Areas etc. - erstellt und wie diese technologisch so aufbereitet werden können, dass sie such- und recherchierbar, nach internationalen Daten- und Metadatenstandards gespeichert und nachgenutzt werden können. Im Mittelpunkt steht die Frage, wie sich aus Bildern maschinenlesbare Informationen gewinnen lassen und diese zugleich weiter verarbeitet werden können. Hierbei wird grundsätzlich von der These ausgegangen, dass - unabhängig von der je spezifischen fachwissenschaftlichen Fragestellung - bestimmte Methoden, Verfahren und Technologien des digitalen Annotierens auf unterschiedlichstes Datenmaterial angewandt werden können.

Nach einem einleitenden Vortrag, der diesen Spannungsbogen thematisiert, wird in drei 20minütigen Vorträgen der Einsatz von digitalen Annotationstools in unterschiedlichsten Fachdisziplinen und im Kontext besonders Bilddaten-intensiver Forschungsvorhaben beleuchtet:

1. Einleitung (Harald Lordick, Stefan Schmunk, Sibylle Söring)
2. Tools und Standards für die Bilderflut: Image Services und Annotationen mit IIIF, OpenAnnotation, Mirador, digilib und TextGrid (Robert Casties, Ubbo Veentjer)
3. Anwendungsbeispiel / Projektwerkstatt: Text – Bild – Inschrift. Hieroglyphenschrift und Sprachen der Maya annotieren (Christian Prager, Frauke Sachse)
4. Kodikologie meets Topologie - topologisches Retrieval als innovative, nicht-textuelle Form der Bedeutungsextraktion aus automatisch gewonnenen oder manuell erstellten Bildannotationen (Jochen Graf)

Aus unterschiedlicher, sowohl technischer als auch semantisch-methodischer Perspektive, gehen die einzelnen Beiträge dabei der Frage nach dem Wandel, aber auch nach den Möglichkeiten der Annotation als geisteswissenschaftlichem Verfahren im digitalen Medium nach. Welche (auch fächerübergreifenden) Methoden und Verfahren werden mit den entsprechenden Tools unterstützt bzw. erleichtert? Welche Anwendungsszenarien werden bereits praktiziert, wo sind derzeit noch Desiderate zu formulieren? Wie können diese - auch disziplinen-unabhängig - umgesetzt werden? Und nicht zuletzt: Wie funktioniert der Brückenschlag zwischen Text und Bild, zwischen Information und Interpretation? Dabei werden verschiedene Lösungsansätze von TextGrid und DARIAH-DE aufgezeigt, um die disziplinübergreifenden Synergien herauszuarbeiten, die eine digitale Forschungsinfrastruktur ermöglicht.

Inhalt

1. Tools und Standards für die Bilderflut: Image Services und Annotationen mit IIIF, OpenAnnotation, Mirador, digilib und TextGrid

Robert Casties, Max Planck-Institut für Wissenschaftsgeschichte, Berlin
Ubbo Veentjer, SUB Göttingen

Digitale Bilddaten fallen im Forschungsprozess nahezu aller geisteswissenschaftlicher Disziplinen an – auch solcher, die primär textbasierte Quellen zum Gegenstand haben. Das Spektrum reicht dabei von Buch- und Manuskriptdigitalisaten über digitalisierte Gemälde und Zeichnungen oder Abbildungen von Sammlungsobjekten bis hin zu Fotografien z.B. historischer Inschriften. Im Vortrag sollen technische Standards und Lösungen für Image Services und Annotationen präsentiert werden, die neue Arbeitsweisen mit Bildern im Netz ermöglichen, indem sie erlauben, Bilder unabhängig von ihrem Speicherort zu betrachten, detailliert zu referenzieren, zu annotieren und eigene Präsentationen und Editionen vorhandener Bilder zu erzeugen und zu publizieren.

Am Beispiel der Virtuellen Forschungsumgebung TextGrid[1] und dem hier angebotenen Image-Dienst digilib[2] soll gezeigt werden, wie Bildservices die Weiter- und Nachnutzbarkeit von Bilddaten erhöhen, und wie durch die Integration interoperabler Standards für Bild- und Annotationsserver bessere Werkzeuge bereitgestellt werden können. In diesem Zusammenhang werden auch die aktuellen Standards des International Image Interoperability Framework (IIIF)[3] für Bild- und Metadaten-Server und die verwandten Standards OpenAnnotation[4] und SharedCanvas[5] vorgestellt.

Bilder, die in digitalen Archiven abgelegt werden, sind oft nicht für die Darstellung im Internet geeignet. Wunschformate für die Archivierung von Bilddateien sind - im Kontext der DFG-Praxisregeln „Digitalisierung“[6] - TIFF oder JPEG-2000, welche von Webbrowsern jedoch nicht dargestellt werden. Des weiteren werden möglichst hoch aufgelöste Bilder für die Archivierung angestrebt, während an mobile Geräte Bilder mit geringerer Auflösung ausgeliefert werden sollen. An dieser Stelle kommt ein Image-Service ins Spiel, der – wie das in TextGrid integrierte Werkzeug digilib - die originalen, hochaufgelösten Bilddaten in angepassten Auflösungen und web-tauglichen Bildformaten ausliefern kann.

Um Programme zur Darstellung digitaler Objekte nicht für jeden eingesetzten Image-Service anpassen zu müssen, kann der interoperable Standard IIIF-Image-API genutzt werden, der festlegt, wie von einem IIIF-konformen Image-Service Bilder in verschiedenen Größen oder Formaten abgerufen werden können. Der IIIF-Standard beschreibt zudem, in welchem Metadatenformat Kollektionen von Bildern beschrieben werden; außerdem wird ein Format für Bildannotation spezifiziert. Somit können IIIF-konforme Programme - wie etwa der Mirador Viewer[7] - Bildsammlungen verschiedener Institutionen, die Daten im IIIF-Format bereitstellen, darstellen, z.B. aber auch Bilder und Metadaten verschiedener Institutionen im selben Workspace anzeigen und eine Annotation dieser Bilder ermöglichen.

DHd2015 Graz · Sektion „Das Wissen der Bilder. Spielarten des digitalen Annotierens“ · Abstracts Vorträge

[1] http://textgrid.de
[2] http://digilib.sf.net
[3] http://iiif.io
[4] http://www.openannotation.org/
[5] http://iiif.io/model/shared-canvas/
[6] http://www.dfg.de/formulare/12_151/12_151_de.pdf
[7] https://github.com/DMSTech/mirador

Durch die jeweils spezifische Trennung und Standardisierung der Formate und Schnittstellen für Bilder und Metadaten ist es heute für Forscher aller Disziplinen möglich, eigene Bildkollektionen, Editionen und Präsentationen zu erstellen und zu publizieren, aber auch, die Bilder aus unterschiedlichen Repositorien aus aller Welt im Rahmen neuer Forschungskontexte neu zu kombinieren und zu präsentieren.

Im Vortrag soll der Stand der Integration von IIIF bei TextGrid und digilib vorgestellt und ein Ausblick auf die zukünftige Integration von Annotationsmöglichkeiten gegeben werden.

2. Projektwerkstatt / Anwendungsbeispiel: Text – Bild – Inschrift. Hieroglyphenschrift und Sprachen der Maya annotieren

2.1 Digitale Epigraphik - Die Erforschung der Hieroglyphentexte und Bildbotschaften der Maya in der Virtuellen Forschungsumgebung TextGrid

Dr. Christian Prager, Rheinische Friedrich-Wilhelms-Universität Bonn

Die nur teilweise entzifferte Hieroglyphenschrift und Sprache der Mayakultur steht im Mittelpunkt eines Forschungsprojekts der NRW Akademie der Wissenschaften, das in Kooperation zwischen den Universitäten Bonn und Göttingen durchgeführt wird. Ziel ist die Erstellung einer Textdatenbank und ein darauf basierendes Wörterbuch des Klassischen Maya, das zwischen 250 bis 950 n.Chr. auf rund zehntausend Text- und Bildträger überliefert ist. Sie ermöglichen einzigartige Perspektiven auf Sprache, Kultur und Geschichte der vorspanischen Maya. Bis heute fehlen allerdings eine systematische Dokumentation sowie die umfassende Analyse dieser Quellen. Diese erlaubten eine präzise Untersuchung des Maya, indem Textpassagen verglichen werden, Bildinhalte mit Textpassagen korreliert oder die Beschaffenheit oder Funktion eines Textträger in der Inschrift erfasst und damit rätselhafte Textpassagen verständlich werden. Bislang war ein derartig systematisches und vernetztes Arbeiten mit Text, Bild und Informationsträgern nicht möglich, da die notwendige Technologie noch nicht existierte. Im Rahmen des Projekts werden die Quellen systematisch und nach einheitlichen Standards beschrieben, das Ausgangsmaterial auf der Basis von XML maschinenlesbar gemacht und auf diese Weise die Grundlagen für die Kompilation des Wörterbuchs geschaffen. Zu diesem Zweck werden in der VRE TextGrid Tools und Workflows entwickelt, welche I. die Dokumentation der Schrift- und Bildträger mit Aufarbeitung des Forschungsstandes, II. die epigraphisch-linguistische Auswertung der Hieroglyphentexte sowie III. die Edition der Texte mit Transliteration, Transkription und Übersetzung in einem einzigen System ermöglichen. Der Textträger erhält dadurch eine ‚Biographie‘, die eng mit dem Textinhalten verwoben ist und bei der Bedeutungsanalyse von Wörtern berücksichtigt wird.

2.2 Digitale Erschließung und systematische Annotation kolonialer Lexikographien am Beispiel der Mayasprache K'iche'

jun. Prof. Dr. Frauke Sachse (Rheinische Friedrich-Wilhelms-Universität Bonn)
Prof. Dr. Michael Dürr (Zentral- und Landesbibliothek Berlin / Freie Universität Berlin)
Christian Klingler M.A. (Rheinische Friedrich-Wilhelms-Universität Bonn)

Gegenstand des hier vorgestellten Forschungsvorhabens ist die Entwicklung eines Tools zur korpusorientierten Erfassung kolonialzeitlicher Wörterbücher amerindischer Sprachen. An der Abteilung für Altamerikanistik der Universität Bonn wurde mit dem *Tool for Systematic Annotation of Colonial K'iche'* (TSACK) der Prototyp einer Software-Anwendung entwickelt, mit dem sich koloniale Lexikographien der Mayasprache K’iche’ in ein maschinenlesbares Korpus im XML-Standard überführen lassen. Kernproblem der Erfassung sind die nicht standardisierten Orthographien der Wörterbücher, die vereinheitlicht werden müssen, um Lexeme einzeln suchbar zu machen und multiple semantische Korrelationen aufzuzeigen.

Dieser Transkriptionsprozess setzt die semantische Zuordnung und morphologische Analyse der kolonialen Lexikon-Einträge voraus. Da Mayasprachen agglutinierend sind, müssen Wortformen bis auf die Wurzel heruntergebrochen werden, um einzelne Lemmata zu isolieren. TSACK unterstützt den Prozess von Transkription, Lemmatisierung und Glossierung im Rahmen eines halbautomatisierten Auszeichnungsverfahrens. Das Tool vereinfacht den Arbeitsprozess, minimiert die Fehlerquote und beschleunigt im Vergleich zu herkömmlichen XML-Editoren die Auszeichnung größerer Datenmengen. Das hier vorgestellte Forschungsvorhaben will einheitliche Erfassungskriterien für koloniale Lexikographien definieren und das Tool in der TextGrid-Umgebung so weiterentwickeln, dass es für die Auszeichnung vergleichbarer Wörterbücher zu anderen Sprachen nutzbar wird.

3. Kodikologie meets Topologie — Topologisches Retrieval als innovative, nicht-textuelle Form der Bedeutungsextraktion aus automatisch gewonnenen oder manuell erstellten Bildannotationen in DARIAH-DE

Jochen Graf, Universität Köln, Historisch-Kulturwissenschaftliche Informationsverarbeitung

In den letzten Jahren sind in der (Kunst-)Geschichte zahlreiche Systeme zur Bildannotation entstanden. Das Projekt *Meta-Image* [War11], zum Beispiel, betrachtet Bildannotation im Web als ein Problem von Bildern und Hyperlinks. Während es für textbasierte wissenschaftliche Arbeit Standardtechniken gibt, mit denen Bezüge zwischen Textstellen hergestellt werden können (Fußnoten, Hyperlinks), fehlt ein adäquater Mechanismus für die kollaborative Arbeit mit Bildern [Hel07]. Die Plattform *Sachsenspiegel_online*[8] [1] integriert einen graphischen Editor, der es erlaubt, Bilddetails mit Rechtecken zu markieren, die wiederum auf die textuelle Information eines Normvokabulars verweisen. Bei diesen und ähnlichen Systemen zur Bildannotation ist eine enge Orientierung am Medium Text zu erkennen: Bilder werden vorzugsweise durch Text erklärt und die Methoden der Bildannotation orientieren sich an denen der Textedition.

Hingegen gehört eine nicht-textuelle, am Medium Bild orientierte Herangehensweise aus Sicht der (kunst-)geschichtlichen Hermeneutik zu den neueren methodischen Entwicklungen der Bildinterpretation [Bohn08]. Der ursprünglich aus der Mathematik stammende und von den Medien- und Kulturwissenschaften übernommene Begriff der Topologie, zum Beispiel, vermeidet einen text-orientierten Zugang zum Bild. Ein kürzlich erschienener Tagungsband behandelt Topologie im Hinblick auf barocke Rechts-Links-Symbolik bis hin zu poetologischer Topologie [Guen07]. Topologie begegnet uns auch in den Methoden vieler geisteswissenschaftlicher Disziplinen. Die Kodikologie, beispielsweise, beschäftigt sich mit der räumlichen Disposition von Büchern. Sie kann aus dem strukturellen Aufbau von Buchseiten—aus der Symmetrie oder Asymmetrie der Textspalten, aus der Breite der Seitenränder oder aus der Größe und dem Format von Buchmalereien—auf den Buchtyp oder die Herkunft von Handschriften schließen [Bar14].

Stellt man die thematische und methodische Vielfalt des Raum-Topologie-Aspekts in den Geisteswissenschaften den aktuellen softwaretechnischen Bestrebungen im Bereich des Semantic Web entgegen, fällt auf, dass uns topologische oder eigentlich topographische Informationssysteme bisher lediglich in Form von Geographischen Informationssystemen zur Verfügung stehen. Mit Geographischen Informationssystemen können wir den raum-zeitlichen Kontext von Artefakten analysieren und visualisieren—die Systeme erschließen uns jedoch nicht die innere räumliche Struktur der Artefakte selbst.

eCodicology[9] ist ein Bildanalysetool, welches mithilfe von Algorithmen aus der Mustererkennung semantisch zusammengehörige Teilflächen digitalisierter, mittelalterlicher Handschriftenseiten erkennen und annotieren, und dadurch Annahmen über die kodikologischen Gestaltungsmerkmale der Seiten treffen kann. *Semantic Topological Notes* (SemToNotes)[10] ist ein manuelles

[8] http://www.sachsenspiegel-online.de/
[9] http://www.ecodicology.org/
[10] http://hkikoeln.github.io/SemToNotes/

Bildannotationstool, welches neben einem *Shared Canvas*[11] RDF Editor auch ein topologisches Retrievalsystem integriert. Beide Werkzeuge werden derzeit im Rahmen von DARIAH-DE entwickelt. Die Projekte verfolgen gemeinsam ein "nicht-textuelles Ziel". Sie untersuchen, wie viel Semantik eigentlich schon in den graphischen Annotationen selbst und deren Koordinaten steckt, noch vor einer textuellen Erschließung und Beschreibung der Digitalisate.

Besonders prägnant wird die Frage, wenn es sich um illuminierte mittelalterliche Handschriften handelt, in denen Figuren mit ganz bestimmten Gesten und Gebärden abgebildet sind. Hier spielen Abstände, Richtungen und Winkel, welche die Figuren und ihre Körperteile zueinander einnehmen, eine entscheidende Rolle. Die Analyse von Gesten in mittelalterlichen Handschriften ist schon in einer Monographie von Karl von Amira aus dem Jahre 1905 als ein topologisches Problem aufgefasst worden [Amir05]. Mit einem topologischen Retrievalsystem wird eine ikonographische Analyse der Bilderhandschriften nicht nur über textuelle Information ermöglicht, sondern zusätzlich auf Basis von polygonalen graphischen Koordinatendaten.

Literatur:

[Bar14] Vladimir Baranov, Kateřina Horníčková, Elena Lemeneva, Dóra Sallay, Gerhard Jaritz. Medieval Manuscript Manual. URL=http://web.ceu.hu/medstud/manual/MMM/ [2014-10-30]

[Schm14] Ruth Schmidt-Wiegand. Gebärden. In: Handwörterbuch zur deutschen Rechtsgeschichte (HRG), URL=http://www.HRGdigital.de/HRG.gebaerden [2014-08-01]

[Pan11] Pandel, Hans-Jürgen. 2011. Bildinterpretation. Die Bildquelle im Geschichtsunterricht. Wochenschau Verlag, Schwalbach, 2011

[Schl11] Joseph Schlecht, Bernd Carqué, Björn Ommer. 2011. Detecting Gestures in Medieval Images. In: IEEE International Conference on Image Processing (ICIP 2011)

[War11] Martin Warnke. 2011. Motive vernetzen: Meta-Image als Bild-Zettelkasten. Bilddiskurse in Zeiten des Internets. URL=http://www2.leuphana.de/meta-image/Material/Texte/KUNSTMAGAZIN_11121201_web.pdf

[Bohn08] Ralf Bohnsack. 2009. The Interpretation of Pictures and the Documentary Method. In: Historical Social Research, Vol. 34, No. 2 (128), 2009

[Guen07] Stephan Günzel (Hrsg.). 2007. Topologie. Zur Raumbeschreibung in den Kultur- und Medienwissenschaften, transcript Verlag, Bielefeld, 2007

[Hel07] Sabine Helmers, Heinz-Günter Kuper. 2007. HyperImage – Bildorientierte E-Science Netzwerke. URL=http://www.uni-lueneburg.de/hyperimage/hyperimage/downloads/helmers.pdf

[Mras06] Marcus Franz Wilhelm Mrass. 2006. Gesten und Gebärden - Begriffsbestimmung und -verwendung in Hinblick auf kunsthistorische Untersuchungen. Verlag Schnell & Steiner, Regensburg 2005

[Schm97] Mathias Schmoeckel. 1997. Karl von Amira und die Anfänge der Rechtsärchäologie. Die rechtsarchäologische Sammlung Karl von Amiras am Leopold-Wenger-Institut. In: Zeitschrift für Rechtsäarchäologie und Rechtliche Volkskunde 17 (1997), S. 67-81. URL=http://bsbdipriorkat.bsb.lrz.de/amira/projekt/amira-schmoeckel.pdf

[11] http://iiif.io/model/shared-canvas/1.0/index.html

[Wohl91] Reiner Wohlfeil. 1991. Methodische Reflexionen zur Historischen Bildkunde. In: Historische Bildkunde. Probleme - Wege - Beispiele, hrsg. von Brigitte Tolkemitt und Reiner Wohlfeil. Zeitschrift für historische Forschung, Berlin, Duncker und Humblot, 1991

[Amir05] Karl von Amira. 1905. Die Handgebärden in den Bilderhandschriften des Sachsenspiegels. URL=http://digi.ub.uni-heidelberg.de/diglit/amira1905

Forschungsdaten in Theorie und Praxis. Das DARIAH-DE Repository und die DARIAH-DE Collection-Registry

Sektionsvorschlag - DHd2015 „Von Daten zu Erkenntnissen", eingereicht von:
Peter Andorfer, Johanna Puhl, Stefan Schmunk

Abstract zur Sektion

Das Thema „Forschungsdaten" ist auch innerhalb DARIAH-DEs von zentraler Bedeutung. Dies gilt sowohl für die theoretisch-methodische Verortung dieses Begriffes als auch hinsichtlich des praktischen Umgangs mit Forschungsdaten in den kultur- und geisteswissenschaftlich arbeitenden Disziplinen. Die konkrete Arbeit kreist dabei vor allem um folgende Fragestellungen und Aufgabengebiete:

(1) Was sind Forschungsdaten in den Kultur- und Geisteswissenschaften? Kann angesichts der hohen Heterogenität in den einzelnen Disziplinen, deren vielfältigen Forschungsinteressen, -materialen und Methoden überhaupt eine allgemein verbindliche Definition des Begriffes Forschungsdaten gefunden werden und falls ja, was sind deren Kriterien und welche technisch-praktischen Konsequenzen lassen sich daraus wiederum für die Generierung, Sicherung und Distribution von Forschungsdaten ableiten.

(2) In engem Zusammenhang dazu stehen die Fragen zum Lebenszyklus von Forschungsdaten: So können Forschungsdaten in unterschiedlichen Phasen eines Projektes auf unterschiedliche Art und Weise erzeugt, gesammelt, aufbereitet und/oder analysiert werden. Forschungsdaten können dabei Ergebnis und/oder Quelle eines Forschungsprojektes sein. Diese Dynamik soll in einem eigenen, speziell auf die Eigenschaften kultur- und geisteswissenschaftlicher Forschungsdaten abgestimmten Modell abgebildet werden. Gleichzeit soll dieses Modell eines Forschungsdatenzyklus auch (technische) Anforderungen an Storage- und Publikationssysteme für digitale geistes- und kulturwissenschaftliche Forschungsdaten beschreiben können. Besonders interessant ist an dieser Stelle die Frage, inwiefern sich mithilfe von automatischen Prozessen in einer den Forschungsdatenzyklus unterstützenden Infrastruktur Erkenntnisse gewinnen lassen.

(3) Die in den Punkten eins und zwei ausgearbeiteten Anforderungen werden bei der Entwicklung des DARIAH-DE Repositoriums und der DARIAH-DE Collection Registry aufgegriffen und realisiert. Forschungsdaten, die in das Repository zur langfristigen Archivierung hochgeladen werden, werden in der Collection Registry kontextualisiert, als Sammlung von Forschungsdaten einem konkreten Forschungsprojekt zugeordnet und somit für die Nachnutzung durch andere aufbereitet.

Die Vorträge der Sektion „Forschungsdaten in Theorie und Praxis“ folgen diesen eben skizzierten Themenkomplexen. **Der erste Vortrag mit dem Titel „Forschungsdaten – Versuch einer Definition“ (Peter Andorfer, HAB Wolfenbüttel)** versucht in einem ersten Schritt eine generische Definition von „digitalen geistes- und kulturwissenschaftlichen Forschungsdaten“ und testet die Funktionalität dieser Definition anhand eines konkreten (geschichts)wissenschaftlichen Forschungsprojektes bzw. der darin gesammelten, beschrieben und/oder erzeugten (Forschungs?)Daten. **Im zweiten Vortrag „Definition *des* DARIAH Research Data LifeCycle“ (Johanna Puhl, HKI Köln)** werden der „DARIAH Research Data LifeCycle“ vorgestellt, die Besonderheiten und Spezifika gegenüber bereits bestehenden Modellen herausgearbeitet und die technische Anforderungen an eine Infrastruktur zur Speicherung und Publikation von Forschungsdaten formuliert. Eine solche von DARIAH-DE entwickelte Infrastruktur wird **im dritten Vortrag „Die Nutzung von Geistes- und kulturwissenschaftlichen Forschungsdaten - Das DARIAH-DE Repository“ (Stefan Schmunk, SUB Göttingen)** vorgestellt. Neben den technischen und administrativen Aspekten (wer kann, wie, unter welchen Voraussetzungen, ab wann und wie lange Repository und Collection Registry nutzen) soll anhand der bereits aus dem ersten Vortrag bekannten Forschungsdaten der Vorgang des Dateningest in das Repository und deren Registrierung in der Collection Registry exemplarisch vorgeführt werden.

Mit Hilfe der hier vorgestellten Sektionen sollen vornehmlich zwei Ziele erreicht werden. Einerseits geht es darum, die DARIAH-DE Collection Registry sowie das DARIAH-DE Repository in der DH-Community und über diese hinaus bekannt zu machen. Andererseits sollen die innerhalb von DARIAH-DE erarbeiten Konzepte und Definitionen zum Forschungsdatenbegriff und zum Research Data Lifecycle mit Vertretern der unterschiedlichen geistes- und kulturwissenschaftlichen Disziplinen diskutiert werden.

1. Abstract zu: „Forschungsdaten – Versuch einer Definition"

Peter Andorfer, HAB Wolfenbüttel

Schon 1998 empfahl die Deutsche Forschungsgemeinschaft in ihren „Vorschlägen zur Sicherung guter wissenschaftlicher Praxis",[1] dass: „Primärdaten als Grundlage für Veröffentlichungen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden [sollen]."[2] Dieser Passus findet sich unverändert auch in der 2013 veröffentlichten „ergänzten Auflage" wieder. Als Primärdaten[3] nennt die DFG dabei Daten, die in allen „experimentellen Wissenschaften" aus „Einzelbeobachtungen", „Experimenten" und „numerischen Rechnungen" gewonnen würden. In den „Sozialwissenschaften" wäre es außerdem mehr und mehr üblich, „Primärdaten nach Abschluss ihrer Auswertung durch die Gruppe, die die Erhebung verantwortet, bei einer unabhängigen Stelle zu hinterlegen." Weitere Primärdaten wären darüber hinaus noch: „Messergebnisse, Sammlungen, Studienerhebungen, Zellkulturen, Materialproben, archäologische Funde, Fragebögen". Wie zu sehen ist, bleibt der große Bereich der Geistes- und Kulturwissenschaften bei diesen Überlegungen aber weitgehend ausgespart. Dies dürfte nicht zuletzt auch daran liegen, dass Begriffe wie Primär- oder Forschungsdaten in den geistes- und kulturwissenschaftlichen Disziplin schlichtweg nicht gebräuchlich sind. Womit gearbeitet, woran geforscht wird, sind im Selbstverständnis wohl in erster Linie Quellen. Quellen, die in Publikationen dann im Regelfall auch im Fußnotenapparat bzw. im Quellen- und Literaturverzeichnis in Form von Verweisen nachgewiesen werden. Als besonders wichtig erachtete oder nur schwer zugängliche Quellen können darüber hinaus noch meist in einem Anhang als Reproduktion (z. B. Faksimile oder Transkript) der Publikation beigefügt werden. Vor dem Hintergrund dieser fest etablierten Tradition des Publizierens geistes- und kulturwissenschaftlicher Ergebnisse ist es wenig verwunderlich, dass sich die Frage nach dem Umgang mit Primär- oder Forschungsdaten und das Problem der Aufbewahrung dieser Materialen nicht stellt.

---

[1] DFG, Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft", Bonn 1998, S. 21f, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

[2] DFG, Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Ergänzte Auflage. Bonn 2013, S. 21f, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf.

[3] Kritisch zum Begriff Primärdaten siehe Jens Klump, Digitale Forschungsdaten, in: Heike Neuroth u.a. (Hg.), nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3, 2010, S. 523-535, hier v.a. S. 524, http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949.

Für den Nachweis und Beleg der eigenen Forschungsergebnisse mag dieses System in den allermeisten Fällen ausreichen. Die effiziente Weiterarbeit an den in Fußnotenapparat und Quellenverzeichnis benannten Materialen wird dadurch aber kaum unterstützt, höchstens vielleicht dadurch, dass anderen Forschern vielleicht das mühevolle Suchen der genannten Quellen etwas erleichtert wird. Der Gang ins Archiv, das Ausheben des gesuchten Materials, die Anfertigung einer Reproduktion oder das Transkribieren relevanter Passagen bleibt nicht erspart, und dass, obwohl ein Fachkollege genau dieselbe Quelle schon selbst im Archiv gesucht, davon eine Reproduktion angefertigt, eine Kopie, eine Fotographie oder einen Scan, und womöglich gar den gesamten Inhalt der Quelle in Form eines Regestes oder einer Transkription erfasst hat. In die Publikation fließt davon aber vielleicht nur eine knappe Paraphrase eines Teiles der Quelle samt der entsprechenden Archivsignatur in der Fußnote. Die Faksimiles und Transkripte hingegen liegen auf der privaten Festplatte des Forschers, wodurch diese Forschungsdaten zu einer bestimmten Quelle faktisch für Andere nicht mehr zugänglich sind.

Geht man wie eben von einem konkreten Beispiel aus, so gestaltet sich die Unterscheidung von Quellen und Forschungsdaten als einfach und nachvollziehbar. Als ungleich schwieriger, vor allem vor dem Hintergrund der großen Heterogenität geistes- und kulturwissenschaftlichen Arbeitens, erweist sich hingegen die Formulierung einer möglichst generischen Definition des Begriffs Forschungsdaten. Insbesondere dann, wenn diese Definition nicht nur der erwähnten Diversität der einzelnen Disziplinen gerecht werden soll, sondern auch die praktisch-technischen Rahmenbedingungen für den Aufbau eines Forschungsdatenrepositoriums vorzugeben hat.

Innerhalb DARIAH-DEs wurde an einer solchen Definition gearbeitet. Ein weit fortgeschrittener Entwurf einer Definition von „digitalen geistes- und kulturwissenschaftlichen Forschungsdaten innerhalb DARIAHs" wird im Rahmen des Vortrages vor- und natürlich auch zur Diskussion gestellt. Bei dieser Vorstellung werden sukzessive die einzelnen Bauteile der Definition näher beleuchtet und die Entscheidungsschritte, die zur letztendlich vorliegenden Formulierung geführt haben, beschrieben.

In einem zweiten Schritt wird diese Definition einem Praxistext unterzogen. Ausgehend von einem konkreten Forschungsprojekt werden jene Materialen, die sich im Zuge des Forschungsvorhabens auf der Festplatte angehäuft haben, dahingehend untersucht, ob es sich darum um Forschungsdaten im Sinne der DARIAH-Definition handelt. Die Ergebnisse dieses Praxistestest sollen dabei einerseits zur schärferen Abgrenzung von Begriffen wie „Quelle", „Primärdaten", „Rohdaten", „Forschungsdaten" und „Publikation" dienen. Andererseits soll anhand dieses Fallbeispiels auch die Problematik der Vielfalt an verwendeten Programmen respektive Dateiformaten reflektiert werden. Behandelt werden Transkripte von archivalischen Quellen im docx-Format, die Auswertung eines Steuerkatastars als

Excel-Dokument im xlsx-Format, Fotographien eines Manuskriptes als jpg, eine in Zotero erstellte Bibliographie, Bilddateien die METS/MODS beschrieben sind, wie auch die Edition dreier Briefe aus dem 18. Jahrhunderts nach den Regeln der TEI.

Vor dieser Vielfalt an Daten und Dateiformaten gilt es dann Konzepte und Workflows zu entwickeln und erproben,[4] um aus den wenig strukturierten und stark idiosynkratrischen Datenmengen, die sich im Laufe eines Forschungsprojektes auf einer Festplatte ansammeln, Forschungsdaten im Sinne der DARIAH-Definition zu generieren. Forschungsdaten, die dann in die ebenfalls von DARIAH-DE entwickelten Infrastruktur, dem DARIAH-Repository und der DARIAH-Collection Registry, eingespeist werden können um, ganz im Sinne des Research Data Lifecycles, Ausgangspunkt für andere Forschungen werden können. Wie ein solcher Workflow aussehen könnte, wird in der Vorstellung von Repository und Collection Registry anhand der in diesem Vortrag konkret benannten Daten demonstriert.

2. Abstract zu: Der Forschungsdatenzyklus in DARIAH-DE. Automatische Erkenntnisse durch Automatisierung von Methoden?

Johann Puhl, HKI Köln

Auf dem großen Feld wissenschaftlicher (Teil- und Unter-) Disziplinen existiert eine breite Vielfalt an Definitionen für einen Forschungsdatenzyklus. Dabei kommen viele dieser Definitionen aus dem Bereich der Informations- und Bibliothekswissenschaften[5] und beziehen sich ganz generisch auf Forschungsdaten und ihre Bereitstellung ohne dabei diese näher gemäß ihrer Zugehörigkeit zu einer Disziplin zu untersuchen. Andere Ansätze stammen eher aus spezifischen (oft naturwissenschaftlichen) Fachbereichen, wie den Lebenswissenschaften[6] oder der Physik.[7]

In DARIAH-DE wird eine Infrastruktur mit einem digitalen Repository als Kernbestandteil implementiert, die einen solchen Lebenszyklus speziell für Forschungsdaten und Fragestellungen aus den Geistes- und Kulturwissenschaften konzeptionell ermöglichen und mithin automatisiert unterstützen soll. Das entscheidende Motiv bei der Konzeption von Forschungsdatenzyklen ist die

[4] Für einen systematischen Überblick zu den "im Feld" verwendeten Dateiformaten samt deren Evaluierung in Bezug auf Archivierbarkeit siehe: IANUS, IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften, http://www.ianus-fdz.de/it-empfehlungen/.

[5] http://www.lib.ua.edu/wiki/sura/index.php/Data_Life_Cycle_Models.

[6] Joyce M. Ray (Hg.), Research Data Management: Practical Strategies for Information Professionals, USA 2014.

[7] UKOLN, I2S2 Idealised Scientific Research Activity Lifecycle Model, UK 2011, http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf.

grundsätzliche Gewährleistung, dass Forschungsergebnisse reproduziert und damit überprüft werden können. Daneben soll speziell durch den in einem solchen Zyklus getriebenen Dokumentations- und Publikationsaufwand auch die Nachnutzbarkeit von Forschungsdaten erhöht werden, sodass einmal erhobene Daten oder digitalisierte Dokumente häufiger und in breiterem Kontext gefunden und genutzt werden können.[8]

Die Allianz der deutschen Wissenschaftsorganisationen verweist in ihrer Empfehlung dabei dezidiert darauf, dass „Formen und Bedingungen des Zugangs zu Forschungsdaten […] gesondert für die jeweiligen Fachdisziplinen unter Berücksichtigung der Art und Weise der Datenerhebung, des Umfangs und der Vernetzbarkeit des Datenmaterials sowie der praktischen Brauchbarkeit der Daten entwickelt werden [müssen]."[9]

Für die Bedarfserhebung in den digitalen Geisteswissenschaften und hier insbesondere für den Aufbau einer Infrastruktur in DARIAH-DE gelten daher folgende Modelle als besonders geeignet und können als Grundlage für ein geeignetes eigenes Modell dienen:

Geistes- und sozialwissenschaftliche Ansätze für Forschungs(daten)zyklen

Ein recht differenziertes Modell für einen Forschungsdatenzyklus stammt in den Sozialwissenschaften von der Data Documentation Alliance.[10] Dabei wird folgender Ablauf beschrieben: Aufbauend auf einem „Study Concept” erfolgt die Sammlung von Daten („Data Collection”). Die Sammlung der Daten wird gemäß dem „Study Concept” verarbeitet („Data Processing"), archiviert („Data Archiving") und verbreitet („Data Distribution"). Auf diesen verbreiteten und veröffentlichten Daten kann hernach die Suche und Nachnutzung („Data Discovery”) zur erneuten Analyse („Data Analysis”) und der damit einhergehenden Umwidmung („Repurposing”") für eine veränderte Forschungsfrage der Daten erfolgen. Die umgewidmeten Daten werden hernach erneut verarbeitet und der Forschungsdatenzyklus kann erneut beginnen.

Schon vor dem sozialwissenschaftlichen Ansatz wurden von John Unsworth eine Liste typischer Methoden ("primitives") eines Geisteswissenschaftlers veröffentlicht.[11] Auch diese lässt sich zyklisch (oder wie Unsworth es nennt: rekursiv) betrachten, sodass auf die Tätigkeit des „Representing”

[8] DFG, Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Ergänzte Auflage, Bonn 2013. S. 21f, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf.

[9] Allianz der deutschen Wissenschaftsorganisationen, Grundsätze zum Umgang mit Forschungsdaten, 2010, http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze.htm

[10] DDI Structural Reform Group, DDI Version 3.0 Conceptual Model, DDI Alliance 2004, Figure: "Combined Life Cycle Model", S. 8, http://opendatafoundation.org/ddi/srg/Papers/DDIModel_v_4.pdf.

[11] John Unsworth, Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?, London 2000, http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

erneut „Discovering” folgt. Jedoch betont er in der dazugehörigen Veröffentlichung den nicht erschöpfenden Charakter seines Modells. Die Liste enthält folgende Tätigkeiten: Discovering, Annotating, Comparing, Referring, Sampling, Illustrating, Representing.

In den Niederlanden wurde speziell für die historischen Informationswissenschaften ein Zyklus beschrieben, der konkret den Fluss und die Nachnutzung von „historischer Information” darstellt.[12] Hier werden im Forschungsprozess sechs Schritte identifiziert, die zyklisch wiederholt werden können: „Creation", „Enrichment”, „Editing”, „Retrieval”, „Analysis” and „Presentation”. Dabei stellt „Creation” nicht nur den Erhebungs- sondern auch den Anreicherungsprozess dar. Darunter können also auch Tätigkeiten, wie Scannen oder die Verarbeitung eines Scans mit OCR-Methoden fallen.

Zur Bezeichnung „Enrichment” zählen Funktionen wie die Verwendung von Annotationen und die Extension der Daten mit Metadaten. „Editing” hingegen beschreibt als Erweiterung von Enrichment komplexere Tätigkeiten, wie Markup und Erweiterung um intellektuelle Inhalte. „Retrieval” schließlich macht die erweiterten und annotierten Daten such- und nutzbar. Darauf erfolgt der Schritt der „Analysis”, welcher sich eher auf den qualitativen Vergleich von Daten aber auch die quantitative Analyse von Datensätzen bezieht. Die Funktion der „Presentation” ist nun als Abschluss der Schritt, der der Veröffentlichung und Zugänglichmachung der Forschung dient und zur Nachnutzung einladen soll.

Das DARIAH-DE Modell

Eine besondere Herausforderung besteht in DARIAH-DE darin, dass der erarbeitete Forschungsdatenzyklus auch tatsächlich in einer technischen Infrastruktur umgesetzt werden soll. Zusätzlich zu den oben geschilderten geisteswissenschaftlichen Modellen für einen Forschungsdatenzyklus ist hier explizit sowohl die Veröffentlichung der Daten als auch ihre Langzeitarchivierung und Kuration vorgesehen, sodass der nachhaltige Zugang nicht nur zu einer Publikation sondern auch zu den hierfür verwendeten geisteswissenschaftlichen Forschungsdaten sicher gestellt werden kann. Auf diese Weise soll sowohl die Nachnutzung der Forschungsdaten als solche erhöht werden als auch der zu einer Publikation führende Forschungsprozess transparent und nachvollziehbar werden.

[12] Boonstra, Breure und Doorn, Past, present and future of historical information science, Amsterdam 2006, Kapitel 2.2: The life cycle of historical information, S.21 f.

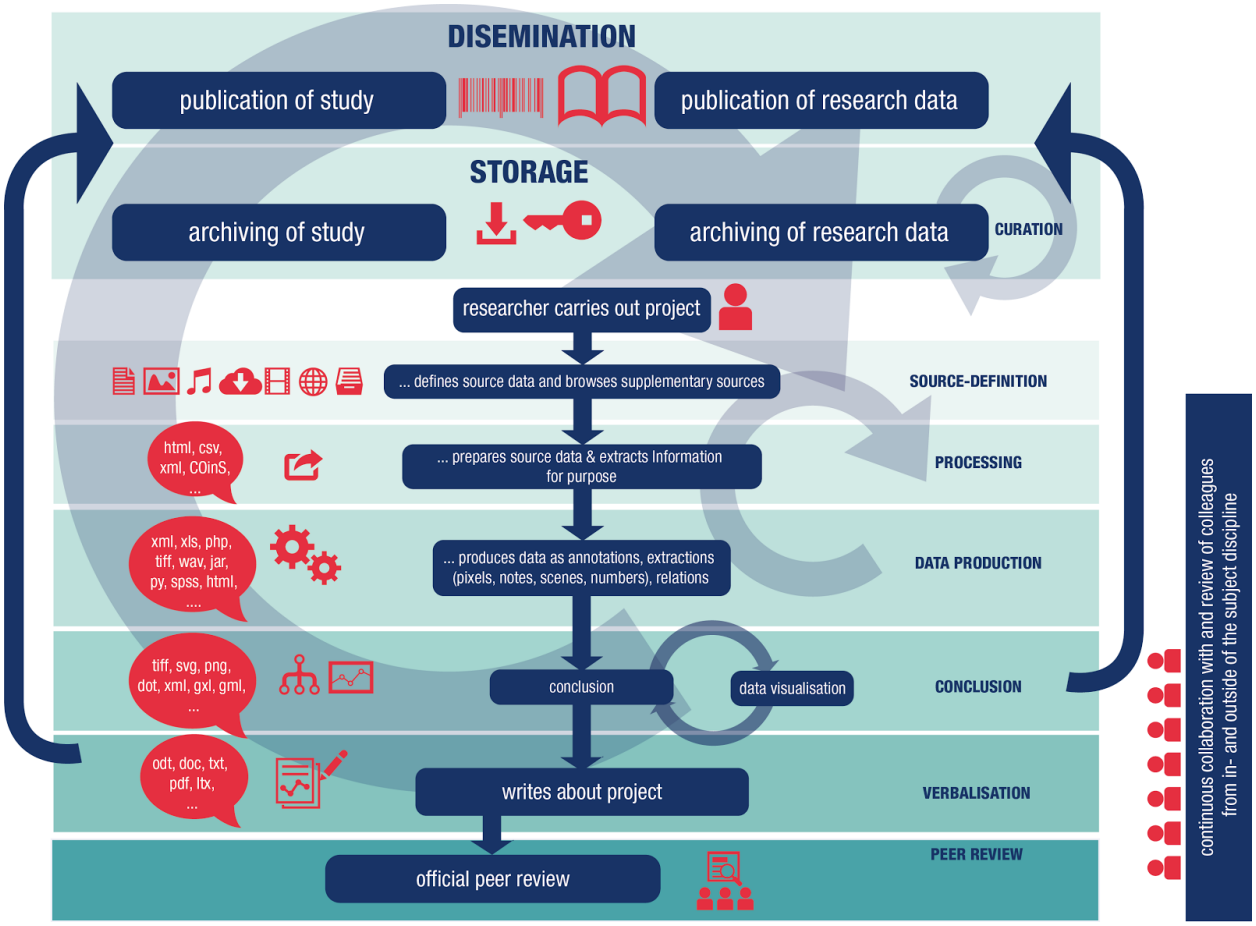


Abb. 1: Referenzmodell für einen Research Data LifeCycle in DARIAH

Bei der Entwicklung eines Referenzmodells in DARIAH-DE konnten eine ganze Reihe zyklischer Prozesse beobachtet werden, die sich nicht in einem linearen System abbilden lassen:
Insbesondere der Peer-Review als auch die Kuration der Daten im Archiv sind Aufgaben, die unbedingt in einer Infrastruktur implementiert werden sollten. Auch kleine und teilweise recht technische Zwischenschritte[13], welche in den großen abstrakten Modellen häufig undefiniert bleiben, bedürfen hier der konkreten Spezifikation.
Eine direkte Folge aus solchen Überlegungen ist der Bedarf nach einem Metadatenmodell, welches in der Lage ist, komplexe Arbeitsflüsse abzubilden, zu jedem Arbeitsschritt Informationen zu speichern oder zu erweitern und so Datei- und Projektversionierung und damit Transparenz über einen Forschungsprozess zu ermöglichen.

[13] Z.B. die Vergabe von persistenten Identifiern, die Spezifikation und Standardisierung von fachspezifischen Dateiformaten, welche in einem Forschungsdatenzyklus verwendet werden, die Angabe akzeptierter auslesbarer Metadatenformate und -felder etc.

Bedingt durch die Wahl eines Datenmodells oder Modellierung eines eigenen Schemas können eine ganze Reihe von Optionen eröffnet werden, die Auswirkungen auch auf die zukünftige Anwendung von Forschungsmethoden in den digitalen Geisteswissenschaften haben können.
Durch die Möglichkeit, Akteure und Ereignisse[14] in den Metadaten einer Infrastruktur zu modellieren, können Tools nicht nur im Routinebetrieb verwendet werden (z.B. Tools für OCR-Scans von Bildern, automatische Lemmatisierung anhand standardisierter Wörterbücher) sondern diese Tools auch als Akteure spezifiziert und in den Metadaten verankert werden. Dabei können die Ergebnisse dieser Prozesse in standardisierten Dateiformaten exportiert, referenziert und später sogar miteinander verglichen werden. Je exakter und gleichzeitig komplexer nun ein solches Metadatenmodell aufgebaut und in einer Infrastruktur verankert ist, desto mannigfaltiger sind die Fragen, die sich mithilfe einer solchen Infrastruktur automatisiert anhand einer Forschungsdatensammlung beantworten lassen.
Entscheidende Kriterien bei der Implementation eines Research Data LifeCycle sind also die Spezifikation der in den Geistes- und Kulturwissenschaften verwendeten Dateiformate und die Beschreibbarkeit und Automatisierbarkeit der darauf anwendbaren Methoden. Hier leistet DARIAH-DE mit der Arbeit an einem Referenzmodell für einen Forschungsdatenzyklus der digitalen Geisteswissenschaften Pionierarbeit, welche sich insbesondere in der praktischen Implementation in einer Infrastruktur auszahlen wird. Der Vortrag soll die hier geschilderten Fragestellungen, insbesondere einzelne Definitionen und Spezifikationen beleuchten und das daraus resultierende Referenzmodell für einen Forschungsdatenzyklus in den digitalen Geisteswissenschaften einer interessierten Öffentlichkeit vorstellen.

3. Abstract zu: „Die Nutzung von Geistes- und kulturwissenschaftlichen Forschungsdaten – Das DARIAH-DE Repository"

Dr. Stefan Schmunk, SUB Göttingen

Im Rahmen von DARIAH-DE widmet sich das Cluster „Wissenschaftliche Sammlungen und Forschungsdaten" nicht nur methodischen und konzeptionellen Fragen des Umgangs, der

[14] Das im Bibliothekswesen eingesetzte Metadatenformat für die Langzeitarchivierung, PREMIS, hält eine sehr komplexe Struktur für „events" und „agents" vor. Vgl. Library of Congress: PREMIS Data Dictionary for Preservation Metadata, Version 2.2, USA 2012, http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

Generierung, der Nutzung[15] und des Enrichments von digitalen Forschungsdaten, ein zentraler Teil der Tätigkeiten besteht insbesondere auch in der Entwicklung und Realisierung einer Repository-Lösung für geistes- und kulturwissenschaftliche Forschungsdaten.[16]

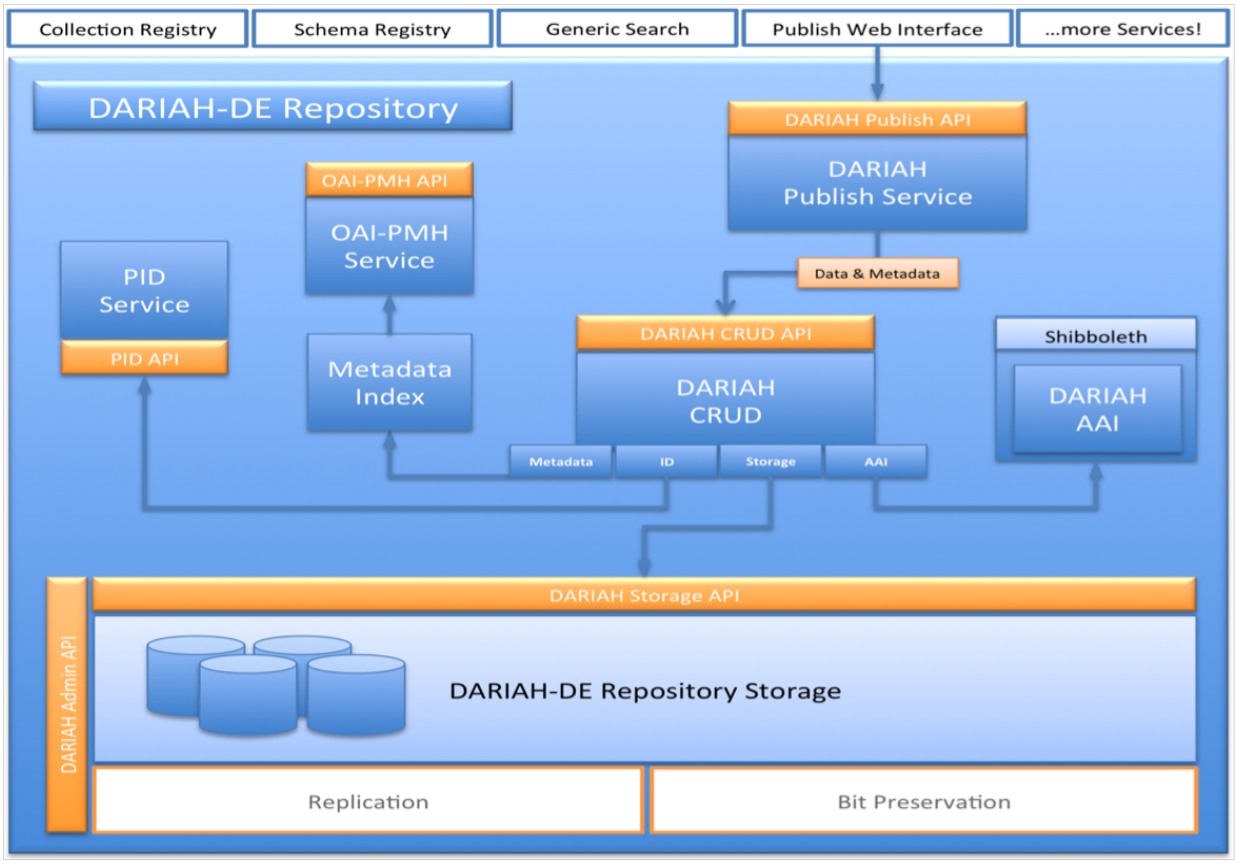


Abb. 2: DARIAH-Repository

Das DARIAH-DE Repositorium wird zukünftig nicht nur assoziierten Forschungsprojekten zur Verfügung stehen, wie derzeit beispielsweise TextGrid,[17] sondern auch EinzelforscherInnen und Forschungsprojekten, die ihre Forschungsdaten persistent, referenzierbar und langzeitarchiviert speichern und Dritten zur Verfügung stellen wollen.

Um dies zu erreichen arbeiten TextGrid, das aus der Virtuellen Forschungsumgebung TextGrid Laboratory[18] und TextGrid Repository[19] besteht, und DARIAH-DE zusammen. Das DARIAH-DE Repositorium stützt sich auf die Codebasis des TextGrid Repository und wird mit verschiedenen

[15] Aber auch Nutzungsmöglichkeiten wie z.B. lizenzrechtlichen Fragen, siehe: https://de.dariah.eu/lizenzen
[16] https://de.dariah.eu/forschungsdaten
[17] https://www.textgrid.de
[18] https://www.textgrid.de/registrierungdownload/download-und-installation/
[19] http://www.textgridrep.de

Service-Instanzen und unterschiedlichen, an das DARIAH-DE Repositorium angepassten Modulen für Funktionen wie Speicher- und AAI-Zugriff, implementiert.

Im Projekt DARIAH-DE wurde beispielsweise in den vergangenen Jahren u.a. eine Authentifizierungs-und Autorisierungsinfrastruktur (AAI) und die DARIAH-DE Storage API für die Speicherung von Forschungsdaten auf Bit Preservation Level aufgebaut, sodass Forschungsdaten zwischen den beteiligten Rechenzentren repliziert werden können. Dadurch ist sichergestellt, dass die Infrastruktur nicht nur als Speicherort für statische Daten verwendet werden kann – diese also öffentlich zugänglich, zitierfähig und langzeitarchiviert sind – sondern ebenso die Möglichkeit gegeben ist, dynamische Daten – die gegebenenfalls durch eine AAI gesichert sind und die aufgrund andauernder aktiver Nutzung aktualisiert werden müssen – dort abzulegen.

Auf die Forschungsdaten kann mithilfe von APIs zugegriffen werden und zugleich werde alle Forschungsdaten mit EPIC-PIDs[20] versehen, sodass andere Tools und Service diese nachnutzen können. Zu diesen Tools gehört beispielsweise die DARIAH-DE Collection Registry.[21] Sie enthält Informationen über beliebige Forschungsdaten-Repositorien und deren Sammlungen. Die in DARIAH-DE entwickelte Generische Suche[22] indiziert die Sammlungen der Collection Registry und bietet so einen userfreundlichen und zudem konfigurierbaren Zugriff auf die Inhalte. Die dritte Komponente bildet die DARIAH-DE Schema Registry, die eng mit der Generischen Suche vernetzt ist und das Mapping unterschiedlichster Metadatenbeschreibungen von Sammlungen ermöglicht. Diese stellt die XML-Schemata für das Mapping und für Metadata Crosswalks zur Verfügung.

Neben dem technischen Aufbau und der Vorstellung des technischen Frameworks steht die Präsentation des Zusammenspiels der technischen Komponenten und deren modularer Struktur - und damit auch nachnutzbarer Integration in DH-Forschungsprojekte – im Vordergrund. Neben den technischen und administrativen Aspekten – wer kann, wie, unter welchen Voraussetzungen, ab wann und wie lange Repository und Collection Registry nutzen und Forschungsdaten nachnutzbar für Dritte speichern – soll anhand der bereits im ersten Vortrag vorgestellten Forschungsdaten der Vorgang des Dateningest in das Repository und der Registrierung von Sammlungsbeschreibungen in der Collection Registry exemplarisch vorgeführt werden. Hierbei sollen insbesondere die Nutzung, der Zugriff und die zugrunde liegenden Arbeits- und Forschungsprozesse beleuchtet werden, um exemplarisch die Möglichkeiten und zugleich die Grenzen eines Forschungsdaten-Repositorys aufzuzeigen.

[20] https://de.dariah.eu/pid-service
[21] Eine Übersicht der verzahnten Applikationen, die zur Speicherung, zur Suche und Recherche und den Zugang zu Forschungsdaten ermöglichen, findet sich hier: https://de.dariah.eu/forschungsdatensammlungen
[22] http://search.de.dariah.eu/search/

Warum sollen wir unseren Daten trauen? Soziale Erkenntnistheorie und die 'rechnenden Geisteswissenschaften'

Wenn wir über die Beziehung zwischen Daten und Erkenntnis nachdenken, übersehen wir leicht, dass Daten selbst implizite Erkenntnisansprüche enthalten können. Daten sind nicht immer gegeben, sie können auch das Ergebnis menschlichen Handelns sein: ein Bibliothekar gibt Metadaten ein. Oder sie beruhen auf den Rechnungen eines Computers, bspw. in 'topic models', die aus einem elektronisch verfügbaren Corpus abgeleitet werden. Der Glaube an die Korrektheit dieser Daten ist eine Sache des Vertrauens. Wenn wir in eine Bibliothek reisen, um ein Rarum einzusehen, so vertrauen wir auf die professionellen Fähigkeiten des Bibliothekars, der für die Erstellung der Metadaten im Katalog verantwortlich war. Wenn wir einen Computer nutzen, um 'topic models' zu berechnen, vertrauen wir der Korrektheit der Berechnungen und benutzen die Resultate im Glauben, dass sie das entsprechende Corpus wahrheitsgetreu (innerhalb der gegebenen Parameter) repräsentieren.

In den letzten Jahrzehnten ist epistemisches Vertrauen und damit zusammenhängend das Wissen aus dem Zeugnis anderer zu einem wichtigen Thema der Erkenntnistheorie, insbesondere der sogenannten "sozialen Erkenntnistheorie" geworden. Don Fallis hat für die Informationswissenschaft diese Debatten in nützlichen Kriterien zusammengefasst (Follis 2004), von denen ich drei in meiner Präsentation heranziehen werde: wenn wir uns fragen, wieviel Vertrauen wir einer möglichen Wissensquelle entgegenzubringen bereit sind, sollen wir (1) die Autorität unseres Informanten, (2) die Anzahl voneinander unabhängiger Wissensquellen sowie (3) den Inhalt der Information, ihre Plausibilität und den Grad ihrer Bestätigung in Rechnung stellen. Bei der Beurteilung des Computers als eines Informanten ist (1) von überragender Wichtigkeit: wie können wir rechtfertigen, dass wir den Computer als epistemische Autorität für die Erarbeitung geisteswissenschaftlicher Forschung akzeptieren? Der Versuch der Beantwortung dieser Frage wird uns jedoch darauf führen, dass auch (2) von einigem Belang ist: wir können (und sollten wohl auch) unser Vertrauen in die epistemische Autorität des Computers dadurch steigern, dass wir mehr als einen benutzen: die Diskussion des 'Wissens aus dem Zeugnis eines Computers' führt also zu (3), zum Problem einer effektiven Methodologie für die Plausibilitätsprüfung computererzeugter Daten.

(1) Die epistemische Autorität von Computern steht nicht nur in den digitalen Geisteswissenschaften in Frage. Durch einen Computer erzeugte Beweise haben in der Philosophie der Mathematik zu einer intensiven Debatte geführt. In einem mittlerweile klassischen Aufsatz (Burge 1998) argumentiert Tyler Burge für die Unterscheidung zwischen einer widerlegbaren Berechtigung zu epistemischem Vertrauen ('entitlement') und der vollständigen Rechtfertigung von Wissensansprüchen, die auf computergenerierten Daten beruhen. Die widerlegbare Berechtigung zu solchem Vertrauen beruht darauf, dass Wissen aus dem Zeugnis anderer auch in anderen Bereichen der Mathematik vorausgesetzt wird: die Anwendung des Satzes des Pythagoras hat bei der Lösung eines Problems auch dann ihre Berechtigung, wenn wir nicht zuvor seinen Beweis nochmals nachvollzogen haben (Burge 1998, 7). In ähnlicher Weise können Kinder in grundlegenden Wahrnehmungsüberzeugungen gerechtfertigt sein, auch wenn sie zu den Tatsachen keinen kognitiven Zugang haben, die dieses Gerechtfertigtsein ermöglichen, weil sich diese Tatsachen erst in der philosophischen Reflexion erschließen (Burge 1998, 3). Eine solche 'widerlegbare Berechtigung' in Burges Sinn kann jedoch durch vernünftigen Zweifel widerlegt werden, Zweifel, die jedoch dadurch ausgeräumt werden können, dass wir die Quelle des Wissensanspruchs ausklammern (bzw.,

in Burges' Terminologie, die Perspektive der dritten Person einnehmen). In dieser Perspektive wird der Gebrauch des Computers in der Mathematik zu einer Erweiterung unserer rationalen Fähigkeiten (Burge 1998, 31).

(2) Aber selbst wenn wir diesen Überlegungen zustimmen würden, können sie nur eine allgemeine Bereitschaft zum Vertrauen in die epistemische Autorität eines Computers begründen. Sie geben jedoch keine Antwort auf die Frage, ob dieses Vertrauen bedingungslos sein muss. Burge konzediert, dass Programmiersprachen und die in ihnen geschriebenen Programme selbst Ausdruck mathematischer Wahrheiten seien. Sie seien aber so zuverlässig wie vom Menschen erarbeitete mathematische Wahrheiten (Burge 1998, 8). Hardwarefehler können mit Schreibfehlern bei der Abfassung eines Beweises gleichgesetzt werden (Burge 1998, 8f). Weil aber Computer von rationalen Wesen gebaut und programmiert werden, sind es insgesamt deren Rationalität und die Unveränderlichkeit der im Computer ausgenutzten Naturgesetze, die aus der Sicht von Burge die letztendliche Rechtfertigung unseres Vertrauens in die Ergebnisse begründen, die von solchen Maschinen geliefert werden. Aus dem Blickwinkel des Ingenieurs erscheint diese Einschätzung als sehr optimistisch. Arkoudas und Bringsjord (2007) unterscheiden sechs unterschiedliche Ebenen, auf denen bei der Generierung von Computerergebnissen Fehler unterlaufen können. Jenseits der allgemeinen Berechtigung, der epistemischen Autorität von Computern zu vertrauen ist die konkrete Gerechtfertigtheit unserer Wissensansprüche und damit zusammenhängend das Vertrauen, das wir dem Computer entgegenbringen sollten, also auch ein technisches Problem. Ähnlich wie in Experimenten der Naturwissenschaften sind auch die digital humanities gehalten, der Überprüfung der Replizierbarkeit von Ergebnissen erhöhtes Augenmerk zu schenken.

(3) Deswegen werde ich abschließend einen konkreten Anwendungsfall erörtern: die Verwendung von algebraischen Programmen zu heuristischen Zwecken in der Mathematik und die daraus resultierenden Probleme, wie sie in Durán et al. 2014 geschildert werden. Wollen wir computergeneriertes Wissen als gerechtfertigt auszeichnen, sind konkrete methodische Richtlinien zu befolgen: dies betrifft bspw. den Gebrauch mehrerer Implementationen, den Einsatz von Open Source Software, um im Zweifel die Fehlerfindung zu erleichtern oder die Publikation von Code, der in einem Projekt eingesetzt wurde, um die Replizierbarkeit von Ergebnissen zu erleichtern. Nur dann sind wir nicht nur widerlegbar berechtigt, sondern gerechtfertigt, dem in Daten implizierten Wissen zu vertrauen und dürfen es zur Erarbeitung weiterreichender interpretatorischer Überlegungen und Erkenntnisse nutzen.

Literatur

Konstantine Arkoudas, Selmer Bringsjord, "Computers, Justification, and Mathematical Knowledge", in: Minds & Machines (17) 2007, 185–202

Tyler Burge, "Computer Proof, Apriori Knowledge, and Other Minds: The Sixth Philosophical Perspectives Lecture", in: Noûs (32) 1998, supplement, 1-37

Antonio J. Durán, Mario Pérez, and Juan L. Varona, "The Misfortunes of a Trio of Mathematicians Using Computer Algebra Systems: Can We Trust in Them?", in: Notices of the AMS, (61) 2014, 1249-1252

Don Fallis, "On Verifying the Accuracy of Information: Philosophical Perspectives", in: Library Trends (52) 2004, 463-487

Gleiche Textdaten, unterschiedliche Erkenntnisziele?

Zum Potential vermeintlich widersprüchlicher Zugänge zu Textanalyse

Thomas Bögel (Universität Heidelberg)
Michael Gertz (Universität Heidelberg)
Evelyn Gius (Universität Hamburg)
Janina Jacke (Universität Hamburg)
Jan Christoph Meister (Universität Hamburg)
Marco Petris (Universität Hamburg)
Jannik Strötgen (Universität Heidelberg)

1. Einleitung

Dieser Beitrag beleuchtet disziplinäre Errungenschaften, die durch die genaue Betrachtung unterschiedlicher disziplinäre Auffassungen von Daten und Erkenntnissen bzw. Erkenntnisinteressen im Projekt heureCLÉA ermöglicht wurden und die das große Potential interdisziplinärer Zusammenarbeit im *Digital Humanities*-Bereich herausstellen.

heureCLÉA ist ein *Digital Humanities*-Kooperationsprojekt zwischen Literaturwissenschaft und Informatik, in dem eine "digitale Heuristik" zur narratologischen Analyse literarischer Texte entwickelt wird.[1] Mit dieser Heuristik sollen (1) bislang nur manuell durchführbare Annotationsaufgaben bis zu einem bestimmten Komplexitätsniveau automatisiert durchgeführt und (2) statistisch auffällige Textphänomene als Kandidaten für eine anschließende Detailanalyse durch den menschlichen Nutzer identifiziert werden können. Dazu wird ein Korpus literarischer Erzählungen kollaborativ manuell annotiert. Anschließend wird mit regelbasierten NLP-Methoden sowie *Machine Learning*-Verfahren an der Entwicklung der Heuristik gearbeitet, die als zusätzliches Modul in die Textanalyseplattform CATMA implementiert werden wird.[2]

[1] Das Projekt heureCLÉA ist ein vom BMBF gefördertes eHumanities-Projekt, das von 02/2013-01/2016 an den Universitäten Hamburg und Heidelberg als Verbundprojekt durchgeführt wird (vgl. dazu auch www.heureclea.de). Zum aktuellen Projektstand vgl. Bögel et al. (im Erscheinen).
[2] vgl. www.catma.de

Die gemeinsame Frage, wie diese Heuristik erstellt werden soll, und die gemeinsame Betrachtung der literarischen Texte, die als Basis dienen, hat schnell gezeigt, dass es in den beteiligten Disziplinen unterschiedliche Auffassungen über die Qualität von und den Zugang zu Textanalysedaten gibt. So wird etwa der in der Literaturwissenschaft als notwendig geltende Interpretationspluralismus in der NLP als widersprüchlicher *Noise* betrachtet. Die in der NLP gängige Praxis, Verfahren weniger nach ihrer Nachvollziehbarkeit, sondern vielmehr nach der Qualität ihrer Ergebnisse zu beurteilen, wird wiederum in der Literaturwissenschaft abgelehnt, da dort die Qualität von Verfahren über einen inhaltlichen Austausch über die angewendeten Verfahren ausgehandelt wird.

Unser Beitrag will auf die möglichen methodischen und methodologischen Konsequenzen solcher disziplinär unterschiedlicher Zugänge zum Forschungsgegenstand – in unserem Fall: zu Texten und zu Textanalyse – in der Zusammenarbeit im *Digital Humanities*-Bereich hinweisen. Im Fokus stehen dabei zwei exemplarische Konsequenzen in den beteiligten Disziplinen: (1) der narratologische Workflow, für den eine Erweiterung des traditionellen hermeneutischen Zugangs zur Textanalyse entwickelt wurde, sowie (2) der für das *Machine Learning* gewählte Zugang der NLP, der sich durch eine besonders hohe Prozesstransparenz von klassischen *Machine Learning*-Ansätzen unterscheidet.

Beide Beispiele sind aus unserer Sicht exemplarisch für Interferenzen, die von *Digital Humanities*-Projekten erzeugt werden können. Diese Interferenzen bedeuten vorerst Störungen des geplanten Forschungsprozesses und erzeugen teilweise erheblichen Mehraufwand. Gelingt die Lösung der damit verbundenen Probleme, generieren sie aber einen Mehrwert sowohl für den Projekterfolg als auch für den von der Projektzusammenarbeit unabhängigen Fortschritt der beteiligten Disziplinen.

2. Die Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse

Die für die Entwicklung der Heuristik in heureCLÉA eingesetzten NLP-Verfahren werden auf ein Korpus 21 deutschsprachiger Erzählungen um etwa 1900 angewendet, das mit dem Textanalysetool CATMA annotiert wird. Das oben erwähnte *Noise*-Problem wird dadurch abgemildert, dass die Texte von mehreren Annotatorinnen mit Markup versehen werden.[3] Dieser Zugang verändert den traditionellen Prozess der Textanalyse in der Literaturwissenschaft zweifach.

[3] Für eine ausführlichere Beschreibung der durch das Spannungsfeld von Informatik und Hermeneutik bedingten Problematik und ihre Auswirkung auf die Anforderungen an die manuelle Annotation vgl. Gius & Jacke (in Vorbereitung). Dort werden auch die methodologischen Konsequenzen für die narratologische Theorie dargelegt.

Erweiterte Analysegrundlage durch Annotation

Eine offensichtliche Veränderung zum traditionellen Zugang ist die Erweiterung der betrachteten Datenbasis bzw. der Annahmen über diese. Zu den Vorannahmen des Textinterpreten, den Annahmen über Textteile und den Annahmen über das Textganze, die sich immer wieder gegenseitig beeinflussen und dadurch die Annahmen bestätigen oder modifizieren, kommen die in den Annotationen festgehaltenen Annahmen weiterer Interpretinnen. Der traditionelle hermeneutische Zugang zu Texten wird hier also nicht nur durch das Annotieren selbst – wie weiter unten ausgeführt – intensiviert, sondern auch um Annahmen anderer ergänzt.[4] Dadurch wird gewissermaßen die Grundlage für die weitere Analyse erweitert.

close(r) reading durch kollaborative, computergestützte Analysen

Der computergestützte Zugang an sich forciert bereits durch sein Sichtbarmachen der Analysen in den Annotationen ein intensiveres *close reading* als Textanalyseverfahren, in denen Interpretationen ohne eine ausführliche Dokumentation zugrundeliegender Analysen generiert werden. Durch die kollaborative Annotation derselben Texte durch mindestens zwei Annotatoren wird außerdem offensichtlich, an welchen Stellen es keine intersubjektive Übereinstimmung zwischen den Annotatorinnen gibt. Dies führte u.a. dazu, dass schnell deutlich wurde, dass die aus Gius (2013) übernommenen Beschreibungen der narratologischen Analysekategorien in der vorliegenden Form nicht als Arbeitsgrundlage für heureCLÉA ausreichen. Deshalb wurden zusätzlich Annotationsguidelines erarbeitet, die die Beschreibung der narratologischen Kategorien weiter systematisieren: Neben der Beschreibung und Operationalisierung des Phänomens enthalten die Guidelines Angaben zum typischen Umfang der getaggten Textmenge (etwa Wort/Wortgruppe, Satz, Absatz etc.), zu unmarkierten Fällen, die nicht annotiert werden, zu Indikatoren auf der Textoberfläche, zur Taggingroutine sowie Textbeispiele zur betreffenden Tagkategorien (vgl. Abbildung 1).[5] Die Taggingroutine zielt dabei insbesondere darauf ab, die Analyse so zu organisieren, dass die damit verbunden Aktivitäten in einer von einfachen zu komplexeren Aktivitäten geordneten Reihenfolge ausgeführt werden können.[6]

[4] vgl. zu den für das hermeneutische Verfahren relevanten Aspekten z.B. Bühler (2003).
[5] vgl. Gius & Jacke (2014).
[6] Dasselbe Verfahren wird in heureCLÉA auch auf die Reihenfolge der annotierten Phänomen angewendet. Dieses an der Komplexität der Analysekatogerien orientierte Vorgehen wurde bereits in Gius (2013) auf Ebene der narratologischen Phänomenbereiche entwickelt und erfolgreich angewendet.

Tagstring • Textabschnitte – Mindestgröße: Teilsatz	**Unmarkierter Fall** • Chronologisches Erzählen
Indikatoren auf der Textoberfläche • Zeitausdrücke, die Vorzeitigkeit, Gleichzeitigkeit oder Nachzeitigkeit ausdrücken • Tempuswechsel	
Tagging-Routine 1. Annotation aller nicht chronologisch dargestellten Textpassagen als Prolepse, Analepse, Simullepse oder Achronie. 2. Bei Anachronien: Spezifizierung von Umfang und Reichweite. 3. Bei Achronien: Spezifizierung der Verknüpfungsart.	
Beispiele • chronologisches Erzählen: „von ohngefähr erhob sie das Auge und traf mit dem blauesten Strahle in seinen Blick. Er ward wie von einem Blitz durchdrungen. Sie strauchelte, und so schnell er auch hinzusprang, konnte er doch nicht verhindern, daß sie nicht kurze Zeit in der reizendsten Stellung knieend vor seinen Füßen lag“ (Der Pokal) • Analepse: „Jetzt sah man, was geschehen war: der Hansjörg hatte sich am mittleren Gelenk den Zeigefinger der rechten Hand abgeschossen“ (Die Kriegspfeife) • Prolepse: „Zwanzig Jahre lang habe ich den Tod auf den Tag herbeigezogen, der in einer Stunde beginnen wird [...]“ (Der Tod) • Simullepse: „'Ich bin nicht allein', sagte ich […]. Dabei preßte sich mein Arm, der die Decke über ihren Kopf gelegt hatte, krampfhaft auf jene Stelle, wo ich den Mund vermutete [...]“ (Die Schutzimpfung) • Achronie: „Vorliebe empfindet der Mensch für allerlei Gegenstände. Liebe, die echte, unvergängliche, die lernt er – wenn überhaupt – nur einma kennen.“ (Krambambuli)	

Abbildung 1: Annotation von Ordnung, Zusammenfassung aus den Guidelines (vgl. Gius & Jacke 2014)

Das *close reading* wird außerdem durch Diskussionen intensiviert, die zwischen den Annotatoren stattfinden, wenn sie nach ihrem ersten Annotationsdurchgang die gesetzten Annotationen mit denen der anderen vergleichen.

Die der hermeneutischen Textanalyse eigene fortdauernde Bewegung zwischen Text und Analyse/Interpretation des Textes, deren Erkenntnisse wiederum in die erneute Betrachtung des Textes mit einfließen, wird durch die beiden durch die interdisziplinäre Zusammenarbeit notwendigen Erweiterungen des Zugangs sowohl in Bezug auf die Analyse bzw. Interpretation als auch in Bezug auf das zur Verfügung stehende Interpretationsmaterial wesentlich verstärkt (vgl. Abbildung 2).

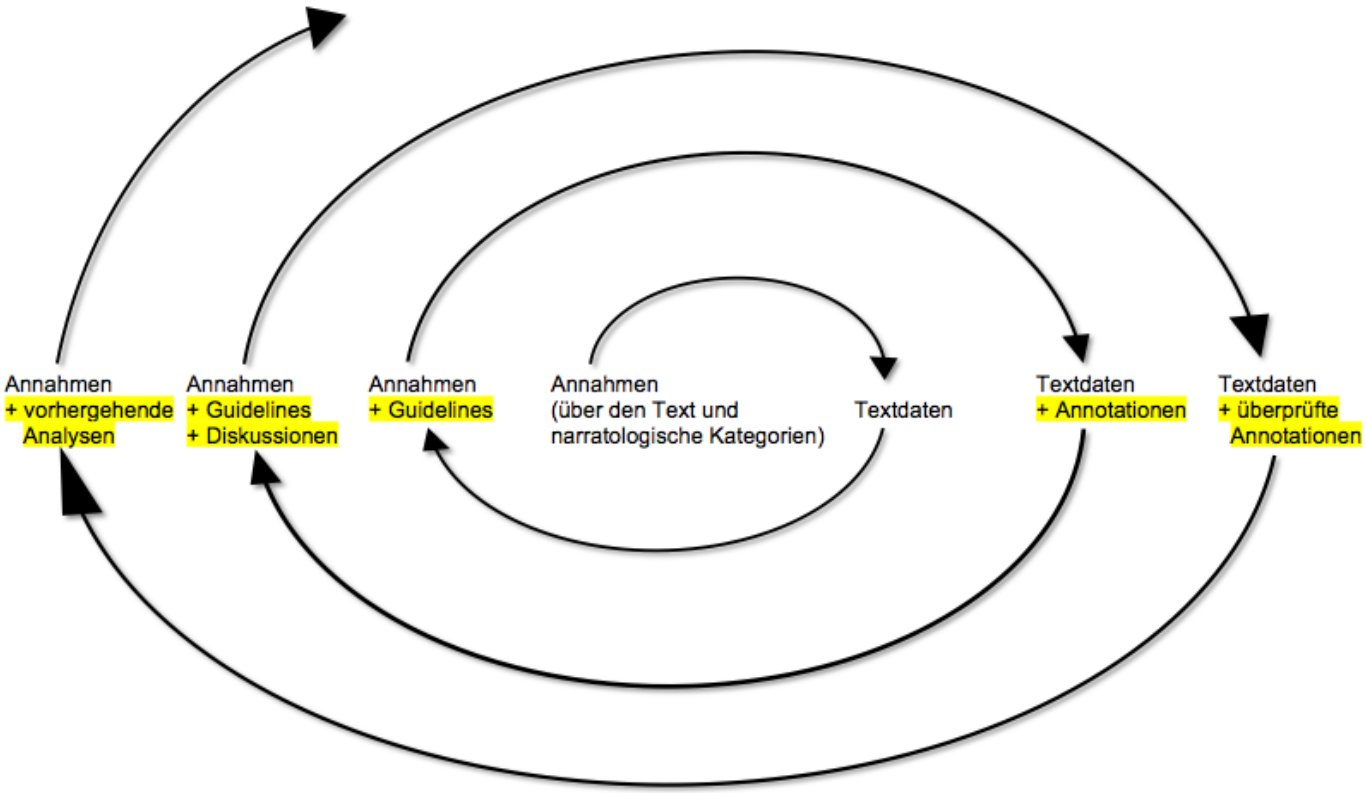


Abbildung 2: Der erweiterte hermeneutische Zirkel

3. NLP vor dem Hintergrund besonderer Textdomänen und notwendiger Transparenz von automatischen Entscheidungsprozessen

Bei der Verarbeitung deutscher literarischer Texte im Kontext einer Zusammenarbeit mit Narratologen stellen sich im Bereich der NLP zwei Hauptaspekte heraus: Zum einen bedingt der Fokus auf eine spezielle Textdomäne die Anpassung und den flexiblen Einsatz von NLP-Komponenten, die zumeist für Zeitungstexte optimiert sind. Auf der anderen Seite ergeben sich im Bereich der Modellbildung insbesondere im Bereich des maschinellen Lernens spezifische Herausforderungen, um die Akzeptanz von automatischen Annotationen sicherzustellen. Beide Aspekte sollen im Folgenden erläutert werden.

Der NLP-Workflow

Zur Erfassung und automatischen Vorhersage linguistischer Oberflächenphänomene entwickelten wir eine flexible und modulare NLP-Pipelinearchitektur auf Basis von UIMA[7], die Annotationen mit steigendem Komplexitätsgrad vornimmt und die Ergebnisse in einem Schichtenmodell speichert.

Die modular aufgebaute Pipeline ermöglicht einen flexiblen Austausch von Komponenten. Diese Flexibilität ist im Kontext unserer Textdomäne, also literarischer Texte, besonders

[7] http://uima.apache.org/

hilfreich und unabdingbare Voraussetzung, wie sich im Verlauf des Projekts gezeigt hat. Da NLP-Komponenten auf der Domäne von Zeitungstexten entwickelt werden, funktionieren viele Systeme nur auf einem Teil der Daten ähnlich qualitativ gut wie auf der Ursprungsdomäne. Details zur Architektur und den verwendeten NLP-Komponenten sind in Bögel et al. (2014) beschrieben.

Sichtbarmachung von Entscheidungsprozessen im maschinellen Lernen

Neben Features, die die Grundvoraussetzung für die Modellierung maschineller Lernverfahren darstellen und aus der oben dargestellten Pipeline gewonnen werden, ergeben sich auch bei der Wahl des konkreten Lern-Algorithmus interessante Aspekte durch das Gesamtprojekt.

In der Theorie des maschinellen Lernens werden Modelle und Gesamtsysteme danach bewertet, welchen empirischen Fehler sie auf ungesehenen Testdaten produzieren (Vapnik, 1998). Ein ideales System würde auf ungesehenen Daten perfekte Ergebnisse liefern und keine Fehler bei der Vorhersage machen. Vor dem Hintergrund unseres Kollaborationsprojektes zeigt sich jedoch, dass die Minimierung des Fehlers von Annotationen nur ein Qualitätsaspekt von Algorithmen ist. Um höhere Akzeptanz von Ergebnissen solcher Systeme zu erreichen, müssen sie einerseits *verlässliche* Vorhersagen produzieren, aber andererseits auch *transparenten, nachvollziehbaren Entscheidungsprozessen* zugrundeliegen. Mit zunehmendem Komplexitätsgrad maschineller Lernverfahren sinkt jedoch die direkte Nachvollziehbarkeit. So ist bei einer *Support Vector Machine* (Hearst et al., 1998), einem Standardverfahren des maschinellen Lernens, nicht ohne Weiteres nachvollziehbar, weshalb eine konkrete Entscheidung getroffen wurde und welche Einzelentscheidungen und Featurekonstellationen konkret zum Endergebnis geführt haben. Derartige Black-Box-Ansätze erschweren jedoch die Akzeptanz automatischer Annotationen.

Ein Beispiel für nachvollziehbare Algorithmen stellen Entscheidungsbäume (*decision trees*) dar, wie sie in Quinlan (1986) erstmalig beschrieben sind. Durch eine Visualisierung des Modells ist es möglich (vgl. Abbildung 3), jede Teilentscheidung, die zur Klassifikation beigetragen hat, nachzuvollziehen und den Einfluss von individuellen Kriterien (*Features*) zu verfolgen.

Abgesehen von der Nachvollziehbarkeit und Transparenz verhindern Black-Box-Ansätze auch direkte Eingriffsmöglichkeiten in den Vorhersageprozess. Für die Vorhersage bestimmter Phänomene (beispielsweise der Erzählgeschwindigkeit in Erzähltexten), die

ambigen Konzepten zugrunde liegen, können verschiedene *Features* als relevant erachtet werden. Bezogen auf Abbildung 3 wäre es so beispielsweise möglich, ein *Feature* zu entfernen und die Auswirkungen auf das neue Modell direkt zu beobachten.

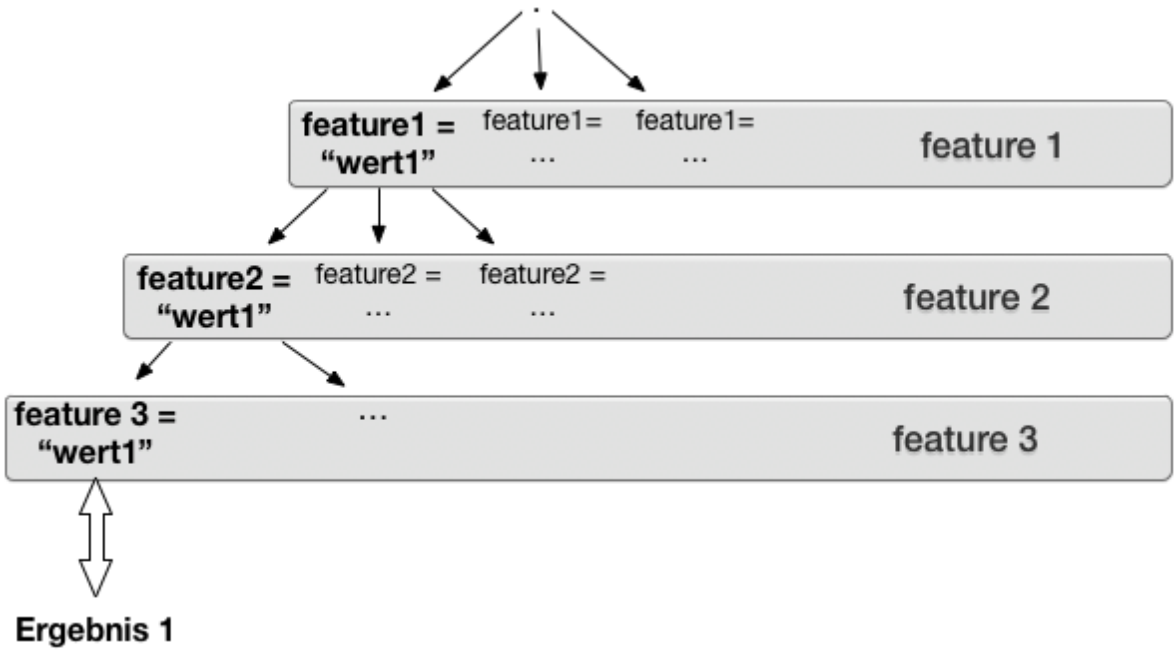


Abbildung 3: Schematische Visualisierung eines Decision Trees.

Dieses dargestellte Transparenz- und damit auch Akzeptanz-Problem stellt sich für maschinelle Lernprozesse grundsätzlich, wenn sie abseits eines reinen Selbstzwecks in einen konkreten Anwendungskontext eingebettet werden, anstatt für synthetische Benchmarks Ergebnisse zu produzieren.

4. Gemeinsame Erkenntnisse aus der interdisziplinären Arbeit

Die hier beschriebenen, durch die Zusammenarbeit veränderten Bedingungen der Textanalyse sind aus unserer Sicht typisch für Ansätze im Bereich der *Digital Humanities* und werden von den dort häufig genutzten kollaborativen Verfahren verstärkt. Damit wird offensichtlich, dass der Einsatz neuer Methoden nicht nur die Bearbeitung neuer Fragestellungen ermöglicht, sondern auch traditionelle Methoden wie etwa die für die Literaturwissenschaft zentrale Methode der hermeneutischen Textanalyse oder den ergebnisorientierten Zugang der NLP ergänzt bzw. modifiziert – und dadurch so weiter entwickelt, dass sowohl die interdisziplinäre als auch die disziplinäre Forschungsarbeit von der Entwicklung profitiert.

In beiden Fällen hat die Erhöhung der Transparenz der genutzten Prozesse gemäß den methodischen Bedarfen der anderen Disziplin maßgeblich zum Erfolg der Weiterentwicklung

beigetragen. Entsprechend wäre es interessant zu prüfen, ob dies generell eine produktive Strategie zur methodischen und methodologischen Verbesserung von in interdisziplinären *Digital Humanities*-Projekten genutzten Forschungsstrategien ist.

Bibliographie

Bögel, T. & Gertz, M., Gius, E. & Jacke, J. & Meister, J.C & Petris, M. & Strötgen, J. (im Erscheinen). Collaborative Text Annotation Meets Machine Learning. heureCLÉA, a Digital Heuristics of Narrative. *DHCommons Journal.*

Bögel, T. & Strötgen, J. & Gertz, M. (2014). Computational Narratology: Extracting Tense Clusters from Narrative Texts. *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC'14)*. Reykjavik, Iceland, S. 950-955.

Bühler, A. (2003). Grundprobleme der Hermeneutik. *Hermeneutik. Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*. Hg. von Axel Bühler. Heidelberg: Synchron, S. 3-19.

Gius, E. (2013). *Erzählen über Konflikte. Eine computergestützte narratologische Untersuchung von narrativen Interviews zu Arbeitskonflikten*. Dissertation, Universität Hamburg.

Gius, E. & Jacke, J. (in Vorbereitung). Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse. *Zeitschrift für digital Humanities.*

Gius, E. & Jacke, J. (2014). *Zur Annotation narratologischer Kategorien der Zeit*. *Guidelines zur Nutzung des CATMA-Tagsets.* Hamburg. http://heureclea.de/publications/guidelines.pdf/

Hearst, M.A. & Dumais, S.T. & Osman, E. & Platt, J. & Scholkopf, B. (1998). Support Vector Machines. *Intelligent Systems and their Applications 13 (4), IEEE,* S. 18-28.

Quinlan, J.R. (1986). *Induction of Decision Trees*. *Machine learning 1 (1)*, S. 81-106.

Vapnik, V. N. (1998). *Statistical Learning Theory*. *Vol. 2*. New York: Wiley.

Status und Probleme der Literaturwissenschaften im Rahmen der Digitial Humanities

In den Digital Humanties nehmen insbesondere (allerdings nicht ausschließlich) im angelsächsischen Raum quantitative Ansätze innerhalb der Literaturwissenschaft eine durchaus prominente Position ein. Allein in den letzten Jahren sind eine Reihe von Studien und Arbeiten in diesem Bereich erschienen (Jockers 2014, Jockers 2013, Siemens/Schreibman 2007, Allison et al. 2011, Moretti 2007, u.a.m.). Bemerkenswerter geht es dabei um ein breites Spektrum unterschiedlichster Fragestellungen und Ansätze, angefangen von der Konzentration auf einzelne literarische Texte bis hin zu umfassenden Fragen der literarischen Evolution. Methodisch reicht die Palette von der Edition und Archivierung über die Visualisierung von komplexen Datenstrukturen bis hin zur Anwendung quantitativer Verfahren sowohl auf synchrone als auch auf diachrone Probleme der Literaturwissenschaft. Dabei ist – so wird im Vortrag argumentiert – aus wissenschaftstheoretischer Sicht u.a. ein vergleichsweise geringer Grad an Reflektion hinsichtlich der Anwendung von quantitativen Methoden zu bemerken, was nicht zuletzt in einer fehlenden Thematisierung der grundlegenden Positionierung einer „digitalen" Literaturwissenschaft zum Ausdruck kommt.
Im Vortrag wird im Detail auf den in den genannten Studien verwendeten Methodenapparat einzugehen sein, der sowohl aus dem Bereich der Linguistik als auch der Statistik schöpft. Insbesondere wird aber trotz der zum Teil durchaus elaborierten Methodik auf die fehlende Modellbildung im Sinne der modernen quantitativen Sprach- und Textanalyse zu thematisieren sein, die einen entscheidenden Schritt bei der Überwindung von deskriptiven Ansätzen darstellen könnte. Darüber hinaus wird abschließend versucht, zumindest kursorisch auf ältere deutsche und russische Konzepte einer exakten Literaturwissenschaft zu verweisen, aus der sich methodologische Querbeziehungen zum gegenwärtigen Status der (digitalen) Geisteswissenschaften herstellen lassen. Im zweiten Teil des Vortrages werden ausgewählte Probleme der quantitativen Analyse von literarischen Texten anhand der lexikalischen Häufigkeit diskutiert. Hierbei ist vor allem auf das Problem von textübergreifenden sprachlichen Regulationsmechanismen und die unterschiedliche Länge von Texten näher einzugehen – beides zentrale Probleme für die entsprechende Lösungsansätze zu besprechen sind.

Literatur:

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore (2011). *Quantitative Formalism: An Experiment.* Stanford: Stanford Literary Lab.

Jockers, Matthew L. (2014): *Text Analysis with R for Students of Literature.* Cham: Springer (Quantitative Methods in the Humanities and Social Sciences, 2).

Jockers, Matthew (2013). *Macroanalysis: Digital Methods and Literary History.* UIUC Press, 2013.

Moretti, F. (2007) *Graphs, Maps, Trees: Abstract Models for a Literary History.* London: Verso.

Siemens, Ray; Schreibman, Susan (ed.) (2007): *A Companion to Digital Literary Studies.* Malden, Mass.

Wissenschaftsgeschichte und digitale Methoden. Eine datengestützte Untersuchung zu wissenschaftlichen Öffentlichkeiten*

Martin Fechner
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin-Brandenburgische Akademie der Wissenschaften

I. EINLEITUNG

In dem sich schnell entwickelnden Gebiet der digitalen Geisteswissenschaften gibt es viele verschiedene Projekte aus dem Bereich der Linguistik und Lexika [1], viele digitale Editionen [2] oder Forschungen zu anderen Gebieten wie Archäologie, Alterumswissenschaft oder Musikwissenschaft [3,4]. Diese werden unterstützt durch die Möglichkeiten der Digitalisierung [5], durch neue Werkzeuge [6,7], Datenanalyse [8] oder Visualisierungen [9]. Auch für andere Wissenschaften bieten die digitalen Methoden neue Ansätze für die Forschung [10]. Die Wissenschaftsgeschichte ist allerdings bisher wenig vertreten.

Wie im Rahmen der Wissenschaftsgeschichte offene Probleme anhand großer Datenmengen auf neue Weise behandelt werden können, wird in der hier präsentierten Arbeit gezeigt. Dort wurden die Strukturen moderner Kommunikation in der Wissenschaft untersucht. Dabei wurde die Methode der »Data Adaptation« [11] eingesetzt, um auf gezielte und transparente Weise zu Ergebnissen zu gelangen. Die so generierte Datensammlung ermöglichte eine Vielzahl von Analysen, die in diesem Fall halfen Kommunikationsstrukturen aufzudecken.

II. DEFINITION VON WISSENSKOMMUNIKATION

In der Wissenschaftsgeschichte ist es für die Forschung notwendig, den Einfluss der verschiedenen Akteure aufeinander zu untersuchen. Doch in der neueren Forschung sind sehr unterschiedliche Ansichten darüber entwickelt worden, was Wissenskommunikation bedeutet. Das fängt bei der Erweiterung der Definition des Wissensbegriffes an, der nicht ausschließlich auf wissenschaftliches Wissen angewendet wird, und setzt sich fort bei der Hinzuziehung aller möglichen Kommunikationsformen von oralen Medien über technische Medien bis hin zu Museen als lokalisiertes Medium.

Das führt zu einem unklaren Blick darauf, was Popularisierung ist. Kretschmann [12] bemerkt dazu, eine „einheitliche, allgemeinverbindliche Popularisierungsdefinition ist folglich nicht in Sicht“ und da die vielen Einzelstudien enge Grenzen hätten, würden „längerfristige Entwicklungen und epochenübergreifende Zusammenhänge kaum wahrgenommen“. Auch Ash [13] fragt sich, „was an die Stelle des obsoleten linearen, diffusionistischen Modells eines einzigen Verhältnisses von »Wissenschaft« und »Öffentlichkeit« treten soll“. Es gibt verschiedene Lösungsansätze [14,15], Whitley etwa definiert Popularisierung und Wissensverbreitung einfach als „Übertragung intellektueller Produkte von ihren Produktionskontexten hin zu anderen Kontexten“.

III. FORSCHUNGSANSATZ

In dieser Arbeit wurden die verschiedenen Ansätze zusammengeführt und um eigene Datenerhebungen erweitert, es wurden zunächst die Eigenschaften der Kommunikation, der Öffentlichkeit und die Entwicklung der Wissenschaft betrachtet. Darauf aufbauend konnten dann zwei wissenschaftliche Kommunikationsräume untersucht werden, um mögliche Strukturen aufzudecken und neue Ansätze für die weitere Forschung zu erhalten.

Um für eine Analyse der Kommunikationsstrukturen in der Wissenschaft zu entscheiden, worauf bei einer Untersuchung geachtet werden müsste, wurden für diese Arbeit zunächst die Modelle von Öffentlichkeiten weiterentwickelt und mit einem Datenmodell kombiniert. Da in einem Kommunikationsprozess immer nur ein bestimmtes Publikum erreicht wird und da die Definition einer allgemeinen Öffentlichkeit daher schwierig erscheint, werden hier statt gestuften Öffentlichkeiten Kommunikationsräume definiert und deren Eigenschaften beschrieben.

IV. DIE FALLBEISPIELE

Um über eine oberflächliche Analyse hinausgehen zu können, wurden zwei Kommunikationsräume exemplarisch ausgewählt, für welche es dann möglich war, eine detaillierte, datengestützte Untersuchung durchzuführen. Bei Entwicklung der Spektralanalyse, sowie des ersten Lasers handelt

* Eine genauere Beschreibung wird demnächst in der Dissertationsschrift des Autors gegeben [10].

es sich um genuin wissenschaftliche Laborforschung in ähnlichen Themenbereichen; beide gehören zu den optischen Phänomenen in der Physik. Durch den Abstand von 100 Jahren war es möglich Unterschiede und Kontinuitäten zwischen dem 19. Jahrhundert und dem 20. Jahrhundert festzustellen.

A. Die Entwicklung der Spektralanalyse

Die Spektralanalyse wurde Mitte des 19. Jahrhunderts entwickelt, als der Buchdruck durch die technischen Innovationen günstiger geworden war und neue Techniken wie die Fotografie andere Darstellungsformen zuließen. Gleichzeitig entstanden gesellschaftliche Öffentlichkeiten, die an Kommunikationsprozessen teilnehmen konnten. An den Universitäten emanzipierten sich die Geistes- und Naturwissenschaften, die Organisation der Wissenschaft wurde professioneller und die Zahl der ausgebildeten Personen wuchs, ebenso wie die Zahl der Professuren, was den wissenschaftlichen Diskurs veränderte.

In dieser Zeit machten der Physiker Gustav Kirchhoff und der Chemiker Robert Bunsen in Heidelberg zusammen Spektralbeobachtungen und entwickelten 1859 die Spektralanalyse [16]. Die von ihnen präzisierte Untersuchungsmethode und die zugehörige mathematische Theorie führten in der Folge zu einer großen Resonanz bei Wissenschaftlern, als auch beim interessierten Publikum. An die Spektralanalyse schlossen sich viele wissenschaftliche Entdeckungen, sowie konkrete Anwendungen an.

B. Die Erfindung des Lasers

In der Mitte des 20. Jahrhunderts wurde der Laser erfunden. Sowohl die Kommunikationsmittel, wie auch die Ausgestaltung der Öffentlichkeiten und der Wissenschaften hatte sich über die letzten hundert Jahre stark verändert. Es waren neue sekundäre technische Medien hinzugekommen und Massenmedien vergrößerten die Verfügbarkeit von überregionaler Kommunikation. Gleichzeitig hatte sich die Wissenschaft weiter professionalisiert und der innerwissenschaftliche Diskurs konzentrierte sich zunehmend auf die gestiegene Zahl an Fachwissenschaftlern.

In diesem Umfeld wurde 1960 in Kalifornien der erste Laser von Theodore Maiman entwickelt und in einer Pressekonferenz der Öffentlichkeit präsentiert [17]. Das Gerät entfaltete bei den Fachwissenschaftlern wie in der allgemeinen Presse eine große Wirkung. Obwohl sich die Spekulationen über militärische Anwendungen nicht erfüllten, gab es für den Laser viele Anwendungsmöglichkeiten in Forschung und Gesellschaft.

V. DATA ADAPTATION

Die Datenerhebung dieser Arbeit konzentrierte sich auf die beiden Fallbeispiele und dort auf Buchpublikationen, sowie auf eine Vollerhebung ausgewählter Zeitschriften innerhalb definierter Zeiträume. Erweitert wurden die Daten, um eine Detailanalyse der Zeitschriftenartikel und einzelner Lehrbücher.

Für die recherchierten Publikationen wurde ein Datenmodell entwickelt, welches verschiedene Anforderungen miteinander kombinierte. Darin sollten die allgemeinen Angaben, sowie eine Klassifizierung des Kommunikationsraumes der Publikation festgehalten werden können, desweiteren sollte darin auch eine Detailanalyse ohne Mehraufwand notiert werden können. Es wurden etwa die Darstellungsformen, sowie die im Text erwähnten Zitierungen notiert. Durch die Verwendung von XML [18] konnten diese Auflagen erfüllt werden und zusätzlich wurde dadurch eine computergestützte Auswertung über die Verwendung von XSL und Analyseprogrammen möglich.

Es wurden insegesamt über 1.500 Publikationen zur Spektralanalyse und zum Laser recherchiert und in das XML-Format gebracht. Weiterhin wurden mehr als 600 Zeitschriftenartikel aus drei Zeitschriften im Detail analysiert und in Datensätzen festgehalten. Die XML-Daten wurden mithilfe von selbst gefertigten Analyseskripten zu Tabellen und komplexen Visualisierungen aufbereitet. Das erlaubte die Auswertung der Datenmenge als Übersicht, aber auch im Detail, um die Vielschichtigkeit der Daten zu begreifen und das wesentliche zu erkennen.

VI. ERGEBNISSE

Die statistischen und detaillierten Betrachtungen der Quellen konnten bestimmte Kommunikationsstrukturen aufzeigen. So lassen sich Kontinuitäten und Dominanzen von Wissensorganisationen zeigen. Ebenso gibt es deutliche Hinweise auf Ausbreitungsmechanismen von neuen wissenschaftlichen Themen. Der Weg durch die wissenschaftlichen Zeitschriften beeinflusst Wissenschaftler zu weiteren Forschungen, nach den Differenzierungen innerhalb eines Themas und der Entwicklung von Anwendungen gibt es eine weitere Verarbeitung in zusammenfassenden Lehr- und Handbüchern. Auch konnten verschiedene Rollen identifiziert werden,

die Wissenschaftler während des wissenschaftlichen Forschungs- und Kommunikationsprozesses einnehmen und es konnte gezeigt werden, wie in den Medien auf verändertes Wissen reagiert wird und wie Wissen selbst während des Transferprozesses angepasst wird.

Die hier präsentierte Arbeit demonstriert, wie mit einer systematischen Vorgehensweise und computergestützten Verfahren neue Erkenntnisse gewonnen werden können. Die Ergebnisse fügen sich gut in den Forschungskontext zur Entwicklung der Wissenschaft ein und können einen Wandel der Wissenschaft, aber auch bleibende Strukturen zwischen dem 19. und dem 20. Jahrhundert anhand der Publikationsdaten belegen.

VII. LITERATUR*

[1] Das Digitale Wörterbuch der deutschen Sprache (DWDS) an der Berlin-Brandenburgischen Akademie der Wissenschaften. – <http://www.dwds.de>

[2] Patrick Sahle, A catalog of Digital Scholarly Editions. – <http://www.digitale-edition.de/>

[3] Berliner Antike-Kolleg und das Excellence Cluster TOPOI. – <http://berliner-antike-kolleg.org>, <http://www.topoi.org/>

[4] Digitale Mozart-Edition (DME) an der Stiftung Mozarteum Salzburg (ISM). – <http://dme.mozarteum.at/DME/main/?>

[5] Digitalisierung und Digitale Sammlungen, Münchner Digitalisierungszentrum (MDZ) an der Bayerischen Staatsbibliothek. – <http://www.muenchener-digitalisie rungszentrum.de>

[6] DARIAH-DE Tools für die Geistes- und Kulturwissenschaften. – <https://de.dariah.eu/tools>

[7] Werkzeuge und Dienste, CLARIN-D. – <http://de.clarin.eu/de/sprachressourcen/werkzeuge-und-dienste.html>

[8] Stéfan Sinclair und Geoffrey Rockwell: *Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies*. In: Digital Humanities Pedagogy: Practices, Principles and Politics, hg. v. B. D. Hirsch, Cambridge 2012. – <http://dx.doi.org/10.1164 7/OBP.0024>

[9] Marian Dörk, Heidi Lam und Omar Benjelloun: *Accentuating Visualization Parameters to Guide Exploration*. In: CHI 2013: Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems, ACM, May 2013, S. 1755-1760. – <http://mariandoer k.de/accentuation/>

[10] Martin Fechner: *Kommunikation von Wissenschaft in der Neuzeit*, Diss. (in Arbeit).

[11] Martin Fechner: *»Data Adaptation« als Analysemethode für geisteswissenschaftliche Forschung*. In: 1. Jahrestagung DHd Konferenz, Passau, 25.-28.3.2014, <https://www.conftool.pro/dhd2014/index.php?page=browseSessions&path=adminSessions&form_session=2 4>.

[12] C. Kretschmann: *Einleitung: Wissenspopularisierung - ein altes, neues Forschungsfeld*. In: Wissenspopularisierung, hg. v. dems., Berlin 2003, S. 7-21.

[13] M. G. Ash: *Wissenschaft(en) und Öffentlichkeit(en) als Ressourcen füreinander. Weiterführende Bemerkungen zur Beziehungsgeschichte*. In: Wissenschaft und Öffentlichkeit als Ressourcen füreinander, hg. v. S. Nikolow u. A. Schirrmacher, Frankfurt/New York 2007, S. 349-362.

[14] Es gibt Ansätze von R. Whitley: *Knowledge Producers and Knowledge Acquirers. Popularisation as a Relation between Scientific Fields and their Publics. In: Expository Science*, hg. v. T. Shinn u. R. Whitley, Dordrecht 1985, S. 3-28

[15] J. Renn u. M. D. Hyman: *The Globalization of Knowledge in History: An Introduction*. In: The Globalization of Knowledge in History, hg. v. J. Renn, Berlin 2012, S. 15-44.

[16] Jochen Hennig: Der Spektralapparat Kirchhoffs und Bunsens (=Deutsches Museum: Wissenschafts- und Technikgeschichte. Originale, Modelle und Rekonstruktionen, Band 1), München 2003.

[17] Jeff Hecht: *Beam. The race to make the laser*, Oxford 2005.

[18] Tim Bray, Extensible Markup Language (XML) 1.0 (Fourth Edition) – Origin and Goals, veröffentl. vom W3C am 29. September 2006. – <http://www.w3.org/TR/2006/REC-xml-2006081 6/#sec-origin-goals>

* Alle aufgeführten Webseiten wurden am 8.11.2014 abgerufen.

Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung

1. Einleitung

Wenn Forschungsdaten in mehreren geisteswissenschaftlichen Disziplinen Verbreitung und Anwendung finden sollen, dann kann dies über einen gemeinsamen Zugriff realisiert werden.[1] So entstehen Synergien, z.B. in Hinblick auf die nicht doppelt zu leistende Digitalisierung oder Annotation von Quellen. Repositorien müssen damit in der Lage sein, den Zugriff auf eine heterogene Menge an Forschungsdaten zu ermöglichen. Die vorliegende Arbeit zeigt, wie ein Forschungsdatenmodell für Metadaten, das in der Korpuslinguistik entwickelt wurde, auch in anderen Geisteswissenschaften für textuelle Daten einen solchen Zugriff in Verbindung mit der technischen Umsetzung realisiert. Dabei soll während des Zugriffs der Entstehungsprozess der Daten überblickt und verstanden werden, um so die Daten korrekt referenzieren oder wiederverwenden zu können.[2]

2. Forschungsdatum Korpus

In der Korpuslinguistik wird ein Korpus als ein Digitalisat unterschiedlichster sprachlicher Primärquellen, die strukturiert mit weiteren Informationen – Annotationen – angereichert sind, verstanden (vgl. Lemnitzer & Zinsmeister 2006, 7). Der Erkenntnisgewinn erfolgt bei korpuslinguistischen Studien über eine durch Annotationen gestützte, qualitative oder quantitative Analyse natürlichen und authentischen Sprachmaterials.[3] Es ist dabei nicht klar, was jeweils unter sprachlichen Primärquellen verstanden wird (vgl. z.B. Claridge 2008, Himmelmann 2012). Die Ausweisung, was in einem Korpus ein Primärdatum ist, wird über die Forschungsfrage und Theorie zur Forschung motiviert.[4] Auch die Bedeutung der Annotationen und deren Zuweisung und Auswertung kann nicht von der jeweiligen Forschungsfrage der Korpusersteller getrennt werden.[5] Beide Konzepte werden technisch in den Korpora, konkreter in diversen Annotationstypen und deren Formaten umgesetzt (vgl. dazu Zipser 2014).
Die Korpuslinguistik steht demnach in einem Spannungsfeld zwischen dem technischen Verständnis der Korpora und der theoretisch-wissenschaftlichen Motivation der Korpuserstellung.

> ‘On an abstract technical level, there are no categorical differences between a large corpus for a well-researched language with many resources and a standardized orthography and a corpus of an endangered language or small variety without codified standards: In both cases one needs to represent a source text and annotations to it.’(Lüdeling 2012, 32)

Genau diese technisch-abstrakte Perspektive ist die Grundlage in dieser Arbeit. Wenn beispielsweise die Korpuseinheit „Token“ theoretisch mit einem Konzept von „Wort“ motiviert wird, dann kann sie als eine konzeptionelle Größe eines Korpus betrachtet werden. Das „Token“ kann auch als kleinste technische und zu annotierende Einheit in einem Korpus verstanden werden und besitzt so keinen theoretisch motivierten Wert (vgl. Krause et al. 2012). Gleiches gilt für Annotationen: Die Bedeutungen der linguistischen Annotation wie Wortartenannotationen sind insofern relative Größen, als dass ihre Bedeutungen immer im Bezug zur Forschungsfrage stehen.[6] Damit gelten solche Definitionen in der Regel für nicht mehr als ein Korpus. Darüber ist ableitbar, dass Korpora allgemein nicht über ein festes Set von Annotationen und eine eigenständig identifizierbare Primärtextebene definiert werden können.

---

[1] Im Gegensatz zu einem nicht einheitlichen Zugriff auf einzelne Forschungsdaten wie Projektwebseiten oder Anwendungen für einen Typ Korpus (z.B. TIGERSearch Suchwerkzeug Lezius 2002).
[2] Unter Wiederverwendung von Korpora wird eine erneute Analyse der vorhandenen Korpora oder die Anreichung dieser mit weiteren Annotationen verstanden.
[3] Dabei kann es sich um mündliche oder schriftliche, synchrone oder diachrone, moderne oder historische Sprache handeln. Je nach Fragestellung entsteht ein Korpus zusammengestellt bspw. nach Register, Datum, Ort etc.
[4] Diese Vielfalt zeigt sich beispielsweise anhand der Korpora, die über das Such- und Visualisierungstool ANNIS (Krause & Zeldes erscheint, Zeldes et al. 2009 http://www.sfb632.uni-potsdam.de/annis/) zur Verfügung stehen.
[5] Es gibt wenige häufig genutzte Quasistandards für Annotationen wie das STTS (Schiller et al. 1999). Was genau in den jeweiligen Studien unter Wortarten, Satzgliedern, etc. und deren Annotationen verstanden wird, ist bereits Teil der Forschung und damit immer Interpretation (vgl. Lüdeling 2011).
[6] Natürlich können „Wortarten“ oder „Satzglieder“ als kategoriale Größen definiert und so annotiert werden.

Ein Forschungsdatenmodell für Korpora muss demnach theorieneutral diese Konzepte erfassen und abbilden können. Zu berücksichtigen sind dennoch die theoretischen Fragen nach dem Primärdatum sowie nach Werten und Bedeutungen von Annotationen, ohne jedoch eine feste Aussage darüber treffen zu müssen.

3. Forschungsdatenmodell

Wenn nicht die Bedeutung der Primärquelle oder die der Annotationen Grundlage für ein Forschungsdatenmodell sein können, dann sind es die technischen Eigenschaften. Dies wird in dem hier vorgestellt Modell aufgenommen. Dabei gilt folgendes (Odebrecht 2014): Ein *Korpus* ist die Summe seiner *Dokumente*. Ein *Dokument* ist die Summe seiner *Annotationen* (Abbildung 1).

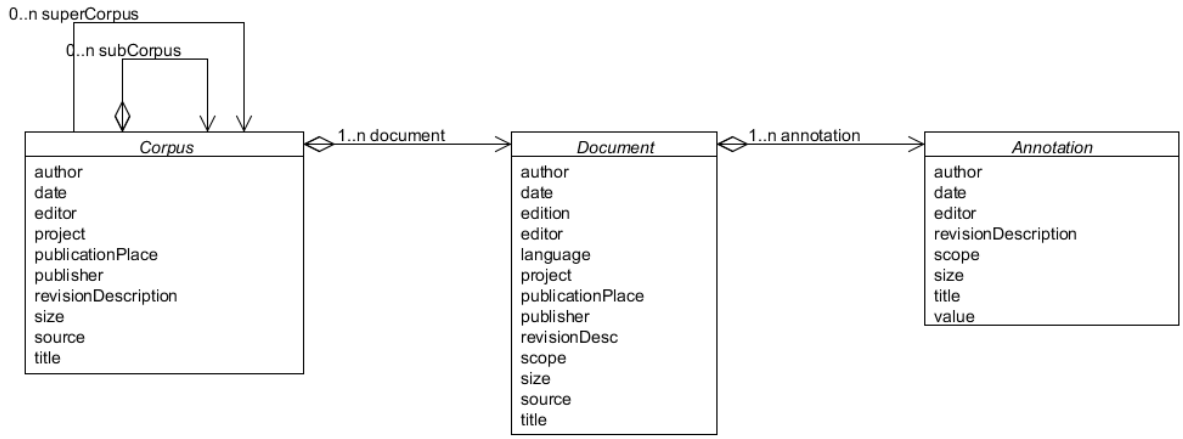


Abbildung 1 Forschungsdatenmodell für Metadaten

Die *Dokumente* definieren sich nicht zwingend nur über ihre konzeptionellen bibliographischen Eigenschaften, aber immer über ihre technische Aufbereitung. Sie bestehen wiederum aus der Summe ihrer *Annotationen*, die von *Dokument* zu *Dokument* unterschiedlich sein kann. So können mehrere *Dokumente* dieselbe Textvorlage wie bspw. ein historisches Buch des 18. Jahrhundert besitzen, aber dennoch als getrennte *Dokument* verstanden werden, da die *Annotationen* getrennt voneinander vorhanden sind. Die Einheit *Annotation* definiert sich aus dem Annotationsschlüssel und dessen Werten, die typischerweise in einem konkreten Format abgebildet sind und sich so wiederum durch ihre technischen Eigenschaften ausweisen lassen.[7] Weiterhin können Subkorpora abgebildet werden, die dann in ihrer Summe ein Superkorpus bilden. Durch die Summenmodellierung kann die Versionsgeschichte eines *Korpus*, mit steigender oder fallender Anzahl von Subkorpora, *Dokumenten* und *Annotationen* modelliert werden.

Für alle Instanzen, die mit diesem Modell abgebildet werden sollen, muss so nicht einheitlich abgebildet werden, welche theoretischen Konzepte hinter den einzelnen Summen stehen oder wie diese aus der Fachwissenschaft heraus motiviert werden. Zum Beispiel: Das diachrone *Korpus* RIDGES Herbology Corpus[8] beinhaltet neben einer Transkription auch mehrere Normalisierungen und linguistische sowie Markup-Annotationen. In diesem Fall werden die Transkription und die Normalisierungsebenen als *Annotation* neben anderen *Annotationen* verstanden und fallen unter dieselbe Summenregel. Damit stellt sich nicht die Frage nach einer Ausweisung des Primärtextes und die Bedeutungen der einzelnen Annotationsebenen sind nicht im Modell verankert. Dies wäre gänzlich irreführend, wie folgendes Minimalbeispiel zeigt: Die Annotation „lb“ wird häufig mit der Bedeutung Zeilenumbruch (‚line break‘) aus den TEI Guidelines (Burnard & Bauman 2008) verbunden – so auch in RIDGES. Sie kann jedoch auch für ein eigenständiges Konzept stehen, nämlich für eine textlinguistische Kategorie der Sachverhaltsdarstellung wie bspw. im Kasseler Junktionskorpus[9]. Das

[7] Alle Annotationen (Token-, Spannen, Baumannotationen) besitzen jeweils mindestens einen Annotationsschlüssel und -wert und können über diese Gemeinsamkeit im Modell zusammengefasst werden.
[8] http://hdl.handle.net/11022/0000-0000-2D32-6
[9] http://hdl.handle.net/11022/0000-0000-2102-8

Modell abstrahiert über alle *Annotationen*, ist damit eben nicht an Konzepte[10] gebunden, sondern überlässt letztere der jeweiligen Forschung.[11]
Zusammen mit einer Arbeitsgruppe aus der Musiksoziologie wird das Modell nun erstmals für Forschungsdaten der Editionswissenschaft getestet. So entsteht eine digitale Edition in TEI-XML für den Nachlass des „Vereins für musikalische Privataufführungen“[12]. Das Korpus wird damit Konzepte dieses Fachbereichs tragen, die sich in allen Klassen des Modells wiederfinden. Neben der nicht linguistisch motivierten Transkription werden Kategorien vergeben, die für die Erforschung der Vereinsstruktur relevant sind. Gerade die Arbeit mit Personen- oder Publikationsreferenzlisten in der Annotation ist hierfür essenziell.[13] Wir werden zeigen, wie diese Daten ebenfalls im Modell abgebildet werden können und somit für eine technisch-abstrakte Modellierung von geisteswissenschaftlichen Korpora argumentieren.

4. Forschungsdatenmodell und Metadaten

Dieses Modell regelt den Zugriff auf Korpora für das Open-Access-Forschungsdatenrepositorium LAUDATIO (Krause et al 2014) mittels Metadaten: Die Klassen *Korpus*, *Dokument* und *Annotation* werden jeweils mit Metadaten beschrieben, um ein Korpus in einer Menge von Korpora zu finden oder zu dokumentieren. Der Fokus liegt weniger auf der korpuslinguistischen Forschung sondern mehr auf der Entstehung dieser Daten und ihren Lebenszyklen (vgl. Rümpel 2011).
Zur Beschreibung wird ein Metadatenschema genutzt, das aus einer Anpassung der TEIp5 via ODD generiert wird (Burnard & Rahtz 2004). [14] Das Metadatenschema gibt vor, welche Metadatenattribute zu welcher Klasse beschrieben werden müssen. Die Werte der Metadatenattribute sind dann spezifisch für jedes Korpus. So kann dokumentiert werden, dass mehrere *Dokumente* eine gemeinsame textuelle Vorlage besitzen, wenn sie dieselben bibliographischen Metadaten teilen und somit konzeptionell als Abschnitte eines Buches interpretiert werden können. Für die Ausweisung des Primärtextes für ein einzelnes Korpus gibt es für alle *Annotationen* ein Metadatum, das besagt, ob sie jeweils technisch gesehen eigenständige Segmentierungen besitzen. Wenn dies der Fall ist, dann kann das auf eine primäre Textebene hinweisen (Krause et al 2012). Wenn es mehrere *Annotationen* gibt, die so ausgewiesen sind, kann es sich um eine Parallelkorpus handeln. Weiterhin werden die Annotationen mithilfe der Metadaten in Kategorien eingeteilt, die wiederum im Repositorium die Menge strukturieren und durchsuchbar machen. Diese Kategorien werden post hoc und für den Anwendungsfall LAUDATIO gebildet. Wenn andere Disziplinen neue Kategorien benötigen, dann können diese ohne Änderung des Modells hinzugefügt werden. So können linguistische Annotationen gemäß dem TIGER Schema[15] gleichwertig zu Markupannotationen, welche in den Editionswissenschaften benutzt werden, beschrieben werden. Die Spezifizierung erfolgt zweckgebunden immer an der Oberfläche und ist für ein breites Spektrum anderer Fachwissenschaften offen.

Referenzen

Broeder, D., Kemps-Snijders, M., et al. (2010). A data category registry- and component-based metadata framework. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC ’10) , Valletta, Malta. ELRA.
Burnard, L., Bauman, S. (Ed.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.*
Burnard, L., Rahtz, S. (2004) RelaxNG with Son of ODD. *Extreme Markup Languages Proceedings 2004*. Montréal, Québec.

[10] Einige Formate wie das EXMRALDA-Format identifizieren Primärebenen, hier verschiedene Sprechebenen (Schmid & Wörner 2009).
[11] Einen ähnlichen Ansatz besitzt das Projekt FREEBANK, das französische Korpora frei zur Verfügung stellt (Salmon-Alt 2006).
[12] Seminar „Der Nachlass des Vereins für musikalische Privataufführungen - digitale Edition in der Musikwissenschaft“ Humboldt-Universität zu Berlin, geleitet von Katrin Bicher.
[13] Solche Referenzlisten sind auch in der Linguistik gängig. So wird bspw. ISOCAT für die Referenzierung von Annotationsbedeutungen verwendet (vgl. Wright, Kemps-Snijders & Windhouwer 2007, http://www.isocat.org/).
[14] Frei verfügbar unter https://github.com/korpling/LAUDATIO-Metadata. Die Text Encoding Initiative ist in vielen Geisteswissenschaften bereits etabliert und findet viel Abdeckung. Deswegen wurde sie aus anderen Frameworks zur Beschreibung von Metadaten wie bspw. CMDI (Broeder et al. 2010) gewählt.
[15] https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-annot.pdf

Claridge, C. (2008) Historical Corpora. In Lüdeling, A., Kytö, M. (Hg.) *Corpus Linguistics. An International Handbook*. Vol 1. De Gruyter, Berlin. 242–259.
Himmelmann, N. P. (2012) Linguistic Data Types and the Interface between Language Documentation and Description. In *Language Documentation & Conservation* 6. 187-207.
Krause, Th., Zeldes, A**.** (2014) *ANNIS3: A new architecture for generic corpus query and visualization.* In Literary and Linguistic Computing. http://llc.oxfordjournals.org/content/early/2014/10/24/llc.fqu057.abstract
Krause, Th., Lüdeling, A., Odebrecht, C., Romary, L., Schirmbacher, P., Zielke, D. (2014) LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository. *Digital Humanities 2014 Conference. Poster Session. 8.7.-12.7.2014, Lausanne.* http://www.laudatio-repository.org/
Lemnitzer, L., Zinsmeister H. (2006) *Korpuslinguistik. Eine Einführung.* Gunter Narr Verlag, Tübingen.
Lezius, W. (2002) *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
Lüdeling, A. (2012) A corpus-linguistics perspective on language documentation, data, and the challenge of small corpora. In Seifart, F., Haig, G., Himmelmann, N. P., Jung, D.,; Margetts, A. & Trilsbeek, P. (Hg.) *Potentials of Language Documentation: Methods, Analyses, and Utilization.* Language Documentation & Conservation Special Publication No. 3 at the University of Hawai'i Press. 32-38.
Lüdeling, A. (2011) Corpora in Linguistics: Sampling and Annotation. In Grandin, K. (Hg.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* [Nobel Symposium 147]. Science History Publications/USA, New York. 220-243.
Rümpel, St. (2011)Der Lebenszyklus von Forschungsdaten. In Büttner, St., Hobohm, H. & Müller, L. (Hg.) Handbuch Forschungsdatenmanagement. Bock und Herchen Verlag. Bad Honnef. 25-31.
Salmon-Alt, S., Romary, L., Pierrel, J. (2006) Un modèle générique d'organisation de corpus en ligne : application à la FReeBank. Traitement Automatique des Langues, ATALA, 2006, 45, 145-169. <hal-00110970>
Schiller, A., Teufel, S., Stöckert, Ch., Thielen, Ch. (1999) Guidelines für das Tagging deutscher Textkorpora mit STTS. Technischer Report. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung & Universität Tübingen, Seminar für Sprachwissenschaft.
Schmidt, Th., Wörner, K. (2009) EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics 19/4.* 565-582.
Wright, S.E., M. Kemps-Snijders, M., Windhouwer, M. (2007) ISO-Cats: The Revised and Future TC 37 Data Category Registry. Presentation at the *Pragmatic Applications for TC 37 Standards* (TC37 2007), Provo, UT USA, August 13, 2007.
Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009) ANNIS: A search tool for multi- layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*. Liverpool. http://www.sfb632.uni-potsdam.de/annis/
Zipser, F. (2014) SaltNPepper und das Formatpluriversum. LAUDATIO-Workshop 07.10.2014. Berlin.

Digitale Geisteswissenschaften und Informatik – Modelle der Zusammenarbeit

Der hier eingereichte Abstract versteht sich als ein Beitrag zum Rahmenthema der DHd 2015 *Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation* und als ein Beitrag zu einer der drei Leitfragen der Tagung, „... welche disziplinübergreifenden Synergien für die Theoriebildung aus den in den Digitalen Geisteswissenschaften entwickelten Methoden, Techniken und Forschungsinfrastrukturen zu erwarten sind." (vgl. http://www.dig-hum.de/jahrestagung-dhd-2015). Die Frage nach disziplinübergreifenden Synergien setzt eine disziplinübergreifende Fundierung der Digitalen Geisteswissenschaften implizit voraus bzw. ist durch eine solche Fundierung in jedem Fall leichter zu beantworten.

Eine disziplinübergreifende Fundierung der Digitalen Geisteswissenschaften muss zwei Anforderungen gerecht werden: (i) sie muss von Methoden und Techniken einzelner geisteswissenschaftlichen Disziplinen hinreichend abstrahieren sowie (ii) in Hinblick auf Methoden und Techniken der Informatik hinreichend anschlussfähig sein. Die Frage nach dem Verhältnis der Geisteswissenschaften und speziell der Digitalen Geisteswissenschaften zur Informatik zu stellen, erscheint gerade zum jetzigen Zeitpunkt äußerst sinnvoll. Der Bereich der Digitalen Geisteswissenschaften hat in den letzten Jahren einen großen Aufwuchs erfahren. Dies wird durch den Anstieg der Teilnehmerzahlen bei nationalen und internationalen Konferenzen und der Gründung von Fachverbänden eindrucksvoll dokumentiert. Noch wichtiger sind das wachsende Angebot an fachlich einschlägigen Studiengängen, die steigende Anzahl an Forschungsprojekten und der damit verbundene Einzug von Methoden der Digitalen Geisteswissenschaften in einen immer größeren Kreis geisteswissenschaftlicher Fächer. Beide Entwicklungen haben sicherlich auch damit zu tun, dass mit einer neuen Generation von NachwuchswissenschaftlerInnen als *Digital Natives* das Spektrum an wissenschaftlicher Neugier an und Kompetenz im IT Bereich kontinuierlich zunimmt.

Die Informatik befindet sich ebenfalls in einer Umbruchsphase. Traditionell hat sich die Informatik in ihren Grundlagen auf die Mathematik und Logik berufen, speziell hinsichtlich der Theorie der Berechenbarkeit und hinsichtlich der Komplexität und der Spezifikation von Algorithmen. Durch jüngste und aktuelle Entwicklungen im sozialen und mobilen Computing und die technologischen Möglichkeiten und ethischen Herausforderungen von Big Data stellt sich die Frage nach den theoretischen Grundlagen des Fachs und nach dem Begriff des Computing selbst in einer neuen Weise.

Beide Bereiche befinden sich somit in einer Übergangsphase. Dies sollte beim Nachdenken über Grundsatzfragen nicht außer Acht gelassen werden. In einer Situation des Umbruchs werden mögliche Antworten immer vorläufig und weiter klärungsbedürftig sein. Die Gefahr besteht darin, sich im Grundsätzlichen zu verlieren und dabei vergangene und gegenwärtige Praxis zu ignorieren. Viel folgenreicher ist allerdings das Risiko, die Grundsatzdiskussion allein an vergangener und gegenwärtiger Praxis zu orientieren und dadurch Entwicklungsmöglichkeiten für zukünftige oder zukunftweisende Kooperationen zwischen den Geisteswissenschaften und der Informatik zu behindern oder gar gänzlich abzuschneiden.

Eine weitere Gefahr, der sich Grundsatzdiskussionen aussetzen, besteht darin, dass sie bisherige Bemühungen in dieser Richtung bewusst oder unbewusst außer Acht lassen. Es erscheint daher sinnvoll, zunächst auf zwei solcher bisheriger Versuche aufmerksam zu machen.

Der eine Versuch einer Grundsatzdiskussion erscheint darin zu bestehen, bestimmte geisteswissenschaftliche Disziplinen aus den Digitalen Geisteswissenschaften sozusagen

auszubürgern und die Digitalen Geisteswissenschaften eher abgrenzend und implizit negativ zu definieren. Solche Versuche sind wenig zielführend, zumal sie zumeist nur von einer Außensicht auf die auszubürgernden Disziplinen gekennzeichnet sind und daher häufig nur mittelbar mit der tatsächlichen Forschungstradition und aktuellen Praxis in diesen Fächern zu tun haben. So wird die Sprachwissenschaft häufig mit dem sicherlich einflussreichen Paradigma der generativen Grammatikforschung gleichgesetzt. Dies wird weder der langen Forschungstradition noch den aktuellen Forschungsentwicklungen des Fachs gerecht, die nicht zuletzt durch die digitale Revolution in den Geisteswissenschaften angestoßen wurden und werden.

Der Versuch, die Digitalen Geisteswissenschaften ausgrenzend zu charakterisieren, erscheint auch insofern wenig hilfreich, als sich die Digitalen Geisteswissenschaften selbst dem untauglichen Versuch ausgesetzt sieht, sie sozusagen aus den Geisteswissenschaften auszubürgern. Bei diesen Versuchen spielen Hinweise auf vermeintliche Dichotomien von quantitativer versus qualitativer Forschung und von der hermeneutischen Tradition des Verstehens und des Erklärens einerseits und der rein beschreibenden, empirischen Wissenschaft anderseits eine wesentliche Rolle.

Ein zweiter und wesentlich aussichtsreicherer Versuch, die Digitalen Geisteswissenschaften inhaltlich zu charakterisieren, ist von Willard McCarty, einem der Pioniere der die Digitalen Geisteswissenschaften und Preisträger des 2013 Father Busa Awards, unternommen worden. McCarty und Ko-autoren (vgl. u.a. [1],[2]) charakterisieren die Digitalen Geisteswissenschaften nicht durch den Rekurs auf einen Fächerkanon, sondern inhaltlich über das Konzept der Modellierung. McCarty bezieht sich beim Begriff *Modellierung* auf Marvin Minskys Modellbegriff ([3]). Minsky betrachtet Objekte A* als Modelle eines Objekts A, wenn sie es einem Betrachter B ermöglichen, Antworten auf interessante Fragen zu einem Objekt A zu geben. Anknüpfend an die Anforderung Minskys, dass Modelle Einsichten vermitteln, versteht McCarty Modellierung nicht als etwas Statisches, sondern als einen heuristischen Erkenntnisprozess. Modellierung in diesem Sinne verstanden hat für ihn zwei Aspekte oder auch Phasen: die Konstruktion und die Verarbeitung. Die Verarbeitungsphase bezeichnet er auch als Manipulation.

McCartys Modellierungsbegriff erscheint außerordentlich hilfreich, um das Verhältnis der Digitalen Geisteswissenschaften zur Informatik zu klären – und in diesem Sinne wird er auch von McCarty verwendet. Er ist disziplinenunabhängig und hat ein sinnvolles disziplinübergreifendes Abstraktionsniveau. Er ist anschlussfähig sowohl hinsichtlich aktueller Entwicklungen in den Digitalen Geisteswissenschaften als auch in der Informatik. Er erlaubt unter anderem auch eine Antwort auf eine Frage, die bereits bei der Kontroversdiskussion zur *Beziehung zwischen der Computerlinguistik und den Digital Humanities* (http://www.dhd2014.uni-passau.de/programm/) auf der DhD 2014 in Passau gestellt wurde: Ist die Informatik Diener oder Partner der Digitalen Geisteswissenschaften? In der Dienerrolle finden sich all diejenige wieder, denen mit dem Hinweis, sie seien doch Informatiker, einmal die Frage gestellt wurde, ob sie nicht bei der Installation eines Fonts, einer Datenbank und – noch schlimmer – bei der Installation der neuen MS Office Version helfen könnten. Dies ist für die Informatik kaum eine attraktive Perspektive der Zusammenarbeit. Wie eine produktivere Zusammenarbeit zwischen den die Digitalen Geisteswissenschaften und der Informatik aussehen kann, skizzieren wiederum McCarty und Kollegen, indem sie fordern: „*If we wish to have a computing of the humanities, we need to be asking a rather different question: ‚how best can we integrate automated processing with human thinking and acting?‘*.“ ([1], S. 142).

Der Begriff der Modellierung und die anzustrebende Synthese aus automatischer Verarbeitung und menschlichem Erkenntnisprozess stellen für McCarty und Kollegen den Versuch dar, die unterschiedlichen Bedeutungsbegriffe zwischen Natur- und Geisteswissenschaften zueinander in Beziehung zu setzen. Beim Minskyschen Modellbegriff ist der Betrachter, der *Observer*, ein integraler Bestandteil der Definition. Modelle erlauben

es **dem Betrachter**, interessante Antworten auf Fragen zu einem Gegenstand zu formulieren. Dadurch fließt der Aspekt der Interpretation und des Verstehens, der für die geisteswissenschaftliche Forschung konstitutiv ist, in den Prozess der Modellierung im McCartyschen Sinne unmittelbar ein. Da empirisch fundierte Modelle immer auf Daten bezogen sind, kommt solchen Modellen genau die Mittlerfunktion zwischen Information und Interpretation zu, von der im Rahmenthema "*Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation*" der DHd 2015 die Rede ist**.**

Wenn man die Integration von automatischer Verarbeitung in den geisteswissenschaftlichen Erkenntnisprozess zum Ziel erhebt, dann rückt dadurch der geisteswissenschaftlicher Forscher in den Blickpunkt der Zusammenarbeit mit der Informatik. Es lassen sich dann verschiedene Optionen für eine solche Zusammenarbeit unterscheiden.

Digitale Geisteswissenschaften	**Informatik**
naive Anwender	generische Standard-Modelle und Standard-Verfahren
aufgeklärte Anwender	
kompetente Anwender und Modellierer	Domänenadaptierte Standard-Modelle und Standard-Verfahren
kreative und innovative Modellierer	neuartige Modelle und Methoden

Abbildung 1: Optionen der Zusammenarbeit

Naive Anwender haben keinerlei Detailkenntnisse über den technischen Hintergrund der von ihnen eingesetzten Modelle und Verfahren. Sie sind daher darauf angewiesen, Standardmodelle und -verfahren anzuwenden. Bei naiven Benutzern kommt es häufig zu Frustrationen und zur Abkehr von automatischen Verfahren, wenn diese sich als fehleranfällig erweisen und nicht perfekt funktionieren. Aufgeklärte Anwender hingegen wissen um solche Schwächen und können das automatischen Verfahren inhärente Rauschen beim Interpretationsprozess der Ergebnisse solcher Verfahren berücksichtigen, indem sie entweder die Analysedaten manuell überprüfen oder nachkorrigieren oder die Aussagekraft solcher Analysen zu bewerten wissen. Naive und aufgeklärte Benutzer haben gemein, dass sie Standardmodelle und –verfahren einsetzen. Um möglichen Missverständnissen vorzubeugen: man sollte den Einsatz von Standardmodellen und –verfahren in den Digitalen Geisteswissenschaften keineswegs kleinreden oder geringschätzen. So können mit Hilfe von Standardverfahren aus der Informatik -- zu nennen sind hier zum Beispiel die Bereiche der Datenvisualisierung, des Data Minings oder Methoden des maschinellen Lernens -- sehr wohl neue Erkenntnisse zu bekannten Fragestellungen gewonnen werden.

Kompetente Anwender und Modellierer kennen nicht nur die technischen Details der von ihnen eingesetzten Verfahren. Sie verfügen auch über die Fähigkeit, Standardmodelle und –verfahren nach ihren Bedürfnissen bzw. auf die Eigenschaften der von ihnen untersuchten Forschungsgegenstände hin zu adaptieren oder zu optimieren. Dadurch entstehen nicht nur aussagekräftigere Modelle, sondern auch eine engere Zusammenarbeit zwischen Informatikern und Forschern in den Digitalen Geisteswissenschaften. Ein wahrhaft partnerschaftliches Verhältnis entsteht dann, wenn beide als kreative und innovative Modellierer neuartige Modelle und Methoden für geisteswissenschaftliche Fragestellungen gemeinsam entwickeln. Ein solches Szenario stellt sicherlich den Idealfall für ein Digital

Computing dar und wird gleichzeitig zu neuen Bindestrich-Informatiken oder -- vielleicht noch wichtiger -- neuen Bindestrich-Geisteswissenschaften führen.

Literaturangaben:

[1] M. Beynon, S. Ross, und W. McCarty (2006). Human Computing – Modelling with Meaning, *Literary and Linguistic Computing* Vol. 21.2, pp. 141 – 157.

[2] W. McCarty (2004). Modelling: a study in words and meanings. In: Schreibman, S., Siemens, S. und Unsworth, J. (Hrsg.) *Companion to Digital Humanities*. Oxford: Blackwell. pp. 254 – 270.

[3] M. Minsky (1968). *Matter, Mind, and Models*. In: Minsky, M. (Hrsg.). *Semantic Information Processing*. Cambridge, Mass: MIT Press.

Herausforderung "Big Data" in der historischen Forschung

Kruse, Sebastian; Schmaltz, Florian; Stiller, Juliane; Wintergrün, Dirk

Max-Planck-Institut für Wissenschaftsgeschichte

Datenintegration und die Verfügbarmachung großer Textkorpora als Quellen für die zeitgeschichtliche Forschung, insbesondere in der Wissenschaftsgeschichte, stellen immer noch eine große Herausforderung dar. In unserem Beitrag stellen wir Methoden und Tools vor, die dieser Herausforderung begegnen und Lösungsansätze aufzeigen sollen. Vorgestellt wird dieses am Beispiel des auf 7 Jahre angelegten Forschungsprogramms zur Geschichte der Max-Planck-Gesellschaft (MPG), das im Juni 2014 begonnen wurde.[1] Ziel des Forschungsprogramms ist die Geschichte der Max-Planck-Gesellschaft von ihrer Gründung im Jahre 1948 bis zum Ende der Präsidentschaft von Hubert Markl 2002 aufzuarbeiten. Ziel ist es hierbei nicht eine additive Geschichte der 80 existierenden und 20 geschlossenen Institute der Max-Planck-Gesellschaft zu schreiben, sondern im Zentrum stehen institutsübergreifende Fragestellungen zu Themenfeldern wie Periodisierungen, Innovationen, Internationalisierung, Forschung und Wirtschaft, Gender und Wissenschaft sowie Konkurrenz und Kooperation.
Ein weiteres Ergebnis des Forschungsprogramms wird es sein, konzeptionelle und epistemologische Perspektiven aufzeigen, wie aus der elektronischen Quellen- und Datenüberlieferung ein digitales Forschungsarchiv generiert werden kann. Damit werden der MPG als Wissenschaftsorganisation neue selbstreflexive Erkenntnismöglichkeiten aus diesen digitalen Wissensspeichern eröffnet.

Für dieses Forschungsprogramm sollen große Aktenbestände und Publikationen, wie Tätigkeitsberichte und Jahrbücher der Max-Planck-Gesellschaft, digital erschlossen werden. Darüber hinaus sollen vorhandene digitale Quellen in einer virtuellen Forschungsumgebung integriert werden und kollaboratives Arbeiten unter den Historikern mit digitalen Tools ermöglicht und unterstützt werden.
Der Umfang der zu digitalisierenden Quellen mit unterschiedlichsten Provenienzen und zu grundliegenden Datenstrukturen erfordert technische Lösungen für Erfassung, Speicherung, Zugriff, Verwaltung und Analyse der Daten.

---

[1] Siehe http://www.mpiwg-berlin.mpg.de/en/research/projects/DEPT1_458_HistMPS

Die erste Herausforderung ist es, eine auf einem flexiblem Datenmodell basierende Infrastruktur zu schaffen, die sich im weiteren Projektverlauf an sich verändernde Anforderungen anpassen lässt und zugleich ermöglichen soll, Datenbestände unterschiedlicher Provenienz zu integrieren. Diese Datenbestände umfassen Normdaten, bibliographische und biographischen Datenbanken, eine Bestandsdatenbank relevanter Archive, sowie als zusätzliche Herausforderung "digital born" Daten verschiedener Disziplinen und Wissenschaftsbereichen der Institute der Max-Planck-Gesellschaft.

Die zweite Herausforderung ist die quantitative Dimension des Forschungsprogramms zur Geschichte der MPG. Es muss eine große Menge zur Zeit noch analog vorliegender, sehr heterogener Quellen digital verfügbar gemacht werden. Der Umfang der Quellen, in der Größenordnung von 4.500 laufenden Regalmetern Akten, macht es unmöglich, diese Quellen im Zeitraum der vorgegebenen Projektdauer mit traditionellen archivfachlichen Methoden zu erschließen. Aus zeit- und arbeitsökonomischen Gründen müssen daher digitale Methoden entwickelt werden, welche die Wissenschaftler bei der Quellenauswahl und -analyse unterstützen.

In unserem Paper werden wir den Ansatz zur Modellierung der Datenstrukturen genauer diskutieren, den Umgang mit den Datenmengen beschreiben und Tools der Digital Humanities vorstellen, die die Auswertung der Datenbestände erleichtern.

Bei der Modellierung der Personendatenbank muss mit widersprüchlichen Informationen und sich verändernden Werten umgegangen werden können. Dazu gehören beispielsweise sich im Zeitverlauf ändernde Namen und Aufenthaltsorte und institutionelle Zugehörigkeiten von Personen. Zugleich müssen strittige oder unsichere Informationen über die in der Datenbasis verwalteten Objekte dokumentiert werden können. Quelle und Urheber der Informationen und Einträge müssen wissenschaftlich nachvollziehbar bleiben. Schließlich sollen unterschiedlichen Versionen eines Eintrages verwaltet werden können.

Zur Umsetzung wurde ein RDF/OWL (Resource Description Framework/Web Ontology Language) Modell für die Daten entwickelt, das angelehnt an CRM (Conceptual Reference Model)[2] auf Grundlage einer ereignisbasierten Ontologie die unterschiedlichen Datenbestände integrieren soll. Hierbei benutzen wir OWL zunächst einmal pragmatisch als optimales Hilfsmittel zur Beschreibung der Daten und Datenstrukturen und sehen es nur als sekundär an, eine formale Konsistenz zu erreichen.

Dabei muss jedoch mit dem Problem umgegangen werden, dass ein Großteil der vorliegenden externen Daten, wie etwa die GND[3] zwar als RDF vorliegen,

---

[2] http://erlangen-crm.org/

[3] http://datahub.io/dataset/dnb-gemeinsame-normdatei

aber nicht kompatibel mit CRM sind. Zur Lösung dieses Problem wird ein hybrider Ansatz analog zu EDM (Europeana Data Model) verfolgt.

Zusätzlich muss der Kontext der Datenerstellung in der Datenbasis nachvollziehbar sein. Zu diesem Zwecke sollen Named Graphs eingesetzt werden, analog zu den Vorschlägen für eine Ontologie und eine API zur Verwaltung von Versionen und Provenienzen, wie sie von Kai Eckert[4] vorgeschlagen wurden. Die konkrete Implementierung erfolgt hierbei in einem Triple Store, das mit einer angepassten API und einem Frontend in Django realisiert wird.

Im zweiten Teil des Vortrages werden wir das Vorgehen zur Digitalisierung der für das Forschungsprogramm notwendigen Quellen näher darstellen. Der Umfang der Quellen, die historisch auszuwerten sind, macht es unmöglich diese in der klassischen Art und Weise durch „Close Reading“ zu sichten. Deshalb müssen Methoden gefunden werden, die mittels computer-gestützter Instrumente es ermöglichen, zu bestimmten historischen Fragestellungen eine Vorauswahl aus dem Gesamtkonvolut der digitalisierten Quellen zu treffen, bzw. Auswertungen mit Hilfe neuer Werkzeuge der Digital Humanities direkt vorzunehmen.

So sind von den einzelnen Akten im Archiv nur die Beschriftung der Aktendeckel bekannt. Mehr als 46.000 Aktenordnerrücken wurden dazu digitalisiert und in einer Datenbank erfasst. Diese Datenbasis dient zur Vorauswahl der jetzt komplett zu digitalisierenden Akten und ermöglicht eine grobe Zuordnung des Inhaltes zu einem Themenfeld.

Dazu sollen Methoden aus der Korpuslinguistik und statistische Modelle herangezogen werden um thematische Zuordnungen in großem Maßstab automatische zu ermöglichen. Um diese Methoden zu testen wurden in einem Vorprojekt mehr als 100.000 Seiten digitalisiert und der Text automatisch mit einer OCR-Software erschlossen. Die Ergebnisse werden in unserem Vortrag diskutiert, um aufzuzeigen, wie sie auf das Gesamtprojekt, das ein Umfang einer quantitativ sehr großen Datenmenge digitalisierter Akten bewältigen muss, angewandt werden können.

Das Vorprojekt hat gezeigt, dass trotz hoher Fehlerraten im jetzigen OCR-Verfahren mittels Named Entity Recognition erhebliche Fortschritte zur Erschließung der Akten und Jahrbücher erzielt werden konnten. Erste Ergebnisse mit Topic Modeling zeigen, dass dieses dazu beitragen kann, die Dokumente näher zu klassifizieren.

Eines der konkreten Ergebnisse der Digitalisierung der Jahrbücher der MPG ist, dass mit dem Aufbau einer möglichst umfassenden Bibliographie der MPG

---

[4] Kai Eckert, „Metadata Provenance in Europeana and the Semantic Web“, Masterarbeit 2012, http://edoc.hu-berlin.de/docviews/abstract.php?id=39630

seit 1949 begonnen werden konnte, die bisher nicht elektronisch vorliegt. Dazu werden die digitalisierten und mit Texterkennung bearbeiteten Jahrbücher mittels eines eigens für das Forschungsprogramm zur Geschichte der MPG entwickelten Annotationswerkzeuges ausgezeichnet. Die aus den erstellten Annotationen resultierenden Referenzen werden in eine Datenbank übertragen. Zurzeit basiert diese Annotationsmethode noch auf einem halb-automatischen Prozess. In unserem Vortrag werden wir über die Ergebnisse der künftig vorgesehenen automatischen Auswertung berichten.

Schließlich kommen unterschiedliche Methoden zur Netzwerk- und Zitationsanalyse in dem Forschungsprogramm zur Anwendung, um die Stellung der MPG innerhalb verschiedener wissenschaftlichen Forschungsgebiete zu analysieren.[5] Die dazu eingesetzten Tools werden wir in unserem Vortrag vorstellen.

Zur Visualisierung von geo-temporalen Abhängigkeiten, wie z.B die Neugründung, Verlagerung und Aufspaltung von Instituten oder die Flüsse von Fördermitteln, wird der Geo-Browser PLATIN[6] eingesetzt, der auf der Grundlage von Stefan Jähnickes GeoTemCo[7] in Zusammenarbeit mit DARIAH und dem Exzellenzcluster TOPOI weiterentwickelt werden konnte.

---

[5] Laubichler MD, Maienschein J, Renn J. 2013. Computational Perspectives in the History of Science. Isis. 104:119-130.

[6] https://github.com/skruse/PLATIN

[7] https://github.com/stjaenicke/GeoTemCo

Gespielte Geschichte: Digital Games als Medium und Fokus der Forschung

Gernot HAUSAR, Uni Wien/FH Campus Wien

Abstract

Spiele begleiten quer durch die Geschichte die Entwicklung von Zivilisationen. Neben der simplen Freude am Spiel selbst ist der Akt des Spielens eng mit der Entwicklung sozialer Bindungen, dem Erlernen grundlegender Fähigkeiten und Fertigkeiten sowie der Erziehung von Kindern im Rahmen der komplexen Zusammenhänge des jeweiligen zivilisatorischen Alltags verbunden. Für spätere Generationen bieten Spiele so einen wichtigen Einblick in diese Vorgänge.

Während dies in besonderer Weise auf digitale Spiele zutrifft, sind sie bisher kaum Fokus der Forschung. So fehlt es teilweise immer noch an gemeinsamen Methoden, Standards und multi-disziplinären Kooperationen. Auch kuratorische Probleme sind bisher noch nicht gelöst. So sollten neben der Soft- auch die Hardware, physische Beigaben und Spielkonsolen erhalten und der individuelle Spielprozess dokumentiert werden. Auch der Austausch zwischen der Spiel- und der realen Welt, die Meta-Diskussionen der Spieler, Fan Kunst und viele andere Aspekte sollten im Fokus von Kuratoren stehen.

So erweisen sich digitale Spiele, die neue Technologien bis an die Grenzen ausreizen, auch als Härtefall für die digitalen Geisteswissenschaften. Als solche bieten sie Forschern auch die Möglichkeit, viele wertvolle Erkenntnisse und Herangehensweisen für alle anderen Bereiche der digitalen Geisteswissenschaften zu destillieren.

Dieser Vortrag bietet einen facettenhaften Überblick über aktuelle kuratorische Bemühungen, die laufenden Diskussionen zu Methoden und Kooperationen mit anderen Fachgebieten außerhalb der Geschichtswissenschaften sowie zu Themen der historischen Game Studies. Es fließen hier auch die Ergebnisse der Fachkonferenz zu History and Games ein, die 2013 in Düsseldorf stattfand.

Weiters werden Stolpersteine wie der rechtliche Rahmen und die spezifischen Herausforderungen für Archiveinrichtungen behandelt. Der Vortrag schließt – so es der zeitliche Rahmen erlaubt – mit einer kurzen Game-Demo, die von Teilnehmern in der Pause auch persönlich erlebt werden kann.

Zum Vortragenden

Gernot Hausar ist (digitaler) Historiker. Er beschäftigt sich u.a. mit Fragen der Information Studies, der Digital Humanities sowie Game und Media Studies. Weitere Interessen inkludieren historische Digitalisierung und OCR, Visualisierung und digitale Gemeinschaften (z.B. Hacker). Er ist Medida Prix Preisträger und in seiner Freizeit Mitarbeiter von Creative Commons Österreich.

Publikationen und Vorträge zum Thema

- Buchbeitrag: Der Stadt ihre Spieler. Wahrnehmung und Wirkung historischer Metropolen in der Assassin's Creed Serie. Transcript 2014.
- Buchbeitrag: Players in the Digital City. Historical Metropoleis in the Assassin's Creed Series. Cambridge Scholars 2014.
- Vortrag: Discarded Toys – Excavating, Documenting and Reviving Abandoned Digital Games in Databases and Platforms. Conference on Cultural Heritage and New Technologies, Vienna 2013.
- Fachartikel: Gespielte Geschichte – Die Bedeutung von Lore im Massive Multiplayer Spiel EVE Online. VGS 2013.
- Fachartikel: Sid Meier als Geschichtsphilosoph? Die Strategiespiele der Civilisation-Serie als Herausforderung für die Geschichtswissenschaft. VGS 2013

Über mehrsprachige Metadaten zu den versteckten Quellen: Das Beispiel deutschsprachiger Täterquellen in der European Holocaust Research Infrastructure

Dr. Veerle Vanden Daelen, Centre for Historical Research and Documentation on War and Contemporary Society (Cegesoma, Brüssel) & Giles Bennett M.A., Institut für Zeitgeschichte (IfZ, München)

Die European Holocaust Research Infrastructure (Europäische Holocaust Forschungsinfrastruktur, abgek. EHRI) ist ein EU-gefördertes FP7 Projekt, dessen Hauptziel es ist, Holocaustforscherinnen und Holocaustforscher durch die Schaffung einer virtuellen Forschungsumgebung zu unterstützen, die online Zugang zu verstreuten Holocaustbezogenen Quellen bietet sowie kollaborative Forschungsprojekte durch die Entwicklung von Tools unterstützt. Im EHRI Portal werden Beschreibungen von Repositorien (Archive, Museen, Gedenkstätten, usw.) sowie Archivbeständen, die für die Holocaustforschung relevant sind, zusammengebracht. Mehrsprachige Suchtools und ein zehnsprachiger Thesaurus eröffnen mehr als 1.800 Repositorienbeschreibungen mit tausenden von Bestandsbeschreibungen. Gemäß unserer Prämisse „von Daten zu Erkenntnissen" kann EHRI der Holocaustforschung neue Forschungsergebnisse und -zugänge eröffnen, indem Metadaten (also hier: Informationen über Quellen) in einer mehrsprachigen Forschungsumgebung zur Verfügung gestellt werden, in der Informationen unterschiedlicher Herkunft zusammengebracht und verknüpft werden.

Wir wollen dies anhand deutschsprachiger Quellen und den durch EHRI neu denkbar und möglich gewordenen Forschungszugängen illustrieren. Die in der Hauptsprache der Täter, der deutschen Sprache, verfassten Quellen über den Holocaust sind über alle vom Zweiten Weltkrieg berührten Länder (und darüber hinaus) verstreut. Obwohl sich in fast allen von EHRI abgedeckten Repositorien deutschsprachige Quellen befinden, ist der Zugang zu diesen Dokumenten für deutschsprachige (bzw. an deutschen Quellen interessierte) Forscherinnen und Forscher hinter einer Vielzahl von anderen Sprachen versteckt. Da deutsche Quellen auf Estnisch, Lettisch, Litauisch, Kroatisch, Hebräisch, Griechisch, Niederländisch, Französisch, Russisch, Polnisch oder Ukrainisch katalogisiert wurden – um nur einige wenige Sprachen und die dazugehörigen Schriftsysteme (Lateinisch, Kyrillisch, Hebräisch, Griechisch) zu nennen – ist die Identifizierung von deutschsprachigen Quellen nicht selbstverständlich. Außerdem ist es eine Herausforderung für die Archivarinnen und Archivare vor Ort, mit oftmals begrenzten Deutschkenntnissen und wenig ausgeprägtem Wissen über den Holocaust die deutschen Quellen adäquat zu beschreiben.

Die EHRI-Datenbank verbindet Daten aus den verschiedenen Sprachen und schreibt den Datensätzen Thesaurusbegriffe und englischsprachige Normdateieinträge (z.B. Institutionen) zu. Darüber hinaus ist das Projekt bemüht, mehrere Beschreibungen für einen Bestand zur Verfügung zu stellen, sowie Original- und Kopienbestände über ihre jeweiligen Beschreibungen zu verbinden. Verschiedene Beschreibungen für denselben Bestand (oftmals, aber nicht zwangsläufig in verschiedenen Sprachen) bieten oft unterschiedliche Zugänge zum gleichen Material. Eine bedeutende, aus dieser Vorgangsweise resultierende Schlussfolgerung ist, dass das EHRI-Projekt und sein Portal das Bewusstsein über Subjektivität in den Forscherinnen und Forschern zur Verfügung stehenden Findmitteln und Archivführern steigert. So beschreibt das EHRI-Projekt nicht nur deutschsprachige Holocaustquellen (z.B.) in baltischen oder ukrainischen Archiven, es zeigt auch, dass in einigen Ländern oder Repositorien Schlagworte wie „Holocaust" oder „Judenverfolgung" nicht in den Metadaten vorkommen, auch, weil z.B. die Bestände umfassend Fragen der deutschen Besatzung berühren. In diesem Kontext können die Quellen nur „aufgedeckt" werden, indem die Beschreibung der Einrichtung selbst mit Findmitteln aus dem Kontext der Holocaustforschung oder der jüdischen Geschichte verbunden wird, oder indem die Beschreibung der Einrichtung, in der sich die Originale befinden, mit den Beschreibungen der

Kopienbestände in Institutionen wie Yad Vashem oder dem United States Holocaust Memorial Museum verbunden werden (um nur die zwei größten Repositorien mit einer großen Zahl an kopierten Archivalien mit Holocaustbezug zu nennen). Dabei können Forscherinnen und Forscher, die sich für Bestände zu einem bestimmten Land interessieren, auch unerwartete Funde in dritten Ländern mit Metadaten in anderen Sprachen machen (z.B. deutschsprachige Quellen zur Judenverfolgung in Frankreich mit polnischen Metadaten).

Wir wollen daher in unserem Beitrag zeigen, dass die digitale Welt nicht nur Werkzeuge für den Umgang mit Daten zur Verfügung stellt, die die Forschung zu neue Methodologien führen, sondern dass die digital verfügbaren Metadaten auch Möglichkeiten für neue Erkenntnisse eröffnen. Indem Informationen über deutschsprachige Bestände in den von EHRI abgedeckten Ländern zugänglich gemacht werden, kann das Projekt unser Wissen vergrößern und Interpretationen für deutschsprachige Forscherinnen und Forscher ermöglichen, die vor der Einrichtung von Plattformen wie EHRI nicht vorstellbar waren.

Quanticod revisited.
Neue Ansätze zur quantitativen Analyse mittelalterlicher Handschriftenbestände

Hannah Busch (Universität Trier)
Swati Chandna (Karlsruher Institut für Technologie)
Celia Krause (Technische Universität Darmstadt)
Philipp Vanscheidt (Universität Trier / Technische Universität Darmstadt)

Die Kodikologie hat sich als Hilfsdisziplin der historischen Wissenschaften seit den fünfziger Jahren etabliert. Ihr Interesse gilt in erster Linie der Beschreibung und dem Studium der äußeren Merkmale des Kodex, darunter fällt neben den materiellen Bestandteilen auch das Studium des sogenannten *Mise-en-page* – des Layouts der handgeschriebenen Seite, der Größe und des Umfangs der gebundenen Handschrift.

Ziel von kodikologischen Studien ist es, mehr über die Arbeitstechniken und den Ablauf der mittelalterlichen Buchherstellung zu erfahren sowie regionale und soziokulturelle Normierungstendenzen im Hinblick auf ihre Existenz, räumliche Verbreitung und zeitliche Überdauerung zu erforschen. Ein besonderes Augenmerk gilt dabei dem Verhältnis zwischen der zur Verfügung stehenden Oberfläche einer Seite und der Fläche, die durch Primärtext, Bilder und Marginalia eingenommen wird.

Relativ schnell kristallisiert sich heraus, dass sich solche kodikologischen Fragestellungen weniger gut an einzelnen Objekten überprüfen lassen, es vielmehr quantitativer Verfahren bedarf. Dafür wird jedoch eine kritische Masse an Kodizes benötigt, die mindestens über eine gemeinsame Eigenschaft verfügen. Diese Eigenschaft kann sich in Form der Provenienz, wie der Zugehörigkeit zu einer mittelalterlichen Bibliothek oder einem Skriptorium als zugeschriebenem Entstehungsort, der Textgattung, der Sprache oder eines bestimmten Layouts manifestieren.

Quantitative Kodikologie

Bereits seit den siebziger Jahren verfolgte eine Gruppe von italienischen und französischen Wissenschaftlern[1] der noch relativ jungen Disziplin der Kodikologie diesen Ansatz einer quantitativen Forschungsmethode an mittelalterlichen Handschriften. Durch manuell erhobene Zähl- und Messergebnisse und deren statistische Auswertung konnten so Entwicklungslinien und -tendenzen auch graphisch dargestellt werden. Auf diese Weise lassen sich beispielsweise Aussagen über die Zeichendichte auf einer ein- oder zweispaltig aufgeteilten Handschriftenseite oder über die Bedeutung von Marginalflächen treffen und

1 Als Mitglieder der Gruppe Quanticod seien hier namentlich zu erwähnen: Carla Bozzolo, Ezio Ornato, Denis Muzerelle, Dominique Coq.

räumlich und zeitlich bedingte Layouttendenzen feststellen. Durch geometrische Berechnungen kann erforscht werden, ob sich das Seitenverhältnis an Normen, wie dem aus der Malerei bekannten Goldenen Schnitt, orientiert. Durch den Aufwand der händischen Datenerhebung und die Verteilung der Handschriften auf viele verschiedene Standorte waren die Untersuchungen aber nur auf bereits erschlossene und zum Teil vermessene Handschriftenbestände und damit sehr limitierte Korpora anwendbar.

Der dargelegte Ansatz einer quantitativen Kodikologie scheint seit den frühen neunziger Jahren nicht intensiver verfolgt worden zu sein, obwohl das Potential computergestützter Analysemöglichkeiten seitdem stetig gewachsen ist. Zwei technische Neuerungen haben nämlich inzwischen den Umgang mit historischen Buchbeständen in Universitäten, Bibliotheken und Archiven nachhaltig verändert: die Entwicklung von handlichen Geräten für die Bilddatenerfassung sowie die Einrichtung des World Wide Web. Im Internet erschließen sich Publikationswege, die den Vorteil haben, dass Inhalte für einen breiten Nutzerkreis zur Verfügung gestellt werden können. Die zunehmende Verfeinerung digitaler Reproduktionsmöglichkeiten ermöglicht die Digitalisierung alter Buchbestände mittels Scannern und Digitalkameras und deren allgemeine Freigabe im Internet. Im Zuge großer Digitalisierungsprojekte sind mittlerweile viele mittelalterliche Handschriften in Archiven und Bibliotheken ins digitale Medium überführt worden. Schon lange werden dabei nicht mehr nur berühmte und reich ausgeschmückte Handschriften der breiten Öffentlichkeit zur Verfügung gestellt. Zahlreiche Projekte erfassen inzwischen den gesamten Bestand von Bibliotheken und Skriptorien. In Folge dessen geraten nun auch Handschriften ins Blickfeld, die bisher nicht mehrfach und detailliert katalogisiert und beschrieben wurden.

eCodicology

Genau an diesem Punkt setzt das Projekt „eCodicology“ (http://www.ecodicology.org) an. Als Korpus dient zunächst der digital verfügbare Bestand des rekonstruierten „Virtuellen Skriptorium St. Matthias“ (http://www.stmatthias.uni-trier.de). Dabei handelt es sich um circa 450 Handschriften aus dem mittelalterlichen Bestand der Benediktinerabtei St. Matthias in Trier aus dem 8. – 18. Jahrhundert. Insgesamt liegen etwa 170.000 Seiten als Bilddateien vor, die ein Datenvolumen von über 5 TB umfassen.

Durch die gezielte Vorverarbeitung der Bilddateien können deutlich größere Datenmengen bewältigt werden, ohne dabei einen Qualitätsverlust zu erleiden. Dies wird ermöglicht, indem eine Farb- und Größenkalibrierung, eine Skalierung und eine Reduzierung eventuellen Bildrauschens durchgeführt werden. Erst dann findet der Vorgang der Merkmalsextraktion statt, bei dem die einzelnen Layoutmerkmale der handgeschriebenen Seite erkannt, isoliert und vermessen werden. Zu diesen Merkmalen zählen die Seitengröße, Schrifträume, Bildräume, freie Flächen und die Anzahl der Zeilen. Neben der Anzahl der jeweiligen Elemente und ihrer Ausdehnung soll auch ihre exakte Position auf der Seite festgehalten werden. Die dadurch neu gewonnen Metadaten ergänzen die Beschreibungen aus früheren Katalogen und werden in XML-Dateien nach TEI P5 gespeichert.

Fallstudien „Die Beobachtung des Banalen“

Nachdem zunächst die Vorverarbeitung der Bilddateien im Vordergrund stand, können nun die ersten vollständig automatischen Vermessungen des Seitenlayouts präsentiert und statistisch ausgewertet werden. Anhand eines Abgleiches mit manuellen Messungen, die mit Lineal und Geodreieck an einigen (analogen) Originalhandschriften durchgeführt wurden sowie vorhandenen Informationen über die Maße aus Handschriftenkatalogen, werden wir damit erste Ergebnisse zur Genauigkeit der Algorithmen für die automatische Merkmalsextraktion vorstellen können. Hierbei wird es um die Frage gehen, mit welchen Herausforderungen der quantitativ arbeitende Kodikologe bei der Vermessung konfrontiert wird. Hierzu gehört etwa die Einschätzung des Einflusses der mit der Bindung einhergehenden Wölbung der Seite auf die Messergebnisse und der daraus resultierenden Vor- und Nachteile von semi- und vollautomatischen Messungen durch den Computer. Wie wirken sich die Erkenntnisse auf die Ansprüche und die bisherigen Konventionen der Digitalisierungsverfahren aus? Der Vortrag widmet sich darüber hinaus der Frage, welche Möglichkeiten der Analyse von digitalen Sammlungen uns durch eine digitale Kodikologie eröffnet werden. Außerdem soll diskutiert werden, welche neuen Fragestellungen sich damit aufwerfen lassen und wie sich die Ergebnisse aufbereiten und visualisieren lassen.

Neben der Ergänzung und Auswertung der neu gewonnenen Bildmetadaten bietet sich die statistische Auswertung auch für die bereits in früheren Katalogen erfassten bibliografischen Metadaten an. Dies ist eine Methode, die unter dem Begriff der *Bibliometrie* in der Buchwissenschaft bereits eine längere Tradition hat. Im Zuge der Erstellung virtueller Handschriftenbibliotheken wurden die Metadaten nicht nur in klassischen Datenbankstrukturen veröffentlicht. Durch Initiativen wie der TEI (Text Encoding Initiative) ist ein XML-Schema entwickelt worden, das sich auch speziell an die Erfassung der Metadaten von Handschriften richtet. Dieses Schema findet auch im Projekt „eCodicology“ in adaptierter Form Verwendung.

Zunächst erschien es sinnvoll, die alten Daten aus den Handschriftenkatalogen und die neugewonnen Daten zu den Maßen der Layoutmerkmale gemeinsam statistisch auszuwerten, also zu den Wurzeln der quantitativen Kodikologie zurückzukehren. Darüber hinaus wollen wir uns in einer kleinen Fallstudie die gemeinsame Art der Aufbereitung der Daten zunutze machen, um den Bestand der Abtei St. Matthias mit einem weiteren, in XML aufbereiteten Handschriftenbestand zu vergleichen: neben einem quantitativen wird somit auch ein komparatistischer Ansatz verfolgt. In einem ersten Schritt sollen dabei Katalogdaten, wie Blatt-/Seitenzahl und Größe der Kodizes anderer Bestände mit dem Skriptorium von St. Matthias verglichen werden. So ließen sich Aussagen über regionale Tendenzen treffen und beispielsweise Vergleiche zum norddeutschen Raum oder dem benachbarten romanischen Raum ziehen.

Die neuen Möglichkeiten einer digital betriebenen Handschriftenkunde können um ein weiteres kodikologisches Werkzeug bereichert werden, das es erlaubt, einzelne Exemplare aus einem Handschriftenbestand oder größere Handschriftengruppen im wahrsten Sinne des Wortes aus der Vogelperspektive zu betrachten. Hierbei handelt es sich um die Erstellung von Bildmontagen nach der Methode von Lev Manovich. Ein Kodex wird virtuell entblättert und alle Seiten der Reihe nach rasterförmig neben- bzw. untereinander angeordnet. Eine solche Bildmontage erlaubt das Aufspüren von Gestaltungsmustern innerhalb repräsentativer Handschriften.

Die Kombination aller vorgestellten Verfahren gestattet dem Kodikologen einen neuartigen Zugang zu digital erschlossenen Handschriftenbeständen. Durch die Untersuchung scheinbar banaler Merkmale lassen sich Forschungsfragen beantworten, auf die allein in Detailstudien kaum eingegangen werden könnte.

Von Daten zu Erkenntnissen
Zweite Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)
25. – 27. Februar 2015, Graz

Zur visuellen Analyse und Synthese historischer Daten in Raum und Zeit

Florian Windhager
Department für Wissens- und Kommunikationsmanagement
Donau-Universität Krems | florian.windhager@donau-uni.ac.at

Methoden der Informationsvisualisierung generieren grafische Repräsentationen zur Unterstützung von Prozessen der Kognition und Kommunikation. Ob als einfache Diagramme oder interaktive Interfaces – Techniken der visuellen Repräsentationen zielen auf die primäre Synthese von Information, sowie auf deren vertiefende visuelle Analyse und ihre Vermittlung. Wenn in diesem Kontext insbesondere „umfassende, dynamische, mehrwertige und oftmals widersprüchliche Datenmengen" zum Bezugsproblem erhoben werden (Thomas & Cook, 2005, S.4), so ist dies ein Kriterium das in den Themenfeldern der Kultur- und Geisteswissenschaften immer schon als ausreichend erfüllt betrachtet werden darf. Erst mit dem Einzug von digitalen Methoden und Technologien emergierte in selbigen Feldern aber die Möglichkeit, ihre schriftbasierten Wissensbestände und Diskurse als komplexe Konstellationen von „Daten" zu rekontextuieren - und unter anderen den Methoden der visuellen Synthese und Analyse zuzuführen. Vor diesem Hintergrund trägt die Erschließung von grafischen Repräsentationen und Verfahren seit geraumer Zeit zur Diversifizierung des Spektrums an Vermittlungs- und Forschungsmethoden bei. Von Standardmethoden der statistischen Datenvisualisierung zur kartografischen Lokalisierung im physischen Raum, von genealogischen Dendrogrammen zu heterarchischen Netzwerktopologien, ob statisch-aggregiert oder dynamisch und interaktiv – der Mehrwert digitaler Methoden erschließt sich nicht zuletzt in einem neuen Spektrum von Bildern, die den multimodalen Aufbau und die Vertiefung von geisteswissenschaftlichem Wissen in Lehre und Forschung unterstützen (Jessop, 2008; Elson, Dames, & McKeown, 2010; Moretti, 2013; Staley 2013).

Während der aktuelle Beitrag die Existenz der entsprechenden Verfahren sowie ihre zunehmende fachliche Reflexion und Integration voraussetzen kann, richtet sich sein Fokus auf eine Folgewirkung der Einführung von zahlreichen eigenlogischen Verfahren auf einer metamethodischen Ebene. Aus dieser synoptischen Perspektive treten neben dem Mehrwert der einzelnen Visualisierungsmethoden auch eine Reihe von ungenützten Synergien hervor, sowie der gelegentliche Mangel an Reflexion von konzeptuellen Designentscheidungen, der sich bis zum konstanten Konflikt auf der Ebene visueller Repräsentationen und Schemata fortsetzen kann. Probleme der Interpretation ergeben sich hierbei durch multiple möglichen Layouts desselben Datensatzes (z.B. als geo-basierte, relationale, hierarchische oder temporale Visualisierungen), durch unterschiedliche Techniken der Visualisierung von Dynamik, sowie durch die nur in Ausnahmefällen mögliche Rückbettung und Verlinkung von mehreren Visualisierungen in einen umfassenden visuellen Bezugsrahmen. Während solche ungenützten Potentiale der Interoperabilität im Rahmen lokaler Applikationen vernachlässigt werden können, ist ihre Erschließung und Ausschöpfung unverzichtbar wenn es um Systeme mit erhöhter NutzerInnenfreundlichkeit, kohärenter visueller Syntax, sowie um kollaborative Systemen der Wissensrepräsentation geht. Unter dieser Perspektive stellt der Beitrag ein methodisch-theoretisches Konzept zur Diskussion, das die synoptische Verknüpfung einer Mehrzahl von Visualisierungtechniken unter Erhalt ihrer jeweiligen Eigenlogiken erlaubt.

Als konzeptueller Rahmen dient hierzu der chronogeographische Ansatz der Lund-School, der sogenannte Raum-Zeit-Kuben entfaltete und zur visuellen Analyse von Innovationsdynamik in der geographischen Raumzeit nutzbar machte (Parkes & Thrift, 1980). Während dieses Verfahren seit geraumer Zeit auch in Tools zur visuellen Analyse implementiert ist (Kapler & Wright, 2004), werden seine umfassenden Möglichkeiten zur konzeptuellen Integration diverser dynamischer Darstellungsformen erst aktuell neu erschlossen (Bach, Dragicevic, Archambault, Hurter, & Carpendale, 2014).

Um zu zeigen wie dieses geo-basierte Verfahren in instruktiver Form zu Formen der nicht-raum-basierten Informationsvisualisierung in Wechselwirkung treten kann, illustriert der Beitrag die mögliche Verknüpfung von geographischen Karten zu statistischen Diagrammen, Clusterdarstellungen, sowie zu sozialen und semantischen Netzwerkgraphen in einem frei skalierbaren kohärent-dynamischen Rahmenwerk. Auf diese Weise können verschiedenartigste Dynamiken von historisch oder geisteswissenschaftlicher Relevanz in ihrer Co-Evolution in semantischen, sozialen und geographischen Bezugsräumen parallel sichtbar gemacht werden (Windhager, 2013) (vgl. Abbildung 1). Durch die resultierende Verlinkung und Verschränkung mehrerer Perspektiven werden räumliche und zeitliche Mikro- und Makrodimensionen ebenso vermittelbar wie diverse Granularitäten der visuellen Analyse. Ein Ausblick dient der geplanten Entwicklung von kollaborativen Projekten und Tools, wobei ein nächstliegendes Anwendungsszenario in der visuellen Exploration von physischen und digitalen Sammlungen des kulturellen Erbes zu finden ist (Windhager & Mayr, 2012).

Referenzen:

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., & Carpendale, S. (2014). A Review of Temporal Data Visualizations Based on Space-Time Cube Operations. In *EuroVis-STARs* (S. 23–41). The Eurographics Association.

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* (S. 138–147). Association for Computational Linguistics.

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing, 23*(3), 281–293.

Kapler, T., & Wright, W. (2004). GeoTime information visualization. In *INFOVIS '04 Proceedings of the IEEE Symposium on Information Visualization* (S. 25–32).

Moretti, F. (2013). *Distant Reading*. London: Verso.

Parkes, D., & Thrift, N. (1980). *Times, Spaces and Places: A Chronogeographic Perspective*. Chichester: John Wiley & Sons Ltd.

Staley, D. J. (2013). *Computers, Visualization, and History: How New Technology Will Transform Our Understanding of the Past.* (2nd edition). Armonk: M.E. Sharpe.

Thomas, J., & Cook, K. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.

Windhager, F., & Mayr, E. (2012). Cultural Heritage Cube. A Conceptual Framework for Visual Exhibition Exploration. In *2012 16th International Conference on Information Visualisation (IV)* (S. 540 –545).

Windhager, F. (2013). On Polycubism. Outlining a Dynamic Information Visualization Framework for the Humanities and Social Sciences. In M. Fuellsack (Ed.) *Networking Networks. Origins, Applications, Experiments.* Wien: Turia + Kant.

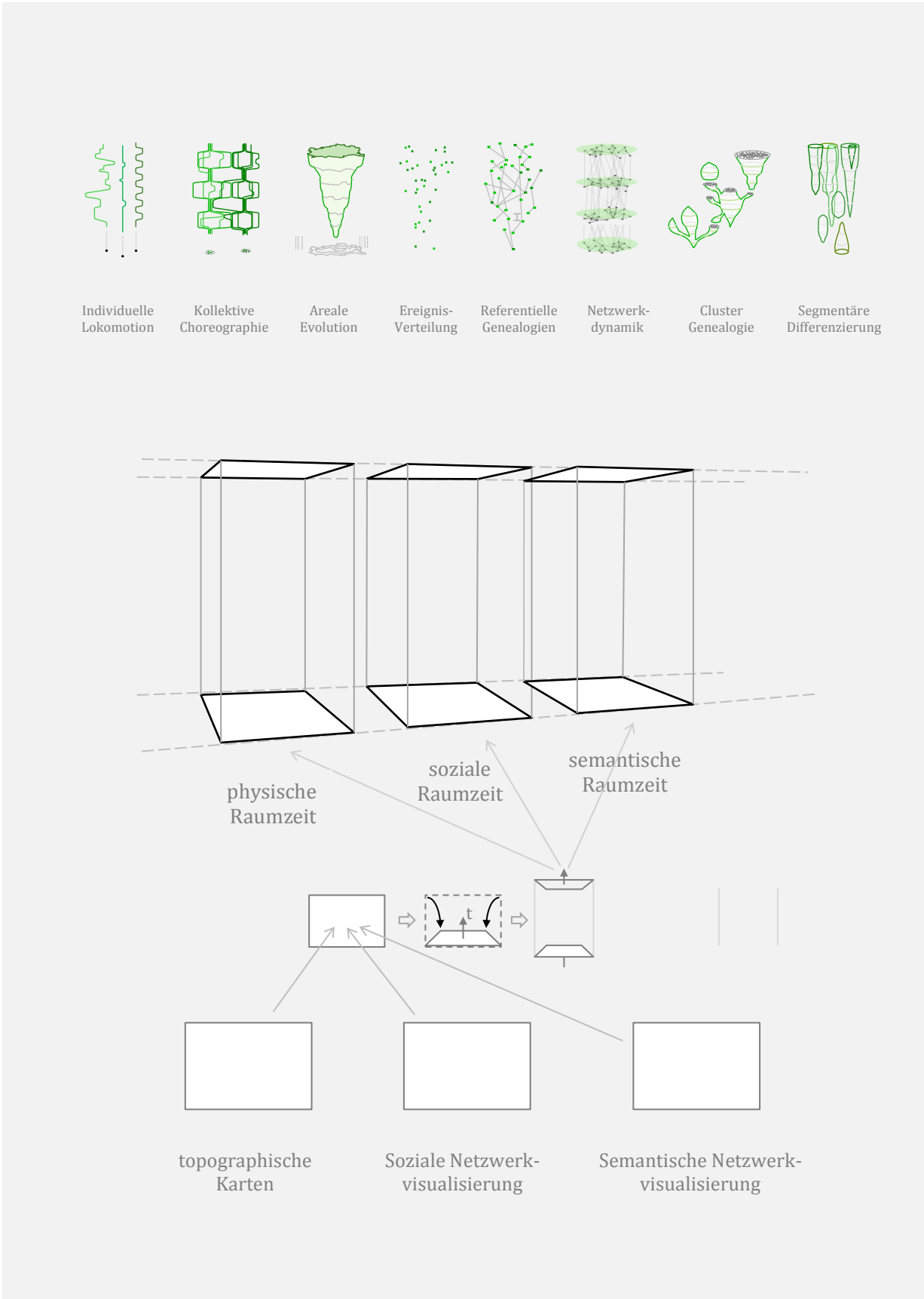


Abbildung 1: Etablierte zweidimensionale Visualisierungsmethoden (unten), kohärente Dynamisierung im visuell-analytischen Rahmen paralleler Space-Time-Cubes (Mitte) und vergleichende Illustration von verschiedenartigen Phänomenen der Translation, Strukturgenetik und Co-Evolution (oben).

Bernadette Biedermann und Nikolaus Reisinger

Abstract

Vom Repositorium zum virtuellen Museum: UserInnen versus BesucherInnen im Spannungsfeld zwischen Erkenntniswunsch und Selbst(er)findung

Im Rahmen des Vortrags sollen sowohl die theoretische wie auch die praktische Relevanz der digitalen Geisteswissenschaften für die auf den Bedeutungswert von kulturellem Erbe konzentrierten Disziplin Museologie untersucht werden. In diesem Zusammenhang soll der Frage nach Möglichkeiten und Herausforderungen eines „virtuelles Museum“ nachgegangen werden, das sich im museologischen Kontext im Spannungsfeld zwischen der Aura der authentischen Objekts und seiner digitalen Abbildung befindet.

Bereits seit dem Zeitalter der Reproduktionsmöglichkeiten durch Fotografie und Film stellt sich die Frage nach der Authentizität des originalen Objekts. Demnach würden Digitalisate zum Verlust der nur durch den direkten Kontakt mit dem Ding an sich zu vermittelnden Objektaura führen. Aus Sicht der klassischen Museologie schließt dies auf den ersten Blick die Präsentation eines musealen Sammlungsbestandes auf virtuelle Weise aus und rechtfertigt lediglich die Abbildung von Informationen zu Konservierungszwecken.

Dementsprechend soll den Fragen nach dem Mehrwert und den Problemfeldern der digitalen Geisteswissenschaften auf dem Weg zum „virtuellen Museum“ („digitales Museum“, „on-line museum“, „electronic museum“, „hypermuseum“, „cybermuseum“) am Beispiel des Projekts „Repositorium steirisches Wissenschaftserbe“ und insbesondere des Bereichs der Universitätsmuseen der Karl-Franzens-Universität Graz nachgegangen und dabei mögliche interdisziplinäre Methoden zur Generierung von Erkenntnissen untersucht werden.

Nach einer theoretischen Einleitung über das System der Museologie, ihren Erkenntnisgegenstand und den Musealisierungsprozess sowohl als Alltagsphänomen wie auch als Theorem der Museologie wird auf den Erkenntnis- sowie den Selbst(er)findungsprozess eingegangen, der durch die digitale Erfassung (Dokumentation) und Darstellung (Präsentation) musealer Bestände für BesucherInnen und UserInnen geleistet werden kann.

Manfred Thaller, Universität zu Köln

Panel: Digital Humanities als Beruf – Fortschritte auf dem Weg zu einem Curriculum

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. Nach wie vor ist die Situation der deutschsprachigen Länder ungewöhnlich dadurch, dass die Anzahl der hier als durchstrukturierte Studiengänge angebotenen Abschlüsse – zum Unterschied von kursartig angebotenen Zusatzqualifikationen - deutlich über denen anderer Länder liegt, was nicht zuletzt auch auf der Digital Humanities 2014 in Lausanne sehr deutlich wurde. Schon auf der ersten Jahreskonferenz der Digital Humanities der deutschsprachigen Länder im März 2014 in Passau wurde deshalb eine Arbeitsgruppe der DHd gegründet, die, aufbauend auf dem Ergebnis eines seit 2009 laufenden Prozesses zur Mitarbeit bei weiteren curricularen Planungen, einlud, mit dem Ziel eines „Referenzcurriculums". Damals wurde festgehalten, dass es bereits eine lange zurückreichende Tradition der Beschäftigung mit curricularen Vorstellungen in unterschiedlichen Ausprägungen der Digital Humanities gegeben habe, die aber eben über die Auflistung und die Feststellung dass an unterschiedlichen Hochschulen Unterschiedliches unterschiedlich unterrichtet würde nie hinausgekommen sind.

Die Proponenten der Arbeitsgruppe schlugen daher vor eine gezielte Anstrengung zu unternehmen, um über das Stadium „Digital Humanities Kurse unterrichten, was die an den jeweiligen Universitäten Digital Humanities Unterrichtenden unterrichten" hinaus zu kommen und von persönlichen Forschungsrichtungen und lokalen Gegebenheiten zu abstrahieren. Dass dies nicht einfacher sei, als einer der aktuellen Versuche, die Digital Humanities additiv als solche zu definieren war unstrittig. Der Aufwand sei aber notwendig, aus pragmatischen Gründen:

- Je größer die Zahl einschlägiger Studiengänge wird, desto schwieriger ist es zu vermitteln, warum der Übergang von einem zum anderen Probleme bereiten sollte. Die wechselseitige Anerkennung von Studienleistungen wird erheblich vereinfacht, wenn sie studiengangsunabhängig definiert sind.

- Die Akkreditierung von Studiengängen wird umso einfacher, je einfacher es ist, sich bei dem Studiengang auf einverständlich über einzelne Institutionen hinaus definierte Referenzwerte zu beziehen.

- Definieren die DH Studiengänge ihre eigenen Orientierungspunkte nicht selbst, ist durchaus zu erwarten, dass andere versuchen, dies für sie zu tun.

Dabei konnte – und durfte – es nicht darum gehen, in einem sich nach wie vor sehr dynamisch weiter entwickelnden Bereich verbindliche Details, etwa im Sinne einer verpflichtenden Studienordnung, festzuschreiben: Der Begriff eines „Referenzcurriculums" versteht sich bewusst im Sinne einer Referenzarchitektur, nach dem Gebrauch des Begriffs in der Softwaretechnologie. Es soll also einerseits ein Modell beschreiben, mit dem einzelne konkrete Curricula verglichen werden können, andererseits ein Vokabular definieren, mit dessen Hilfe Umsetzungen möglichst präzise definiert werden können.

Zu diesem Zweck wurde am 2. Oktober 2014 in Köln ein erster Workshop eingeladen, bei dem die Teilnehmerinnen und Teilnehmer gezielt so ausgewählt worden waren, dass möglichst unterschiedliche disziplinäre Hintergründe vorlagen. Das Schwergewicht lag dabei auf Einrichtungen, bei denen schon Erfahrungen mit der Umsetzung von Studiengängen bestehen, es wurden aber auch gezielt Vertreterinnen und Vertreter von Einrichtungen eingeladen, die in fortgeschrittenen Stadien der Planung von Studiengängen eingebunden sind. Dabei ging es bewusst nicht um die Diskussion sich aus lokalen institutionellen Randbedingungen ergebende Sachzwänge, sondern um die abstrakte Definition curricularer Anforderungen. Daran beteiligt waren (* = Mitglied der Arbeitsgruppe „Curricula" der DHd): Tara Andrews, Bern; Sabine Bartsch*, Darmstadt; Stephan Büttner, Potsdam; Andreas Henrich*, Bamberg; Matthias Lang*, Tübingen; Andy Lücking, Frankfurt / Main; Patrick Sahle*, Köln; Walter Scholger*, Graz; Caroline Sporleder, Trier; Heidrun Stein-Kecks*, Erlangen; Manfred Thaller*, Köln; Gabor Mihaly Toth, Passau; Thorsten Vitt, Würzburg.

Aus den dortigen Diskussionen entstehen derzeit gerade Unterlagen, die die existierenden Studiengänge besser vergleichbar machen sollen und einem größeren Kreis von curricular Interessierten im Laufe des November vorgelegt werden. In einem weiteren Workshop im Januar wird schließlich ein Dokument redigiert, das einen Entwurf für ein Referenzcurriculum mit umfangreichen Hintergrundüberlegungen zu den unterschiedlichen

Studiengängen verbinden wird, gleichzeitig aber auch einen ergänzten und auf Vergleichbarkeit angelegten Katalog bestehender Studiengänge enthalten wird.

Der erreichte Stand dieser Überlegungen wird in Graz präsentiert werden und eine Gruppe der an seiner Vorbereitung beteiligten Kolleginnen und Kollegen wird in persönlichen Statements einzelne Positionen dazu vertreten und diskutieren, bevor die Diskussion für das Publikum geöffnet wird. Wir gehen von einem Zeitverhältnis Präsentation : Paneldiskussion : Publikumsdiskussion von 2 : 1 : 2 aus.

Der oben definierte Begriff eines „Referenzcurriculums“ macht deutlich, dass jeder derartige Versuch zwischen zwei Gefahren steht: Die Vorgaben müssen konkret genug sein, um keine totale Beliebigkeit zuzulassen; sie müssen aber auch flexibel genug sein, um auf real existierende Studiengänge und die Bedingungen für deren Einbindung in die Fakultäten anwendbar zu sein. Ob dies vom dann erreichten Stand des Referenzcurriculums erreicht wird, wird bei Präsentation und Diskussion im Zentrum stehen.

Die TeilnehmerInnen am Panel sind noch nicht abschließend bestimmt. Sie werden in den noch ausstehenden Stufen des beschriebenen Arbeitsprozesses aus den oben angeführten TeilnehmerInnen am Workshop am 2. Oktober in Köln ausgewählt.

Was sind und was sollen Datenzentren in den Geisteswissenschaften?

Panel der AG Datenzentren im Verband DHd

Aus Gründen der “guten wissenschaftlichen Praxis”, wegen der Nachnutzbarkeit von Forschungsdaten und hinsichtlich eines fortlaufenden wissenschaftlichen Diskurses, besteht für Forscher ein dringender Bedarf an der nachhaltigen Zugänglichkeit vertrauenswürdiger geisteswissenschaftlicher Forschungsdaten. Aus der Sicht der Forschungsförderer geht es bei der Sicherung und anhaltenden Bereitstellung von Projektergebnissen um die Effizienz von Projektfinanzierungen. Datenzentren scheinen die institutionelle Antwort auf die anstehenden Aufgaben zu sein. Hier gibt es bereits einige erfahrene Vorreiter und viele aktuelle Projekte, dennoch fehlt ein übergreifendes Bild und ein gemeinsames Verständnis, was ein Datenzentrum eigentlich ausmacht. Die *AG Datenzentrum* des DHd-Verbandes stimmt ein solches gemeinsames Verständnis der Ziele und Aufgaben von Datenzentren ab, und identifiziert offene Entwicklungspotentiale. Mitglieder der AG Datenzentren sind Wissenschaftler, die ihr Forschungsdatenmanagement bisher schon selbst durchgeführt haben, sowie bestehende und entstehende Datenzentren aus Deutschland, Österreich und der Schweiz.

Datenzentren gewährleisten die Zugänglichkeit und die Nachnutzbarkeit geisteswissenschaftlicher Forschungsdaten jenseits der aktiven Projektlaufzeit. Aufgaben der Datenkuration (z.B. Dokumentation des Projektkontextes, Migration von Datenformaten) werden in enger Zusammenarbeit mit Wissenschaftlern durchgeführt. Existierende Werkzeuge der Langzeitarchivierung decken dabei nicht immer die spezifischen Anforderungen geisteswissenschaftlicher Forschungsdaten ab; so entstehen z.B. erst langsam Ansätze für die Langzeitverfügbarkeit von Datenbanken, Präsentationssystemen und interaktiven Visualisierungen. Insgesamt scheint es auch erhebliche Unterschiede zwischen Modellen der Research Life Cycles in den Naturwissenschaften und den Geisteswissenschaften zu geben, so dass Lösungsvorstellungen aus dem einen Bereich nicht ohne weiteres in den anderen zu übertragen sind.

Der Aufbau von Datenzentren hat nicht nur wissenschaftliche und technische Aspekte, sondern vor allem auch organisatorische Stabilität und finanzielle Nachhaltigkeit müssen gesichert sein. So war unter den Vorreitern geisteswissenschaftlicher Datenzentren das AHDS, das mit den AHDS-Spezialzentren zu Archäologie, Geschichte, Literatur, darstellende und bildende Künste hunderte Kollektionen mit Millionen an Datensätzen und Datenobjekten beherbergt hat. Nach 10-jährigem Betrieb stellte der Förderer (AHRC, und schließlich auch JISC) die Finanzierung für das AHDS im Jahr 2008 kurzfristig ein, trotz eines Aufschreis der Geisteswissenschaften weltweit.

Ein Datenzentrum deckt insgesamt viele Bereiche ab: von technischen Fragen, über konzeptionelle Herausforderungen bis hin zu organisatorisch-institutionellen Problemen. Es bewegt sich dabei zwischen teils widersprüchlichen Anforderungen, um die langfristige Zugänglichkeit und Nachnutzbarkeit von Forschungsdaten mit möglichst geringen Kosten und Aufwänden zu ermöglichen.

Antworten auf diese Herausforderungen müssen in lokalen Initiativen gefunden, und an die jeweiligen Kontexte und Zielgruppen angepasst werden. Gleichzeitig sind sie relevant für die gesamte DH-Community, denn Vertrauen in das Angebot eines Datenzentrums kann nur auf Basis eines gemeinsamen Verständnisses und gemeinsamer Werte entstehen. Letztlich kann und muss die gesamte DH-Community zu Entwicklung, Betrieb und Nachhaltigkeit der Angebote von Datenzentren beitragen. Dies impliziert insbesondere auch die Notwendigkeit einer stärkeren Vernetzung oder gar Föderation zwischen den Datenzentren, die zu einer gemeinsamen Konzeptentwicklung, einer besseren Lastverteilung und einer insgesamt stabileren Forschungsdatenlandschaft beiträgt.

Gegenwärtig scheinen Fragen aus drei Bereichen zu diskutieren zu sein ...

- **Grundbegriffe, Paradigmen und Definitionen** im Forschungsdatendiskurs. Was sind Daten und Forschungsdaten? Gibt es spezifische Datentypen sind in den Geisteswissenschaften? Wie sieht der Research Data Life Cycle in diesem Bereich aus? Müssen wir zwischen “Daten” und “Ressourcen” unterscheiden? Was bedeutet “Archivierung” und “Kuratierung”? Was sind “Datenzentren” - als konzeptioneller und als institutioneller Begriff? Wie werden allgemeine Begriffsdiskussionen in lokale Konzepte übersetzt und welche Auswirkungen auf das jeweilige Angebot und die organisatorische Nachhaltigkeit haben sie? Welche Profile von Datenzentren lassen sich hinsichtlich ihrer Ausrichtungen unterscheiden? Wie lässt sich das Anforderungsprofil an Datenzentren und das Angebotsspektrum von Datenzentren sinnvoll strukturieren?
- **Lokale Umsetzung und Organisation der Nachhaltigkeit** geisteswissenschaftlicher Daten. Welche Strategien zur Sicherstellung von Nachhaltigkeit können erfolgreich sein? Wie unterscheiden sich Datenzentren in Bezug auf ihre Struktur und Zielgruppe (z.B. institutionelle, disziplinspezifische, nationale Datenzentren). Wie können Angebote von Datenzentren strukturiert werden

(z.B. nach Gigabyte, Datentypen, Projekten, Forschungsmethoden, etc.), so dass Vergleichbarkeit hergestellt werden kann? Wie können die Sicherungsprozesse im Einzelnen organisiert werden: die Übernahme der Daten und Ressourcen, ihre Beschreibung und Dokumentation durch Metadaten, sowie ihre Langzeitarchivierung und ggf. der dauerhafter Betrieb von einzelnen Systemen. Braucht man jenseits der Archivierung auch transparente Lösch-Prozesse für ungenutzte Daten? Wie lassen sich für diese Prozesse und Leistungen geeignete langfristige Finanzierungsmodelle denken?

- **Zusammenarbeit mit den FachwissenschaftlerInnen**. Müssen Forschungsprojekte bereits ab Antragstellung durch erfahrene Datenkuratoren beraten und begleitet werden, um das Datenmanagement bereits von Beginn an auf Nachhaltigkeit auszurichten und um gleichzeitig den Gesamtaufwand von Wissenschaftlern für Datenmanagement-Aufgaben möglichst gering zu halten? Wie kann eine gute Balance zwischen Standardisierung und innovativen projektspezifischen Lösungen gefunden werden? Wie können Standardsierungbemühungen in Fachcommunities unterstützt werden, die gleichzeitig übergreifende Sicherungsprozesse in den Datenzentren erst realistisch machen? Wie kann DH-weit die Publikation von kuratierten Daten aus der Forschung als wissenschaftliche Leistung gefördert werden, geleitet von der Vision eines Impact Factors für solche veröffentlichten Ressourcen?

Das Panel wird in drei “Runden” verlaufen. Zunächst werden einige Vertreter der verschiedenen Institutionen und Projekte kurz über den jeweiligen Entwicklungsstand und Ziele berichten. Danach sollen die oben genannten Themenfelder vorgestellt und thesenartig beleuchtet werden. Schließlich soll die Diskussion der aufgeworfenen Fragen mit dem Publikum geführt werden.

Als Teilnehmer des Panels sind vorgesehen: Matej Durco; Daniel Kurzawe; Lukas Rosenthaler; Patrick Sahle (Convenor der AG); Johannes Stigler und ggf. weitere AG-Mitglieder

Zur Integration computerbasierter raum-zeitlicher Visualisierungen in die Methodik der historischen Wissenschaften

Kartographische Visualisierungen bis hin zu komplexen Geoinformationssystemen sind heute ein fester Bestandteil computerbasierter Anwendungen in den Geisteswissenschaften. Mit den neuen Technologien, welche im Allgemeinen unter dem Schlagwort „Web 2.0“ zusammengefasst werden, ist die Anzahl an frei verfügbaren Grundkarten ebenso sehr gestiegen, wie die Schwierigkeit der Integration in ein System gesunken ist. Doch gerade diese Einfachheit der Visualisierung räumlich bezogener Daten macht es notwendig, den Erkenntniswert einer solchen Nutzung für die Geisteswissenschaften zu hinterfragen. Aus diesem Grund thematisiert dieses Panel die Chancen und Risiken visualisierender Anwendungen für die historischen Wissenschaften.

Ein erster Beitrag von Wolfgang Spickermann behandelt auf einer theoretischen Ebene die Adaption computerbasierter Arbeitsweisen für die historisch-kritische Methode. Hier soll besonders auf die Problematik einer Vereinbarkeit von „naturwissenschaftlicher“ Empirie „geisteswissenschaftlicher“ Hermeneutik sowie deren Auswirkungen auf den Einsatz computerbasierter Systeme in den historischen Wissenschaften eingegangen werden. Ferner gilt es, die digitalen historischen Wissenschaften nicht als reine Textwissenschaften zu verstehen, sondern vielmehr als thematisch bestimmt und sich einer großen Varianz an Quellen bedienend. Räumliche Visualisierungen können hier als Bindeglied der zum Teil sehr divergenten Quellen dienen, wobei eine Beschränkung auf die Darstellung von Verbreitungen bei weitem zu kurz greift. Vielmehr müssen Systeme entwickelt werden, die es erlauben, Wege durch die Gesamtheit der bereitgestellten Quellen auf zu bauen und so historische Argumentationen zu visualisieren.

Ein zweiter Beitrag von Susanne Rau thematisiert das Thema 'Raum' als Forschungsgegenstand in den Geistes- und Kulturwissenschaften. Die in den letzten Jahren interdisziplinär verhandelten und zunehmenden Studien zu Raumwahrnehmungen, Raumpraktiken und Raumnutzungen in historischen und gegenwärtigen Gesellschaften haben in aller Deutlichkeit gezeigt, dass Raum in lebenweltlichen Kontexten weder auf seine Dreidimensionalität reduziert werden kann, noch vorgegeben, noch unbeweglich ist. Ähnliches gälte für den Zeitbegriff. Die meisten computerbasierten raumzeitlichen Visualisierungsinstrumente aber bieten uns bislang nur Werkzeuge zur Visualisierung homogener, euklidischer Räume oder linearer Zeiten. Es stellt also ein dringendes Desiderat dar, die Kompatibilität des "sozialen" und des "technischen" Raum- und Zeitverständnisses zu prüfen, wenn man Geisteswissenschaften und digitale Technologien in der Zukunft besser zusammenzubringen möchte.

Nach diesen theoretischen Überlegungen folgt eine Vorstellung der neuesten Entwicklungen zu raum-zeitlichen Informationssystemen, welche im Rahmen des ICE (Interdisciplinary Center of eHumanities in History and Social Sciences) entwickelt werden.

Einen Einblick in die Entstehung eines computerbasierten raum-zeitlichen Werkzeuges gibt René Smolarski, der anhand eines sich derzeit in der Entwicklung befindlichen prototypischen Virtuellen Kartenlabors aufzuzeigen versucht, mit welchen Problemen die Durchführung eines solchen Projektes im Hinblick auf die Kommunikation zwischen Informatik und Fachwissenschaft verbunden

ist. Das Ziel dieses Projektes ist es zum einen, die Kartenbestände der Sammlung Perthes Gotha einem breiten und ortsunabhängigen Publikum online zur Verfügung zu stellen, zum anderen, auf der Basis konkreter wissenschaftlicher Anforderungen seitens verschiedener geisteswissenschaftlicher Disziplinen entsprechende Werkzeuge zu entwickeln, um den Erkenntniswert der zugrundeliegenden Sammlungsobjekte, neben Karten auch eine sehr heterogene Masse von Archivalien, zu erhöhen und explizite Forschungsfragen an diese richten zu können. Dabei spielt gerade der Dialog zwischen Geisteswissenschaft und Technik eine herausragende Rolle, die für das Gelingen eines solchen Projektes von elementarer Bedeutung ist.

Das Leipziger Projekt eXChange (http://www.exchange-projekt.de), gefördert von 2012-2015 durch das BMBF, widmet sich der Entwicklung einer konzeptbasierten Suche für die digital vorliegenden Textkorpora der griechischen und lateinischen Literatur der Antike. In Kooperation mit der Informatik wird ein Recherchesystem aufgebaut, das Bedeutungsverschiebungen semantischer Räume nach Ort und Zeit visualisiert und durch das Begriffe auf ihren Kontext in argumentativen Strategien hin analysiert werden können. Mit dem Recherchefrontend wird es möglich sein, sowohl einzelne Wörter und Paraphrasen ausfindig zu machen, als auch Texte nach vordefinierten Konzepten durchsuchen können. Inhaltlich untersucht das Projekt dies exemplarisch am Verhältnis von Wissenschaftssprache und alltäglichem Handeln insbesondere in der Politik von der Antike bis in die frühe Neuzeit.

Ferner stellt Martin Dreher (Universität Magdeburg) die TheDeMa (Thesaurus Defixionum Magdeburgensis), eine relationale Datenbank, die sämtliche antike Fluchtafeln (defixiones) nach äußeren und inneren Merkmalen erfasst. Auf dieser Grundlage wird die transkulturelle Entwicklung der antiken Fluchformeln als Mittel der Durchsetzung subjektiven Rechts in verschiedenen Regionen und in verschiedenen Perioden analysiert.

Die vorgestellten Projekte verstehen sich noch nicht als Umsetzungen der in den theoretischen Überlegungen formulierten Anforderungen, sondern vielmehr exemplarisch und als Mosaikteile, welche zukünftig zu einem Ganzen zusammengeführt werden müssen. Die Vereinbarkeit und Erweiterung der Einzelprojekte zu einer virtuellen Arbeitsumgebung, welche den spezifischen Anforderungen eines historisch-kritischen Arbeitens entspricht, steht im Zentrum des letzten Beitrags.

Wie kann man nun erreichen, dass die vielfältigen verteilten Aktivitäten der Digital Humanities interoperabel werden und durch ungeahnte Vergleiche, Verbindungen und bislang undenkbare Interaktionen die Wissenschaftler(innen) zu unerwarteten Sichten und neuen Hypothesen treiben? Das Fraunhofer IDMT stellt innovative Schnittstellentechnologien vor - sogen. Webbles - mit denen bisher monolithische Systeme vernetzt werden können.

‘Over-tagging’ with XML in Digital Scholarly Editions

Elise Hanrahan, hanrahan@bbaw.de

NOTE: Obwohl ich mein Abstrakt auf Englisch geschrieben habe, kann ich den Vortrag gern auf Deutsch halten.

This talk looks at the phenomenon of over-tagging (term created here) in XML, which consists of exaggerated and unfocused tagging that concentrates on diplomatic characteristics. These tags are used for the display of the digital edition text on the computer screen--an especially questionable result when digital facsimiles exist.

To what degree computer technology affects practices and theories in scholarly editing is an open question. But whether or not we are in the midst of a revolution or simply experiencing a change of tools, there are certain influences that can already be observed. One is the availability of digital facsimiles combined with the use of XML.

How could digital facsimiles change online scholarly editions? Digital facsimiles challenge one of the claimed purposes of a scholarly edition—the recreation of the original manuscript.[1] This aim, often strived for by means of a diplomatic transcription, is particularly prevalent in recent editions. Elements of diplomatic transcribing can for instance be found in most German critical editions from the last twenty-five years.[2]

Yet the question ‘How *could* digital facsimiles change online scholarly editions?’ was posed because astonishingly the diplomatic method continues to be dominant. It is thus the argument of this talk that online editions do not reflect this significant alteration in the relationship between researcher and original source. Instead the same editorial method is being used, and the primary new development is the use of XML instead of Microsoft Word. Indeed not only do digital editions continue to be diplomatic, but the tendency has even increased.[3]

---

[1] It should be noted that this very strong focus on an ‘authentic’ recreation of the original handwriting is fairly new in Editionswissenschaft, starting in the 1970s and becoming very dominant in the 1990s with editors like Hans Zeller.

[2] For a very concise summary of the take-over of the material-paradigma, see: Rasch, Wolfgang, Wolfgang Lukas, and Jörg Ritter. "Gutzkows Korrespondenz – Probleme Und Profile Eines Editionsprojekts." *Brief-Edition Im Digitalen Zeitalter (Beihefte Zu Editio)* 34 (2013): 99.

[3] Elena Pierazzo addresses the popularity of digital documentary editions in: Pierazzo, Elena. "Digital Documentary Editions and the Others." *Scholarly Editing: The Annual of the Association for Documentary Editing* 35 (2014). Accessed November 10, 2014. http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html.

Structure of the Talk

I. Arguments for why diplomatic transcriptions are no longer necessary when digital facsimiles are available
II. An examination of why diplomatic aspects in digital editions have surprisingly increased instead of decreased
III. Arguments and counter-arguments for continuing to transcribe diplomatically in digital editions
IV. Suggestions for alternative editorial priorities

I. Digital facsimiles are a game-changer

There are two main arguments for recreating the original manuscript in the edited text (thus using a diplomatic transcription), both of which online facsimiles challenge. The first is to bridge the gap between the original manuscript and the researcher. Before digital facsimiles it was quite possible that the researcher never set eyes on the original manuscript, which was carefully stored away in a library or archive. The edition thus strove to offer the researcher an objective depiction of the handwriting. Now, however, the researcher can look at the image of the handwriting online, thus the edition must no longer bridge this gap.

The second argument for a diplomatic transcription is to preserve the manuscript. If anything happened to the original source, the text of the edition could be used as a replacement. Additionally the edition protects the manuscript from being over-handled, because it functions as an authoritative substitute. There is however no longer a need to create a substitute for the manuscript, because a high-quality image exists.

Despite these arguments, in praxis one finds digital editions still ruled by the diplomatic trend. Why is this?

II. The diplomatic-tradition and XML

There are two reasons for the prevalence of diplomatically-influenced digital transcriptions. The first is that digital editions emerged at the same time diplomatic editing was the dominant method for scholarly editions.[4] It is therefore not surprising that the current method for print editions was transferred to the newly emerging digital editions.

The second reason is due to the very nature of XML. In XML, unlike in Microsoft Word, specifications for the visual presentation of the edited text are completely separated from the documentation of the original source. Hence in XML the editor is no longer limited by the

[4] That is, the end of the 1990s/start of the 2000s

space on the page for recording textual phenomena and can enter as many XML tags as desired, allowing a theoretically endless documentation of the characteristics of the manuscript. Combine this opportunity with an already diplomatic trend, and the result is a lot of diplomatic XML-tagging, sometimes to an incredibly minute degree. This very real phenomenon shall be called 'over-tagging'.

Over-tagging refers to an exaggerated amount of XML tags that do not pursue a specific research question, but are in praxis only used for the display of the edited text on the screen. Over-tagging could be for example using a character in the line below an indentation to mark the length of an indentation, tagging the exact location and angle of marginalia, tagging orthographical elements like the long *s* in German, or tagging line breaks that are not semantically meaningful.

While there is nothing inherently wrong with tagging these kinds of textual phenomena, it is important to ask, what is the purpose of these tags? Such tagging is especially problematic when it comprises a large part of the XML schema. And one must add, no matter how detailed a transcription is, it can't recreate the image of the handwriting and the vast amount of data that the image carries. And not only is the usefulness of the results debatable, over-tagging takes a great deal of time and energy. Other fundamental editorial tasks fall by the wayside--tasks such as editorial commentary, to name only one of many. There are many other gaps to be bridged between the reader and original source besides the material gap. An essential benefit to be gained from more reflective tagging practices is the time to focus on these other editorial tasks.

III. Counter-arguments: machine searchable and a reader-aid

One argument for over-tagging is machine readability. This means that tagged textual information can be found in automated searches. Yet does anyone truly search for aspects such as indentation size? An editor might argue: ‘Perhaps not now, but someone could in the future. And what’s more, someone could discover that this seemingly insignificant textual characteristic actually carries a semantic meaning’. Such an answer reveals the editorial tradition that still strongly underlies digital editions today and is inseparable from diplomatic transcribing. This is the philosophy that literally everything on the page could be significant.[5]

There are two responses to this. First of all search masks are currently not being made to search for diplomatic characteristics. In praxis these tags are used for the display of the text only. It is true that, if desired, search masks could be made for this purpose. It is also true that

[5] This of course leads back again to the authenticity/materiality movement from Hurlebusch, Zeller and others. For an example of such a perspective that does not directly lead to the solution of a diplomatic transcription (but instead to digital facsimiles), see: Richter, Elke. "Goethes Briefhandschriften digital – Chancen und Probleme elektronischer Faksimilierung." *Brief-Edition Im Digitalen Zeitalter (Beihefte Zu Editio)* 34 (2013): 53-75.

it is impossible to prove that a certain aspect of a page is not meaningful. However, the ‘everything could be significant’ position is not a feasible editorial method. It is definitely not the basis on which to build a well-functioning XML schema. In reality, at the end of the project there exists a very large amount of information that is solely used for the display of the text on the computer screen.

Another argument for over-tagging is that the resulting text is a kind of facsimile-reading-aid.[6] This is however a problematic stance. Firstly, critical editions are not made primarily to be reading-aids for facsimiles, although they can be helpful for this purpose. Critical editions are editorial arguments and offer a readable edited text according to that argument. A facsimile-reading tool is potentially very useful, but is something different than a critical edition. Secondly, a diplomatic transcription was not conceived for this aim and is probably not the best method. There are certainly much better ways to help guide a researcher through a facsimile than simply mirroring the facsimile in the edited text, especially in consideration of technological possibilities.

IV. Different priorities for digital editions

Relinquishing over-tagging means more time for editors to concentrate on other aspects of editing, such as commentary and semantic tagging (including not just person or place names, but also more abstract themes, such as concepts found in the texts). There would also be more time to tag the creative process of the author (such as capturing the layers of the text's development by tagging crossed out/added words). There would be more time to enter meta data, to link through standard IDs like VIAFs for persons and geonames for places, and to simply to think about how to use the digital space to the researcher's best advantage.

In many ways, instead of reflecting on what doors technology opens for critical editions and thus shaping technology to this end, editors have let technology define them, losing sight of priorities in today’s digital world. For instance, an essential current challenge for digital editions is to avoid the 'island' problem—single editions floating in the internet without a real connection to one another. Minute diplomatic tagging does not address this problem (standard IDs and meta data to some degree does). Yet it isn’t a question of what is important and what is not--all editorial tasks are important--it is a questioning of appreciating what an edition has to offer and carefully considering the energy invested and the benefits gained. ‘Over-tagging’ is perhaps a very small piece of the debate on digital editions, but it could point to a general direction and is therefore worthwhile to consider in this context.

[6] This idea is touched on in Pierazzo(2014), 4.

Erkennung und Visualisierung attribuierter Phrasen in Poetiken

Andreas Müller (1), Markus John (2), Steffen Koch (2), Thomas Ertl (2) und Jonas Kuhn (1)

(1) Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
(2) Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart

Einleitung

In wissenschaftlichen Werken über Literatur (zum Beispiel Poetiken) spielen Referenzen zu Autoren, fiktiven Charakteren aus literarischen Werken und anderen Arten von Personen eine wichtige Rolle. Zum Beispiel bildet Personenerkennung eine der Grundlagen für die Erkennung der Sprecher von direkter Rede in literarischen Werken (Elson and McKeown 2010). Diese Information kann weiter benutzt werden um, zum Beispiel, soziale Netzwerke zu extrahieren (Elson et. al. 2010).

In diesem Abstrakt stellen wir eine Erweiterung der in John et. al. 2014 präsentierten Technik zum Vergleich von Textdokumenten vor. Diese Erweiterung basiert auf der Erkennung von Personennamen und Personen zugeordneten Konzepten. Ein Beispiel für ein einer Person zugeordnetes Konzept ist „Schillers Poesie“ oder „Klopstocks Messias“. Im ersten Fall wird mit der Phrase die gesamte Poesie Schillers referenziert, im zweiten Fall das konkrete Werk „Messias“ von Klopstock. Im Folgenden wird gezeigt wie man mit einem simplen, auf morphologischer Analyse, Nominalphrasenerkennung und der Erkennung von Personennamen basierendem Suchmuster solche Phrasen extrahieren kann. Nach der Extraktion werden Personennamen und Phrasen in die bereits erwähnte Technik integriert. Auf dieser Basis haben wir eine benutzerbasierte Evaluation durchgeführt, die zeigt, dass die Erweiterung der Technik durch Personennamen und Personen zugeordneten Konzepten bei literarischen Vergleichen und Analysen von Texten hilfreich ist.

Als Grundlage für unsere Untersuchung verwenden wir Poetiken aus einem Korpus von 20 Texten, die im Rahmen des Projekts ePoetics untersucht werden. Das Korpus wurde aus 1000 Poetiken ausgewählt, die von Richter (2010) analysiert wurden. Für unsere Analyse verwenden wir die vier Poetiken der Autoren Staiger, Scherer, Kleinpaul und Engel. Diese wurden von dem Experten für Literaturwissenschaft in unserem Projekt als sehr interessant für literarische Textvergleiche eingestuft.

Methode

Für die linguistische Vorverarbeitung verwenden wir die OpenNLP Tools[1] für automatische Satz- und Worterkennung. Des weiteren benutzen wir die mate tools[2] (Bohnet 2010) für Lemmatisierung, automatische morphologische Analyse und Wortartenerkennung und die StanfordCoreNLP library (Finkel et. al. 2010) mit den Modellen fürs Deutsche von Faruqui and Pado (2010) zur Erkennung von Personennamen. Für die Erkennung von Nominalphrasen benutzen wir die in MuNPEx[3] enthaltenen JAPE-Grammatiken.

Nach diesen linguistischen Vorverarbeitungsschritten suchen wir alle Personennamen, die im Genitiv auftreten. Diese signalisieren Vorkommen von den Personen zugeordneten Konzepten. Anschließend extrahieren wir für jeden Personennamen die ununterbrochene Sequenz von Nominalphrasen die dem Personennamen am nächsten ist als Konzept, das dem Personennamen zugeordnet wird. Wir extrahieren die Sequenz von Nominalphrasen statt nur der am nächsten

1 https://opennlp.apache.org/
2 https://code.google.com/p/mate-tools/
3 http://www.semanticsoftware.info/munpex

stehenden Nominalphrase um auch komplexe Nominalphrasen zu erfassen.

Integration in das System

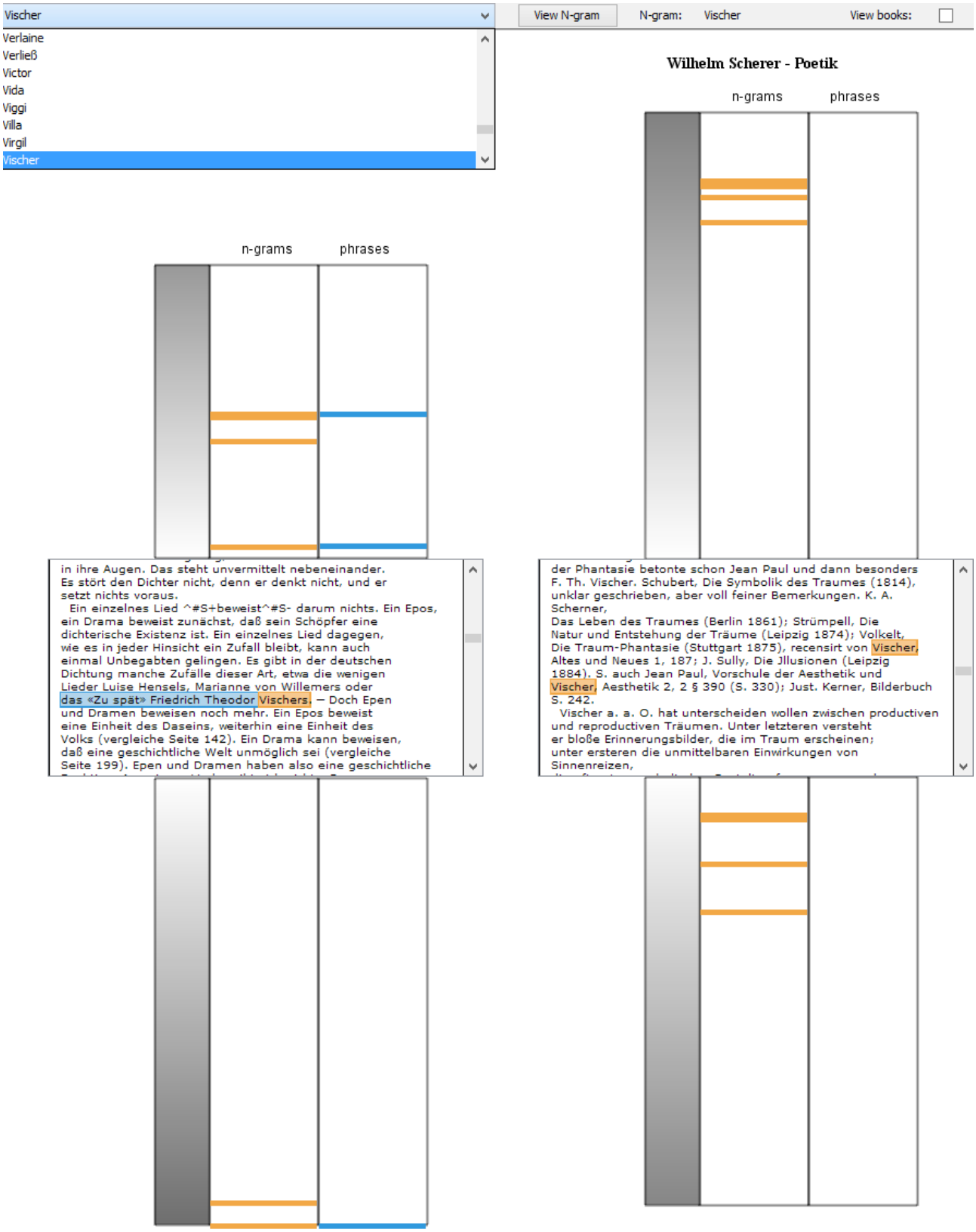


Abbildung 1. Zwei ausgewählte Dokumente werden als Band dargestellt. Die Vorkommen von Personen (Orange) und Phrasen (Blau) werden als Balken dargestellt.

In Abbildung 1 ist ein Screenshot des genannten Systems dargestellt. In der linken oberen Ecke befindet sich eine Liste von Personennamen. Nach Auswahl einer Person, werden die Vorkommen der Personen (Orange) und die der Person zugewiesenen Konzepte (Blau) in den Dokumenten als Balken dargestellt. Durch die Selektierung der Vorkommen, können Textpassagen weiterführend untersucht werden.

Literaturwissenschaftliche Evaluation

Um den Vorteil für diese Erweiterung aus literaturwissenschaftlicher Sicht zu zeigen, führten wir eine Evaluation mit einem Literaturexperten durch. Nach einer kurzen Einführung in das System, begann der Experte die 4 ausgewählten Poetiken im Hinblick auf die vorkommenden Personen zu analysieren. Es ist noch zu erwähnen, dass der Experte schon mit den Poetiken vertraut war, was die Nähe zu einem realistischen Analyse-Szenario erhöht, da eine literaturwissenschaftliche Analyse genau damit beginnt, sich mit dem Untersuchungsgegenstand vertraut zu machen, einen Text ggf. also auch mehrfach zu lesen.

Beim Durchgehen der Liste bemerkte der Analyst den Namen „Aristoteles“, der ihn interessierte. Bei der Auswahl von „Aristoteles“ fiel dem Analysten sofort auf, dass dieser Name in der Poetik von Staiger nur selten und in der Poetik von Scherer sehr häufig vorkommt, was er so nicht erwartet hatte. Solche Frequenz-basierten Eigenschaften lassen sich durch die Visualisierung sehr schnell erkennen. In der Poetik von Scherer fiel ihm weiterhin auf, dass die meisten Aristoteles zugeordneten Phrasen „Aristoteles Poetik“ oder „Aristoteles Rhetorik“ referenzierten. Dies untermauert die These, dass diese beiden Werke für Scherer eine hohe Bedeutung haben empirisch.

Eine weitere Frequenz-basierte Eigenschaft erkannte der Experte, als er „Homer“ auswählte. Im Kapitel über Epik bei Staiger (linkes Dokument mittig, siehe Abbildung 2) ist ein häufiges Vorkommen erkennbar. Dies bestärkte ihn in seiner Annahme, dass Staiger einen Hauptvertreter für jede der 3 Gattungen Epik, Lyrik und Dramatik benennt. Durch Auswahl der anderen beiden Hauptvertreter, „Goethe“ und „Schiller“, lässt sich diese Annahme empirisch untermauern, da man erkennen kann, dass „Goethe“ im Kapitel über Lyrik und „Schiller“ im Kapitel über Dramatik besonders häufig vorkommt.

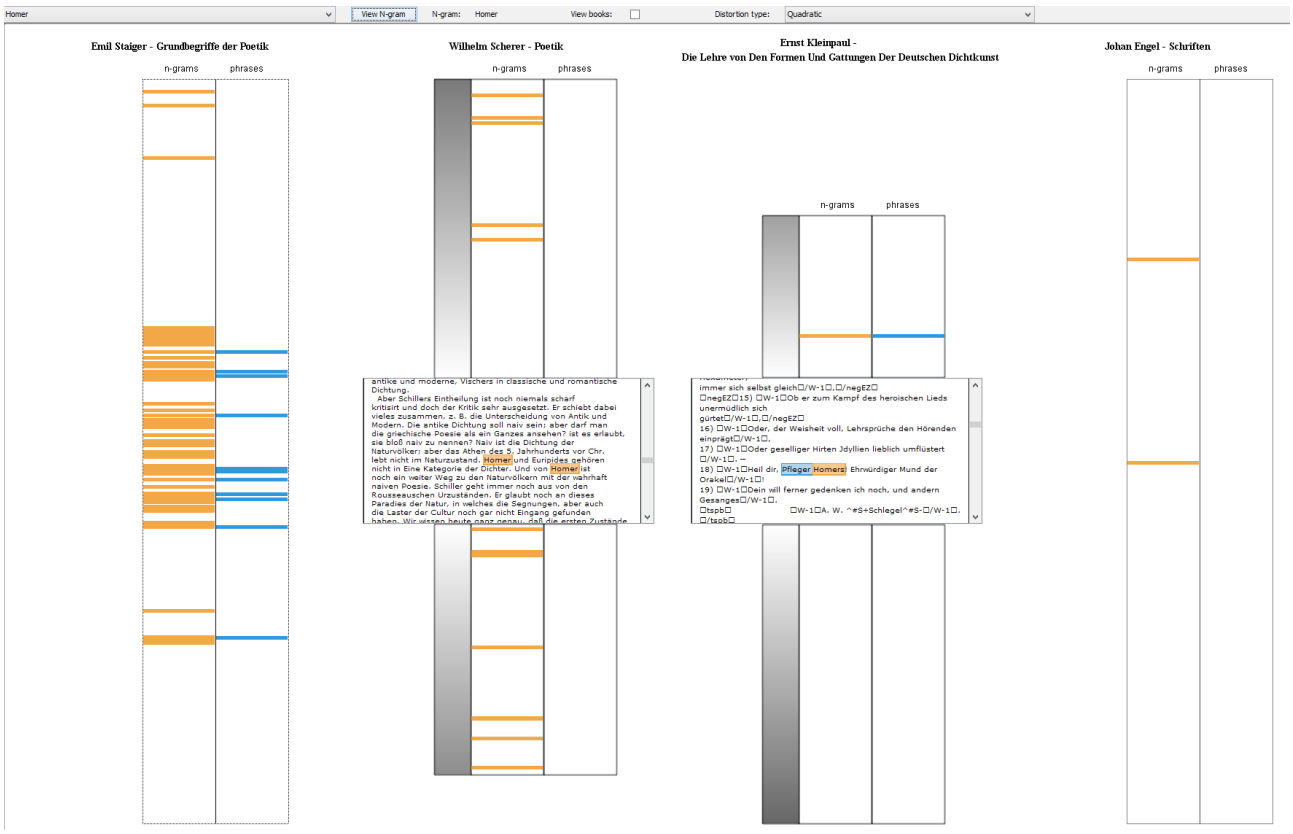


Abbildung 2: Häufiges Vorkommen von „Homer“ in dem Kapitel über Epik von Staiger (linkes Dokument mittig)

In diesem Abschnitt wurde exemplarisch gezeigt, dass die Stärken des Systems in der Übersicht über Frequenz-basierte Eigenschaften von Dokumenten und der Verwendung von Personennamen in Dokumenten liegen. Außerdem enthalten die Phrasen unter anderem Referenzen auf Personen zugewiesene Werke, wie zum Beispiel, „Aristoteles Poetik“ oder „Aristoteles Rhetorik“. Dadurch lässt sich schnell ein Überblick über Diskussionen über diese Werke und den Stellenwert der Werke in einem Dokument gewinnen. Durch die Ansicht mehrerer Dokumente, wird ein einfacher Vergleich von Textstellen über Personen oder Phrasen ermöglicht.

Technische Evaluation

Um die Präzision der Phrasenerkennung einschätzen zu können, haben wir auf Basis der Poetik von Staiger jedes erkannte attribuierte Konzept in der Poetik in eine von 4 Klassen eingeteilt, die im Folgenden aufgelistet sind. Diese Evaluation dient hauptsächlich dazu für weitere Forschungen Fehlertypen zu finden. Die 4 Klassen sind:

1. Vollständig korrekt: Personenname und attribuiertes Konzept werden korrekt erkannt

2. Teilweise korrekt weniger: Es wird mindestens ein Wort des Personennamens und des attribuierten Konzepts erkannt, aber es wird auch mindestens ein Wort des Personennamens oder des attribuierten Konzepts nicht erkannt

3. Teilweise korrekt mehr: Personenname und attribuiertes Konzept werden korrekt erkannt, aber es wird auch linguistisches Material erkannt, das weder zum Personennamen noch zum attribuierten Konzept gehört

4. Inkorrekt: Entweder wird kein Wort des Personennamens oder kein Wort des attribuierten

Konzepts erkannt. Dieser Klasse werden auch Annotationen zugewiesen bei denen es sich nicht um ein attribuiertes Konzept handelt

Diese Einteilung der erkannten Konzepte, sowie die Fehleranalyse im nächsten Abschnitt, wurden bisher nur von dem Erstautor des Abstrakts vorgenommen. Deshalb sollten die Zahlen als Schätzung der Qualität der Phrasenextraktion, nicht als definitive Evaluation angesehen werden. Eine Verifizierung der Zahlen durch einen zweiten Annotator ist geplant.

In der Poetik von Staiger erzielen wir folgende Resultate:

Vollständig korrekt: 72
Teilweise korrekt weniger: 25
Teilweise korrekt mehr: 6
Inkorrekt: 25

Es werden 56% der Instanzen komplett richtig und 20% komplett falsch erkannt. 24% der Instanzen werden nicht komplett richtig erkannt. Allerdings sind einige der inkorrekten Instanzen auf Fehler in der Erkennung von Personennamen zurückzuführen. So wird zum Beispiel „Gott“ als Personenname erkannt, was zu einem Fehler führt der mit der Erkennung der Nominalphrasen nichts zu tun hat.

Wir haben auf der Basis der 24% nicht komplett richtig erkannter Instanzen eine Fehleranalyse durchgeführt und zeigen im nächsten Abschnitt exemplarisch eine häufige Art von Fehler.

Fehleranalyse

Eine häufige Art von Fehler ist, dass oft einer komplexen Nominalphrase zugehörige Phrasen nicht erkannt werden. So wird „Goethes Forderung“ als attribuierte Phrase erkannt, nicht aber die vollständige Phrase „Goethes Forderung an ein gutes Gedicht“. Ein anderes Beispiel ist die Phrase „ein Gedicht Hebbels“. Diese Phrase wird als attribuierte Phrase erkannt, die komplette Phrase wäre aber „Ein Gedicht Hebbels, das «Lied» überschrieben ist“. Die Phrase „das Lied überschrieben ist“ beinhaltet zusätzliche Informationen, die es dem Leser erlauben zu erkennen, welches Gedicht Hebbels gemeint ist. Diese Art von Fehler lässt sich wahrscheinlich durch Einbindung eines Dependenz- oder Konstituentenparsers und darauf aufbauender Erkennung komplexerer Phrasen erkennen.

Für eine weitergehende automatische Verarbeitung der extrahierten Phrasen wären diese Fehler schwerwiegender als für die Integration in das vorgestellte Visualisierungssystem, da die Phrasen in ihrem Kontext dargestellt werden. Dadurch können Fehler schnell gefunden und vom Analysten leicht korrigiert werden. Außerdem lassen sich auch bei einer fehlerhaften automatischen Analyse zum Beispiel Frequenz-basierte Eigenschaften erkennen, solange der Fehler nicht darin besteht, dass eine Stelle markiert wird an der keine attribuierte Phrase vorliegt. Dadurch kann das System auch auf ältere Varianten von Sprachen, bei denen automatische Methoden oft nicht so gut funktionieren wie bei moderner Sprache für die sie entwickelt wurden, angewendet werden. Dies ist vor allem in literarischer Textanalyse hilfreich, da in diesem Bereich auch mit älteren Dokumenten gearbeitet wird.

Ausblick

Die Erweiterung des Ansatzes wurde im Rahmen des ePoetics Projekt entwickelt und evaluiert. Wir haben damit begonnen, das System so zu erweitern, dass die oben beschriebenen Fehler vermieden werden. Dazu verwenden wir einen Dependenzparser (Bohnet, 2010), der syntaktische Analysen

der Sätze bereitstellt. Letzten Endes sollen neben attribuierten Konzepten auch sonstige Äußerungen, Zitate und Referenzierungen von Dritten automatisch erkannt und über eine entsprechende Visualisierung zur Verfügung gestellt werden.

Referenzen:

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 89-97.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 138-147.

Elson, David K. ; McKeown, Kathleen ; Fox, Maria (Bearb.) ; Poole, David (Bearb.): Automatic Attribution of Quoted Speech in Literary Narrative.. In: AAAI : AAAI Press, 2010

M. Faruqui and S. Pado. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. Proceedings of Konvens 2010, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Koch, Steffen; John, Markus; Wörner, Michael; Ertl, Thomas: VarifocalReader – In-Depth Visual Analysis of Large Text Documents. In: IEEE Transactions on Visualization and Computer Graphics (TVCG) (2014) (Noch nicht erschienen).

S. Richter. 2010. A History of Poetics: German Scholarly Aesthetics and Poetics in International Context, 1770- 1960. De Gruyter.

PD Dr. Friedrich Michael Dimpel

Ausgezeichnete Mären analysieren – ein Werkstattbericht

1.

Digitale Korpora können dank der Fortschritte im Bereich der Digital Humanities mit zahlreichen quantitativen Analyseverfahren untersucht werden: Im Rahmen der Stilometrie beschäftigen sich Computerphilologen etwa mit Fragen der Autorschaftsattribution oder der Werkchronologie sowie mit Fragen nach Spezifika von Gattungen oder Epochen.[1] Auch wenn digitale Texte mit zahlreichen quantitativen Analyseverfahren untersucht werden können: In der Regel sind vollständige Texte oder zusammenhängende Textabschnitte wie Buchkapitel der Gegenstand etwa von stilometrischen Studien. Die Möglichkeiten für automatische Analysen enden jedoch dort, wo solche spezifische Textebenen in den Blick zu nehmen wären, die repräsentieren, welche Eigenschaft auf welche Figur bezogen wird, und welche Figuren oder Erzähler diese Zuschreibung vornehmen: Wer spricht? Zudem: Über wen wird gesprochen – welche Terme werden auf welche Figuren bezogen? Gibt es dabei bspw. genderspezifische Distributionen? Auf welche Weise wird gesprochen? Es ist ein erheblicher Unterschied, ob eine Figur über ihre Liebe spricht und damit ihre Emotion in der erzählten Welt öffentlich macht, oder ob nur eine Gedankenrede einer Figur erzählt wird. Wenn man auf solche Informationen zugreifen möchte, dann ist es nötig, die Texte mit einer entsprechenden Annotation zu versehen.

Quantitative Analysen von Textsamples, die spezifische Textdaten enthalten wie Figurenrede bestimmter Aktanten, fokalisierte Textpassagen oder Textpassagen, die sich auf bestimmte Aktanten beziehen, sind derzeit nicht möglich ohne eine aufwendige manuelle Textaufbereitung für eine konkrete Fragestellung.

Wie wichtig jedoch eine narratologische Textauszeichnung in einem diachron angelegten Korpus wäre, unterstreichen Fotis Jannidis, Gerhard Lauer und Andrea Rapp: „Wer hat, wann und wo zum ersten Mal die Form der erlebten Rede eingesetzt, wer die episodische Reihung zu psychologischer Figurenzeichnung verdichtet? Wo verknüpfen sich populäre Novellenstoffe mit hochkulturellen Erzähltechniken [...]? Denn nur mit Hilfe des Computers lässt sich ein hinreichend großes Korpus über einen langen Zeitraum hinweg untersuchen und damit etablierte literaturhistorische Methoden um serielle, computerbasierte Verfahren so ergänzen, dass eine Geschichte des Erzählens überhaupt erst geschrieben werden kann."[2]

---

1 Vgl. exemplarisch JOHN F. BURROWS: ‚Delta'. A Measure of Stylistic Difference and a Guide to Likely Authorship. In: Literary and Linguistic Computing 17, 2002, S. 267–287, FRIEDRICH MICHAEL DIMPEL: Computergestützte textstatistische Untersuchungen an mittelhochdeutschen Texten. Tübingen 2004, FOTIS JANNIDIS: Methoden der computergestützten Textanalyse. In: Vera Nünning / Ansgar Nünning (Hrsg.), Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen. Stuttgart 2010, S. 109–132, CHRISTOF SCHÖCH: Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In: Christof Schöch / Lars Schneider (Hrsg.), Literaturwissenschaft im digitalen Medienwandel. 2014, S. 130–157.

2 FOTIS JANNIDIS / GERHARD LAUER / ANDREA RAPP: Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des

2.

Bei narratologisch annotierten Korpora zur deutschen Literatur handelt es sich um ein Desiderat.[3] Daher soll ein Korpus von 100 Kurzerzählungen narratologisch annotiert werden. Das Korpus soll historisch relativ ausgewogen angelegt werden: 30 mittelhochdeutsche und frühneuhochdeutsche Mären, 10 Boccaccio-Novellen, 10 Chaucer-Tales, 30 neuhochdeutsche Novellen und 20 neuhochdeutsche Kurzgeschichten werden aufgenommen. Als Ebenen der Auszeichnung sind vorgesehen:

1. Welche Figuren befinden sich an dem Ort, von dem erzählt wird?
2. Welche Figurenbewegungen im Raum finden statt?
3. Welche temporalen Abweichungen ereignen sich (Prolepsen etc.)?
4. Welche Figur ist wie fokalisiert?
5. Um welche Art von Redewiedergabe (direkte Rede, Bewusstseinsdarstellung etc.) handelt es sich – und welche Figur denkt oder spricht?
6. Um welche Art von Erzählerrede (Descriptio, Bericht Figurenaktivität, Erzählerreflexion etc.) handelt es sich?
7. Auf welche Figur bezieht sich eine Figuren- oder Erzählerrede?
8. Auf welche Figur bezieht sich eine wertende Äußerung einer anderen Figur oder des Erzählers?
9. Steht eine Äußerung in Negation?
10. Liegt eine uneigentliche Rede vor (metaphorisch, ironisch etc.)?
11. Inwieweit besteht Unsicherheit hinsichtlich der Eindeutigkeit der Textdaten?

Die nötigen XML-Elemente sind noch nicht im TEI-Standard enthalten, daher muss ein geeignetes Tagset entwickelt werden.[4] TEI-Kompatibilität wird jedoch angestrebt: Die Elemente können über das Roma-Tool in eine ODD-Datei integriert werden.[5]

Romans in Deutschland (1500-1900). In: Lucas Marco Gisi / Jan Loop / Michael Stolz (Hrsg.), Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien. germanistik.ch 2006. (= online unter http://www.germanistik.ch/publikation.php?id=Hohe_Romane_und_blaue_Bibliotheken).

3 Allerdings kann in vielfältiger Weise auf die wichtige Grundlagenstudie von ANNELEN BRUNNER: Automatische Erkennung von Redewiedergabe in literarischen Texten. Diss. masch. Würzburg 2012, aufgebaut werden. Brunner hat in ihrer Dissertation ein Annotationsverfahren für Redewiedergabe entwickelt und ein Korpus, das aus Texten von 1787 bis 1913 besteht, manuell annotiert. Auch wenn die automatische Erkennung der Redewiedergabe (regelbasiert und via Maschinelles Lernen) noch nicht eine Fehlerrate erreichen kann, die narratologische Auswertungen erlaubt, kann das vorliegende Projekt in konzeptioneller Hinsicht von Brunners Studie erheblich profitieren.

4 Zu narratologischen Desideraten von TEI-P5 vgl. FOTIS JANNIDIS: TEI in a Crystal Ball. In: Literary and Linguistic Computing 24, 2009, S. 253–265, hier S. 261f.

5 Vgl. http://www.tei-c.org/Guidelines/Customization/odds.xml.

Als XML-Elemente werden vorgestellt:

1. `<FigurFokusort> (@Bezeichnung, @Figurengruppe)`
2. `<BewegungLokal> (@Typ)`
3. `<Chronologie> (@Typ)`
4. `<Fokalisiert> (@Bezeichnung, @Typ)`
5. `<Redewiedergabe> (@Typ, @Bezeichnung, @non-fact, @level)`
6. `<Erzählerrede> (@Typ, @Bezeichnung)`
7. `<Figurenbezug> (@Unmittelbar, @Mittelbar)`
8. `<Wertung> (@BezeichnungWertende, @BezeichnungGewertete, @Typ)`
9. `<Negation> (@Typ)`
10. `<UneigentlicheRede>`
11. `<certainty> sowie @cert als Attribut zu allen Elementen` (> TEI P5)

3.

Zentral für den Workflow ist die Entwicklung von Annotationsrichtlinien. Es wird angestrebt, dass verschiedene Versuchspersonen beim gleichen Text zu homogenen Ergebnissen kommen. In ähnlicher Weise, wie man sich bei der Herstellung einer Edition über Kollationierungsregeln verständigen muss, sind hier Annotationsregeln zu erarbeiten. Solche Regeln sind nötig, weil selbst scheinbar eindeutig definierte narratologische Phänomene sich oft nicht eindeutig in Texten wiederfinden lassen. Zudem ist Ambiguität ein charakteristisches Merkmal von literarischen Texten.

Mit Blick auf das Homogenitätsziel werden Annotationsregel und Bearbeitungsdatum direkt im XML-Code dokumentiert, damit bei einer Weiterentwicklung der Annotationsrichtlinien einerseits rasch auf eine einschlägige Fallsammlung zugegriffen werden kann; anderseits können bei einer Regelrevision Entscheidungen gezielt aufgesucht und revidiert werden. Dabei hilft ein eigener Projekteditor, der in Perl/TK implementiert wurde, der neben dem Annotationsfenster in einem zweiten Fenster in Kurzform Informationen zu bereits ausgezeichneten Elementen einblendet.

4.

Annotiert wurden bislang sechs Texte. Vorgestellt werden einige Probleme, die sich bei der Auszeichnung von ‚Sperber' und das ‚Häslein'[6] etwa durch Segmentierung oder durch Ambiguitäten ergeben haben. Anhand von Auswertungsdaten soll exemplarisch aufgezeigt werden, in welch vielfältiger Weise ein entsprechend annotiertes Korpus Analysen möglich macht; etwa in Bezug auf

a) multiple Methoden:

i. Das Korpus kann wie andere Korpora auch mit einer Vielzahl an statistischen Methoden analysiert werden – etwa in Hinblick auf Heterogenität oder Homogenität.

ii. Eine besondere Stellung nimmt das Korpus jedoch dadurch ein, dass auf Basis der Textauszeichnung eine Sample-Erstellung für spezifische Fragestellung möglich ist, die

---

6 Ausgabe: KLAUS GRUBMÜLLER: Novellistik des Mittelalters. Märendichtung. Frankfurt/Main 1996 (=Bibliothek des Mittelalters 23)

nicht nur chronologisch-lineare Zugriffe auf Korpussegmente erlaubt, sondern systematische Zugriffe auf gleichartig annotierte Korpussegmente. So lässt sich ein Korpussegment bspw. mit Bewusstseinsdarstellung von weiblichen Figuren mit einem Korpussegment vergleichen, das aus Erzählerrede besteht; die Figurenrede von Antagonisten lässt sich mit Figurenrede von Protagonisten vergleichen, u.v.m.

b) multiple Fragestellungen, beispielsweise:

i. Wie steht es um die diachrone Entwicklung von Fokalisierung, um die Eigenschaften von Erzähler- und Figurenrede, um temporale Alternationen, wie verteilt sich der Redebezug auf verschiedene Figurentypen wie Protagonist oder Antagonist, wie steht es um quantitative Parameter bei uneigentlicher Rede?

ii. Korrelation kulturwissenschaftlich relevanter Terme und aktantieller Rolle. Hier werden bspw. zahlreiche gender-bezogene Auswertungen möglich, indem eine Sample-Analyse mit Figuren- oder Erzählerrede möglich wird, die jeweils auf weibliche oder männliche Figuren bezogen ist oder durch eine Sample-Analyse mit Figurenrede, die jeweils von weiblichen oder männlichen Figuren stammt.

iii. Studien zur Wertungstheorie: Wie sind evaluative Äußerungen auf Erzählerrede und Figurenrede verteilt? Welche Aktanten bewerten bevorzugt, welche werden bevorzugt bewertet?

iv. Lassen sich für diese oder für andere Fragestellungen epochenspezifische Verteilungen ausmachen? Es werden Studien ermöglicht, die einen Beitrag zur Gattungsgeschichte leisten. So wären beispielsweise Theoriebildungen zu überprüfen, inwieweit und inwiefern das Verhältnis vom Märe zur Novelle in Anschluss an die Unterscheidungskriterien von Hans-Jörg Neuschäfer unter dem Gesichtspunkt eines „Noch-Nicht" beschrieben werden kann.[7]

---

7 Vgl. HANS-JÖRG NEUSCHÄFER: Boccaccio und der Beginn der Novelle. Strukturen der Kurzerzählung auf der Schwelle zwischen Mittelalter und Neuzeit. München 1969 (=Theorie und Geschichte der Literatur und der schönen Künste 8). Kritisch dazu FRIEDRICH MICHAEL DIMPEL: Sprech- und Beißwerkzeuge, Kunsthandwerk und Kunst in Kaufringers ‚Rache des Ehemanns'. In: Daphnis 42, 2013, S. 1–27 (im Erscheinen).

corrospSearch. Ein zentraler Service zum Vernetzen von Briefeditionen und -repositorien

Stefan Dumont, Marcel Illetschko, Sabine Seifert, Peter Stadler

Als am 24. Februar 1848 die Revolution in Paris der Herrschaft des ‚Bürgerkönigs' Louis-Philippe von Orléans ein Ende setzte und ihre alles verändernden Wogen über Europa aussandte, schrieb Gustave Flaubert an seine Geliebte Louise Colet, er vergnüge sich „höchlichst bei der Betrachtung all der zunichte gewordenen Ambitionen". Zwar wisse er nicht, ob die neue Form der Regierung und der gesellschaftliche Zustand, der daraus hervorgehen werde, für die Kunst günstig sei. Man könne allerdings kaum bürgerlicher und belangloser als die alte werden. „Und noch dümmer - ist das möglich?"[1] Etwa gleichzeitig meinte Charles Dickens gegenüber seinem Freund, dem Schauspieler William Macready, dass er den neuen französischen Regierungschef für „one of the best fellows in the world" halte und die große Hoffnung hege, „that great people establishing a noble republic."[2] Heinrich Heines Reaktionen auf die republikanischen Umbrüche in Frankreich und im restlichen Europa waren weniger positiv. Seine Mutter ließ er wissen: „Eben weil es jetzt so stürmisch in der Welt und hier besonders tribulant hergeht, kann ich Dir wenig schreiben. Der Spektakel hat mich physisch und moralisch sehr heruntergebracht. Ich bin so entmuthigt, wie ich es nie war. [...] Sollten die Sachen sich hier, wie ich fürchte, noch düsterer gestalten, so gehe ich fort, mit meiner Frau, oder auch allein."[3]

Die Februarrevolution 1848 steht hier als ein willkürlich herausgegriffenes Beispiel mit einer zufällig getroffenen Anzahl an Briefen und Briefeditionen. Dabei wird deutlich, dass Briefe zu den wertvollsten Quellen historischer Forschung zählen. Unterschiedliche Themen und Ereignisse aus der Lebenswelt der Korrespondenten werden angesprochen[4] und soziale Netzwerke abgebildet: „Auf Grund der Tatsache, dass ein Individuum nie nur mit einem einzigen Gegenüber korrespondiert, ist jeder Brief immer auch ein kommunikativer Akt in einem größeren, interpersonalen Zusammenhang. So wie das einzelne Subjekt in seiner Rolle als Schreiber-Adressat Teil eines Beziehungsgeflechts ist, so nimmt jeder einzelne Brief eine Mehrfachposition im kommunikativen Gesamtgefüge epistolaren Austauschs ein [...]".[5] Und so stellen sich Fragen an die Briefnetzwerke, wie z.B.: Welche politischen oder künstlerischen Kreise sind abgrenzbar, wie intensiv ist die Kommunikation innerhalb solcher Zirkel, welche Schlüsselfiguren gibt es, wie ist es um die Kommunikation mit anderen Kreisen bestellt? Etc.

---

[1] Gustave Flaubert. Briefe. Hg. von Helmut Scheffel. Zürich: Diogenes 1977, S. 110.

[2] Letters of Charles Dickens: 1833-1870. Hg. von Georgina Hogarth und Mary Dickens. Cambridge: Cambridge University Press 2011, S. 182.

[3] HSA Bd. 22, S. 270, Brief Nr. 1215, Online-Abruf Heine-Portal, 3.11.2014.

[4] Einschlägig zum Thema etwa Wolfgang Frühwald et al. (Hg.): Probleme der Brief-Edition: Kolloquium der Deutschen Forschungsgemeinschaft. Boppard 1977. – Irmtraut Schmid, Was ist ein Brief? Zur Begriffsbestimmung des Terminus 'Brief' als Bezeichnung einer quellenkundlichen Gattung. In: editio 2 (1988), S. 1–7. – Reinhard Nickisch: Brief. Stuttgart 1991. – Roloff, Hans-Gert (Hrsg.): Wissenschaftliche Briefeditionen und ihre Probleme. Editionswissenschaftliches Symposion. Berlin 1998. - Waltraut Wiethölter: Der Brief – Ereignis und Objekt. Frankfurt am Main 2010. – Anne Bohnenkamp und Elke Richter (Hg.): Brief-Edition im digitalen Zeitalter (=Beihefte zu editio Bd. 34). Berlin/Boston 2013.

[5] Wolfgang Bunzel: Briefnetzwerke der Romantik. Theorie – Praxis – Edition. In: Anne Bohnenkamp und Elke Richter (Hg.): Brief-Edition im digitalen Zeitalter (=Beihefte zu editio Bd. 34) Berlin/Boston 2013. S. 109-131, hier S. 113. - Im Folgenden zitiert als: Bunzel (2013).

Allerdings werden meist nur Briefeditionen *eines* Autors oder *einer* speziellen Korrespondenz (z.B. Verlegerbriefwechsel, Familienbriefwechsel) erstellt. Lediglich in wenigen neueren Projekten sind Briefe im Hinblick auf z. B. ein historisches Thema ediert.[6] Übergreifende Forschungsprojekte, die auf Basis von Korrespondenz(en) größere thematische Zusammenhänge (wie das eingangs genannte Beispiel) oder größere Korrespondenznetzwerke in den Blick nehmen wollten, mussten die Datenbasis meist aufwendig aus vorhandenen gedruckten oder online verfügbaren Editionen zusammentragen und aufbereiten.

Prinzipiell gilt also: „Because of the everyday occurrence of the epistolary exchange, the [...] large stocks of letters render—without the support of computers—a scientific analysis of the correspondence networks spanned within them nearly impossible."[7] Bekannt ist diese Problematik in der deutschsprachigen Literaturwissenschaft seit Langem v.a. in Bezug auf Romantikernetzwerke. So meint etwa Wolfgang Bunzel: „Die schiere Masse der überlieferten Briefe ist selbst für Experten nicht zu überschauen, viel weniger für Forscher, die sich nur mit einzelnen Personen beschäftigen oder an Detailaspekten interessiert sind", was zur Folge habe, „dass schon das gedruckt vorliegende Material nur höchst selektiv benutzt wird und die Mehrzahl der Briefe unzitiert, meist sogar ungesichtet bleibt."[8] Er fordert deshalb

> „die Schaffung einer dezentralen, möglichst offenen, auf (bis auf weiteres) html/xml-Grundlage basierenden und mit TEI-Minimalstandards operierenden digitalen Plattform, die nach vielen Richtungen hin erweiterbar ist und es den bereits bestehenden Portalen und Homepages erlaubt, sich mit denkbar geringem Zusatzaufwand daran zu beteiligen. Nötig ist dafür keine Superstruktur, welche die – ohnehin nicht exakt bezifferbare – Gesamtheit aller Briefe der Romantik überwölbt, sondern vielmehr ein intelligentes Verknüpfungssystem, das vorhandene Dokumente in Konnex zueinander bringt. Mit dem Aufbau eines solchen Nexus gehen natürlich Recherchemöglichkeiten einher, die von der Personen- über die Datums- und Orts- bis hin zur gezielten Stichwortsuche (und dies natürlich in beliebiger Kombination der Suchparameter) reichen."[9]

Wir möchten mit „correspSearch" einen Webservice vorstellen, der einen ersten Schritt in diese Richtung geht, indem er aus verteilten Editionen und Repositorien die Briefmetadaten aggregiert und über offene Schnittstellen, auf TEI-XML-Grundlage sowie unter einer freien Lizenz zentral zur Verfügung stellt. Die Initiative zu diesem Webdienst entstand im Februar 2014 im Workshop „Briefeditionen um 1800: Schnittstellen finden und vernetzen", der von Anne Baillot (Nachwuchsgruppe „Berliner Intellektuelle 1800–1830" an der HU Berlin) und Markus Schnöpf (TELOTA, BBAW) organisiert worden war.

Das Fundament des Webservices ist ein Austauschformat, das auf dem Modul „correspDesc" (Correspondence Description)[10] für die Richtlinien der Text Encoding Initiative[11] basiert. Das Modul

---

[6] So z.B. das Projekt „Letters 1916", http://dh.tcd.ie/letters1916/, das Briefe aus der Zeit um den Osteraufstand 1916 in Irland sammelt und transkribiert, oder das Projekt „Vernetzte Korrespondenzen | Exilnetz33", http://exilnetz33.de/.

[7] Martin Andert, Frank Berger, Paul Molitor and Jörg Ritter: An optimized platform for capturing metadata of historical correspondence. In: Lit Linguist Computing (2014), doi: 10.1093/llc/fqu027

[8] Bunzel (2013), S. 117.

[9] Ebd., S. 123.

[10] https://github.com/TEI-Correspondence-SIG/correspDesc

[11] Burnard, Lou; Bauman, Syd (Hg.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Charlottesville, Virginia, USA 2014. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>

wurde von der TEI Special Interest Group Correspondence[12] entwickelt, um die Metadaten einer einzelnen Korrespondenz (d.h. insbesondere eines Briefes) standardisiert in TEI-XML notieren zu können. Derzeit befindet sich das Modul „correspDesc“ im Antragsverfahren[13]. Auf der Grundlage dieses TEI-Moduls wurde (und wird) ein Austauschformat entwickelt, das nach Abschluss des Prozesses von der TEI SIG Correspondence empfohlen werden soll. Dieses Austauschformat ist ebenfalls in TEI-XML definiert und bietet die Möglichkeit, das Briefverzeichnis einer Edition standardisiert zu notieren – also Absender, Empfänger, Datums- und Ortsangaben eines jeden Briefes. Neben dieser einheitlichen TEI-XML-Kodierung wird der Austausch auch durch die Verwendung von Normdaten ermöglicht.[14] Absender, Empfänger, Schreib- und Empfangsorte werden durch Normidentifikationsnummern (wie z.B. die GND-Nummer der Deutschen Nationalbibliothek[15]) identifiziert. Datumsangaben werden ebenfalls standardisiert erfasst. Dadurch werden die Metadaten eines Briefes über Projekt- und Sprachgrenzen hinweg operabel gemacht. Schließlich wird im Austauschformat auf den einzelnen Brief referenziert. Während es sich bei ausschließlich gedruckten Editionen um die Briefnummer und die bibliografische Angabe handelt, können digitale Editionen zusätzlich die URL hinterlegen. So kann jedes Editionsvorhaben sein Briefverzeichnis digital bereitstellen.

Um das Potential dieser digitalen Briefverzeichnisse zu nutzen, wurde von der TELOTA-Arbeitsgruppe an der BBAW in Zusammenarbeit mit der TEI SIG Correspondence und weiteren Wissenschaftler(inne)n der Webservice „correspSearch” (http://correspSearch.bbaw.de) entwickelt, der digitale Briefverzeichnisse aggregiert und abfragbar bereitstellt. Die Verzeichnisse werden dabei jeweils vom Anbieter unter einer CC-BY-Lizenz[16] vorgehalten und vom Webservice in periodischen Abständen neu bezogen. Jedes Verzeichnis muss also lediglich mit seiner permanenten URL im Webservice registriert werden und wird dann automatisch ausgelesen. Dadurch können weitere Briefverzeichnisse ganz leicht dem Webservice hinzugefügt werden. Beim Einlesen der Verzeichnisse wird für jede Norm-ID eines Korrespondenten bei der Virtual International Authority File[17] nach den gängigsten Norm-IDs gefragt, so dass unterschiedliche Normdatensysteme aufeinander abgebildet werden. Derzeit unterstützt der Webservice GND, VIAF, BNF, LC und NDL. Für die Ortsnamen wird „GeoNames“[18] unterstützt.

Die aggregierten Briefverzeichnisse kann man nun nach Korrespondenzpartner (auf Wunsch eingeschränkt auf dessen Rolle als Absender oder Empfänger), nach Schreibort und Datum durchsuchen. Als Ergebnis werden die Kopfdaten der einzelnen Briefe mitsamt bibliografischer Angaben ausgegeben. Briefe aus digitalen Editionen werden zusätzlich direkt verlinkt. Diese Recherchen können zum einen über eine grafische Benutzeroberfläche ausgeführt werden. Zum anderen wurde auch ein Application Programming Interface (API)[19] implementiert, wodurch man den Webservice automatisiert abfragen kann. Das Ergebnis wird dann ebenfalls unter einer CC-BY-Lizenz im beschriebenen Austauschformat ausgegeben, d.h. als TEI-XML, und kann von Programmen zur Anzeige in der eigenen Webapplikation weiter verwendet werden. Mit Hilfe der

[12] http://www.tei-c.org/Activities/SIG/Correspondence/
[13] Antrag: http://sourceforge.net/p/tei/feature-requests/510/
[14] Zur Verwendung von Normdaten in Editionen vgl. Stadler, Peter: Normdateien in der Edition. In: editio 26, 2012, S. 174-183.
[15] http://www.dnb.de/gnd
[16] Creative Commons Attribution 3.0 Unported: https://creativecommons.org/licenses/by/3.0/de/
[17] http://www.viaf.org
[18] http://www.geonames.org/
[19] http://correspSearch.bbaw.de/api/tei-xml.xql

API können zukünftige digitale Editionen daher auch automatisiert auf verwandte Briefe aus anderen Editionsvorhaben hinweisen oder direkt auf diese verlinken.

Obwohl schon funktionsfähig und frei zugänglich, befindet sich der Webservice noch in der Aufbauphase. Zum einen werden sowohl das Austauschformat als auch der Funktionsumfang noch weiter entwickelt. Zum anderen fällt der Datenbestand im Moment noch recht klein aus. Derzeit werden digitale Verzeichnisse aus drei Editionen ausgewertet: der Weber-Gesamtausgabe[20], der Edition „Briefe und Texte aus dem intellektuellen Berlin um 1800"[21] sowie dem Soemmerring-Briefwechsel 1792–1805[22]. Während letzteres ein retrodigitalisiertes Briefverzeichnis einer gedruckt vorliegenden Edition ist, stammen die ersten beiden Verzeichnisse aus digitalen Editionen und wurden direkt aus deren Datenbestand generiert. Mehrere Editionsvorhaben haben bereits zugesagt, ein digitales Briefverzeichnis bereit zu stellen, so dass der Datenbestand demnächst weiter wachsen wird.

Trotz der Aufbauphase macht der Webservice schon jetzt deutlich, dass er – mit einer vergrößerten Datenbasis – eine wertvolle Ressource für die weitere Forschung sein kann. So aggregiert er Daten, die ansonsten im jeweiligen Projektkontext verbleiben würden. Im Fall der gedruckt vorliegenden Editionen werden Textinformationen überhaupt erstmalig digital aufbereitet und als Daten der maschinellen Verarbeitung zugänglich gemacht. Zudem beschränkt sich correspSearch weder auf einen thematischen noch auf einen zeitlichen Schwerpunkt, so dass die Daten auch für bisher noch nicht entwickelte Forschungsfragen genutzt werden können.[23] Da correspSearch über eine API verfügt und die digitalen Briefverzeichnisse transparent abgelegt sowie frei nachnutzbar sind, können Forscher den Datenbestand auch mit Technologien abfragen, die neuartig sind oder für die der Webservice selbst keine technische Basis bietet. So wird mit einer ausreichenden Datenmenge und einer entsprechenden Software auch die Erforschung von sozialen Netzwerken möglich sein. Darüber hinaus könnten mit einem weiteren angedachten Ausbau des Webservices auch die thematischen Aspekte eines Netzwerkes untersucht werden: Wie diffundieren Themen und politische oder gesellschaftliche Ereignisse durch persönliche Netzwerke? Welche lokalen Zentren gibt es in Bezug auf bestimmte Fragestellungen? Wie werden veröffentlichte Werke oder Zeitschriftenartikel bewertet und diskutiert? Mit correspSearch wurde nun zumindest der Grundstein dazu gelegt, um diese Fragen eines Tages beantworten zu können.

[20] http://www.weber-gesamtausgabe.de
[21] http://tei.ibi.hu-berlin.de/berliner-intellektuelle/
[22] Dumont, Franz (Hrsg.): Samuel Thomas Soemmerring. Briefwechsel November 1792 – April 1805. Basel 2001 (= Samuel Thomas Soemmerring. Werke, begr. v. Gunter Mann, hrsg. v. Jost Benedum u. Werner Friedrich Kümmel, Bd. 20).
[23] Der Schwerpunkt des derzeitigen Datenbestandes auf der ersten Hälfte des 19. Jahrhunderts ist rein zufällig und der Aufbauphase geschuldet.

Abstract: *Topic Modelling* des *Letters of 1916* Briefkorpus

Roman Bleier, Trinity College Dublin, Irland
Email: bleierr@tcd.ie

Einführung und Kontext

Das Jahr 1916 ist in der irischen Geschichte und Erinnerung eng mit dem Unabhängigkeitskampf gegen Grossbritannien verbunden und dem sogenannten Osteraufstandes, oder *Easter Rising*. Das hundertjährige Jubiläum des Osteraufstandes im Jahre 2016 wird national und international große Aufmerksamkeit erregen und mehrere Projekte zur Aufarbeitung der Geschichte des Jahres 1916 versprechen neue Erkenntnisse rund um den Osteraufstand. Das *Letter of 1916: Creating History* Projekt ist eines der umfangreichsten Unternehmen dieser Art.

Das *Letters of 1916: Creating History* Projekt hat als Ziel private Briefe, welche in den Monaten vor und nach dem Osteraufstand geschrieben wurden, zu sammeln und zu digitalisieren. Angestrebt wird ein Korpus von bisher unveröffentlichten, privaten Briefen aufzubauen, das die Studie der Situation in Dublin und Irland aus der Sicht der Bevölkerung ermöglicht. In den Worten der Projektleiterin Professor Schreibman:

> *'Through these letters we will to bring to life to the written word, the last words, the unspoken words and the forgotten words of ordinary people during this formative period in Irish history. All too often our emphasis is on the grand narrative focusing on key political figures. But as we approach the centenary of the Easter Rising we want to try to get a sense of how ordinary people coped with one of the most disruptive periods in contemporary Irish history...'*[1]

Ein weiteres Bestreben des Projekts ist es die Bevölkerung Irlands in das Sammeln und Transkribieren der Briefe einzubinden. Durch die Methode des *Crowdsourcing* wird nicht nur Arbeit ausgelagert, sondern auch Interesse am Projekt geweckt. Es ist das erste Crowdsourcing Projekt dieser Art in Irland und hat als Vorbild das Projekt *Transcribe Bentham* (UCL).

[1] Letters of 1916 Press Release, 27 September 2013, http://dh.tcd.ie/letters1916/wp-content/uploads/2013/09/1916-Letters-Press-Release-27-September-20131.pdf (09.11.2014).

Email: bleierr@tcd.ie

Die Webseite des *Letters of 1916* Projekts wurde im September 2013 eröffnet. Bilder von Briefe können dort hochgeladen, gelesen und transkribiert werden. Im Laufe des vergangenen Jahres sind auf diesem Wege über 1350 Briefe gesammelt worden, und in den nächsten zwei Jahren wird erwartet, dass das Korpus um das Doppelte oder Dreifache wachsen wird. Für 2016 ist geplant, dass das Korpus in Form eines Bild- und Textarchives der Öffentlichkeit zugängig gemacht wird. In Vorbereitung auf diese zweite Phase des Projekts experimentieren die Projektmitarbeiter mit Methoden zur Korpusanalyse und visuellen Veranschaulichung des Korpus. Eine dieser Methoden ist *Topic Modelling*.

Topic Modelling ist eine Technik die seit einigen Jahren in den Digitalen Geisteswissenschaften verwendet wird. Die Methode ist eng verknüpft mit dem Begriff des *Distant Reading* und *Macroanalysis*, dem Bestreben großer Textkorpora durch automatisierte Analyse und visuelle Aufbereitung mittels Diagrammen und Graphen Herr zu werden.[2] Vereinfacht ausgedrückt wird bei *Topic Modelling* versucht computergenerierte Themen aus einem Korpus von (meist) Textdokumenten zu gewinnen. Die einzelnen Dokumente des Korpus können je nach ihrer probabilistischen Nähe einem oder mehreren dieser Themen zugeordnet werden. In den Digitalen Geisteswissenschaften wird vor allem das LDA Model von Blei, Ng und Jordan verwendet, da es standardmäßig in Tools wie Mallet oder Gensim[3] implementiert ist.[4]

Das *Letters of 1916* Korpus wurde bisher zweimal durch *Topic Modelling* untersucht. Im Juni 2014 wurde eine erste Untersuchtung mit knapp 700 Briefen durchgeführt. Das Verfahren wurde im Januar 2015 mit 1350 Briefen wiederholt. Zur automatisierten Generierung von Themen wurde das *Topic Modelling Tool Mallet* verwendet. Sechzehn computergenerierte Themen wurden erstellt und die Briefe im Korpus je nach ihrer probabilistisch Nähe zu den einzelnen Themen in einem Graphen eingezeichnet. Für die graphische Aufbereitung wurde das Programm Gephi verwendet. Ziel dieser periodischen Untersuchung ist es, vor allem die Korpuszusammensetzung und Korpusentwicklung zu studieren und zu dokumentieren.

---

[2] Besonders Franco Moretti und Mat Jockers haben in den letzten Jahren zur Verbreitung dieser Methode in den Digital Humanities beigetragen. Franco Moretti, Distant Reading (2013). Mat Jockers, Macroanalysis: Digital Methods and Literary History (2013).

[3] Mallet: MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/ (09.11.2014); Gensim: topic modelling for humans, http://radimrehurek.com/gensim/ (09.11.2014).

[4] Grundlegender Artikel zu LDA ist: Blei, et al., Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022, http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf (09.11.2014).

Email: bleierr@tcd.ie

Beim Hochladen eines Briefes müssen bestimmte Metadaten angegeben werden. Name und Geschlecht des Absenders, das Datum des Briefes, und eine von sechzehn Kategorien, in die der Brief eingeordnet werden kann. Diese Kategorien sind von den Editoren zu Beginn des Projekts festgelegt worden und beinhalten Themen wie: Easter Rising, World War 1, Official Documents, Love Letters, Family Life, etc. Die von Menschenhand zugeordneten Themen sind besonders interessante Metadaten, da sie einen Vergleich mit den computer-generierten Themen ermöglichen.

Mein Vortrag wird einen Überblick über das Letters of 1916 Projekt und die Untersuchung des Korpus durch *Topic Modelling* bieten. Die Resultate der Untersuchen werden vorgestellt, und etwaige Probleme, die bei der Korpusreinigung und der Datenanalyse aufgetreten sind, werden diskutiert. Die Untersuchen des Teilkorpus soll feststellen, inwieweit eine Lesung des entgültigen Briefkorpus durch *Topic Modelling* möglich und sinnvoll ist und welche Erkenntnisse dadurch gewonnen werden können.

VORTRAG / ABSTRACT

Claudia Resch, Daniel Schopper, Barbara Krautgartner
Österreichische Akademie der Wissenschaften / Austrian Centre for Digital Humanities

Von A wie *Abraham a Sancta Clara* bis U wie *Unbekannter Verfasser* Annotation und Repräsentation barocker Literatur

Der bekannte Prediger und Augustinermönch Abraham a Sancta Clara (1644-1709) gilt in der Literaturgeschichtsschreibung als einer der sprachmächtigsten Autoren seiner Zeit. Dass seine Schriften bisher in keiner Gesamtausgabe vorliegen, ist darauf zurückzuführen, dass der Umfang seines Œuvres nicht feststeht und sich seine Autorschaft besonders im Spätwerk verunklart: „Je leuchtender der Name und Stil des populären Autors in Erscheinung trat, desto fragwürdiger wurde der Bezug des realen Ordensmannes und Schriftsteller zu seinem Werk."[1]

Das Projekt[2], das im Vortrag vorgestellt werden soll, hat daher im Rahmen des Vorhabens *ABaC:us – Austrian Baroque Corpus* ein digitales Korpus erstellt, das u.a. eine Auswahl von Abraham a Sancta Clara zugeschriebenen Texten in ihren ersten zu identifizierenden Ausgaben enthält. Es eröffnet der Forschung damit einen neuen, unvoreingenommenen Blick auf diese Quellen[3] und verfolgt folgende Ziele:

1 Erstellung einer verlässlich annotierten Textgrundlage

Die diplomatischen Transkriptionen der Frakturdrucke (u.a. von *Mercks Wienn, Lösch Wienn, Grosse Todtenbruderschaft, Augustini Feurigs Hertz* und *Besonders meubliert- und gezierte Todtencapelle*) wurden im XML-Format erstellt, folgen dem international empfohlenen de-facto Standard der Text Encoding Initiative (Version P5) und bilden den historischen Sprachstand der Texte unverändert, d.h. zeichen-, zeilen- und seitengetreu ab. Die linguistische Annotation der historischen Texte im Umfang von 200.000 Token erfolgte semi-automatisch, d.h. die Wortklassenzuordnung[4] und Lemmatisierung[5] auf Basis des *Treetagger* (der für die deutsche Gegenwartssprache entwickelt wurde und daher mit den graphematischen Varianten, wie sie in Barocktexten vorkommen, größtenteils nicht umgehen

[1] Franz M. Eybl: Wissenslücken um Abraham a Sancta Clara – Zur Problematik populärer Autorschaft. In: Unterhaltender Prediger und gelehrter Stofflieferant. Abraham a Sancta Clara. Eggingen 2012, S. 104.

[2] Das Projekt „Texttechnologische Methoden zur Analyse österreichischer Barockliteratur" (Laufzeit 2012-2014) wird durch den Jubiläumsfonds der Österreichischen Nationalbank, Projektnummer: 14738 gefördert.

[3] Die Vielzahl von populären Auswahlausgaben oder "Blütenlesen", die zu Abraham a Sancta Clara bis heute publiziert werden, hat sich für wissenschaftliche Fragestellungen als unzureichend erwiesen.

[4] Die Wortklassenzuordnung erfolgte auf Basis des 54-teiligen Stuttgart-Tübingen TagSets (STTS), das für die historische Sprachstufe des Älteren Neuhochdeutsch adaptiert und um weitere Tags ergänzt wurde, etwa bei kontrahierten Formen wie *wirstu* (VVFIN_PPER) oder *mans* (PIS_PPER).

[5] Als Referenzwerke für die Lemmatisierung wurden der „Duden" sowie das „Deutsche Wörterbuch" von Jacob und Wilhelm Grimm herangezogen bzw. für lateinische Belege das „Lateinisch-deutsche Schulwörterbuch von Stowasser". Sogenannte „out-of-vocabulary words", die in keinem der genannten Wörterbücher vorkommen, tragen einen entsprechenden Vermerk.

konnte[6]) wurden durchgehend von zwei AnnotatorInnen mit Hilfe des an der Österreichischen Akademie der Wissenschaften entwickelten *token_editor* verifiziert beziehungsweise manuell korrigiert.

2 Verbesserung des automatischen Taggings historischer Texte

Die bereits annotierten Daten der ersten Texte wurden in der Folge dafür verwendet, die Erfolgsquote des automatisch generierten Taggings weiterer Texte zu verbessern, indem die bereits identifizierten und systematisierbaren Fehleinträge der ersten Werke in den noch nicht annotierten Textstrecken berücksichtigt wurden. Das verlässlich vollannotierte Korpus könnte in Zukunft maßgeblich dazu beitragen, weitere Texte aus dieser Zeitperiode effizienter zu annotieren bzw. die Leistung verschiedener Tagger daran zu messen.

3 Erweiterung des Wissens über abrahamische Spezifika

Mit dem Aufbau der multifunktionalen Sprachressource hat die Projektgruppe Voraussetzungen für sprach- und literaturwissenschaftliche Fragestellungen geschaffen, die sie selbst exemplarisch beantwortet: Erstmals ist man in der Lage, den Wortschatz des Autors, dessen „Sprachmächtigkeit“ in Literaturgeschichten ausdrücklich gewürdigt wird, zu fassen: Mit Textanalysetools wie der *Sketch Engine* oder *Voyant* können systematische Wortschatzanalysen, Konkordanzen und Type-Token-Relationen erstellt, Frequenzen, Kollokationsprofile und Wortklassenverteilungen ermittelt, sowie musterbasierte Abfragen generiert werden, wodurch sich ausgewählte sprachliche Phänomene und barocke musterhafte, stilbildende Elemente (wie Doppelformeln, Wiederholungs-, Häufungs- und Steigerungserscheinungen) in abrahamischen Texten identifizieren und quantifizieren lassen. „Abrahamischen Stil“ anhand ausgewählter Beispiele zu beschreiben, war bereits das Anliegen einiger älterer Untersuchungen[7], deren Ergebnisse nun überprüft, auf größere Textmengen bezogen, systematisch ausgewertet und damit auf ein begründetes empirisches Fundament gestellt werden können. Was die bis heute ungeklärte Autorenschaft des angeblich letzten Werks von Abraham a Sancta Clara angeht, möchte der Vortrag erstmals eine Reihe von Argumenten vorbringen, die für bzw. gegen eine tentative Zuschreibung sprechen.

---

[6] Dass die Texte von Abraham a Sancta Clara im Vergleich mit anderen (sogar älteren) Texten die höchste Fehlerquote aufwiesen, haben Erhard Hinrichs und Thomas Zastrow bereits festgestellt – vgl. ihre Studie „Linguistic Annotations for a Diachronic Corpus of German“. In: Linguistic Issues in Language Technology, Volume 7 (2012), S. 1-16, hier S. 11.

[7] Vgl. etwa Curt Blanckenburg: Die Sprache Abrahams a S. Clara. Ein Beitrag zur Geschichte der deutschen Drucksprache. Halle an der Saale: Ehrhardt Karras 1897; Hans Strigl: Einiges über die Sprache des P. Abraham a Sancta Clara. In: Zeitschrift für Deutsche Wortforschung. Hrsg. v. Friedrich Kluge. Band 8 (1906), S. 206-312; Margaretha Stiassny: Das Wortspiel bei Abraham a Sancta Clara. Phil. Diss. Wien 1939 und Norbert Bachleitner: Form und Funktion der Verseinlagen bei Abraham a Sancta Clara. Frankfurt am Main / Bern / New York: Peter Lang 1985.

4 Publikation in einem webbasierten Interface und Integration in europäische Forschungsinfrastrukturen

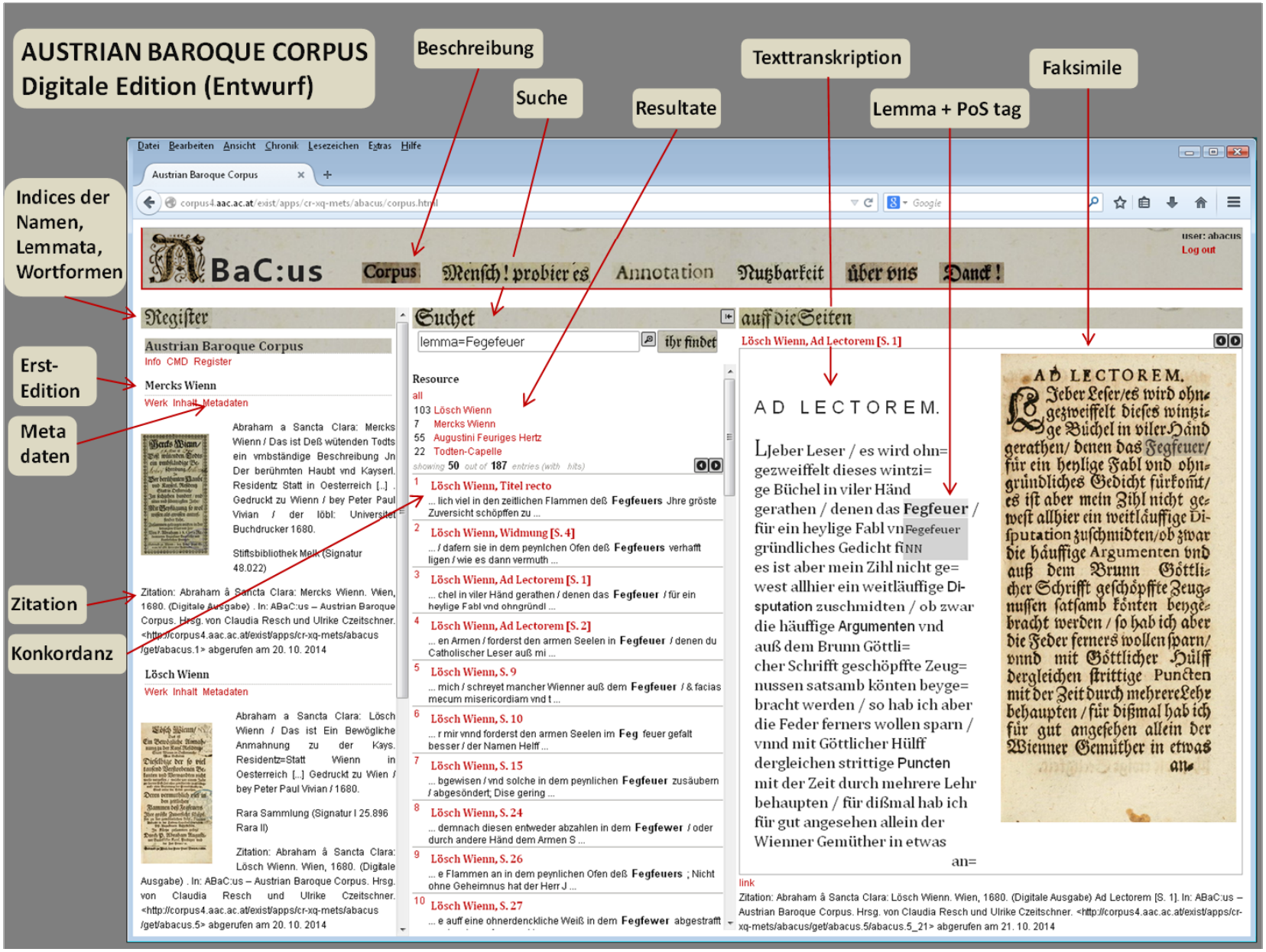
Das vorläufige Ergebnis des Projektes stellt die Integration von fünf abrahamischen Texten in ein webbasiertes Interface dar, das das Korpus über unterschiedliche Wege des Zugriffs nutzbar macht (vgl. Abbildung). Besonders im Zusammenhang mit zeitentfernten Texten wie diesen ist die Frage nach der Art und Weise ihrer Repräsentation im digitalen Kontext eine wesentliche. Das trifft auf das vorliegende Projekt besonders zu, als es sich zum erklärten Ziel gesetzt hat, die Texte nicht nur anderen, textbezogenen Wissenschaften und der universitären Lehre zur Verfügung zu stellen, sondern sie auch einer interessierten nicht-fachlichen Öffentlichkeit näherzubringen.

Wert gelegt wurde insbesondere darauf, die Texte als Werke in ihrer Eigenständigkeit zugänglich zu machen, ohne jedoch das Korpus als ihren Verbund in den Hintergrund zu rücken. Daher ermöglicht das Interface zum einen den „lesenden" Zugang zum Material über eine seitenweise synoptische Ansicht von Volltext und Faksimile, die durch Inhaltsverzeichnisse, Personennamen- und Ortsregister erschlossen ist, und auch die Einbeziehung von graphischen und typographischen Eigenschaften der Drucke bei der Textinterpretation ermöglicht. Zum anderen bietet die Plattform auch die Möglichkeit, *ABaC:us* auf Wortformen, Lemmata und Part-of-Speech-Tags einzeln und in Kombination zu durchsuchen sowie Frequenzlisten zu generieren.

Die Funktionalität des digitalen Suchens und Navigierens basiert technisch auf einfachen XPath-Pfadausdrücken, die zur Adressierung von Dokumentteilen dienen. Eine frei konfigurierbare Gruppe solcher Indizes bildet beispielsweise das Gerüst des hierarchischen Inhaltsverzeichnisses oder der synoptischen Ansicht, die dynamisch aus den Gesamttexten extrahiert wird. Mit dem vorliegenden semantischen Markup der Texte wäre es etwa ein Leichtes, mittels eines entsprechenden Pfadausdruckes ein Register von Bibelstellen zu erstellen. Somit ist das Interface für die flexible Erweiterung des Korpus sowohl hinsichtlich seines Umfangs als auch seiner Funktionalität ausgerichtet und ermöglicht, das in den Daten kodierte Wissen über die Texte entsprechend sich verändernden Anforderungen nutzbar zu machen.

Die Weboberfläche von *ABaC:us* setzt auf dem am ACDH entwickelten *cr-xq Content Repository* auf, das ein Teil des modularen *corpus_shell* Frameworks[8] ist und auf dem *CLARIN Federated Content Search*-Standard[9] basiert. Durch die Einbettung in eine wachsende europäische Infrastruktur an Forschungsdaten ist die Verfügbarkeit der Projektdaten auch in Zukunft gewährleistet, so dass künftige NutzerInnen der Korpusdaten die Annotationen zeitsparend, zweckmäßig und gewinnbringend für ihre Erkenntnisinteressen einsetzen können.

[8] http://www.oeaw.ac.at/icltt/node/4
[9] http://www.clarin.eu/content/federated-content-search-clarin-fcs


AUSTRIAN BAROQUE CORPUS
Digitale Edition (Entwurf)
Beschreibung
Suche
Resultate
Texttranskription
Lemma + PoS tag
Faksimile
Indices der Namen, Lemmata, Wortformen
Erst-Edition
Meta daten
Zitation
Konkordanz
Austrian Baroque Corpus
corpus4.aac.ac.at/exist/apps/cr-xq-mets/abacus/corpus.html
ABaC:us
Corpus
Mensch! probier es
Annotation
Nutzbarkeit
über uns
Danck!
user: abacus
Log out
Register
Austrian Baroque Corpus
Info CMD Register
Mercks Wienn
Werk Inhalt Metadaten
Abraham a Sancta Clara: Mercks Wienn / Das ist Deß wütenden Todts ein vmbständige Beschreibung Jn Der berühmten Haubt vnd Kayserl. Residentz Statt in Oesterreich [..] . Gedruckt zu Wienn / bey Peter Paul Vivian / der löbl: Universitet Buchdrucker 1680.
Stiftsbibliothek Melk (Signatur 48.022)
Zitation: Abraham â Sancta Clara: Mercks Wienn. Wien, 1680. (Digitale Ausgabe) . In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <http://corpus4.aac.ac.at/exist/apps/cr-xq-mets/abacus/get/abacus.1> abgerufen am 20. 10. 2014
Lösch Wienn
Werk Inhalt Metadaten
Abraham a Sancta Clara: Lösch Wienn / Das ist Ein Bewögliche Anmahnung zu der Kays. Residentz=Statt Wienn in Oesterreich [..] Gedruckt zu Wien / bey Peter Paul Vivian / 1680.
Rara Sammlung (Signatur I 25.896 Rara II)
Zitation: Abraham â Sancta Clara: Lösch Wienn. Wien, 1680. (Digitale Ausgabe) . In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <http://corpus4.aac.ac.at/exist/apps/cr-xq-mets/abacus/get/abacus.5> abgerufen am 20. 10. 2014
Suchet
lemma=Fegefeuer
ihr findet
Resource
all
103 Lösch Wienn
7 Mercks Wienn
55 Augustini Feuriges Hertz
22 Todten-Capelle
showing 50 out of 187 entries (with hits)
1 Lösch Wienn, Titel recto
... lich viel in den zeitlichen Flammen deß Fegfeuers Jhre gröste Zuversicht schöpffen zu ...
2 Lösch Wienn, Widmung [S. 4]
... / dafern sie in dem peynlchen Ofen deß Fegfeuers verhafft ligen / wie es dann vermuth ...
3 Lösch Wienn, Ad Lectorem [S. 1]
... chel in viler Händ gerathen / denen das Fegfeuer / für ein heylige Fabl vnd ohngründl ...
4 Lösch Wienn, Ad Lectorem [S. 2]
... en Armen / forderst den armen Seelen in Fegfeuer / denen du Catholischer Leser auß mi ...
5 Lösch Wienn, S. 9
... mich / schreyet mancher Wienner auß dem Fegfeuer / & facias mecum misericordiam vnd t ...
6 Lösch Wienn, S. 10
... r mir vnnd forderst den armen Seelen im Feg feuer gefalt besser / der Namen Helff ...
7 Lösch Wienn, S. 15
... bgewisen / vnd solche in dem peynlichen Fegfeuer zusäubern / abgesöndert; Dise gering ...
8 Lösch Wienn, S. 24
... demnach diesen entweder abzahlen in dem Fegfewer / oder durch andere Händ dem Armen S ...
9 Lösch Wienn, S. 26
... e Flammen an in dem peynlichen Ofen deß Fegfeuers ; Nicht ohne Geheimnus hat der Herr J ...
10 Lösch Wienn, S. 27
... e auff eine ohnerdenckliche Weiß in dem Fegfewer abgestrafft
auf die Seiten
Lösch Wienn, Ad Lectorem [S. 1]
AD LECTOREM.
LJeber Leser / es wird ohn= gezweiffelt dieses wintzi= ge Büchel in viler Händ gerathen / denen das Fegfeuer / für ein heylige Fabl vnFegefeuer gründliches Gedicht fiNN es ist aber mein Zihl nicht ge= west allhier ein weitläuffige Di= sputation zuschmidten / ob zwar die häuffige Argumenten vnd auß dem Brunn Göttli= cher Schrifft geschöpffte Zeug= nussen satsamb könten beyge= bracht werden / so hab ich aber die Feder ferners wollen sparn / vnnd mit Göttlicher Hülff dergleichen strittige Puncten mit der Zeit durch mehrere Lehr behaupten / für dißmal hab ich für gut angesehen allein der Wienner Gemüther in etwas an=
link
Zitation: Abraham â Sancta Clara: Lösch Wienn. Wien, 1680. (Digitale Ausgabe) Ad Lectorem [S. 1]. In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <http://corpus4.aac.ac.at/exist/apps/cr-xq-mets/abacus/get/abacus.5/abacus.5_21> abgerufen am 21. 10. 2014

Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind

Max Hadersbeck, Alois Pichler, Florian Fink, Daniel Bruder, Ina Arends
Maximilian.Hadersbeck@lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München,
Wittgenstein Archives at the University of Bergen (WAB).

1 Einleitung

In dem Vortrag berichten wir über Erfahrungen, Erkenntnisse und Erweiterungen unserer schon seit 2 Jahren im Einsatz befindlichen FinderApp WiTTFind, die mit Hilfe von computerlinguistischen Verfahren den Open Access zugänglichen Teil des Nachlasses von Ludwig Wittgenstein (Wittgenstein Source, 2009) nach Wörtern, Phrasen, Sätzen und semantischen Begriffen im „Zusammenhang des Satzes“[1] durchsucht.

Im Sommer 2014 gewannen wir mit WiTTFind den EU-AWARD, der vom EU-Projekt Digitised Manuscripts to Europeana (DM2E) ausgeschrieben wurde, verbunden mit der expliziten Aufforderung zur Öffnung unseres Finders für andere Projekte der Digital Humanities. Darauf hin entwarfen wir in der disziplinübergreifenden Wittgenstein Sommerschule am CIS im Juni 2014 und in Diskussionen mit Fachleuten der Philosophie und Digital Humanities Verbesserungsmöglichkeiten, die mittlerweile in der neuen Version implementiert sind. Die Weboberfläche unseres Finders wurde optimiert, („rich-client“), jetzt können mehrere Dokumente parallel durchsucht werden, eine lemmatisierte symmetrische Vorschlagssuche und ein Faksimile E-Reader sind integriert. Der Faksimile E-Reader erlaubt es nun, dass die Faksimiles der Edition durchblättert und gefundene Textstellen automatisch visuell hervorgehoben werden. Neben den Weiterentwicklungen der FinderApp setzten die Wittgensteinforscher unseren Finder für semantische Untersuchungen ein und gewannen aus dieser Arbeit wichtige Erkenntnisse z.B. zum Thema des Verstehens in Wittgensteins Big Typescript.[2]

Der wichtigste Mehrwert unseres Finders besteht allerdings darin, dass wir die vom EU-AWARD geforderte Öffnung unseres Finders für andere Projekt konsequent umsetzten. Für die Texte der Edition, die unser Finder durchsucht, gibt es eine XML-TEI P5 kompatible Document Type Definition (DTD). Die Programme, Faksimile E-Reader und Tools sind unter der Bezeichnung „Wittgenstein Advanced Search Tools“ (WAST) in einem „docker“-Softwarecontainer zusammengefasst und werden „open source“ verfügbar sein. Somit ist unsere FinderApp mit ihren WAST-Tools in anderen Projekten der Digital Humanities einsetzbar.

Die folgende Abbildung zeigt eine Suchanfrage an unseren Finder WiTTFind:

http://wittfind.cis.uni-muenchen.de:

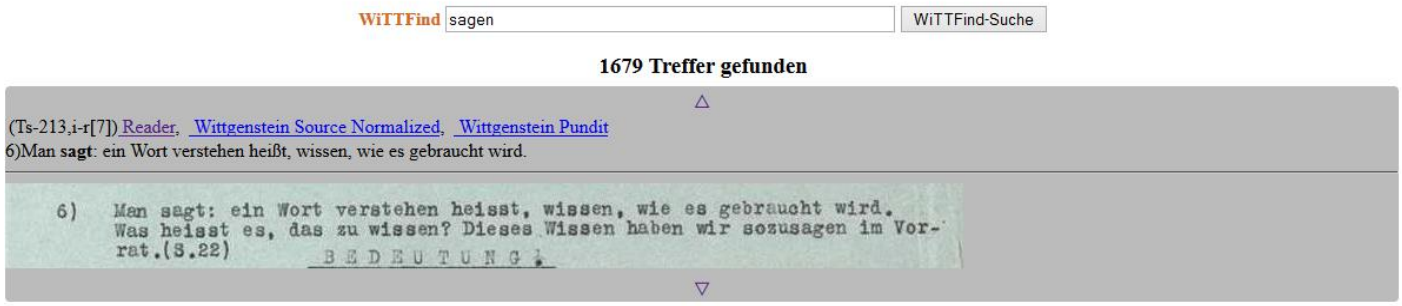


Bild 1: Suchanfrage bei WiTTFind

[1] http://www.wittgensteinsource.org/Ts-213,1r[4]_n

[2] http://www.wittgensteinsource.org/Ts-213_n

2 Erkenntnisse aus der Zusammenarbeit Computerlinguistik und Philosophie

2.1 Verbesserte Benutzeroberfläche unseres Finders

Eine der ersten Erkenntnisse unserer Zusammenarbeit war, dass die Benutzeroberfläche unserer FinderApp auf die Bedürfnisse der jeweiligen Forschergruppe abgestimmt sein muss: die Forscher sollen sich auf der Webseite „wiederfinden“. Nur dann ist die Einstiegshürde nicht zu hoch, und die Bereitschaft mit dem Finder zu arbeiten steigt. Erst für fortgeschrittene Benutzer werden in einer tieferen Schicht globale Einstellungsmenus sichtbar und spezielle Parameter einstellbar. Als Kompromiss zwischen Komplexität und gewohnter Suchmaschinenarbeit können die Nutzer verschiedene Suchumgebungen auswählen (siehe Bild 1): „Regelbasiertes Finden“, „Semantisches Finden“, „Graphisches Finden“, „Statistische Suche“ und „Geheimschriftübersetzer“.

Bild 2: Suchumgebungen bei WiTTFind

Damit die zahlreichen Suchmöglichkeiten bei WiTTFind auf einen Blick sichtbar sind, programmierten wir fachspezifische Hilfeseiten mit Beispielen:

Bild 3: Hilfeseiten bei WiTTFind

2.2 Video-Tutorials zur Nutzung von WiTTFind

Zum erleichterten Einstieg bei WiTTFind gibt es jetzt zwei Video-Tutorials in deutscher und englischer Sprache unter folgendem Link:

http://wittfind.cis.uni-muenchen.de/tutorial

2.3 E-Reader für die Faksimile

Gerade bei komplexen Editionen mit vielen handschriftlichen Einfügungen und Streichungen, wie der des Nachlasses von Ludwig Wittgenstein, ist es für die Editionswissenschaftler eine Herausforderung, den Editionstext in der niedergeschriebenen Form als HTML-Text in einem Browser darzustellen. In der neuen Version unserer FinderApp programmierten wir einen eigenen Faksimile E-Reader, der es erlaubt, komplementär durch die Faksimile der Edition zu blättern und gleichzeitig die gefundenen Textstellen im Bild hervorhebt.

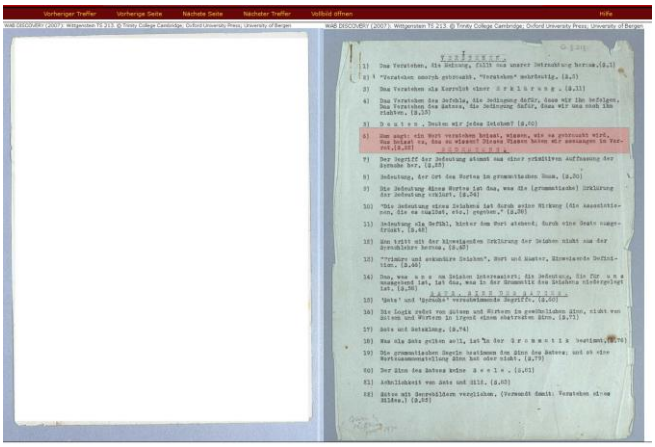


Bild 4: Faksimile Reader bei WiTTFind

2.4 Lemmatisierte Vorschlagssuche mit statistischen Angaben

Die Arbeit mit WiTTFind zeigte, dass eine komfortable Vorschlagssuche, die den gesamten Wortindex der Edition mit Frequenzlisten im Hintergrund hält, einen sehr guten Einstieg in die eigentliche Suche darstellt. Hierhin zielt unsere neueste Erweiterung von WiTTFind, eine komfortable Index-Suchfunktion, die auf einen symmetrischen Suchindex basiert. Dieser Index greift auf Einträge des zugrunde liegenden Lexikons und Wort-Frequenzlisten der Texte der Edition zurück. Dem Anwender werden nach Eingabe von wenigen Buchstaben alle Wörter mit der Häufigkeit des Auftretens im Text automatisch aufgezeigt, in denen die eingegebenen Buchstaben vorkommen; dazu werden auch noch die morphologischen Varianten dieser Wörter angezeigt. Diese Art der Autovervollständigung ist eine völlig

neue Technologie, da bisherige Autovervollständigungen die eingegebenen Buchstaben nur um die Wörter ergänzen, die mit diesen Buchstaben beginnen.

3 Von Daten zu Erkenntnissen

3.1 Semantisches Suchen: Wortfelder

Ein großes Problem semantischer Untersuchungen mit Wortfeldern stellt die Disambiguierung der Wortfeldbegriffe dar. Mit Hilfe unseres elektronischen Lexikons, der syntaktischen und semantischen Disambiguierung über Part of Speech Tagging und lokale Grammatiken können neben Einzelwörter auch Wortphrasen einem Wortfeld zugeordnet und disambiguiert werden.
Ein einfaches Beispiel wurde um das semantische Feld von "Verstehen" ausgearbeitet. Welches Interesse an Verstehen hat Wittgenstein im Big Typescript? Eine Suche nach *<N> verstehen* [Substantiv + „verstehen“] im Big Typescript ergibt, dass dort ganz klar das Verstehen von Wörtern, Sätzen, Sprachen, Befehlen ... allgemein: das Verstehen von sprachlichen Zeichen, im Vordergrund steht. Daneben gibt es aber auch bereits eine gewisse Aufmerksamkeit auf das Verstehen von Menschen und Menschlichem: von Handlungen, Gebärden, Gesten. Diese Aufmerksamkeit nimmt in Wittgensteins Spätwerk beständig zu, was eine Suche nach *<HUM> verstehen* [Substantiv für Menschliches + „verstehen“] bestätigt.
Ein zweites, komplexeres Beispiel wurde um das semantische Feld von "Grammatik" ausgearbeitet. Zuerst baten wir Wittgensteinexperten, uns eine Liste von 10-15 Wörtern zu geben, welche ihrer Ansicht nach im Wortfeld von "Grammatik" zentral sind. Dazu gehören z.B. "Anwendung", "Regel", "Kalkül" und "System". Daraufhin wurden diese Wörter im Lexikon über den Begriff "Grammatik" vernetzt. Eine WiTTFind-Suche nach *Grammatik* wird dann nicht nur Stellen mit "Grammatik" ergeben können, sondern auch Bemerkungen, welche eine Bündelung von Begriffen aus dem Wortfeld aufweisen. Erste Anwendungen ergaben, dass Wittgenstein im Big Typescript tatsächlich einen regelfixierten Begriff von Grammatik verfolgt, während dieser Aspekt später abgeschwächt werden wird (vgl. Szeltner 2013).

4 Synergien: Unsere FinderApp für andere Digital Humanities Projekte

4.1 Vorbemerkung

Wie vom DM2E Projekt bei der Preisverleihung gefordert, öffneten wir unsere FinderApp für andere Projekte der Digital Humanities. Editionsprojekte müssen ihre Dokumente in unser reduziertes XML-TEI P5 Format (CISWAB) konvertieren und die Open-Source Software *docker*[3] auf ihrem Rechner installieren. Dann können sie unseren Finder bei ihren Editionstexten anwenden. Zur Darstellung und Highlighting der Treffer im Faksimile sind allerdings umfangreiche OCR-Arbeiten notwendig. In den nächsten Unterkapiteln beschreiben wir im Detail, wie unser Finder einsetzbar wird.

4.2 Die Texte der Edition

Unsere FinderApp findet Wörter, semantische Begriffe und Satzphrasen über mehrere Dokumente hinweg, sofern die Dokumente in unserem XML-TEI-P5 Format vorliegen. Wir nennen dieses XML-Format CISWAB und beschreiben es in einer eigenen Document Type Definition (DTD). Die einzelnen Dokumente sind bis auf Satzebene über Siglen eindeutig zu spezifizieren:
(z.B. `<s n="Ts-213,i-r[7]_1" ana="facs:Ts-213,i-r abnr:7 satznr:15">6)Man sagt: ein Wort verstehen heißt, wissen, wie es gebraucht wird.</s>`)

4.3 Elektronisches Vollformenlexikon

Zu den Texten einer Edition benötigt unsere FinderApp ein elektronisches Lexikon im DELA Format (Laboratoire d'Automatique Documentaire et Linguistique, Paris). Bei der Erstellung des Lexikons können wir behilflich sein, da wir am CIS das größte deutsche Vollformenlexikon erstellt haben.

4.4 Syntaktische Disambiguierung: Part of Speech Tagging

Grundvoraussetzung für die syntaktische Disambiguierung ist es, dass die Texte mit einem Part of Speech Tagger bearbeitet werden. Zu unseren WAST-Tools gehört das automatische Taggen der Texte. Dazu verwenden wir den *treetagger* von Dr. Helmut Schmid, der am CIS entwickelt wird. Der *treetagger* konvertiert

[3] siehe: https://www.docker.com/

die Textdatei in eine getaggte XML Datei, die die Eingabedatei für unsere FinderApp darstellt.

4.5 Darstellung der Treffer im Faksimile Reader

Um die Treffer in unserem Faksimile-Reader darzustellen, müssen die Faksimile mit der open source Software *tesseract* bearbeitet werden, und je nach Qualität der Faksimiles manuell nachbearbeitet werden. Wir entwickelten Tools, die diese manuelle Arbeit erleichtern.

4.6 Praktische Voraussetzung zur Verwendung unserer FinderApp

Wir haben unser Ziel, dass die FinderApp WiTTFind und die WAST-Tools möglichst auf jedem Rechner lauffähig sind, erreicht. Mit Hilfe der neuesten Open Source Software Technologie *docker* werden die unterschiedlichen Programmiersprachen und Libraries, die wir einsetzen, in einem Softwarecontainer, genannt WAST-dockerimage, zusammengefasst. Jeder Anwender, der auf seinem Rechner die *docker*-Serversoftware installiert hat, kann das WAST-dockerimage herunterladen und virtualisiert läuft die FinderApp WiTTFind unter dem Dockerserver auf dem Rechner. Die Dockerserversoftware funktioniert nahezu unter jedem Betriebssystem (Linux, Windows, MACOS).

4.7 Vorstellung und Vorführung unseres Finders auf der Tagung

Neben diesem Vortrag wollen wir auf der Tagung in einem Poster den Aufbau und den Einsatz der FinderApp WiTTFind als Open Source Tool vorstellen: Die optimierte Browseroberfläche, zugrunde liegende Texte der FinderApp, Faksimile mit OCR, Faksimile Reader und den Einsatz des Finders als Open Source Programm. Für Interessierte wird die FinderApp unter verschiedenen Betriebssystemen an Laptops vorgeführt.

5 EU-AWARD und Publikationen

EU AWARD 2014: http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal: Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST). Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014) Madrid: 91-96

Szeltner, Sarah: 'Grammar' in the Brown Book. Papers of the 36th International Ludwig Wittgenstein-Symposium, vol 21. Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society; 2013.

Wittgenstein Source: Bergen Text and Facsimile Edition. In: Pichler A., collaboration with, Krüger H.W., Lindebjerg A., Smith D.C.P., BruvikT.M., Olstad V., editors. Bergen: Wittgenstein Archives at the University of Bergen; 2009. http://www.wittgensteinsource.org/

Klaus Kastberger/Katharina Pektor: Information/Kommentar/Interpretation: Handke*online*

Einreichung zu einem Vortrag für die DHd-Konferenz 2015 (Graz 23. bis 27. 2. 2015)

In einem Vortrag soll das Projekt www.handkeonline.onb.ac.at als eine paradigmatische Anwendung der DH aus dem Bereich Literaturwissenschaft vorgestellt und hinsichtlich der allgemeinen Fragen spezifiziert werden, die die Konferenz stellt. Dies betrifft vor allem auch die Kombination aus Datenanzeige, Kommentierung und Interpretation, die die Website leistet:

I. Kurzbeschreibung

Die Website www.handkeonline.onb.ac.at schafft einen schnellen und unkomplizierten Zugang zu den Werkmaterialien des österreichischen Autors Peter Handke. Nach Art eines kommentierten digitalen Archivs werden Bestände aus öffentlichen und privaten Sammlungen verzeichnet, aufeinander bezogen, inhaltlich beschrieben und durch zahlreiche Abbildungen anschaulich gemacht. Die Seite ist frei zugänglich und bietet dem interessierten Lesepublikum eine spezifische Hinführung zum Werk des Autors und tiefe Einblicke in die ihm zugrundeliegende Arbeitsweise. Für materialzentrierte Forschungen der Literatur- und Kulturwissenschaft hält die Seite viele neue Ansatzpunkte bereit. Ausgewählte Werkfassungen und einige Notizbücher Peter Handkes werden als Gesamtfaksimiles erstveröffentlicht. Eine integrierte Open-Access-Plattform macht aktuelle Ergebnisse der internationalen Handke-Forschung frei zugänglich.

II. Projektinhalt

Das Projekt www.handkeonline.onb.ac.at weist die werkgenetischen Materialien, die zu Peter Handkes Büchern bis hin zu dem 2013 erschienenen Band *Versuch über den Pilznarren* vorliegen, laut den „Regeln zur Erschließung von Nachlässen und Autographen“ (RNA) in ihrer Gesamtheit nach. Die verstreuten Einzelmaterialien (frühe Notizen, Fassungen in Handschriften und Typoskripten, Quellen wie annotierte Bücher, Landkarten und Fotos) werden zu werkgenetischen Konvoluten zusammengestellt und in anschaulicher Weise präsentiert. Die inhaltliche Beschreibung der Konvolute umfasst eine Darlegung des jeweiligen Entstehungskontextes und eine ausführliche werkgenetische Kommentierung der Materialien. Das umfangreiche Korpus von Handkes unveröffentlichten Notizbüchern findet dabei punktuelle Berücksichtigung. Eine wissenschaftliche Edition von Fassungen und Varianten ist im Rahmen des Projektes explizit nicht vorgesehen, allerdings schafft www.handkeonline.onb.ac.at wesentliche Grundlagen einer künftigen historisch-kritische Ausgabe.

Die Website richtet sich nicht nur an Forscher und Experten, sondern auch an ein breiter interessiertes Lesepublikum. Die hohen und konstant steigenden Zugriffszahlen seit Freischaltung der Seite (in einer Probeversion ab Januar 2013) und Umfragen zum Nutzungsverhalten zeigen, dass die übersichtliche Struktur und die einfache Bedienung intuitiv erschlossen werden. Die zahlreichen Informationen, die die Seite bietet, werden oft nachgefragt und in vielfacher Weise genutzt. *Die Welt* schreibt in einer *Versuch über das Internet* betitelten Besprechung des Projekts: „Ob Peter Handke Handke*online* kennt, kann nur gemutmaßt werden. Er lebt ja eher offline. Auf der üppigen Website können Fans jetzt lauter digitale Devotionalien entdecken, die die Handke-Forschung umtreiben.“

Sechs Module bieten unterschiedliche Zugriffsmöglichkeiten auf den Content: Der Bereich "Wege durchs Material" ist eine virtuelle Ausstellung, die erste Einblicke in Handkes Themen und die Art seine Schreibens gibt. Das Modul "Notizbücher von 1972 bis 1990" verzeichnet die heute zugänglichen 76 Notizbücher des Autors bibliothekarisch und erschließt die mehr als 10.000 Einzelseiten in seinen Inhalten punktuell. Orte, Werkbezüge und Lektürenotizen finden sich aufgeschlüsselt. Das Kernmodul "Werke & Materialien" bietet eine genaue Beschreibung der werkgenetischen Konvolute zu den mehr als 80 Büchern, die bislang vom Autor vorliegen. Notizen, Fassungen und Quellen aus verschiedenen Archiven und Sammlungen werden zusammengeführt und in eine textgenetische Ordnung gebracht. Ausgewählte Faksimiles einzelner Seiten vermitteln von den Materialien ein anschauliches Bild. Im Bereich unerschlossener Archivbestände und privater Sammlungen wurden umfangreiche Recherchearbeiten unternommen, auch diese Materialien konnten in ihrer Gesamtheit auf die Seite gestellt werden.

Mit Zustimmung von Peter Handke veröffentlicht das vierte Modul eine Auswahl von "Gesamtfaksimiles". Die Open Access-Plattform im Modul "Forschungsbeiträge" enthält derzeit (Stand September 2014) mehr als 50 wissenschaftliche Aufsätze zum Werk des Autors, ein Fünftel davon sind Originalbeiträge. Eine umfangreiche Gesamtbibliografie zu Peter Handke mit mehr 2600 Einträgen schließt sich im sechsten Modul an. Als technischer Hintergrund wurde für die Webapplikation das CMS Drupal als Framework verwendet. Die Inhalte der Seite sind untereinander vernetzt und in mehr als 1000 Datensätzen strukturiert abgelegt und können auch über eine einfache Suchfunktion erschlossen werden.

Finanzierung und Laufzeit

www.handkeonline.onb.ac.at ist das Ergebnis eines vom Fonds zur Förderung der wissenschaftlichen Forschung (Austrian Science Fund) finanzierten Einzelprojekts, das seit Mai 2011 läuft. Nach Abschluss der Arbeiten im April 2015 wird die Seite auf der Website der Österreichischen Nationalbibliothek verfügbar gehalten. Es ist geplant, die integrierte Open-Access-Plattform über das Projektende hinaus nach Art einer digitalen Zeitschrift fortzuführen.

Projektleitung und Projektmitarbeiter

Projektleiter des vom FWF (Austrian Science Fund) finanzierten Projekts ist PD Dr. Klaus Kastberger, wissenschaftlicher Mitarbeiter am Literaturarchiv der Österreichischen Nationalbibliothek und Privatdozent an der Universität Wien. Projektmitarbeiter sind Mag. Christoph Kepplinger und Mag. Katharina Pektor. Beratend steht dem Projekt eine internationaler wissenschaftlicher Beirat zur Seite.

Christian Thomas
Humboldt-Universität zu Berlin, Institut für Kulturwissenschaft
Projekt *Hidden Kosmos*, www.culture.hu-berlin.de/hidden-kosmos

Exposé zu einem Vortrag, angenommen für die DHd-Tagung 2015 „Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation“, http://dhd2015.uni-graz.at/

– eingereicht am 10.11.2014, korrigierte Version vom 20.12.2014 –

Hidden Kosmos – Humboldts ‚Kosmos-Vorträge‘ als Probe der Digital Humanities

Abstract

Im Projekt *Hidden Kosmos – Reconstructing Alexander von Humboldt's »Kosmos-Lectures«* (www.culture.hu-berlin.de/hidden-kosmos) der Humboldt-Universität zu Berlin (HU) werden seit Juni 2014 Quellen zu den sogenannten Kosmos-Vorträgen digitalisiert. Darin stellte Alexander von Humboldt 1827/28 in zwei teilweise parallel verlaufenden Vortragsreihen an der Berliner Universität und der benachbarten Singakademie das naturwissenschaftliche Wissen seiner Zeit dar.

Das Projekt *Hidden Kosmos* wird sämtliche bislang bekannte Hörernachschriften aus Bibliotheken, Archiven und Privatsammlungen in Deutschland, Polen und der Türkei virtuell zusammenführen. Insgesamt elf Nachschriften wurden bisher ermittelt, davon beziehen sich acht auf den 61-stündigen Universitäts-Zyklus, drei auf den nur 16 Vorträge umfassenden Singakademie-Zyklus. TEI-kodierte Transkriptionen der ca. 3500 Manuskriptseiten werden als standardkonform aufbereitetes, tief annotiertes und vielseitig vernetztes Volltextkorpus unter einer offenen Lizenz zur Verfügung gestellt. Damit wird auf einer umfassenden Quellenbasis die noch ausstehende, intensive Erforschung der Humboldtschen Vortragszyklen überhaupt erst ermöglicht.

Der Vortrag veranschaulicht anhand erster Ergebnisse des Forschungsprojekts dessen konzeptionelle Grundlagen. Um dem breiter angelegten Themenkreis der DHd-Tagung gerecht zu werden, wird dabei weitgehend von Humboldt-spezifischen Inhalten abstrahiert und der Fokus auf die Aufbereitung der Daten sowie auf die dabei zum Einsatz kommenden Methoden und Verfahren der ‚Digital Humanities‘ (DH) gerichtet.

Korpusaufbau: Drei Wege der Volltexterstellung

Das zu bearbeitende Korpus umfasst insg. ca. 3500 Seiten aus elf handschriftlichen Nachschriften. Die Online-Publikation *Hidden Kosmos* soll jedes Dokument dieser vielstimmigen Überlieferung im Volltext wiedergeben, anstatt aus den voneinander abweichenden Nachschriften einen ‚idealen' Text zu konstruieren. Dabei werden drei unterschiedliche Wege der Volltexterstellung verfolgt.

1) Kollation: Zwei der Nachschriften, je eine aus der Universität und der Singakademie, sind bereits als Druckausgaben erschienen (Anonym 1934, Hamel/Tiemann 1993; vgl. Literaturverzeichnis im Anhang). Der per OCR gewonnene und manuell korrigierte Volltext dieser Ausgaben wird mit der jeweiligen Vorlage verglichen. Die teilweise gravierenden, im Druck nicht vermerkten Abweichungen vom Manuskript – ‚Normalisierungen', ‚stillschweigende Korrekturen' sowie Transkriptionsfehler – werden in editorischen Kommentaren vermerkt.

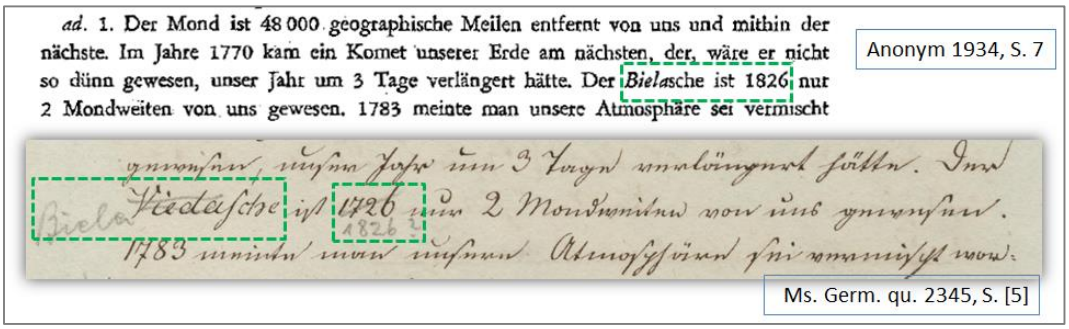


Abb. 1: *‚Stillschweigende Korrekturen' gegenüber der handschriftlichen Vorlage in Anonym 1934.*

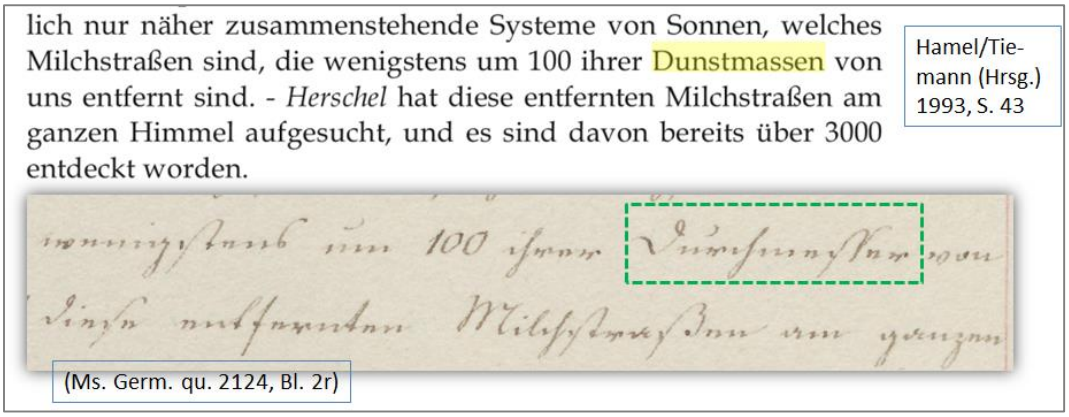


Abb. 2: *Transkriptionsfehler in Hamel/Tiemann 1993: „Dunstmassen" st. „Durchmesser".*

Die Kollation der beiden Druckausgaben sowie einer späteren Abschrift eines Manuskripts mit ihren jeweiligen handschriftlichen Vorlagen ist abgeschlossen; die Texte werden zum Zeitpunkt der Konferenz über das Deutsche Textarchiv (DTA) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und die HU verfügbar sein (s. Abschnitt Parallele Publikation). Jede der entstehenden Transkriptionen dokumentiert allein die Handschrift, die sie wiedergibt, d. h. Abweichungen der Manuskripte voneinander werden *nicht* innerhalb der Transkription vermerkt, sondern nach Fertigstellung der beiden Volltexte automatisch ermittelt und visualisiert. Im Vortrag werden die Vorzüge dieses DH-Verfahrens gegenüber dem ‚klassischen' Variantenapparat dargestellt.

2) Texterfassung durch Dienstleister: Fünf der bislang nicht erfassten Nachschriften aus der Universität (insg. ~2600 Seiten) weisen ein größtenteils regelmäßiges Schriftbild und vergleichsweise wenige Ergänzungs- und Überarbeitungsspuren auf.

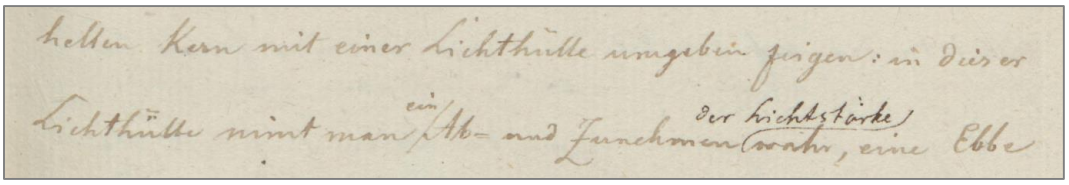


Abb. 3: *Leichte Transkription: SBB-PK, Ms. Germ. qu. 1711, Bl. 2r*

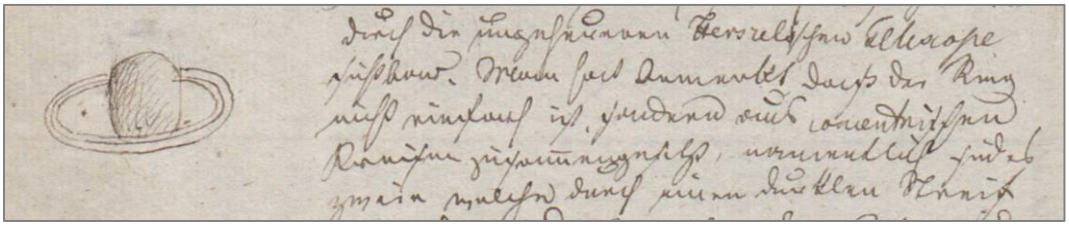


Abb. 4: *Mittelschwere Transkription: Biblioteka Jagiellońska Kraków, Handschrift 6623 II, S. 144*

Wenngleich es – anders als bei der Erfassung von Drucktexten – im Bereich handschriftlicher Überlieferung noch keine etablierte Praxis ist, lag es aus unserer Sicht nahe, diese Manuskripte durch einen Dienstleister erfassen zu lassen. Das wissenschaftliche/studentische Personal des *Hidden-Kosmos*-Projekts konzentriert sich auf die Qualitätssicherung, die tiefere Annotation, Kommentierung, intra- und intertextuelle Verknüpfung, (normdatengestützte) Referenzierung und Publikation der Volltexte. So kann trotz geringer Personalausstattung in nur zweijähriger Laufzeit die verhältnismäßig große Zahl von ca. 3500 Manuskriptseiten bewältigt werden.

Mit *textloop* wurde ein bei der Erstellung elektronischer Volltexte, insbesondere XML-annotierter Volltexte gemäß den Richtlinien der *Text Encoding Inititiative* (TEI) erfahrener Partner gefunden. *textloop* übernimmt die Texterfassung (Zeichengenauigkeit mind. 98%) samt Basisannotation, die entsprechende Einweisung und Anleitung der Texterfasser. Weiterhin ist *textloop* für die Erstellung und fortlaufende Anpassung des RNG-Schemas und der ODD-Dokumentation[1] nach den Bedürfnissen des Projekts verantwortlich.[2] Im Vortrag sollen unsere Erfahrungen mit dem externen Dienstleister dargestellt werden, auch, um Handschriften-Digitalisierungsvorhaben mit ähnlichen Beständen über die Möglichkeiten einer solchen Zusammenarbeit zu informieren.

3) Texterfassung durch Projektteam: Schließlich werden Nachschriften, die aufgrund des Schriftbilds und zahlreicher Überarbeitungsspuren für die Erfassung durch einen Dienstleister nicht geeignet sind, durch wissenschaftliches/studentisches Personal im Projekt transkribiert.

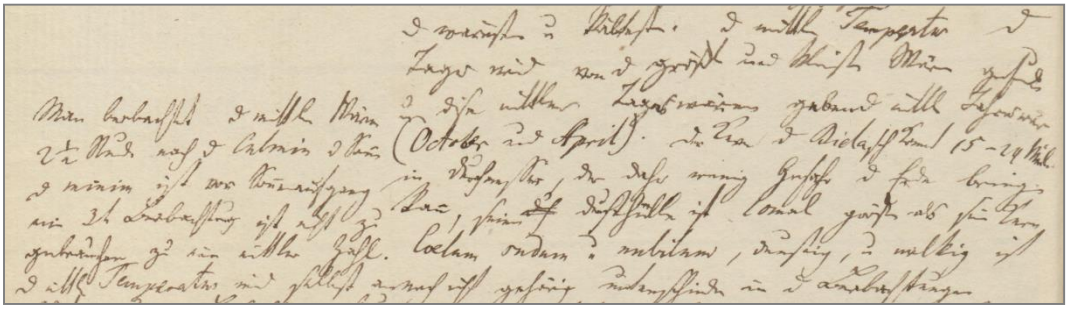


Abb. 5: *Schwere Transkription: SBB-PK, Slg. Darmstaedter, F 2c 1853_Riess, Peter Theophil , S. 67*

[1] RNG = Relax NG, Regular Language for XML; vgl. http://relaxng.org/. ODD = One Document Does it All; vgl. http://www.tei-c.org/Guidelines/Customization/odds.xml. Vgl. auch TEI P5 Guidelines, Ch. 23: "Using the TEI", http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html, zum Einsatz von RNG und ODD.

[2] Die Erfassung einer bisher unveröffentlichten, 800-seitigen Nachschrift aus der Universität wurde im November 2014 abgeschlossen; bis zum Zeitpunkt der Konferenz werden dieser und weitere von *textloop* erfasste Bände publiziert sein.

Annotation in TEI P5

Die Volltexterstellung erfolgt von Beginn an in TEI-XML, so dass bereits während der Transkriptionsarbeit die Basisannotation der Dokumente vorgenommen werden kann. Die Kodierung der Dokumente geschieht vollständig TEI-konform unter Verwendung des projekteigenen Schemas, eines TEI-Subsets *ohne* projektspezifische Erweiterungen. Das *Hidden-Kosmos*-Tagset folgt weitgehend dem Basisformat des Deutschen Textarchivs (DTABf), das eine minimale Anzahl an Tags für die konsistente Kodierung eines heterogenen Korpus definiert, um die größtmögliche Einheitlichkeit und Interoperabilität der darin enthaltenen Dokumente sicherzustellen (vgl. Haaf et al. forthcoming). Das DTABf-Tagset wurde für das *Hidden-Kosmos*-Projekt lediglich um Annotationsmöglichkeiten für Handschriften[3] erweitert.

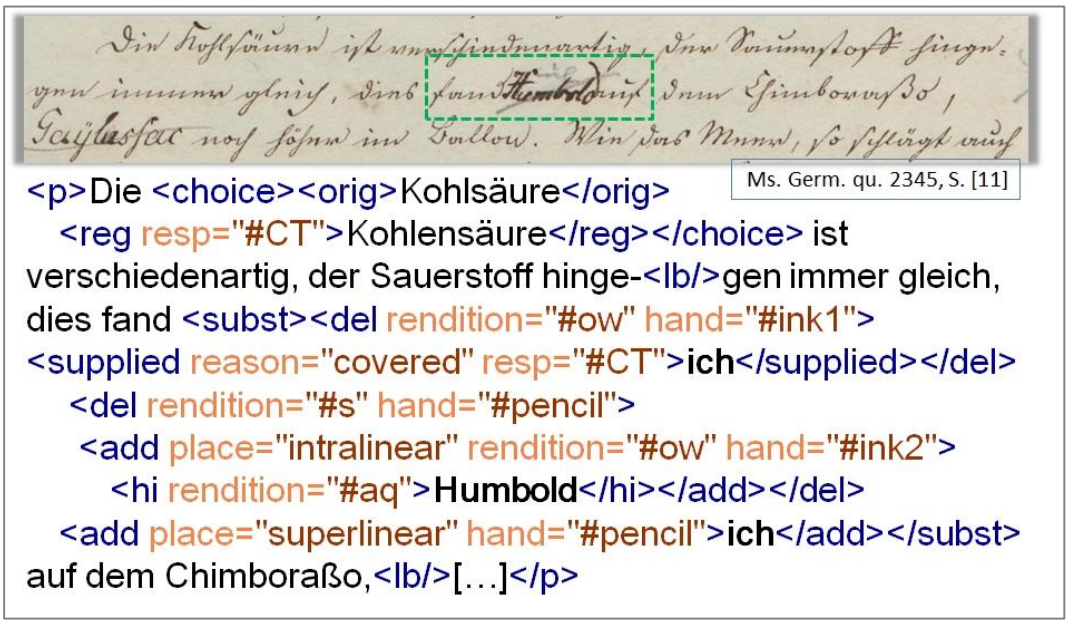


Abb. 6: *TEI-/DTABf-konforme Manuskript-Annotation; SBB-PK, Ms. Germ. qu. 2345, S. [11]*

Orientierungspunkte für diese Erweiterung des Tagsets in Richtung Manuskript-Annotation boten die Encoding Guidelines der Nachwuchsgruppe Berliner Intellektuelle 1800–1830 der HU und das elaborierte „Genetic encoding" des Shelley-Godwin Archive. Das in dieser Weise erstellte *Hidden-Kosmos*-Tagset und das dazugehörige RNG-Schema stellt die einheitliche, TEI-konforme Kodierung sämtlicher Dokumente durch mehrere Bearbeiter auf Seiten des Dienstleisters und der Projektgruppe sicher.

Parallele Publikation

Das derart aufbereitete Korpus wird auf der projekteigenen *Hidden-Kosmos*-Webseite[4] präsentiert. Die vielfältigen inhaltlichen Verbindungen der Kosmos-Vorträge zu anderen, bereits existierenden Online-Ressourcen machen eine Publikation der Daten auch auf anderen Plattformen wünschenswert – ein Desiderat, das sich auf etliche andere digitale Bestände übertragen ließe. Im Projekt *Hidden Kosmos* kann dies dank der interoperablen Kodierung der Dokumente in TEI-XML und der von Beginn an engen Kooperation mit verschiedenen Vorhaben realisiert werden.

---

[3] Notwendig waren vor allem Erweiterungen um msDesc/<header>-Elemente sowie im <text>-Bereich um Elemente für Überschreibungen, Hinzufügungen über/unter/neben der Zeile, Schreiberwechsel etc.

[4] Die Webseite befindet sich im Aufbau; zum geplanten Funktionsumfang siehe Abschnitt DH-Verfahren.

Die Re-Integration der für *Hidden Kosmos* vorgenommenen Erweiterungen in das DTABf-Tagset erlaubt die unmittelbare ‚Parallelpublikation' der Nachschriften im DTA. Dort sind bereits mehrere Werke Humboldts, so der *Kosmos* (1845–62) und eine wachsende Zahl seiner unselbstständigen Schriften,[5] sowie zahlreiche Werke seiner Zeitgenossen verfügbar.

Die linguistische Erschließung der Texte im DTA (inkl. automatischer Normierung historischer Schreibweisen) und die darauf basierende Suchmaschine ermöglichen eine optimale Nutzung der Nachschriften im Kontext der DTA-Korpora. Die Suchfunktion des DTA wird über eine Schnittstelle auch auf der *Hidden-Kosmos*-Webseite eingebunden, womit dort die Entwicklung einer eigenen (zwangsläufig basaleren) Volltextsuche nicht notwendig ist. Ein weiterer Vorzug der Publikation der Texte im DTA ist die Möglichkeit, die Dokumente mit Hilfe der webbasierten Qualitätssicherungs-Plattform DTAQ iterativ und möglichst kollaborativ zu optimieren: Transkriptions- und Auszeichnungsfehler können mit DTAQs Online-Editor direkt in der XML-Quelle behoben werden; ebenso können Normdaten-Referenzen für Personen- und Ortsnamen sowie weitere Verlinkungen und Kommentierungen vorgenommen bzw. ergänzt werden.

Als weitere Publikationsplattform dient das Projekt *Briefe und Texte aus dem intellektuellen Berlin um 1800*. Die Nachschriften der Kosmos-Vorträge werden dort den Vorlesungen F. A. Wolfs, Karl Solgers u. a. zur Seite gestellt. Der Datenaustausch wird durch die Orientierung beider Projekte am DTABf erleichtert.

Die jeweils aktuelle Version der XML-Basis im DTA dient als Referenzobjekt für alle weiteren Veröffentlichungsplattformen, deren Datenstand möglichst automatisiert mit der DTA-Version synchronisiert werden soll. Auf diese Weise lässt sich das Problem divergierender Versionen bei der ‚Spiegelung' von Daten auf verschiedenen Plattformen umgehen. Über das DTA bzw. das CLARIN-D-Repository der BBAW ist die Integration des Textkorpus in die web- und zentrenbasierte Forschungsinfrastruktur CLARIN-D gewährleistet, womit eine weitere Dissemination sowie die langfristige Archivierung und Bereitstellung der Daten sichergestellt wird.

DH-Verfahren

Neben den bereits beschriebenen DH-Verfahren und -Methoden[6] bieten sich vor allem die in den DH inzwischen etablierten automatischen Kollationsverfahren für die Analyse des Korpus an. Diese dienen dazu, Ähnlichkeiten und Differenzen zwischen Textzeugen zu ermitteln und zu visualisieren.

Traditionellerweise, d. h. vor allem in Print-Editionen, werden Varianten zweier oder mehrerer Textzeugen in Kommentaren beschrieben und/oder in einem (notwendigerweise hochkomplexen) textkritischen Apparat ausgewiesen.[7] Die automatische Kollation durch Programme wie *CollateX* und *juXta* macht die manuelle Auszeichnung von Varianten weitgehend überflüssig, ohne auf deren Mehrwert verzichten zu müssen.

---

5 Werke A. v. Humboldts im Deutschen Textarchiv siehe www.deutschestextarchiv.de/api/pnd/118554700.

6 Siehe dazu z.B. Reiche et al. 2014.

7 Deren Nutzung in der geisteswissenschaftlichen Forschungen ist wiederholt diskutiert worden; kürzlich wies Rüdiger Nutt-Kofoth auf der Tagung der AG germanistische Edition (vgl. Vanscheidt 2014) anhand von ca. 540 einschlägigen Fachartikeln nach, dass nur 40% die (historisch-)kritischen Editionen heranzogen, davon wiederum nur 10% die Erläuterungen der Ausgaben und *lediglich 5%* die Textvarianten zitierten. Demnach wird der am aufwendigsten herzustellende Editionsbestandteil am wenigsten genutzt.

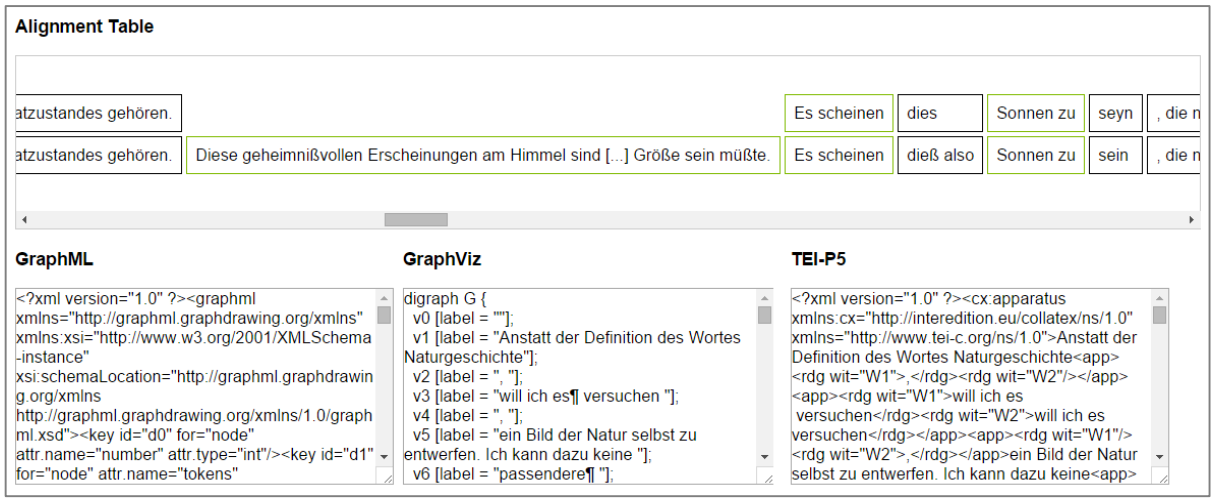


Abb. 7: *CollateX Alignment Table: SBB-PK, Ms. Germ. qu. 2124 vs. Abschrift Hufeland, Privatbesitz C. Şengör.*

Tabellarische Darstellungen wie die hier abgebildete, Variantengraphen, Parallelisierungen von Textzeugen (siehe Abb. 8 und 9 im Anhang) und ‚Critical Apparatus' bieten explorative Zugänge zu den unterschiedlichen Fassungen der Kosmos-Vorträge. Stets auf der aktuellen Version des Dokuments basierend, werden je nach Kollationsparameter Unterschiede auf der Ebene einzelner Zeichen, Worte und/oder ganzer Absätze sichtbar.

Die TEI-Annotation der Dokumente eröffnet noch weitere computergestützte Auswertungsmöglichkeiten: Die per `<div type="session" n="[Zaehler]">` einheitlich annotierte Gliederung nach Vorlesungsstunden erleichtert dem Nutzer den Einstieg in parallele Lektüren der Nachschriften. Durch die Referenzierung von Personen und Orten kann direkt aus dem TEI ein normdatenbasiertes Gesamtregister aller Nachschriften generiert werden, was den systematischen Zugang zu den Quellen erleichtert. Eine Gesamtbibliographie der in den Vorträgen mit `<bibl>`-Tags versehenen Literaturangaben wird ein weiterer Zugangsweg zu den Kosmos-Vorträgen sein.

Ausblick

Über die beschriebene Nutzung als eigenständiges Forschungskorpus hinaus sollen die Nachschriften in einem Folgeprojekt als ‚Werkzeug' zur Erschließung der eigenhändigen Vortragsmanuskripte aus Humboldts Nachlass in der SBB-PK dienen. Die ursprünglichen Manuskripte wurden von Humboldt im Anschluss an die Vorträge während seiner Arbeit am *Kosmos* und anderen Publikationen stark überarbeitet und reorganisiert (vgl. Erdmann/Thomas 2014). Daher kann die im *Hidden-Kosmos*-Projekt angestrebte ‚Rekonstruktion' der Kosmos-Vorträge nicht aus Humboldts Manuskripten allein, sondern nur durch eine computergestützte Auswertung der umfangreichen Gesamtbasis aus Sekundärquellen (= Hörernachschriften) und Primärquellen (= Humboldts Vortragsnotizen) gelingen.

Abbildungen

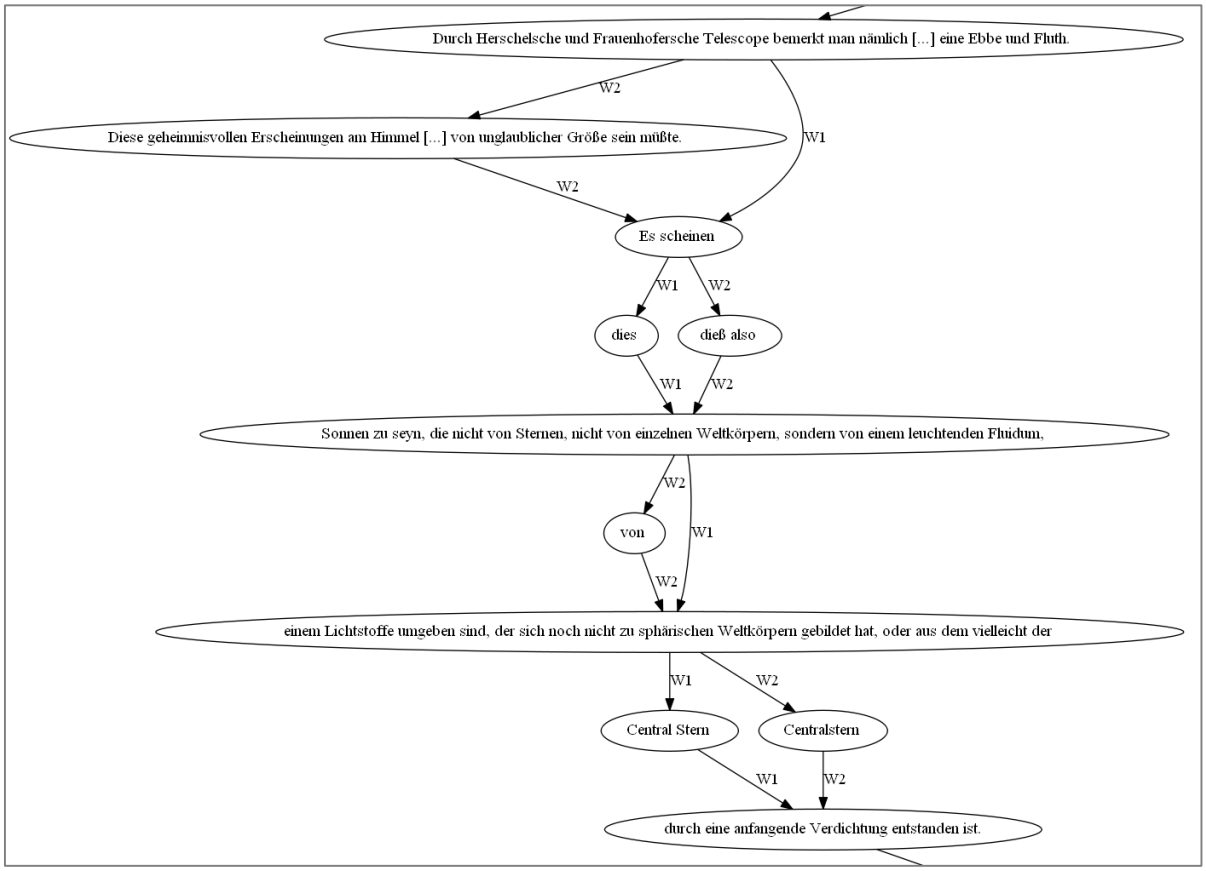


Abb. 8: *CollateX Variant Graph*

(SBB-PK, Ms. Germ. qu. 2124 vs. Abschrift Hufeland, Privatbesitz C. Şengör)

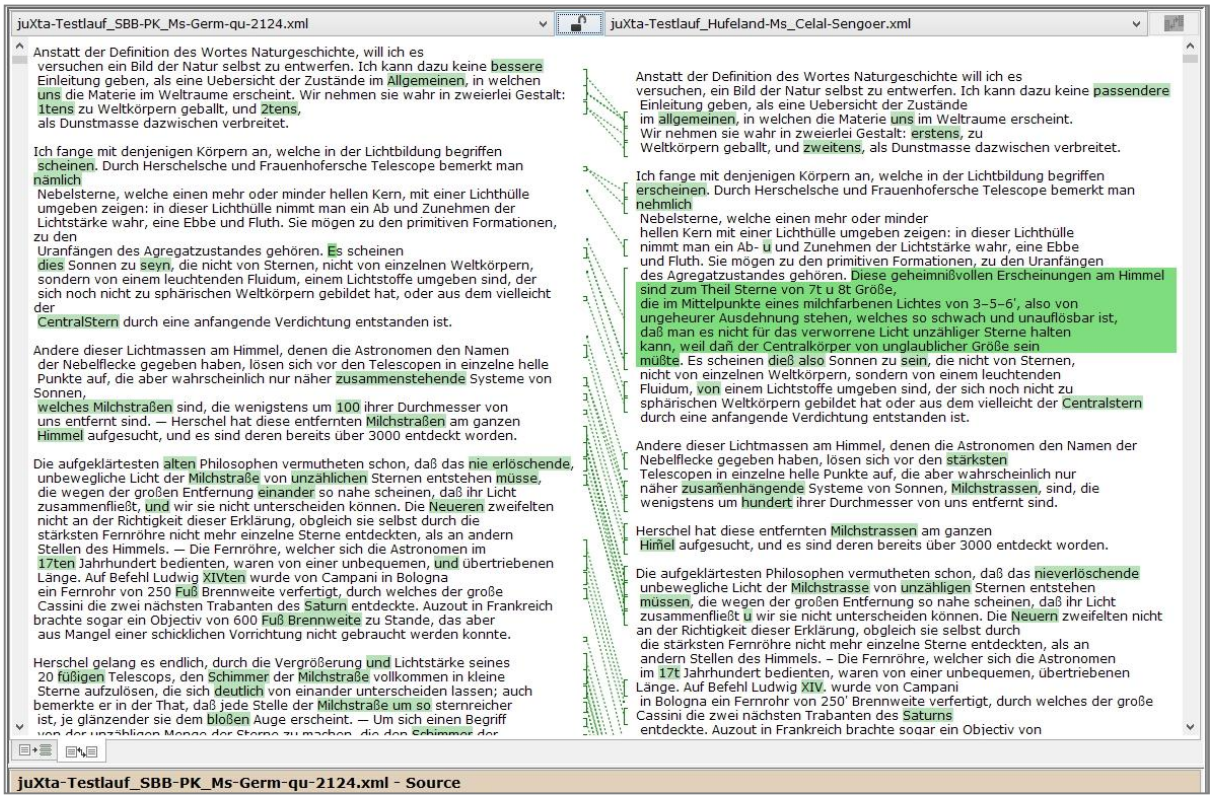


Abb. 9: *Juxta Side-by-Side view*

(SBB-PK, Ms. Germ. qu. 2124 vs. Abschrift Hufeland, Privatbesitz C. Şengör)

Bibliographie

Literatur

Anonym (1934): Alexander von Humboldts Vorlesungen über physikalische Geographie nebst Prolegomenen über die Stellung der Gestirne. Berlin im Winter von 1827 bis 1828. Erstmalige (unveränderte) Veröffentlichung einer im Besitze des Verlages befindlichen Kollegnachschrift. Berlin: Miron Goldstein.

Dove, Alfred (1872): „Alexander von Humboldt auf der Höhe seiner Jahre. (Berlin 1827–59.)“, in: Bruhns, Karl (Hg.) (1872): Alexander von Humboldt: Eine wissenschaftliche Biographie. Leipzig: Brockhaus, 3 Bde, hier Bd. II., S. 93–484.

Erdmann, Dominik und Christian Thomas (2010): Aussicht vom Zettelgebirge – Zur Datenverarbeitung in Alexander von Humboldts Manuskripten der Kosmos-Vorlesungen. In: Trajekte 20 (2010), S. 30–36.

Erdmann, Dominik und Christian Thomas (2014) »… zu den wunderlichsten Schlangen der Gelehrsamkeit zusammengegliedert«. Neue Materialien zu den ›Kosmos-Vorträgen‹ Alexander von Humboldts, nebst Vorüberlegungen zu deren digitaler Edition. In: HiN – Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien (Potsdam – Berlin) XV, 28, S. 34-45. Online verfügbar unter http://hin-online.de/hin28/erdmann-thomas.htm [zuletzt abgerufen 27.10.2014].

Haaf, Susanne, Alexander Geyken and Frank Wiegand (2014/15): The DTA ›Base Format‹: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. To appear in: Journal of the Text Encoding Initiative (jTEI), Issue 8 [forthcoming].

Hamel, Jürgen und Klaus-Harro Tiemann (Hg.) (1993): Alexander von Humboldt: Über das Universum. Die Kosmosvorträge 1827/28 in der Berliner Singakademie. Hrsg. von Jürgen Hamel u. Klaus-Harro Tiemann in Zusammenarbeit mit Martin Pape. Frankfurt a. M.: Insel.

Harnack, Adolf (1900): Geschichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin, Bd. I.2: Vom Tode Friedrichs des Großen bis zur Gegenwart. Berlin: Reichsdruckerei.

Humboldt, Alexander von (1845–62): Kosmos. Entwurf einer physischen Weltbeschreibung. 5 Bände. Stuttgart (u.a.): Cotta. [Der Kosmos und weitere Schriften Alexander von Humboldts sind im Deutschen Textarchiv der BBAW verfügbar, siehe die Übersicht unter www.deutschestextarchiv.de/api/pnd/118554700 (zuletzt abgerufen 27.10.2014).]

Nutt-Kofoth, Rüdiger (2014, bislang nicht publizierter Vortrag): „Wie werden Editionen für die literaturwissenschaftliche Interpretation genutzt? Versuch einer Annäherung aufgrund einer Auswertung neugermanistischer Periodika“ auf der 15. internationalen Tagung der Arbeitsgemeinschaft für germanistische Edition „Vom Nutzen der Editionen“, 19.–22.2.2014, Universität Aachen.

Pierazzo, Elena (2011): „A Rationale of Digital and Documentary Editions.“ In: Literary and Linguistic Computing (LLC) 26(4), S. 463–77. doi:10.1093/llc/fqr033, http://llc.oxfordjournals.org/content/26/4/463 [zuletzt abgerufen 27.10.2014].

Pierazzo, Elena (2014): „Digital Documentary Editions and the Others.“ In: Scholarly Editing: The Annual of the Association for Documentary Editing, 35 (2014), http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html [zuletzt abgerufen 27.10.2014].

Reiche, Ruth, Rainer Becker, Michael Bender, Matthew Munson, Stefan Schmunk und Christof Schöch: "Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften." (= DARIAH-DE Working Papers Nr. 4). Göttingen, 2014. URN: urn:nbn:de:gbv:7-dariah-2014-2-6, http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2014-4.pdf.

Vanscheidt, Philipp (2014): Bericht zur Tagung „Vom Nutzen der Editionen“, in: Scriptorium. Digitale Rekonstruktionen mittelalterlicher Bibliotheken, http://scriptorium.hypotheses.org/364.

Webressourcen [zuletzt abgerufen 28.10.2014]

Anne Baillot (ed.): "Briefe und Texte aus dem intellektuellen Berlin um 1800". Berlin: Humboldt-Universität zu Berlin, http://tei.ibi.hu-berlin.de/berliner-intellektuelle/;
Edition-specific TEI encoding guidelines, http://tei.ibi.hu-berlin.de/berliner-intellektuelle/encoding-guidelines.pdf.

CollateX, http://collatex.net/;
collateX Console, http://collatex.net/demo/.

Deutsches Textarchiv, http://www.deutschestextarchiv.de/;
DTABf: DTA-Basisformat, http://www.deutschestextarchiv.de/doku/basisformat;
DTAE: DTA-Erweiterungen, http://www.deutschestextarchiv.de/dtae;
DTAQ: Deutsches Textarchiv – Qualitätssicherung, http://www.deutschestextarchiv.de/dtaq/about.

Hidden Kosmos — Reconstructing Alexander von Humboldt's »Kosmos-Lectures«, http://www.culture.hu-berlin.de/hidden-kosmos.

Juxta Collation Software for Scholars, http://www.juxtasoftware.org/.

Shelley-Godwin Archive, http://shelleygodwinarchive.org/;
Encoding the S-GA, http://shelleygodwinarchive.org/about#encodingthesga.

Text Encoding Inititiative (TEI), http://www.tei-c.org/;
TEI: P5 Guidelines, http://www.tei-c.org/Guidelines/P5/.

textloop Martina Gödel, http://textloop.de/.

exploreAT!
Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts

Im vorliegenden Paper wird erstmals das interdisziplinäre, internationales Projekt „exploreAT! exploring austrias culture through the language glass“ vorgesellt, das ab 2015 an der Österreichischen Akademie der Wissenschaften im Bereich „Digital Humanities“ für 4 Jahre umgesetzt wird.
Die beteiligten Kolleg.innen arbeiten in den Bereichen Soziologie, Softwarentwicklung, Mensch-Maschine-Interaktion, Lexikographie und Digital Humanities.

Auf Basis des Jahrhundertprojekts „Wörterbuch der bairischen Mundarten in Österreich (WBÖ)“ (1911-) und seines digitalen Schwesterprojekts „Datenbank der bairischen Mundarten in Österreich (DBÖ)“ [1993-; samt Weiterentwicklungen wie: „Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)“ 2007-] wird ein Beispiel für einen Transformationsprozess eines geisteswissenschaftlichen Projekts diskutiert.
Die Datensammlung ist quantitativ umfassend:
Wir arbeiten mit rund 200,000 ungedruckten, nicht-lexikographisch finalisierten Stichwörtern,; geschätzten 4 Mio. Dateneinträgen (Wörter, Kontexte); 5 Wörterbuchbänden, mit über 50,000 Stichwörtern; jeweils von den Anfängen der deutschen Sprache bis heute, unter besonderer Berücksichtigung des bairischen Dialekts in Österreich (sei es die Habsburgermonarchie, wie es für den Großteil des Sammelzeitraums essentiell ist, oder das Staatsgebiet des heutigen Österreich).

Die Zusammenarbeit ist motiviert im Bestreben der Kolleg.innen, folgende Aspekte einem zu einem synergetischen Ganzen zusammenzufügen:

1. Das Vorhandensein einer unikalen Wörtersammlung für die ländliche, österreichische Kultur des letzten Jahrhunderts.
2. Das Vorhandensein darauf aufbauender wissenschaftlicher Arbeiten und Dokumentationen (z.B: WBÖ, DBÖ, dbo@ema).
3. Das Bewusstsein um die identitätsstiftende Verankerung des Dialekts in Österreich (z.T. noch vermehrt durch die neuen Medien und das Web 2.0, z.B. Facebook, SMS, WhatsApp)-

4. Das umfassende, steigendende lexikographische Interesse breiter Bevölkerungsteile (Beispiel: Wikipedia, Regionalwörterbücher).
5. Das wissenschaftliche Interesse der Kolleg.innen an (der Weiterentwicklung) zeitgemäßer Lexikographie und ihren Produkten bzw. an der kritischen Analyse des lexikographischen Prozesses und seiner Einbettung in einen aktualisierten realweltlichen Kontext, markiert durch Cyberscience und Web 2.0- / Web 3.0 im Alltag).

Die grundlegenden Aspekte der Zusammenarbeit werden vorgestellt:

1. Infrastruktur:
   Weiterentwicklung einer bestehenden Datenbank zu einer web-basierten, kollaborativen Infrastruktur für Archivierung, Edition, Publikation und Analyse multilingualer Non-Standard-Daten, deren lexikographische Output (diverse Wörterbuchprodukte) und deren Wissensquellen, sowohl für Wissenschaftler als auch für Laien, basierend auf Standards und Up-to-date-Technologien.

Weitere Schwerpunkte der technischen Neuentwicklungen stellen dar:

- Semantic Web:
  Nutzung der Vorteile des Semantic Web für die Lexikographie: Modellierung und Datenpublikation (WBÖ+DBÖ) in Linked Open Data; Semantische Erschließung des Datenkorpus durch Ableitung von Basiskonzepten aus dem Fragebogen (20,000 Detailfragen).

- Visual Analytics:
  Entwicklung und Weiterentwicklung von Visual Analytics Werkzeugen für Non-Standard Daten. Es wird darauf abgezielt, Tools zu entwickeln, die für analoge Projekte eingesetzt und zum Vergleich von Datenstrukturen angewendet werden können.

- Serious Games:
  Entwicklung spielerischer Anwendungen für Lehre und Laien.
  Damit werden Werkzeuge zur Verfügung gestellt, welche die direkte Kommunikation mit neuen Usern gewährleisten sollen.

2. Daten Enrichment / Re-Use:
   Die Zusammenarbeit verfolgt Open Science Prinzipien.
   Die Anreicherung der eigenen Daten durch externe Daten (z.B. mittels Linked Open Data) stellt einen Kernpunkt der Zusammenarbeit dar; ebenso wird die inter- und transdiziplinäre Wiederverwendung der Daten in unterschiedlichen Kontexten gefördert.
   Die Zusammenarbeit mit Firmen bzw. das Einbringen lexikographischen Knowhows in den Business Kontext wird erprobt.

3. Gesellschaft:
   Basierend auf der Rolle und Funktion, die Non-Standard Sprache in Österreich einnimmt, werden Ansätze zur gewinnenbringenden Zusammenarbeit und Einbindung der Gesellschaft in den Wissenschaftsprozess diskutiert. Citizen Science Modelle werden aktiv erprobt und umgesetzt.

4. Metalexikographie:
   Ein Schwerpunkt der Zusammenarbeit widmet sich der methodischen Neuorientierung und somit der Diskussion der Fragestellung, welche Rolle Lexikographie als Teilbereich von „Digital Humanities“ derzeit einnimmt bzw. wo lexikographisches Wissen in anderen Bereichen festgemacht werden kann / könnte.
   Welche methodischen Veränderungen bedingen Linked Open Data für die klassische Lexikographie bzw. umgekehrt.
   Die Folgen und Auswirkungen lexikographischer Arbeit im Open Science Paradigma wird reflektiert und thematisiert.

Die Zusammenarbeit ist eingebettet in enge Zusammenarbeit mit Stakeholdern in einzelnen Bereichen, wie DARIAH-EU, COST IS 1305 European Network of electronic Lexicography, EUROPEANA, WIKIMEDIA.AT , OPEN KNOWLEDGE FOUNDATION, oder beispielsweise den Projekten LIDER, opendataportal.at, SOCIENTIZE.

Der Beitrag versteht sich als zusammenfassender Überblick über Perspektiven eines Transformationsprozesses für ein lexikographisches Großprojekt. Aufgrund der Ausgangslage – im frühen 20. Jahrhundert sind nach dem Beispiel des Schweizerischen Idiotikons viele analog gestaltete Projekte konzipiert und umgesetzt worden – erscheinen

Analogieentwicklungen und Anwendung diverser Tools auf andere bestehende Projekte unterschiedlichen Digitalisierungsgrads möglich und wünschenswert.
Im Rahmen der COST Aktion IS 1305 werden entsprechende Übertragungen beispielhaft umgesetzt.

Der Beitrag fokussiert neben dem beispielhaften Umreißen einzelner Möglichkeiten vor dem Hintergrund der aktuellen Zusammenarbeit, den Mehrwert für ein lexikographisches Projekt durch einen Transformationsprozess Richtung Digital Humanities.

Referenzen:

Agosti, M. et. al. (Eds.): Digital Libraries and Archives. 2013: 195-206.

Agosti, M. et al. "An Evaluation of the Involvment of General Users in a Cultural Heritage Collection." In: *Digital Humanities* (2013): 75-77.

Agosti, M. et al. "Digital Libraries and Archives. 8th Italian Research Conference, IRCDL, 2012, Bari, Italy, February 9-10, 2012, Reveised Selected Papers". *CCIS* Vol. 354.

"Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch". Wien.1913.

Bailey, E. et al: "CULTURA: Supporting Professoinal Humanities Researchers." In: *Digital Humanities* (2013): 99-101.

Burigat, S., Chittaro, L. "Interactive visual analysis of geographic data on mobile devices based on dynamic queries." *Journal of Visual Languages & Computing* 19.1 (2008): 99-122.

Catarci, T. et al. "Evaluating Cultural Heritage Information Access Systems. Bridging Between Cultural Heritage Institutions." Berlin Heidelberg. 2014: 7-16.

Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Ed. by Wandl-Vogt, E. Wien. 2010.

Dear, M., Ketchum, J., Luria, S. "GeoHumanities: Art, History, Text at the Edge of Place. " London, New York. 2014.

De Gasperis, G, Florio, N. "Opensource Gamification of a Computer Science Lecture to Humanities Students." *Methodologies and Intelligent Systems for Technology Enhanced Learning.* 2014. 119-126.

Declerck, T., et al. "Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data. " Francopoulo, G. (Ed.) LMF Lexical Markup Framework. London. 3/2013.

Declerck, T., Wandl-Vogt, E., Mörth, K. "A SKOS-based Schema for TEI encoded Dictionaries at ICLTT. " In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), ed. Nicoletta Calzolari u. a., Reykjavik, Iceland, ELRA, Paris, 5/2014.

Declerk, T., Wandl-Vogt, E. "How to semantically relate dialectal Dictionaries in the Linked Data Framework. " In: Kalliopi Zervanou u. a. (Ed.), Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Siences, and Humanities (LaTech 2014), Gothenburg, Sweden, ACL, 4/2014.

Deutscher, G. "Through the Language Glass. How Words Colour Your World. " London: 2010.

Esteban, A., Therón, R. "Corpusexplorer: supporting a deeper understanding of linguistic corpora. " *Smart Graphics.* Vol. 6815 (2011): 126-129.

Europäische Kommission: "Digital Science in Horizon 2020." 2013.

Ferro N. et al. "Fostering Interaction with Cultural Heritage Material via Annotations: The FAST-CAT-Way." In: T. Catarci et al. (Eds.): Bridging between Cultural Heritage Institutions, Proceedings of the 9th Italian Research on Digital Libraries (IRCDL 2013), CCIS Vol. 385, Berlin, Heidelberg. 2014. 41-52.

Finke, P. "Citizen Science: Das unterschätze Wissen der Laien." München. 2014.

Heath, T., Bizer, Ch. "Linked data: Evolving the web into a global data space." Synthesis lectures on the semantic web: theory and technology1.1 (2011): 1-136.

van Hooland,S., Verborgh, R. "Linked Data for Libraries, Archives and Museums. How to clean, link and publish your metadata." London: 2014.

Hoque, F., Bear, D. "Everything Connects. How to TRANSFORM and LEAD in the Age of Creativity, Innovation and Change." New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto. 2014.

Jagoda, P. "Gaming the Humanities." *differences* 25.1 (2014): 189-215.

Keim, D.A., Oelke, D. "Literature fingerprinting: A new method for visual literary analysis." *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*. IEEE, 2007.

McCrae, J., Declerck, T. et al. "Interchanging lexical resources on the Semantic Web." In: Language Resources and Evaluation. Vol. 46, Issue 4, 2012:701-719.

Nentwich M., König R. "Cyberscience 2.0: Research in the Age of Digital Social Networks." Frankfurt, New York: 2012.

Novotny, H., Scott, P., Gibbons, M. "Mode 2 Revisited: The New Production of Knowledge. " *Minerva* 41 (2003): 179-194.

Pink, S. "Interdisciplinary agendas in visual research: re-situating visual anthropology." *Visual studies* 18.2 (2003): 179-192.

Raddick, M. J., Bracey, G., Carney, K., Gyuk, G., Borne, K., Wallin, J., Jacoby, S., et al. (2009). Citizen Science: Status and Research Directions for the Coming Decade. AGB Stars and Related Phenomena Aastro2010 The Astronomy and Astrophysics Decadal Survey, 2010, 46P.

Reder, C. (Ed.). "Kartographisches Denken." New York 2012.

"Straffungskonzept für das Wörterbuch der bairischen Mundarten in Österreich (WBÖ)". Wien. 1998.

The Royal Society. "Science as an Open Enterprise". 2012.

Therón, R. , Fontanillo L. "Diachronic-information visualization in historical dictionaries." (2014).

Therón, R. et al. "Visual analytics: A novel Approach in corpus linguistics and the Nuevo Diccionario Histórico del Español. *Proc. of III Congreso Internacional de Lingü{\i}stica de Corpus.* 2011.

Therón, R., Wandl-Vogt, E. "The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually". LREC-Proceedings 2014.

Wandl-Vogt, E., Declerck, T. "Mapping a Traditional Dialectal Dictionary with Linked Open Data. " In: Kosem, I. et al. (Eds.): Electronic lexicography in the 21st century: thinking outside the paper. Proceedings. Ljubljana / Tallinn. 2013: 460-471.

Wiggins, A., Crowston K. "From Conservation to Crowdsourcing: A Typology of Citizen Science." Proceedings of the Forty-fourth Hawai'i International Conference on System Science (HICSS-44). 2011.

Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Wien. 1963-. online: 2012-.

Für eine pan-europäische Lexikologie und Lexikographie mittels des Linked Open Data Frameworks

Thierry Declerck, DFKI GmbH, Saarbrücken, Deutschland

Eveline Wandl-Vogt. ÖAW, Wien, Österreich

Einleitung

Das rasche Wachstum an Webinhalten erfordert innovative Lösungen für die automatische Analyse von solchen Inhalten. Nur mit derartigen Lösungen können Herausforderungen in Szenarien wie die umfangreiche Analyse und Interpretation von heterogenen Datenmengen verschiedener Sprachen, Medien und aus diversen Organisationen bewältigt werden.

Für die inhaltliche Analyse von täglich neu produzierten, heterogenen, mehrsprachigen und multimedialen Inhalten greifen sprachtechnologische Anwendungen immer häufiger auf sprach- und medienunabhängige Datenanalysen und Repräsentationsmethoden wie etwa Linked Data[1] und Semantic Web Technologien zurück.

Notwendig dafür ist die Darstellung von sprach- und medienspezifischen linguistischen Informationen auf einer semantischen Ebene. Nur so wird eine Form von Analytics möglich, welche auf die zunehmende Bandbreite von Medien und menschlichen Sprachen angewendet werden kann, welche sich heute im Web findet.

Dabei spielen lexikalische Ressourcen eine wesentliche Rolle. Und weiterdenkend, stellt sich die Frage inwiefern lexikalische Ressource nicht nur als Basis für die semantische Analyse von Webinhalten sich eignen, sondern ob sie nicht im gleichen Format im Web zu stehen haben als die Wissensobjekte, die sie ja auch beschreiben. Dies ist auch eine zentrale Frage der modernen digitalisierten Lexikographie: wie werden Wörterbücher konzipiert, in einer Zeit in der Sprachdaten unterschiedlichsten Typs in digitaler Form vorliegen und zugreifbar sind? Wie soll ein Wörterbuch in diesem neuen Kontext aussehen?

Wir beschreiben in diesem Beitrag, wie Vertreter von zwei Wissenschaftsgemeinden – Sprach- und Semantic Web Technologie auf der einen Seite, und Lexikologie und Lexikographie auf der anderen Seite, sich diese Fragestellung annehmen, auch im Rahmen von zwei Europäischen Projekten, die wir auch kurz beschreiben.

[1] S. http://linkeddata.org/ für Details.

Das LIDER Projekt

LIDER (http://www.lider-project.eu/) schafft die Grundlage für ein Ökosystem aus frei verfügbaren, verlinkten und semantisch interoperablen Ressourcen. LIDER untersucht, wie Sprache („Linguistic Linked Data"-Repräsentationen[2] von Korpora, Wörterbüchern, lexikalischen und syntaktischer Metadaten etc.) und multimediale Daten (Bild, Video etc.) als Basistechnologie für die Analyse unternehmensweiter, mehrsprachiger und cross-medialer Inhalte im Netz fungieren können.

LIDER hilft dabei, eine Community zur Linguistic Linked Licensed Data (3LD) zu gründen, in der unter Linguistic Linked Data sowohl frei verfügbare linguistische Ressourcen als auch lizensierte linguistische Daten verstanden werden.

LIDER arbeitet auch an einer Referenzarchitektur für das Erstellen von Linguistic Linked Data auf Basis von bereits existierenden und zukünftigen Plattformen sowie von frei verfügbaren Quellen.

Das ENeL Projekt

ENeL (http://www.cost.eu/domains_actions/isch/Actions/IS1305) ist eine so-genannte COST Aktion der Europäischen Union und der Kommission zur Unterstützung von paneuropäischer Forschung. ENeL zielt auf dem Aufbau eines Europäischen Netzwerks für e-Lexikographie (Enel).

Die Arbeitsgruppen von ENeL setzen sich mit der Tatsache auseinander, dass Computer und die Verfügbarkeit des World Wide Web (WWW) die Bedingungen für die Produktion und Rezeption von Wörterbüchern deutlich verändert haben. Für Redakteure wissenschaftlicher Wörterbücher ist das WWW nicht nur eine Quelle der Inspiration, sondern auch eine neue und große Herausforderung. Zum Beispiel wenn es darum geht, die Lücke zwischen der Öffentlichkeit und wissenschaftlichen Wörterbüchern zu schließen und dabei den Benutzern einfacher Zugang zu wissenschaftlicher Wörterbüchern zu gewährleisten. Es wird auch versucht, einen breiteren und systematischen Austausch von Know-how und gemeinsamen Standards und Lösungen zu schaffen und ein gemeinsames Konzept zu entwickeln, wie die Grundlage für eine neue Art der Lexikographie auszusehen hat. Darüber hinaus wird der paneuropäische Charakter der lexikographischen Arbeit in Europa in den Mittelpunkt gesetzt. Wie kann man von nationalen Wörterbücherprojekten zu supranationalen Wörterbüchern gelangen? Wie kann die Mehrsprachigkeit des Europäischen Bürgers optimal berücksichtigt werden? Gibt es paneuropäische Wörter oder Begriffe? Gibt es gemeinsame Neologismen, die in Wörterbücher berücksichtigt werden müssen? So dass sich die Frage stellt, ob es neben paneuropäischen Korpora (wie Europarl, s. http://www.statmt.org/europarl/), auch paneuropäische Wörterbücher geben kann, oder soll?

In diesem Beitrag präsentieren wir einen implementierten Ansatz zu einer Linked Data konformen Modellierung von lexikographischen Daten, die eine Vernetzung und

[2] Siehe auch http://linguistics.okfn.org/resources/llod/.

Integration von bestehenden multilingualen lexikalischen und enzyklopädischen Ressourcen erlaubt.

Linked (Open) Data und Linguistic Linked Open Data

Wir geben hier die Definition, die in Wikipedia steht: **„Linked Open Data (LOD)** bezeichnet im World Wide Web frei verfügbare Daten, die per Uniform Resource Identifier (URI) identifiziert sind und darüber direkt per HTTP abgerufen werden können und ebenfalls per URI auf andere Daten verweisen. Idealerweise werden zur Kodierung und Verlinkung der Daten das Resource Description Framework (RDF) und darauf aufbauende Standards wie SPARQL und die Web Ontology Language (OWL) verwendet, so dass Linked Open Data gleichzeitig Teil des Semantic Web ist. Die miteinander verknüpften Daten ergeben ein weltweites Netz, das auch als „Linked [Open] Data Cloud" oder „Giant Global Graph" bezeichnet wird. Dort wo der Schwerpunkt weniger auf der freien Nutzbarkeit der Daten wie bei freien Inhalten liegt (Open Data), ist auch die Bezeichnung **Linked Data** üblich." (http://www.wikiwand.com/de/Linked_Open_Data, konsultiert am 2014-11-10)

Die ersten solchen Datenmengen im LOD waren ursprünglich klassische Wissensobjekte, die enzyklopädischer Natur sind. Aber in den letzten Jahren sind Bestrebungen aktiv gewesen, auch linguistisches Wissen so zu kodieren, und eine „Linguistic Linked Open Data" Gemeinschaft ist entstanden (s. http://linguistics.okfn.org/resources/llod/). Eine graphische Darstellung des aktuellen Standes des „Linguistic Linked Open Data cloud diagram" ist auf einer Extraseite am Ende dieses Beitrages, im Anhang, wiedergegeben. Und genau in diesem Rahmen findet die Kooperation zwischen den ENeL und LIDER Projekten statt. Als Repräsentationsformalismus für die LOD konformen Modellierung von lexikographischen Daten, die ENeL Teilnehmer uns zu Verfügung gestellt haben, verwenden wird das Ontolex Modell, das im nächsten Abschnitt eingeführt wird.

Das Ontolex Modell

Das Ontolex Modell wird im Rahmen eines W3C Vorhabens[3], einer so-genannten Community Group, diskutiert und steht kurz vor der offiziellen Veröffentlichung. Es basiert auf dem Modell *lemon* (s. J. McCrae et al., 2012) und auch auf dem ISO Modell Lexical Markup Framework (LMF; http://www.lexicalmarkupframework.org/). Diese Modelle beschreiben einen modularen Ansatz zur Lexikonbeschreibung, und dadurch verliert die traditionelle Sicht an Bedeutung, dass der „Einstieg" zu einem Lexikoneintrag beim sogenannten „Headword" stattzufinden hat. Alle Elemente eines Wörterbucheintrages sind gleichwertig und können unabhängig voneinander beschrieben und durch expliziten Relationsmarker miteinander verbunden werden. Das neue mit *lemon* und Ontolex ist, dass die Komponente eines Lexikoneintrages im Netz verteilt werden können und durch RDF Relationen („properties") miteinander verlinkt werden. Praktisch heißt das, dass ein Wörterbuchautor nicht alle

[3] Siehe http://www.w3.org/community/ontolex/.

Komponente oder Elemente eines Eintrages detailliert beschreiben und an einem „Ort“ halten muss, aber dass sie/er auch auf bestehende Elemente (zum Beispiel die Etymologie eines Wortes) zurückgreifen kann, und einfach darauf verweisen kann. Wir sind überzeugt, dass diese Eigenschaften des Modells die Zusammenarbeit zwischen verschiedenen wissenschaftlichen Lexikographen ermöglichen und unterstützen können, und dass dadurch virtuelle Forschungsumgebungen im lexikographischen Bereich entstehen können.

Die RDF basierte Verlinkung gewährleistet, dass der Eintrag dennoch eine Einheit bleibt, in der Form eines genannten Graphs. Abbildung 1 unten zeigt eine graphische Darstellung von Ontolex.

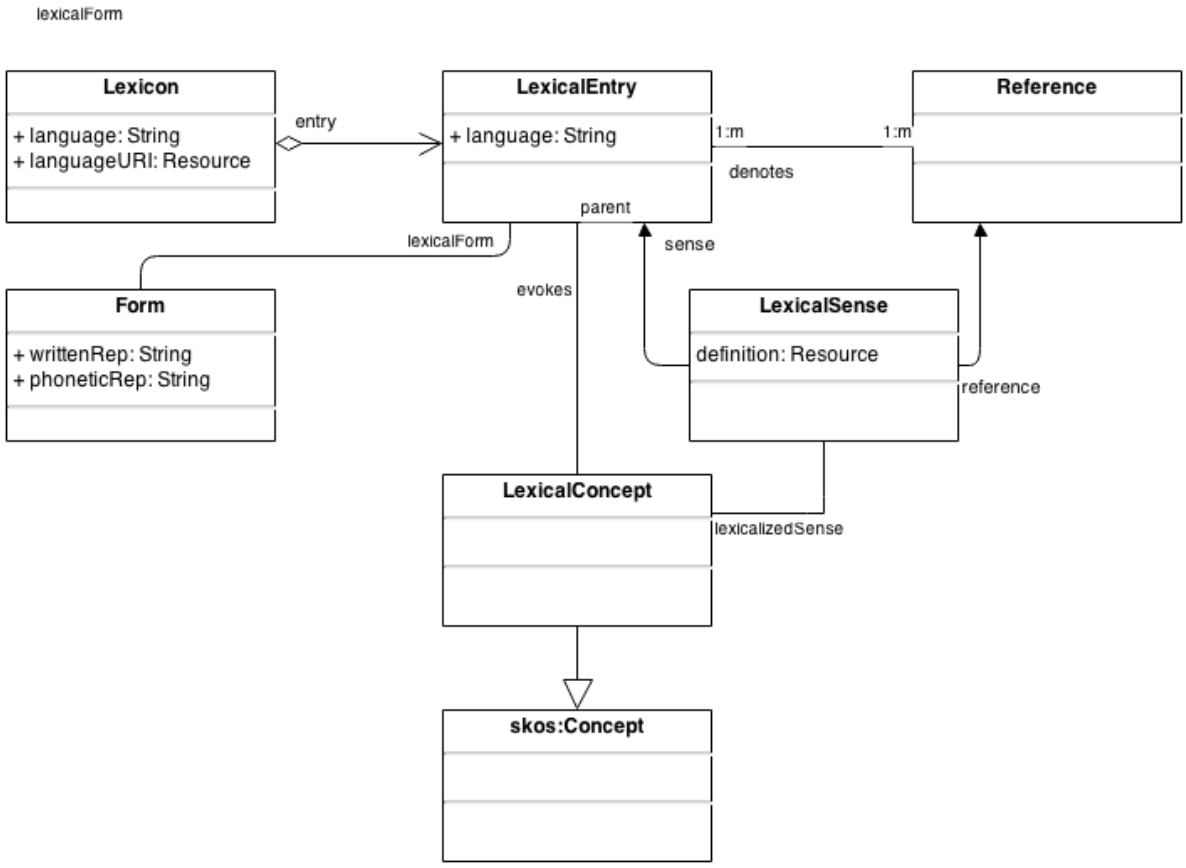


Abbildung 1: Graphische Darstellung des Ontolex Modells

Unsere Experimente

Von ENeL Teilnehmern haben wir lexikographische Daten erhalten, die wir dann in Ontolex umgewandelt haben. Es handelt sich momentan um:

- 2 Österreichische Dialektwörterbücher (in Tustep/XML und Word)
- Beispiele aus einem Slowakischen Wörterbuch (in XML, + PDF/Word)
- Ein Slowenisches Wörterbuch (in XML, basiert auf dem LMF Standard)
- 2 Wörterbücher von Arabischen Dialekten (in TEI kodiert)
- 1 Beispiel aus einem Baskisch-Deutsch Wörterbuch (in XML)
- 1 Beinspiel aus eine Französischen Wiktionary-basierten Wörterbuch

- 1 Konzept-basiertes Wörterbuch des Limburgischen (in Excel)
- 1 Beispiel aus dem multilingualen KDictionary (in XML)
- Beispiele aus dem Digital Scottisch Lexicon (Old Scottisch, html + 1 Beispiel in TEI)

Wir haben alle Daten zunächst “manuell” analysiert und einzelne Einträge manuell in ein Ontologie Editierwerkzeug nach den Ontolex Modell eingetragen. Waren wir mit der Modellierung zufrieden, sind dann Skripte geschrieben worden, die die einzelnen Quellen dann vollständig in das RDF Format von Ontolex übertragen haben. In manchen Fällen haben wir gesehen, dass Lemmata von Einträgen eine Bedeutung mit anderen Lemmata (auch aus anderen Lexikons) teilen, so dass direkte (auch multilinguale) Relationen zwischen solchen Elemente automatisch etabliert werden können.

Für einzelne Fälle haben wir dann (manuell) einen Link zwischen Lexikonelementen und externen enzyklopädischen Quellen eingefügt, um zu zeigen, wie lexikalischen Daten mit Daten eines anderen Typs effizient verlinkt werden können. Abbildung 2 unten zeigt ein Screenshot aus unserem Ontologie Editor, aus dem der Leser einige Elemente des Ontolex Modell sehen kann.

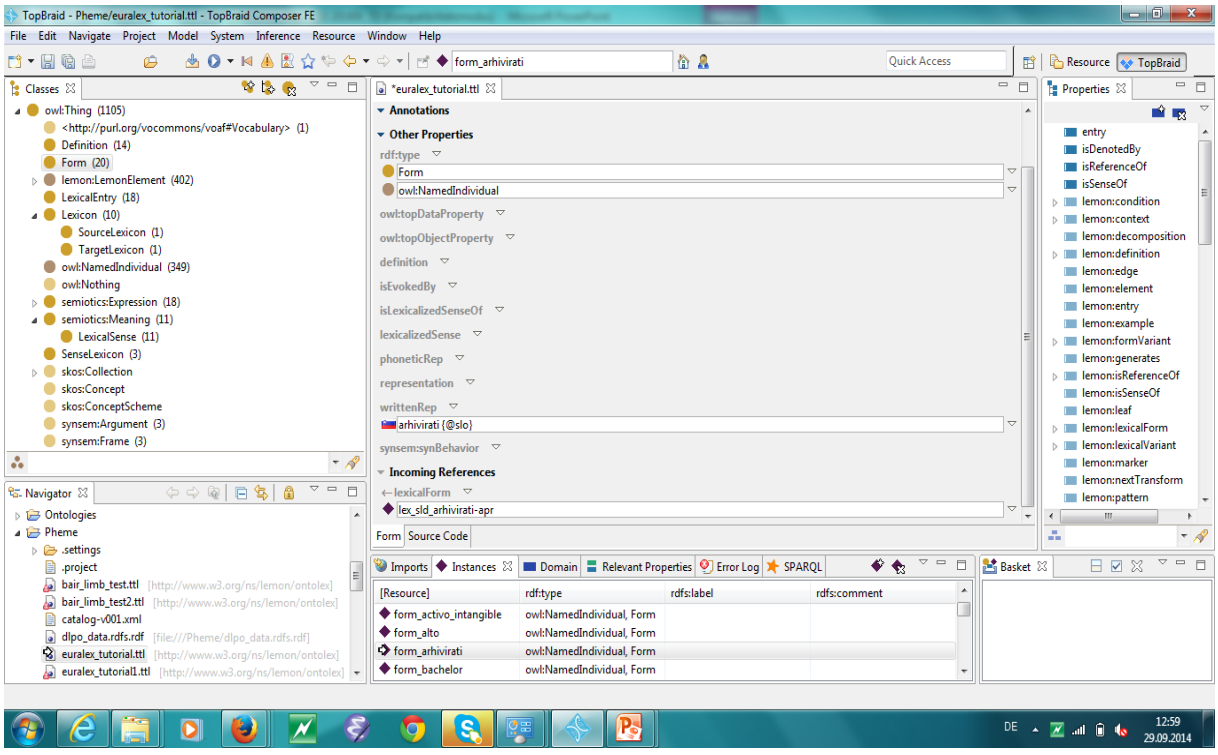


Abbildung 2: Ein Schnappschuss aus dem Ontologie Editor

Referenzen

Thierry Declerck, Eveline Wandl-Vogt. Cross-linking Austrian dialectal Dictionaries through formalized Meanings. in: Andrea Abel, Chiara Vettori, Natascia Ralli (eds.): *Proceedings of the XVI EURALEX International Congress, Pages 329-343.*

Thierry Declerck, Eveline Wandl-Vogt. How to semantically relate dialectal Dictionaries in the Linked Data Framework. *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014), Gothenburg, Sweden, ACL, 4/2014*

Maud Ehrmann, Francesca Cecconi, Daniele Vannella, John P. McCrae, Philipp Cimiano, and Roberto Navigli. A Multilingual Semantic Network as Linked Data: lemon-BabelNet**.** *Proceedings of the 3rd Workshop on Linked Data in Linguistics*

Philipp Cimiano and Christina Unger**. Multilingualität und Linked Data.** *In: Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien* (Editors: Tassilo Pellegrini, Harald Sack, and Sören Auer)

J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. Interchanging lexical resources on the Semantic Web**.** *Language Resources and Evaluation* (2012).

Georg Rehm and Felix Sasaki**.** Semantische Technologien und Standards für das mehrsprachige Europa (Editors: B. Humm Ege, B. and A. Reibold eds. Corporate Semantic Web)

Anhang

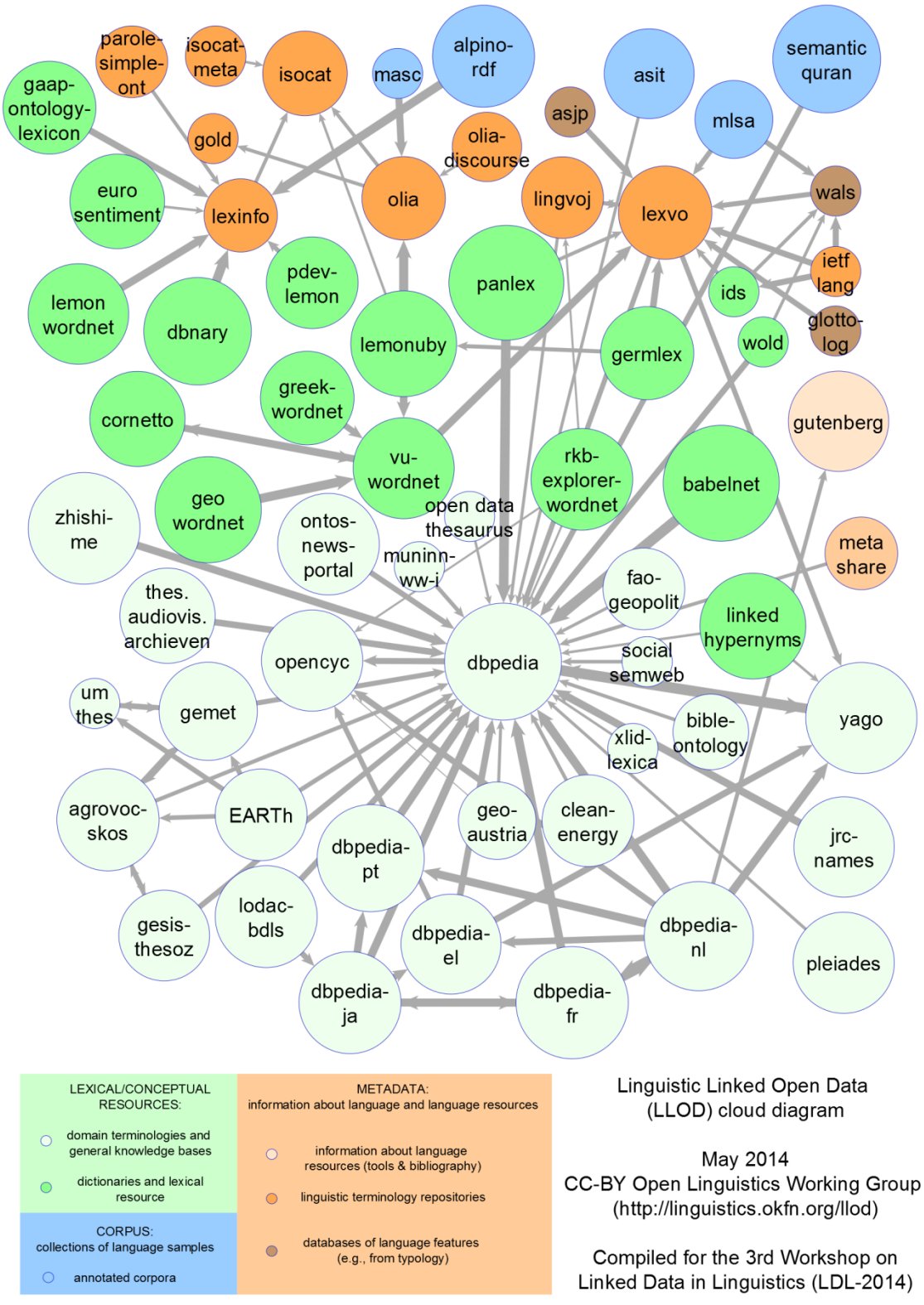


Abbildung 3: Die graphisch Darstellung des Linguistic Linked Open Data clouds

Integrierte Lexikographische Dienste zur Unterstützung der digitalen Geisteswissenschaften (aiLEs)

Karlheinz Mörth, Hannes Pirker

Austrian Center for Digital Humanities (ACDH) der Österreichischen Akademie der Wissenschaften

Digitale Wörterbücher stellen ein wichtiges Hilfsmittel für die geistes- und sozialwissenschaftliche Forschung dar, denn sie bieten nicht nur Unterstützung bei der schriftlichen Formulierung von Forschungsresultaten, sondern werden insbesondere als grundlegende Informationsquellen für die automatische Annotation und Analyse unterschiedlicher Textdaten benötigt. Indem sie als unabdingbare Basis verschiedener Verfahren der automatischen Textanalyse fungieren, bilden sie die Grundlage für die Erschließung textueller Informationsquellen, einem der zentralen Beiträge der *Digital Humanities* im gesamtwissenschaftlichen und -gesellschaftlichen Kontext.

Während in der Lexikographie, von der hier die Rede sein soll, digitale Methoden seit langer Zeit Anwendung finden, stehen frei verfügbare und qualitativ hochwertige Lexika nach wie vor nur in sehr begrenztem Maß zur Verfügung. Es werden immer mehr digitale Sprachdaten verfügbar, vertrauenswürdige lexikographische Daten, die von digital arbeitenden Linguisten und Lexikographen für ihre Forschungen verwendet werden können, gibt es nach wie vor kaum. Der Zugang zu existierenden Lexika wird sowohl durch technische als auch durch kommerzielle Hürden behindert. Der größte Teil derartiger Materialien wurde unter kommerziellen Gesichtspunkten erzeugt, und ist, wenn im Netz verfügbar, nur über Schnittstellen zugänglich, die es nicht erlauben, direkt mit diesen Daten zu arbeiten. Selbst für gut dokumentierte und digital gut erschlossene Sprachen wie Deutsch, Französisch, Spanisch usw. steht es um die Verfügbarkeit von lexikalischen Ressourcen nicht gut.

In diesem Beitrag wird die Erstellung einer neuen lexikalischen Datensammlung diskutiert, des Weiteren werden die organisatorischen und technischen Strategien aufgezeigt, durch die die so entstandenen Daten dauerhaft verfügbar gemacht werden.

AiLEs – *Austrian Integrated Lexicographic System*

aiLEs steht für *Austrian Integrated Lexicographic System* und ist ein Versuch, die eingangs beschriebene Situation zumindest ansatzweise zu verbessern. Im Rahmen von *aiLEs* sollen lexikalische Daten zur Verfügung gestellt werden (*xBaffle*). Die *technische* Verfügbarkeit der Lexika wird durch die Verwendung standardisierter XML-Formate für die Datenrepräsentation garantiert, der programmatische Zugriff auf die Daten durch die Bereitstellung von REST basierten Schnittstellen ermöglicht werden. Auf organisatorischer Ebene sollen die Daten gemäß der *open access* Philosophie barrierefrei zugänglich sein. Durch die Einbettung des Projekts in den Kontext der europäischen Infrastrukturkonsortien CLARIN und DARIAH soll sowohl die *technische* Langzeitstabilität – konkret durch die Bereitstellung der Ergebnisse im *Language Resources Portal* des CLARIN-Zentrums Wien[1] – als auch die nicht minder wichtige *institutionelle* Stabilität des Unterfangens garantiert werden.

Im Rahmen von *aiLes* ist mit *xBaffle* eine Reihe lexikalischer Ressourcen für unterschiedliche Sprachen konzeptioniert, wobei *Baffle* für *Basic Austrian Fullform Lexicon* steht, und das *x* im Namen jeweils durch den entsprechenden zweibuchstabigen Sprachidentifikator zu er-

1 http://www.oeaw.ac.at/icltt/ccv

setzen ist. Das erste digitale Wörterbuch, das aus diesem Projekt hervorgehen soll, ist *deBaffle*, ein deutsches Vollformenlexikon mit morphologischen Basisdaten, das eine möglichst gute Abdeckung der deutschen Gegenwartssprache zum Ziel hat.

Es wurden für das Deutsche in der Vergangenheit bereits umfangreiche morphologische Datenbestände (Morphy, Canoo) aufgebaut. Die meisten dieser Ressourcen sind jedoch nicht zugänglich, unvollständig oder nur gegen teures Geld verfügbar. Die Dokumentation der für die Erzeugung von *deBaffle* angewendeten Methoden und Werkzeuge soll es ermöglichen, in der Zukunft auch über das Deutsche hinaus ähnliche Sprachressourcen aufzubauen.

Lexikalischer Ausgangspunkt: Wiktionary

Als Ausgangspunkt für *deBaffle* wurde das deutschsprachigen Wiktionary[2] gewählt. Wiktionary ist ein Schwesterprojekt zur freien Enzyklopädie Wikipedia und arbeitet an Wörterbüchern, die wie die Wikipedia selbst, in kollaborativer Art und Weise manuell von lexikographischen Enthusiasten editiert werden. Die deutschsprachige Version ist im Vergleich zu anderen Wiktionaries verhältnismäßig umfangreich. Die Wörterbucheinträge verfügen oft über bemerkenswert umfangreiche linguistische Informationen. Neben Angaben zur Orthographie finden sich auch Daten zur Flexion, Morphologie, Phonologie und Semantik. Trotz zahlreicher Lücken im Material stellte sich das Material für unsere Zwecke als erstaunlich brauchbar heraus.

Eine Hürde für die Verwendung von Wiktionary-Daten in automatisierten Sprachverarbeitungsprozessen stellt das allen Wikiprodukten zugrunde liegende Repräsentationsformat dar. Hierbei handelt es sich um eine sogenannte wiki Sprache, die je nach Domäne und natürlichsprachlichem Kontext unterschiedliche Vokabulare verwendet und deren Syntax den Einsatz von Standardtools erschwert, in der Regel unmöglich macht. Um zeitgemäße Verfahren des Datenmanagements, wie z. B. automatisches Validieren zu erlauben, ist es nötig, derartige Daten in ein XML Format zu konvertieren. Aus einer Reihe von hierfür in Frage kommenden Zielformaten (LMF, RDF, …) wurde für das konkrete Projekt TEI P5[3] gewählt, da die gesamte verwendete Infrastruktur auf dieses Format zugeschnitten ist und langjährige Expertise mit der Applikation dieses Formats vorliegt (vgl. Budin et al. 2013).

Korpusbasierte Lexikonkonstruktion

Bei der Erstellung von *deBaffle* kommt eine korpusbasierte inkrementelle Methode zur Anwendung. Die Grundidee dabei ist, Informationen aus bereits verfügbaren lexikalischen Ressourcen – in diesem Fall dem Wiktionary – mit Daten aus möglichst umfangreichen Textkorpora zu verknüpfen.

Der Abgleich der Lexikoninhalte mit der tatsächlichen Sprachrealität in Texten bietet die Möglichkeit, den bestehenden Abdeckungsgrad durch das Lexikon laufend zu evaluieren, und so den Prozess der Lexikonerweiterung durch die Identifikation systematischer Lücken zu optimieren. Insbesondere kann aber das Korpus helfen, sowohl die Korrektheit bestehender lexikalischer Informationen zu überprüfen, als auch Informationen über noch unbekannte Wortformen abzuleiten.

deBaffle ist als Vollformlexikon konzipiert, d. h., alle Flexionsformen eines Lemmas sind explizit im Lexikon aufgeführt. Dieses Lexikondesign wird in der automatischen Sprachverarbeitung bevorzugt verwendet, weil es erlaubt, die morphosyntaktische Kategorie einer Wortform durch einfaches Nachschlagen im Lexikon zu ermitteln, und somit auf eine morphologische Analysekomponente zu verzichten. Eine zentrale Aufgabe bei der automatischen

2 http://de.wiktionary.org/

3 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Erstellung eines neuen Lexikoneintrags für *deBaffle* ist somit die Festlegung der Wortklasse und des Flexionsparadigmas eines noch unbekannten Wortes, um in der Folge alle Flexionsformen zu einem Lemma zu generieren.

Die korpusbasierte Methodik in *deBaffle* soll anhand der Substantivflexion beispielhaft demonstriert werden. Zur Ermittlung des Flexionsparadigmas bei Substantiven muss immer zumindest das grammatische Geschlecht und eine Pluralform bekannt sein.

Unser Ansatz nutzt nun den Umstand, dass bestimmte morphosyntaktische Muster den Kasus eines Substantivs determinieren können. Beispielsweise kann aus dem Auftreten eines Musters wie „(eines|des) *Subst*“ im Korpus abgeleitet werden, dass das Substantiv männliches oder sächliches Geschlecht aufweist und im Genitiv steht. Ein Muster wie „(viele|einige) *Subst*“ identifiziert Substantive im Plural usw. Durch die inkrementelle Anwendung und schrittweise Verfeinerung solcher Regeln können Hypothesen über die Wortart und das Flexionsparadigma eines Wortes gebildet werden. So können Informationen für neue Lexikoneinträge automatisch identifiziert, aber auch bestehenden Einträge verifiziert werden.

Korpusdaten: Austrian Media Corpus (AMC)

Als Basis für die korpusbasierte Methodik von *deBaffle* dient das Austrian Media Corpus (AMC), eine von der Austria Presse Agentur (APA) kompilierte Sammlung von Texten aus österreichischen Zeitungen und Magazinen der letzten 20 Jahre (Ransmayr et al. 2013). Mit einem Umfang von derzeit ca. 6 Mrd. Wörtern und seinem kontrollierten Genre-Repertoir bietet das Korpus einen repräsentativen Querschnitt durch den aktuellen schriftlichen Sprachgebrauch in Österreichs Medienlandschaft. Das Korpus wurde mit zwei unterschiedlichen *Part-of-Speech (PoS) Taggern* (RFTagger und TreeTagger) mit Wortarteninformation und auch Flexionsinformation angereichert. Diese automatisch generierten Informationen sind zwar naturgemäß fehlerbehaftet, sollen aber dennoch im oben beschriebenen iterativen Verarbeitungszirkel miteinbezogen werden. Durch Abgleich der Annotationsentscheidungen der beiden PoS-Tagger untereinander, mit Informationen aus dem Wiktionary und den Ergebnissen der morphosyntaktischen Testmuster sollen die verschiedenen Informationsquellen evaluiert, und in der Folge verbessert werden. Das annotierte Korpus bietet also eine Grundlage zur Konstruktion des Lexikons, ist aber gleichzeitig über mittelfristig verbesserte Annotationen wiederum Nutznießer dieses Prozesses.

Statistische Informationen für das Lexikon

Durch den Abgleich zwischen lexikalischen Einträgen und Textkorpus liegen statistische Informationen zur Auftretenshäufigkeit einer Wortform oder eines Lemmas vor. Diese Daten werden in *deBaffle* zur Verfügung gestellt, und dienen sowohl den Lexikographen bei der Erstellung und Pflege des Lexikons, als auch den künftigen Anwendern der Lexika.

Durch den Abgleich von Lexikon und Korpus lassen sich etwa die quantitativ vordringlichsten Lücken in der lexikalischen Abdeckung identifizieren, die freilich immer auch einen korpusspezifisch *bias* aufweisen. So zählen beispielsweise im AMC die Akronyme österreichischer politischer Parteien zu den häufigsten Wörtern, die von den PoS-Taggern nicht identifiziert werden konnten. Auffällige Lücken finden sich auch bei – insbesondere österreichischen – Toponymen. Hier zeigen sich systematische Verzerrungen, verursacht durch die unterschiedliche geografische Herkunft der lexikalischen Ressourcen einerseits und des AMC andererseits. Das Beispiel der mangelhaften Abdeckung der Toponyme legt die Integration zusätzlicher externer Wissensquellen – etwa Ontologien mit Geodaten – zur Ergänzung der Lexika als sinnvollen zusätzlichen Schritt nahe.

Manuelle Verifikation

Alle automatischen Schritte produzieren auch fehlerhafte Daten, die schlussendlich nur durch manuelle Verifikation eliminiert werden können. Für diesen notwendigen Bearbeitungsschritt steht mit dem am ICLTT entwickelten Viennese Lexicographic Editor (VLE)[4] zumindest ein effizientes Werkzeug zur Verfügung, das für die Produktion von TEI-konformen Lexikoneinträgen optimiert wurde (Budin et al. 2013).

Status und Ausblick

Das gegenwärtig im Aufbau begriffene digitale Wörterbuch gehört zu einer Reihe von Tools die im Rahmen des Österreichischen Zentrums für digitale Geisteswissenschaften (ACDH) als Ergänzung des bestehende Toolinventars geplant ist. Es soll in Zukunft als Teil des österreichischen Engagements in den europäischen Infrastrukturkonsortien CLARIN und DARIAH weitergepflegt werden.

Im Moment enthält Baffle nur deutschsprachige Daten. Es wird am ACDH aber bereits an ähnlichen Sprachressourcen für andere Sprachen gearbeitet.

Referenzen

Budin, G., Moerth, K., Ďurčo, M. (2013). European Lexicography Infrastructure Components. In: *Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estland.*

Ransmayr, J., Moerth, K., Ďurčo, M. (2013). Linguistic variation in the Austrian Media Corpus. Dealing with the challenges of large amounts of data. In: *Proceedings of International Conference on Corpus Linguistics (CILC), Alicante, Spanien.*

4 https://clarin.oeaw.ac.at/ccv/vle

Neue Erkenntnisse durch digitalisierte Geschichtswissenschaften? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern.

Während mit computerlinguistischen und auf Statistik beruhenden Methoden in der Verarbeitung natürlicher Sprache interessante neue Ergebnisse erzielt und neue Forschungsfragen aufgeworfen worden sind (Autorenschaftszuweisungen, Plagiatserkennung etc.), scheint die digitale Geschichtswissenschaft jenseits der Computerlinguistik bislang recht konventionell daher zu kommen. Die Mehrzahl der Angebote konzentriert sich auf Information Retrieval sowie digitale Editionen und soll mithin den klassischen Forschungsprozess erleichtern. Diese Entwicklung lässt sich direkt auf die Forschungsfragen und die Methoden dieser Fächer zurückführen: der historischen (Re-)Konstruktion historischen Geschehens auf der Grundlage von Einzelinformationen, die oft mühsam aus verschiedenen Quellen zusammengetragen werden. Welchen Mehrwert bietet die kritische digitale Edition in diesem Zusammenhang über die größere Freiheit der Darstellung der Textgenese hinaus, wenn die gewonnen Einzelinformationen am Ende erneut im Kopf des Forschers neu zusammengesetzt werden? Und wenn eine computerbasierte, (teil)autonome Form der Informationsverknüpfung schon möglich wäre, würde man diese überhaupt wünschen?

Peter Haber hat in seinen Arbeiten die Entwicklung, aber auch zukünftige Potentiale der digitalen Geschichtswissenschaft mit Weitblick beschrieben. Bislang sind digitale Werkzeuge für Historiker oft nur die moderne Form des analogen Zettelkastens, mit dessen Hilfe schon im 19. Jahrhundert erstaunliches geleistet wurde. Welche Bereiche der digitalen Geschichtswissenschaft kann man also heute von wirklich neuen Methoden sprechen, die auch grundsätzlich neue Erkenntnisse erwarten lassen? Drei Bereiche erscheinen mir dabei besonders vielversprechend:

In den aktuellen Methoden der **Netzwerkanalyse** lassen sich in der Tat neue Einsichten in soziale Beziehungsstrukturen gewinnen, die aufgrund ihrer Komplexität früheren Forschergenerationen grundsätzlich verschlossen blieben. Nicht zuletzt aus diesem Grund einer qualitativen Neubewertung von seriellen prosopographischen Quellen findet diese Methode immer breitere Akzeptanz in den Geschichtswissenschaften.

Noch mutet OCR für Manuskripte ein wenig wie Science Fiction an, aber spätestens seitdem Frederic Kaplan den Plan einer Digitalisierung der Venezianischen Handschriften im dortigen Staatsarchiv (Venice Time Machine) verkündete, scheinen Optical Character and Structure Recognition von Manuskripten greifbarer geworden zu sein. Hier geht es um die Beziehung von **Digitalen Methoden** zu den **historischen Hilfswissenschaften** – ein bislang nicht ausreichend beleuchteter Bereich der digitalen Geschichtswissenschaft, der, ähnlich wie bei google books, aufgrund der reinen Quantität des so verfügbaren Materials einen qualitativen Durchbruch ermöglichen könnte. Dies gilt übrigens grundsätzlich auch für die born digital Quellen der Zeitgeschichte, die gerade erst entstehen und in Zukunft ganz neue Formen digitaler Hilfswissenschaft mit Blick auf die langfristige Lesbarkeit von Datenformaten notwendig machen werden.

Im Rahmen des Konzepts eines **Semantic Web** auf Grundlage von maschinenlesbaren Beschreibungen von Ressourcen rücken die Möglichkeit der künstlichen Intelligenz und des automatisierten „reasoning“ zumindest theoretisch in erreichbare Nähe. In der Realität wird die Extraktion impliziten Wissens aber in den Geschichtswissenschaften bislang kaum genutzt. Facettierungen bieten Hilfe bei der Beurteilung von Suchergebnissen, aber wirkliche „künstliche Intelligenz“ ist dies noch nicht. Offensichtliche Fehler in Schemata können erkannt werden, aber können Computer schon kreativ Informationen verknüpfen und daraus einen neuen „Gedanken“

generieren, der Ausgangs- oder Endpunkt einer Forschungsfrage sein könnte? Immerhin bringt das gemeinsame Datenformat die Möglichkeit mit sich, Informationen in neuer Weise miteinander automatisiert zu verknüpfen und so neue Zusammenhänge aufzudecken.

Der Beitrag diskutiert anhand der drei oben skizzierten Beispiele Chancen und Grenzen sowie die hermeneutische Reichweite digitaler Methoden in den Geschichtswissenschaften vor einem erkenntnistheoretischen Hintergrund. Besondere Aufmerksamkeit liegt dabei auf den Erkenntnismöglichkeiten, die in der Zusammenschau und automatischen Verknüpfung von verteilten Daten bzw. Informationen unter einer gemeinsamen Oberfläche liegen. Linked Data kann heute schon aktiv genutzt werden, um digitale Ressourcen miteinander zu verknüpfen und damit Zusammenhänge evident zu machen, die bislang erst nach serieller Rezeption im Kopf des oder der Forschenden entstanden. Mit sog. „**Mashups**“ können so qualitativ neue Ergebnisse aus schon in standardisierten Datenformaten digital vorliegenden Informationen gewonnen werden, die das Potential besitzen, bislang unbekannte Zusammenhänge sichtbar(er) werden zu lassen. In diesem Kontext spielen Verknüpfungspunkte wie die Normdaten zu Personen und Orten aber auch die dbpedia oder Ontologien als (vereinfachte) Konzepte des Weltwissens eine entscheidende Rolle.

Literatur:

Alves, Daniel (Hg.) (2014): Digital Methods and Tools for Historical Research: A Special Issue. International Journal of Humanities and Arts Computing, Bd. 8, Heft 1 (April).

Düring, Marten & Ulrich Eumann (2013): Historische Netzwerkforschung: Ein neuer Ansatz in den Geschichtswissenschaften, in: Geschichte und Gesellschaft, Bd. 39, Heft 3, S. 369-390.

Dougherty, Jack & Kristen Nawrotzki (Hg.) (2013): Writing history in the digital age / Ann Arbor: University of Michigan Press.

Endres-Niggemeyer, Brigitte (Hg.) (2013): Semantic Mashups. Intelligent Reuse of Web Resources, New-York: Springer.

Haber, Peter (2011): Digital Past. Geschichtswissenschaft im digitalen Zeitalter, München.

Knap, Tomáš, Michelfeit J., Nečaský M. (2012): Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality, in 36th Annual IEEE Computer Software and Applications Conference Workshops, Izmir, Turkey, IEEE Computer Society, S. 106-111.

Meroño-Peñuela, Albert, Schlobach, Stefan, van Harmelen, Frank (2013): “Semantic Web for the Humanities”, In: Proceedings of the 10th Extended Semantic Web Conference, ESWC 2013, Montpellier, France, May 28-30, 2013. Philipp Cimiano et al. (Eds.), Lecture Notes in Computer Science 7882, Berlin Heidelberg: Springer, S. 645-649.

Reichert, Ramón (2014): Einführung, in: Big Data, hg. von Ramón Reichert, Bielefeld, S. 9-31.

Rosenthaler, Lukas, Fornaro, Peter, Clivaz, Claire (2014): National Data Curation and Service Center for Digital Research Data in the Humanities, Conference Proceedings DH 2014, Lausanne, S. 338-340.

Sacramento, Eveline R.; Casanova, Marco A.; Breitman, Karin K.; Furtado, Antonio L.; Macĕdo, José Antonio F. de; Vidal, Vănia M. P. (2012): Dealing with inconsistencies in linked data mashups, Proceedings of the 16th International Database Engineering & Applications Symposium [http://delivery.acm.org/10.1145/2360000/2351496/p175-sacramento.pdf].

Sahle, Patrick (2013): Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels, 3 Bände, Norderstedt: Books on Demand. (Schriften des Instituts für Dokumentologie und Editorik 7-9).

Sequeda, Juan , Marcelo Arenas, Daniel P. Miranker (2012): On directly mapping relational databases to RDF and OWL. WWW 2012, S. 649-658.

Sequeda, Juan, Miranker, Daniel P. (2013): Ultrawrap: SPARQL execution on relational data. Journal of Web Semantics, Bd. 22, S. 19-39.

Wettlaufer, Jörg & Sina Westphal (2014): "Digital Humanities", Der Archivar, Heft 3, S. 270-277.

Zaagsma, Gerben (2013): On Digital History, Low Countries Historical Review, Bd. 128, Heft 4, S. 3-29.

Dr. Jörg Wettlaufer

Göttingen Centre for Digital Humanities (GCDH)

Akademie der Wissenschaften zu Göttingen (ADWG)

Papendiek 16

37073 Göttingen

Germany

Tel. +49 551 39 20477

email: jwettla@gwdg.de

http://www.gcdh.de/en/people/researchers-project-staff/joerg-wettlaufer/

Big Data und Data Mining in den Digital Humanities

Big Data in den Geisteswissenschaften ermöglicht die Validierung und Extrahierung von Hypothesen aus großen Datenmengen wie sie es nie zuvor möglich war. In Sprachwissenschaften zum Beispiel können so Aussagen überprüft werden die sich nicht nur auf einzelne Textbeispiele beziehen sondern auf eine Verteilung über große Textmengen. So können Aussagen über Genres oder zeitspezifische Phänomene überprüft werden. In diesen Beitrag beschreiben wir Ergebnisse aus einem Projekt aus den Bereichen Korpuslinguistik und Data Mining. Auf Basis großer Datenmengen unterstützen wir mittels Data Mining Methoden linguistische Analysen.

In der Analyse von Sprache und allgemein in der Sprachforschung spielen große Textkorpora immer größere Bedeutung. Bei einem Korpus handelt es sich um eine Kollektion von Texten mit zusätzlichen Annotationen. Als Annotation verstehen wir eine Klassifizierung oder Beschreibung von Textelementen. Diese Annotationen können auf Textebene, Satzebene oder Wortebene vorliegen. Annotationen für Wörter können deren syntaktische Klasse sein oder verschiedene andere Schreibweisen des Wortes sein. Annotationen für Sätze können deren syntaktische Struktur als Parse-Baum darstellen. Annotation von ganzen Texten können Metainformationen wie deren Autoren, die Quelle, die Textsorte oder das Datum der Veröffentlichung sein.

Im Vergleich zu Anfragen an einen Korpus stellen Anfragen an eine Suchmaschine im Internet wie Google ungenügend Informationen zur Verfügung. So ist die Quellenlage bei einem Korpus geklärt. Dies ist bei der Sprachforschung besonders wichtig, da die Glaubwürdigkeit der Ergebnisse von den Quellen anhängt. Ferner können spezielle Features wie die oben genannten Annotationen berücksichtigt werden. Zum Beispiel kann man häufig auf linguistischen Korpora spezielle linguistische Features bei Anfragen berücksichtigen. So kann mit der Anfrage: http://www.dwds.de/?qu=bringen+with+%24p%3DVVFIN nach vorkommen des Verbs „bringen“ als finites Verb in den Texte gefragt werden.

Eine wichtige Aufgabe in den Sprachwissenschaften ist die Lesartendisambiguierung zur Erfassung von Mehrdeutigkeiten in der deutschen Sprache. Dabei werden für ein gegebenes Wort unterschiedliche Bedeutungen auf Basis des Kontextes in dem dieses Wort vorkommt bestimmt. Nachdem mögliche Bedeutungen ermittelt wurden, kann einem Vorkommen dieses Wortes dann eine Bedeutung, oder Lesart, zugewiesen werden. Ein Beispiel für solch eine Disambiguierung ist der Webservice Babelnet. Das Wort „Ampel“ kann man hier auf seinen unterschiedlichen Bedeutung hin untersuchen, wie man hier sehen kann: http://babelnet.org/exploreResult?word=Ampel&lang=DE

Diese Bestimmung der unterschiedlichen Bedeutungen bassiert auf den Vorkommen und dem Kontext des Wortes „Ampel“ in Wikipediaartikeln zum Beispiel. Dieser Datenbestand ist jedoch für sprachwissenschaftliche Untersuchungen zu ungenügend. So wird die Bedeutung des Wortes „Ampel“ als Blume nicht gefunden. Dies liegt an der Tatsache, dass „Ampel“ in dieser Bedeutung eher früher verwendet wurde als heute. In Wikipediaartikeln findet man

„Ampel“ hingegen nur in den heute gängigsten Bedeutungen als Lichtsignalanlage und als Lebensmittelampel.

Weiterhin sind die Entwicklung der Bedeutungen und die Verteilung dieser auf unterschiedlichen Genres hier nicht enthalten. Große Textkorpora wie sie von der Berlin Brandenburger Akademie der Wissenschaften (www.dwds.de) angeboten werden bieten hingegeben Texte aus unterschiedlichen Genres und Zeiten an. So kann man die Verteilung einer bestimmten Bedeutung eines Wortes in Zeitungsartikeln im Vergleich zur Belletristik untersuchen. Auch die Entwicklung der Bedeutungen über die Zeit ist so möglich.

In dem vom BmBF (Bundesministerium für Bildung und Forschung) geförderten Projekt KobRA http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Hauptseite werden unterschiedlichen Methoden zur Ermittlung und Zuweisung der verschiedenen Bedeutungen bestimmter Wörter mit Hilfe von Data Mining und Maschinellen Lernen entwickelt und evaluiert. Hierbei werden große Textkorpora wie das Deutsche Textarchiv (www.deutschestextarchiv.de) zusammen mit Informationen über die Textsorten und Veröffentlichungzeitpunkt mit berücksichtigt.

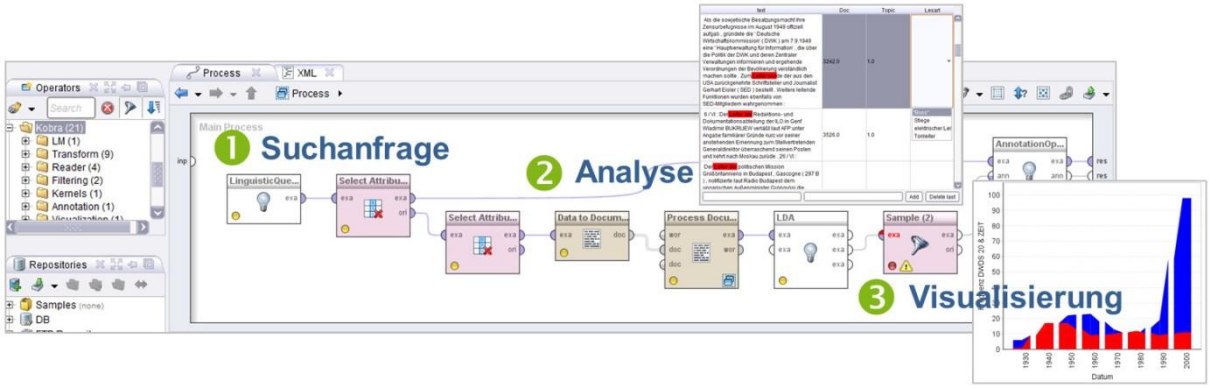


Abbildung 1: Visualisierung des Prozesses der Lesartendisambiguierung

Zur Bestimmung der unterschiedlichen Lesarten verwenden wir Topic Modelle wie Latent Dirichlet Allocation (Blei et al., 2002). Dabei wird ein probabilistisches Model ermittelt welches die Wörter auf eine vorgegebene Anzahl an möglichen Topics oder Lesarten verteilt. Ein im Rahmen des Projektes entwickeltes Tool ermöglicht es den Prozess der Disambiguierung eines Wortes komplett durchzuführen. Dieses Tool ist ein Plugin in dem bereits erfolgreich bestehenden Data Mining Werkzeug Rapidminer (https://rapidminer.com/). In Abbildung 1 ist ein Prozess zur Disambiguierung eines Wortes dargestellt. Zuerst werden Texte, die ein bestimmtes Wort enthalten (zum Beispiel Ampel), von einem Textkorpus extrahiert. Hierbei können auch die oben erwähnt linguistische Features mit angefragt werden. Die erhaltenden Textbeispiele aus dem Korpus können nun visuell untersucht und eventuell schon per Hand annotiert werden, so dass sie einer bestimmten Lesart angehören. Dies ist wichtig, da man so einen Goldstandard erhält an dem man die später automatisch ermittelten Lesarten bewerten kann. Nach einer Umwandlung der Texte in eine interne Repräsentation wird ein Topic Model berechnet. Dieses Model ermittelt die Wahrscheinlichkeiten, dass ein

bestimmtes Wort oder der ganzer Text zu einem bestimmten Topic (Lesart) gehört. Ein wichtiges Merkmal unserer Methoden ist, dass man die Möglichkeit hat die Entwicklung dieser gefundenen Lesarten über die Zeit und über Textsorten zu ermitteln und zu visualisieren.

Neben der visuellen Darstellung der Ergebnisse ermitteln wir ferner auch Gütemaße die die automatisch extrahierten Lesarten mit dem per Hand vorgegebenen Goldstandard vergleichen. Wir benutzen Normalized Mutual Information (NMI) und den F1 Score um die Lesarten aus dem Topic Model zu bewerten. NMI misst wie viele Texte mit gleicher Lesart im Goldstandard auch von unseren Topic Model der gleichen Lesart zugeordnet werden (Manning et al. 2008, p. 357f). Der F1 Score ist der relative Mittelwert aus den richtig der gleichen Lesart zugewiesenen Texte und der Anzahl aller dieser Lesart zugewiesenen Texte (Navigli et al. 2010).

Mit unseren Software Tool haben wir Experimenten zur Disambiguierung des Wortes „Leiter“ durchgeführt. Wir haben aus dem DWDS Kernkorpus (www.dwds.de) Sätze extrahiert, die das Wort „Leiter“ enthalten. Davon wurden 30 Prozent per Hand annotiert und zur Evaluierung verwendet. In der Tabelle in Abbildung 2 haben wir für die extrahierten Lesarten die Wörter aufgelistet, die die höchste Wahrscheinlichkeit haben in dieser Bedeutung zusammen mit „Leiter“ aufzutreten. Fernen haben wir in der Tabelle in Abbildung 3 die Wahrscheinlichkeiten aufgelistet, dass in einem bestimmten Genre eine dieser Lesarten verwendet wird.

Es wird jedem Text eine Bedeutung zugeordnet, die am wahrscheinlichsten ist, gegeben die Wörter in diesem Text. Mit dieser Zuordnung berechneten wir die NMI und den F1 Score. Für das ermittelte Topic Model ergibt sich eine NMI von 0,2573 und ein F1 Score von 0,7416. Wenn wir die Tabelle in Abbildung 2 anschauen, sehen wir dass die Bedeutung 2 wahrscheinlich „Leiter“ in der Bedeutung von Trittleiter meint. Die Bedeutungen 3 und 4 hingegen beinhalten eher Begriffe die auf den „Leiter“ als politischen Leiter deuten. Wenn wir nun in der Tabelle in Abbildung 3 die Verteilung der Bedeutungen über die Genres anschauen, sehen wir dass „Leiter“ in der Bedeutung als Trittleiter eher in Belletristik auftaucht als zum Beispiel in Zeitungsartikeln.

Bedeutung 1	**Bedeutung 2**	**Bedeutung 3**	**Bedeutung 4**
Musik	Stehen	DDR	Regierung
Berlin	Sehen	SED	Haben
Professor	Oben	Partei	Berlin
Komposition	Oberhalb	Politisch	ZK

Abbildung 2: Häufigste Wörter der ermittelten Bedeutungen

Leiter	**Bedeutung 1**	**Bedeutung 2**	**Bedeutung 3**	**Bedeutung 4**
Belletristik	0,0117707368	0,5777281878	0,0781636158	0,0706102457
Gebrauchsliteratur	0,059760642	0,1284352487	0,6671553638	0,0917891152

Wissenschaft	0,8194956525	0,0792926965	0,012734051	0,010994093
Zeitungen	0,1089729687	0,2145438669	0,2419469694	0,826606546

Abbildung 3: Wahrscheinlichkeit, dass ein Wort in mit ein bestimmten Bedeutung in einem Genre vorkommt

Mit den von uns entwickelten Methoden wird es künftig möglich sein quantitative empirische Untersuchungen auf großen Textkorpora durchzuführen, zu visualisieren und zu evaluieren. Die Integration von speziellen linguistischen Features und Metainformationen über die Texte ermöglichen es komplexe linguistische Analysen durchzuführen. In der Zukunft wollen wir eine noch nahtlosere Anbindung an externe Informationsquellen wie Wortnetzen (www.wordnet.de) und Baumdatenbanken wie die Tüba-DZ (http://www.sfs.uni-tuebingen.de/de/ascl/ressourcen/corpora/tueba-dz.html) in die Lesartendisambiguierung ermöglichen. Dadurch wollen wir nicht nur große Datenmengen nutzen sondern auch heterogene Datenquellen einbauen.

Blei, David M., Ng, Andrew Y., and Michael I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 116-126.

Title

Exploring the reuse of data in the humanities by means of asynchronous collaboration and authorship in nodegoat.

Abstract

nodegoat is a web-based research environment that facilitates an object-oriented form of data management with an integrated support for diachronic and spatial modes of analysis. This research environment has been designed to allow scholars to determine and design custom relational database models. The environment can be used in self defined collaborative configurations with varying clearance levels for different groups of users. Due to the focus on relations and associations between heterogeneous types of objects, the platform is equipped to perform analyses spanning multitudes of objects.

In order to facilitate a practice of reuse of datasets in the humanities, an ecosystem has to exist in which scholars can publish their datasets, correctly attribute this data according to the roles played by each author, share these datasets and allow for various scenarios of reuse. Current publication channels do not allow for complex authorship attribution (Nyhan and Duke-Williams, 2014). In this paper we will explore reuse scenarios by means of an object-oriented referencing system in which datasets, data selection, entities and records are all referenceable objects with uniquely identifiable authors. Once a reference has been made to any of these objects, a citation is automatically determined based on the position of the referenced object in the network and all their corresponding authors. This object-oriented referencing system paves the way for various scenarios of reuse and processes of asynchronous collaboration.

Publishing research data transcends traditional citation practices on three levels. Firstly, publishing data may happen before any synthesised text is in sight. Secondly, research outcomes in the form of data can have an extended life cycle that stretches far beyond the reach of a static text. Thirdly, research data that would not have been included in the final syntheses can still be published as data and find its way to a wider audience. These opportunities show the potential of publishing data in the humanities. Still, a number of challenges has to be overcome to arrive at the position in which scholars in the humanities will directly publish their data.

One of the most prominent challenges we still face is the awarding of academic credit for publishing datasets (Nowviskie, 2011). As Cloudine Mouline has stated, we need a 'change of publication cultures and recognition of these new publication cultures as equal to traditional ones'. Next to the monograph and the article, results and achievements in the form of the database, data visualisation, the scientific blog and micropublications in different forms should be recognised as well (Mouline, 2013).

In 2014, Dutch research institute Huygens ING together with the University of Amsterdam (UvA), the Free University (VU), the Royal Dutch Institute in Rome (KNIR) and LAB1100 led by Charles van den Heuvel ran a project that relied on asynchronous collaboration.[1] For this project, 'Mapping Notes and Nodes in Networks', multiple existing

[1] https://www.huygens.knaw.nl/mapping-notes-and-nodes-in-networks/?lang=en

datasets were brought together and manually enriched in order to map meaningful relationships between artists and intellectuals by combining biographical data with relevant contextual information for the history of the creative industry. Three complementary, but heterogeneous datasets Biographical Reference Works (Huygens ING), Ecartico (UvA) and Hadrianus (KNIR) were integrated in nodegoat.[2]

In the course of the project a number of researchers carried out individual research projects within the research environment that contained the three datasets. This led to a productive form of asynchronous collaboration as all the biographical data about artists and intellectual available in the existing datasets was used as context for new research questions. This prosopographical information was subsequently enriched with information about society membership in Italy (the Accademie). By adding this data, research questions regarding weak ties between these societies could be explored.

Lisa Spiro has developed a comprehensive overview of collaborative practices in the digital humanities (Spiro, 2012, 2009). She has identified three scenarios in which collaboration takes place: "(1) communicating and exchanging knowledge through participatory online environments; (2) building digital collections of primary and/or secondary scholarly resources; and (3) developing computational methods for analyzing humanities data" (Spiro, 2012. p. 45). In her work, she has mainly focused on synchronous forms of collaboration in which research groups or participatory projects work together on a set of resources. Although these challenges are closely related to the concept of asynchronous collaboration, they only apply on a closed environment in which the project team, project data and collaborators all work together. We propose a form of asynchronous collaboration that is platform independent. Platform independency ensures the sustainability of the datasets and fosters an extensive applicability of the data.

The effective reuse of the data functions as the dividing line between asynchronous collaboration and traditional citation practices. Whereas traditional citation practices also reference to other scholarly resources and in doing so extend their lifespan and validity, the underlying data is never reused. Although we can cite *The Waning of the Middle Ages* of Dutch historian Johan Huizinga, we will never *reuse* his research notes or card catalogue. Since the emergence of digital research tools, historians and other scholars in the humanities have the ability to create digital card catalogue systems (databases). Asynchronous collaboration aims to open up these vast resources of rich data in order to establish an ecosystem of reuse and multiple forms of authorship.

In traditional forms of scholarship in the humanities, the claim on authorship is closely connected to the composition of a narrative in which the syntheses of the research project are brought together. We propose new forms of asynchronous authorship that are connected to the publication of datasets. These forms of authorship are in essence hybrid as the creation process of a dataset is often a collaborative process. Moreover, once reuse of these datasets takes place, new forms of authorship emerge that can span multiple layers of conceptualisation, creation, selection and publication processes. The process of

[2] http://www.biografischportaal.nl/, http://www.vondel.humanities.uva.nl/ecartico/, http://hadrianus.it/.

asynchronous collaboration is to be regarded as an additional collaborative methodology for the humanities and poses new opportunities for scholarly communication.

References

MOULINE, Claudine, 2013, Je t'aime, moi non plus. Career, Financing and Academic Recognition in the Digital Humanities (#dhiha5) [online]. 12 June 2013. [Accessed 6 November 2014]. Available from: http://annotatio.hypotheses.org/303

NOWVISKIE, Bethany, 2011, Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Profession* [online]. 2011. p 169–181. DOI 10.1632/prof.2011.2011.1.169. Available from: http://www.mlajournals.org/doi/pdf/10.1632/prof.2011.2011.1.152

NYHAN, Julianne and DUKE-WILLIAMS, Oliver, 2014, Is Digital Humanities a collaborative discipline? Joint-authorship publication patterns clash with defining narrative [online]. 10 September 2014. [Accessed 6 November 2014]. Available from: http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/10/joint-authorship-digital-humanities-collaboration/

SPIRO, Lisa, 2009, Collaborative Authorship in the Humanities [online]. 21 April 2009. [Accessed 6 November 2014]. Available from: http://digitalscholarship.wordpress.com/2009/04/21/collaborative-authorship-in-the-humanities/

SPIRO, Lisa, 2012, Computing and Communicating Knowledge: Collaborative Approaches to Digital Humanities Projects in : *Collaborative Approaches to the Digital in English Studies.* Old Main Hill : Computers and Composition Digital Press. p 44-82. ISBN: 9780874218879. Available from: http://ccdigitalpress.org/cad/CollaborativeApproaches.pdf

Prof. Dr. Josef Focht

Museum für Musikinstrumente der Universität Leipzig

Vorschlag eines Vortrags oder eines Statements in einem Panel

Die getriebenen Geisteswissenschaften – oder: Warum der Schritt von der Daten- zur Wissensproduktion nicht immer gelingt

In den DH-Forschungsprojekten geisteswissenschaftlicher Fächer agiert jüngst mit zunehmender Häufigkeit ein ungleiches Paar: ein schwacher Partner aus den Humanities und ein starker aus der Informatik. Dies hat verschiedenartige Gründe, von denen einer in der Wissenschaftsgeschichte des 20. Jahrhunderts liegt. In dieser Zeit erlebten die Geisteswissenschaften eine deutlich stärkere Differenzierung als etwa die Natur- oder Sozialwissenschaften. In diesem Differenzierungs- und Emanzipationsprozess galt die Abgrenzung von allen Nachbarn stets als existentiell und identitätsstiftend.

In der Folge zeigen interdisziplinäre Kooperation und partnerschaftliche Methoden in den Geisteswissenschaften häufig eine schwache Tradition; und so existieren vergleichsweise wenige Datenrepositorien, die fachübergreifend Verwendung finden. Dies gilt auch für die Musikwissenschaft. Ein positives Beispiel bieten etwa biographische Ressourcen, die mit bedeutendem Mehrwert fachübergreifend gepflegt und genutzt werden (vgl. etwa BMLO http://bmlo.de/Q/GND=116263431). Dagegen bieten instrumentenkundliche Datenpools ein negatives Exempel, obwohl sie neben der musikwissenschaftlichen Organologie etwa auch die Kunst- oder Buchwissenschaften, die Physik oder Werkstofftechnik beschäftigen könnten.

Diesem Mangel möchte man heute in der instrumentenkundlichen Forschung offensiv begegnen. Dazu müssen aber zunächst hinreichende Ressourcen für DH-Projekte angelegt werden, bevorzugt große. Die Informatik erweist sich hier als der stärkere Partner, der üppige Datenpools verspricht. In den vergangenen Jahren entstehen dabei zunehmend korpusbasierte Repositorien, die – auf einer einzigen oder mehreren gleichartigen Sammlungen analoger Quellen beruhend – den Charakter von Monokulturen aufweisen.

Zu den Stärken geisteswissenschaftlicher Methoden zählt aber nun gerade die Berücksichtigung verschiedener Quellengattungen, die abwägende Fallunterscheidung, die erkenntnisgeleitete Modellierung komplexer Fragestellungen, die bedarfsgerechte Berücksichtigung ergänzender Methoden in Sonderfällen. Weil aber nur wenige Datenpools digital verfügbar sind, und diese möglicherweise in ungleichem Gewicht oder mangelhafter Kompatibilität, entstehen oft facettenarme Forschungsdesigns oder simplifizierte Methoden. DH-Projekte verharren dann im Anfangsstadium einer Dokumentation oder Digitalisierung, und sie verflachen, ehe der eigentliche Prozess der Wissensgewinnung beginnt. Oder es erhält eine schwächere Untersuchungsmethode den Vorzug vor einer mutigeren, nur weil sie die

Erwartung eines geschmeidigen Berichts an den Projektförderer nährt, der zum nächsten Quartalsende wieder fällig ist. Speziell bei der Musik treten darüber hinaus branchenspezifische Einschränkungen auf den Plan, etwa forschungsungünstige Verwertungs- und Nutzungsrechte an Audio- und Video-Medien in der virtuellen Öffentlichenkeit.

Die spezifisch geisteswissenschaftliche Fachtraditionen (wie etwa in der Musikwissenschaft) erforderen die selbstbewusste Kombination innovativer digitaler und etablierter analoger Methoden. Sie fordern ggf. auch den mutigen Verzicht auf einen großen Datenpool, wenn er keinen Erkenntnisgewinn verspricht. Auch in DH-Projekten muss das Erkenntnisinteresse den Vorrang vor der Datenproduktion behalten. Oder wieder zurückgewinnen? Wenn dies gelingt, dann entstehen in den grundsätzlich begrüßenswerten und unbedingt notwendigen DH-Projekten der Musikwissenschaft voraussichtlich (oder hoffentlich) auch Ressourcen, die fachübergreifendes Interesse und Relevanz finden.

Panel-Diskussion

Digital Humanities aus der Sicht der Informatik

mit Günther Görz, Andreas Henrich, Gerhard Heyer und Martin Warnke

Zahlreiche Informatikerinnen und Informatiker arbeiten in DH-Projekten mit und kooperieren mit Geisteswissenschaftlern – und das zum Teil seit vielen Jahren und nicht erst seit dem jüngsten „DH-Hype".

Während sich aus den Geisteswissenschaften heraus die DH nun organisieren und ein klareres Bild des „Faches" zeichnen, bleibt die Perspektive der Informatik jedoch weiterhin unscharf. Um hier zu einer ersten Bestandsaufnahme zu kommen, wurden im Rahmen eines von der DHd veranstalteten Workshops am 3. November 2014 in Leipzig die folgenden Fragen bzw. Themenkomplexe in Form von ausgewählten Beiträgen und einem an dem Instrument des *Knowledge Café* orientierten offenen Gesprächs diskutiert (vgl. auch die Webseite des Workshops und den Call for Papers http://informatik-dh-workshop2014.topicmapslab.de/):

1) Zur institutionellen Verortung der Digital Humanites (Moderator: Günther Görz)

Etablierte institutionelle Strukturen können förderlich sein, aber mittelfristig sind für Digital Humanities–Projekte dauerhafte Infrastrukturen, z.B. in Zentren, notwendig. Nur so kann eine effektive Kommunikation von Projektpartnern aus verschiedenen Fakultäten, die u.U. im Rahmen desselben Projektes unterschiedliche Ziele verfolgen, erreicht werden. Essentiell ist die Entwicklung eines fächerübergreifenden gemeinsamen Methodenkatalogs; wichtige Themen sind Datendiversität, Annotation und Archivierung. Aus pragmatischen Gründen wäre die Einrichtung eigener Studiengänge sinnvoll; deren Absolventen könnten eine Vermittlungsfunktion zwischen verschiedenen Denkstilen und Forschungsparadigmen (informatikaffine vs. traditionell humanistische Forschung) wahrnehmen. Dem gegenüber steht die Schaffung einer „Transdisziplin" oder die Absorption der DH in neue Fachinformatiken. Im Hinblick auf das Publikationswesen sind traditionelle Anerkennungsmechanismen zu hinterfragen und geeignete Publiaktions-orte und -standards zu entwickeln.

2) Auf welche Weise muss sich die Informatik ändern oder öffnen, damit DH erleichtert, verbessert oder sogar erst möglich werden? (Moderator: Martin Warnke)

In der Wahrnehmung von Informatikern bleibt die Arbeits- und Funktionsweise eingesetzter informatischer Verfahren und entwickelter Softwareartefakte jenseits der Benutzeroberfläche für geisteswissenschaftliche Projektpartner oft opak. Umgekehrt fehlt Informatikern oft das Verständnis dafür, dass z.B. scheinbar triviale Veränderungen an Systemen einen bedeutenden Unterschied für die tägliche Arbeit geisteswissenschaftlicher Projektpartner machen können. Ungenügend ist unter Umständen auch die Bereitschaft sich mehr auf fachliche Inhalte der Geisteswissenschaften einzulassen. Inwiefern ist es nötig, bei Geisteswissenschaftlern Akzeptanz für gewisse Vorgehensweisen und Methoden der Informatik zu schaffen? Inwiefern gehen solche Vorgehensweisen vielleicht aber am Wesen und Selbstverständnis geisteswissenschaftlichen Arbeitens grundsätzlich vorbei und müssen angepasst und erweitert werden um eine erfolgreiche Zusammenarbeit zu gestalten?

3) Usability, Human Computer Interaction und Visualisierung im Kontext der DH (Moderator: Andreas Henrich)

Wie können in den Bereichen in denen Geisteswissenschaftler als Anwender von Software oder informationstechnischen Lösungen auftreten, Erkenntnisse zur Benutzerfreundlichkeit, Human Computer Interaction und dem differenzierten Feld von Visualisierung und Visualisierungsforschung auf die Bedürfnisse von Geisteswissenschaftlern oder bestimmten Geisteswissenschaften zugeschnitten werden? Welche neuen Erkenntnisse, welche neuen Forschungsfragen ergeben sich daraus für die Informatik? Können die Traditionen statischer und manuell erzeugter Visualisierung und hermeneutischer Interpretation für die Informatik fruchtbar gemacht werden und wenn ja, welche Formen könnte das annehmen?

4) Was kann die Informatik im Sinne methodologischer Reflexion von den Humanities lernen? (Moderation: Manfred Thaller)

Jenseits der Frage in welchem institutionellen Rahmen DH betrieben werden soll und welche methodologischen Konsequenzen gezogen werden müssen, um inter- bzw. transdisziplinäre Zusammenarbeit erfolgreicher zu gestalten, stellt sich bei der Betrachtung des Verhältnisses von Informatik und Geisteswissenschaften auch die Frage nach einer methodologischen Befruchtung der unter Umständen weniger selbstreflektierten Informatik durch die geschichtsbewussten und um eine Abgrenzung und Beschreibung der eigenen Fähigkeiten, Zuständigkeiten und (Neben-)wirkungen bemühten Geisteswissenschaften. Kann die Informatik die ihren Methodenkanon mithin formaler und statischer begreift als die Geisteswissenschaften und ihre Entwicklungslinien vielleicht mehr im Sinne monoton steigender, oftmals quantifizierbarer Verbessungen zeichnet hier von den Geisteswissenschaften lernen und wenn ja auf welche Weise?

Im Panel sollen die Ergebnisse dieser Diskussion von den beteiligten Moderatoren vorgestellt werden (allerdings mit Gerhard Heyer anstelle von Manfred Thaller) und aus eigener Sicht kritisch zusammengefasst und bewertet werden.

Bildungspotentiale digitaler Musik-Editionen zwischen Demokratisierung und Ungewissheit. Ein Theoretischer Verortungsversuch

Medien und Bildung stehen jeher in einer komplexen, symbiotischen Verbindung zueinander. So sind Medien bereits durch orale Tradierung von Wissen, die Entwicklung der Schrift oder des Buchdrucks tiefgehend mit Inhalten, Konzepten und Prozessen von Bildung verbunden und prägen diese maßgeblich. Mit der fortschreitenden Digitalisierung sind im öffentlichen Diskurs vor allem entweder euphorische Erwartungen (Prensky 2001) oder aber apokalyptische Szenarien (Spitzer 2012) populär. Diese Positionen stellen jedoch nur zwei Pole einer simplifizierenden Betrachtungsweise dar. Eine reflektierte Analyse digitaler Medien sollte jedoch spezifischer auf die jeweiligen Phänomene bezogen sein, um Chancen und Risiken adäquat beschreiben zu können. Dementsprechend werden in diesem Beitrag eingehend die Implikationen digitaler Musikeditionen analysiert, um erste theoretische Anknüpfungspunkte zwischen Bildung, Demokratisierung und Ungewissheit zur Diskussion zu stellen.

Medien und Bildung sind einerseits eng miteinander verknüpft, andererseits sind damit jedoch verschiedene Prozesse, Aneignungsperspektiven und Ziele verbunden. So plädiert bereits Humboldt für eine Auffassung von Bildung, die den Menschen als Ganzes umfasst und Bildung als eine Form der Subjektwerdung betrachtet (Humboldt 1966). Eine Auseinandersetzung mit Bildung und deren Aneignung leistet Klafki mit seiner Forderung nach Allgemeinbildung. Der Mensch soll in die Lage versetzt werden selbstbestimmt, partizipativ und solidarisch sein eigenes Leben zu vervollkommnen, gesellschaftlich und kulturell mitzugestalten sowie sich solidarisch für die Belange anderer einsetzen zu können. Diese Ziele können nur erreicht werden, in dem Bildung für alle Menschen zugänglich ist und somit grundsätzlich demokratisiert ist. Sie muss die vorherrschenden gesellschaftlichen Fragestellungen historisieren, aber auch in die Gegenwart transferieren und darüber hinaus eine Zukunftsperspektive beinhalten. In Bezug auf Medien fordert er ein kritisches Reflexionswissen, welches sowohl technologisches Wissen, eine Einführung in die Nutzung als auch das Bedenken möglicher Auswirkungen umfasst (vgl.: Klafki 1990). Darüber hinaus gilt es zu bedenken, dass Bildung nicht allen Menschen gleich zur Verfügung steht, und es immer noch ausgeprägte soziale Ungleichheiten gibt, die auf Basis von Bourdieus Kapitalsorten tief im gesellschaftlichen Leben verankert sind und sich fortschreiben (vgl. Bourdieu 1983). Die von Klafki identifizierten Schlüsselfragen der Gegenwart sind angesichts zunehmender Mediatisierung immer auch medial vermittelt, weshalb den Medien eine enorme Relevanz zugesprochen wird (vgl. Krotz 2007). Medien und Medieninhalte sind jedoch nicht singulär zu betrachten, sondern sind immer schon in individuellen und sozialen Kontexten eingebettet, die es gilt integrativ zu bearbeiten.

Mit diesen Überlegungen wird eine erste Einordnung der digitalen Musik-Editionen möglich, die den Zugang zu diesem Wissen thematisiert. Durch die digitale (Ab-)bildung der Musik-Editionen wird, wenn auch nicht vollständig, aber zumindest teilweise ein Demokratisierungsprozess eingeleitet. Zunächst sind digitale Editionen einfacher und kostengünstiger zugänglich und bieten somit die Möglichkeit, neue und andere Rezipientenkreise zu erreichen. Der Zugang jedoch verspricht noch keine versierte Nutzung (vgl. Bourdieu 1983), dafür sind erklärende Anleitungen in Editionen notwendig, vor allem aber musikdidaktische Aufbereitung, Vermittlung und Schulung in universitären oder schulischen Kontexten. Mit der sukzessiven Verbreitung der grundlegenden Kenntnisse über Editionsarbeit, vor allem im schulischen Kontext, ließen sich jedoch Nutzerkreise erschließen, die sonst kaum Berührungspunkte zu diesen Erkenntnisinhalten hätten. Diese einfachen Konsequenzen können dazu beitragen Themengebiete der sogenannten klassischen Kulturelite gesellschaftlich relevanter und zugänglicher zu gestalten.

Um die Bildungsanknüpfungspunkte von Medien grundsätzlicher zu betrachten sind Überlegungen von Reinhard Keil (2010) dienlich. Er analysiert das Verhältnis von Didaktik und Medien und kritisiert, dass Medien in didaktischer Perspektive weitreichend lediglich als Mittler von Informationen interpretiert werden. Jedoch spricht er sich für eine spezifischere Sichtweise aus, indem er die Implikationen von Medien als Bildungsprozesse auffasst. So argumentiert er für die eingehende Betrachtung der technischen Möglichkeiten der Medien und ihrer Bildungsanknüpfungspunkte. Er stellt Persistenz als ein entscheidendes Merkmal von Medien heraus, das Differenzerfahrungen ermöglicht, da Gedankengänge, Rechenschritte und Operationen für den Einzelnen und Dritte nachvollziehbar, wiederholbar und somit überprüfbar werden (vgl. ebd., S. 127 ff.). Somit stellt sich hinsichtlich der digitalen Editionen die Frage, was abgebildet und wie dies dargestellt wird. Noch eindringlicher stellt sich diese Frage, wenn grundsätzlich über die Rolle der Medien reflektiert wird: "Alles was wir über die Welt sagen, erkennen und wissen können, das wird mit Hilfe von Medien gesagt, erkannt und gewusst" (Assmann/Assmann 1990, S. 2).

Digitale Musikeditionen sind zunächst ein Konglomerat aus digitalen und analogen Praktiken. Die Quellen werden ebenso gesichtet wie bei der konventionellen Editionsarbeit. Die Möglichkeiten der digitalen (Re-) präsentation sind jedoch vielfältiger als beim gedruckten Werk. Digitale Musik-Editionen bilden nicht DAS Werk ab, sondern mit der Prozessabbildung der Werksgenese Variationen von möglichen Lesarten. Durch fast archäologischen Auseinandersetzungen mit den Originalschriften (vgl. Beethoven Werkstatt, Bonn) wird die Perfektion von Musikwerken brüchiger, sobald alle Informationen der Werksgenese abgebildet, dokumentiert

und von jedem Rezipienten und Nutzer nachvollzogen werden können. Es wird also nicht das Werk in seiner vermuteten Einheit repräsentiert, sondern vielmehr die Rekonstruktionen der Werksgenese (vgl. Bohl/Kepper/Röwenstrunk 2011), die unter Umständen nicht einmal einem linearen Fortschrittsgedanken folgen muss, sondern auch aus Zufällen, Übertragungsfehlern oder Missverständnissen resultieren kann. Dies wirkt nicht nur ermächtigend auf der Rezipientenebene, welche nun nach eingehender Recherche der aufbereiteten Informationen eine eigene Lesart und Variation des Werkes markieren können. Andererseits sind angesichts der angebotenen Informationsfülle auch Überforderungen seitens der Rezipienten möglich. Darüber hinaus hat diese Form der Darstellung auch Rückwirkungen auf die Betrachtung des musikalischen Genies: Auch dieser Typus ist dem Entstehungsprozess seines Werkes unterworfen. Die Komposition ist nicht einfach gegeben, sondern entsteht sukzessiv, wird überarbeitet, korrigiert, optimiert. In dieser Perspektive wird das Ideal des musikalischen Genies in Kontexte lebensweltlicher Erfahrungen zurückgeholt und bieten "normalen" Musikern Potenziale sich nicht an Idealvorstellungen abarbeiten zu müssen, sondern vielmehr einen Prozess durchlaufen zu können. Dadurch dass Editionsprojekte kostengünstiger realisiert werden können, ist es möglich auch unbekanntere Komponisten mit einer Edition zu bedenken. Dies bewirkt eine weitere Demokratisierung, da der traditionelle Kanon durchbrochen und somit viele verschiedene Komponisten in den Kreis der Editionen aufgenommen werden können. Mit Rekurs auf Walter Benjamin (1974) sind sowohl auf der Ebene der (Re-)Präsentation des Materials als auch in der Wirkung des Kunstwerks andere Blicke auf Werk, Genese und Komponist möglich, die in der technischen Reproduzierbarkeit einen demokratischen Charakter aufweisen.

Letztlich sollen die vorangegangenen Überlegungen noch einmal verdichtet werden, um eine übergreifende Lesart der digitalen Musik-Editionen zu ermöglichen. Wie skizziert wird also nicht mehr die Einheitlichkeit und Sicherheit des musikalischen Werkes hervorgehoben, sondern dessen Entstehung, Überarbeitung, Optimierung. Damit einhergehend werden Ungewissheiten abgebildet, wie etwa das Fehlen von Noten oder unklare Annotationen. Ungewissheit in diesem Sinne bedeutet also dass "vermeintliche Selbstverständlichkeiten, vormalige Evidenzen, scheinbare Sicherheiten, begründete Erwartungen und gewisse Grenzen verflüssigt und zersetzt" (Liesner/Wimmer 2005, S. 23) werden. Das Werk wird nicht mehr als totale Einheit dargestellt, sondern mit unbeantworteten Fragen und Unsicherheiten der Auslegung (ab-)gebildet. Die vermeintliche Gewissheit des Werkes steht mit den digitalen Editionen zur Disposition. Jedoch gilt es zu bedenken, dass diese Ungewissheiten immer schon in den Originaldokumenten eingeschrieben waren, aufgrund der begrenzten Abbildungsmöglichkeiten jedoch nicht gezeigt wurden/werden sollten. Auf einer übergreifenden Ebene sind mit den digitalen Editionen somit Ungewissheiten verbunden, die zu neuem Wissen führen. Einerseits die Sicht auf das Werk als entstehender Prozess und andererseits die reflektierte Sicht auf das musikalische Genie und damit einhergehend die Erkenntnis, dass individuelle Interpretationen ebenso legitim sind, wie diejenigen der Experten. Diese individuelle Freiheit wird in Bildungsdiskursen auch als Last des Individuums zur eigenverantwortlichen Selektion und Aneignung des richtigen, nutzbringenden Wissens gelesen (vgl. Höhne). Im Sinne der Möglichkeiten der digitalen Musik-Editionen überwiegt jedoch der Demokratisierungsprozess. Das einmalige Kunstwerk der einheitlichen und durch Experten legitimierten sicheren Lesart ist brüchiger geworden und gibt Raum für individuelle Interpretationen und Projektionen.

Die theoretischen Anknüpfungspunkte sind vorläufige Szenarien der Wirkung, Nutzung und Interpretation digitaler Musik-Editionen. Um diese Überlegungen zu konkretisieren bedarf es eingehender empirischer Forschung, die wichtige Einblicke über die Aneignung, Nutzung und Interpretation sowie damit verbundene Erkenntnismotivationen und –ziele ermöglicht.

Literatur

Assmann, Aleida; Assmann, Jan (1990): Schrift – Kognition – Evolution. In Havelock, Eric A. (Hrsg.) Schriftlichkeit: Das griechische Alphabet als kulturelle Revolution. Weinheim: Wiley-VCH, S. 1-35.

Benjamin Bohl, Johannes Kepper, Daniel Röwenstrunk. (2011). Perspektiven digitaler Musikeditionen aus der Sicht des Edirom-Projekts. In: DIE TONKUNST, Juli 2011, Nr. 3, Jg. 5 (2011), S. 270–276
Bourdieu, Pierre (1983): Ökonomisches Kapital, kulturelles Kapital, soziales Kapital In: Kreckel, Reinhard (Hg.): Soziale Ungleichheiten. Göttingen: Otto Schwartz & Co., S. 183-198.

Benjamin, Walter. (1974). Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit. In: Gesammelte Schriften. Frankfurt a. M.: Suhrkamp.

Humboldt, Wilhelm von. (2007). Vorschläge zur Organisation des preußischen Bildungssystems. In: Baumgart, Franzjörg (Hrsg.). Erziehungs- und Bildungstheorien. Bad Heilbrunn: Julius Klinkhardt, S. 111-116.

Keil, Reinhard. (2010). E-Learning vom Kopf auf die Füße gestellt. In: Herzig, B.; Meister, D. M.; Moser, H.; Niesyto, H. (Hrsg.): Jahrbuch Medienpädagogik 8. Medienkompetenz im Zeitalter des Web 2.0. Wiesbaden: VS.

Klafki, Wolfgang. (2007). Abschied von der Aufklärung? In: In: Baumgart, Franzjörg (Hrsg.). Erziehungs- und Bildungstheorien. Bad Heilbrunn: Julius Klinkhardt, S. 267-279.

Krotz Friedrich (2007) Mediatisierung: Fallstudien zum Wandel von Kommunikation. Wiesbaden, VS.

Liesner, Andrea; Wimmer, Michael. (2005). Der Umgang mit Ungewissheit. Denken und Handeln unter Kontingenzbedingungen. In: Helsper, Werner; Hörster, Reinhard; Kade, Jochen. (Hrsg.) Ungewissheit. Pädagogische Felder im Modernisierungsprozess. Weilerswist: Velbrück Wissenschaft, S. 23-49.

Meinungen in Twitterdiskursen

Potenziale der automatisierten Inhaltsanalyse aus der Computerlinguistik für Fragestellungen der Kommunikationswissenschaft

Problemstellung

Die manuelle Inhaltsanalyse, die in der Kommunikationswissenschaft umfassend eingesetzt wird, ist grundsätzlich geeignet, tiefer gehende Kenntnisse über Inhalte und Beziehungen von Textfragmenten in Social Media zu liefern. Sie ist jedoch zeit- und kostenaufwändig und kann deshalb nur auf kleine Stichproben angewandt werden. Um dieser Einschränkung zu begegnen, wird in der vorliegenden Studie in Zusammenarbeit mit der Computerlinguistik untersucht, inwiefern Meinungsäußerungen auf Twitter durch automatisierte Inhaltsanalysen erhoben werden können. Erweist sich die automatisierte Inhaltsanalyse als valide Methode zur Erfassung von Meinungsäußerungen, wären damit forschungspragmatische Vorteile wie Zeit- und Kostenersparnis verbunden. In der Kommunikationswissenschaft könnten öffentliche Diskurse bzw. Teildiskurse, d.h. Meinungsäußerungen zu politischen Streitfragen auf Twitter und anderen Internetplattformen breiter und kontinuierlich erhoben werden. Um dies zu erreichen, muss zunächst die automatisierte Inhaltsanalyse durch Daten der manuellen Inhaltsanalyse von Meinungsäußerungen auf Twitter validiert werden.

Theoretische Relevanz der Problemstellung

Die Analyse von Meinungsäußerungen ist u.a. aus der Perspektive einer integrierten Netzwerköffentlichkeit von Interesse (Benkler 2006). Dabei wird angenommen, dass im Internet verschiedene Öffentlichkeitsebenen (Gerhards/Neidhardt 1990: 19-26; Habermas 1992: 452) in der *vertikalen Dimension* stärker miteinander vernetzt und durchlässiger sind, als dies in den traditionellen Massenmedien der Fall ist. Dies soll auch nicht-organisierten Bürgern und zivilgesellschaftlichen Akteuren ermöglichen, sich folgenreich über Social Media wie z.B. Twitter an öffentlichen Diskursen zu beteiligen. Meinungsbildungsprozesse sollen eher „von unten nach oben“ verlaufen. Bürger sollen eher die Chance haben, den Diskursverlauf (durch größere Resonanz und weiter reichende Diffusion ihrer Beiträge) und letztlich auch politische Entscheidungen zu beeinflussen. In der *horizontalen Dimension* des öffentlichen Raums wird das Spektrum der artikulierten Meinungen betrachtet. Durch die vereinfachte Partizipation im Internet soll sich, so die gängige Annahme, die Meinungsvielfalt im Vergleich zu den traditionellen Massenmedien erweitern. Es wird aber auch befürchtet, dass es zu einer Fragmentierung der Öffentlichkeit im Internet kommt (Marr 2002; Sunstein 2007; Habermas 2008: 162). Nach dieser These wird der Meinungsstreit nicht mehr ausgetragen, weil sich die Internetnutzer in homogene Interessen- und Meinungsgruppen aufspalten. Solche Annahmen können leichter überprüft werden, wenn sich die automatisierte Inhaltsanalyse als valides Instrument für die Erfassung von Meinungen und Akteurstypen erweist.

Forschungsfragen

Wir legen unserer Untersuchung zwei Forschungsfragen zugrunde:

1. Inwieweit ist die manuelle Inhaltsanalyse von Twitterdiskursen aus der Kommunikationswissenschaft durch geeignete Verfahren der Computerlinguistik automatisierbar? (FF1)
2. Wie lässt sich die automatisierte Inhaltsanalyse der Computerlinguistik in der Kommunikationswissenschaft anwenden? (FF2)

Untersuchungsanlage

Auswahl des Untersuchungsmaterials

Es wird eine manuelle Inhaltsanalyse von Meinungen mit einer automatisierten Inhaltsanalyse kombiniert, wobei die Kombination auf die einseitige Validierung der automatisierten Inhaltsanalyse abzielt (Loosen/Scholl 2012). Dies wird am Beispiel von Tweets untersucht, die Teil von Diskussionen auf Twitter sind. Diese Diskussionen wurden beispielhaft aus einem Twitterdiskurs zur Energiewende extrahiert. Dieser Diskurs wurde vorab mit mehr als 180 energiewende-relevanten Keywords definiert und über die Twitter-API getrackt. Ein Tweet ist dann Bestandteil einer Diskussion, wenn er an mindestens einen anderen Akteur adressiert ist und anschließend mindestens eine Antwort auf diesen adressierten Tweet folgt. Solche Diskussionen wurden mit der in-reply-to-Funktion aus allen getrackten Tweets zwischen 20. November und 01. Dezember 2013 extrahiert. Insgesamt wurden 3101 Diskussionen (bestehend aus 11.587 Tweets) in diesem Zeitraum für den Energiewende-Diskurs extrahiert und rückwärts vervollständigt.

Erhebungskategorien für manuelle und automatisierte Inhaltsanalyse

Das Untersuchungsmaterial wurde zunächst manuell auf zwei Ebenen annotiert: (1) auf Ebene kompletter Diskussionen (z.B. Relevanz einer Diskussion für das Thema „Energiewende“) und (2) auf Ebene einzelner Tweets. Auf der zweiten Ebene wurden folgende Aspekte zunächst manuell erhoben:

1. formale Kategorien: Stellung eines Tweets in der Diskussion, Relevanz des Tweets für die Energiewende
2. geäußerte Meinungen: Vorhandensein einer/mehrerer Meinung pro Tweet, Objekt bzw. Gegenstand der Meinung, positive oder negative Polarität der Meinung, Intensität der Meinung
3. Kontext geäußerter Meinungen: Autorentyp für Tweeturheber und für adressierten bzw. erwähnten Autor im Tweet (z.B. private Einzelakteure, nicht-profitorientierte Interessengruppen, profitorientierte Interessengruppen, politische Akteure und Journalisten)

Insgesamt wurden 2.655 Tweets in 729 Diskussionen manuell annotiert. Darin wurden 1.243 polare Meinungen identifiziert (positiv: $n = 330$; negativ: $n = 896$).

Validierungsmaße für den Vergleich der manuellen und der automatisierten Inhaltsanalyse

Für die Validierung jeder erhobenen Kategorie wurde entweder das traditionelle F1-Maß oder zwei weniger restriktive Varianten dieser Metrik verwendet – nämlich das binäre und das proportionale F1-Maß (vgl. Johansson/Moschitti 2010). Der Unterschied zwischen diesen drei Varianten besteht haupt-

sächlich darin, wie korrekt Übereinstimmung zwischen manueller Annotation und automatisierter Klassifizierung berechnet werden soll. Zudem wurden mehrere automatisierte Klassifikationsalgorithmen (Naive Bayes, SVM, LibLinear, AdaBoost und Logistic Regression) auf einem Set Trainingsdaten trainiert und deren Leistungsfähigkeit getestet.

Befunde und Diskussion

Der Vergleich zwischen manueller und automatisierter Inhaltsanalyse hat gezeigt, dass sowohl für die automatisierte Klassifikation von Meinungen als auch für die von Akteurstypen die besten Ergebnisse mit Hilfe des maschinellen Klassifikationsverfahrens LibLinear (Fan et al., 2008) erzielt wurden. Das F1-Maß für die Erkennung von Meinungen betrug 66 Prozent, das für die Erkennung der Akteurstypen des Autors und ggf. des adressierten Benutzers lag bei 41 bzw. 59 Prozent. Wurden neben rein linguistischen auch kommunikationswissenschaftliche Aspekte in den Vergleich einbezogen (z.B. Anzahl der Sprecher in der Diskurssion, Akteurstyp des Tweeturhebers oder des Adressaten, wurden sogar 69 Prozent Übereinstimmung zwischen manueller und automatisierter Inhaltsanalyse erreicht (FF1).

Die genannten Befunde der automatisierten Inhaltsanalyse in der Computerlinguistik bergen enormes Potential für die Erforschung von öffentlichen Meinungen in Social Media wie z.B. auf Twitter im Rahmen der Kommunikationswissenschaft. So lassen sich dynamische Meinungsbildungsprozesse großer Textmengen zeit- und kostengünstig in Social Media untersuchen. Dadurch erfahren wir zum Beispiel, inwiefern sich nicht-organisierte Bürger und nicht-profitorientierte Interessengruppen an öffentlichen Diskursen beteiligen, ob Meinungsbildungsprozesse „von unten nach oben“ verlaufen und inwieweit diese Akteurstypen die Chance haben, den Diskursverlauf und letztlich auch politische Entscheidungen zu beeinflussen. Momentan können wir dies nur für kleine Stichproben zeigen (FF2). So illustriert zum Beispiel Tabelle 1 im Anhang, dass 19% aller Akteure als einfache Bürger einen Tweet absetzen, während 25% als einfache Bürger erwähnt oder adressiert werden. Darüber hinaus sieht man zum Beispiel, dass einfache Bürger untereinander weniger negativ bzw. kritisch miteinander diskutieren als Bürger mit Politikern oder Journalisten (Tabelle 2, Anhang). Dies spricht für eine kritische Rolle der Bürger im öffentlichen Diskurs. Um die prozentuale Übereinstimmung zwischen manueller und automatisierter Inhaltsanalyse weiter zu erhöhen, hat die Zusammenarbeit zwischen Kommunikationswissenschaft und Computerlinguistik auch zukünftig großes Potential.

Literatur

Fan, R.. E., Chang K. W., Hsieh Ch. J., Wang X. R., Lin Ch. J. (2008): LIBLINEAR: A Library for Large Linear Classification. In: Journal of Machine Learning Research 9, S. 1871-1874.

Gerhards, J., Neidhardt, F. (1990): Strukturen und Funktionen moderner Öffentlichkeit. Fragestellungen und Ansätze. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (= FS III 90-101).

Habermas, J. (1992): Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats. Frankfurt a. M.: Suhrkamp.

Habermas, J. (2008): Hat die Demokratie noch eine epistemische Dimension? Empirische Forschung und normative Theorie. In: Habermas, Jürgen: Ach, Europa. Frankfurt a. M.: Suhrkamp, S. 138-191.

Johansson, R., Moschitti, A. (2010): Reranking Models in Fine-grained Opinion Analysis. In: 23rd International Conference on Computational Linguistics, Proceedings of the Conference, COLING 2010, S. 519-527.

Landis, J. R., Koch, G. G. (1977): The measurement of observer agreement for categorical data. In: Biometrics. 33, S. 159-174.

Loosen, W., Scholl, A. (2012): Theorie und Praxis von Mehrmethodendesigns in der Kommunikationswissenschaft. In: Loosen, W., Scholl, A. (Hrsg.): Methodenkombinationen in der Kommunikationswissenschaft. Methodologische Herausforderungen und empirische Praxis. Köln: Herbert von Halem, S. 9-25.

Marr, M. (2002): Das Ende der Gemeinsamkeiten? Folgen der Internetnutzung für den medialen Thematisierungsprozess. In: Medien und Kommunikationswissenschaft 50(4), S. 510-532.

Sunstein, C. R. (2007): Republic.com 2.0. Princeton, NJ: Princeton University Press.

Anhang

Tabelle 1: Verteilung der Akteurstypen zwischen Tweet-Autor und Tweet-Adressat/-Erwähnung

	Tweet-Adressat oder Tweet-Erwähnung						
Tweet-Autor	Private Personen	Non-Profit Organisationen	Profitorient. Organisationen	Politische Akteure	Journalisten	Sonstige Akteure	**Gesamt**
Private Personen	270 **10,8%**	50 **2,0%**	27 **1,1%**	140 **5,6%**	81 **3,2%**	53 **2,1%**	**621** **24,7%**
Non-Profit Organisationen	31 **1,2%**	97 **3,9%**	41 **1,6%**	101 **4,0%**	52 **2,1%**	162 **6,5%**	**484** **19,3%**
Profitorient. Organisationen	18 **0,7%**	45 **1,8%**	25 **1,0%**	34 **1,4%**	23 **0,9%**	38 **1,5%**	**183** **7,3%**
Politische Akteure	74 **2,9%**	64 **2,5%**	20 **0,8%**	208 **8,3%**	45 **1,8%**	248 **9,9%**	**659** **26,3%**
Journalisten	22 **0,9%**	23 **0,9%**	16 **0,6%**	38 **1,5%**	19 **0,8%**	167 **6,7%**	**285** **11,4%**
Sonstige Akteure	61 **2,4%**	26 **1,0%**	5 **0,2%**	89 **3,5%**	20 **0,8%**	77 **3,1%**	**278** **11,1%**
Gesamt	**476** **19%**	**305** **12,2%**	**134** **5,3%**	**610** **24,3%**	**240** **9,6%**	**745** **29,7%**	

Basis: n = 2510 Meinungen (in 2.655 Tweets)

Tabelle 2: Meinungspolarität zwischen Tweet-Autor und Tweet-Adressat/-Erwähnung (Mittelwerte)

	Tweet-Adressat oder Tweet-Erwähnung						
Tweet-Autor	Private Personen	Non-Profit Organisationen	Profitorient. Organisationen	Politische Akteure	Journalisten	Sonstige Akteure	**Gesamt**
Private Personen	**–0,1**	**–0,1**	**0,0**	**–0,3**	**–0,4**	**–0,2**	**–0,2**
Non-Profit Organisationen	**–0,3**	**–0,2**	**–0,1**	**–0,3**	**–0,4**	**–0,3**	**–0,3**
Profitorient. Organisationen	**–0,1**	**0,0**	**–0,2**	**–0,1**	**+0,2**	**–0,1**	**0,0**
Politische Akteure	**–0,1**	**–0,2**	**–0,2**	**–0,2**	**–0,1**	**–0,2**	**–0,2**
Journalisten	**–0,1**	**–0,3**	**–0,3**	**–0,1**	**+0,1**	**–0,2**	**–0,2**
Sonstige Akteure	**–0,1**	**–0,3**	**–0,4**	**–0,3**	**–0,2**	**–0,3**	**–0,3**
Gesamt	**–0,1**	**–0,2**	**–0,2**	**–0,2**	**–0,2**	**–0,2**	

Basis: n = 2510 Meinungen (in 2.655 Tweets)

Identifikation kognitiver Effekte in Online-Bewertungen

Valentina Stuß & Michaela Geierhos, Universität Paderborn

Wie und ob die Meinungsbildung das Verhalten der Verbraucher[1] und Dienstleistungsnehmer beeinflusst und wie Wissensverarbeitung bei Menschen abläuft, sind die Fragen, mit denen sich Sozial- und Kognitionspsychologen beschäftigen (Stürmer, 2009; Wolff, 1993). Fehlerhafte Urteile bei der Meinungsbildung, auch kognitive Effekte genannt, werden in bestimmten Situationen wie der Bewertung von Produkten oder in Anspruch genommenen Leistungen besonders deutlich. Sie lassen sich in der Wissensorganisation und Gedächtnisstruktur des menschlichen Gehirns begründen. Mögliche Gründe für verzerrte Meinungen sind falsche Wahrnehmungen, die auf Basis kognitiver Informationsorganisation und -verarbeitung, aber auch aufgrund emotionaler Betroffenheit und/oder persönlicher Voreingenommenheit entstehen können.

Im Web 2.0 bieten Bewertungsportale Internetnutzern eine Plattform für den multi-direktionalen Erfahrungsaustausch an. Zahlreiche Rezensionen der Konsumenten, Kunden, Patienten oder anderer Zielgruppen bilden einen riesigen Datenbestand und stellen eine wertvolle Informationsquelle dar, die eine wissenschaftliche Auseinandersetzung von mehreren Seiten erfordert. Nicht nur Sozial- und Kognitionspsychologen, sondern auch Computerlinguisten, haben ein wissenschaftliches Interesse daran, Erkenntnisse mittels Informationsextraktion und Stimmungserkennung (Kim & Hovy, 2004; Hatzivassiloglou & McKeown, 1997) aus diesen digitalen Textsammlungen zu gewinnen. Eine interdisziplinäre Analyse psychologischer Phänomene, auch kognitive Effekte genannt, im Kontext schriftlich geäußerter Online-Meinungen wurde jedoch bisher nicht durchgeführt. Darum ist es das Ziel, mögliche kognitive Effekt zu definieren und computergestützt zu identifizieren.

Die Abweichung von der Norm

Je nach Art und Weise der dargebotenen Informationen, können diese unterschiedlich wahrgenommen werden, was zu den sogenannten Wahrnehmungsdefekten oder Informationspathologien (kognitiven Effekten) und somit zu fehlerhaften Meinungsbildern führen kann (Schneider, 2013). Ein kognitiver Effekt liegt immer dann vor, „wenn relevante Informationen nicht beschafft, nicht (korrekt) übermittelt, nicht produziert oder nicht (korrekt) verarbeitet werden, obwohl dies eigentlich möglich wäre." (Schneider, 2013:13f.). Es gibt eine Reihe von kognitiven Effekten, wie z. B. den Framing-Effekt, der dann auftritt, wenn inhaltsgleiche, aber auf verschiedene Weise dargestellte Informationen unterschiedlich bewertet werden. Insbesondere in Online-Bewertungen werden allgemeine Verfälschungen wie der Framing-Effekt oder auch Negativitätsverzerrungen auftreten.

„It does not appear possible today to group all of the phenomena that have been qualified as cognitive biases under one and the same definition. [...] As such, a bias is detected when derivation from norm is observed. [...] Occasional and accidental errors are obviously not part of the issue of cognitive biases." (Caverni et.al., 1990:7f.)

Um kognitive Effekte letztendlich automatisiert in Online-Bewertungen identifizieren zu können, stellt sich die Frage, was als Norm in einer Bewertung gilt. Da sich diese meist aus einem Kommentar und einem Notenwert zusammensetzt, ist zu vermuten, dass Textinhalt und Note kongruent sind und damit in ihrer Polarität (positiv, neutral, negativ) übereinstimmen. Jedoch können aufgrund von zufälligen Individualfehlern (wie z. B. Verständnisprobleme beim Benotungssystem eines Bewertungsportals) nicht alle Nichtübereinstimmungen grundsätzlich als kognitive Effekte interpretiert werden. Logisch für

[1] Aus Gründen der leichteren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Entsprechende Begriffe gelten im Sinne der Gleichbehandlung für beide Geschlechter.

eine Norm wäre ebenfalls, dass die Mehrheit an Bewertungen diese Norm erfüllen sollte. Das bedeutet, dass nur eine kleine Anzahl Bewertungen von kognitiven Effekten überhaupt betroffen sein können.

Als kognitive Effekte bei Online-Bewertungen werden somit diejenigen Bewertungsfehler bezeichnet, die sowohl von der Erwartungsnorm abweichen als auch bewusst von den Rezensenten gemacht wurden, um ein bestimmtes Ziel zu erreichen. Diese führen zu einer verzerrten Meinung, die sich durch einen Widerspruch im Freitext zur entsprechend vergebenen Note aufspüren lässt. Insbesondere um Inkonsistenzen im Bewertungsverhalten aufzudecken, ist es erforderlich ausreichend repräsentative Meinungen als Referenzkorpus zu haben. Nur so kann sichergestellt werden, dass ein Vergleich der zu untersuchenden Meinungen mit der zu erwartenden Norm für jede Ausprägung eines kognitiven Effekts durchgeführt werden kann.

Die Identifikation kognitiver Effekte

Datenbasis

Dieser Studie liegen zufällig ausgewählte Datensätze der Portale jameda und DocInsider aus den Jahren 2009 bis 2013 zugrunde. Das Korpus umfasst 217.841 individuelle Erfahrungsberichte, die allesamt nach einem Arztbesuch in Deutschland verfasst und online gestellt wurden. Jede Arztbewertung besteht dabei aus einer Überschrift, dem eigentlichen Text und bis zu 17 verschiedenen numerischen Bewertungskategorien (u.a. Behandlung, Vertrauensverhältnis, Wartezeit, Barrierefreiheit).

Methodik

Bei der Identifikation kognitiver Effekte geht es zunächst darum, bestimmte Muster, die Meinungen zu den vordefinierten Kategorien ausdrücken, und deren Polarität (Turney, 2002) zu erkennen. Diese Muster sind aus den frei formulierten Texten der Nutzer zu extrahieren. Im Rahmen einer Sentiment Analyse werden alle in der jeweiligen Meinung angesprochenen Themen identifiziert und entsprechend ihrer Polarität klassifiziert (Kim & Hovy, 2006). Diese Muster werden mittels regelbasierter, morpho-syntaktischer Verfahren, sogenannten lokalen Grammatiken (Gross, 1997) extrahiert und mithilfe von Transduktoren annotiert. Lokale Grammatiken sind stets modular aufgebaut, was in Abbildung 1 durch eine Subgrammatik zur Erkennung von Zeitangaben mit positiver Polarität der Kategorie „Wartezeit (Praxis)“ illustriert wird.

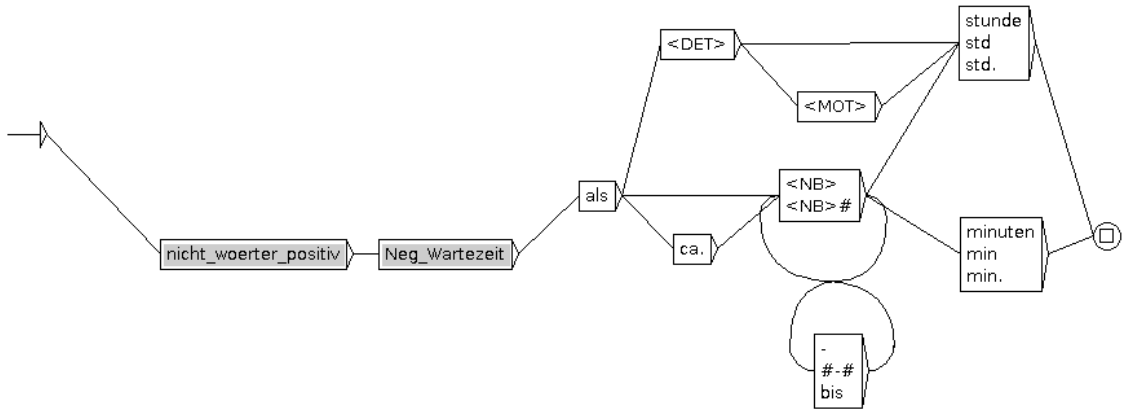


Abbildung 1: *Lokale Grammatiken zur Erkennung der Wartdauer in der Arztpraxis*

Durch Suche der in dieser Grammatik beschriebenen Muster entsteht auf der hier verfügbaren Datenbasis folgende Konkordanz (vgl. Abbildung 2). Hierfür werden unter anderem Lexika miteinbezogen, die beispielsweise in der in Abbildung 1 dargestellten lokalen Grammatiken über <DET> eingebunden werden, und hier auf alle Determinatoren zugreifen können. Darüber hinaus werden reguläre Ausdrücke zur Erkennung von Zahlen mittels <NB> und Wörter mittels <MOT> eingesetzt. Zudem stellen Lexika wichtige Ressourcen bei der Analyse der Sentiments dar und enthalten Polaritätswörter wie z. B. positive oder negative Adjektive, oder das aus dem Korpus selbst gewonnene Fachvokabular.

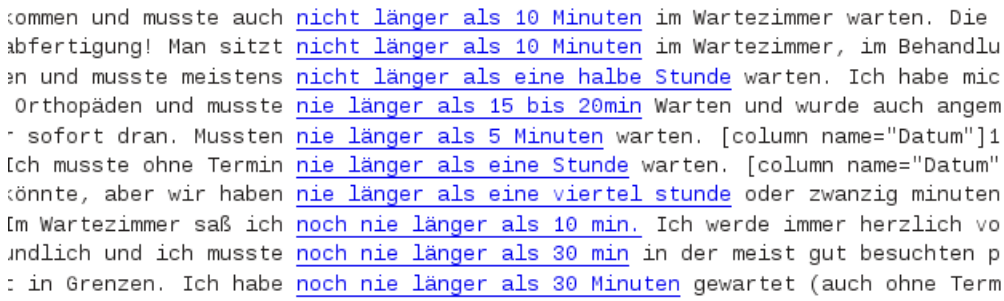


Abbildung 2: *Auszug aus der Konkordanz positiver Äußerungen über Wartezeiten in der Praxis*

Iterative Evaluation

Begleitend zur Entwicklung des Verfahrens wird eine iterative Evaluation durchgeführt, die als Basis für Entwurfsentscheidungen oder Korrektur derselben dient. Deshalb wird sowohl die Erkennung der Bewertungskategorien in den Bewertungskommentaren als auch die Identifikation der kognitiven Effekte in mehreren Schritten erfolgen. Nach jeder Etappe werden Zwischenevaluationen durchgeführt, um so die Erkennungsprobleme zu identifizieren und die lokalen Grammatiken zu verbessern.

Überbewertung: Positiv + Negativ = Positiv ?

Am Beispiel eines kognitiven Effekts, der aufgrund der Überbewertungen einer Kategorie durch die Hervorhebung einer anderen entsteht, soll illustriert werden, inwiefern dieser automatisiert in Online-Bewertungen mittels oben genannter Methodik identifiziert werden kann. Wie die Bezeichnung des Effekts bereits sagt, geht es hier um eine Bewertungskategorie, deren Benotung möglicherweise positiver ausfällt als dies in der Wirklichkeit der Fall ist. Als exemplarische Kategorie wird „Wartezeit (Praxis)“ herangezogen. In diesem Auszug aus der Datenbasis wurden alle Aussagen und Noten der Patienten zu den Wartezeiten in Arztpraxen durch lokale Grammatiken annotiert. Ein kognitiver Effekt der Überbewertung liegt in diesem Fall vor, wenn die Wartezeit von Patienten besser benotet als die Situation beim Warten in der Praxis von ihnen tatsächlich wahrgenommen wird.

Herr Dr. Brachvogel nimmt sich für seine Patienten ausreichend Zeit, was gelegentlich auch dazu führen könnte, dass man <N Kategorie="WZP_4">länger warten</N> muss!

In dieser Bewertung war die Note 1,0 für die Kategorie „Wartezeit (Praxis)“. Dabei ist auffällig, dass die Wartezeit zwar als „lang“ wahrgenommen wird, jedoch vorher eine andere Kategorie (in diesem Fall „Genommene Zeit“) positiv hervorgehoben wird. So ist es meist bei diesem Effekt der Fall, dass entweder im gleichen Satz oder in unmittelbarer Nähe von dem betreffenden negativen Muster zur Erkennung der verbrachten Zeit im Wartezimmer eine andere Kategorie (oft „Genommene Zeit“, „Behandlung“ oder „Freundlichkeit“) positiv hervorgehoben wird. In anderen Fällen werden auch positive Äußerungen zu mehreren Kategorien aufgelistet und damit zum Schluss die langen Wartezeiten begründet, was ebenfalls teilweise zum genannten Effekt führen kann.

Obwohl die angesprochene Wahrnehmung der Wartezeitdauer und somit die Polarität der obigen Aussage offensichtlich zu sein scheint, könnten Kritiker behaupten, dass die Polarität der Kategorie „Wartezeit (Praxis)“ in dem genannten Kontext nicht negativ (N), sondern positiv (P) wäre. Somit würde ebenfalls die Existenz des hier angesprochenen Effekts infrage gestellt werden. Als Gegenargument sei angemerkt, dass sich bei der Korpusanalyse zeigt, dass es eine Vielzahl ähnlicher Äußerungen gibt, bei denen sich die Patienten zu langen Wartezeiten bereiterklären und diese durch die Fachkompetenz oder Freundlichkeit des Arztes rechtfertigen, jedoch die „Wartezeit (Praxis)“-Kategorie selbst negativ benoten. Dies spiegelt wiederum die bei der allgemeinen Definition von kognitiven Effekten angesprochene Erwartungsnorm wider, von der die Bewertungsfehler abweichen.

Insgesamt konnten in ungefähr 22% der Bewertungen, in denen Äußerungen zur Wartezeit in der Arztpraxis gemacht werden, kognitive Effekte identifiziert werden. Dementsprechend handelt es sich bei 78% um keine kognitiven Effekte, wie in folgendem Beispiel illustriert wird:

Immer freundlich, kompetent, empathisch [column Name="Bewertung"] Die <N Kategorie= "WZP_4">langen Wartezeiten</N> nimmt man da gerne in Kauf. [...] [column Name="b_WartezeitPraxis"]5.0[/column]

Bestehende Herausforderungen

Der hier vorgestellte interdisziplinär ausgerichtete Forschungsansatz schlägt eine Brücke zwischen den Kognitionswissenschaften und der Computerlinguistik, indem sozialpsychologische Phänomene aus sprachwissenschaftlicher Perspektive maschinell erschlossen werden. Das Ziel dieser Studie ist es, kognitive Effekte über sprachliche Muster zu definieren und in Patientenbewertungen automatisch identifizieren zu können.

Die ersten Probleme und Erkenntnisse der durchgeführten Studie lassen sich wie folgt zusammenfassen:

- Die Genauigkeit der Mustererkennung ist besser als deren Abdeckung; bei der Identifikation von kognitiven Effekten ist dies umgekehrt der Fall. Je höher die Extraktionsgenauigkeit, desto besser und präziser ist die Identifikation der kognitiven Effekte.
- Die Problematik der Mustererkennung hat hauptsächlich mit der Sprachspezifik der Kundenbewertungen zu tun, wobei die Sprachstile von einer gesprochenen Sprache bis hin zu literarischen Ausdrücken reichen.
- Die Verbesserung der Extraktionsqualität kann durch die Erweiterung bzw. Aufnahme zusätzlicher Muster in lokale Grammatiken erreicht werden. Zahlreiche Phänomene müssen kategorienübergreifend behandelt werden, wie z. B. die Extraktion ironischer Aussagen, Emoticons oder Smileys, Tippfehler, Polaritätsbestimmung auf der Satzebene usw.

Literatur

Caverni, J. P., Fabre, J. M., & Gonzalez, M. (Hrsg.) (1990): *Cognitive biases*. New York u.a.: Elsevier. S. 7–12, 59–68.

Gross, M. (1997): *The Construction of Local Grammars*. In E. Roche und Y. Schabès (Hrsg.): Finite-State Language Processing. Language, Speech und Communication, Cambridge, Mass.: MIT Press. S. 329–354.

Hatzivassiloglou, V., & McKeown, K. R. (1997): *Predicting the semantic orientation of adjectives*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. S. 174–181.

Kim, S. M., & Hovy, E. (2004): *Determining the sentiment of opinions*. In Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics. S. 1367–1374.

Kim, S.-M. & Hovy, E. (2006): *Automatic identification of pro and con reasons in online reviews*. In Proceedings of the Poster Session at the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17.-21. Juli 2006, S. 483–490.

Schneider, S. (2013): *Das 3-Dimensionsmodell der Wissensrekonstruktion: A priorische Sicherstellung der Güte generierten Wissens.* Forschungsbericht. S. 5–22. http://www.fhkiel.de/fileadmin/data/wirtschaft/dozenten/schneider_stephan/Science/ResearchReport/Schneider_2013_Informationspathologien.pdf (16.09.2014).

Stürmer, S. (2009): *Sozialpsychologie*. München: Reinhardt. S. 11–16, 69–90, 165–217.

Turney, P. (2002): *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn., 7.-12. Juli 2002, S. 417–424.

Wolff, D. (1993): *Der Beitrag der kognitiv orientierten Psycholinguistik zur Erklärung der Sprach- und Wissensverarbeitung.* In: Gienow, W. / Hellwig, K. (Hrsg.): Prozeßorientierte Mediendidaktik im Fremdsprachenunterricht. Frankfurt am Main u.a.: Lang. S. 27–41.

Beziehung und Bedeutung. Soziale und semantische Netzwerkanalyse religionshistorischer Korpora

Abstract für die Dhd-Tagung 2015 in Graz

Frederik Elwert, Ruhr-Universität Bochum

Simone Gerhards, Ruhr-Universität Bochum

Sven Sellmer, Ruhr-Universität Bochum

Mit der fortschreitenden Digitalisierung historischer Textsammlungen steht ein immer größerer Quellenfundus für computergestützte Analysen zur Verfügung. Auch wenn in den letzten Jahren neue Analysetechniken wie etwa das *topic modeling* zunehmend in den Geisteswissenschaften angewandt worden sind, ist das Potenzial der verfügbaren Daten für die Gewinnung neuer Erkenntnisse – um das Tagungsthema zu zitieren – längst noch nicht ausgeschöpft. Das BMBF-geförderte eHumanities-Projekt „Semantisch-soziale Netzwerkanalyse als Instrument zur Erforschung von Religionskontakten“ (SeNeReKo) ist mit der Anwendung und Weiterentwicklung von Methoden textbasierter Netzwerkanalyse befasst. Am Beispiel von religionshistorischen Korpora wie dem buddhistischen Pali-Kanon, dem altindischen Epos *Mahābhārata*und dem Thesaurus Linguae Aegyptiae, einer Sammlung altägyptischer Quellen, sollen die Erkenntnisse aus der dreijährigen Projektarbeit präsentiert werden.

Methoden der Netzwerkanalyse haben sich mittlerweile im Methodenkanon der *digital humanities* etabliert. Dafür lassen sich insbesondere zwei Gründe anführen: Zum einen entsprechen sie theoretischen Entwicklungen in den Geistes- und Sozialwissenschaften, die das Relationale gegenüber isolierten Entitäten betonen. Diese Tendenz findet sich etwa in der Geschichtswissenschaft (Verflechtungsgeschichte, Beziehungsgeschichte u.a., etwa Osterhammel 2001), in der Soziologie (Relational Sociology, Emirbayer 2007) oder in der Linguistik (Fuzzy Semantics, Rieger 1989). Zum anderen stellt die Graphentheorie ein formales Modell bereit, das relationale Systeme unterschiedlichster Provinienz repräsentieren und quantitativen Analysen zugänglich machen kann. Was die zentralen Netzwerkkomponenten, die Knoten und Kanten des jeweiligen Modells sind, unterscheidet sich dabei je nach Anwendungsfall. Netzwerkansätze in Geschichts- und Literaturwissenschaft beschäftigen sich zumeist mit (historischen oder literarischen) Akteuren und weisen damit eine gewisse nähe zu *social network analysis* auf, wie sie sich im Kontext der Soziologie entwickelt hat (etwa Gramsch 2013). Ansätze in Linguistik und Computerlinguistik stellen dagegen eher auf sprachliche Strukturen ab, die als Netzwerke repräsentiert werden können (etwa Ferrer i Cancho, Solé, and Köhler 2004). Das SeNeReKo-Projekt hat zum Ziel, beide Perspektiven zu verbinden, um die Netzwerkanalyse für die Religionsforschung fruchtbar zu machen. Die interdisziplinäre Religionsforschung zeichnet sich dabei durch eine Pluralität sowohl der Gegenstände (nach Zeit, Ort und Genre) als auch der Zugänge (philologische, historische und soziologische) aus. Sie ist daher ein fruchtbares Feld, um neue methodische Ansätze zu erproben.

Der Vortrag stellt anhand ausgewählter, im SeNeReKo-Projekt bearbeiteter religionshistorischer Korpora zentrale Ergebnisse des Projekts vor. Im Zentrum der Methode stehen dabei Verfahren der textbasierten Netzwerk-Erzeugung, die Ausgangspunkt für unterschiedliche Analysestrategien sind.

Die Grundlage für die Untersuchung der altägyptischen Textzeugen bildet die digitale Datenbank des Thesaurus Linguae Aegyptiae (TLA) der Berlin-Brandenburgischen Akademie der Wissenschaften. Ihre Sammlung umfasst Textkorpora der unterschiedlichsten Gattungen, wie beispielsweise literarische Erzählungen, Jenseitsliteratur, königliche Dekrete oder magische Sprüche, und beinhaltet dabei mehr als 1.200.000 Textworte aus knapp 2.000 Jahren Geschichte.

Aus dem indischen Bereich beschäftigt sich das Projekt zum einen mit der als *Pāli-Kanon* bekannten Sammlung altbuddhistischer Texte (ca. 4.-1. Jh. v. u. Z). Aus diesem wurden die aus religionswissenschaftlicher Sicht interessantesten Teile (im Umfang von ca. 1,7 Mio. Wörtern) ausgewählt. Zwar lagen diese Daten schon in digitalisierter Form vor, mussten jedoch für die durchgeführten Analysen in umfangreicher Weise aufgearbeitet werden. Zum anderen wird auch das in Sanskrit verfasste indische Nationalepos *Mahābhārata* (ca. 3. Jh. v. – 3. Jh. n.. u. Z.) untersucht. Dieser umfangreiche Text (ca. 1 Mio. Wörter) ist ein dankbares Studienobjekt, da er eine Art Gründungsdokument des Hinduismus darstellt und daher verschiedene religiöse Strömungen seiner Entstehungszeit widerspiegelt. Die computergestützte Analyse des *Mahābhārata* wird dadurch erleichtert, dass ein großer Teil semantisch und syntaktisch annotiert vorliegt.

Um nun von den Textdaten zu Erkenntnissen zu gelangen, werden auf Basis linguistischer Annotationen alle relevanten Untersuchungseinheiten (Entitäten und Lemmata) und ihre relationale Struktur als Graph dargestellt und so für die Netzwerkanalyse erschlossen. Relationen von Einheiten werden dabei durch Kookkurrenzen definiert, wobei die Kontextfenster je nach Untersuchungsziel unterschiedlich weit gefasst werden können. Dies erlaubt dann unterschiedliche Analysen. Für die Bedeutungsanalyse einzelner Worte kann das Netzwerk auf ein zu erforschendes Lemma und alle mit diesem in Verbindung stehenden Begriffe reduziert werden. Auf diese Weise können sowohl kookkurrente Worte mit einer hohen Zentralität als auch mögliche Cluster engverwandter Begriffe aufgedeckt werden. Dieses Verfahren ermöglicht es, relevante Informationen aus einem umfangreichen Textkorpus (1) zu ermitteln und sie (2) durch verschiedene Visualisierungstechniken schneller und leichter interpretierbar zu machen. Für eine soziale Netzwerkanalyse kann für das Korpus (oder ausgewählte Teilkorpora) ein Netzwerk nur der sozialen Akteure erstellt werden. Ein besonderer Ansatz ergibt sich aus der Kombination beider Ansätze: In den jeweiligen Netzwerken können nicht nur alle mit einem bestimmten Lemma in Verbindung stehenden Begriffe als solche abgebildet, sondern auch die in diesem Kontext agierenden sozialen Akteure dargestellt werden. Darüber hinaus können über Kookkurrenzen einer bestimmten Entität nicht nur alle verbundenen Akteure aufgedeckt werden, sondern zudem gezeigt werden, mit welchen semantischen Begriffen diese verbunden sind. Anhand von konkreten Beispielen sollen die Verbindung aus semantischer und sozialer Netzwerkanalyse vorgestellt und ihre Möglichkeiten für die Fächer Religionswissenschaft, Ägyptologie und Indologie diskutiert werden.

Literatur

Emirbayer, Mustafa. 2007. ‘Manifesto for a Relational Sociology.’ *American Journal of Sociology* 103 (2): 281–317.

Ferrer i Cancho, Ramon, Ricard V. Solé, and Reinhard Köhler. 2004. ‘Patterns in Syntactic Dependency Networks’. *Physical Review E* 69 (5): 051915. doi:10.1103/PhysRevE.69.051915.

Gramsch, Robert. 2013. *Das Reich als Netzwerk der Fürsten: politische Strukturen unter dem Doppelkönigtum Friedrichs II. und Heinrichs (VII.) 1225-1235*.

Osterhammel, Jürgen. 2001. *Geschichtswissenschaft jenseits des Nationalstaats: Studien zu Beziehungsgeschichte und Zivilisationsvergleich*. Kritische Studien zur Geschichtswissenschaft ; 147. Göttingen: Vandenhoeck & Ruprecht.

Rieger, Burghard B. 1989. *Unscharfe Semantik: Die Empirische Analyse, Quantitative Beschreibung, Formale Repräsentation Und Prozedurale Modellierung Vager Wortbedeutungen in Texten*. Frankfurt am Main u. a.: Lang.

Auf der Suche nach dem erfüllten Raum: Digitale Korpusanalyse in der Literaturwissenschaft am Beispiel von Ilse Aichinger

Christine Ivanovic (Universität Wien)
Andrew U. Frank (Technische Universität Wien)

Universität Wien, Abt. für Vergleichende Literaturwissenschaft
Sensengasse 3A A-1090 Wien
christine.ivanovic@univie.ac.at

TU Wien, Geodäsie und Geoinformation
Gusshausstrasse 27-29/E120.2
A-1040 Wien
frank@geoinfo.tuwien.ac.at

1 Einleitung

Literarische Topographien sind mehr als geographisch verifizierbare Ortsangaben. In Literatur kann auf komplexe Weise das Profil historisch-kulturell identifzierbarer Räume musterhaft gestaltet, zu „Topoi“ verdichtet und diese als Kommunikationsformen sui generis eingesetzt werden. Ihre Analyse ermöglicht Aussagen über den Funktionszusammenhang von Gesellschaften im historischen Wandel. Sie bedingt notwendigerweise eine genaue Bestimmung der sprachlich-literarischen Formalisierung, welche die dichte Raumgestaltung im Rahmen literarischer Topographien erst ermöglicht.

2 Literarische Topographien in Digitaler Analyse

Um die Anlage und Wirkungsweise literarischer Topographien zuverlässig und umfangreich zu erfassen, versucht man seit einiger Zeit digitale Methoden für die Analyse genau begrenzter Korpora einzusetzen. Dabei werden verschiedene Ansätze verfolgt: Der historiographisch-diskursanalytisch orientierte Ansatz von

Franco Moretti beruhte auf statistischen Methoden, wie sie in der Soziologie Anwendung finden [1999], während eine Forschergruppe der ETH Zürich um Barbara Piatti eng in Verbindung mit der Kartographie arbeitete, um präzise Profile für die literarische Gestaltung konkreter geographischer Räume in eng umgrenzten Zeiträumen zu erstellen[2008, 2009]. Todd Presner begründete mit dem Projekt „HyperCities“ [2009] ein *Thick Mapping* insbesondere im Bereich der Großstadtliteratur.

Diese Ansätze entfernen sich dezidiert von traditionellen Methoden philologischer Textanalyse und deren Darstellungsmodi: Moretti vernachlässigte vor allem in seinen frühen Arbeiten vorsätzlich die Analyse der Textstrukturen zugunsten einer Erhebung all jener Daten, die sich auf die Schauplätze von Literatur beziehen (intra- wie extratextuell); Piatti und Presner konzentrieren sich auf digitale Visualisierungen zur Veranschaulichung der von ihnen aus den Texten gewonnenen topographischen Daten, um damit neue Analyseergebnisse zu generieren.

In dem hier vorzustellenden Projekt hingegen wird ein geschlossenes Œuvre untersucht im Hinblick auf die Generierung literarischer Topographien, auf die den Text strukturierenden Verfahren und auf die allgemeine Funktion der Raumdarstellung.

3 Raumreferenzen im Werk von Ilse Aichinger

Das Gesamtwerk der österreichischen Autorin Ilse Aichinger (geb. 1921 in Wien) bietet aus mehreren Gründen ein dafür geeignetes Untersuchungsbeispiel:

3.1 Günstige Korpusgröße

Aichingers Werk entstand innerhalb von sechs Jahrzehnten (1946-2006) und umfasst insgesamt 2100 Seiten. Die Texte liegen in einer leicht greifbaren Werkausgabe in 8 Bänden [1991] sowie in vier danach erschienenen Einzelbänden vor. Alle Bücher wurden gescant und in eine OCR-Textdatei umgewandelt.

3.2 Klare Profilierung literarischer Topographien

Aichinger bezieht sich in einem großen Teil ihrer Werke explizit oder implizit auf den Raum der Stadt *Wien*, in der sie geboren wurde, wo sie die Judenverfolgungen miterlebte und in die sie nach jahrzehntelanger Abwesenheit 1988 zurückkehrte[Fässler, 2014].

Des weiteren profiliert sie den Raum *England* als Sehnsuchtsort [Ivanovic, 2011a] sowie weitere Räume, an denen sie sich kürzere Zeit aufhielt (Frankreich, USA) oder wo sie länger lebte (österreichisch-bayerisches Grenzland).

Schließlich elaboriert sie *Heterotopien* wie das Kino [Ivanovic, 2011b], den Friedhof, das Meer sowie *typische/topische Räume*.

4 Arbeitshypothese

Aichinger „füllt“ einen (geographisch identifizierbaren, realen) Erfahrungsraum mit diversen, oftmals heterogenen Erinnerungsstücken unterschiedlicher geographischer, historischer und medialer Provenienz und verkettet diese mittels gleichbleibender sprachlicher Verfahren. Sie generieren keine neue Bedeutung im Sinne syntagmatischer Verknüpfung, stellen aber auf paradigmatischer Ebene neue Zusammenhänge her.

Dieser Vorgang entspricht grundsätzlich der Verknüpfung von Daten im www: die Anhäufung von Daten wird durch Verweise neu gruppiert und generiert dadurch neue Datensätze, die von sich aus aber noch keine in sich abgeschlossene, diskursiv organisierte Aussage bilden.

Bei Aichinger entsteht dadurch im Text - und nur hier - ein „erfüllter Raum“. Dieser bildet weder einen realen Raum noch ein einzelnes (historisches) Ereignis, das hier stattfand (oder stattfindet), ab. Der Text stellt vielmehr dar, wie das Bewusstsein an diesem konkreten Ort im Durchgang durch verschiedene Zeitschichten Geschichte bearbeitet.

Die digitale Analyse setzt bei der systematischen Erfassung der Raumreferenzen an. Sie zielt auf die Erarbeitung der strukturellen Merkmale und textuellen Strategien zur Generierung literarischer Topographien.

5 Erfassung

Die digitale Erfassung der Texte soll dementsprechend Analysen in zwei Untersuchungsbereichen ermöglichen:

1. Welche Orte werden mit welchen „Erinnerungsstücken“ verknüpft?
2. Welche wiederkehrenden sprachlich-literarischen Konstruktionen lassen sich dabei feststellen?

Mit dem gewählten Verfahren lassen sich die in Frage 1 fokussierten Informationen - Nennung von Orten, Zeiten, Personen, Medien - systematisch und vollständig erfassen. Das Erfassungssystem ist so aufgebaut, dass damit auf einfache Weise Abfragen auch in Bezug auf die in Frage 2 fokussierten sprachlichen Konstruktionen - Wiederholungen, Paarungen, Kombinationen, Kontextbildung etc. - möglich werden.

6 RDF als Hilfsmittel zur Codierung: Aus Informationen werden Daten

Die in Aichingers Texten explizit genannten Personen sowie die Orts- und Zeitangaben werden systematisch, einheitlich und vollständig erfasst. Dazu wird das gesamte publizierte Werk von Ilse Aichinger mit Hilfe einer speziellen Markupsprache kodiert und danach in RDF (Ressource Description Framework)[Manola

et al., 2004, Hitzler et al., 2008] codiert: Aus Informationen werden digital analysierbare Daten.

Die Erfassung dieser Informationen im Gesamtkorpus ermöglicht eine Auswertung, die zuverlässige Aussagen über das Gesamtwerk der Autorin erlaubt. Die computergestützte Auswertung ersetzt die literaturwissenschaftliche Analyse nicht: RDF und SPARQL werden lediglich als zeitgemäße Werkzeuge zur Sichtung und Aufbereitung der in den Texten vorhandenen Informationen eingesetzt. Und von allen Informationen, die ein literarischer Text enthält, wird in unserem Ansatz lediglich eine einzige Kategorie in drei unterschiedlichen Bereichen berücksichtigt: zunächst geht es nur und allein um die Erfassung von Personennamen sowie Orts- und Zeitangaben.

7 Digitale Analyse

Der annotierte Text wird in RDF (im Turtle Format) formatiert und in einen SPARQL endpoint eingebracht (wir verwenden 4store). Einfache Abfragen zur Kontrolle der Erfassung und zum Finden von Text-Abschnitten nach den üblichen Kriterien (Seitenzahl, Titel, Textstellen) werden von einem derartigen System erwartet.

Beispiele für quantitative Abfragen:

- Wie viele (und welche) Texte enthalten Bezugnahmen auf geographisch referenzierbare Orte, wie viele (und welche) Texte enthalten nur Bezugnahmen auf typische Orte? Welche Text enthalten keine Bezüge zu Orten?
- Welche sind die am häufigsten von Aichinger in Wien genannten Orte, welche kommen nicht vor?
- In welchem Zeitraum ihres Schreibens spricht sie von Wien, in welchem Zeitraum gibt es keine Bezugnahmen auf die Stadt?
- Wie verhalten sich Nennungen öffentlicher Räume zu Nennungen von privaten Lokalitäten?
- Wie verhalten sich Bezugnahmen auf Aussenräume zu Bezugnahmen auf Innenräume?

Die Darstellung von Orten in einer Karte erlaubt den Überblick über den durch den Text beschriebenen Raum und zeigt allenfalls auch Räume, die ausgespart werden („Negative Räume“) und Ziel einer möglicherweise lohnenden Analyse sein könnten. Es kann untersucht werden, ob sich der örtliche Schwerpunkt der Bezüge im Laufe der Schreibzeit oder im Laufe der Erzählzeit verändert. All dies kann in Diagrammen und Chronologien dargestellt werden.

```
»Haben schon gewählt?« heißt es bei Demel. Aber welche Wahl
im Leben ist offen?
.propText
        lit:ort         :Demel.
:Demel  lit:name        "K. & K. Hofzuckerbäcker Demel";
        lit:google      "Kohlmarkt 14 , 1010 Wien";
        lit:wiki        "Demel".
```

Abbildung 1: Ausschnitt aus codiertem Text; ein Ort muss nur einmal im Detail (*name*, *google*) beschrieben werden, später genügt der eingeführte Bezeichner (*Demel*).

8 Wahl der technischen Lösung

Unsere Ziele für die technische Realisierung waren: (1) Langfristige Nutzung: Die erzeugten Daten können in spätere Anschlusssysteme migrieren; (2) Quelloffen; (3) Standardisierung.

Die von der W3C, dem Steuerungsgremium des World Wide Web, in den Richtlinien "Semantic Web" [Berners-Lee et al., 2001] vorgeschlagenen Methoden zur Erfassung von Bedeutung in Form von strukturierten Daten und einer zugehörigen Abfragesprache, scheinen für die vorliegende Aufgabe passend. Die Trennung in eine Markup Sprache, die Kodierungen im Text erlaubt, und die Speicherung und Verarbeitung mit RDF hat sich soweit bewährt.

Die technische Lösung und insbesondere die Codierung muss ausbaufähig bleiben; die gesuchten Eigenschaften können möglichst objektiv bestimmt werden. Das erlaubt es, die Codierungs-Arbeit z.B. in einem Seminar auf alle Teilnehmer aufzuteilen (crowdsourcing) und dennoch einen einheitlichen Datenbestand zu generieren.

Die Codierung muss kontextrelevant angelegt sein, wobei der relevante Kontext der Abschnitt (nicht ein Wort, nicht ein Satz) ist, in dem ein Ort erwähnt wird. Bei Abfragen werden die angesprochenen Abschnitte ausgegeben um die Interpretation zu gewährleisten.

8.1 Methode der Annotation

In die mittels OCR erzeugten Textdateien werden nach jedem Abschnitt in RDF Turtle Syntax die erwähnten Orte, Personen und Zeiten codiert und mit anderen Datenbanken (z.B. Google Maps oder Wikipedia) verknüpft (Fig. 1). Dieser annotierte Text wird dann von einem eigenen Programm in RDF Syntax umgewandelt und diese in einen SPARQL endpoint geladen. Damit kann aus dem Web mit einem üblichen Browser gesucht und die Ergebnisse angezeigt werden.

Abbildung 2: Ein Ausschnitt mit einigen Textstellen zu Wien

8.2 Kartographische Ausgabe

Karten, die die gewonnene Information räumlich darstellen, sind ein wichtiger Teil des Projektes und lassen sich leicht mit bereits bestehenden Werkzeugen erstellen:

1. Textstellen, mit bestimmten Eigenschaften und die darin erwähnten Orte werden mittels einer SPARQL Abfrage herausgesucht und als CSV Datei dargestellt (CSV - Comma Separated Values, ein übliches Format für die Übermittlung von Tabellen). Zum Beispiel, um alle Abschnitte, die einen konkreten Ort beschreiben, herauszusuchen und mit der Ortsbeschreibung auszugeben genügt:

```
Select ?absatz ?titel ?googleName where
 { ?absatz lit:ort ?ort .
        ?ort lit:google ?googleName.
        ?absatz lit:titel ?titel .
  }
```

2. Die Datei mit den drei Kolonnen *Absatz*, *Titel* und *Location* wird in Google Docs hochgeladen und dort mittels Fusion Tables angezeigt. Dabei werden die im Feld *Location* (= *googleName*) verwendeten Adressen in Koordinaten umgesetzt und als rote Punkte angezeigt; die Karte kann im Massstab verändert und der Ausschnitt frei gewählt werden; beim Klicken auf einen Punkt wird die Textstelle geöffnet (Fig. 2):

Auf ein Beispiel kann unter

https://www.google.com/fusiontables/DataSource?docid=
1BCoyc2Pho9GkVTTc9gfKbfQJ4Mf4nHiy6-33rQww#map:id=3

zugegriffen werden.

9 Zusammenfassung und Ausblick

Ziel ist die Erarbeitung einer zuverlässigen, anwendungsneutralen und anpassungsfähigen digitalen Analysemethodik, die auch bei anderen Textcorpora und veränderten Fragestellungen eingesetzt werden kann. Besonders interessant ist die Möglichkeit der Verknüpfung mit bestehenden Datenbanken; bei Aichinger z.B. die Verbindung zwischen den erwähnten Filmnamen und der Filmdatenbank, die Darsteller etc. bereitstellt und in SPARQL Abfragen nutzbar macht. Es wird also möglich, alle Abschnitte, in denen ein Film mit einem bestimmten Schauspieler erwähnt wird, zu finden.

Literatur

Ilse Aichinger. *Werke*. Fischer-Taschenbuch-Verl., 1991.

Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

Simone Fässler. *Von Wien her, auf Wien hin. Ilse Aichingers "Geographie der eigenen Existenz"*. Böhlau, 2014.

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, and York Sure. Semantic web. *Berlin, Heidelberg*, 2008.

Christine Ivanovic. Nach England! Zur Geschichte einer Sehnsucht. In Rüdiger Görner, Christine Ivanovic, and Sugi Shindo, editors, *Wort-Anker Werfen. Aichinger und England*. Königshausen & Neumann, 2011a.

Christine Ivanovic. Masse. Medien. Mensch. Ilse Aichingers bioskopisches Schreiben. In Christine Ivanovic and Sugi Shindo, editors, *Absprung zur Weiterbesinnung. Geschichte und Medien bei Ilse Aichinger*. Stauffenburg, 2011b.

Frank Manola, Eric Miller, Brian McBride, et al. RDF primer. *W3C recommendation*, 10(1-107):6, 2004.

Franco Moretti. *Atlas of the European novel, 1800-1900*. Verso, 1999.

Barbara Piatti. *Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien*. Wallstein, 2008.

Barbara Piatti, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright. *Mapping literature: Towards a geography of fiction*. Springer, 2009.

Todd Samuel Presner. Hypercities: Building a web 2.0 learning platform. *Teaching Literature at a Distance*, page 171, 2009.

Elisabeth Burr
elisabeth.burr@uni-leipzig.de
Universität Leipzig

Julia Burkhardt
jburk@rz.uni-leipzig.de
Universität Leipzig

Elena Potapenko
potep@rz.uni-leipzig.de
Universität Leipzig

Rebecca Sierig
rebecca.sierig@uni-leipzig.de
Universität Leipzig

Arámis Concepción Durán
acduran@uni-leipzig.de
Universität Leipzig

DAS DUISBURG-LEIPZIG KORPUS ROMANISCHER ZEITUNGSSPRACHEN UND SEIN TEXTMODELL

1. EINLEITUNG

Ziehen wir das 574 Seiten starke *Book of Abstracts* der Internationalen ADHO Konferenz *Digital Humanities 2014* vom Juli 2014 in Lausanne heran, dann wirft eine Suche nach *newspaper(s)* / *magazine(s)* / *periodical(s)* eine Vielzahl von Belegstellen aus. Eine Suche nach Zeitung(en) in den Abstracts von *DHd 2014*, der ersten Konferenz des deutschsprachigen Fachverbandes, die im März desselben Jahres in Passau stattgefunden hat, lässt sich leider nicht so komfortabel durchführen, da die Abstracts nicht zu einem digitalen Band zusammengeführt wurden, Zeitungen werden aber zumindest in einem Abstract genannt (cf. Fischer / Kirsten / Witt 2014). Schauen wir uns dann die Projekte, bei denen Digitalisierung und Erschließung von Zeitungen / Magazinen / Zeitschriften eine Rolle spielen, genauer an, so werden wir feststellen, dass es dabei entweder um Kuration oder um Sprachkorpora geht. So versuchen die einen Sammlungen historischer Zeitungen / Zeitschriften / Magazine in Form von digitalen Facsimilies (vgl. *Modernist Magazines Project*) oder hochwertigen digitalen Editionen (vgl. z. B. *The Modernist Journals Project* oder *Die Fa-*

ckel) verfügbar oder große Mengen an Zeitungsseiten durchsuchbar zu machen (vgl. etwa *Europeana Newspapers*). Die anderen nehmen dagegen Zeitungen / Magazine / Zeitschriften in Referenzkorpora auf, die der Untersuchung des schriftlichen Gebrauchs der jeweiligen Sprache dienen (vgl. z. B. *Das Deutsche Referenzkorpus* – DeReKo oder *Coris/Codis*) oder transformieren eine Vielzahl von Ausgaben eines bestimmten Blattes zu einem Zeitungskorpus (vgl. z.B. das *La Repubblica Corpus* oder *The Time Magazine Corpus*). Wird eine Verbindung zwischen beiden Formaten versucht, so beschränkt sich diese in der Regel auf die Integration von mehr oder minder ausgefeilten Werkzeugen, die eine Suche nach sprachlichen Phänomenen (zumeist lexikalischen) in den digitalen Editionen und das Aufrufen der den einzelnen Belegstellen entsprechenden Texte erlauben (vgl. etwa *Die Fakel*, *Der Brenner* oder *The Modernist Journals Project*).

Solche Werkzeuge sind allerdings nicht neutral, sondern es liegt ihnen eine bestimmte Konzeption bzw. Modellierung der jeweiligen Artefakten zugrunde. So spricht Harro Biber etwa von Magazinen als „container of texts“ (Ermolaeva et al. 2014: 52) und als erstes Ziel von *Europeana Newspapers* wird angegeben „Europeana Newspapers [...] will create full-text versions of about 10 million newspaper pages. It will also detect and tag millions of single articles with related metadata and named entities (information identifying people, locations etc.).“ (EurNews 2014). Die *Austrian Academy of Sciences* ist sich zwar bewusst, dass historische Zeitungen, wie der Sozialdemokrat „nicht nur anderer technischer Handhabung [bedürfen], sondern auch anderer Auszeichnungskriterien“ als die „literarischen Texten, die im AAC bearbeitet werden“, beschränkt die Unterschiede aber auf „Textsortenzuordnung, [...] spezifische[...] Annotierung von Pseudonymen, chiffrierten Autoren- und anderen Personennamen“, obwohl die Illustration zu den genannten Kriterien mit einer Zeitungsseite eine viel komplexere Realität abbildet (AAC o.J. c). Konzeptionen wie Textcontainer, Ansammlung individueller Texte oder Datenkonserve liegen auch vielen Korpora zugrunde, die Tageszeitungen als Komponenten aufweisen oder insgesamt aus Tageszeitungen bestehen. Wulfman und Ermolaev kontern solche Konzeptionen zurecht mit: „In order to discuss the rela-

tionship of content elements in a magazine, for example – such as the relationship of advertisements to articles – one must have a common language for expressing layout. Simple text transcription of a magazine's content is insufficient for many kinds of research; thus our ontology is based on an understanding of the historical language of page composition (columns, paragraphs, various forms of headings, publication metadata, and so on) that is vital to the useful encoding of magazine structure and the analysis of a magazine's meaning." (Ermolaev et al. 2014: 53).

In unserem Beitrag werden wir ein Textmodell vorstellen, das auf eine originalgetreue Abbildung nicht nur der Struktur von Tageszeitungen und ihrer einzelnen Seiten zielt, sondern auch der komplexen Beziehungen, die auf einzelnen Seiten oder über ganze Teile der Zeitung hinweg angelegt sind.

2. DAS KORPUS ROMANISCHER ZEITUNGSSPRACHEN

Mit der Erstellung des heute als *Korpus romanischer Zeitungssprachen* vorliegenden Korpus wurde 1989 in Duisburg mit einem allein der Untersuchung der italienischen Zeitungssprache gewidmeten Korpus begonnen (cf. Burr 1993). Mitte der neunziger Jahre des letzten Jahrhunderts kam, wiederum in Duisburg, ein Korpus der italienischen, französischen und spanischen Zeitungssprache hinzu (cf. Burr 1997). Seit 2011 arbeitet eine Projektgruppe am Leipziger Lehrstuhl für französische, frankophone und italienische Sprachwissenschaft an der Erstellung eines Korpus aus französischen, québequer und italienischen Tageszeitungen (cf. Burkhardt et al. 2014).

2.1 Zusammensetzung und Größe des ausgezeichneten Korpus

In das Korpus gehen seit Beginn grundsätzlich ganze Zeitungsausgaben ein. Während die ersten beiden Korpora nach COCOA ausgezeichnet wurden, folgt das Markup des sich derzeit in der Entstehung befindlichen Korpus dem TEI-Standard P5 (cf. TEI Consortium 2014). Das ausgezeichnete Korpus setzt sich aktuell wie folgt zusammen:

- **Italienische Zeitungen - “Die Wende 1989” (724.517 Wortformen)**

Blatt	**Ausgabe**	**Wortformen**
Corriere della Sera	19, 20. und 21.10.1989	258.287
Il Mattino	20. und 21.10.1989	171.501
La Repubblica	20. und 21.10.1989	174.958
La Stampa	20. und 21.10.1989	119.771

- **Französische, italienische und spanische Zeitungen – “Europawahlen 1994” (801.010 Wortformen)**

Blatt	**Ausgabe**	**Wortformen**
Le Monde	12./13., 14. und 15.06.1994	236.236
Corriere della Sera	13., 14. und 15.06.1994	303.641
La Vanguardia	13., 14. und 15.06.1994	261.133

- **Französische, québequer und italienische Zeitungen – “Frauenfußballweltmeisterschaft 2011” (836.977 Wortformen)**

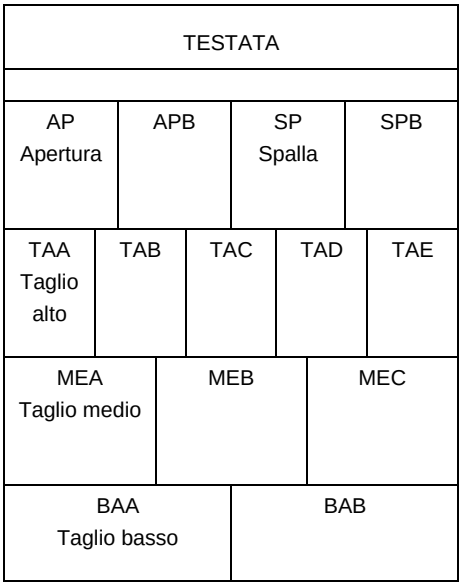
Blatt	**Ausgabe**	**Wortformen**
Le Monde	06. u. 20.07.2011	104.926
Libération	06. u. 20.07.2011	85.520
Le Parisien	20.07.2011	38.894
La Repubblica	06. u. 20.07.2011	260.277
La Stampa	06. u. 20.07.2011	347.360

2.2 Das Textmodell

Das Textmodell, das wie gesagt auf eine originalgetreue Abbildung der textuellen Inhalte und Strukturen der Quellen, also hier der Tageszeitungen zielt, wurde in seiner ersten Form (vgl. das *Korpus italienischer Zeitungssprache*) auf der Basis einer intensiven Auseinandersetzung mit der in den 70er und 80er Jahren nicht nur in der italienischen und deutschen Sprachwissenschaft, sondern gerade auch in der Medienwissenschaft und Publizistik geführten Diskussion um die Tagespresse, ihre Struktur und Sprache sowie ihre interne

Ausdifferenzierung mit Blick auf ein Massenpublikum erarbeitet (cf. Burr 1993: 125-174). Bei dem im Folgenden dargestellten Textmodell handelt es sich also um ein theorie- und forschungsbasiertes und das Medium als solches sowie auch seine Produktion (vgl. das Layoutschema unten) fokusierendes Modell:

Bezug	Kodierung	Beispiel
Zeitung als Fragment des Korpus	**<Z>**	<Z Stampa>
Ausgabe	**<E>**	<E 211089>
Sparte	**<S>**	<S Politica>
Autorenschaft	**<A>**	
unterschieden werden:		
a) signiert		<A firmato>
b) anonym		<A Non firmato>
c) Redaktion		<A Redazione>
Name des Autors	**<N>**	<N Ferrara Giovanni>
Positionierung des Textes	**<C>**	<C MEA01>



Textart	**<T>**	
unterschieden werden:		
a) die Typen von Überschriften		
Vorzeile		<T Occhiello>
Schlagzeile		<T Titolo>
Untertitel		<T Sottotitolo>
Zusammenfassung		<T Sommario>
Zwischenüberschrift		<T Catenaccio>
b) die festen journalistischen Formen		
Leitartikel		<T Fondo>
'Aufmacher'		<T Spalla>

Glosse		<T Corsivo>
Leserbrief		<T Lettera>
Antwort auf eine Leseranfrage		<T Risposta>
Kolumne		<T Rubrica>
Wetterbericht		<T Tempo>
Filmbesprechung		<T Film>
Kurznachricht		<T Breve>
Kurzmeldung		<T Flash>
Agenturmeldung		<T Agenzia>
Ankündigung		<T Riassunto>
c) die fließenden Formen		
Nachricht		<T Notizia>
Artikel		<T Articolo>
Kritik		<T Critica>
Spielbericht		<T Partita>
Interview		<T Intervista>
Darstellungsarten	**<P>**	
fortlaufender Text		<P Prosa>
direkte Rede		<P Discorso>
Zitat von schriftlich Geäußertem		<P Citazione>
Frage des Journalisten (Interview)		<P Domanda>
Antwort des Interviewten		<P Risposta>

(cf. Burr 1993:464-465)

Dieses Textmodell wurde im Zusammenhang mit der Erstellung des zweiten Korpus (vgl. das Korpus *Französische, italienische und spanische Zeitungen*) auf der Grundlage einer ausgiebigen Beschäftigung mit der englischen Korpuslinguistik und ihrer Diskussion um *Corpus Design* und *Sampling frames* weiter begründet und an einzelnen Stellen modifiziert (cf. Burr 1997).

Auch das Textmodell des aktuell im Entstehen begriffenen Korpus wurde unter Heranziehung der Forschung entwickelt. Das es umsetzende TEI Markup sollte nämlich zum einen den Forschungsinteressen, die den in den letzten 20 Jahren erschienenen korpusbasierten Untersuchungen zur französischen Pressesprache zugrunde liegen, gerecht werden, zum anderen den Elementen, die in den dem Zeitungsdesign gewidmeten linguistischen, pressewissenschaftlichen oder kommunikationswissenschaftlichen Arbeiten als konstitutiv für Tageszeitungen betrachtet werden und die ihrerseits den Sprachgebrauch determinieren oder für die Interpretation von Untersuchungsergebnissen bedeutsam sein können: Titelseite, Rubriken und Textsorten, Strukturierungselemente von Artikeln, Gestaltung des Textkörper, mögliche Komponenten des Anlaufs, Illustrationen und Legenden, synoptische Texte, Übersichtstexte und Textcluster (cf. Sierig

2013: 49-72). In dieses soll zu gegebener Zeit auch das COCOA-Markup der beiden früher schon erstellten und online verfügbaren Korpora (cf. Burr 1997-2004) überführt werden.

3. AUSBLICK

Die Entwicklung dieses Textmodells und des es umsetzenden TEI-Markups werden wir in unserem Beitrag begründen und diskutieren. Dass eine einfache Texttranskription nicht ausreicht und wir stattdessen Textmodelle brauchen, die der komplexen Struktur von Zeitungen und ähnlichen Artefakten Rechnung tragen, wenn wir Sprache tatsächlich in ihrem Gebrauch, das Wissen, das bei den Rezipierenden vorausgesetzt wird, oder die Semantik solcher Artefakten untersuchen wollen, können allein schon die folgenden Sierig (2013) entnommenen Beispiele zeigen:

Bausteine und Ebenen der *Une*


VOIX EXPRESS

Quel est votre pire souvenir avec un voisin de bureau ?

Alain Guichaoua
52 ans, chef de projet
Penmarc'h (29)

« Je partageais mon bureau avec un gros fumeur. Il était à plus d'un paquet par jour, alors que moi, je n'ai jamais fumé. A la maison, je demande à ma femme de fumer dehors. On avait un passé un accord avec ce collègue pour ouvrir la fenêtre. Mais il était frileux… Cela a duré deux-trois ans. A la fin nous avons eu des bureaux séparés. Et lorsque la loi antitabac a été votée, j'étais très heureux ! »

Nathalie Dupond
32 ans, en recherche d'emploi
Meaux (77)

« J'ai été victime de harcèlement moral de la part d'une collègue qui refusait de me former alors que je devais la remplacer pendant ses congés. Elle m'a accusé de faits que je n'avais pas commis. Elle était jalouse et se sentait menacée. Elle ne voulait pas qu'on montre qu'on bossait, alors qu'elle ne faisait rien. Tout le monde était pétrifié car elle avait plus de vingt-cinq ans de boîte. J'ai fini par partir. »

Marc Dumas
55 ans, chef d'entreprise
Versailles (78)

« J'avais une stagiaire qui sentait mauvais. Elle transpirait beaucoup et nous étions trois dans un petit bureau. Je l'adorais et je voulais l'embaucher. J'ai donc demandé à sa tutrice de stage de lui faire comprendre le problème en douceur. Mais elle lui a dit que la remarque venait de moi. C'était très gênant. Finalement, je l'ai embauchée et on a travaillé ensemble pendant quinze ans. »

Christiane Frazier
53 ans, assistante de direction
Aubervilliers (93)

« Je travaillais à côté de quelqu'un qui parlait tout le temps. Cela ne s'arrêtait jamais. C'était non-stop. Elle nous racontait sa vie, même si on cela ne nous intéressait pas. Il devenait impossible de se concentrer. Parfois, je perdais patience et je lui demandais d'arrêter de me parler, mais elle changeait d'interlocuteur. Comme on était dans un espace sans cloisons en open space… »

synoptische Texte

Text-Cluster

Bei der Begründung und Diskussion des Textmodells sollte auch deutlich werden, dass die Zeitungen nicht als Textcontainer, Ansammlungen individueller Texte oder Datenkonserven verstanden werden, sondern das Medium mit samt seinen komplexen Strukturen und internen wie externen Relationen als Artefakt betrachtet wird und das Korpus romanischer Zeitungssprachen deshalb letztendlich dem Humanities Computing / den Digitalen Humanities näher steht als der Korpuslinguistik computerlinguistischer Prägung.

AUSGEWÄHLTE QUELLEN

[AAC] = Austrian Academy Corpus (o.J. a): *Der Brenner* <http://corpus1.aac.ac.at/brenner/> [08.11.2014].

[AAC] = Austrian Academy Corpus (o.J. b): *Die Fakel* <http://corpus1.aac.ac.at/fackel/> [08.11.2014].

[AAC] = Austrian Academy Corpus (o.J. c): „Der Sozialdemokrat", in: *Austrian Academy Corpus* <http://www.aac.ac.at/apps_digied_sozial.html> [08.11.2014].

AA.VV. (2014): *Book of Abstracts*. Digital Humanities 2014. Lausanne <http://dh2014.org> [08.11.2014].

Brooker, Peter / Thacker, Andrew (2006-2014): *Modernist Magazines Project* <http://www.modernistmagazines.com/about.php> [08.11.2014].

Brown University / The University of Tulsa (o.J): *The Modernist Journals Project* <http://modjourn.org/> [08.11.2014].

Burkhardt, Julia / Concepción Durán, Aramis / Potapenko, Elena / Sierig, Rebecca (2014): „FrItZ: Le corpus de la langue des journaux français et italiens de Leipzig – Développement et possibilité d'application d'un corpus de la presse écrite", 9. Frankoromanistenkongress *Schnittstellen / Interfaces*. Sektion *Les interfaces numériques* (Elisabeth Burr / Christoph Schöch). 24. bis 27. September 2014, Universität Münster.

Burr, Elisabeth (1993): *Verb und Varietät*. Ein Beitrag zur Bestimmung der sprachlichen Variation am Beispiel der italienischen Zeitungssprache (= Romanische Texte und Studien 5). Hildesheim: Olms.

Burr, Elisabeth (1997): *Wiederholte Rede und idiomatische Kompetenz*. Französisch, Italienisch, Spanisch. Habilitationsschrift, Gerhard-Mercator-Universität GH Duisburg, Fachbereich 3: Sprach- und Literaturwissenschaften (Manuskript 429 Seiten).

Burr, Elisabeth (1997-2004): *Korpus Romanischer Zeitungssprachen*. Duisburg / Bremen / Leipzig <http://www.uni-leipzig.de/~burr/CorpusLing/> [09.11.2014].

Ermolaev, Natalia / Wulfman, Clifford E. / Biber, Hanno / Crombez, Thomas (2014): „Remediating 20th-Century Magazines of the Arts: Approaches, Methods, Possibilities", in: AA.VV.: Book of Abstracts. Digital Humanities 2014. Lausanne: 52-55 <http://dh2014.org> [08.11.2014].

Davies, Mark (2007-): *TIME Magazine Corpus: 100 million words, 1920s-2000s* <http://corpus.byu.edu/time/> [08.11.2014].

[EurNews] = Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (2014): *Europeana Newspapers* <http://www.europeana-newspapers.eu/> [08.11.2014].

Fischer, Jens / Kirsten, Jens / Witt, Andreas (2014): „Aufbau eines Korpus zur Beobachtung des Schreibgebrauchs im Deutschen", in: *DHd 2014*. Universität Passau <https://www.conftool.pro/dhd2014/index.php/Fischer-Aufbau

_eines_Korpus_zur_Beobachtung_des_Schreibgebrauchs-2711152.pdf> [08.11.2014].

Institut für Deutsche Sprache (o. J.): *Das deutsche Referenzkorpus – DeReKo* <http://www1.ids-mannheim.de/kl/projekte/korpora/> [10.11.2014].

Rossini Favretti, Rema / Grandi, Nicola / Nissim, Malvina / Tamburini, Fabio / Gagliardi, Gloris (1998-): *Coris / Codis*. Università di Bologna <http://corpora.dslo.unibo.it/coris_ita.html> [08.11.2014].

Sierig, Rebecca (2013): *Die Erstellung eines Zeitungskorpus*. Sampling und Markup. Leipzig: unveröffentlichte Masterarbeit.

SSLMIT (2004): *La Repubblica Corpus*. Università di Bologna <http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica> [08.11.2014].

TEI Consortium (16.09.2014): *TEI P5: Guidelines for Electronic Text Encoding and Interchange* 2.7.0. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> [10.11.2014].

Zu den für die Erstellung des Textmodells relevanten Quellen vgl. Burr (1993 u.1997) sowie Sierig (2013).

1486 Wörter (ohne Bibliographie und Autor_innen)

Bearbeitung großer digitaler Korpora mit Topic Modelling

Bei „Welt der Kinder“ handelt es sich um den geschichtswissenschaftlichen Versuch, große Bestände digitaler Korpora für die historische Arbeit nutzbar und zugänglich zu machen. Durch die Stärkung einer engen Zusammenarbeit zwischen Historikern, Informationswissenschaftlern und Informatikern zielt es darauf, neue Erkenntnisse über die Periode zwischen 1850 bis 1918 zu gewinnen: Eine Zeit beschleunigter Wissensproduktion, die geprägt war von gleichzeitigen Prozessen der Globalisierung und der Nationalisierung.

Die Forschung will Zugang zu deutschsprachigen Massenquellen aus der Zeit zwischen 1850 und 1918 ermöglichen. Dieses Material spiegelt auf der einen Seite zeitgenössische Interpretationsmuster der Welt sowie Elemente eines kulturellen Gedächtnisses wieder, formte sie aber gleichzeitig auf der anderen Seite. Aber schon aufgrund ihrer reinen Menge können diese Quellen nicht mit klassisch heuristischen Methoden bearbeitet werden. Daher werden in interdisziplinärer und explorativer Arbeit digitale Werkzeuge entworfen, welche eine Analyse großer (digitaler) Korpora ermöglichen. Dieser Entwicklungsprozess implementierte User-zentrierte Methoden, um die Forschungsfragen der Historiker zu unterstützen. Die so bereitgestellten Werkzeuge helfen, semantische Strukturen und Muster in einer Vielzahl von Bildungsmedien des 19. Jahrhunderts zu erkennen. Dies ermöglicht den Historikern einen innovativen Ansatz zur Analyse digitalen Quellenmaterials umzusetzen; vorher musste sich gewissermaßen auf Volltextsuche beschränkt werden. Als Grundstock dazu dienen annähernd 3500 Schulbücher aus dem Bestand des Georg-Eckert-Instituts für Internationale Schulbuchforschung in Braunschweig mit einem Erscheinungsdatum vor 1919. Schulbücher wurden gewählt, da diese zum einen den quasi-offiziellen Diskurs des Kaiserreiches widerspiegeln und es sich zum anderen um eine weitverbreitete und viel rezipierte Quellengattung handelt. Allerdings geraten beim Umfang des Quellenmaterials (momentan über 600.000 Seiten) die klassischen historischen Herangehensweisen schnell an ihre Grenzen. Für die Zukunft des Projektes ist diese Methodenfrage umso bedeutsamer, da der digitale Quellenbestand noch ausgebaut werden wird. Hier wendet nun das Projekt für die Geschichtswissenschaft neue digitale Methoden an, um die Menge an Quellenmaterial bewältigen zu können.

Drei Projektziele spielen für die hiesige Präsentation eine wesentliche Rolle:

1.) Historische Forschung über Repräsentationen und Interpretationen der Welt in der oben stehenden Periode, in der Wissen über die Welt normalerweise nicht durch eigene Anschauung wie Reisen oder durch audiovisuelle Medien gesammelt werden konnte. Daher sind Schulbücher und andere gedruckte Medien die Hauptinformationsquelle für junge Erwachsene.
2.) Die Erforschung eines spezifischen Quellentypus (Schulbücher), die Millionen von späteren Bürgern prägten, die aber bisher noch nicht mit einem Ansatz, der Medientyp, Zirkulation und Wissenstransformation zusammenbringt, untersucht wurden.
3.) Grundlagenforschung in Computerlinguistik; die Entwicklung und Adaptierung verschiedener Methoden der semantischen Analyse und Opinion Mining, die an die Sprache des 19. Jahrhunderts und diesen spezifischen Quellentyp angepasst werden. Wir werden die Daten und Methoden, wie sie bisher im Projekt genutzt wurden, präsentieren sowie Herausforderungen, Erfahrungen und Probleme, die sich im Vorlauf der bisherigen Arbeit ergaben, vorstellen.

Mapping the Words

Übersetzungsstrukturen zwischen
Altgriechisch und Hocharabisch
in visuellen Formen

Abstract des Vortrags auf der DHd 2015 in Graz
von Torsten Roeder (Würzburg) und Yury Arzhanov (Bochum)

Einleitung

Das Feld der Linguistik ist in den Digital Humanities seit dessen Anfängen von großer Bedeutung. Ob die abstrakten Strukturen der Informationstechnik den Sprachstrukturen besonders leicht nahe kamen, ob die Datenmengen die Verwendung von Computern interessant machten, oder ob hier noch ganz andere Faktoren im Spiel waren, darf eine offene Frage bleiben. Als ein Ertrag der Computerlinguistik stellte sich in jedem Falle heraus, dass das Sammeln von Daten nicht nur mit dem Ziel, ein elektronisches Nachschlagewerk zu erhalten, verfolgt werden könnte, sondern dass durch systematische Weiterverarbeitung und Darstellung der Daten bestimmte, meist quantitative, aber auch strukturorientierte Fragen gestellt werden könnten, deren Ergebnisse dem Forscher Hinweise auf bislang nicht erkannte Phänomene geben könnten.[1]

Während monolinguale Lexika sehr zahlreich und divers aufgestellt sind, füllen die bilingualen Lexika derzeit noch eine Lücke. Dabei konnten Übersetzungsbewegungen stets Impulse für Sprachentwicklung liefern und sind insofern für Sprachgeschichten und infolge dessen auch für aktuelle Sprachstrukturen möglicherweise von großer Bedeutung. Das Beispiel dafür sollen in diesem Paper die arabischen Übersetzungen altgriechischer Texte sein, welche im Raum von Bagdad während einer bedeutenden Blütezeit arabischer Sprache und Kultur im 9.–11. Jh. n. Chr. angefertigt wurden. Die Erforschung dieser Periode wurde im Projekt „Glossarium Graeco-Arabicum“ unternommen, welches zu diesem Zweck unter anderem eine Datensammlung von Übersetzungen auf Wort- und Kontextebene angefertigt hat.[2]

1 Vgl. dazu den ausführlichen Band von Oakes/Ji 2012.

2 Glossarium Graeco-Arabicum, European Research Council / Ruhr-Universität Bochum / Berlin-Brandenburgische Akademie der Wissenschaften, <http://telota.bbaw.de/glossga> (08.11.2014); aktuelle Entwicklungsversion unter <https://telotadev.bbaw.de/glossga> (08.11.2014). Siehe dazu auch Endress/Arnzen/Arzhanov 2013; Arzhanov/Roeder 2013.

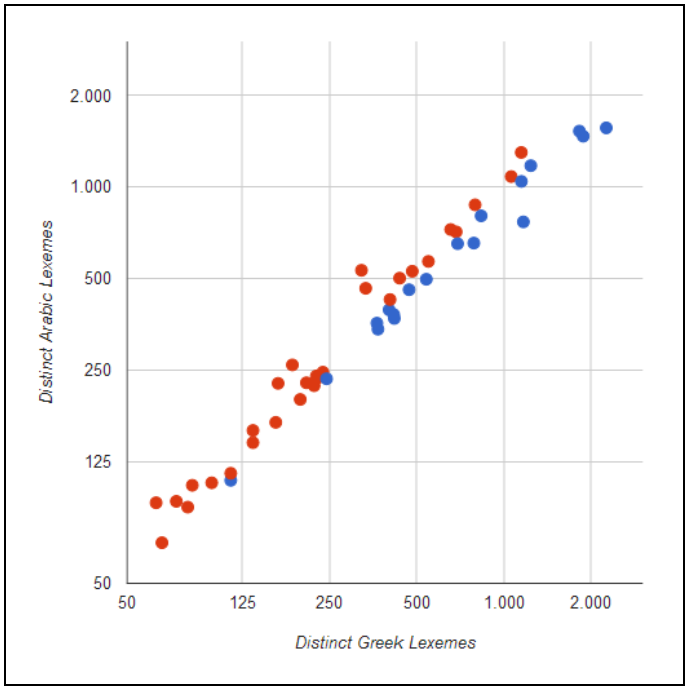
Diese Datenbank bestand anfangs noch aus handgeschriebenen Karteikarten, die später digitalisiert und Stück für Stück in eine Datenbank übertragen wurden.[3] Die Datensammlung hat sich seitdem von einer Datenverwaltung zu einem digitalen Lexikon weiterentwickelt und ist auf dem Wege, zu einem komplexen Forschungsinstrument weitergestaltet zu werden.[4] Mittlerweile enthält die Datenbank über 100.000 Wortpaare, die aus über 80 unterschiedlichen Quellen stammen (die nebenstehende Abbildung zeigt eine „Treemap“ des Bestandes, die die proportionalen Anteile der jeweiligen Quellen am Gesamtkorpus darstellt). Quantitative und visuelle Auswertungsverfahren werden somit nicht nur immer naheliegender, sondern auch immer notwendiger, um einerseits Fragen an das Material zu stellen und andererseits über die schiere Fülle von Informationen die Übersicht zu behalten. Das vorliegende Paper möchte diese Entwicklungen beobachtend nachzeichnen und zeigen, welche Potenziale in den Daten und Datenstrukturen stecken, Impulse für die Erforschung von Übersetzungsprozessen zu generieren.

„Mapping the Words“?

Da sich der Gegenstand räumlich auf die Gegend von Bagdad beschränkt, zielt der Titel nicht auf die geographische Dimension der Sprache, sondern vielmehr im allgemeinen Sinne für visuelle Abbildungen von Übersetzungsstrukturen. Zudem sind es weniger einzelne Wörter, sondern Paare aus griechischen und arabischen Wörtern, die das Grundelement dieser Visualisierungen bilden sollen. Entscheidend ist die Gegenüberstellung von Ausgangs- und Zielstrukturen, und dies möglichst in einer Art und Weise, die es dem Betrachter erlaubt, die Strukturen ähnlich wie auf einer Landkarte erkennen und lesen zu können.[5] Diese „Karten“ bilden dann im günstigsten Falle eine Referenz für die Sprachforschung. Wie dies aussehen könnte, soll anhand von drei Beispielen vorgeführt werden.

3 Siehe Arnzen/Arzhanov/Endress 2012.
4 Vortrag von Yury Arzhanov / Gerhard Endreß / Torsten Roeder; Internationaler Workshop „Plotinus East and West. The Enneads in Arabic and Latin“, Pisa, 3.–6. November 2014.
5 Dies wurde bereits für ein historisches Sprachkorpus des Englischen demonstriert; siehe dazu Alexander 2010.

(1) Differenziertheit der Sprache



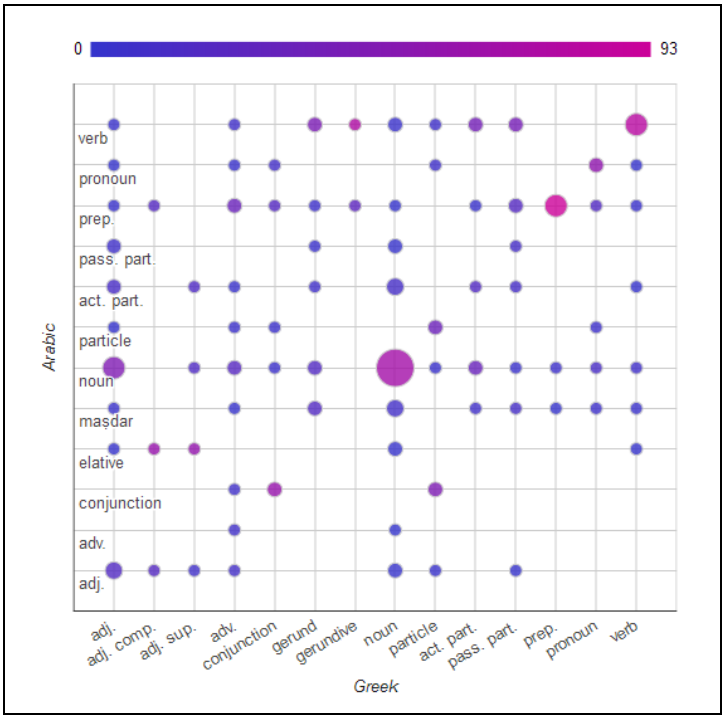
Vergleicht man Texte mit ihren Übersetzungen, stellt man möglicherweise fest, dass in manche Fällen eine Tendenz zur Ausdifferenzierung des Vokabulars besteht, während in anderen Fällen eine verallgemeinernde Sprache gewählt wird. In welchem Maße liegt eine ganz allgemeine Tendenz vor, wenn vom Griechischen ins Arabische übersetzt wird, und in welchem Maße ist dies abhängig von Übersetzer und Ursprungstext? Diese Frage kann durch einen Mengenvergleich von distinkten griechischen und arabischen Lexemen in den jeweiligen Quellen beantwortet werden. Die nebenstehende Grafik vergleicht die Quellen auf einer logarithmischen Skala, wobei die Anzahl der griechischen Lexeme auf der Abszissenachse und die der arabischen Lexeme auf der Ordinatenachse festgehalten sind.

Zunächst wird an der Grafik deutlich, dass es keine klare Tendenz gibt, welche der beiden Sprachen differenzierter ausgeschöpft wurde. Im unteren Bereich überwiegt im Arabischen die Ausdifferenzierung, wobei die Auswahl möglicherweise nicht repräsentativ ist, da hier erst wenige Lexeme erfasst wurde. Im oberen, repräsentativeren Bereich besteht hingegen eine leichte Tendenz zur Verallgemeinerung des Vokabulars, so insbesondere an den Quellen *Oneirocritica* (Artemidorus Daldianus), *De generatione animalium* (Aristoteles) und *Placita Philosophorum* (Pseudo-Plutarchus). Im mittleren Bereich treten die zwei Texte von Pseudo-Aristoteles *De virtutibus et vitiis* sowie *Divisiones quae vulgo dicuntur Aristoteleae* durch eine hohe Wortvielfalt im Arabischen hervor. Diese Texte, in deren Übersetzungen sich die Vielfalt des Vokabulars deutlich veränderte, würden eine nähere Betrachtung lohnen.

(2) Wortarten

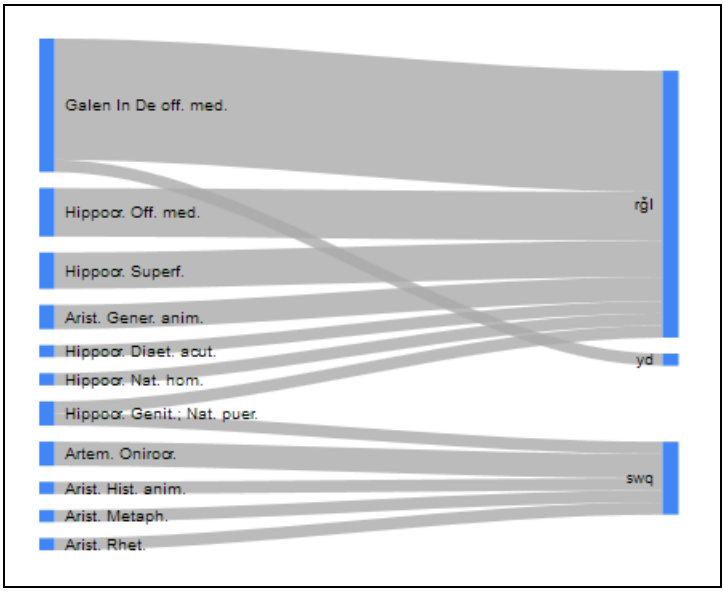
Das zweite Beispiel geht näher auf die unterschiedlichen Sprachstrukturen ein. Die Grammatiken des Griechischen und des Arabischen sind voneinander so verschieden, dass bei Übersetzungen manchmal Wortarten zwingend verändert werden müssen (sofern nicht bereits die Freiheit des Übersetzers einen Einfluss ausübt). So besitzt das Arabische keine Entsprechung des griechischen Gerundivs, weshalb dafür in der arabischen Übersetzung eine andere Wortart gefunden werden

muss. Die dahingehende Kreativität des Übersetzungsprozesses kann visualisiert werden, ohne dass die Expertise in einer oder beiden Sprachen notwendig ist. Die nebenstehende Abbildung zeigt, welche griechischen Wortarten (Abszissenachse) mit welcher Häufigkeit in eine arabische Wortart (Ordinatenachse) übertragen wurden. Die Größe des Schnittpunktes repräsentiert die absolute Häufigkeit, während die Farbe die Häufigkeit in Relation zur jeweiligen Wortart widerspiegelt. Erkennbar ist, dass Gerundive als Verb-Präpositions-Kombination umgesetzt wurden; außerdem ist ablesbar, dass Adjektive häufiger umgeformt als beibehalten wurden, ganz anders als z. B. bei Nomen, die in den meisten Fällen erhalten blieben. Über die Auswahl des Schnittpunktes gelangt man zu einer Liste der einzelnen Wortpaare, die man nun genauer unter die Lupe nehmen kann.



(3) Übersetzungsvariaten

Ist es auch möglich, Visualisierungen zu entwerfen, die über die quantitative Ebene hinausgehen und mehr über einzelne Wörter verraten? Schließlich ist auch bei Erhalt der Wortart längst nicht zu erwarten, dass ein Wort stets mit dem gleichen Begriff übersetzt wird. Die Ursache dafür kann wiederum von Übersetzern, aber auch schon von den Ursprungstexten selbst abhängen. Mit dieser Frage beschäftigt sich das dritte Beispiel. Die Relationen zwischen Quelltexten und der Wahl eines Wortes lassen sich beispielsweise durch ein Sankey-Diagramm[6] sichtbar machen. Am Beispiel des griechischen Wortes



6 Benannt nach dem irischen Ingenieur Matthew Henry Phineas Riall Sankey (1853–1925), der eine graphische Darstellung von simultanen Mengenflüssen erfand, in denen Proportionen und Flussrichtung gleichzeitig sichtbar werden.

σκέλος (Bein) zeigt die Grafik, dass dieses Wort in Texten von Hippokrates und Galen in fast allen Fällen erwartungsgemäß mit ساق (Bein) übersetzt wurde, bei Texten von Aristoteles aber fast durchgehend رجل (Mann) bevorzugt wurde, und in einem weiteren, anscheinend besonderen Falle, يد (Hand) verwendet wurde. Welche Ursachen dieser Korrelation zwischen Autor und Übersetzung zugrunde liegen, ist nun durch eine Betrachtung der Einzelfälle zu untersuchen.

Zusammenfassung und Ausblick

Die Beispiele zeigen auf verschiedenen Ebenen, wie Übersetzungsstrukturen visualisiert werden können. Je nach Fokussierung auf Quellen, Grammatik oder Vokabular können unterschiedliche Formen zum Einsatz gebracht werden, mit denen sich die jeweiligen Themen angehen lassen. Die Darstellungsarten sind im Prinzip grenzenlos und hängen vor allem davon ab, welcher Aspekt in den Vordergrund gestellt werden soll. Zudem werden die Strukturen auch für diejenigen sichtbar, die keine Kenntnis des Griechischen und/oder Arabischen besitzen. Hier kann eine grafische Form deutlich mehr Aufschluss geben als der Volltext der lexikographischen Einträge. Auch die sprachkundigen Fachleute können die Visualisierungen nutzen, um den Überblick über ein großes Korpus zu erhalten.

Jedoch wird ebenso deutlich, dass Visualisierungen Sachverhalte zwar aufzeigen, nicht aber erklären können. Hier ist in der Tat die philologische Prüfung der Einzelfälle gefragt. Für die Zukunft besteht zudem noch reichlich Bedarf an komplexeren Formen der Visualisierung, die z. B. auch Wortfelder oder syntaktische Strukturen miteinbeziehen.[7]

7 Vgl. z. B. Shneiderman/Plaisant 2009; Leydesdorff/Welbers 2011.

Literatur

<Alexander 2010> Marc Alexander: The Various Forms of Civilization Arranged in Chronological Strata. Manipulating the Historical Thesaurus of the OED. In: Cunning passages, contrived corridors. Unexpected Essays in the History of Lexicography, hrsg. von M. Adams, Monza: Polimetrica, 2010.

<Arzhanov/Roeder 2013> The Glossarium GraecoArabicum. Linguistic Research and Database Design in Polyalphabetic Environments, Vortrag im Digital Classicists Berlin Seminar, 19. November 2013, <http://hdl.handle.net/11858/00-1780-0000-0022-D548-B> (Permalink).

<Arnzen/Arzhanov/Endress 2012> Rüdiger Arnzen; Yury Arzhanov; Gerhard Endress: Griechische Wissenschaft in arabischer Sprache. In: RUBIN Wissenschaftsmagazin, Frühjahr 2012, S. 14–21, <http://rubin.rub.de/de/griechische-wissenschaft-arabischer-sprache> (10.11.2014).

<Endress/Arnzen/Arzhanov 2013> Gerhard Endress; Rüdiger Arnzen; Yury Arzhanov: Griechische Wissenschaft in arabischer Sprache. Ein griechisch-arabisches Fachwörterbuch der internationalen Wissensgesellschaft im klassischen Islam. In: Studio graeco-arabica 3 (2013), S. 143–156.

<Leydesdorff/Welbers 2011> Loet Leydesdorff; Kasper Welbers: The semantic mapping of words and co-words in contexts. In: Journal of Informetrics, Juli 2011 (Volume 5, Issue 3), S. 469–475.

<Oakes/Ji 2012> Michael P. Oakes; Meng Ji (Hg.): Quantitative Methods in Corpus-Based Translation Studies. A practical guide to descriptive translation research (= Studies in Corpus Linguistics 51), Amsterdam/Philadelphia: John Benjamins, 2012.

<Shneiderman/Plaisant 2009> Ben Shneiderman; Catherine Plaisant: Treemaps for space-constrained visualization of hierarchies, 2009–2014, <http://www.cs.umd.edu/hcil/treemap-history/index.shtml> (10.11.2014).

Das artifizielle Manuskriptkorpus TASCFE

Armin Hoenen

13. Januar 2015

1 Abstrakt

In diesem Paper soll das *Teheran Artificial Shahname Corpus with Frankfurt Extension (TASCFE)* digitalisierter handschriftlicher Texte zur Evaluation automatisch generierter Stemmata vorgestellt werden. Ein *Stemma codicum* oder kurz Stemma ist eine Visualisierung der genealogischen Zusammenhänge innerhalb eines Manuskriptkorpus oder einfacher gesagt ein Manuskriptstammbaum. Die Generierung solcher Stammbäume verfolgt generell zwei Hauptziele: ein genaueres Verständnis der Überlieferungsgeschichte und die Rekonstruktion eines Urtextes. Bis in die 90er Jahre hinein wurden Stemmata vornehmlich manuell erstellt, sind aber seitdem zunehmend auch automatisch generiert und analysiert worden, siehe u.a. Spencer et al. (2004), Roos and Heikkilä (2009) und Roelli and Bachmann (2010).[1] Technologisch ist die bio-informatische Phylogenie Donordisziplin, wie die Nutzung phylogenetischer Programme und Algorithmen zur automatischen Manuskriptstammbaumerstellung zeigt. Dabei fehlt es in der biologischen Phylogenie an Möglichkeiten, erzeugte Stammbäume zu evaluieren, da die Aufspaltungsvorgänge der Spezies, die durch die Verzweigungen symbolisiert werden nicht beobachtet und aufgezeichnet werden konnten, lagen sie doch z.T. Millionen von Jahren in der Vergangenheit. Im Gegensatz dazu ist es in der Stemmatologie durchaus möglich, sowohl die Vorlage als auch die Kopie im Korpus vorzufinden. Mehr noch, es ist möglich neue Korpora zu erzeugen und gleichzeitig die Kopiergeschichte der Manuskripte aufzuzeichnen. Diese Daten können dann in einem klassisch informatischen Evaluationsszenario der Beurteilung von stem-

[1] Für eine detaillierte Darstellung der historischen Entwicklung der Stemmatologie siehe O'Hara (1996), Robinson and O'Hara (1996), van Reenen et al. (1996) und van Reenen et al. (2004).

Text	Sprache	Anzahl Manuskripte	Anzahl Worte	Publikation
Parzival	Englisch	21	957	Spencer et al. (2004)
Notre Besoin	Französisch	13	1029	Ph.V. Baret (2004)
Heinrichi	Altfinnisch	64	1208	Roos and Heikkilä (2009)
Shahname	Persisch	50	107	Hoenen (hic ipsum)

Abbildung 1: Die artifiziellen Traditionen

magenerierenden Methoden genutzt werden. Artifizielle Korpora wurden bisher drei Mal erzeugt, siehe Ph.V. Baret (2004), Spencer et al. (2004) und Roos and Heikkilä (2009). Nur das letztgenannte Paper evaluierte mehrere stemmagenerierende Algorithmen, darunter auch die händische Rekonstruktion, mittels einer Distanzfunktion zwischen dem echten Stemma und den erzeugten. Diese Distanz nannten die Autoren Average Sign Distance (ASD). Sie misst die Ähnlichkeit der Topologien des korrekten und des erzeugten Stammbaums anhand der Ähnlichkeit der inneren Abstände aller Knotentripel (von vorhandenen Manuskripttexten) im erzeugten mit deren *shortest-path* Abständen im echten Stammbaum. Abbildung 1 fasst Kennwerte der drei artifiziellen Korpora zusammen.

Alle bisher bekannten artifiziellen Traditionen sind im lateinischen Alphabet verfasst. Hier wird das TASCFE Korpus vorgestellt, welches in persischer Sprache (Farsi) im arabischen Alphabet vorliegt. Neben der Sprache besteht seine Besonderheit für eine Bereicherung der Landschaft artifizieller Korpora darin, orale Variation zu approximieren. Orale Variation ist solche Variation, die nicht aufgrund von Fehlern im Kopierprozess, sondern aufgrund der Dynamik mündlicher Überlieferung entstanden ist und die zum Teil stark von erstgenannter Variation abweicht. Die Oral Formulaic Theory (OFT) wurde in den 30er bis 60er Jahren des vorigen Jahrhunderts durch Parry and Parry (1987) und Lord (1960) im Zusammenhang mit der *Homerischen Frage* erarbeitet. Ergebnis dieser Theorie war u.a. die Erkenntnis, dass Texte wie die Odyssee keinen Urtext, d.h. keine Originalversion besitzen. In der Zeit vor Erfindung der Schrift wurden Texte ausschließlich oral tradiert. Dabei war zur konkreten Textmanifestation ein Aufführender und (mindestens ein) Zuhörer notwendig. Da die Umstände jeder Aufführung jedoch unterschiedlich waren, war es so auch der Text selbst. Z.B. nutzte ein Barde bei derselben Geschichte viele Ausschmückungen (z.B. Adjektive), wenn er viel Zeit hatte, erzählte sie jedoch ein anderes Mal, wo die Zeit drängte, ohne Ausschmückungen. Dazu kommen Fehler des menschlichen Erinnerungsapparates, der andersartige Variationen erzeugt, als solche, die beispiels-

weise durch Buchstabenverwechslung beim Abschreiben zu Stande kommen. Zu Beginn der Schrifteinführung wurden Texte via "Pseudo-Aufführungen" vor einem Schreiber (Diktate) erstmals in schriftliche Form überführt. Da derselbe Text mehrfach in solchen Diktaten aufgezeichnet worden sein kann, da weiterhin jede Aufführung ganz wie in der rein oralen Welt dieselbe Geschichte in unterschiedlicher Textform (mehr Ausschmückungen/weniger Ausschmückungen u.a. Arten oraler Varianz) hervorbrachte, können am Beginn mancher (meist der frühesten) Manuskripttraditionen Varianten stehen, die sich nicht mit den Arten an Variation aus rein literarisch überlieferten (d.h. als schriftliche Texte entstandenen) Texten decken. Solch eine Variation ist für die Stemmagenerierung wichtig, da sie determiniert, ob ein einziges oder mehrere Stemmata und ob mehrere oder nur ein Urtext angenommen werden müssen. Das TASCFE Korpus trägt dieser Art der Variation zuminndest teilweise Rechnung, da an seinem Anfang vier verscheidene Versionen stehen. Neben der Sprache ist dies die zweite stemmatologierelevante Besonderheit des TASCFE.

Der Text ist ein Auszug (Strophe) aus dem persischen Nationalepos Shahname (*Buch der Könige*), (l Qasim Ferdoussi, 1967, p.55). Es entstand um das Jahr 1000. Die Autorenschaft wird generell *Abu l-Qasim Ferdoussi* zugerechnet, wobei orale Einflüsse im Werk bereits seit längerem diskutiert werden, siehe u.a. Yamamoto (2003) und Rubanovich (2011) . Die ca. 6.500 Token des Korpus wurden 2014 in Teheran (43 Manuskripte) und in Frankfurt (7 Manuskripte) von Freiwilligen entweder von einer gedruckten oder einer handschriftlichen Vorlage durch Abschreiben produziert. Anschließend wurde das Korpus digitalisiert und aligniert. Ein des Persischen nicht mächtiger Freiwilliger kopierte zusätzlich eines der handgeschriebenen Manuskripte, um der These nachzugehen, dass in historischer Zeit Analphabeten oder Schreiber anderer Schriften Manuskripte kopiert haben könnten, was sich aber deshalb als unwahrscheinlich erwies, da es einer persischen Muttersprachlerin aufgrund der partiellen Unlesbarkeit nicht möglich war von dem so kopierten Manuskript eine weitere Kopie anzufertigen.

Kein einziges Manuskript entsprach genau der Vorlage. Eine qualitative Analyse der Phänomene, die denen historischer Korpora ähnelten (so z.B. Zeilensprünge oder Wortsprünge, aber auch synonymische Ersetzungen) konnte zeigen, dass aufgrund des Schriftsystems und der daraus teils zur lateinischen Schrift unterschiedlichen Fehler eine andere Differenzierung in Fehlerklassen je nach

Schriftsytsem notwendig sein kann. Dies knüpft an Andrews and Macé (2013) an, die zeigen konnten, dass Variationsklassen je nach Sprache variieren können. Das digitale Zeitalter eröffnet dahingehend die Möglichkeiten einer schriftsystemübergreifenden Analyse von durch Abschreibefehler verursachter Variation, die dann zur Abstraktion der dort wirkenden universalen Prinzipien beitragen wird, siehe Abbildung 2. Dies setzt die Schaffung geeigneter Ressourcen voraus.

Auf die Daten wurden im Weiteren stemmatologische Algorithmen angewandt (die dann mittels der oben angesprochenen ASD evaluiert wurden). Hierbei konnte gezeigt werden, dass ein Ansatz zur Feststellung von Oralität durch Gruppenbildung im Stemma besteht, wobei die Levenshtein Distanz, Levenshtein (1965), bei hoher Gewichtung von Lücken im Alignment eine besonders geeignete und gleichzeitig leicht zugängliche algorithmische Basis darstellt, siehe Abbildung 3. Dabei wurde ein wortpaar-basierter Vergleich aller Manuskriptpaare durchgeführt. Die Levenshtein Distanz aller Wortpaare des jeweiligen Manuskriptpaares wurde (auch als Baseline für den Vergleich mit weiteren Algorithmen) zu einer Manuskriptpaargesamtdistanz aufsummiert.

Die Matrix der Manuskriptpaardistanzen wurde mittels des Neighbor Joining Algorithmus, Saitou and Nei (1987), wiederum eine geeignete Baseline, aus dem 'ape' Packet (Paradis et al. (2004), Paradis (2012)) der Programmiersprache R in einen Stammbaum überführt, der dann visualisiert und evaluiert wurde, siehe Abbildung 4.[2]

[2]www.r-project.org/

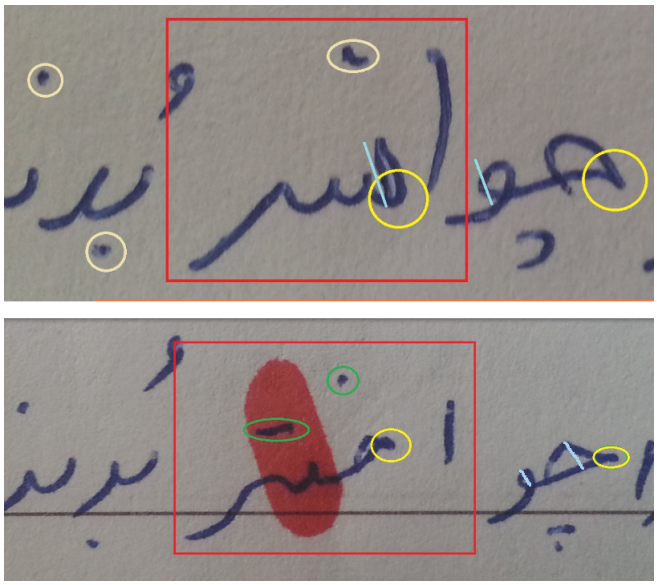


Abbildung 2: Die durch ungewöhnliche Buchstabenform ausgelöste Fehlkopie von افسر (oben) nach اختر (unten) in roten Rechtecken. Die Kennstellen, die den Abschreibefehler ausgelöst haben, sind farblich markiert. Das ف in افسر ist nicht so rund wie und länger als erwartet (blau). Zudem ist der einzelne Punkt auf dem ف versehentlich breiter (hautfarben). Dennoch ist im oberen Rechteck eindeutig nur ein Punktmuster erkennbar, unten jedoch zwei (grün). Des Weiteren hat das ـسـ zwei deutliche Hacken, wobei der mit rotem Stift unterlegte untere entsprechende Buchstabe nur einen aufweist. Außerdem ist das ف in der unteren rechten Ecke rund, was auf das vorausgehende چ jedoch nicht zutrifft (gelb). Obgleich der Abschreibefehler höchstwahrscheinlich durch die ungewöhnliche Form des ف ausgelöst wurde, passte die Ersetzung gut in den Kontext, vielleicht sogar besser als das Original. Genau diese Interaktion von kontextuellem Priming und ungewöhnlichen Buchstabenformen ist ein idealer Kandidat für schriftsystemübergreifende Prozesse beim Abschreiben.

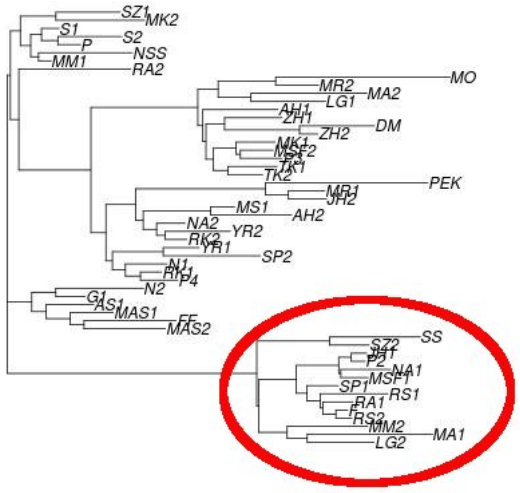


Abbildung 3: Automatische Erkennung einer durch orale Variation gekennzeichneten Gruppe.

Version	**ASD**
Shahname(V1)	55, 59
Shahname(V2)	55, 13
Shahname(V3)	57, 93
Shahname(V4)	55, 83
Shahname(Durchschnitt V1-V4)	56, 12
Shahname(als eine Tradition)	38, 31

Abbildung 4: Evaluation der erzeugten Stemmata (ASD). Das Stemma der Gesamttradition unter der Annahme nur einer Wurzel evaluiert mit diesen Algorithmen deutlich schlechter als der Durchschnitt der einzelnen Versionen.

Literatur

Andrews, T. L. and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmas. *Literary and Linguistic Computing*, 28(4):504–521.

l Qasim Ferdoussi, A. (1966-1967). *The Shahname - the book of kings.* The Great Islamic Encyclopaedia.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. english in: Soviet Physics Doklady 10 (8) (1966) 707–710.

Lord, A. B. (1960). *The Singer of Tales.* Harvard University Press.

O'Hara, R. J. (1996). Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano*, 27(1):81–88.

Paradis, E. (2012). *Analysis of Phylogenetics and Evolution with R.* Springer, New York, 2nd edition.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20:289–290.

Parry, M. and Parry, A. (1987). *The Making of Homeric Verse: The Collected Papers of Milman Parry.* Oxford University Press.

Ph.V. Baret, C.Macé, P. (2004). Testing methods on an artificially created textual tradition. In *Linguistica Computationale XXIV-XXV*, volume XXIV-XXV, pages 255–281, Pisa-Roma. Instituti Editoriali e Poligrafici Internationali.

Robinson, P. M. and O'Hara, R. J. (1996). Cladistic analysis of an old norse manuscript tradition. *Research in Humanities Computing (4).*

Roelli, P. and Bachmann, D. (2010). Towards generating a stemma of complicated manuscript traditions: Petrus alfonsi's dialogus. *Revue d'histoire des textes*, 5(4):307–321.

Roos, T. and Heikkilä, T. (2009). Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24:417–433.

Rubanovich, J. (2011). *Medieval Oral Literature*, chapter Orality in Medieval Persian Literature, pages 653–680. De Gruyter.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Spencer, M., Davidson, E. A., Barbrook, A., and Howe, C. J. (2004). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 227:503–511.

van Reenen, P., den Hollander, A., and van Mulken, M. (2004). *Studies in Stemmatology II*. Studies in Stemmatology. John Benjamins Publishing Company.

van Reenen, P., van Mulken, M., and Dyk, J. (1996). *Studies in Stemmatology I*. Studies in Stemmatology. John Benjamins Publishing Company.

Yamamoto, K. (2003). *The Oral Background of Persian Epics*. Brill, Leiden.

Digitale Analyse Graphischer Literatur

Alexander Dunst[1], Rita Hartel[2], Sven Hohenstein[3], Jochen Laubrock[3]
[1]Institut für Anglistik und Amerikanistik, Universität Paderborn
[2]Institut für Informatik, Universität Paderborn
[3] Department Psychologie, Universität Potsdam

Zusammenfassung

Der hier vorgeschlagene Vortrag stellt erste Ergebnisse der vom deutschen Bundesministerium für Bildung und Forschung (BMBF) finanzierten Nachwuchsgruppe „Hybride Narrativität: Digitale und Kognitive Methoden zur Erforschung Graphischer Literatur“ vor. Der erste Teil des Vortrages wird das Projekt der Nachwuchsgruppe und zentrale Forschungsfragen vorstellen. Im zweiten Teil wird eine kurze Einführung in die wichtigsten Merkmale der Beschreibungssprache „Graphic Narrative Markup Language“ (GNML) gegeben sowie die wesentlichen Funktionen des Editors zur Erfassung und Analyse graphischer Erzählungen demonstriert. Außerdem präsentieren wir erste Ergebnisse, die mit Hilfe von Netzwerkgraphen und Beziehungsmatrizen zu Paul Austers *City of Glass* (1985) und dessen graphischer Adaptation durch David Mazzuchelli und Paul Karasik (1994) strukturelle Ähnlichkeiten und Differenzen zwischen den narrativen Systemen des literarischen und graphischen Romans darstellen. Der abschließende dritte Teil stellt, wiederum am Beispiel von *City of Glass*, eines der interdisziplinären Anwendungsgebiete der digitalen Annotation graphischer Literatur vor: anhand von Blickbewegungsmaßen können Rückschlüsse auf das Leseverständnis graphischer Literatur gewonnen werden.

1. Forschungsfragen

Die digitale Annotation und Analyse literarischer Text-Korpora kann in den vergangenen Jahren auf enorme Fortschritte verweisen und hat neue Erkenntnisprozesse etabliert, die mittlerweile Eingang in die Forschungsbestrebungen einer breiteren Literaturwissenschaft und Literaturgeschichte finden (1). Im Gegensatz dazu steckt die Analyse visueller Kultur erst in den Anfängen und stellt in den Digitalen Geisteswissenschaften aus mehreren Gründen oft eine Randerscheinung dar: zu der institutionellen Verortung in traditionell text-fokussierten Disziplinen und den Forschungsinteressen der Computerphilologie gesellen sich urheberrechtliche Fragen, sowie der vergleichsweise hohe technische Aufwand und niedrigere Entwicklungsstand von Methoden der Bildanalyse.

Ziel dieses interdisziplinären Projektes ist die empirische Erforschung graphischer Literatur, insbesondere des Genres des graphischen Romans („graphic novel“). Durch die Entwicklung von empirischen Methoden für graphische Literatur sollen Ansätze aus dem „Distant Reading“ für multimediale Kulturformen erschlossen werden. Die durch die Nachwuchsgruppe entwickelten Annotations-Werkzeuge, insbesondere die XML-Sprache GNML und der GNML-Web-Editor (siehe 2.), sind in weiterer Folge nicht nur für die Beschäftigung mit graphischer Literatur sondern auch für die Analyse von Handschriften, Film und Fernsehen von Interesse. In erster Linie zielt die Nachwuchsgruppe jedoch darauf ab, grundlegende Fragen zur spezifischen Narrativität und dem formalen Aufbau des graphischen Romans zu beantworten, die im Rahmen qualitativer Methoden nicht empirisch überprüft werden können oder vollständig außerhalb des Forschungsradius hermeneutischer Fragestellungen in den Geisteswissenschaften liegen. Folgende zentrale Forschungsfragen sind hier beispielshaft zu erwähnen: Beschreibt der Terminus graphischer Roman,

ursprünglich ein Begriff aus der Verlagswerbung, tatsächlich strukturelle Ähnlichkeiten mit dem *literarischen* Roman? Welche Charakteristika unterscheiden den graphischen Roman vom Comicbuch? Lassen sich strukturelle Innovationen isolieren und historisch verfolgen, die das Genre erfolgreich haben werden lassen? Sind diese narratologischer oder thematischer Natur, oder handelt es sich um eine Kombination beider? Aus welchen Sub-Genres besteht der graphische Roman, und wie interagieren diese im System des Genres? Welche gesellschaftlich relevanten Fragen werden im graphischen Roman kulturell verarbeitet und tragen so zu seiner Popularität bei?

2. Editor als Erfassungs- und Analysewerkzeug

Die auf der „Comic Book Markup Language (CBML)“ (2) und damit auf der „Text Encoding Initiative“ (TEI) (3) basierende und im Rahmen dieses Projektes entwickelte XML-Sprache GNML erlaubt dem Bearbeiter nicht nur das Erfassen textueller sondern insbesondere auch visueller Aspekte graphischer Erzählungen. Mit Hilfe von GNML können unter anderem Seiten, Panel-Anordnungen, Texte, Sprechblasen, Charaktere und andere Objekte erfasst werden, sowie Interpretationen, z.B. zu Panel-Übergänge und Texttypen, abgelegt werden. GNML bietet somit eine abstrakte Sicht auf visuelle und textuelle Aspekte, die so effizient analysiert werden können. Basierend auf GNML können z.B. Eyetracking-Experimente ausgewertet werden, und Fragestellungen wie „Wie oft wechselt die Aufmerksamkeit des Lesers vom Text zum Bild“ oder „Was ist der relative (visuelle) Anteil eines Charakters an der gesamten Erzählung“ effizient beantwortet werden.

Eine zentrale Rolle dieses Projektes nimmt der GNML-Editor ein. Er erlaubt ein effizientes, benutzerfreundliches Erfassen der visuellen und textuellen Aspekte, ohne dass der Bearbeiter XML oder GNML beherrschen muss. Neben der Erfassung visueller Aspekte, bei der Objekte durch den Bearbeiter nachgezeichnet und annotiert werden können, bietet der Editor auch die automatisierte Erkennung verschiedener Aspekte. Derzeit bietet der Editor z.B. eine automatische Erkennung der Panels. Hierbei werden nicht nur regelmäßige Formen (Rechtecke), sondern nahezu beliebige Umrandungsformen automatisch erkannt. Eine weiterführende automatisierte Erfassung, etwa mit Hilfe von Texterkennungssystemen und Handschriftenerfassung oder das automatische Erkennen von Charakter-Objekten, befinden sich derzeit in Entwicklung. Verfahren aus dem Bereich des maschinellen Lernens sollen dafür sorgen, dass die Erkennung neuer Charaktere mit zunehmendem Training zuverlässiger funktioniert.

Ein zusätzliches Analysetool ermöglicht die Analyse bereits erfasster GNML Dokumente. So kann der Benutzer z.B. sich die Beziehungen der Charaktere untereinander in Form von Netzwerkgraphen und Beziehungsmatrizen anzeigen lassen. Auch erweiterte Statistiken über die Objekte und Charaktere der Erzählung können berechnet und dem Benutzer in Form von Diagrammen und Tabellen angezeigt werden, basierend z.B. auf den folgenden Fragestellungen:

- Wie oft erscheint ein Charakter?
- Was ist der relative visuelle Anteil eines Charakters?
- Mit wem zusammen erscheint der Charakter auf derselben Seite oder in demselben Panel?

3. Empirische Ergebnisse: Eyetracking-Maße für die Aufmerksamkeitszuwendung des Lesers

Die Methode der Blickbewegungsmessung (Eyetracking) liefert Einblicke in die größtenteils unbewusste und in gewissem Maße kulturspezifische Verteilung der Aufmerksamkeit bei der visuellen Rezeption von Informationen. Für die Rezeption von Texten hat die psycholinguistische Forschung hier bereits viele grundlegende Erkenntnisse gewonnen, und die Methode wird auch

erfolgreich bei der Erforschung der kognitiven Verarbeitung von Bildern oder visuellen Szenen angewandt. Sequenzielle Kunst, Comics und graphische Literatur sind jedoch bisher fast völlig vernachlässigt, obwohl sie idealtypisch textuelle und graphische Elemente kombinieren. Die Nutzung per GNML annotierten Materials eröffnet neue Möglichkeiten, die psychologische Wirkung graphischer Literatur zu untersuchen.

Wie gelingt es einem Zeichner, die Aufmerksamkeit des Lesers von einem Panel zum nächsten zu lenken? Wie interagieren textuelle und graphische Elemente bei der Rezeption eines Comics? Scott McCloud (4) hat dazu erste theoretische Überlegungen angestellt und visualisiert; Neil Cohn (5) hat die theoretische Analyse weiterentwickelt zu einer formalen „visuellen Sprache“, die vergleichbar einer generativen Grammatik die narrativen Elemente einer graphischen Geschichte kategorisiert. Haben diese Ordnungsschemata eine psychologische Realität? Erste Eyetracking-Studien aus unserem Labor zeigen, dass das von McCloud postulierte Ausmaß an „Closure“, das zwischen unterschiedlichen Arten von Panel-Übergängen variiert, sich in unterschiedlich langen Betrachtungszeiten niederschlägt.

Wir zeigen außerdem erste Ergebnisse aus einem empirischen Eyetracking-Corpus graphischer Literatur. Im Kontext des Projektes wird ein R-Paket zum Import von GNML-Daten und zur statistischen Analyse und Visualisierung von Blickbewegungen und Corpusdaten entwickelt, das in diesen Analysen zur Anwendung kommt.

Literaturverzeichnis

1. Siehe etwa: **Moretti, Franco**. *Distant Reading*. London: Verso, 2013.

2. **Walsh, John.** Comic Book Markup Language: An Introduction and Rationale. *Digital Humanities Quarterly.* 2012, Volume 6, Number 1.

3. Text Encoding Initiative - P5: Guidelines for Electronic Text Encoding and Interchange. [Online] 2.7.0, 09 16, 2014. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/.

4. **McCloud, Scott**. Understanding Comics: The Invisible Art. Northampton, MA: Kitchen Sink Press, 1993.

5. **Cohn, Neil**. The Visual Language of Comics. London: Bloomsbury, 2013.

Wann findet die deutsche Literatur statt? Zur Untersuchung von Zeitausdrücken in großen Korpora

Frank Fischer[1] und Jannik Strötgen[2]

[1] Göttingen Centre for Digital Humanities, Universität Göttingen, Deutschland
[2] Institut für Informatik, Universität Heidelberg, Deutschland

1 Einleitung

Exakte Datumsangaben sind ein Merkmal vieler Prosatextsorten. In der Literatur finden sich dagegen bevorzugt ungefähre Datumsangaben, die Interpretationsräume öffnen. So sind etwa alle 19 Monatsnennungen in Theodor Storms *Schimmelreiter* ungefährer Natur („an einem October-Nachmittag", „Ende März" usw.). Ausnahmen von dieser Regel bilden Tagebuch-, Brief-, Abenteuer- und historische Romane: In Goethes *Werther* oder Jules Vernes *Tour du monde en quatre-vingts jour* wimmelt es genretypisch von exakten Datumsangaben, im letztgenannten Roman sind sie in Form eines Countdowns gar unentbehrlicher Teil der Handlung. Ansonsten lässt sich als Hypothese formulieren: Kommt in der erzählenden Literatur ein exaktes Datum vor, ist das eine narrative Setzung, die der näheren Analyse lohnt. Davon ausgehend lassen sich dezidiert literaturwissenschaftliche Fragen ins Blickfeld nehmen:

- Ist die Vermeidung exakter Datumsnennungen tatsächlich ein durchgehendes Merkmal bestimmter literarischer Genres? In welchem Verhältnis stehen exakte zu ungefähren Datumsangaben?
- Kann die Frequenzanalyse von Datumsangaben zu Genreuntersuchungen genutzt werden?
- Welche andere Bedeutung haben Datumsangaben neben der zeitlichen Verortung der Handlung? ('Semiotisierung' eines Datums)
- Gibt es zeitliche Konjunkturen für bestimmte Datumsnennungen? Wenn ja, warum?
- Welche Rolle spielen fiktive Daten? (vgl. etwa Erich Kästners *35. Mai* und Shakespeares 80. April in *The Winter's Tale*, Autolycus' Ballade im 4. Akt)

In unserem Vortrag widmen wir uns den Datumsangaben als einem isolierbaren *feature* literarischer Korpora im Sinne Matthew Jockers: „Indeed, the very object of analysis shifts from looking at the individual occurrences of a feature in context to looking at the trends and patterns of that feature aggregated over an entire corpus" (Jockers 2013, S. 24). Anhand dieses Features (der expliziten Datumsnennung) als einzeln betrachtbare Einheit soll die Praxis der literaturwissenschaftlichen Makroanalyse methodisch bereichert werden. Die Extraktion der Zeitangaben erfolgt automatisiert mithilfe eines Temporal Taggers. Indem es dabei um die Untersuchung der Repräsentanz eines außerliterarischen Phänomens (Zeit, Datumsangaben) in literarischen Texten geht, wird auch ein Beitrag zur Erzählforschung geleistet. Die mit den Methoden der Digital Humanities erzielten Erkenntnisse werden dadurch in die Fachdisziplin (in diesem Fall die Literaturwissenschaften) zurückgespielt.

2 Vorgehen

Unser Vorgehen bestand aus vier meist parallel ablaufenden Schritten: 1. Zusammenstellung geeigneter (deutschsprachiger) Korpora. 2. Erhebung der Daten durch Einsatz des Temporal Taggers HeidelTime (Strötgen & Gertz 2012) zur automatischen Extraktion zeitlicher Ausdrücke im Sinne der Temporal Markup Language TimeML (Pustejovsky et al. 2003). 3. Analyse der Daten (von Heatmaps zu Einzelfällen). 4. Entwicklung einer Android-App zur explorativen Analyse des „literarischen Jahres".

Zunächst haben wir mit dem TextGrid Repository[1] und Gutenberg-DE[2] zwei große literarische Korpora zusammengebracht, mit HeidelTime auf Datumsstrukturen untersucht und anhand der expliziten (und damit sehr sicher richtigen) Ausdrücke eine kalendarische Heatmap erstellt ('1' bedeutet 0–9 Vorkommen, '2' bedeutet 10–19 Vorkommen usw., '+' bedeutet 90 oder mehr Vorkommen). Dabei zeigten sich erwartete und unerwartete Konjunkturen:

```
JAN:  +566554435455576554574445555455
FEB:  65655546454635554446554666762
MAR:  84553334565364646+4644435444465
APR:  +555553446644665544655556464 47
MAY:  +777765557566565674836454565466
JUN:  95748657464665665 7586656444576
JUL:  947948655446897567 6654555565465
AUG:  +6+565555+5665+6974755775555556
SEP:  9745735555446575+7695554546457
OCT:  +375564536445665+76734555645555
NOV:  +685575466565496655554645545456
DEC:  774557554644555476445 54+5533457
```

Wie man sieht, kommen Monatserste und fixe Feiertage (Neujahr, Weihnachten) besonders häufig vor. Ansonsten fiel die Konjunktur weiterer Tage auf, etwa des '18. März'. Die Vermutung lag nahe, dass die Nennung dieses Datums in Werken nach 1848 und damit nach der Märzrevolution ansteigt, da dieser Tag wegen der blutigen Ereignisse in Berlin eine eigene Semantik annahm. Die Erhebung und Analyse solcher Datumskonjunkturen soll systematisch ausgebaut werden.

3 Untersuchung des DTA-Korpus

Dieser ersten quantitativen Analyse lagen die beiden erwähnten Korpora zugrunde, in denen aber auch nichtliterarische Werke und (v. a. bei Gutenberg-DE) auch Übersetzungen fremdsprachiger Literatur vorkommen. Um mit dieser Methode belastbare Ergebnisse zu gewinnen, bedurfte es eines qualifizierteren Textkorpus. Das Deutsche Textarchiv (DTA) ist zwar weit weniger umfangreich, aber es beinhaltet (nomen est omen) nur original deutschsprachige Texte und ist sowohl grob zeitlich – nach Jahrhunderten – als auch thematisch – nach Genres – sortiert, sodass dort Hinweise auf Konjunkturen in literarischen Texten zu erwarten waren. In Tabelle 1 sind Informationen zu den DTA-Teilkorpora dargestellt.

Die Anzahl der enthaltenen Texte stellt zwar letztlich keine kritische Masse für large-scale Untersuchungen wie die unsere dar, wir versprachen uns aber weitere Hinweise auf Chancen und Grenzen der Methode. Eine Analyse des einzeln abrufbaren Belletristik-Korpus ergab eine weit weniger stabile Heatmap als oben und als häufigste Daten (zwischen ca. 20 und 70 Nennungen) die folgenden: 1. 1., 1. 4., 20. 4., 1. 5., 10. 8., 3. 11.

Die DTA-Stichprobe hebt wieder die Bedeutung des 1. Mai heraus, unter den unerwarteten Daten sei der '10. August' herausgegriffen, der neben der zeitlichen Verortung fiktionaler Handlungen wiederum auch auf ein historisches Datum zurückzuführen ist, den Tuileriensturm am 10. August 1792 (vgl. in Büchners *Dantons Tod*: „Erster Bürger: Danton war unter uns am 10. August,

[1] http://www.textgridrep.de/
[2] http://gutenberg.spiegel.de/

	Dokumente	Sätze	Tokens	TimeML Zeitausdrücke	explizite Tagausdrücke
1600–1699	124	957,950	15,779,536	34,425	635
1700–1799	341	1,866,106	30,077,557	97,219	1,531
1800–1899	559	3,565,758	60,750,795	289,697	18,644
Belletristik	401	1,774,209	29,303,975	106,988	3,366
Gebrauchsliteratur	103	624,829	11,041,708	35,079	787
Wissenschaft	532	4,103,679	68,268,553	298,195	17,170

Tabelle 1: Informationen zu den Teilkorpora des DTA-Korpus.

Danton war unter uns im September.“). Das DTA-Korpus ist aber, wie gesagt, relativ klein (nur 401 belletristische Werke, wobei Belletristik hier neben Fiktionalem auch Reiseberichte und Lebenserinnerungen einschließt) und nicht sehr belastbar für Korpusanalysen.

4 Bedeutung für die Literaturwissenschaft

Es wäre analog zu Hans Ulrich Gumbrechts Untersuchung *1926* denkbar und wünschenswert, dass man die Literaturgeschichte einzelner Tage schriebe. Dass jedes Datum seine eigene Literaturgeschichte hat, dies also in Ansätzen schon untersucht wird (allerdings noch ohne Korpusanalysen o. Ä.), zeigt das Beispiel Paul Celan und der ‘20. Jänner’.

In Celans Prosagedicht *Gespräch im Gebirg*, 1960 in der *Neuen Rundschau* (Heft 2) erschienen, wird auf Georg Büchners Erzählung *Lenz* angespielt, in der ebenfalls ein Gang durchs Gebirge geschildert wird. Büchners Text beginnt mit dem Satz: „Den 20. [Jänner] ging Lenz durchs Gebirg.“ Dass Lenz’ Wanderung durchs Gebirge am 20. Jänner stattfindet, darauf weist auch Celan in seiner Büchner-Preis-Rede *Der Meridian* hin. Er erweitert den Anspielungsraum noch, indem er auf einen weiteren 20. Januar verweist, nämlich den des Jahres 1942, als in Berlin die Wannseekonferenz stattfand. Und er folgert: „Vielleicht darf man sagen, daß jedem Gedicht sein ‘20. Jänner’ eingeschrieben bleibt?“ (vgl. Sieber 2007, S. 114ff).

Die computergestützte Erhebung von Zeitangaben aus großen Korpora macht solche Gleichzeitigkeiten sichtbarer und ermöglicht so auch deren systematische Erforschung.

5 Android-App zur Exploration expliziter Zeitausdrücke in der Weltliteratur

Um eine Idee für die Jahreszeitlichkeiten der Literatur zu entwickeln, haben wir parallel zum Projekt eine Android-App entwickelt, die als Kalender funktioniert und für jeden Tag des Jahres Passagen aus der Weltliteratur anzeigt, die an diesem Tag spielen. Beispiele sind in Abbildung 1 dargestellt.

Die von uns entwickelte und mit HeidelTime belieferte App soll die einfache Exploration expliziter Zeitausdrücke in der Weltliteratur ermöglichen. Dass James Joyce’ *Ulysses* etwa am 16. Juni 1904 stattfindet (‘Bloomsday’), ist allgemein bekannt, allerdings im Text nicht sofort ersichtlich. In der App wird die einzige Stelle des überlangen Romans, an dem das Datum erwähnt wird, zitiert (die Sekretärin Miss Dunne tippt es auf ihrer Schreibmaschine ein). Weitere Beispiele für Passagen sind der 12. Juni in der *Blechtrommel* (Oskar Matzeraths erklärter Sohn Kurt wird geboren), der 29. Februar in Thomas Manns *Zauberberg* (der als spezielle Variante der Walpurgisnacht eine zentrale Rolle spielt, siehe Abbildung 1) oder der 27. Juli in Stefan Zweigs *Schachnovelle* (der Protagonist Dr. B. eignet sich an diesem Tag das Schachbuch an). Die App stellt somit ein erweiterbares Korpus mit Datumsnennungen in der Weltliteratur dar, das der weiteren Forschung zur Verfügung steht. Um allerdings das gesamte „literarische Jahr“ abzubilden, müssen auch die *ungefähren* zeitlichen Verortungen in den Blick genommen werden, was im nächsten Abschnitt versucht werden soll.

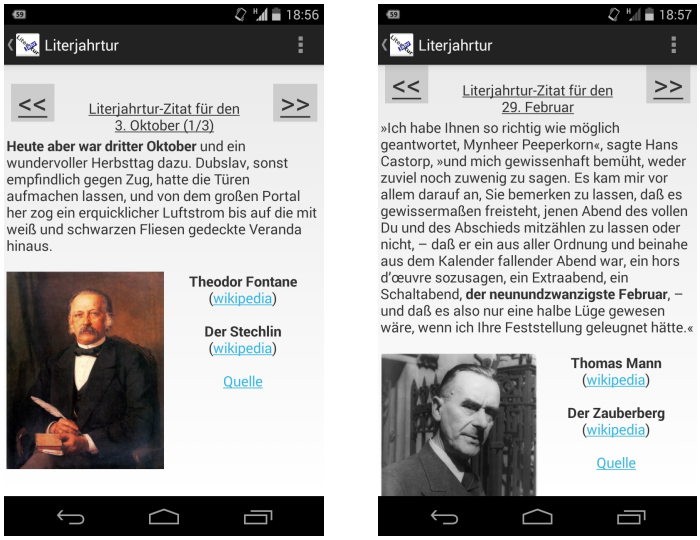


Abbildung 1: Screenshots unserer Android-App *Literjahrtur*.

	Fontane	Storm
JAN	30	5
FEB	13	3
MAR	18	7
APR	13	7
MAY	28	11
JUN	17	9
JUL	16	6
AUG	17	10
SEP	36	10
OCT	41	11
NOV	27	13
DEC	25	1

Tabelle 2: Fontane vs. Storm.

6 Die Jahreszeiten der Literatur

Von dem für literarische Texte sehr spezifischen Verhältnis zwischen exakten und ungefähren Datumsangaben war schon die Rede. Die Analyse des DTA-Korpus nur mit (auch nicht-literarischen) Texten aus dem 19. Jahrhundert hat eine besondere Konjunktur für die Monate März bis Juli erbracht:

```
JAN:  +5755445554544535544443444444667
FEB:  535534435334244443434456453 52
MAR:  +6667676789669+49+8+6877888699+
APR:  +78878876+76597+75699+89668869
MAY:  ++6+968758+899+8886+9+98987+9++
JUN:  +8489768+++++978697+976++9988+
JUL:  +789987+758978+7765887565767755
AUG:  655654445845555555454455445345
SEP:  64433343443334534534322333334
OCT:  722222222223333223124232323332 4
NOV:  84555345443443644553434444343 4
DEC:  24232222162222221122222421411 23
```

Bei der Suche nach Abweichungen im Werk einzelner Autoren des 19. Jahrhunderts stießen wir etwa auf Theodor Fontane und Theodor Storm. Eine Erhebung nur der Monatsnennungen in ausgewählten fiktionalen Texten beider Autoren ergab das in Tabelle 2 verzeichnete Bild.

Analog zur Popularität des 1. Mai ist auch der Gesamtmonat bei beiden stark repräsentiert. Doch für die Sommermonate gilt das nicht. Fontanes Romane und Erzählungen scheinen vor allem von September–Januar stattzufinden, Storms Texte von August–November. Auch unter der Maßgabe, dass der Monatsname als sprachlich-klangliches Zeichen einen stilistischen Effekt hat, scheinen beide Autoren herbstlich-winterliche Settings und Stimmungen zu bevorzugen.

Mit den vorgestellten Methoden zur Ermittlung von Datumskonjunkturen, zur Beschreibung des Verhältnisses zwischen ungefähren und exakten Datumsangaben, zum Aufbau eines Korpus mit exakten Datumsnennungen und zur Jahreszeitlichkeit der Literatur und bestimmter Autoren kann die im Titel gestellte Frage „Wann findet die deutsche Literatur statt?“ tatsächlich makroanalytisch beantwortet werden. Damit die getroffenen Aussagen tatsächlich literaturhistorisch belastbar sind, soll für die weitere Forschung ein größeres und besser (v. a. mit Entstehungs-/Veröffentlichungsdaten) ausgezeichnetes Korpus zusammengestellt werden.

Literatur

Matthew Jockers. *Macroanalysis. Digital Methods and Literary History.* Chicago: University of Illinois Press, 2013.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, S. 28–34, 2003.

Mirjam Sieber. *Paul Celans "Gespräch im Gebirg". Erinnerung an eine versäumte Begegnung.* Tübingen: Niemeyer, 2007.

Jannik Strötgen, Michael Gertz. Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, S. 3746–3753, 2012.

Abstract

Erzählen im Computerspiel. Der Text als „Designer's Narrative Discourse" und über Möglichkeiten seiner Präsentation.

Im Rahmen des Vortrags werden (1) Besonderheiten und Herausforderungen einer Annäherung an das Computerspiel als narratives Medium diskutiert sowie (2) die Möglichkeit der Textpräsentation problematisiert.

(1) Die zugrundeliegende Forschungsarbeit schlägt ein Analysemodell vor, welches die Adaption und Weiterentwicklung bestehender Modelle forciert und damit im Sinne eines interdisziplinären Zugangs möglichst viele Anschlussmöglichkeiten für die Literaturwissenschaft an die Computerspielforschung und umgekehrt schaffen möchte. Hierbei erweist sich der besondere Nutzen kommunikationswissenschaftlicher Zugänge und des Signals als Mittelpunkt der Analyse.

Das Computerspiel ist als Hybrid aus Spiel und Erzählung durch die Interdependenz von spiel- und erzählrelevanten Elementen charakterisiert. Unter dieser Voraussetzung werden unterschiedliche Modelle der Spiel- und Erzählforschung herangezogen, um einen Textbegriff zu entwickeln, der zur Beschreibung von Narrativität im Computerspiel verwendet werden kann. Der Vortrag thematisiert hierfür die vielfältigen Diskursebenen im und um das Medium und unterscheidet die individuelle Rezeption der Spiele von intrinsischen narrativen Elementen. Der entwickelte Textbegriff als „Designer's Narrative Discourse" präsentiert sich letztlich als Interaktion von intrinsischen narrativen Elementen und der potentiellen narrativen Funktion der Spielmechanik. Darauf aufbauend wird ein Überblick über das dramaturgisches Repertoire des Mediums geboten, das Besonderheiten des Computerspiels wie die Spielperspektive und die Spielmechanik berücksichtigt. Erzählen wird als Kommunikationsprozess zwischen Spieler_in, Spielfiguren, Requisiten und der Spielwelt angenommen, wenn das Signal, seine Übertragungskanäle und seine Qualitäten im Mittelpunkt der Analysemethode stehen. Um der besonderen Bedeutung der Räumlichkeit im Computerspiel Rechnung zu tragen, wird sie ebenfalls als absolute und potentielle Größe diskutiert.

Anhand ausgewählter Beispiele aus der zugrundeliegenden Forschungsarbeit wird die vorgeschlagene Analysemethode demonstriert und die These von der Notwendigkeit, das Computerspiel verstärkt als narratives Medium zu begreifen, weiter argumentiert. Nach Auffassung des Vortragenden zeigt das Computerspiel nicht nur: „Daten sind mehr als Text", sondern auch, dass der Text im Computerspiel oft mehr als die Summe seiner Daten ist.

(2) Die Beschreibung des „Designer's Narrative Discourse" im Computerspiel steht allerdings vor der methodischen Herausforderung, dass sie sich individueller Rezeptionserfahrungen bedienen muss (Player's Narrative Discourse). Diese repräsentieren nur einen Bruchteil des im „Designer's Narrative Discourse" angelegten potentiellen Textes. Vor diesem Hintergrund muss die inhaltliche Präsentation – auch wenn sie durch die Erstellung von In-Game-Videos, die auf Videoplattformen wie „Youtube" zugänglich gemacht werden können, erheblich verbessert wird – kritisch reflektiert werden.

Vgl: Bernhard Friedl: „Erzählen im Computerspiel. Methodik und Beispiele für eine kommunikationswissenschaftliche Betrachtung des Computerspiels als narratives Medium." Diplomarbeit (82S.). Wien: 2014

Literatur (Auswahl)

- Aarseth, Espen: A narrative theory of games. In: FDG '12 Proceedings of the International Conference on the Foundations of Digital Games. New York: 2012, S. 129–133
- Backe, Hans-Joachim: Strukturen und Funktionen des Erzählens im Computerspiel. Eine typologische Einführung. Würzburg: 2008
- Carr, Diane; Buckingham, David; Burn, Andrew; Schott, Gareth: Computer games : text, narrative and play. Cambridge: 2006
- Günzel, Stephan: The Spatial Turn in Computer Game Studies. In: Exploring the edges of gaming: proceedings of the Vienna Games Conference 2008 - 2009; future and reality of gaming. Wien: 2010, S. 147–156
- Jenkins, Henry: Game Design as Narrative Architecture. In: Noah Wardrip-Fruin, Pat Harrigan (Hg.): First Person. New Media as Story, Performance, and Game. Cambridge, London: 2004, S. 117–130
- Juul, Jesper: Games Telling Stories?–a Brief Note on Games and Narratives. In: Games Studies Vol. 1 (2001), URL: http://www.gamestudies.org/0101/juul-gts/
- Krämer, Sybille: Medium, Bote, Übertragung : kleine Metaphysik der Medialität. Frankfurt: 2008
- Nohr, Rolf F.: Raumfetischismus. Topographien des Spiels. In: Klaus Bartels, Jan-Noël Thon (Hrsg.): Computer/Spiel/Räume. Materialien zur Einführung in die *Computer Game Studies* (Hamburger Hefte zur Medienkultur Band 5). Hamburg: 2007, S. 61–81
- Pfister, Manfred: Das Drama: Theorie und Analyse. München: 2000
- Rouse, Richard: Game design. Theory & practice. Plano: 2005
- Ryan, Marie-Laure: The Interactive Onion: Layers of User Participation in Digital Narrative Texts. In: Ruth Page, Bronwen Thomas (Hg.): New Narratives : stories and storytelling in the digital age. Lincoln, London: 2011, S. 35–62
- Ryan, Marie-Laure: Narrative as virtual reality. Immersion and interactivity in literature and electronic media. Baltimore: 2001

Kontakt:

Bernhard Friedl

Koppstraße 54/34

1160 Wien

Mobil: 0043 650 828 78 21

Email: bernhard.friedl@gmail.com

DHd2015

Von Daten zu Erkenntnissen

23. bis 27. Februar 2015, Graz

Poster

CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

Der Gang durch die Domänen

zur Erfassung, Aufbereitung und Präsentation von Audiodaten im BMBF-Projekt „Freischütz Digital“

Benjamin W. Bohl · Thomas Prätzlich · Meinard Müller · Joachim Veit

Abb. 1 Exemplarische Darstellung von Archivinhalten – Noten- und Librettotexte (jeweils in Faksimile und Transkription) sowie Audioaufnahmen.

Das BMBF-Projekt *Freischütz Digital*[1] (FreiDi), widmet sich der paradigmatischen Konzeption und Umsetzung eines genuin digitalen Editionskonzepts am Beispiel von Carl Maria von Webers Oper *Der Freischütz*.

Die International Audio Laboratories Erlangen sind eine gemeinsamen Einrichtung der Friedrich-Alexander-Universität Erlangen-Nürnberg und des Fraunhofer-Instituts für Integrierte Schaltungen IIS.

Benjamin W. Bohl und Joachim Veit
Musikwissensenschaftliches Seminar Detmold/Paderborn
Gartenstr. 20
32756 Detmold
Tel.: +49 5231 975-876
E-Mail: thomas.praetzlich@audiolabs-erlangen.de

Thomas Prätzlich und Meinard Müller
International Audio Laboratories Erlangen
Am Wolfsmantel 33
91058 Erlangen – Tennenlohe
Tel.: +49 9131 85–20 520
E-Mail: thomas.praetzlich@audiolabs-erlangen.de

[1] http://www.freischuetz-digital.de

Leitgedanke ist Wierings *Multidimensional Model* [6], das im Sinne eines *kritischen Archivs* eine dichte Verknüpfung und Annotierung von Quellen unterschiedlicher medialer Ausprägung vorsieht (siehe Abb. 1). Ein für *FreiDi* wichtiger Aspekt ist in diesem Zusammenhang der erstmalige Einbezug von Audio-Daten in den Kontext einer Digitalen Edition (Detmold). Hierfür werden Algorithmen zur automatisierten Audiosegmentierung eingesetzt und entwickelt (Erlangen) [4,5].

In dieser Kombination von Musikinformatik und Musikwissenschaft entstehen (für ausgewählte Aufnahmen des *Freischütz*) strukturell und inhaltlich reiche Metadaten, die die graphische, logische und akustische Domänen in Bezug setzen [1]. So werden zum Beispiel Pixelpositionen (graphische Domäne) mit Noteninformationen (logische Domäne) und Zeitpositionen (akustische Domäne) verknüpft. Die Codierungsrichtlinien der *Music Encoding Initiative*[2] (MEI) liefern hierfür einen umfassenden Rahmen.

Während für die Musikwissenschaft die Kombination von graphischer und akustischer Domäne eine perspektivische Weitung im Kontext einer Digitalen Edition ermöglicht [2], und damit etwa neue Möglichkeiten für die Rezeptions- und Interpretationsforschung bietet, kann auch die Musikinformatik von der engen Verknüpfung der logischen mit der akustisch-performativen Domäne profitieren. So können zum Beispiel die mit den Zeitpositionen verknüpften Notentextdaten zur Verbesserung der Ergenisse von Quellentrennungsalgorithmen genutzt werden [3]. Hierbei ist das Ziel einzelne Instrumentenstimmen aus einer Audioaufnahme, die verschiedene gleichzeitig erklingende Stimmen enthält, abzutrennen.

Dieses Poster soll zunächst die Anforderungen der beiden Disziplinen an die Datenmodellierung und die jeweilige Umsetzung in *MEI* vorstellen. Schließlich sollen unter dem Gesichtspunkt der Präsentation und

[2] http://www.music-encoding.org

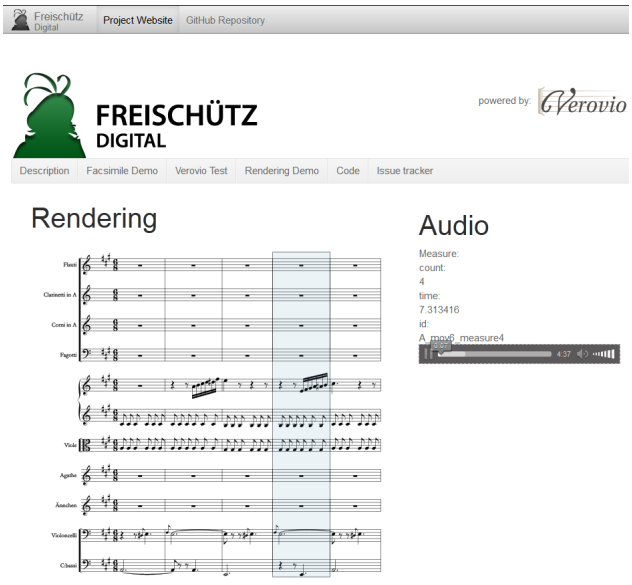


Abb. 2 *FreiDi-scoreFollowsAudio* Screenshot des Demonstrators zur synchronisierten Anzeige gerenderter Noten zur Audiowiedergabe. Der blaue Kasten hebt den aktuell erklingenden Takt automatisch hervor.

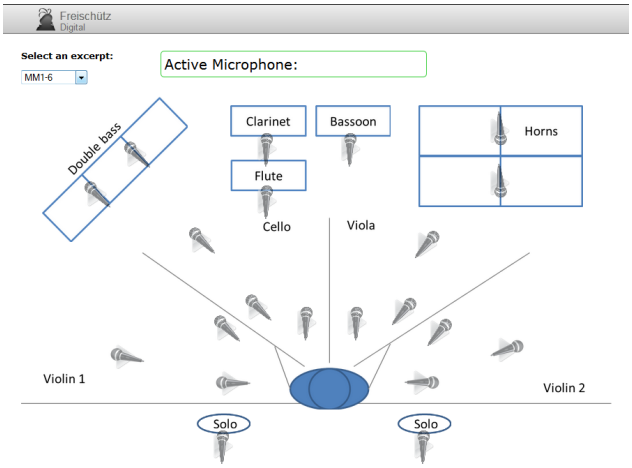


Abb. 3 Screenshot des Demonstrators zum Abspielen von Multitrack Aufnahmen. Die Mikrofon-Symbole sind entsprechend der Aufnahmesituation angeordnet – ein Klick darauf spielt das jeweilige Signal ab.

Nachnutzung der Daten die im Projekt entwickelten Demonstratoren und Web-Applikationen vorgestellt werden, etwa zum automatischen Blättern eines Notentexts zu einer laufenden Aufnahme (siehe Abb. 2), oder zum veranschaulichenden Abspielen von Multitrack Aufnahmen (siehe Abb. 3). Dabei gilt es auch der Frage nachzugehen, ob *MEI* für Echtzeit-Anwendungen sinnvoll einsetzbar ist.

Literatur

1. Babbitt, M.: The use of computers in musicological research. Perspectives of New Music **3**(2), 74–83 (1965)
2. Bohl, B., Kepper, J., Röwenstrunk, D.: Perspektiven digitaler Musikeditionen aus der Sicht des Edirom-Projekts. In: Die Tonkunst **5**, 270–276 (2011)
3. Müller, M., Driedger, J., Ewert, S.: Notentext-informierte Quellentrennung für Musiksignale. In: Proceedings of 43th GI Jahrestagung, 2928–2942. Koblenz, Germany (2013)
4. Prätzlich, T., Müller, M.: Freischütz digital: A case study for reference-based audio segmentation of operas. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 589–594. Curitiba, Brazil (2013)
5. Prätzlich, T., Müller, M.: Frame-level audio segmentation for abridged musical works. In: Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR). Taipei, Taiwan (2014)
6. Wiering, F.: Digital critical editions of music: A multidimensional model. In: Modern Methods for Musicology, 23–45 (2009)

Poster

Big, complex, heterogeneous.. Laufende Projekte aus dem Arbeitsbereich *Big Data in den Geisteswissenschaften* in DARIAH-DE

Stefan Pernes, Uni Würzburg

Der Begriff *Big Data* wird in den unterschiedlichsten Kontexten gebraucht, er umfasst unterschiedliche Größenordnungen und Strategien der Datenverarbeitung, und kann aufgrund dieser heuristischen Schwäche im besten Fall als *Buzzword*, mit Sicherheit aber nicht als trennscharfes Konzept bezeichnet werden. Zu Recht wird die Aufmerksamkeit zunehmend auf Fragen der Reliabilität und Validität der Verfahren gelenkt (Jordan 2014) und übergroße Heilsversprechen sowie Ankündigungen eines *Endes der Theorie* (Anderson 2008) kritisch hinterfragt. In geisteswissenschaftlichen Kontexten rückt diese allgemein geführte Diskussion jedoch in den Hintergrund. Hier gilt es, Textbestände in einer zuvor nicht da gewesenen Größe unter Berücksichtigung ihrer Vielschichtigkeit und Heterogenität zu verwalten und festzustellen, welchen Beitrag quantitative Methoden zu hermeneutischen Interpretationsverfahren leisten können. Diese Spezifika führen auch dazu, dass einige Voraussetzungen erst geschaffen werden müssen; so zum Beispiel das Training bestehender Verfahren der Sprachverarbeitung für literarische Textsorten, das Erstellen spezifischer Korpora und Vokabulare, oder die Verbesserung der Texterkennung von mittelalterlichen Handschriften. Das sind einige der Aufgaben, zu denen die *Use Cases* des DARIAH-DE Clusters *Big Data in den Geisteswissenschaften* einen Beitrag leisten und die im Folgenden vorgestellt werden sollen. Die *Use Cases* bearbeiten Fragestellungen aus Literaturwissenschaft, Philologie und Geschichte und werden jeweils in Kooperation eines fachwissenschaftlichen Partners und eines Partners aus dem Bereich der angewandten Informatik durchgeführt.

Narrative Techniken und Untergattungen im Deutschen Roman

Lehrstuhl für Computerphilologie, Uni Würzburg / Ubiquitous Knowledge Processing Lab, TU Darmstadt

Fachwissenschaftlicher Gegenstand des *Use Case* ist es, anhand quantitativer Verfahren die historische Entwicklung narrativer Techniken und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien nachzuvollziehen. Die Textgrundlage bildet ein Korpus 2000 deutschsprachiger Romane aus dem Zeitraum von 1500 bis 1930 und eine Sammlung von 200 französischen Kriminalromanen aus dem 19. und 20. Jahrhundert. Zum Einsatz kommen Verfahren zur automatischen Erkennung bestimmter Merkmale, wie zum Beispiel der Erkennung von Eigennamen oder von Passagen direkter Rede. Sämtliche Merkmale werden im Anschluss als *Features* zueinander in Bezug gesetzt, um die Texte

zu gruppieren und Gattungsbegriffe nachprüfen zu können. Parallel dazu werden Lernmaterialien erstellt, welche die dabei entwickelten Arbeitsabläufe in Form allgemeinverständlicher *Rezepte* zugänglich machen und interessierte ForscherInnen dazu ermächtigen sollen, *state of the art* Werkzeuge der Sprachverarbeitung für ihre jeweiligen Forschungsvorhaben einzusetzen und auf ihre jeweiligen Daten anzupassen.

Identifikation grenzübergreifender Lebensläufe in nationalen Biografien

Leibniz-Institut für Europäische Geschichte Mainz / Lehrstuhl für Medieninformatik, Uni Bamberg

Der *Use Case* erforscht die Verbindungen von individuellen historischen Lebensläufen und Internationalitätskriterien auf Grundlage von *Wikipedia* und mehreren europäischen Nationalbiografien. Dabei werden verschiedene Merkmale von Mobilität - wie zum Beispiel Geburts-, Wirkungs- und Sterbeorte, Tätigkeiten und verwandtschaftliche Beziehungen - miteinander korreliert und durch eine gezielte Erhebung sämtlicher Zusammenhänge mitunter Beobachtungen gemacht, die in den Geschichtswissenschaften noch nicht theoretisch erfasst sind. Die Datengrundlage umfasst strukturierte Daten sowie unstrukturierte Texte in mehreren Sprachen die miteinander verschränkt werden. Zusätzlich zum fachwissenschaftlichen Erkenntnisgewinn, stellt das Vorhaben eine quantitative Grundlage für kontrollierte Vokabulare in der Biografieforschung dar und zeigt auf, welche inhaltlichen und formalen Kategorien für *Semantic Web*-Ansätze in der Biografieforschung erforderlich und nutzbringend sein können.

Spuren der Zitation und Wiederverwendung im OpenMigne Korpus

Lehrstuhl für Digital Humanities, Uni Leipzig / Lehrstuhl für Medieninformatik, Uni Bamberg

Ausgehend von Editionen der Texte frühchristlicher Kirchenväter durch Jacques Paul Migne im 19. Jahrhundert entwickelt der *Use Case* Verfahren zur Erschließung vollständiger diachroner *Zitationsspuren.* Es handelt sich dabei um ein Feststellen chronologisch nachvollziehbarer Verläufe in einem Netzwerk von Zitationen, welches sich über ein gesamtes Korpus spannt. Textgrundlage bildet das *OpenMigne* Korpus, dessen Texte in griechischer und lateinischer Sprache einen Zeitraum vom Ursprung des Christentums bis in das 15. Jahrhundert abdecken. Die technische Umsetzung verläuft schrittweise: Beginnend mit der Erkennung von Zitationen in direkter Rede und in gleichsprachigen Ursprungstexten werden die Verfahren dahingehend erweitert, dass auch eine Erkennung von Paraphrasierungen und Zitationen in sprachlich heterogenen Korpora möglich wird. Die entwickelten Verfahren werden weiters für eine Anwendung über den *Use Case* hinaus aufbereitet und bereitgestellt.

Fazit

Anhand der vorgestellten Projekte wird ein Mal mehr deutlich, wie unterschiedlich die Voraussetzungen und Fragestellungen sein können, die unter dem Begriff *Big Data* verhandelt werden. Dabei tritt jedoch auch in den Vordergrund, was in diesem Feld die gemeinsamen, spezifisch geisteswissenschaftlichen Interessen sein können - ein methodologischer Austausch, von dem alle beteiligten Disziplinen profitieren.

Literatur

Jordan, Michael (2014): *Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts.* Interview by Lee Gomes, IEEE Spectrum. http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts (10.11.2014)

Anderson, Chris (2008): *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired Magazine 16.07. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (10.11.2014)

An End-To-End Integration of Automatic Annotations into CATMA

Thomas Bögel* **Marco Petris†** **Jannik Strötgen*** **Michael Gertz***
* Institute of Computer Science, Heidelberg University
{boegel, stroetgen, gertz}@uni-hd.de
† Institute for German Studies, University of Hamburg
marco.petris@uni-hamburg.de

1 Introduction

Natural Language Processing offers solutions for predicting linguistic annotations at different levels of complexity. Thus, it seems obvious and – in general – a good idea to apply these methods to the Humanities in order to automate laborious manual annotations and to facilitate a deeper text analysis understanding. Apart from the purely technical aspect of developing suitable models, however, additional challenges for NLP in the Humanities arise: in order to be used as part of an analysis tool, humanists often desire justifications and explanations of automatic annotations. Just implementing a black-box approach, evaluating it intrinsically and returning the presumably best results to the user is not sufficient. In this paper, we suggest a transparent way of presenting the results of a NLP pipeline in a collaborative setting. This gives the user the possibility to judge the results directly within an already existing annotation interface and potentially use them for individual analysis tasks.

We will first present individual components that are combined with each other, namely the collaborative annotation tool CATMA and UIMA as a processing pipeline for Natural Language Processing. We will then show our end-to-end integration of UIMA into CATMA and its advantages.

2 CATMA integration

CATMA[1] is a flexible, collaborative annotation tool for literary scholars. So far, it integrates three functional and interactive modules, namely the tagger, the analyzer, and the visualizer. While the tagger module is a graphical interface to allow the easy creation of manual annotations in texts using flexible tag sets (including feature structures, overlapping annotations, etc.), the analyzer component offers a wide range of possibilities to query a document collection or single documents, e.g., for frequently occurring patterns. Finally, the visualizer module can be used to explore a document collection, e.g., by generating distribution charts of the analysis results. In this paper, we present an extension to CATMA, which was developed in the context of the heureCLÉA project[2] - the integration of a UIMA-based text processing pipeline for the automatic creation of tag annotations created by natural language processing tools.

UIMA (Unstructured Information Management Architecture)[3] is a wide-spread framwork for developing and using natural language processing pipelines. One of its key characteristics is that it allows the easy combination of tools that have initially not been built to be used together. All UIMA components rely on the same data structure - the Common Analysis Structure (CAS) - there are three types of components: collection readers, analysis engines, and CAS consumers. The collection readers task is to access the source of the documents that are to be processed and to initialize a CAS object for each document. Then, the analysis engines perform linguistic processing of the data and stand-off add annotations to the CAS object. The subsequently called analysis engines can access the annotation results of the earlier components, i.e., they can perform more complex tasks. Finally, a CAS consumer performs the final processing of the CAS object.

[1] Website: http://www.catma.de/clea

[2] http://heureclea.de/

[3] Website: http://uima.apache.org/

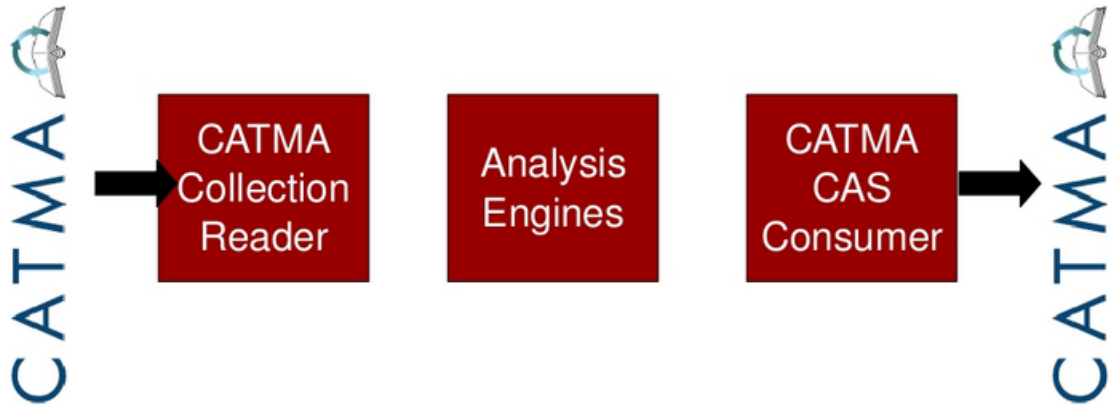


Figure 1: End-To-End architecture of combining the collaborative annotation platform CATMA with the automatic text processing pipeline UIMA.

In our case, the pipeline architecture is set up as depicted in Figure 1. The Collection Reader accesses the documents directly from CATMA and returns annotation information back to CATMA. However, the actual key feature of our development is that the user can directly access the automatic processing feature within the CATMA interface. That is, the user can select the types of annotations that shall be added to her document or document collection automatically. This significantly decreases the boundary for users not familiar with applying NLP tools for automatic processing of textual data, i.e., for typical CATMA users who are often literary scholars or students of the Humanities.

Nevertheless, our implementation is not a black box solution that only adds annotations that the user has to accept. In contrast, we are currently working on integrating a user feedback interface that will allow the initialization of user parameters based on the users feedback in the form of accepted or rejected annotations.

3 Research Workflow within CATMA

The advantage of a direct integration of UIMA into CATMA is best illustrated with an example: in order to analyse the temporal structure of documents (such as order phenomena), many linguistic aspects need to be taken into account. Temporal signals, e.g., calendrical, deictic or relational temporal expressions (Lahn and Meister, 2008), offer a hint for temporal phenomena of order. As manual annotation for these basic linguistic phenomena is laborious, we are currently developing a machine learning system for predicting temporal signals. Figure 2 shows the possibility to create and directly inspect automatic annotations directly within the CATMA interface. With one click, the prediction of our NLP pipeline for temporal signals – or other annotations such as date and time expressions (Strötgen and Gertz, 2013) – can be shown. Note that the system output can easily be compared to any manual annotation as the type systems are completely independent. This flexibility allows scholars to focus on complex phenomena of the text with the possibility of automating simpler annotations. All automatic annotations are, however, non-obtrusive and completely changeable and reversible to give the choice of the level of automation to the user.

References

Lahn, S. and J. C. Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart: JB Metzler.

Strötgen, J. and M. Gertz (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation 47*(2), 269–298.

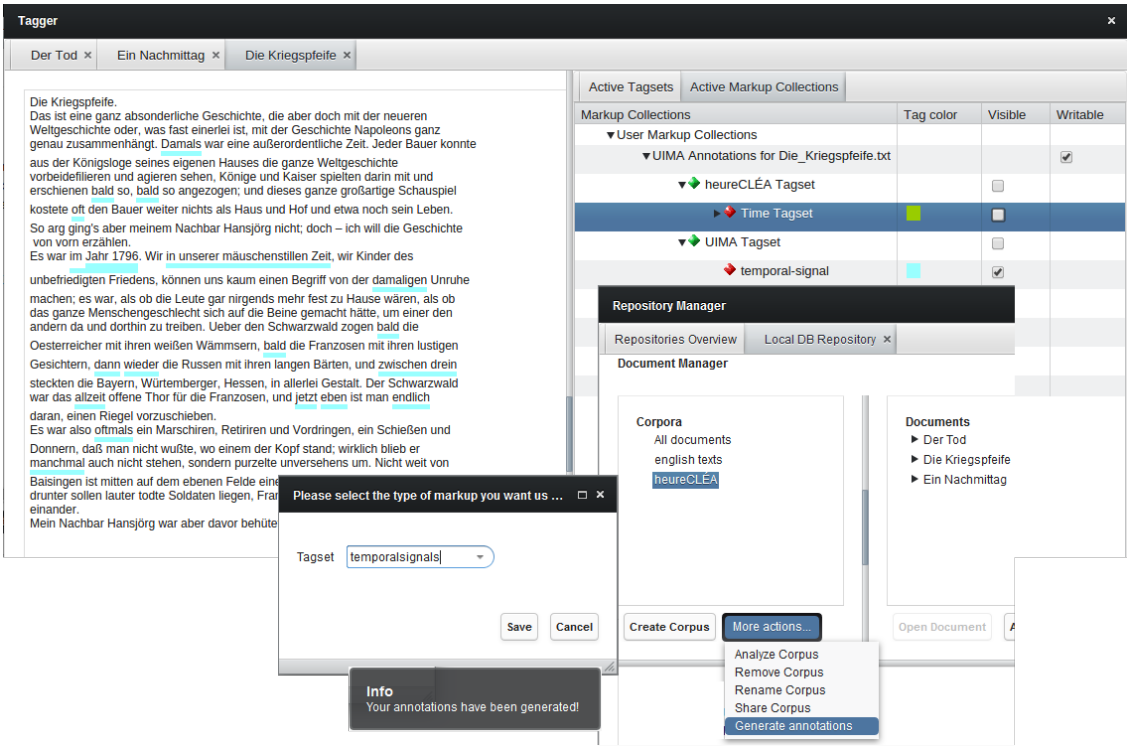


Figure 2: Screenshot showing automatic annotations within CATMA.

Abstract zum Vorhaben „Sprachwissenschaftliche Untersuchungen zum Klagspiegel Conrad Heydens (1436) und zum Laienspiegel Ulrich Tenglers (1511)“

Von Dr. Barbara Aehnlich und Elisabeth Witzenhausen

Das Forschungsvorhaben ist interdisziplinär angelegt und beruht auf einem Korpus von verschiedenen Textzeugen zweier frühneuhochdeutscher Rechtsbücher des 15. und 16. Jahrhunderts, Klagspiegel und Laienspiegel. Der Klagspiegel ist das mit Abstand älteste populärwissenschaftliche Rechtsbuch der Rezeptionszeit und bildet mit dem Laienspiegel zusammen die wichtigste Grundlage an rechtswissenschaftlichen populären Texten des 15. und 16. Jahrhunderts. Ziel ist die Untersuchung der sprachlichen Besonderheiten der Texte und ihrer Auswirkungen auf die Rezeptionsgeschichte des römischen Rechts in Deutschland. Neben der korpusbasierten linguistischen Analyse der Bücher, die eine völlig neue Textsorte begründen, bietet das Projekt auch aus der Perspektive der historischen Rechtssprachenforschung einen innovativen Ansatz. Das Erkenntnisinteresse liegt hierbei auf der Geschichte von Kulturtransferprozessen innerhalb der Jurisprudenz. Durch semantische und linguistische Annotationen wird eine umfassende Forschungsgrundlage geschaffen, die für die Schließung rechts- und sprachhistorischer Forschungslücken einen zentralen Beitrag leistet. Ein weiterer Schritt soll die Digitalisierung mehrerer Ausgaben des Klagspiegels sein, um Prozesse des Schreibsprachwandels im 15. und frühen 16. Jahrhundert nachzuvollziehen. Bisher gibt es kein Korpus frühneuhochdeutscher Rechtstexte.

In einem ersten Schritt zur Vorbereitung des Projektes wurden verschiedene Annotationstools getestet und geeignete Formate für die Speicherung evaluiert. Aktuell werden mit der Jenaer Computerlinguistik Möglichkeiten der Normalisierung und automatischen Annotation erprobt. Ziel ist die Beantragung eines größeren Forschungsprojektes, das bestehende Werkzeuge nutzt und die Technologie auf die Besonderheiten des Rechtskorpus anpasst. Das Poster soll die bisherigen methodologischen Überlegungen und Probleme darstellen und bietet somit gleichzeitig einen Überblick und eine Evaluation der aktuell zur Verfügung stehenden Open Source Software zu Annotationszwecken.

Die Untersuchungen beziehen sich zum einen auf die sprachliche Herkunft des Klagspiegels und des Laienspiegels. Es soll festgestellt werden, welche Textsorte mit welchen spezifischen

sprachlichen Eigenheiten vorliegt. Zudem muss geklärt werden, ob diese Rechtsbücher aufgrund ihrer Herkunft nur im südwestdeutschen Raum oder aber im gesamten hochdeutschen Sprachgebiet verständlich waren. Dabei wird nach möglichen Ausgleichstendenzen gesucht, die vom Oberdeutschen abweichen. Auf der Ebene der Syntax ist zu fragen, welche Strukturen die sprachliche Einfachheit und leichte Verständlichkeit ausmachen, die den Texten in der gesamten (bisher ausschließlich juristischen) Forschung zugeschrieben wird. Im Bereich des Wortschatzes sind besonders die Bezeichnungen juristischer Fachbegriffe oder Tatbestände für die Forschung interessant, denn für diese gab es zuvor im Deutschen keine entsprechenden Termini. Zum anderen soll untersucht werden, inwieweit Klagspiegel und Laienspiegel frühneuhochdeutschen Sprachstandard aufweisen und ob die beiden Bücher durch ihre Verbreitung eine wesentliche Rolle für die Entwicklung des neuhochdeutschen Sprachstandards im Rahmen rechtswissenschaftlicher Prozesse gespielt haben können. Ein Vergleich mehrerer Textzeugen der Rechtsbücher liefert Erkenntnisse des frühneuhochdeutschen Schreibsprachwandels. Der Einfluss der beiden Texte auf die deutsche Standardsprache sowie auf die deutsche Rechtssprache wurde bisher noch nicht analysiert; das Vorhaben soll hierfür eine nutzbare Ausgangsbasis liefern. Eine zentrale Frage ist dabei auch, inwieweit römisches Recht und deutsches Recht sprachlich unterschiedlich vermittelt wurden und ob textintern Varianz zwischen den einzelnen Passagen, die zum Teil auch literarisiert sind, festzustellen ist.

Zwei Textzeugen, jeweils eine Ausgabe des Laien- und des Klagpiegels, liegen bereits in digitalisierten Abbildungen vor und wurden transkribiert. Im nächsten Schritt werden sie in ein XML-Format übertragen und sollen semantisch sowie linguistisch annotiert werden, um eine valide Datenbasis für die Untersuchung zu schaffen und das Korpus in einem standardisierten Format in einer Infrastruktur der Digital Humanities zur Verfügung stellen zu können. Im Sinne eines vielseitig nutzbaren Korpus soll die Transkription diplomatisch, mit allen Sonderzeichen und typografischen Besonderheiten, abgebildet werden. Problematisch ist die heterogene Gestalt der Texte, die Mehrfachannotationen notwendig macht. Alle Annotationen werden deshalb in einem XML-Stand-off Format vorgenommen, um eine leichte Übertragung in andere Formate und einen annotationsfreien Primärtext zu ermöglichen. Das TCF-Format bietet hierfür eine gute Möglichkeit und ist mit vielen anderen Formaten kompatibel.[1] Werkzeuge wie WebAnno[2] oder GATE[3] bieten geeignete Arbeitsoberflächen, deren Vor- und Nachteile es zu diskutieren gilt.

[1] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.(06.10.2014, 13.30 Uhr).
[2] https://code.google.com/p/webanno/. (06.10.2014, 13.46 Uhr).

Die Präsentation stellt somit zum einen den Mehrwert der bisher geleisteten Forschungsarbeit im Rahmen der Digital Humanities für sprachwissenschaftliche Untersuchungen historischer Texte heraus, zum anderen werden Grenzen in der Annotation heterogener und nicht standardisierter Sprachdaten deutlich, die weiterer Forschungsarbeit bedürfen. Die interdisziplinär angelegte Forschungsfrage und die unterschiedlichen Zielgruppen des zu erstellenden Korpus sind Faktoren, die es bei der Aufarbeitung der Daten besonders zu beachten gilt.

[3] https://gate.ac.uk/sale/tao/split.html. .(06.10.2014, 12.51 Uhr).

Poster-Abstract

Erfahrungen aus dem *Bibliotheca legum*-Projekt. Zum Aufbau einer Handschriftendatenbank (*http://www.leges.uni-koeln.de*)

Daniela Schulz und Dominik Trump, M.A. (Universität zu Köln)

Seit 2012 entsteht am Lehrstuhl für die Geschichte des Mittelalters (Prof. Dr. Karl Ubl) in Köln eine Datenbank, welche die handschriftliche Überlieferung des weltlichen Rechts im Früh- und Hochmittelalter in den Blick nimmt. Unter weltlichem Recht werden dabei sowohl das römische Recht als auch die germanischen Volksrechte, die sog. *leges barbarorum*, verstanden. Durch die genaue Erfassung aller relevanten Textzeugen, ihrer Produktion und Verbreitung ist es möglich, Rückschlüsse auf das damalige Rechtswissen zu ziehen.

Das Projekt versteht sich zum einen als Ergänzung zur *Bibliotheca capitularium* von Hubert Mordek[1], der grundlegend alle Handschriften mit fränkischen Herrschererlassen, den sog. Kapitularien, gesammelt hat.[2] Diese bilden eine weitere zentrale Quelle der frühmittelalterlichen Rechtsgeschichte. Zum anderen bietet die *Bibliotheca legum* einen umfassenden Überblick über die Überlieferung und damit Rezeption des römischen Rechts im frühen Mittelalter – ein Gebiet, welches bisher in der Forschung nur relativ wenig Beachtung gefunden hat.

Die *Bibliotheca legum* bietet momentan zu 296 Handschriften Informationen zu Datierung, Entstehungsort, Provenienz, äußerer Beschreibung, Inhalt, Literatur und vor allem zu im Internet frei zugänglichen Ressourcen wie Digitalisaten (z.B. aus *Europeana regia*, *Gallica* oder der Bayerischen Staatsbibliothek München) und Katalogeinträgen (z.B. *Manuscripta Mediaevalia*). Das Projekt ist somit konzeptionell als Portal angelegt, welches nicht nur selbst umfassende Informationen bietet, sondern auch vorhandene Ressourcen nachnutzt und miteinander verknüpfen möchte.

Für das Projekt wurden sowohl die einschlägigen Editionen der Rechtstexte als auch ältere und insbesondere neuere und neueste Forschungsliteratur systematisch ausgewertet. Die gesammelten Informationen, die zunächst nur für eine lehrstuhlinterne Nutzung vorgesehen waren, wurden dabei anfänglich in einer Word-Tabelle gesammelt. Um die Ergebnisse in Form einer Webpräsenz einem breiteren Publikum zugänglich machen zu können, überführte

[1] Hubert Mordek, Bibliotheca capitularium regum Francorum manuscripta. Überlieferung und Traditionszusammenhang der fränkischen Herrschererlasse (MGH Hilfsmittel 15), München 1995.

[2] Aufgrund der freundlichen Genehmigung der *Monumenta Germaniae Historica* (MGH) in München, kann seine Studie bzw. Auszüge aus dieser (auf den einzelnen Handschriftenseiten) zum Download angeboten werden.

man diese Datensammlung nach XML und generierte daraus Handschriftenbeschreibungen nach TEI P5, welche den aktuellen Forschungsstand abbilden und zum Download verfügbar sind.

Zur Verwaltung der Webpräsenz wurde mit Wordpress ein kostenfreies Content Management System gewählt, welches zwar als Blogsoftware weite Verbreitung im World Wide Web gefunden hat, bisher aber nicht für XML-basierte Digital Humanities-Projekte herangezogen wurde. Gerade wegen der breiten Community, die an diesem CMS partizipiert und damit dem Vorhandensein zahlreicher Plugins zur Erweiterung der Funktionalitäten, hat sich Wordpress für das Projekt, welches weder über Drittmittel noch über einen technischen Partner verfügt, insgesamt als sehr geeignet erwiesen. Aufgrund der positiven Erfahrungen folgt die momentan entstehende Webpräsenz der Arbeitsstelle „Edition der fränkischen Herrschererlasse“ (ein Projekt der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste unter der Leitung von Prof. Dr. Karl Ubl; URL: *http://capitularia.uni-koeln.de/*) dem Vorbild der *Bibliotheca legum* und setzt ebenfalls auf Wordpress auf.

Die *Bibliotheca legum* bietet die folgenden *Features*:

- Mehrsprachigkeit (Menüführung und Inhalt in deutscher und englischer Sprache)
- verschiedene Browsingzugänge (z.B. nach Signaturen, enthaltenen Rechtstexten, Entstehungszeit und -ort) zu den fast 300 Handschriftenbeschreibungen
- Volltextsuche und facettierte Suche
- Inter- und Hyperlinking (externe Ressourcen)
- Einleitungstexte
- Übersicht über die in mittelalterlichen Bibliothekskatalogen bezeugten Rechtscodices
- umfangreiche Projektbibliographie sowie Register zu Orten, Personen und haltenden Institutionen unter Verwendung von Normdaten wie VIAF und TGN
- ein Projektblog (deutsch/englisch), der über aktuelle Entwicklung informiert
- umfangreiche Materialsammlung und Visualisierungen (Karten, Stemmata, Transkriptionen)
- Downloads (z.B. die *Bibliotheca capitularium*, Handschriftenbeschreibungen)

Innerhalb recht kurzer Zeit – die Datenbank ist erst seit September 2012 online – konnte sie sich in der historischen und rechtshistorischen Forschung als Arbeitsinstrument etablieren. Neben dem regelmäßigen Einsatz in Lehrveranstaltungen an der Universität zu Köln, fand sie z.B. auch im MOOC „Karl der Große – Pater Europae?“ (URL:

https://iversity.org/de/courses/karl-der-grosse-pater-europae) von apl. Prof. Dr. Rainer Leng (Universität Würzburg) Erwähnung, wo sie als Arbeitsinstrument für die Geschichtswissenschaft präsentiert wird.

Das Poster soll das Projekt nun erstmals auch der deutschsprachigen DH-Community vorstellen. Dabei soll zum einen der aktuelle Stand des Projekts sowie dessen Nutzen für die Wissenschaft, zum anderen dessen – sicher nicht gewöhnliche – Genese dargestellt werden. Die Präsentation kann für all jene von Interesse sein, die digitale Projekte unter ähnlichen Umständen bzw. Voraussetzungen (Fehlen technischer und finanzieller Mittel) realisieren wollen.

Leitung des Projekts: Prof. Dr. Karl Ubl, Lehrstuhl für die Geschichte des Mittelalters, Schwerpunkt Früh- und Hochmittelalter, Historisches Institut der Universität zu Köln

Mitarbeiter: Daniela Schulz (Technische Umsetzung), Dominik Trump, M.A. (Inhaltliche Bearbeitung)

Common Names 4 living organisms @ EUROPEANA

Ein Beispiel für Mehrwert durch interdisziplinäre Kollaboration

Im Zeitalter von digitalisierten Wissenschaften, in denen das Aufbereiten, das zur Verfügung stellen und Austauschen von Daten, Fakten und Erkenntnissen einen mehr als wichtigen Stellenwert einnimmt, ist es somit nicht verwunderlich, dass auch der Aspekt der interdisziplinären Zusammenarbeit und die Gestaltung bzw. Durchführung von Landesgrenzen übergreifenden Projekten mehr und mehr in das Zentrum der Arbeit von Forschern und Forscherinnen rückt und gleichzeitig diese Anforderungen an die Wissenschaften selbst gestellt werden.

Im vorgeschlagenen Poster wird eine interdisziplinäre Zusammenarbeit zwischen den Lexikographen und Lexikographinnen des *Instituts für Corpus Linguistik und Texttechnologie* der *Österreichischen Akademie der Wissenschaften* mit den Kollegen des *Common Names Service* (CNS) des Naturhistorischen Museums Wien (NHM) vorgestellt. Der Mehrwert für beide Disziplinen, die Datenpublikation im Rahmen von Europeana.EU und die Europäisierung der Services stehen im Zentrum der Präsentation.

Die Zusammenarbeit war seitens der Projektpartner wie folgt motiviert:

1) Die *Datenbank der bairischen Mundarten in Österreich* (DBÖ) weist unter anderem eine umfassende Sammlung volkstümlicher Pflanzennamen auf (geschätzt 30,000). Um die Daten lexikographisch im *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ) entsprechend wissenschaftlich dokumentieren zu können, muss der Lexikograph bzw. die Lexikographin den jeweiligen Common Name einer konkreten Pflanze zuweisen (Definition). Aufgrund der Historizität der Daten (Sammelzeitraum „von den Anfängen bis in die Gegenwart“) ist das eine fachspezifische Aufgabe der Botaniker und Botanikerinnen.

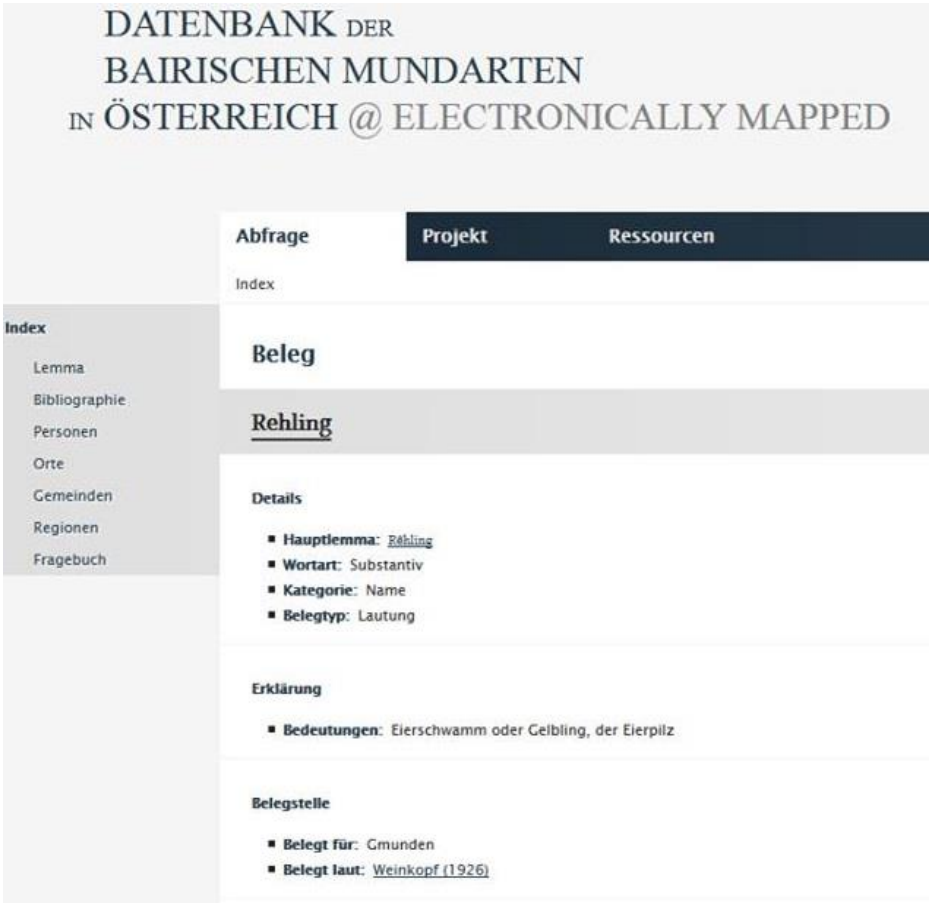
DINAMLEX

Suche nach bellis perennis
Anzahl der Belege: 151

id	katalog lade bereich	quelle	beleg	bedeutung	orig. anmerkung
23275	pflnk	WBÖ	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23276	pflnk	Cat.	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23277	pflnk	Marzell	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23278	pflnk	Flachgau Sa.	Monatsblümchen	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23279	pflnk	Waldviertel NÖ	Gefpllllte Gartenrockal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23280	pflnk	Waldviertel NÖ	Gensbleamal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23281	pflnk	Hall Tir.	Schweizerlan	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23282	pflnk	Knittelfeld Stmk.	Monatsröserl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23283	pflnk	Stmk.	Jägerblumel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23284	pflnk	Stmk.	Monatblümel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23285	pflnk	Stmk.	Ruckerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23286	pflnk	Ennstal Stmk.	Mannerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23287	pflnk	Stmk.	Saubleaml	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23288	pflnk	Mürztal, Wechsel Stmk.	Saublümel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23289	pflnk	NÖ	Angerrösal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23290	pflnk	NÖ	Gensbleamln	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23291	pflnk	NÖ	Goldbleamel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23292	pflnk	Nö	Monatsbleaml	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23293	pflnk	NÖ	Rokal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23294	pflnk	NÖ	Rukerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	

Datenbank der DBÖ

2) Umgekehrt ist es für die Bestimmung einer Pflanze und deren Zuweisung im Zeit-Raum-Kontinuum sehr hilfreich für die Botaniker und Botanikerinnen, auf eine umfassende Sammlung wie die DBÖ zurückgreifen zu können: Daraus können Varianten für Volkssprache, Büchernamen und historische Taxonomien ebenso gewonnen werden wie Rückschlüsse auf Verwendung (Ethnobotanik) und auf Verbreitung (Georeferenzierung) getätigt werden.



Darstellung in der DBÖ

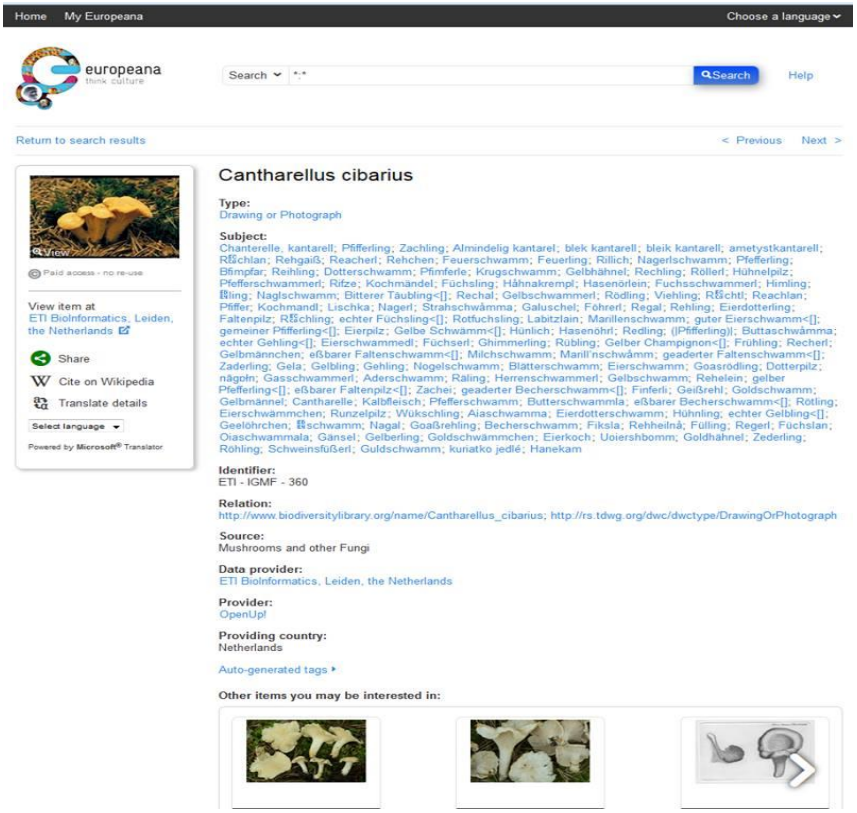
Um die Motivationen der beiden kollaborierenden Einheiten herzustellen, wurden folgende Schritte eingeleitet:

1) Entwicklung eines allgemeinen Schemas zur Modellierung von Pflanzennamen und zugehöriger Informationen.
2) Modellierung der Pflanzennamen in SKOS.
3) Datenaufbereitung in der DBÖ.

   Im Zuge der Datenaufbereitung übernimmt eine überregionale bzw. standardnahe Bezeichnung die Funktion des Hauptlemmas, unter welchem die jeweiligen regionalsprachlichen bzw. dialektalen Entsprechungen in Kombination mit den biologischen Informationen der botanischen Datenbank zusammengefasst werden.

Die dabei entstehenden Datensätze stellen eine sehr umfangreiche, wissenschaftlich fundierte Sammlung sprachlicher sowie sozio-kultureller Phänomene dar, womit sie als Quelle für verschiedenste wissenschaftliche Forschungsfragen herangezogen werden können. Diese bilden demnach die Grundlage für eine institutsübergreifende Zusammenarbeit und Verkettung von Daten, sowie deren Bereitstellung für die öffentliche Verwendung.

4) Entwicklung eines Webservices zur Kommunikation zwischen den existierenden Datenbanken zwecks abfragegesteuertem Datenaustauschs.
5) Publikation der Daten in Europeana im Kontext des Projekts OpenUp!



Darstellung in Europeana

Das Projekt BioLing wird im Kontext der COST Aktion IS1305 weitergeführt:
Durch Berücksichtigung weiterer lexikographischer Ressourcen im Webservice (und somit letztlich in Europeana) sollen Möglichkeiten geschaffen werden, europäisches Kulturgut besser zugänglich zu machen und Zusammenhänge in der Benennungsmotivik zu erarbeiten. Im Zuge dessen wird die interdisziplinäre Kommunikation zwischen den einzelnen wissenschaftlichen Instituten innerhalb Österreichs intensiviert und es wird dazu beitragen, dass es auf internationaler Ebene vermehrt zu Kooperationen kommt und länder-, instituts- und wissenschaftsübergreifende Projekte konzipiert werden.

Als Ausblick basierend auf der kollaborativen Zusammenarbeit und der bestehenden Infrastruktur: Über die Aufarbeitung der jeweiligen Etymologie einer

Pflanzennamenbezeichnung werden tieferliegende Zusammenhänge erarbeitet, die als Konzepte im Europäischen Kontext qualifizierbar und quantifizierbar gemacht werden sollen.

Im Kontext von DARIAH-EU wird auf Basis der etablierten modellhaften Zusammenarbeit zwischen ÖAW und NHM ein Beitrag traditionell lexikographischer europäischer Arbeiten zu Controlled Vocabularies erarbeitet.

Digital Humanities in der Hochschullehre – Erfahrungen aus dem Lern-Lehr-Projekt „Digitale Medien in den Geisteswissenschaften in Lehre und Forschung“

Patrick Pfeil, M.A. (Alte Geschichte, Universität Leipzig), Sabrina Herbst, M.A. (Medienzentrum, TU Dresden), Corina Willkommen, B.A. (Alte Geschichte, Universität Leipzig)

Die zunehmende Bedeutung der Digital Humanities für die Geisteswissenschaften erfordert auch die Vermittlung neuer Kompetenzen an Studierende und damit die Konzeption neuartiger Lehr- und Lernangebote. In diesem Zusammenhang ist das vom BMBF geförderte und vom Projektverbund „Lehrpraxis im Transfer“ betreute Lern-Lehr-Projekt zu sehen (https://www.hds.uni-leipzig.de/index.php?id=projektkohorte-3). Am Vorhaben beteiligt sind die Alte Geschichte der Universität Leipzig (Prof. Charlotte Schubert), die Korpuslinguistik der TU Desden (Prof. Joachim Scharloth) und das Medienzentrum der TU Dresden (Prof. Thomas Köhler). Ziel des Projekts ist die Einbindung von Lehrangeboten der Digital Humanities (konkret hier die Projekte eAQUA, eComparatio und Papyrusportal Deutschland) in die Regellehre des Bachelor Studienganges Geschichte und des Master Studienganges Klassische Antike der Universität Leipzig sowie die Entwicklung von Selbstlernmodulen zur Einführung in die Korpuslinguistik an der TU Dresden.

Im Vortrag wird zunächst das Vorhaben im Einzelnen vorgestellt. Im Anschluss werden die gemachten Erfahrungen mit der Einbindung in die Regellehre aus Sicht der Dozierenden dargestellt. Hierbei wird schwerpunktmäßig mit den im Wintersemester 2014/15 abgehaltenen zwei Digital-Humanities-Modulen an der Universität Leipzig gearbeitet. Auf das Selbstlernmodul der TU Dresden wird schlaglichtartig eingegangen. Zum Abschluss des Vortrages steht die Sicht der Studierenden im Mittelpunkt der Ausführungen.

Durch die Integration der Digital Humanities in die Regellehre der verschiedenen Geisteswissenschaften ändert sich das Profil der Studiengänge. Dies bringt neue Anforderungen an Lehrende und Studierende mit sich. Es gilt zu fragen, wie heutige Studierende, die Angebote aus den Digital Humanities annehmen und für das eigene Studium nutzbar machen. Stellen diese dabei ein Zusatzangebot dar oder gehen die Möglichkeiten der Digital Humanities in das alltägliche Arbeitsgerüst der Studierenden ein, wie es früher beim Wörterbuch oder bei einer Grammatik der Fall war? Wie entwickelt man bei den Studierenden die Bereitschaft sich auf Angebote aus den Digital Humanities einzulassen und welche Methoden sind dabei anwendbar? Darüber hinaus ist von Interesse, in welcher Phase des Studiums man diese Angebote einbringen sollte und ob man durch Digital Humanities die Befähigung der Studierenden zum Forschenden Lernen besonders fördern kann.

In den letzten Jahren konnten an der Universität Leipzig mehrere Seminare im Bereich Digital Classics angeboten werden, die aufeinander aufbauen und seit vier Jahren durch dieselben DozentInnen veranstaltet werden. Die Erfahrungen der DozentInnen bezüglich der didaktischen Umsetzung konnten durch mehrere Förderprojekte verstetigt werden, mit dem Ziel, die sich entwickelnden digitalen Methoden dauerhaft in den Hochschulunterricht einzupflegen. Im Rahmen besagter Projekte konnten außerdem parallel zu den Lehrveranstaltungen weiterbildende Maßnahmen angeboten werden, wie Workshops zu Programmierung, Visualisierung und Digitaler Edition. Aufbauend auf diesen Erfahrungen stehen das derzeit durchgeführte Bachelorseminar zur „Einführung in die antike Numismatik“ und das Masterseminar „Zur kulturellen Praxis des Zitierens“ in der Tradition der Digital Classics-Seminare und profitiert nicht nur aus den Erfahrungen der DozentInnen, sondern auch durch die Studienarbeiten vorangegangener Seminare, die in einer eigenen

Publikationsreihe „eAQUA Working Papers“ erschienen sind (http://journals.ub.uni-heidelberg.de/index.php/eaqua-wp). Aufgrund dieser Arbeitsgrundlage haben die Studierenden die Möglichkeit auf die in den vorangegangenen Veranstaltungen entwickelten Fragestellungen, Lösungsstrategien, Fehleranalysen und methodischen Entwicklungen zurückzugreifen und sich neu zu orientieren.
Ziel des Projektes ist zum einen der selbstständige und praktische Umgang mit digitalen Tools, um alternativ und ergänzend zu den klassischen Fachmethoden Lösungsstrategien zur Bearbeitung historischer Fragestellungen zu entwickeln. Zum anderen steht das Konzept des Forschendes Lernens im Fokus der Seminare.
Die Einführung in die digitalen Tools des Programms eAQUA („Extraktion von strukturiertem Wissen aus antiken Quellen“ - http://www.eaqua.net/) „Kookkurrenzanalyse“, „Zitationsgraph“ und „Mental Maps“ erfolgt dabei durch die DozentInnen anhand fachspezifischer Fragestellungen. In einem weiteren Schritt erlernen Studierenden die praktische Handhabung der Tools mittels eigens dafür erstellter Übungshandbücher im iBook-Format. Vor allem der praktische Umgang mit den vorgestellten Tools unter Einbeziehung eigener Fragestellungen hat sich als überaus erfolgreich erwiesen, als es darum ging, die Studierenden zur aktiven Mitarbeit anzuregen. In diesem Fall konnten die Studierenden als „ExpertInnen“ eines bereits behandelten Themas und mit dem Wissen um das Ergebnis ihrer Fragestellung, selbige durch das Tool verifizieren lassen. Die Erwartungshaltung der Studierenden konnte durch Einsatz digitaler Tools bestätigt und erheblich ergänzt werden. Die Vorteile des Einsatzes von digitalen Hilfsmitteln zeigten sich besonders deutlich im Vergleich verschiedener Arbeitsmethoden, die zur Bearbeitung wissenschaftlicher Fragestellungen herangezogen worden. Die Studierenden entwickeln in einer teils autodidaktischen Atmosphäre nicht nur fachliche Kompetenzen, sondern konnten auch gruppendynamisch in einen Diskurs treten und damit soziale Kompetenzen erwerben. Die Ergebnisanalyse wird direkt im Unterricht von den KommilitonInnen in erster und nachfolgend von den DozentInnen in zweiter Instanz vorgenommen.

Im zweiten Teil des Vortrags wird die Perspektive der Studierenden betrachtet. Dabei gilt es verschiedenste Herausforderungen zu überwinden: Dies ist zum einen die vielfach diskutierte Diskrepanz zwischen dem Nutzungsverhalten Neuer Medien der Generation der sog. „Digital Natives“ (Prensky 2001; 2001a - zur kritischen Auseinandersetzung mit dem Begriff der Digital Natives u. a.: Arnold & Weber 2013) und ihrem Einsatz Neuer Medien für das Studium (vgl. hierzu Weller et al. 2014). Zum anderen müssen Lehrangebote konzipiert werden vor dem Hintergrund einer hohen Diversität der Zielgruppe, hinsichtlich fachlicher Hintergründe und Motivation der Studierenden (Schwerpunktmodul oder Wahlpflichtbereich), Studiengang (BA oder MA) und Fachsemester. Es ist daher in unterschiedlicher Hinsicht von Bedeutung, bei der Konzeption von Lehrangeboten unter Einbezug digitaler Technologien die unterschiedlichen Bedürfnisse der Studierenden in den Blick zu nehmen. Nicht zuletzt handelt es sich bei der Einführung von Digital Humanities-Lehrangeboten um eine Lerninnovation (Kerres 2013) in der geisteswissenschaftlichen Lehre, bei deren Einführung besondere Akzeptanz durch die studentische Zielgruppe notwendig ist, um Lernerfolge zu erzielen und eine Verstetigung zu ermöglichen. Bei der Planung von Lehrangeboten bietet sich daher im Vorfeld die Durchführung einer Anforderungsanalyse an, um möglichst viel über die Zielgruppe der Lernenden und die das Lernen beeinflussenden Rahmenbedingungen herauszufinden. Dabei geht es bei der Konzeption eines Digital Humanities-Lehrangebotes vor allem um unterschiedliche Nutzungsgewohnheiten, Einstellungen und Kompetenzen hinsichtlich Neuer Medien bei den Studierenden, ihre Erfahrungen mit unterschiedlichen Lehr- und Lernformaten sowie strukturelle Einflussfaktoren, wie Studiengang, Modulform, Fachsemester zu identifizieren. Für das im Lehr-Lern-Projekt „Neue Medien in den Geisteswissenschaften in Lehre und Forschung“ zu entwickelnde Lehrangebot im Bereich der

Digital Humanities wurden im Mai 2014 gemeinsam mit den Studierenden Anforderungen und Praktiken des Einsatzes Neuer Medien in den Geisteswissenschaften erhoben. Eine ebenso wichtige Rolle wie die Durchführung einer Anforderungsanalyse spielt außerdem die Evaluierung des Lehrangebotes im Nachgang, so dass das im Wintersemester 2014/2015 erprobte Lehrangebot an der Universität Leipzig daher auch Ende 2014/ Anfang 2015 evaluiert wird.
Sowohl für die Anforderungsanalyse im Vorfeld als auch für die Evaluation wurde die Fokusgruppe als Erhebungsinstrument der empirischen Sozialwissenschaft gewählt. Fokusgruppeninterviews eignen sich aufgrund der breiten kollektiven Wissensbasis der TeilnehmerInnen besonders, um unterschiedliche Facetten einer Problemstellung zu erheben (vgl. Schulz 2010).
Diese erste Fokusgruppe (Anforderungsanalyse) wurde mit 11 Studierenden des Seminars „Digitale Altertumswissenschaft“ im vergangenen Sommersemester an der Universität Leipzig durchgeführt. Hierfür wurden die Studierenden einerseits zu den generellen Rahmenbedingungen ihres Lernens im Studium, andererseits zu ihren Erfahrungen mit der Seminarstruktur des Seminars „Digitale Altertumswissenschaft“ sowie den dort vorgestellten Digital Humanities-Werkzeugen „Perseus-Datenbank“ (www.perseus.tufts.edu) und den eAqua-Tools „Kookkurenzanalyse“, „Zitationsgraph“ und „Mental Maps“ befragt. Die Abschlussphase des Interviews diente dazu, die Bereitschaft der Studierenden, sich weitere Kompetenzen im Bereich der Digital Humanities anzueignen, auszuloten und Verbesserungsvorschläge für die Vermittlung von Inhalten in der Lehre zu erhalten. Die so erhobenen Anforderungen wurden dann für die Entwicklung des Lehrangebotes genutzt. Dabei hat sich unter anderem gezeigt, dass die Studierenden zwar an dem Erwerb von Kompetenzen im Bereich der Digital Humanities interessiert sind, jedoch über ein nur sehr rudimentäres Verständnis des Begriffs der Digital Humanities verfügen. Es wurde ebenfalls deutlich wie wichtig eine Einführung in die verschiedenen Methoden der Digital Humanities ist und welche Rolle bestimmte Lehrformate dabei spielen. Die Durchführung des zweiten Fokusgruppeninterviews zur Evaluation des im Wintersemester 2014/2015 durchgeführten Lehrangebots an der Universität Leipzig ist für Ende 2014/ Anfang 2015 geplant, um zu erheben, wie gut sich dieses in der Praxis bewährt hat.
Im Vortrag werden die Ergebnisse beider Fokusgruppeninterviews vorgestellt und die daraus abgeleiteten Handlungsempfehlungen für die Erarbeitung des Lehrangebots präsentiert. Darüber hinaus sollen die Ergebnisse der Usability-Untersuchung des Lehrangebots Ende 2014 dargestellt und die Frage diskutiert werden, inwieweit eine Berücksichtigung der Anforderungen der Studierenden gelungen ist.

Literatur**:**

Arnold, P. & Weber, U. (2013): Die „Netzgeneration“. Empirische Untersuchungen zur Mediennutzung bei Jugendlichen. In: M. Ebner & S. Schön (Hrsg.), L3T. Lehrbuch für Lernen und Lehren mit Technologien. Online-Dokument: http://l3t.eu/homepage/das-buch/ebook-2013/kapitel/o/id/144/name/die-netzgeneration (06.11.2014)

Kerres, M. (2013): Mediendidaktik, Konzeption und Entwicklung mediengestützter Lernangebote. München, Oldenbourg.

Prensky, M. (2001): Digital Natives, Digital Immigrants. In: On the Horizon 9,5 (2001). Online-Dokument: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf (10.11.2014)

Prensky, M. (2001a) Digital Natives, Digital Immigrants. Part II. Do They Really Think Differently? In: On the Horizon 9,6 (2001). Online-Dokument: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part2.pdf (10.11.2014)

Schulz, M. (2012): Quick and easy?! Fokusgruppen in der angewandten Sozialwissenschaft. In: M. Schulz, B. Mack & O. Renn, Fokusgruppen in der empirischen Sozialwissenschaft. Von der Konzeption bis zur Auswertung. Wiesbaden, S. 9-22.

Ansprechpartner:
Patrick Pfeil, M.A.
Universität Leipzig, Fakultät für Geschichte, Kunst und Orientwissenschaften, Historisches Seminar, Lehrstuhl für Alte Geschichte, Projektleiter Lern-Lehr-Projekt „Digitale Medien in den Geisteswissenschaften in Lehre und Forschung“
GWZ, Beethovenstraße 15, 04107 Leipzig (Raum: H4 2.16)
Tel.: +49 341 9737077, Fax: +49 341 9737071
Email: ppfeil@uni-leipzig.de

Posterpräsentation

Jahrestagung DHd 2015

eComparatio

Editionsvergleich

Oliver Bräckel, Hannes Kahl, Friedrich Meins, Charlotte Schubert

Das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt eComparatio wird seit 2014 als Kooperationsprojekt des Lehrstuhls für Alte Geschichte der Universität Leipzig und des ICE (Interdisciplinary Center of E-Humanities in History and Social Sciences/ Forschungsstelle am Max-Weber-Kolleg für kultur- und sozialwissenschaftliche Studien an der Universität Erfurt) entwickelt. Das Ziel des Projektes ist es, eine modular aufgebaute Anwendung zu entwickeln, die es ermöglicht, verschiedene Versionen eines Textes (aus Handschriften, gedruckten oder digitalen Texteditionen) miteinander zu vergleichen. Das Kernstück der Anwendung ist ein Modul zum Vergleich von Textausgaben, das auch die Erstellung eines Variantenapparates für digitale Editionen antiker Autoren ermöglicht. Die Zahl der Vergleichstexte ist beliebig, ebenso das Eingabeformat (TXT, HTML, XML, JSON, PDF). Die Anwendung wird frei skalierbar sein, so dass der Umfang der zu vergleichenden Texte nicht beschränkt ist, das Ergebnis (Kollationierung) soll in Form von Listen als kritischer Apparat (positiver oder negativer Apparat) oder auch in beliebiger anderer Form ausgegeben werden können. In einem weiteren Modul soll für Autorenreferenzen bei der Abfrage von online-Datenbanken die Anbindung an das Referenzsystem CTS (Canonical Text Services) und die Referenz auf Images von Handschriften (über das Image Citation Tool der CITE Collection Services) ermöglicht werden. Die Ansprechbarkeit für weitere Adressschemata wird ebenfalls implementiert (z.B. für JSON und den im Aufbau befindlichen PID-Service von CLARIN-D).
Im bisherigen Verlauf des Projektes ist es gelungen, die Grundfunktionen des Tools zu implementieren und es in die Lage zu versetzen eine beliebig große Anzahl an Texten miteinander zu vergleichen. Dabei sind drei unterschiedliche Ansichten entstanden, die es dem Benutzer ermöglichen das Ergebnis aus verschiedenen Perspektiven zu betrachten. Die Detailansicht zeigt einen Text und markiert entsprechende Unterschiede zu anderen Texten. Die Parallelansicht (siehe auch Abbildung) zeigt alle Texte nebeneinander und markiert die Unterschiede farbig. Die Buchansicht schließlich zeigt wieder nur einen Text an und visualisiert

die Varianten im Stile traditioneller Printeditionen unter dem betreffenden Abschnitt. Zu betonen ist dabei, dass der Ausgangstext für den Vergleich bei jeder Ansicht frei wählbar ist und sich somit nicht auf einen zu bevorzugenden Haupttext festgelegt bzw. eine Gewichtung der Textzeugen vorgenommen wird.

Die Visualisierung und Ergebnissicherung ermöglicht zum einen, einen schnellen Überblick über die Text- und Editionsgeschichte verschiedener in digitalisierter Form vorliegender Werke zu erlangen. Darüber hinaus eignet sich das Tool als Hilfsmittel zum Kollationieren bei der Erstellung beliebiger kritischer, historischer bzw. genetischer Editionen.

Weitere Funktionen, die das Spektrum von eComparatio noch einmal entscheidend erweitern werden, sind in Entwicklung. So ist die Einbindung von hochauflösenden Images der Handschriften der betreffenden Editionen geplant, um auch diesen Abschnitt der Textgeschichte dem Nutzer zugänglich zu machen. Weiterhin ist ein weiteres Modul in Entwicklung, das für die Abfrage von online-Datenbanken die Anbindung an das Notationssystem CTS (Canocical Text Services) ermöglicht. Beide Erweiterungen des Tools werden in absehbarer Zeit implementiert werden.

Nach seiner Fertigstellung soll das Tool als freier Webservice für Forschung und Lehre zur Verfügung gestellt werden. Davon können Handschriften-Digitalisierungsprojekte, Editionsprojekte sowie Projekte profitieren, die sich Spezialfragen einzelner Textpassagen widmen; es ist auch für Seminararbeiten, d.h. den Einsatz in der Lehre geeignet, da es sowohl von Nicht-Editionsphilologen als auch von Editionsphilologen eingesetzt werden kann. Es ist natürlich auch nicht an den Fachbereich der Alten Geschichte gebunden, sondern kann in verschiedenen Bereichen der Textwissenschaften, unabhängig von der Sprache, eingesetzt werden.

In der Fachcommunity der E-Humanities im Speziellen kann das Tool darüber hinaus in einem Bereich angewandt werden, der in jüngerer Zeit vermehrt ins Zentrum der Aufmerksamkeit gerückt ist, nämlich bei der Qualitätssicherung der digitalen Datengrundlage an sich. Gerade im Falle der Altertumswissenschaften, in denen bereits früh umfangreiche, abgeschlossene Korpora (TLG, BTL u.a.) vorlagen, ist ein nächster Schritt ein Ausbau dieser Datengrundlagen in die Tiefe, d.h. hinsichtlich der zahlreichen verschiedenen Editionen und Textausgaben. Solche Varianten spielen in der herkömmlichen altertumswissenschaftlichen Diskussion oftmals eine zentrale Rolle bei der Erörterung fachwissenschaftlicher Fragestellungen; die Möglichkeit, solche Varianten im Falle auch großer Textmengen schnell zu überblicken, kann als eine wesentliche Grundlage dafür gesehen werden, auch auf „klassischem“ Textmining basierende Untersuchungen mit einer besseren Datengrundlage zu versehen.

Da es sich bei dem Tool in erster Linie um ein Mittel zur Visualisierung handelt, ist es in hohem Maße für die Präsentation in Form eines Posters geeignet. Geplant ist die Darstellung des gesamten Workflows anhand eines Beispiels, von der Eingabe unstrukturierter Textdokumente bis hin zu den drei oben genannten Visualisierungsformen.

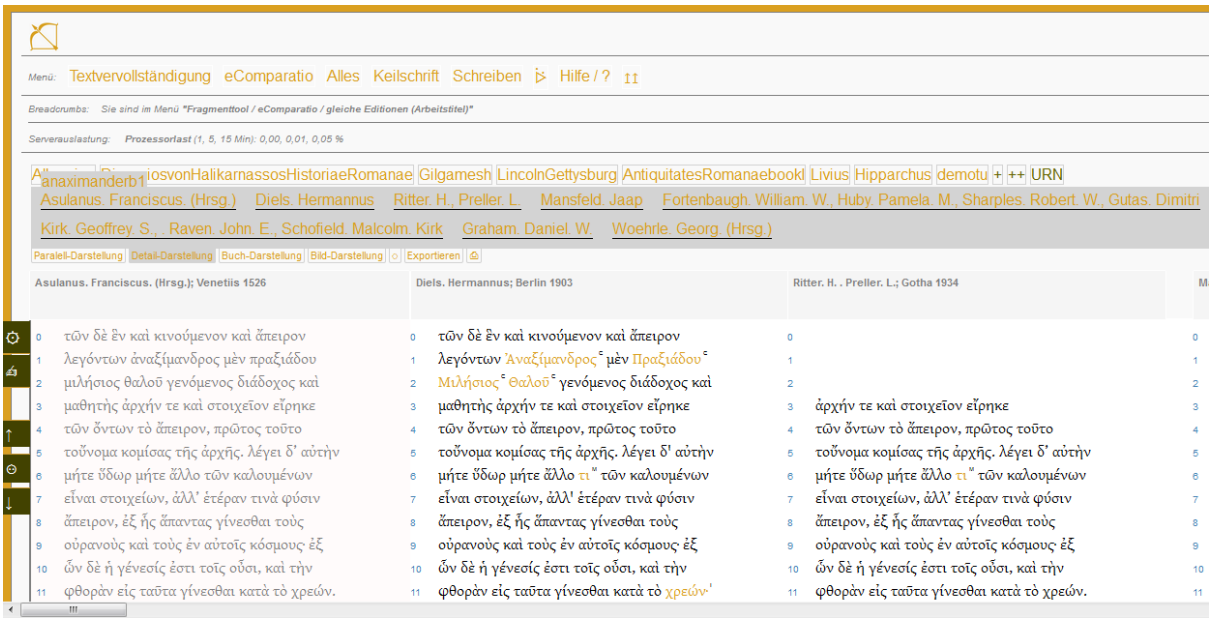


Abb. der Parallelansicht von eComparatio am Beispiel des Fragments B1 des Anaximander.

Kontakt:
Prof. Dr. Charlotte Schubert
Historisches Seminar
Lehrstuhl für Alte Geschichte
Beethovenstraße 15
04107 Leipzig
Raum 3.204
Telefon: +49 341 97 37071
Email: schubert@rz.uni-leipzig.de

»So viele Briefe mit all ihrem Für und Wider...« Die kommentierte Online-Edition des Gesamtbriefwechsels Ludwig von Fickers als wissenschaftlicher Quellenfundus

Markus Ender
Forschungsinstitut Brenner-Archiv
Universität Innsbruck

Ludwig von Ficker (1880–1967) erlangte als Entdecker und Förderer Georg Trakls und als Herausgeber der Zeitschrift »Der Brenner« (1910–1954) einige Bekanntheit; daneben betätigte er sich als Inhaber des Brenner-Verlags, als Literaturkritiker, Juror und Organisator von Lesungen. Aufgrund seiner vielen Tätigkeiten ergaben sich enge briefliche Kontakte mit Personen aus Politik und Kultur, so z.B. mit Else Lasker-Schüler, Martin Heidegger, Karl Kraus, Rainer Maria Rilke oder Ludwig Wittgenstein. Sein Briefwechsel, von dem im Innsbrucker Forschungsinstitut Brenner-Archiv mehr als 16.500 Korrespondenzstücke von über 2200 AdressatInnen erhalten sind, markiert und dokumentiert einen Teil der deutschsprachigen Kulturgeschichte und bietet Forscherinnen und Forschern wie auch interessierten Laien Einblicke in das Geistesleben der ersten Hälfte des 20. Jahrhunderts. Zwischen 1988 und 1996 erschien in vier Bänden eine Auswahl von 1300 Briefen von und an den »Brenner«-Herausgeber.[1]

Im Rahmen des FWF-Projektes »Ludwig von Ficker als Kulturvermittler« (P24283) entsteht seit April 2012 am Brenner-Archiv eine digitale Ausgabe des Briefwechsels Ludwig von Fickers in Form einer kommentierten Online-Edition. Diese Edition versteht sich nicht als bloße Retrokonversion der bereits gedruckten, vierbändigen Auswahlausgabe von Fickers Briefen, sondern als eigenständige Neu-Edition, die in ihrer Konzeption, der methodischen Durchführung, in ihrer Darstellungsform und in ihrem intendierten Gebrauchswert in wesentlichen Punkten von der früheren Buchausgabe abweichen wird.
Die digitale Edition der Fickerschen Korrespondenz wird dabei in mehrfacher Hinsicht als eine integrale Schnittstelle zwischen den Beständen im Brenner-Archiv und den RezipientInnen dienen; im Sinne des Tagungsthemas möchte ich in meinem Beitrag die Leistungsfähigkeit einer solchen Editionsform aufzeigen. Es soll am Beispiel der kommentierten Online-Edition des Briefwechsels Ludwig von Fickers demonstriert werden, dass sie sich sowohl als Medium für die zukünftige Generierung von Wissen als auch für die nachhaltige Nutzung von Daten eignen kann. Diesbezüglich lassen sich drei große Bereiche anführen:

- Zum einen bietet die Form der digitalen Internet-Edition auf *quantitativer* Ebene die Möglichkeit, erstmalig den gesamten Archivbestand im Nachlass Ludwig von Fickers (also Briefe und Gegenbriefe) zugänglich zu machen.[2] Die Online-Edition betrachtet die vorliegenden Briefwechsel (im Sinne Foucaults) wertfrei als Summe von Aussagen, die zu einem bestimmten Zeitpunkt möglich waren, wobei nicht zwischen »wichtigen« und »unwichtigen« BriefpartnerInnen bzw. Briefen unterschieden wird. Die Breite der vorliegenden Daten ermöglicht einen neuen Blick auf das Gesamtkorpus, der in dieser Form bislang nicht möglich gewesen ist.
- Zum anderen spricht ein solch umfangreiches Briefkonvolut, das einen gewichtigen Baustein im kulturellen Erbe darstellt, auf *qualitativer* Ebene durch die Veröffentlichung im Rahmen eines methodisch kontrollierten und dokumentierten

[1] Ludwig von Ficker: Briefe. 4 Bde. Hg. von Franz Seyr, Walter Methlagl u. a. Salzburg; Innsbruck 1988–1996.
[2] Diesem Ansatz wird insofern Rechnung getragen, als dass bereits 14000 Briefe als Transkripte vorliegen.

Editionsprojekts ein breites Spektrum von InteressentInnen an. Die Daten werden über das Internet sowohl einer interessierten Öffentlichkeit als auch der wissenschaftlichen Forschung zugänglich gemacht; der Briefwechsel dürfte dabei aufgrund der inhaltlichen Diversität, der erschließenden Kommentierung sowie der Möglichkeit, über spezifizierte Suchfunktionen personelle und thematische Netzwerkstrukturen auszumachen, nicht nur für Literaturwissenschaftler, sondern für verschiedene Fachrichtungen (so z.B. Soziologie, Geschichtswissenschaften oder Theologie) von beträchtlichem Interesse sein.

- Überdies kommen bei der Erstellung der Edition etablierte digitale Standards (so z. B. das XML-Dateiformat oder die TEI-Codierung) zur Anwendung. Dadurch ist in Folge auch für die Institution Archiv ein Zusatznutzen gewährleistet, denn es kann durch die Digitalisierung der Korrespondenz eine seiner Hauptaufgaben, die nachhaltige Langzeitarchivierung der Bestände, wahrnehmen. Die den Transkripten zugrunde liegende XML-Datenstruktur garantiert bei zukünftigen Bearbeitungen die volle Verfügbarkeit der editorischen Kerndaten (Brieftranskripte) sowie der darauf aufbauenden Metadaten (inklusive Kommentar etc.). Die dem Editionsprojekt zugrundeliegende Open-Access-Policy und die geplante Anbindungen an die vom Austrian Academy Corpus besorgte Online-Version des »Brenner« sowie an die Gemeinsame Normdatei der Deutschen Nationalbibliothek erweitern das mögliche Nutzungsspektrum der Edition.

Kontakt:

Mag. Markus Ender
Forschungsinstitut Brenner-Archiv
Universität Innsbruck
Josef-Hirn-Straße 5-7
A-6020 Innsbruck
Tel. +43 512 507 45022
E-Mail: Markus.Ender@uibk.ac.at
http://www.uibk.ac.at/brenner-archiv/projekte/lfickeralskulturvermittler/

Geometrische Verfahren als Brücke zwischen Text und Objekt

Hubert Mara und Bartosz Bogacz

Universität Heidelberg
IWR – Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
FCGL - Forensic Computational Geometry Laboratory
Im Neuenheimer Feld 368, 69120 Heidelberg, Deutschland
hubert.mara@iwr.uni-heidelberg.de

Keilschrifttafeln gehören zu den ältesten Textzeugen, die im Umfang mit den Texten in lateinischer und altgriechischer Sprache vergleichbar sind. Da diese Tafeln aus dem gesamten Alten Orient über beinahe viertausend Jahre in Verwendung waren [Sod94], lassen sich damit viele interessante Fragestellungen zur Entwicklung von Religion, Politik, Wissenschaft, Handel bis hin zu Klimaveränderungen [Kan13] beantworten. Die aus Ton geformten Tafeln, bei denen Zeichen [Bor10] als keilförmige Abdrücke mit einem eckigen Stylus eingedrückt wurden, erfordern neue informationstechnische Methoden zu der Dokumentation und Analyse als die in Archiven üblichen Flachwaren. Zusätzlich gibt es kaum Verfahren aus dem Bereich der *Optical Character Recognition* (OCR), die für Sprachen in Keilschrift zur Verfügung stehen [Spe81]. Der Arbeitsablauf in der Assyriologie von der Keilschrifttafel als dreidimensionales Objekt bis hin zur Darstellung als Text in einer modernen Sprache – üblicherweise Deutsch – ist in Abbildung 1 dargestellt. Dabei ist der manuelle Zeitaufwand als Kurve dargestellt. In der Zusammenarbeit mit der Heidelberger Assur-Forschungsstelle unter der Leitung von Prof. Stefan Maul konnte festgestellt werden, dass ein erheblicher Teil der Arbeit im Bereich der Identifikation und Extraktion von Zeichen liegt, der stark mit der Dokumentation als Zeichnung verknüpft ist.

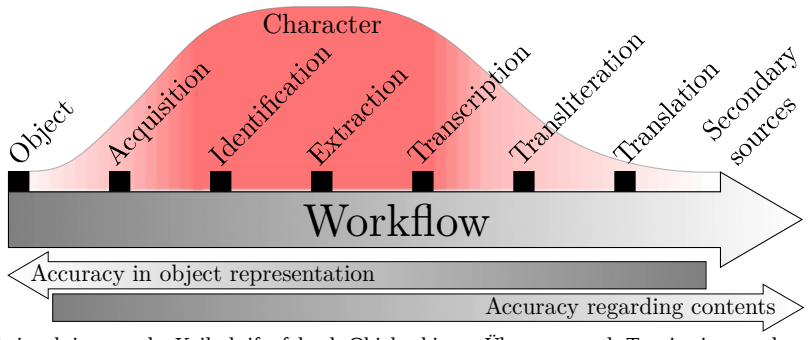


Abb. 1: Arbeitsschritte von der Keilschrifttafeln als Objekte bis zur Übersetzung als Text in einer modernen Sprache.

Das Digitalisieren von Keilschrifttafeln inspiriert durch die Open Data Initiative, wurde bereits vor einigen Jahren von der *Cuneiform Database Library Initiative* (CDLI) des Max Planck Institut für Wissenschaftsgeschichte und der *University of California at Los Angeles* begonnen und entsprechende Datenbanken entwickelt [GWLU05]. Dabei werden üblicherweise Photos und Bilder von Flachbettscannern eingesetzt, die zwar günstig und rasch zu erstellen sind, jedoch bei beschädigten oder gekrümmten Tafeln viele Bereiche unscharf und/oder verschattet sind. Daher werden in Jena, Würzburg [CMFW14] und Heidelberg [MKJB10] moderne 3D-Messgeräte eingesetzt um möglichst exakte digitale Repliken anzufertigen, mit denen entsprechende Visualisierungen berechnet werden. Als Mittel- und Fernziel sind digitale Werkzeuge im Sinne der OCR in Entwicklung.

Da die Datengrundlage keine regelmäßigen Gitter i.e. Rasterbilder wie die in *Digital Humanities* üblichen 2D-Digitalisate sind, sind Methoden notwendig, die aus der Geometrie eines 3D-Modells die Schriftzeichen extrahiert. Dafür kommen Integral Invariante Filter zum Einsatz, die mit Hilfe eines Mehr-Skalen Ansatzes die einzelnen Elemente der Keilschriftzeichen in einer Vektordarstellung extrahieren [MK13]. Mit Hilfe von einer minimalen Anzahl von Orts- und Richtungsvektoren werden mit dem *GigaMesh Software Framework* parametrischen Kurven (i.e. Splines) bestimmt, die als zweidimensionale XML-Dateien im offenen *Scalable*

Vector Graphics (SVG) Format exportiert werden. In Abbildung 3 wird eine schematisches SVG zur Darstellung eines Keils, mit einer minimalen Menge von vier Punkten repräsentiert wird.

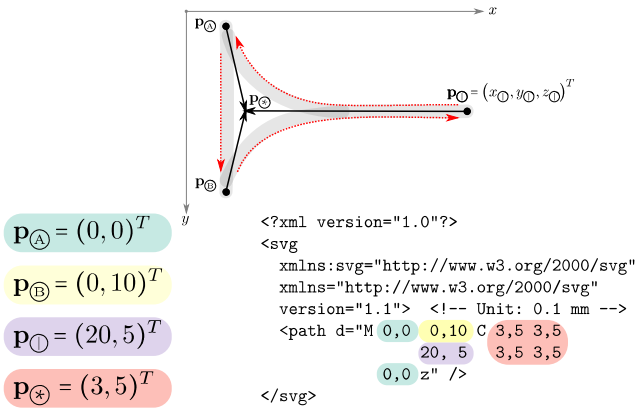


Abb. 2: Minimalbeispiel für einen Keil als XML/SVG-Datei.

Das selbe Format wird von proprietären Zeichenprogrammen und dem *Open Source* Pendant Inkscape verwendet, die beide in der Assyriologie und in der Grabungsdokumentation in der Archäologie zum Einsatz kommen. Damit ist automatisch sichergestellt, dass aus 3D-Modellen berechnete Zeichnungen kompatibel sind zu digitalen Handzeichnungen. Darüber hinaus bietet SVG – wie alle anderen – XML-Dateien die Möglichkeit zur automatischen und manuellen Annotation, wie es in den digitalen Textwissenschaften üblich ist. Abbildung 3 zeigt einen Vergleich zwischen einer digitalen Handzeichnung und einer berechneten Zeichnungen des zugehörigen 3D-Modells.

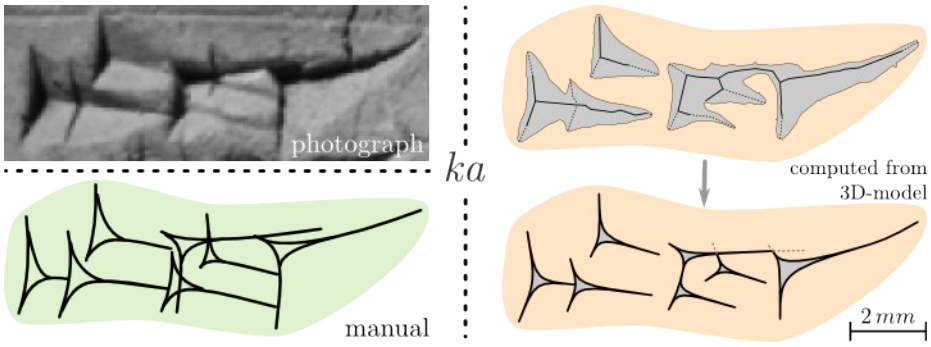


Abb 3:Das Zeichen „ka“ digital mit der Hand gezeichnet und aus einem 3D-Modell berechnet.

Die Vektordarstellung der Keilschriftzeichen sowie die komplexe zweidimensionale Anordnung von Keilabdrücken verhindert die Anwendung von gebräuchlichen OCR Methoden, die Zeichen in Rasterdarstellung [AGFV14] und aufeinander folgende Zeichen [RRF13] erwarten. Die Analyse von Keilschriftzeichen erfordert eine Transformation der SVG Daten in eine vereinfachte aber mathematisch handhabbare Repräsentation als mathematische Graphen mit Knoten und Kanten. Keilschriftzeichen identifizieren sich hauptsächlich durch die Lage und Position ihrer Keilabdrücke, eine Eigenschaft, die sich mit der Zerlegung des Graphen in Teilstrukturen, die den Keilabdrücken entsprechen, nutzen lässt. Der Teilabdruck als Teilstruktur in einem Graphen lässt sich einfach mit Richtungs- und Ortsvektoren beschreiben, die als Features genutzt werden, um Keilschriftzeichen auf Ähnlichkeit zu prüfen. Die vollständigen Graphen der Zeichen werden zudem genutzt, um Methoden aus dem Gebiet der Graphenähnlichkeit, wie den Graph-Kerneln und dem spektralen Embedding [BR10] anwenden. Die ist vor allem vorteilhaft bei komplexen bildhaften Zeichen, die sich nicht in Teilstrukturen von Features zerlegen lassen, aber trozdem auf Ähnlichkeit verglichen werden müssen.

Abbildung 4 zeigt Keilschriftzeichen dargestellt als Graphen, die aus einer SVG Datei extrahiert wurden, und gegenseitig auf Ähnlichkeit verglichen werden. Die Aufgabe besteht darin Keilschriftzeichen, die einem Zeichen (L: 19) ähneln, aufzufinden. Das erste Zeichen von Links ist der gesuchte Prototyp, alle darauf folgende Zeichen

sind die gefundenen Zeichen. Die Zeichen mit grünem Hintergrund wurden korrekt identifiziert, die Unähnlichkeit zum Prototyp wird mit „k:“ unterhalb des Zeichens betitelt. Alle fünf Zeichen, die in dem analysierten Dokument vorhanden waren und dem gesuchten Prototyp ähnelten, wurden erfolgreich gefunden.

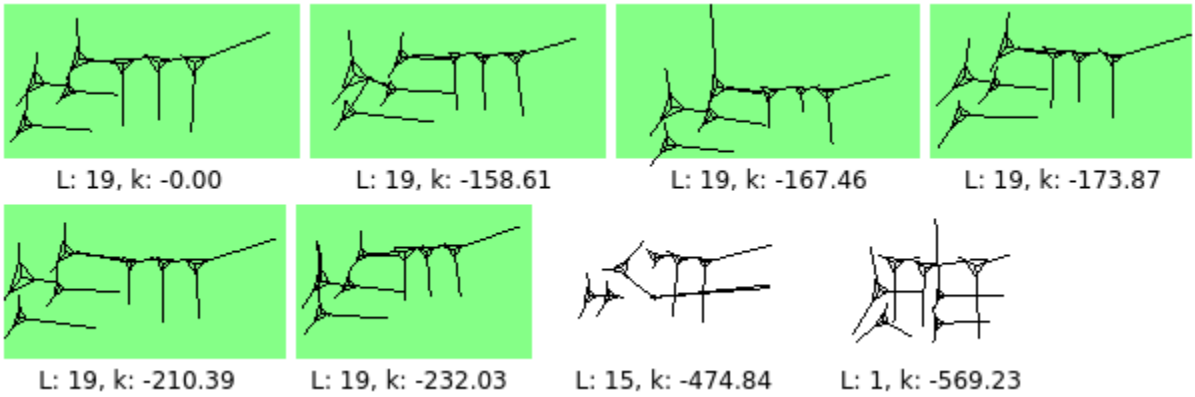


Abb. 4: Keilschriftzeichen in Graphenrepräsentation werden in einem SVG Dokument gesucht und auf Ähnlichkeit verglichen.

Zusammenfassend werden in diesem Beitrag Methoden aus der Geometrie und der Mustererkennung vorgestellt, die frei von lexikographischen / linguistischen Annahmen sind. Damit werden neue Zugänge zur Integration in OCR System für Handschriften geschaffen, die weit über die Anwendung an Keilschrift i.e. Handschrift in 3D hinaus gehen. Eine direkte Anwendung an mittelalterlichen Epitaphen hat bereits Aufnahme in entsprechende online Datenbanken gefunden [Krö12].

Literatur

[AGFV14] Segmentation-free word spotting with exemplar SVMs, J. Almazán, A. Gordo, A. Fornés, E. Valveny, Journal of Pattern Recognition, pp. 3967-3978, Elsevier, 2014.
[Bor10] R. Borger. Mesopotamisches Zeichenlexikon, volume 305 of Alter Orient und Altes Testament – Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments (AOAT). Ugarit-Verlag, 2. edition, 2010.
[BR10] Recent advances in graph-based pattern recognition with applications in document analysis, H. Bunke, K. Riesen, Journal of Pattern Recognition, pp. 1057-1067, Elsevier, 2010.
[CMFW14] M. Cammarosano, G. G.W. Müller, D. Fisseler and F. Weichert. *Schriftmetrologie des Keils: Dreidimensionale Analyse von Keileindrücken und Handschriften*, Die Welt des Orients, Ausgabe: 44.1, 2014.
[GWLU05] B. Groneberg, F. Weiersh#user, T. Linnemann, and D. Ullrich. Jahrbuch der Max-Planck-Gesellschaft, chapter Digitale Keilschriftbibliothek Lexikalischer Listen aus Assur. Gesellschaft für wissenschaftliche Datenverarbeitung mbH , Göttingen, Germany, 2005.
[Kan13] D. Kaniewski, E. Van Campo, J. Guiot, S. Le Burel, T. Otto and C. Baeteman. Environmental Roots of the Late Bronze Age Crisis. PLoS ONE 8(8), 2013.
[Krö12] S. Krömker, Kombinierte 3D-Datenaufbereitung von Schriftfeldern und Gelände des mittelalterlichen Jüdischen Friedhofs ‚Heiliger Sand’, in: Die SchUM-Gemeinden Speyer–Worms–Mainz. Auf dem Weg zum Welterbe. Band zur Internationalen Tagung der Generaldirektion Kulturelles Erbe Rheinland-Pfalz, angenommen, Mainz, Deutschland, 2012.
[MK13] H. Mara and S. Krömker. Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes. Proc. of 12. Int. Conference on Document Analysis and Recognition (ICDAR/IAPR), pp. 62–66, Washington, DC, USA, 2013.
[MKJB10] H. Mara, S. Krömker, S. Jakob and B. Breuckmann. GigaMesh and Gilgamesh - 3D Multiscale Integral Invariant Cuneiform Character Extraction. Proc. VAST Int. Symposium on Virtual Reality, Archaeology and Cultural Heritage, pp. 131-138, Palais du Louvre, Paris, France, 2010.
[RRF13] Bag-of-Features HMMs for segmentation-free word spotting in handwritten documents, L. Rothacker, M. Rusinol, G.A. Fink, Proc. of 12th International Conference on Document Analysis and Recognition, pp. 1305-1309, Washington, DC, USA, 2013.
[Sod94] W. von Soden. The ancient Orient: an introduction to the study of the ancient Near East. Wm. B. Eerdmans Publishing Co., 1994.
[Spe81] G. Sperl. Erkennen von Keilschriftzeichen mit Hilfe Elektronischer Rechenanlagen. PhD thesis, Leopold-Franzens-Universität Innsbruck, Innsbruck, Austria, 1981.

Vernetzte Datenstrukturen als Grundlage philosophischer Erkenntnisse
Technische Umsetzung und eine exemplarische Anwendung anhand des elektronischen Apparats der Wiener Ausgabe

Michael Nedo, Daniel Bruder, Pascal Zambito, Max Hadersbeck, Josef Rothhaupt

The Wittgenstein Project Clare Hall und der Ludwig Wittgenstein Trust, University of Cambridge
Centrum für Informations- und Sprachverarbeitung, in Zusammenarbeit mit dem Lehrstuhl II der Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, LMU München

1. Einleitung

Ludwig Wittgenstein, einer der bedeutensten Denker und Philosophen unserer Zeit, eignet sich aufgrund seiner Themen und seiner komplexen Arbeitsweise besonders gut, um die Vorteile eines elektronischen Apparats auf Grundlage vernetzter Datenstrukturen zu demonstrieren. Da bisher publizierte Editionen ohne solche Werkzeuge auskommen mussten, lassen sich durch geschickte Nutzung der digitalen Möglichkeiten neuartige philosophische Erkenntnisse gewinnen und alte Irrtümer ausräumen.

Auf unserem Poster wollen wir die zugrunde liegenden Datenstrukturen des Apparats erklären und in einer exemplarischen Anwendung zeigen, wie sich konkreter Nutzen aus seiner Anwendung ziehen lässt

2. Struktur von Wittgensteins Werk

Wittgensteins Werk zeichnet sich durch eine Vielzahl an internen und externen Verknüpfungen aus. Das Stemma in Abb. 1 zeigt u.a. die Entstehung der bei Suhrkamp publizierten *Philosophischen Bemerkungen (PB),* an der deutlich wird, dass die Genese des Textes sowohl editionsphilologisch als auch philosophisch relevant ist.

Die Entstehungsgeschichte der *PB* liefert Aufschluss darüber, wie Bemerkungen aus den Manuskripten Eingang in die maschinenschriftliche Synopse TS208 gefunden haben. Die Synopse wurde von Wittgenstein zerschnitten, in der Zettelsammlung (TS209) neu angeordnet und schließlich unter Hinzuziehung weitere Materials aus anderen Manuskripten von seinen Erben als Buch veröffentlicht.

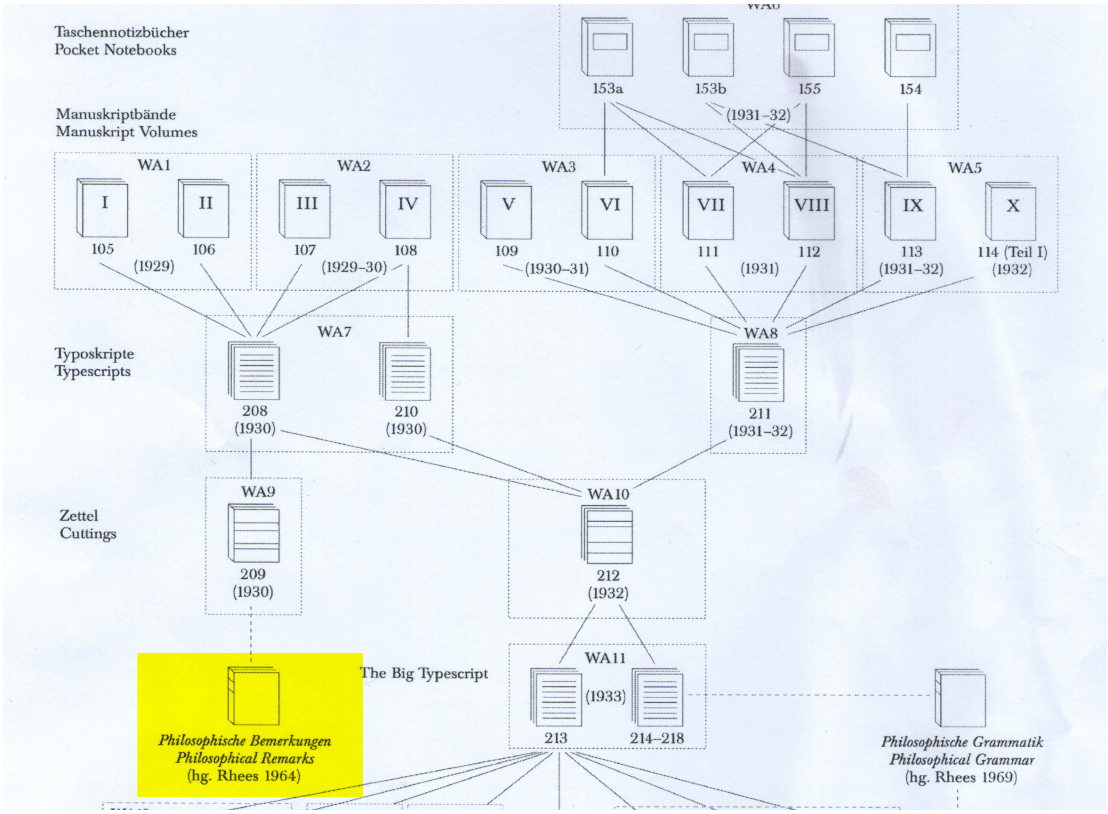


Abbildung 1: Stemma zur Genese der *Philosophischen Bemerkungen*

Da die meisten Interpreten von diesem Werk als fertiges Buch ausgingen, finden sich bei ihnen Missverständnisse, die durch eine angemessene Darstellung der intra- und intertextuellen Verknüpfungen vermeidbar sind. Der Apparat, an dem wir arbeiten, ist ein digitales Werkzeug, das mittels verschiedener Such- und Sortierungsfunktionen diese Interdependenzen dem Nutzer verständlich macht und ihm einen ebenso einfach zu handhabenden wie zuverlässigen Zugang zum Werk erlaubt.

3 Datenstruktur des Apparats

Die Nutzung des elektronischen Apparats beginnt in der Regel mit einer Wortsuche, welche über eine lemmatisierte Wortkonkordanz realisiert wird. Den weiteren Funktionen - Zugriff auf eine elektronische Realkonkordanz und interaktive Nutzung durch den User - liegt eine Matrix dynamisch vernetzter Objekte zugrunde.

3.1 Wortkonkordanz

Mithilfe des Computers lassen sich Suchfunktionen effizienter und benutzerfreundlicher realisieren als in den gedruckten Apparaten zur WIENER AUSGABE. Zur Wortsuche wird ein Vollformenlexikon genutzt, welches es ermöglicht, aus jeder flektierten Form eines Wortes eine lemmatisierte Form

abzuleiten und daraus wiederum sämtliche möglichen Formen des Wortparadigmas zu generieren und in die Suche miteinzuschließen. Indem das Suchergebnis also nicht nur wortidentische Treffer, sondern alle linguistisch verwandten Formen enthält, erlaubt es die Wortkonkordanz, das gesamte Korpus auf ein spezifisches Thema oder einen bestimmten Begriff hin „quer“ zu lesen.

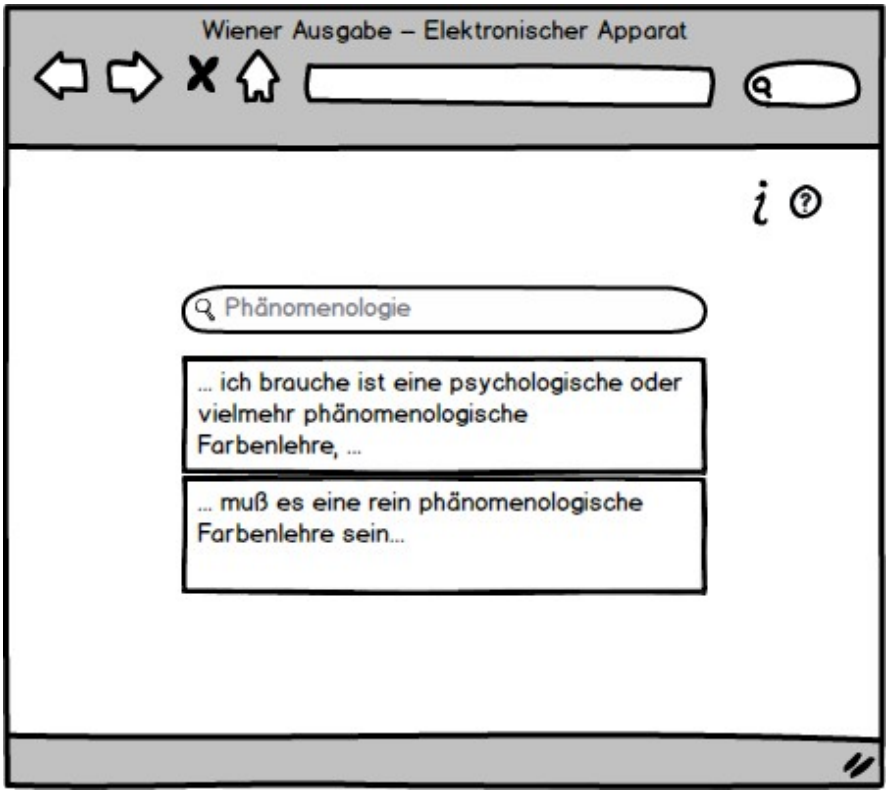


Abbildung 2: Eingabemaske für die Wortsuche

3.2 Realkonkordanz

Die Funktionsstruktur der Realkonkordanz basiert auf einer mit dem Werk dynamisch vernetzten Matrix. Auf der ersten Ebene werden die Objekte, die das Werk repräsentieren, aufgeführt: Wittgensteins Manuskripte, Typoskripte und Mitschriften sowie Diktate und deren jeweilige Veröffentlichungen. Die internen Strukturen der Objekte entsprechen der Nomenklatur der Wiener Ausgabe (vgl. Abb. 3, A); bei den posthumen Veröffentlichungen werden stattdessen die Strukturen der Herausgeber aufgeführt.

Entsprechend dieser Grundstruktur werden auf separaten Ebenen weitere Objekte erfasst, wie z.B. biographische Dokumente, Korrespondenzen, Bilder oder auch Sekundärliteratur und Übersetzungen (B).

Zusätzlich werden über eine interaktive Arbeitsplattform unter den Namen der jeweiligen Nutzer inhaltsbezogene Kommentare mit den Texten verknüpft sowie Fehler und vorgeschlagene Korrekturen in den Editionen (D).

Die Vernetzung der einzelnen Elemente der Objekte erfolgt auf darunter liegenden Ebenen der Matrix. Dabei werden drei Typen von Verknüpfungen unterschieden: ein-eindeutige, quasi festverdrahtete; provisorische, noch endgültig zu bestimmende; und offene, die noch recherchiert werden müssen (C).

Die Matrix ist über Such- und Sortierungswerkzeuge mit den Textdateien und deren Darstellungsformen auf dem Computerbildschirm verknüpft, die wiederum über die Nomenklatur der WIENER AUSGABE auf die gedruckte Edition verweisen.

Abbildung 3: Datenstrukturen der Matrix

4. Anwendung

Im Rahmen einer Masterarbeit an der LMU werden diese Datenstrukturen genutzt, um philosophische Erkenntnisse zu gewinnen. Konkret geht es um das Themenfeld der Phänomenologie bei Wittgenstein in den Jahren 1929-30, in denen der Begriff signifikante Bedeutungsveränderungen erfuhr. Mittels der Wortkonkordanz lassen sich zunächst alle Vorkommen des Begriffs im entsprechend zeitlich eingegrenzten Korpus finden (vgl. Abb. 2). Die einmal gefundene Textstelle lässt sich dann

1. im Rahmen des betreffenden Objektes um den Kontext erweitern, sodass benachbarte Stellen etwa in TS209 angezeigt werden
2. um frühere oder spätere Versionen der gleichen Bemerkung erweitern. Bemerkungen aus den TS209 lassen sich z.B. rückwärts bis zu den Manuskriptbänden oder vorwärts bis zur Publikation in den *PB* nachverfolgen. (s. Abb. 1)
3. um Hintergrundinformationen erweitern, die zu jeder Textstelle Faksimiles, Sekundärliteratur und weitere Daten enthalten können.

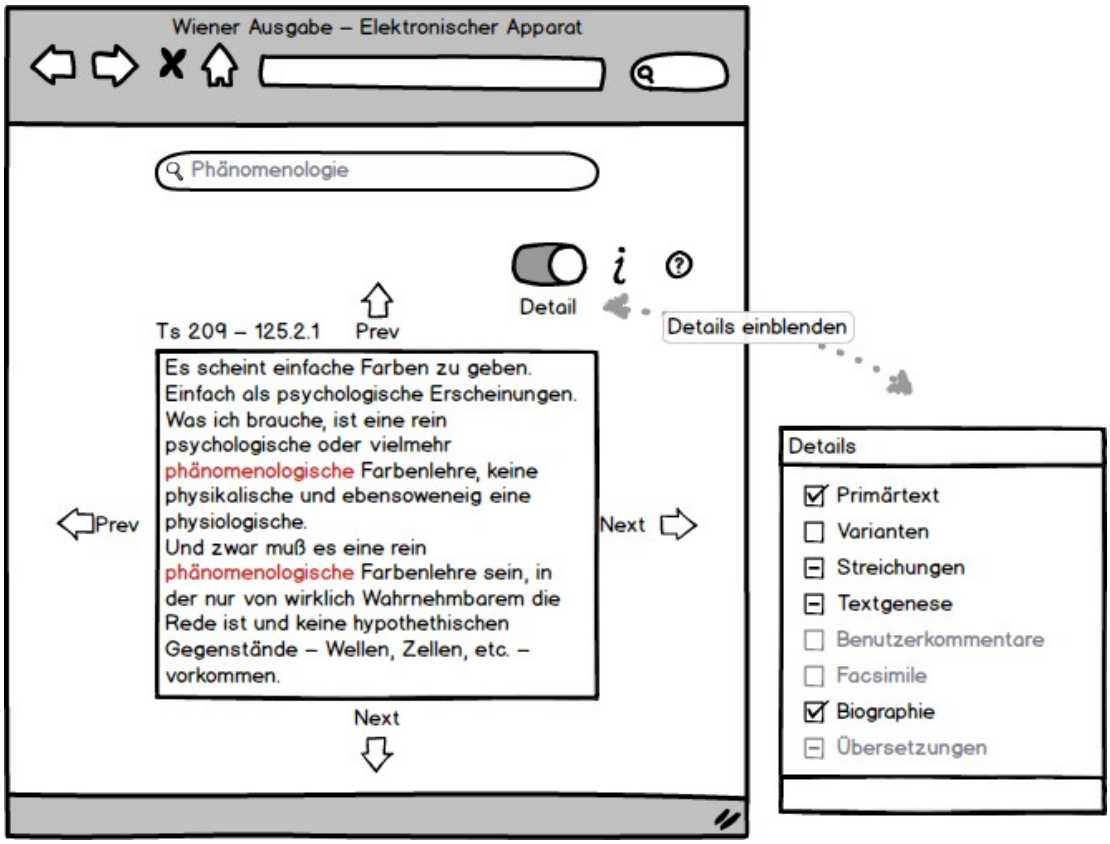


Abbildung 4: Die Bemerkung TS209,125.2.1/2.2 in der Darstellung des Apparats

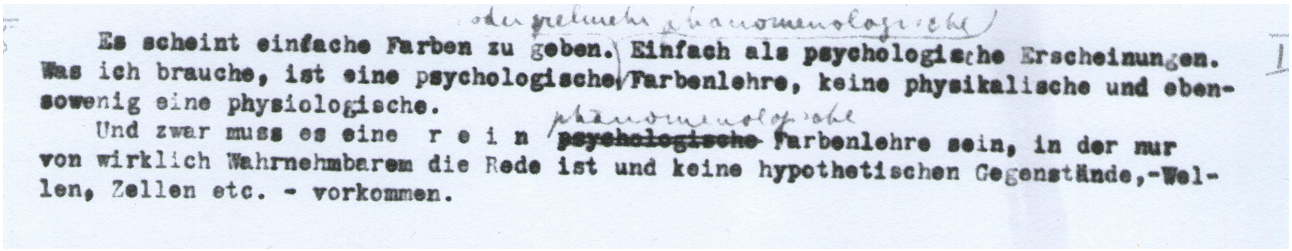

Es scheint einfache Farben zu geben. Einfach als psychologische Erscheinungen. Was ich brauche, ist eine psychologische Farbenlehre, keine physikalische und ebensowenig eine physiologische.
Und zwar muss es eine r e i n psychologische Farbenlehre sein, in der nur von wirklich Wahrnehmbarem die Rede ist und keine hypothetischen Gegenstände,-Wellen, Zellen etc. - vorkommen.

Abbildung 5: Faksimile derselben Bemerkung: man sieht, dass das Wort "phänomenologische" erst nachträglich eingefügt wurde und könnte nun frühere Textstufen untersuchen, um die Begriffsentwicklung zu erforschen.

Auf Grundlage dieses Wissens lassen sich gängige Missverständnisse in der Sekundärliteratur ausräumen, der die Genese der PB nicht in diesem Ausmaß zugänglich war. Schließlich lassen sich die neu gewonnenen Erkenntnisse verknüpft mit dem Namen des Verfassers als Kommentar zurück in die Matrix einspeisen.

5. Anhang- Technische Details

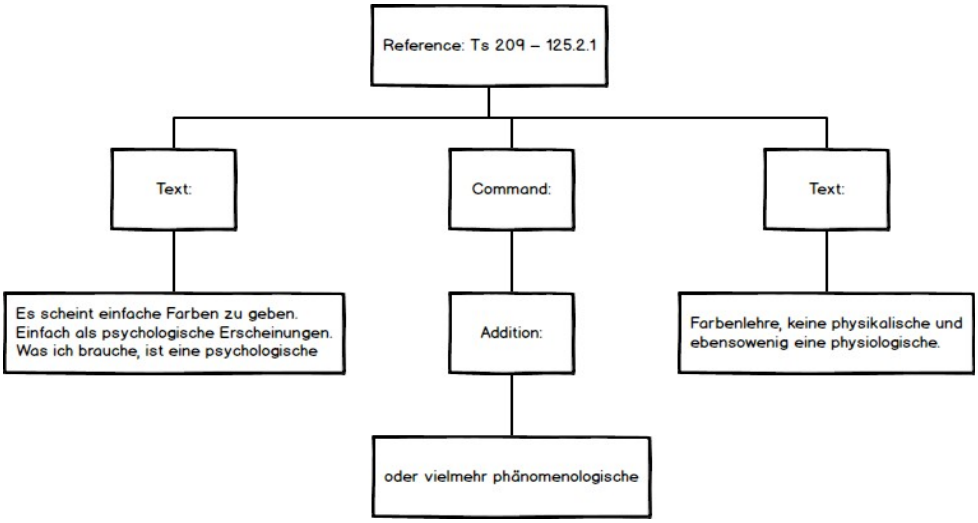


Abbildung 6: Abschnitt eines Abstract Syntax Tree (AST) für Bem. TS209, 125.2.1

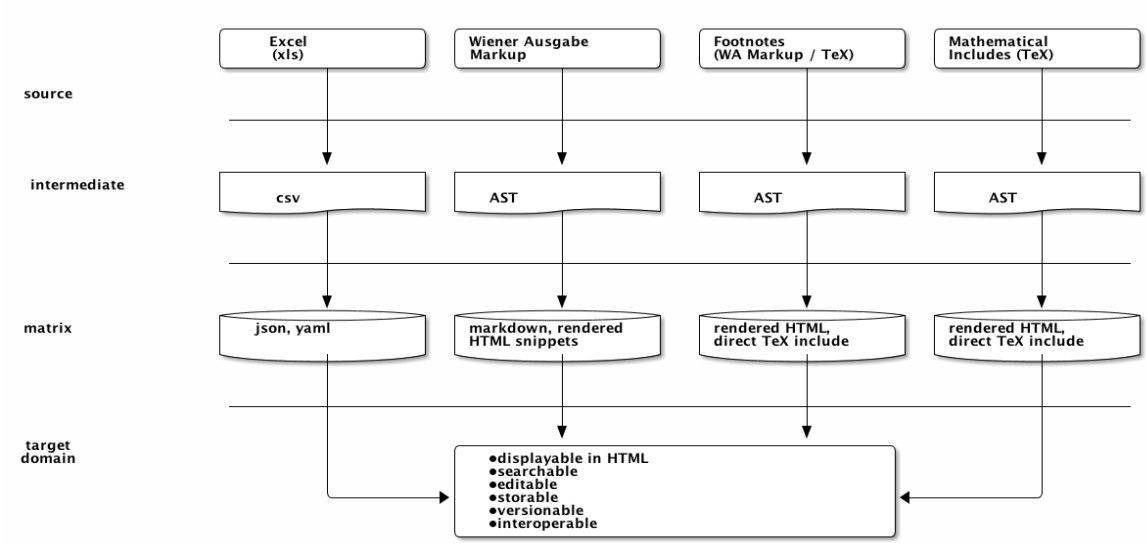


Abbildung 7: Herstellung und Funktionsweise von AST

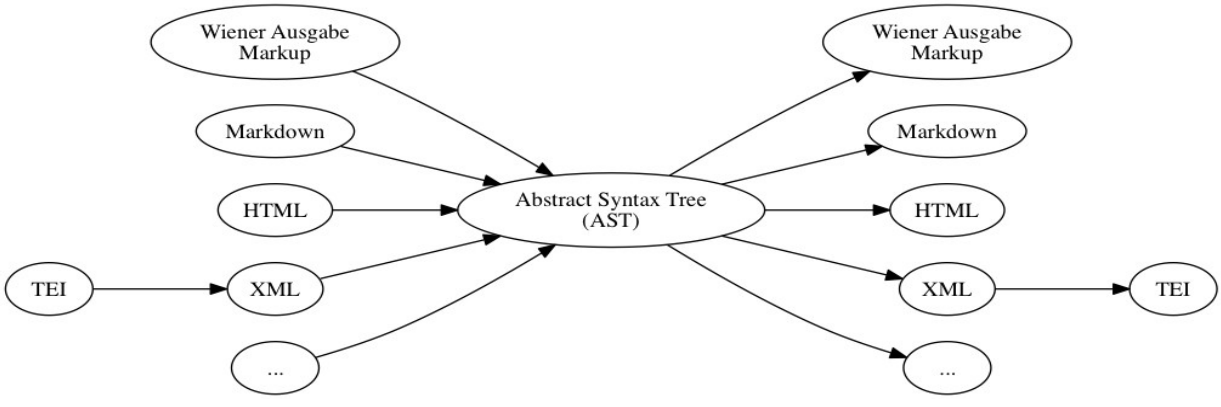


Abbildung 8: Herstellung und Funktionsweise von AST

Abstract für 2. Jahrestagung des Verbandes “Digital Humanities im deutschsprachigen Raum”

Poster

Visualisierung von Kultur im Web

Das Poster problematisiert die softwaregestützte Visualisierung von Kultur im World Wide Web (Web). Bisher konzentrieren sich sozialwissenschaftlich relevante Visualisierungen auf das Web als Netzwerk von Informationen. Untersucht wird vor allem der menschliche Faktor in Verbindung mit Hypertext und Internetinfrastrukturen bei der Entstehung, Verbreitung und Veränderung von verknüpften Informationen und Informationsflüssen. Dazu werden in der Regel Beziehungen, Verteilungen und die Performance auf der Basis von Metadaten durch vereinfachte visuelle Darstellungen (Punkte, Linien, einfache geometirische Figuren etc.) in Häufigkeitsverteilungen, Netzwerkbeziehungen oder Pfadmodellen abgebildet. Eindrucksvolle Studien visualisieren beispielsweise Beziehungsmuster zwischen den NutzerInnen in Facebook oder die NutzerInnen-Performance in Wikipedia.

An solchen Visualisierungen kann mit Manovich (2011) die Reduktionen der Daten auf ein Merkmal (die oben angesprochenen vereinfachten Darstellungsformen) oder die Präferenz, Eigenschaften vor allem räumlich zu visualisieren (z.B. nah/fern, innen/außen), kritisiert werden. Ebenso kann man die dominante Netzwerklogik bei der Beobachtung des Webs problematisieren. Die einseitig erfolgte Zuspitzung ist jedoch legitim, um die Vernetzung von Informationen zu repräsentieren und zu beobachten. Zugleich besteht das Web nicht nur aus verknüpften Informationen, sondern es erzeugt auch spezifische Bedeutungszuschreibungen und kollektive Deutungsmuster. Es handelt sich folglich, kultursoziologisch ausgedrückt, beim Web ebenfalls um ein von Menschen geschaffenes Bedeutungsgewebe (Geertz 1987). Mit Gewebe ist aber eben kein Netzwerk aus Informationen gemeint, sondern kulturspezifische Auslegungen und Deutungen (kurz: Sinnzusammenhänge, Frames, Orientierungsschemata). Für die textbasierte Untersuchung und Analyse von Strukturen und Mustern bieten sich in der Regel direkte Visualisierungen wie Tag Clouds, statistisch gestützte Kookkurrenzanalysen oder Clusterverfahren wie das Topic Modeling an. Für die Analyse von Kultur liefern diese Verfahren Möglichkeiten, thematische Differenzierungen und die Relevanz bestimmter Themen punktuell über die Zeit abzubilden. Insbesondere für online zugängliche Pressemitteilungen und Pressenachrichten liegen dazu bereits verschiedene, überzeugende Studien vor (z. B. Mohr & Bogdanov 2013).

Textproduktionen in Organisationen wie Unternehmen, Medienanstalten oder Nicht-Regierungsorganisationen unterliegen aber häufig hierarchischen Entscheidungsstrukturen. In der Regel existieren Formen institutionalisierter Autorisierung, die über die Herstellung, die

Inhalte und Freigaben von Texten entscheiden. Das Web bietet hingegen auch die Möglichkeit, kulturspezifische Deutungen und Auslegungen jenseits institutioneller Autorisierungen zu beobachten. Zu diesen Sinnzusammenhängen zählen „folksonomies“ (Mathes 2004), welche vor allem in den sozialen Medien entstehen. Derartige „grassroots“-Kategorisierungen in Form von „tags“ (nutzergenerierte Schlagworte), um beispielsweise digitale Medien wie URLs oder Photographien zu teilen und zu organisieren, entsprechen alltagstheoretischen, kommunikativen Typisierungen. Aus einer kultursoziologischen Perspektive bilden solche Typisierungen ein erfahrungsbasiertes Orientierungswissen, welches die alltägliche Praxis strukturiert und anleitet (Bohnsack 2006). Daraus resultieren Gewohnheiten und Routinen, die sich auch in relativ stabilen, kollektiv erzeugten Kategorisierungen im Web dokumentieren (z.B. Golder & Huberman 2005). Kategorisierungen von Bildern auf Plattformen wie Flickr oder Instagram erlauben zudem, die Schlagworte den BildproduzentInnen zuzurechnen, da sie in der Regel die textliche Kategorisierung der Bilder vornehmen. Solche Kategorisierungen des Abgebildeten sollten sich daher im Besonderen dafür eignen, kultur- und webspezifische Wahrnehmungs-, Denk- und Bewertungsweisen aufzuzeigen. Einerseits könnten Bezeichnungen des Abgebildeten in Kombination mit anderen Schlagworten untersucht werden (Stichwort: Topic Modeling). Andererseits könnten die Prozesse der Kategorisierung nachverfolgt werden, um – für gewöhnlich unzugängliche – Prozesse von Neuverknüpfungen und die Entstehung neuartiger Sinnzusammenhänge zu beobachten. Eine entsprechende Visualisierung könnte ermöglichen, theoretisch postulierte Umkehrungen oder die „neuartige Fortsetzung von Eingelebtem“ (Hörning 2004: 33) sichtbar zu machen und empirisch in Feinanalyse zu untersuchen.

Die angesprochenen Kategorisierungsprozesse stellen besondere Anforderungen an automatisierte Visualisierungsverfahren, da sie nicht nur Sinnzusammenhänge in Form von Topics abbilden, sondern ebenfalls Verschiebungen im Orientierungswissen erfassen sollen. Das Poster formuliert kultursoziologische Ansprüche an entsprechende Softwarelösungen und diskutiert Potentiale und Grenzen von technischen Lösungen wie Topic Modeling zu unterschiedlichen Zeitpunkten, Cloudaliscious oder Alluvial Diagramme.

Referenzen

Bohnsack, R. (2006). Mannheims Wissenssoziologie als Methode. In: (Hg.): Neue Perspektiven der Wissenssoziologie (S. 271–291), herausgegeben von D. Tänzler, H. Knoblauch, & H.-G. Soeffner. Konstanz.

Geertz, C. (1987). Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme. Frankfurt a. M.

Golder, S. A., & Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems (No. cs. DL/0508082). cs/0508082.

Hörning, Karl H. (2004). Soziale Praxis zwischen Beharrung und Neuschöpfung. In Doing culture: neue Positionen zum Verhältnis von Kultur und sozialer Praxis (S. 19-39), herausgegeben von K. Hörning, & J. Reuter. Bielefeld.

Manovich, L. (2011). What is visualisation? Visual Studies, 26(1), 36-49.

Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata. Computer Mediated Communication, 47(10), 1-13.

Mohr, J. W., & Bogdanov, P. (2013). Topic models: What they are and why they matter. Poetics: Special Issue on Topic Models and the Cultural Sciences, 41(6).

Neue Wege des Sammelns, Erfassens und Erforschens: Die Datenbank ‚Dialect Cultures'.

Elisabeth Zehetner, Stefanie Edler

Institut für Germanistik, Karl-Franzens-Universität Graz, Heinrichstr. 26, 8010 Graz

Das Projekt ‚Dialect Cultures'

Dialektliteratur erlebte im bairisch-österreichischen Raum bereits im 18. Jahrhundert eine erste Blüte, in der sie in ihrer inhaltlichen und funktionalen Bandbreite weit über jene idyllische, rückwärtsgewandte Heimatdichtung hinausging, mit der Mundartdichtung bis heute – auch in der Forschung – gerne identifiziert wird. Dialektkunst polarisierte wie keine andere Gattung das zeitgenössische Werturteil, wurde von Kaisern geliebt und von Kritikern verachtet, war in aller Munde und wurde von großen Meistern ebenso gepflegt wie von ungebildeten Laien.

Ihre Vielfalt ist jedoch kaum beachtet worden: Dialektliteratur führt in der Wissenschaft nach wie vor ein ungeliebtes Dasein, und die überlieferten Texte wurden bisher nicht systematisch dokumentiert oder kommentiert. Das Projekt ‚Dialect Cultures' will diese Lücke schließen und erforscht die verschiedenen ästhetischen und funktionalen Möglichkeiten der Dialektkunst im 17. und 18. Jahrhundert, indem bestehende Forschungsergebnisse und historische Quellen neu erschlossen und zusammengeführt werden. Die Materialgrundlage bildet dabei eine im Rahmen der ersten Projektphase erstellte umfassende Sammlung von historischen literarischen Texten und Notenmaterialien aus handschriftlichen oder gedruckten Quellen, welche nunmehr in einer Datenbank gebündelt, strukturiert und vernetzt vorliegt. Unter Berücksichtigung von Ansätzen aus unterschiedlichen Disziplinen soll Mundartverwendung auf dieser Basis als künstlerisches Phänomen vor 1800 in ihrer ganzen Bandbreite erfasst werden.

Die Datenbank

Kernstück des Projekts ist die Datenbank, in der im Rahmen der Projektarbeit seit 2010 literarische Texte des 17. und 18. Jahrhunderts gesammelt und kommentiert werden. Die Sammlung umfasst zurzeit ca. 1300 Werke aus den Bereichen Lyrik, Drama und Prosa in mehr als 2000 Varianten.

Werke lassen sich teilweise durch mehrere überlieferte Textzeugnisse belegen. Wenn diese untereinander leichte Abweichungen aufweisen, so sind für ein Werk mehrere Varianten zu verzeichnen, deren Differenzen in den jeweiligen Varianteneinträgen diskutiert werden. Die Varianteneinträge können mit entsprechend zugeordneten Autoren/Komponisten-, Quellen- sowie Literatureinträgen verlinkt werden. Darüber hinaus können den Varianten in gesonderten Dateien auch Digitalisate von Handschriften und Drucken sowie Transkriptionen zugeordnet werden. Für die Benutzer ist über die Variantenansicht auch der Zugriff auf diese Inhalte und somit z.B. ein Vergleich unterschiedlicher Varianten direkt am Originalmaterial möglich.

Auf einzigartige Weise verbindet die Datenbank so eine Datensammlung zu Texten historischer Dialektliteratur mit wissenschaftlicher Kommentierung und Edition. Diese bislang nur verstreut und in der Regel getrennt voneinander verfügbaren Informationen – in Bibliothekskatalogen und Überblicksdarstellungen einerseits, in Einzeleditionen und Artikeln zu spezifi-

schen Themen andererseits – können so gesammelt und systematisch verknüpft werden. Alle Materialien und Informationen von der Quelle bis zur Forschungsliteratur sind damit auf einer gemeinsamen Plattform verfügbar.

Die Struktur der Datenbank ermöglicht auch das Aufdecken neuer Zusammenhänge, indem etwa verschiedene, bislang nicht bekannte Varianten verglichen werden oder thematische Schwerpunkte in der überlieferten Dialektliteratur und innerhalb einzelner Gattungen systematisch recherchiert werden können.

Die Datenbank erfüllt damit zwei wichtige Funktionen:

(1) Unterstützung der Forschung: Die online zugängliche Datenbank ermöglicht das Zusammenführen verschiedener Varianten und die Nachvollziehbarkeit von Quellen und Literatur, und dies insbesondere auch bei der Arbeit im Team. Der Datenbestand ist jederzeit von allen Beteiligten ausweitbar und für alle zeitgleich und übersichtlich nutzbar.

(2) Öffentlicher Zugang: Der Aufbau der Datenbank, der einen einfachen Zugriff über verschiedene Ebenen – Autoren, Werktitel und -incipits, Gattungen etc. – erlaubt, macht die Datensammlung über die Projektarbeit hinaus für ein breites Publikum nutzbar:

- Wissenschaftler aus verschiedenen Disziplinen wie Literaturwissenschaft, Sprachwissenschaft, Geschichte oder Musikwissenschaft, die mit ihren jeweils eigenen Fragestellungen an das Korpus herantreten können
- Studierende, die die Datenbank für Recherche und für das Kennenlernen von Transkriptions- und Editionsmethoden nutzen können
- Interessierte außerhalb des Wissenschaftsbetriebs, für die die Ergebnisse wissenschaftlicher Forschung auf einfache Weise zugänglich werden.

Über eine Rückmeldefunktion können alle drei Gruppen nicht nur die Ergebnisse der Projektarbeit nutzen, sondern auch weiter ausbauen, indem neue Funde oder zusätzliche Kommentare zur Verfügung gestellt und – nach Überprüfung durch die Projektverantwortlichen – wieder in die Datenbank integriert werden können.

Die Datenbank gewährleistet also Offenheit und Austausch sowohl im Team als auch mit einem größeren Publikum und garantiert, dass die im Projektverlauf gesammelten Daten auch nach dem Ende der Projektlaufzeit gesichert und zugänglich bleiben.

Damit erweist sich das Modell der Datenbank mit ihrer Verknüpfung von verschiedenen Daten und Erkenntnissen als wegweisend über unser Projekt hinaus: Wissenschaft ist zunehmend durch eine wachsende Anzahl meist verhältnismäßig kurzfristiger Drittmittelprojekte gekennzeichnet, die häufig nur in geringem Ausmaß in die etablierten institutionellen Strukturen der Universität eingebunden sind. Gerade angesichts dessen scheint die längerfristige, umfassende Sicherung von Daten unerlässlich, um die Weiterverwendung der Ergebnisse und damit die Nachhaltigkeit des Erarbeiteten zu sichern.

neonion – Kollaboratives, semantisches Annotieren von Dokumenten als Mehrwert für das Forschen in den Geisteswissenschaften und der Informatik

Claudia Müller-Birn[1], Florian Schmaltz[2], Tina Klüwer[1], Juliane Stiller[2]
[1] Freie Universität Berlin, Institut für Informatik, Human-Centered Computing
[2] Max-Planck-Institut fuer Wissenschaftsgeschichte, Forschungsprogramm Geschichte der Max-Planck-Gesellschaft

neonion ist eine Webanwendung, die es Benutzer_innen erlaubt, Wörter und Textteile in Dokumenten oder Dokumente selbst semantisch zu annotieren. Das Ziel bei der Softwareentwicklung ist es dabei insbesondere, den Prozess der Erstellung der semantischen Annotationen so intuitiv zu gestalten, dass die Komplexität des zugrundliegenden Datenmodells vor den Nutzer_innen weitestgehend verborgen werden kann. Mit neonion sollen somit vor allem Wissenschaftler_innen angesprochen werden, die nicht mit semantischen Technologien wie RDF, Ontologien oder SPARQL vertraut sind, aber trotzdem von den Vorteilen dieser Technologien profitieren wollen. neonion ermöglicht es Dokumente gemeinschaftlich zu annotieren, d.h. als Mensch-Mensch oder Mensch-Maschine-Kollaboration. Annotationen können privat sein, in einer Gruppe gemeinsam erstellt oder sogar öffentlich sichtbar gemacht werden. So kann das bei der Annotation erstellte Wissen auch Teil des *Webs of Data* werden.

Derzeit wird neonion gemeinsam mit den Nutzer_innen aus dem Bereich der Geschichtswissenschaften entwickelt. In einem Pilotprojekt für das Forschungsprogramm „Geschichte der Max-Planck-Gesellschaft, 1948-2002“, welches am Max-Planck-Institut für Wissenschaftsgeschichte in Berlin angesiedelt ist, finden regelmäßige Treffen mit den potentiellen Nutzer_innen der Software statt.

Parallel zur Entwicklung der Software wird eine Studie durchgeführt, in der Wissenschaftler_innen interviewt werden und ihre Benutzung von neonion experimentell beobachtet wird. Erste Ergebnisse der Interviews zeigen, dass die dem Annotationsprozess zugrundeliegenden mentalen Modelle sehr unterschiedlich ausfallen, d.h. dass die Befragten ein individuelles Verständnis zum Konzept der Annotation besitzen. So wurde der verwendete Annotationsinhalt (z.B. Kategorie, Freitext, Tag) durch das Ziel der Auswertung (z.B. Klassifizierung, Übersetzung) beeinflusst. Die Befragten sehen semantische Annotationen dann als nützlich an, wenn Nutzer_innen ihre Annotationen gemeinschaftlich erstellen und verwenden sowie diese maschinell weiter verarbeiten können. In den Interviews hat sich gezeigt, dass Wissenschaftler_innen um die Nützlichkeit des semantischen Ansatzes wissen, aber sehr unsicher bei der eigentlichen Anwendung sind. Die fehlende Erfahrung in diesem Bereich führt letztlich wieder zur Verwendung von Werkzeugen wie beispielsweise Textverarbeitungsprogrammen oder PDF-Software, obwohl die Nachteile, beispielsweise bezüglich der Weiterverwendung der Annotationen, bekannt sind. Erste Usability-Studien mit neonion haben gezeigt, dass Nutzer_innen sich sehr gut in der Software zurechtfinden und Annotationen schnell und intuitiv durchführen können. Durch die enge Zusammenarbeit von Anwender_innen und Entwickler_innen können Möglichkeiten zur Verbesserung des Interaktionsdesigns frühzeitig im Entwicklungsprozess erkannt und adressiert werden.

Wie bereits dargelegt, ist ein Ziel von neonion die kollaborative Annotation von Dokumenten nicht nur für menschliche Interaktionen, sondern auch Mensch-Maschine-Interaktionen zu unterstützen. Dazu werden zwei Ebenen der manuellen semantischen Annotation unterschieden, die im Folgenden kurz anhand der Personenannotation erläutert werden:

(1) Auf der *Konzeptebene* annotiert der Nutzer_innen ausgewählte Wörter oder Textteile mit vorher festgelegten Begriffen (sog. Konzepte). Zum Beispiel sollen alle in historischen Dokumenten genannten Personen auf ihr gemeinsames Vorkommen (z.B. bezüglich Zeit und Ort) hin untersucht werden. Daher werden von den Nutzer_innen Namen, z.B. „Feodor Lynen“ mit dem Konzept „Person“ verbunden und damit im Dokument eine Annotation erzeugt. Die resultierende RDF-basierte Beschreibung der Instanz enthält die ausgewählten Namen vom Typ Person.

(2) Benutzer_innen können nicht nur die ausgewählten Namen auf ein Konzept in einer Ontologie beziehen, sondern diese Instanzen auch mit einer direkt identifizierbaren Ressource im Web verknüpfen. Wir bezeichnen dies als Annotation auf der *Referenzebene*. So könnte die lokal annotierte Person „Feodor Lynen“ mit dem Wikidata-Eintrag Q44597 zu Feodor Lynen referenziert werden.

Die Annotationen können nun um weiteres Wissen angereichert werden. Derzeit nutzt neonion Wikidata und dies erlaubt Nutzer_innen ebenfalls auf Daten aus der VIAF (Virtual International Authority File) oder der GND (Gemeinsame Normdatei) zuzugreifen. Indem also weitere Informationen aus der Linked Open Data Cloud einbezogen werden, kann auf Basis einer einfachen Annotation ein komplexes Wissensnetzwerk erzeugt werden. Solche Wissensnetzwerke können dann auch zur weiteren Analyse visualisiert werden.

Diese manuellen Annotationsebenen werden durch automatische Annotatoren, derzeit durch einen Named Entity Recognizer, erweitert. Während der manuellen Annotation von Dokumenten werden den Nutzer_innen Vorschläge des automatischen Annotators präsentiert. Die Nutzer_innen können diese annehmen, ablehnen oder editieren. Über diesen Interaktionsprozess werden zukünftig angebotene Empfehlungen schrittweise verbessert. Das entwickelte Mensch-Maschine-Interaktionskonzept wird derzeit evaluiert.

neonion soll als ein Beispiel für eine gelungene Zusammenarbeit zwischen den Geisteswissenschaften und der Informatik dienen, da die Forschung in beiden Fachdisziplinen mit diesem Projekt vorangetrieben werden kann.

Das Labeling System – ein freier Baukasten für kontrollierte Vokabulare

Michael Piotrowski
Florian Thiery
Kai-Christian Bruhn

10. November 2014

Maschinenlesbare Annotationen sind die Voraussetzung für die semantische Verarbeitung von Daten. Diese Aussage gilt unabhängig davon, ob es sich bei den Daten um natürlichsprachigen Text oder um strukturierte Datensätze in einer Datenbank handelt, und unabhängig davon, ob es um einfaches Sortieren und Filtern geht oder um komplexes automatisches Schließen. Kontrollierte Vokabulare (ob in einer einfachen Terminolgieliste oder als Taxonomien, Thesauri oder Ontologien strukturiert) sind dabei unbedingt notwendig, um Annotationen maschinell verarbeitbar zu machen; ohne terminologische Kontrolle sind Annotationen für die maschinelle Verarbeitung kaum nützlicher als an den Rand eines Buches gekrizelte Notizen. Kontrollierte Vokabulare abstrahieren von natürlichsprachlichen Ambiguitäten und Konnotationen; sie sind daher entscheidend für die semantische Verarbeitung von Forschungsdaten. Um projektübergreifende Zusammenarbeit und den semantischen Austausch von Daten zu ermöglichen, müssen Vokabulare nicht nur kontrolliert, sondern auch formell oder informell standardisiert sein. Standardisierte kontrollierte Vokabulare ermöglichen den Austausch, die Kombination und die gemeinsame Analyse annotierter Daten aus verschiedenen Quellen sowie die Implementierung generischer Werkzeuge für die semantische Verarbeitung.

Erstellung und Wartung standardisierter kontrollierter Vokabulare sind jedoch zeitaufwändig und damit teuer. Zu den größten Herausforderungen zählen, dass alle beteiligten Parteien zu einem gemeinsamen Verständnis der Begriffe kommen, und dass die richtige Balance zwischen möglichst breiter Anwendbarkeit einerseits und möglichst präziser Analyse andereseits gefunden werden. Diese Ziele sind insbesondere in den Geisteswissenschaften schwierig zu erreichen: nicht nur sind die Forschungsfragen, die potentiell an einen gegebenen Datensatz gerichtet werden können, extrem weit gefächert, sondern die Kategorisierung der Daten ist häufig ein essenzieller Teil des Forschungsprozesses selbst. Es gibt daher einen eklatanten Mangel an standardisierten kontrollierten Vokabularen in den Geisteswissenschaften, der Digital-Humanities-Projekte letztlich dazu zwingt, eigene, projektspezifische Vokabulare zu definieren. Projektspezifische Vokabulare lösen können den internen Bedarf zwar kurzfristig befriedigen, sind aber nicht interoperabel und verhindern den zukünftigen Austausch und die Nachnutzung der annotierten Daten.

Unser Poster stellt einen neuen konzeptuellen Ansatz zur Lösung dieser Probleme vor und beschreibt die Implementierung dieses Ansatzes in einem Softwarewerkzeug, dem *Labeling System*.

Da es in der geisteswissenschaftlichen Forschung praktisch unmöglich ist, kontrollierte Vokabulare zu definieren, die alle denkbaren Anwendungen abdecken und generell akzeptiert sind, schlagen wir ein anderes Vorgehen vor. Bei unserem Ansatz definieren Projekte ihre eigenen Vokabulare, aber anstelle natürlichsprachlicher

Definitionen werden die Terme mit einem oder mehreren Konzepten in einem Referenzthesaurus verknüpft. Der projektspezifische Term dient also quasi als »Label« für eine Menge gemeinsamer Konzepte. Dieser Ansatz ermöglicht es Projekten Vokabulare entsprechend ihrer Bedürfnisse und unter Verwendung im jeweiligen Forschungsgebiet üblichen Bezeichnungen benutzen, während gleichzeitig die Interoperabilität mit anderen Projekten über den Referenzthesaurus gewährleistet ist.

Das Labeling System ist eine Webanwendung, die es Benutzern ermöglicht, SKOS-Vokabulare zu erstellen und auf einfache Weise deren Terme mit einem oder mehreren Konzepten in einem oder mehreren Referenzthesauri zu verknüpfen. Die Benutzeroberfläche ermöglicht die Visualisierung der definierten Vokabulare in einer hierarchischen Baumstruktur und ermöglicht den Zugriff auf Vokabulare über eine SPARQL-Schnittstelle. Das Labeling System basiert auf ausgereiften Open-Source-Komponenten und ist selbst ebenfalls frei verfügbar.

Learning cuneiform the modern way

Timo Homburg[1], Christian Chiarcos[1], Thomas Richter[2], Dirk Wicke[2]

[1] Institute for Computer Science [2] Institute for Archaeology
Goethe University, Frankfurt, Germany
timo.homburg@stud.uni-frankfurt.de,
{chiarcos|thomas.richter|wicke}@em.uni-frankfurt.de

Keywords: Assyriology, cuneiform, input method engines (IME), flash card learning

With our poster and the accompanying demo, we present current progress on the information-technological support for scholars and students of cuneiform. For a period of about 3000 years, cuneiform was the dominant writing system of the Ancient Near East, with a rich literary tradition in several languages, and an extensive amount of texts preserved in tens of thousands of clay tablets.

Despite this wealth of data and a strong academic tradition in their analysis, the numerous specific challenges of cuneiform have only partially been addressed so far. Here, we propose adapting input method engines (IMEs) and learning strategies commonly used for Asian languages according to the needs of Assyriology.

Typing cuneiform Cuneiform writing for Akkadian, Sumerian and Hittite is ideosyllabic, i.e., combining syllabic and ideographic elements, often for the same sign. Up until this date there is no free and convenient way of typing Unicode cuneiform characters other than utilizing the Unicode code tables directly, i.e., to copy and paste from online dictionaries. Not all online dictionaries, however, use Unicode symbols, some use legacy fonts, some represent signs by images, and some do not provide a cuneiform representation at all.

To accommodate this deficit, we developed an input method which is based on the transliteration concept of Chinese Pinyin, the most common way of typing non-alphabetical languages on a computer. To achieve an equivalent input for the aforementioned languages we utilized a given char transliteration to cuneiform table[1] to create transliteration to cuneiform mappings of Akkadian, Sumerian and Hittite CDLI[2] corpora respectively. Organized as a tree, thus minimizing latency, word and char-based input method engines were created for Java (JIMF,

[1] http://www.acoli.informatik.uni-frankfurt.de/resources/cuneiform/signs-final.xml
[2] http://cdli.ucla.edu/

Fig. 1.a)[3], JQuery(Fig. 2)[4], SCIM[5] and Ibus (Fig. 3)[6], thereby covering the most important input method engines on Linux, Web and Java environments.

Learning cuneiform Because of limited technological support, digital resources in assyriology often focus on transliteration or transcription as means of representation whereas students are required to acquire the necessary knowledge on cuneiform characters on their own. Clearly, none of those practices are satisfying or easily adaptable for text processing and therefore not useful for computer-aided teaching methods. For conveniently typing and learning cuneiform characters, words and phrases for the Akkadian, Sumerian and Hittite language, we present an adaptation of Anki, a common tool for flash card learning (Fig. 1.b). We utilized the existing character/word table to create flash card sets consisting of more than 50000 words for the Anki and AnkiDroid[7] flash card learning program. Subsequent extensions may exploit existing corpora, e.g., the Open Richly Annotated Cuneiform Corpus (ORACC),[8] to create flash cards for words and phrases. Anki schedules learning content according to a spaced repetition learning method having proven its positive learning effect over a longer period of time to maximize learning success. Because of its usability in both mobile and desktop environments and its ability to share flashcards online Anki suits not only the students but also simplifies sharing lecture specified flash card sets for the lecturers.

In our presentation, we demonstrate how an input method engine can act as a suitable tool for solving the mentioned input and compatibility problems while at the same time being useful for education and language learning purposes. With the tools described above, teachers can now easily create their own flash cards according to the pace and content of their lectures. Students may enjoy a convenient and scientifically proven way of learning cuneiform vocabulary, as well as a way to prove their learning by utilizing the input method engine to create their own cuneiform texts. In conclusion, a notable improvement in writing and in learning the concerned languages has been realized and is in general perceived well.

Both tools and the accompanying sign/word table have been created in the context of on-going experiments on word segmentation and transliteration in cuneiform languages. In this regard, we are thus primarily working on *processing cuneiform*. In addition to demonstrating input methods and learning tools, we will include early results with respect to these aspects in demo and presentation, as well.

[3] http://docs.oracle.com/javase/7/docs/technotes/guides/imf/overview.html
[4] Sourcecode: https://github.com/situx/webime
Livedemo: http://www.web-ime.de.vu
[5] http://sourceforge.net/projects/scim/
[6] https://code.google.com/p/ibus/
[7] http://ankisrs.net
[8] http://http://oracc.museum.upenn.edu

(a) (b)

Fig. 1. Learning Cuneiform: (a) Java Input Method Framework based IME for Swing based applications and (b) Anki Flash Cards

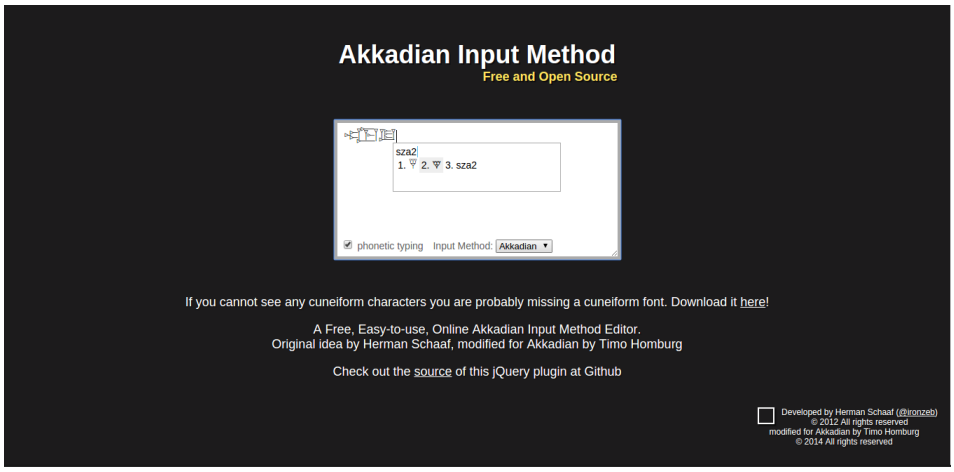


Fig. 2. JQuery based Akkadian Input Method Engine testable on http://www.web-ime.de.vu

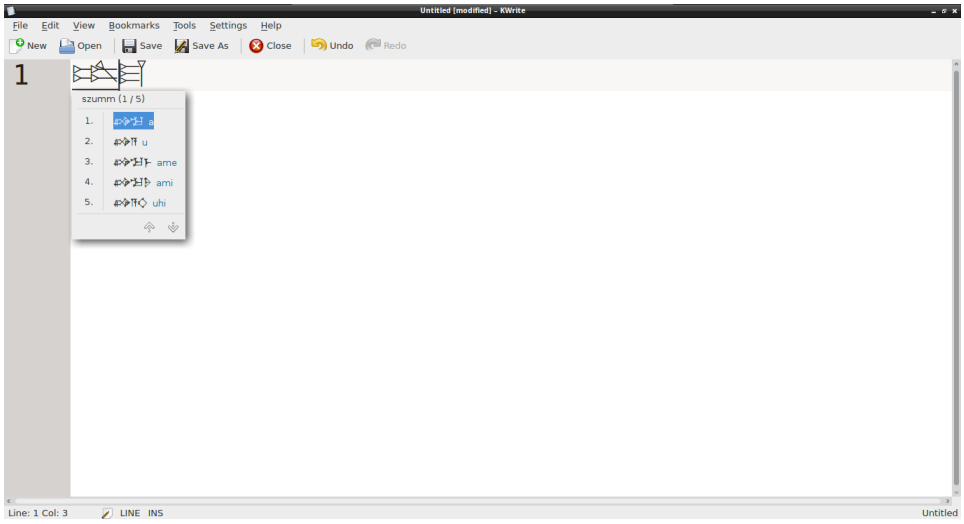


Fig. 3. System-wide Input Method Engine using Ibus for Linux with SCIM giving a similar output

Modellierung eines maschinell lesbaren Lexikons für das Korpus der altäthiopischen Literatur

Alessandro Bausi, Andreas Ellwardt, Cristina Vertan
Universität Hamburg

1. Einführung

Die Entwicklung und ständige Erweiterung des Unicode-Kodierungssytems Unicode[1] sowie der Mark-up-Sprachen XML[2], TEI[3] haben in den letzten Jahren u.a. die digitale textuelle Repräsentation von historischen Dokumenten, die mit unterschiedlichen Alphabeten geschrieben wurden, ermöglicht.

Diese textuelle Repräsentation eröffnet wiederum, im Kontrast zur reinen Speicherung von Bild-Digitalisaten, die Möglichkeit, computergestützte linguistische sowie philologische Untersuchungen auf großen Textmengen durchzuführen. Durch solche Methoden lässt sich beispielsweise eine diachrone Analyse der Sprache gleichzeitig auf mehreren Ebenen (morphologisch, syntaktisch, semantisch) realisieren, vorausgesetzt, die elektronischen Ressourcen wie Lexika oder annotierte Korpora sowie die sprachtechnologischen Prozesse (morphologische Analysierer, Wortart-Tagger, Parser) sind vorhanden.

Während die sprachtechnologischen Ressourcen und Werkzeuge für moderne Sprachen sehr weit entwickelt sind, gelten viele historische Sprachen als stark „under-ressourced“. Laut Krauwer (2003) gibt es ein minimales Set von Ressourcen, die für eine computergestützte Sprachanalyse unabdingbar sind. Dessen Weiterentwicklung stellt die Wissenschaft vor neue Forschungsprobleme, da sich häufig Modelle, die für moderne Sprachen entwickelt wurden, nicht 1:1 auf historische Sprachen übertragen lassen (VertanEtAl.2014)

In diesem Beitrag werden wir die Modellierung und Entwicklung von sprachtechnologischen Ressourcen für das Altäthiopische (Geʿez) erläutern. Die Besonderheiten des Geʿez (s. Sektion 2), bedingen die Entwicklung von neuen Modellen, z.B. im Bereich der Lexika. In Sektion 3 werden wir exemplarisch die Entwicklung eines Lexikon-Modells für Geʿez darstellen, während wir in Sektion 4 die Einbindung des Lexikons in einer Architektur für die diachrone Analyse des Geʿez diskutieren werden.

2. Kurze Darstellung des Altäthiopischen (Geʿez)

Das südsemitische Geʿez ist die Sprache des Königreichs Aksum in der heutigen nordäthiopischen Provinz Tigray, von wo aus die im 4. Jahrhundert beginnende Christianisierung Äthiopiens ihren Anfang nahm. Die in der Folge entstehende reiche Literatur ist in großem Umfang geprägt von Übersetzungen aus dem Griechischen und später, ab dem 13. Jahrhundert, aus dem Arabischen, was durch grammatische Interferenzphänomene reflektiert wird. Während seine Verdrängung als gesprochene Sprache bereits im 9./10. Jahrhundert beginnt, bleibt es als Schriftsprache sehr viel länger erhalten und ist bis in die Gegenwart hinein Liturgiesprache des äthiopischen und eriträischen Klerus.

Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf, außerdem werden die Vokale vollständig geschrieben. Beides unterscheidet das Geʿez von den ihm nächst verwandten Sprachen Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch. Des weiteren sind Grapheme, die ursprünglich distinkten Phonemen zugeordnet waren, schon früh in identischer phonetischer Realisierung zusammengefallen, was sich konkret bereits in den ältesten überlieferten Handschriftzeugnissen (aber noch nicht in den aksumitischen Inschriften) niederschlägt, wo eine beliebige Austauschbarkeit der Laryngale und Sibilanten jeweils untereinander zu konstatieren ist.
Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Hierbei muss das einzelne Lexem als Kombination von zwei Elementen beschrieben werden, nämlich der Wurzel und dem Schema: Die konsonantische Wurzel gibt veränderliche Positionen zwischen

[1] http://www.unicode.org/
[2] http://www.w3.org/XML/
[3] http://www.tei-c.org/index.xml

ihren, zumeist drei, Wurzelkonsonanten vor, die durch die Vokale des Schemas aufgefüllt werden, häufig, jedoch nicht zwingend, ergänzt um (vokalische oder konsonantische) Affixe.

3. Arbeitsschritte zu einer computergestützten Analyse des Altäthiopischen

Wie bereits in Sektion 2 erwähnt, sind Geʿez-Dokumente für die gesamte Geschichte des christlichen Orients extrem wertvoll. Manche Überlieferungen von alten griechischen Texten sind in der Originalsprache verloren und nur im Altäthiopischen erhalten. In der Zeit digitaler Bibliotheken erscheint also die Entwicklung von computergestützten Tools für die Geʿez-Sprache umso dringender. Das primäre Ziel des Projekts TraCES[4] ist die Entwicklung eines digitalen Korpus der Geʿez-Sprache, zusammen mit Annotationen auf morphologischer, syntaktischer und semantischer Ebene. Dieses annotierte Korpus soll einerseits eine diachrone Analyse des Altäthiopischen ermöglichen, anderseits soll es selbst als Ressource für weitere computergestützte Prozesse dienen. Langfristig soll eine vergleichende digitale Analyse von altäthiopischen und griechischen (z.B. die in der digitalen PERSEUS Sammlung[5] verfügbaren) oder arabischen sowie anderen christlich-orientalischen Dokumenten möglich sein.

Mit Ausnahme von einigen wenigen Texten gibt es zur Zeit keine verfügbare elektronische Ressource für das Altäthiopische. Daher haben wir uns als erstes der Entwicklung eines maschinell lesbaren Lexikons des Geʿez gewidmet. Dessen Modellierung wird in der nächsten Sektion erklärt.

4. Ein Lexikon-Modell für Geʿez

Die in Sektion 2 erwähnte Austauschbarkeit der Laryngalen und Sibilanten untereinander stellt uns vor eine erste Modellierungsanforderung. Für einen Lexikon-Eintrag muss nicht nur die Grundform, sondern es müssen auch alle möglichen graphischen Varianten gespeichert werden, wobei wohlgemerkt diese graphische Variationen auch in einigen Fällen als selbständige Lexikon-Einträge mit ganz anderer Bedeutung existieren können.

Das Lexikonmodell muss daher eine starke Modularisierung und Verlinkung zwischen den einzelnen Modulen unterstützen. Wir haben uns für das Lemon-Modell (McCraeEtAl.2012) entschieden. Unserer Kenntnis nach, ist dies der erste Versuch, eine semitische Sprache mit dem Lemon-Modell zu beschreiben.
Die Grundkomponenten eines Lemon-Lexikon-Modells für Geʿez wurden wie folgt angepasst.

Die Zitierform eines Wortes in klassischen Lexika semitischer Sprachen ist in der Regel eine verbale Repräsentation der Wurzel in der 3. Person Perfekt Singular maskulin. Diese Form wird in unserem Lemon-Modell als „Lexical Entry“ gespeichert.

Ein „Lexical Entry“ ist mit den folgenden weiteren Modulen verknüpft:

- Das Lexical Form-Modul beinhaltet alle möglichen graphischen Varianten des Lemmas. Jede graphische Variante wird zusammen mit ihrer Transkription gespeichert.
- Das Morphologie-Modul beinhaltet eine Subkomponente für den lexikalischen Eintrag, die das Paradigma, Ausnahmen der morphologischen Realisierung (z.B. Sonderformen im Imperfekt oder Plural) sowie die jeweiligen anderen morphologischen Kategorien für das Lemma umfasst. Das Semantik-Modul setzt sich aus einer Übersetzungs-, einer Korpusevidenz- und einer semantische-Merkmale-Komponente zusammen. Unter Korpusevidenz verstehen wir Beispiele aus Korpora für dieses Lemma oder eine seiner morphologischen Realisierungen. Die Übersetzungen sind unterteilt in eine Übersetzung ins Englische und semantische Äquivalente in anderen Sprachen wie (falls vorhanden) Arabisch, Hebräisch, Syrisch, Koptisch, Griechisch oder sogar Sanskrit.
- Das Syntax-Modul beinhaltet syntaktische Funktion des Lemmas, zusammen mit Beispielen von syntaktischen Bäumen. Dieses Modul wird in einer späteren Projektphase entwickelt.

[4] European Union Seventh Framework Programme IDEAS (FP7/2007-2013), European Research Council, grant agreement no. 338756, project “TraCES – From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages”, http://www1.uni-hamburg.de/ethiostudies/traces.html
[5] http://www.perseus.tufts.edu/hopper/

4.1. Wurzel-Modellierung

Da die Wurzel eine zentrale Stellung in der semitischen Morphologie hat, haben wir als ersten Schritt ein Wurzel-Sublexikon erstellt. Dieses entspricht dem Wurzel-Submodul im morphologischen Modul.

Die Erstellung des Wurzel-Lexikons wurde vollständig automatisiert. Aus einer digitalen Version des trotz seiner Abfassung im Jahre 1865 unverändert als Standardwerk geltenden „Lexicon linguae aethiopicae“ von August Dillmann (Dillmann1865) (im Unicode-Format) wurden zirka 4000 Wurzel-Einträge mit Hilfe von String-basierten Regeln extrahiert.

Für jede Wurzel wurden:

- die vollständige Transkription
- die auf das konsonantische Gerüst zurückgeführte Transkription
- das konsonantischen Wortbildungsschema
- alle graphischen Varianten zusammen mit deren Transkriptionen

durch regel-basierte Verfahren extrahiert. Die Automatisierung ermöglicht zum ersten Mal die Sammlung aller graphischen Varianten für alle 4000 Wurzeln (wobei hervorgehoben werden muss, dass manche Wurzeln bis zu 50 graphische Varianten haben).

Jede Wurzel wird automatisch mit ihren Homophonen (Einträge mit identischer graphischer Form, aber unterschiedlicher Bedeutung) verknüpft. Erfasst werden durch automatische Prozesse auch alle Lexikoneinträge von graphischen Varianten (falls vorhanden).

Das Wurzel-Lexikon wird im XML-Format gespeichert. Dafür wurde ein eigenes XML-Schema entworfen. Eine Java-basierte graphische Oberfläche wurde implementiert. Diese Oberfläche ermöglicht nicht nur die Visualisierung von den Einzeleinträgen und die Navigation durch das Wurzel-Lexikon, sondern auch manuelle Korrekturen, das Löschen oder das Einfügen von neuen Einträgen.

Nach Korrekturen wird das Wurzel-Lexikon:

- als eine „Authority List“ für das Geʿez-Lexikon und
- als Generierungsquelle für Lexikoneinträge

benutzt.

5. Zusammenfassung und weitere Arbeit

In diesem Beitrag haben wir die Modelle für ein Wurzel- und ein Lemma-Lexikon für die Geʿez-Sprache erklärt. Die Wurzel und Lemma-Akquisition werden weitgehend durch computergestützte Prozesse realisiert. Die erstellte Software wird bei der Präsentation des Beitrags vorgeführt.

Das Projekt TraCES wurde im März 2014 begonnen und hat eine Laufzeit von fünf Jahren. Die Erstellung des Lexikons der Geʿez Sprache ist zurzeit die zentrale Arbeit im Projekt, wobei derzeit die Erstellung von Generierungsparadigmen im Vordergrund steht. Mit deren Hilfe werden durch Computerverfahren Lexikoneinträge generiert.
Ein erster Test hat mehr als 13 000 Einträge generiert. Dies zeigt, dass die Automatisierung eine erhebliche Zeitersparnis für die Lexikon-Akquisition ermöglicht.

Literatur

(Dillmann1865) Dillmann, August, *Lexicon linguæ Æthiopicæ cum indice Latino*, Lipsiae 1865.

(Krauwer2003) Krauwer, Steven, „*The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources*“, http://www.elsnet.org/dox/krauwer-specom2003.pdf (09.11.2014)

(McCraeEtAl2012). McCrae, John und Aguado-de-Cea, Guadalupe und Buitelaar, Paul und, Cimiano, Philipp und Declerck, Thierry und Gómez Pérez, Asunción und Gracia, Jorge und Hollink, Laura und Montiel-Ponsoda, Elena und Spohr, Dennis und Wunner, Tobias, *The Lemon Cookbook*, http://lemon-model.net/lemon-cookbook.pdf (09.11.2014)

(VertanET.AL.2014) Vertan, Cristina und Zervanou, Kalliopi und van den Bosch, Antal und Sporeleder, Caroline (Hrsg.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Götheborg, Sweden, 2014, http://www.aclweb.org/anthology/W14-06 (09.11.2014)

Titel: OpenSource-Bibliotheken und -Tools des SeNeReKo-Projekts

Autoren: Jürgen Knauth, Frederik Elwert

Abstract

Ziel des SeNeReKo-Projektes ist es, durch Techniken der **Se**mantisch-Sozialen **Ne**tzwerkanalyse einen Einblick in Teile von Textkorpora zu erhalten. Damit wird eine Form des „Distant Readings“ realisiert, im konkreten Fall zur Erforschung von **Re**ligions**ko**ntakten in altägyptischen Texten und dem Pali-Kannon. (= SeNeReKo)

Im Kontext des Projekts sind verschiedene Programmierbibliotheken und Werkzeuge entwickelt worden, um den Anforderungen des Projekts gerecht zu werden. Wesentliche Komponenten sind jedoch nicht projektspezifisch, sondern als allgemein verwendbare OpenSource-Komponenten geplant und umgesetzt worden: Wiederverwertbarkeit war von Vorneherein eines der Entwicklungsziele. Da Teilaufgaben von DH-Projekten durchaus öfters ähnlich gelagert sind, ist davon auszugehen, dass die so entstandenen Komponenten und Tools von anderen Wissenschaftlern entweder direkt oder nach geringer Adaption für andere Projekte genutzt werden können: Ziel des vorliegenden Posters ist es daher über genau diese Komponenten und Werkzeuge zu informieren. Da unsere Werkzeuge gerade deswegen entstanden sind, weil bislang noch nichts Vergleichbares zur Verfügung stand um die von uns angetroffenen Probleme effizient zu lösen, hoffen wir so durch unsere Software einen Beitrag für die Wissenschafts-Community zu leisten und so andere Wissenschaftler in ihrer zukünftigen Arbeit unterstützen zu können.

Konkret wurde in SeNeReKo ein Werkzeug zum Tagging von Texten entwickelt. Eine Besonderheit dieses Werkzeugs ist neben der Eigenheiten zur Auflösung von Pali-Sandhis seine besonders gut optimierte Usability: Unser Anliegen war hier, dass möglichst wenige Klicks erforderlich werden, um manuelle Tagging-Aufgaben durchzuführen. Das Fehlen von Tools mit vergleichbar Usability war Motivation der Entwicklung dieses Werkzeugs. Dieses client-server-basierte Standalone-Tool leistete einen wertvollen Beitrag in SeNeReKo für die Erstellung eines Gold-Standards im Pali, um weitere computerlinguistische Arbeitsschritte zu ermöglichen. Das Tool selbst kann jederzeit an die Verwendung für andere Sprachen angepasst werden.

Ferner stellen wir einen NoSQL-basierten Server zur Verwaltung von Wörterbuchdaten vor. Dieser ist als Komponente in einer klassischen Client-Server-Umgebung konzipiert und wird von den gleich nachfolgend erwähnten Werkzeugen verwendet: Ein Tool zur maschinellen Verarbeitung dieser Wörterbucheinträge, sowie einem Tool zur Visualisierung einzelner Datensätze. Der Server verwaltet dabei alle Wörterbucheinträge zentral und erlaubt dank seiner Bulk-Requests das effiziente Durchforsten der Daten auch bei größeren Datenmengen.

Ein Transformationswerkzeug, welches an den Server andockt, ist als IDE (Integrated Development Environment) konzipiert: Es erlaubt die Eingabe von C#-Programmcode-Fragmenten zur Datenverarbeitung. Diese Fragmente werden kompiliert; dann können sämtliche Wörterbucheinträge mit diesem Kompilat verarbeitet werden, um z.B. Muster zu erkennen und darauf basierend einzelne Wörterbucheinträge mit erkannten Informationen anzureichern. Eine Preview-Funktion gibt genauen Einblick darüber, auf welche Einträge sich die aktuell eingegebene Verarbeitungslogik erstreckt. Dadurch entsteht Transparenz: Erst der Einblick in die konkreten Änderungen über alle Datensätze hinweg erlaubt eine effiziente weil fehlerfreie Überarbeitung von Wörterbucheinträgen.

Ebenfalls an den Wörterbuchserver angegliedert ist ein Werkzeug zur Suche und Darstellung einzelner Wörterbuchartikel. Per serverseitig gespeicherter Konfiguration kann festgelegt werden, welche Controls in der GUI angezeigt werden sollen, und mit welchen Datenfeldern der einzelnen Artikel diese verbunden sein sollen: So kann eine Anpassung an Wörterbuchdaten beliebiger Struktur mit wenigen Handgriffen erfolgen. Die graphische Oberfläche erlaubt es, die manuelle Überarbeitungen auch größerer Artikelmengen auf einfachem Weg zu realisieren.

Ein anderes SeNeReKo-Tool unterstützt die Transformation beliebiger XML-Daten nach TEI: Über eine IDE-ähnliche Oberfläche können Umwandlungsregeln in Form eines Skripts eingegeben werden. Diese sind so gestaltet, dass sie fast schon natürlichsprachlich und somit leicht verständlich sind. Eine Erweiterbarkeit durch eigene Regeln ist jederzeit möglich: So können auch projektspezifische und über klassische X-Technologien möglicherweise nur schwer realisierbare Verarbeitungsprozesse durch ein Kommando repräsentiert werden (wie u.a. zur Verarbeitung im Pali in SeNeReKo). Angewandt auf eingelesene XML Datei(en) kann so die Aufbereitung von Daten erleichtert werden.

Des Weiteren stellen wir auf unserem Poster eine Reihe Tools vor, die dazu verwendet werden können, um aus TEI- bzw. TCF-Daten Netzwerke zu erzeugen. Sie stellen die Basis für den Kern des Projekts – die Erzeugung von Netzwerken dar. Da hier Standard-Graph-Datenformate für die Ausgabe verwendet werden, ist es möglich, diese Netzwerke dann anschließend mit verfügbaren Standard-Tools zu visualisieren.

Diese oben genannten Werkzeuge (bzw. Bibliotheken) sind frei nutzbar und helfen, gerade den teilweise sehr schwierigen Prozess der Datenaufbereitung zu adressieren. Wir würden uns freuen, wenn diese Komponenten nicht nur uns in SeNeReKo, sondern zukünftig auch anderen Wissenschaftlern helfen, damit so genau die Brücke geschlagen werden kann, die im Zentrum der Tätigkeit von uns Wissenschaftlern liegt: Die Brücke von Daten zu Erkenntnissen.

Annotationen für die automatisierte Verarbeitung von Märchen

Thierry Declerck, Universität des Saarlandes

(word count: 741)

In diesem Poster- und Demobeitrag fassen wir ältere und aktuelle Arbeiten zur Entwicklung eines Annotationsschemas für Märchen zusammen, das auch die Einbettung von Märchentexten in automatisierten Verarbeitungszenarien erlaubt. Eine Entwicklung unserer Arbeit in diesem Bereich führte zur automatischen Erkennung von Charakteren in Märchen, deren Rolle in Dialogen und deren Emotionen, die als Grundlage eines TextToSpeech Szenarios dient, das Märchentexte „vorliest".

Dieses Ergebnis basiert auf einer Zusammenarbeit mit Studenten der Computerlinguistik an der Universität des Saarlandes, die in den letzten Jahren in Form von Bachelor- oder Masterarbeiten, oder auch in Form eines Softwareprojekts erfolgten.

Angefangen hat es mit der Masterarbeit von Antonia Scheidel zur Annotation von Märchen mit Proppschen[1] Funktionen. Antonia Scheidel entwickelte ein neues Annotationsschemas, nach dem Märchen nach Texteigenschaften, temporalen Strukturen, Charakteren, Dialogen, und Proppschen Funktionen abfragen kann (s. [1]). Ein Annotationsschema ist insofern wichtig, als dadurch automatisierte Systeme ein Ziel haben, in das sie ihre Ergebnisse abbilden können. Wenn dazu auch Märchen mit dem Annotationsschema manuell annotiert werden, können die Ergebnisse der automatischen Verarbeitungen mit den menschlichen Annotationen verglichen werden.

Darauf aufbauend hat Nikolina Koleva an einem automatisierten System gearbeitet, das in Märchentext (sie hat mit 2 Beispielen gearbeitet; „The Magic Swan Geese", eine englische Version eines russischen Märchens, und „Väterchen Frost", eine deutsche Version eines russischen Märchens). Sie hat ein Programm geschrieben, dass der Text nach linguistischen Kriterien analysiert, mit dem Ziel, die darin vorkommenden Charaktere zu erkennen, und in eine Datenbank zu speichern. Diese Datenbank ist von der Sorte „Ontologie": darin können logische Operationen durchgeführt werden. Als Hintergrund fungiert eine formale Beschreibung dessen, was in den genannten Märchen vorkommen kann, inklusive eine Ontologie über Familienverhältnissen. So kann das System erkennen, dass im Text „die Tochter" die gleiche Person wie die „Schwester" ist, wenn der Kontext dies suggeriert. Erkannte Charaktere im Märchen werden somit mit allgemeineren Kategorien

[1] Auszug aus Wikipedia: „Propp gilt als Begründer der morphologischen oder strukturalistischen Folkloristik. Zwischen 1914 und 1918 studierte er russische und deutsche Philologie. Danach unterrichtete er die deutsche Sprache an verschiedenen Hochschulen in Leningrad. Von 1938 bis 1969 war er Professor für Germanistik, russische Literatur und Folklore an der Staatlichen Universität Leningrad.
1928 erschien sein bahnbrechendes Werk *Morphologie des Märchens*. Das Buch wurde 1958 in den USA in englischer Sprache veröffentlicht, was Propp weltweite Anerkennung verschaffte. 1946 erschien das Buch *Die historischen Wurzeln des Zaubermärchens*."
(http://www.wikiwand.com/de/Wladimir_Jakowlewitsch_Propp. Zugriff am 2014.11.10)

semantisch annotiert. Und wir wissen dann in welchen Kontexten (oder Situationen) die Tochter (zum Beispiel) involviert ist (s. hierzu [2]).

Schließlich eine Gruppe von Studenten (Christian Eisenreich, Jana Ott, Tonio Süßdorf und Christian Wilms) im Rahmen eines Softwareprojekts an Erweiterungen der oben genannten Arbeiten gearbeitet. Sie haben zum einem das Annotationsschema erweitert, mit detaillierteren Dialogbeschreibungen, und mit der Kodierung von Emotionen. Die Ontologie wurde auch erweitert, und sie inkludiert jetzt auch eine Beschreibung von Dialogen (Fragen, Antworten, Monologe, etc.), inklusive der Kodierungen der Teilnehmern und der Dialogwechseln. Auch 6 Basisemotionen (Angst, Trauer, Freude, etc) sind in der Ontologie kodiert.

Eine Haupterweiterung der vergangenen Arbeiten besteht darin, dass auch synthetische Stimmen eine Rolle spielen. Ist einmal ein Charakter erkannt worden, zum Beispiel die Prinzessin (im Märchen „Froschkönig"), werden zusätzliche Merkmale kodiert (zum Bsp. Alter, usw.). Dann wird automatisch eine vorher definierte synthetische Stimme zum Charakter addiert. Wenn dann der Text von dem System analysiert wird, kann die Geschichte von den Stimmen „erzählt" werden. Wenn kein Charakter in einer Dialogsituation vorkommt, dann wird angenommen, dass der Erzähler/die Erzählerin „daran" ist. Eine Demo kann hier gehört werden:

https://bytebucket.org/ceisen/apftml2repo/raw/763c5eb533f09997e757ec61652310c742238384/example%20output/audio_output.mp3

Im Anhang sind 2 Screenshots, die (für den ersten Teil der Audiodatei) zeigen wie das System den Text bearbeitet und kodiert, so dass die Sprachausgabe (s. Link oben) erzeugt werden kann. Unser Poster/Demo zeigt die Korrelation zwischen die Annotationen, die zum größten Teil automatisch generiert worden sind, und den verschiedenen Stufen der Verarbeitung bis hin zur Sprachausgabe.

Referenzen

[1] Thierry Declerck, Antonia Scheidel, Piroska Lendvai. **Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales.** *Language Technology for Cultural Heritage. Selected Papers from the LaTeCH Workshop Series,Theory and Applications of Natural Language Processing, Pages 155-169, Springer, Heidelberg, 2011*

[2] Nikolina Koleva, Thierry Declerck, Hans-Ulrich Krieger. **An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales** in: Jan Christoph Meister (ed.): *Digital Humanities 2012 Conference Abstracts, Pages 467-470, Hamburg.*

[3] Christian Eisenreich, Jana Ott, Tonio Süßdorf, Christian Willms, Thierry Declerck. **From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output** *Proceedings of ISWC 2014, Riva del Garda, Italy, Springer.*

Anhang

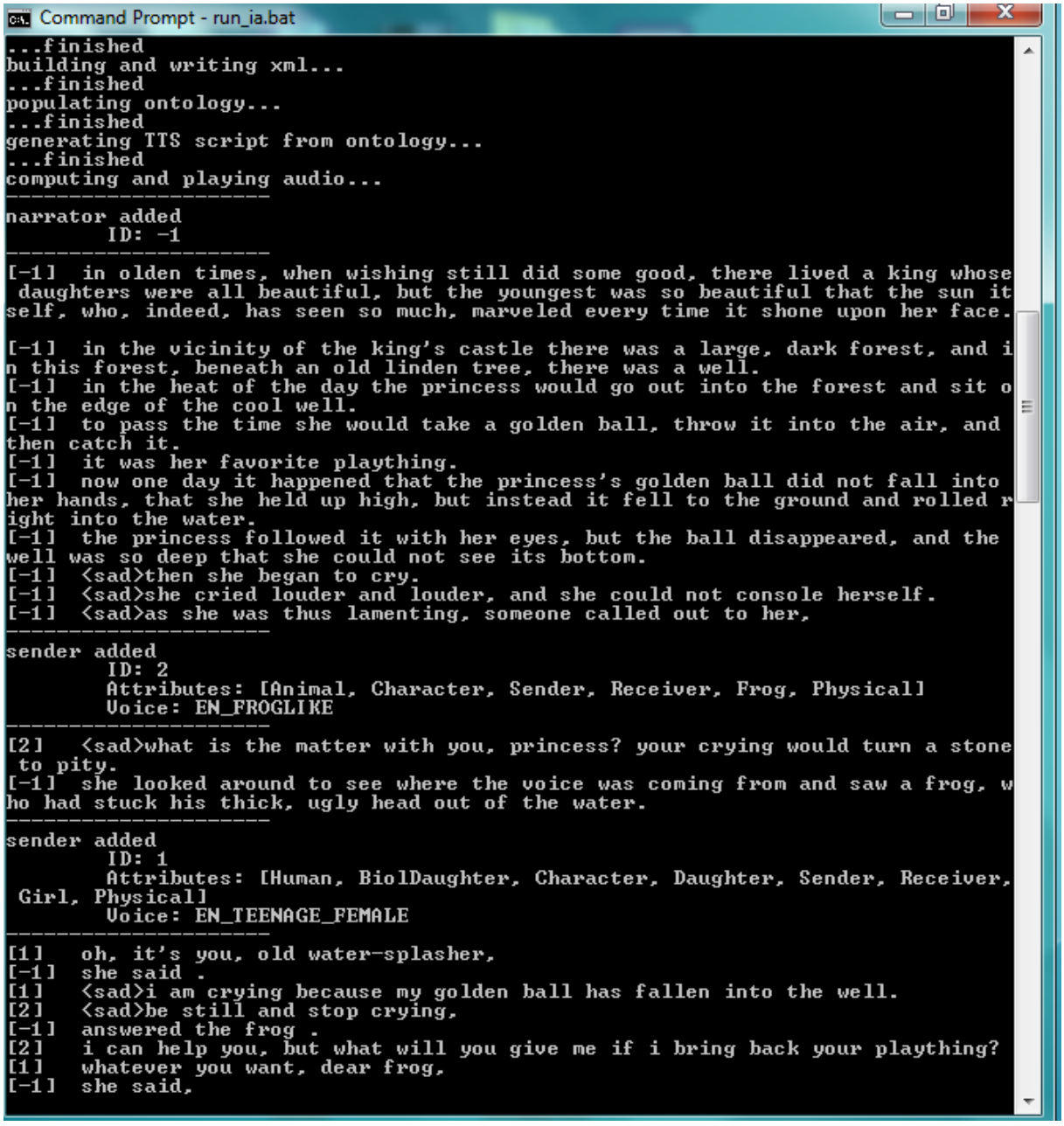


Abbildung 1: Wie der Text analysiert wird, Charaktere erkannt werden, sowie Dialogstrukturen und Emotionen. Die Basis für die Generierung der Sprachausgabe

Abbildung 2 Wie der Text analysiert wird, Charaktere erkannt werden, sowie Dialogstrukturen und Emotionen. Die Basis für die Generierung der Sprachausgabe (Fortsetzung von Abbildung 1)

Extended Abstract, Poster at DHd 2015 (Graz, Austria, February 23-27, 2015)

Musterforschung in den Geisteswissenschaften: Werkzeugumgebung zur Musterextraktion aus Filmkostümen

Johanna Barzen[1], Michael Falkenthal[1], Frank Hentschel[2], Frank Leymann[1]

Institut für Architektur von Anwendungssystemen
Universität Stuttgart
Nachname@iaas.uni-stuttgart.de[1]

Musikwissenschaftliches Institut
Universität zu Köln
Frank.Hentschel@uni-koeln.de[2]

1. Einleitung: Kostümsprache als Mustersprache

In der Literatur zum Filmkostüm findet sich immer wieder der Begriff der „Kostümsprache" als metaphorische Umschreibung der filmisch vestimentären Kommunikation. Wie diese aber funktioniert, welche Mittel das Kostüm nutzt, um Informationen über die Charaktere, deren Gruppenzugehörigkeit, Stimmungen oder Transformationen, sowie die Zeit- und Ortsgegebenheiten eines Films zu geben, ist nur rudimentär untersucht. Um sich den Funktionsweisen und etablierten Konventionen einer Kostümsprache im Film zu nähern, hat sich das Musterkonzept als fruchtbar erwiesen [SBL12].

Das Konzept des Musters, ursprünglich aus der Architektur stammend [AIS85], hat sich im Besonderen in der Informatik etabliert und findet hier vielseitige Anwendung (Cloud-Computing Patterns, Enterprise Integration Patterns etc.). Definiert wird ein Muster als ein einem vorgegebenen Format folgendes Problem-Lösungspaar, welches eine erprobte Lösung zu einem wiederkehrenden Problem abstrakt erfasst und dieses Wissen so effizient für andere nutzbar macht. Diese Muster werden mit anderen Mustern gleichen Formates untereinander in Beziehung gesetzt, so dass eine Mustersprache entsteht. Im Falle der Filmkostüme ist ein Kostümmuster eine abstrakte Beschreibung einer bewährten Lösung eines wiederkehrenden Designproblems einen adäquaten und schnell verständlichen textilen Ausdruck für beispielsweise eine bestimmte Rolle oder einen Charakterzug zu finden.

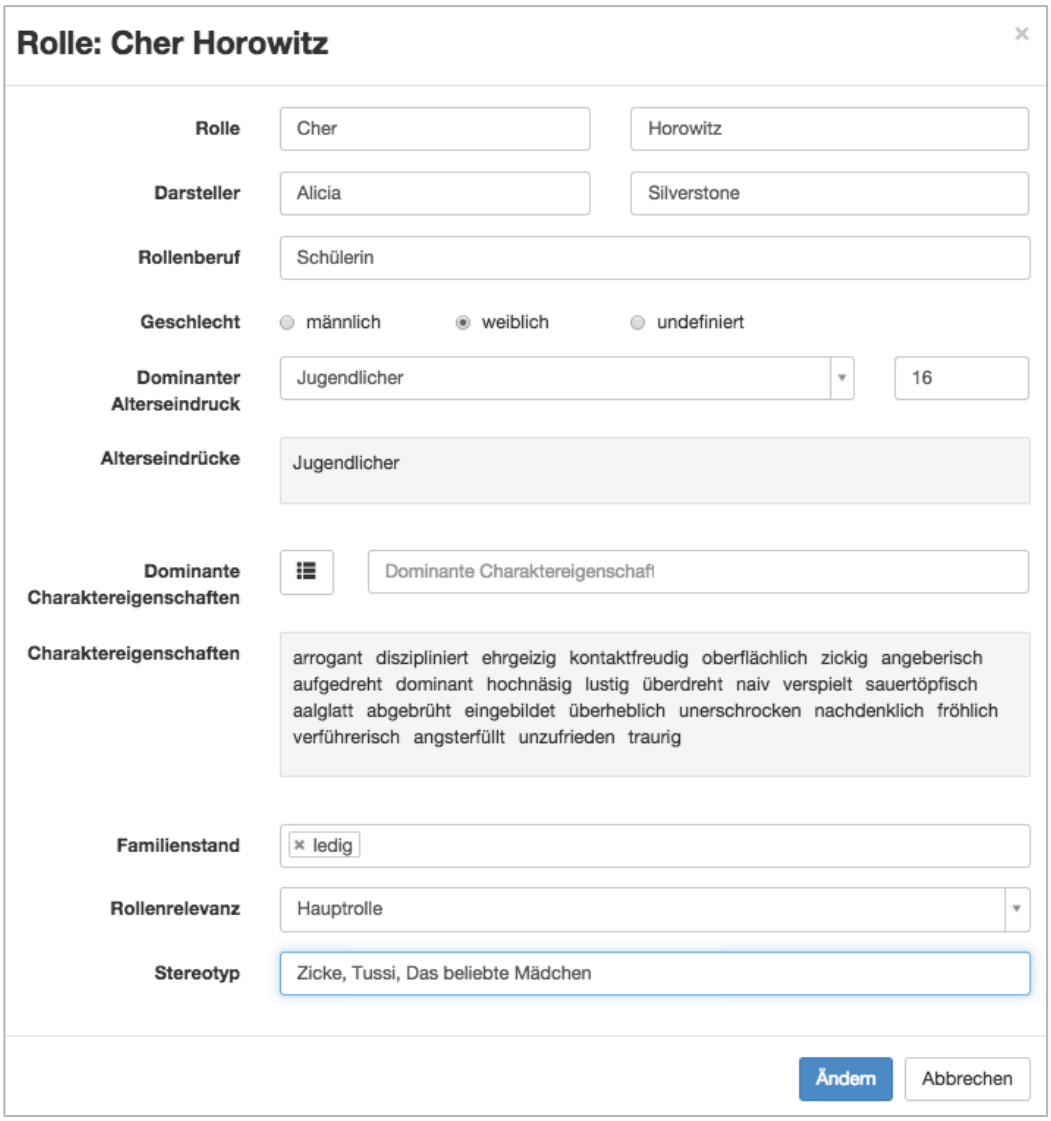
Um diese Kostümmuster als abstrakte Lösungsprinzipien (als Essenz vestimentärer Kommunikation) zu entwickeln, müssen erstens die ganzen konkreten Lösungen, in diesem Fall die konkreten Kostüme in Filmen, detailliert erfasst werden [FBB14]. Hierzu haben wir MUSE (MUster Suche und Erkennen) entwickelt. MUSE ist ein Kostümrepository, welches in Sektion 2 näher erläutert wird. Zweitens müssen die erfassten Daten aufbereitet und

ausgewertet werden, um daraus Muster abstrahieren zu können. Wie eine solche Analyse mittels OLAP Cubes aussehen kann wird in Sektion 3 vorgestellt.

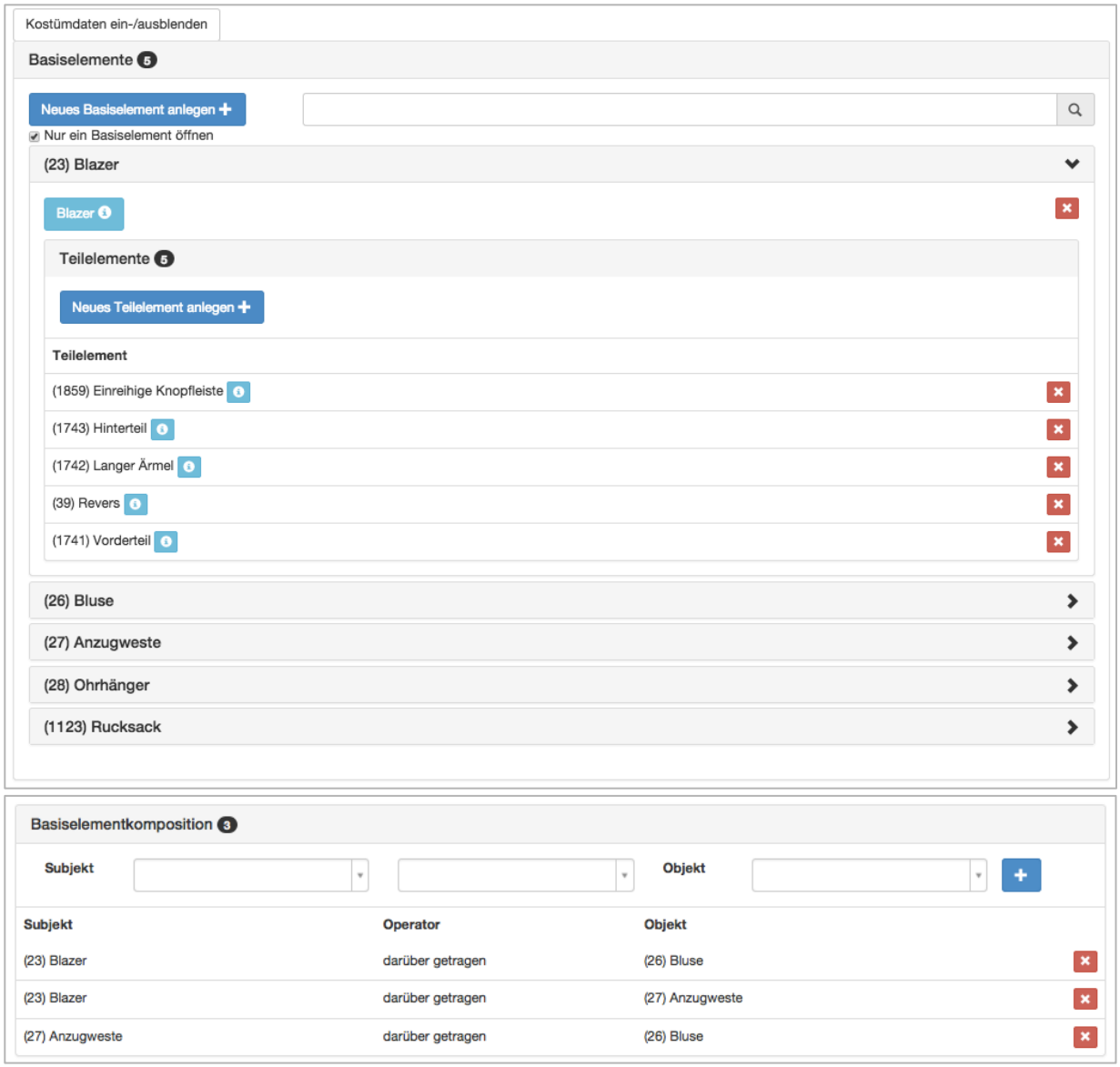
2. MUSE: Kostümrepository zur Kostümerfassung

MUSE ist ein, auf die Erfassung von Kostümen spezialisiertes Kostümrepository, das es ermöglicht, Film- und Rolleninformationen, vor allem aber detailgetreue Kostümbeschreibungen einzupflegen. Um eine strukturierte Erfassung und weiterführende Analyse dieser Daten zu ermöglichen, basiert MUSE auf einer umfassenden Bekleidungsontologie, in welche die konkreten Kostüme während des Einpflegens direkt als Instanzen dieser abgelegt werden [Ba13].

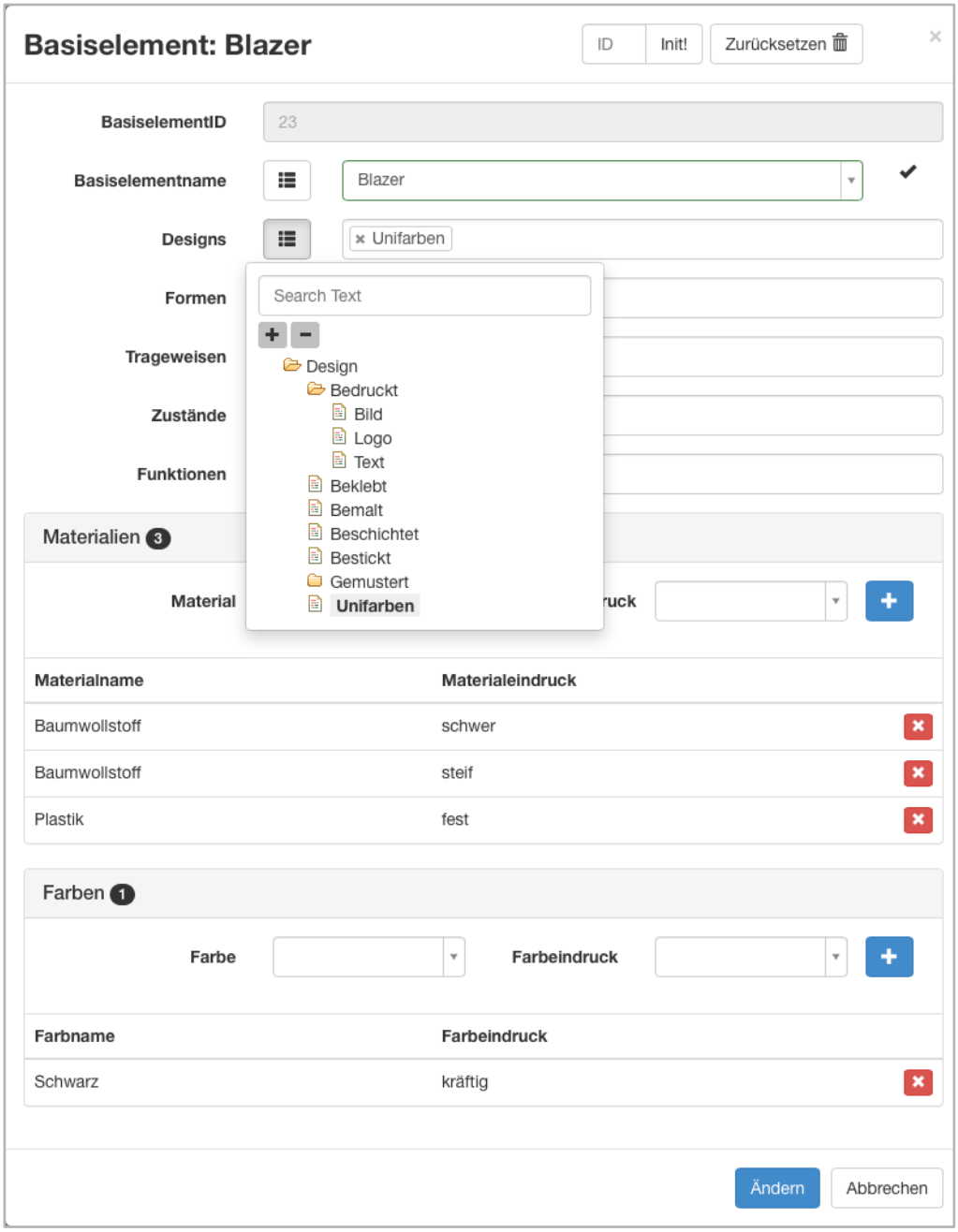
Die folgenden Screenshots sollen einen Eindruck vermitteln, wie MUSE die Erfassung von Kostümen unterstützt. Zur Zeit wird hier ein Filmkorpus von 60 Filmen unterschiedlicher Genres eingepflegt, wobei ein Film ca. 200 Kostüme aufweist, welche sich wiederum aus mehreren Basiselementen (Hose, Bluse, etc.) und deren Teilelementen (Ärmel, Kragen, etc.) zusammensetzen.



Screenshot 1: Eingabemaske zur detaillierten Erfassung von Rolleninformationen



Screenshot 2: Übersicht der Kostümkomposition aus Basis- und Teilelementen und deren Beziehungen zueinander (Operatoren)



Screenshot 3: Eingabemaske zur Basiselementerfassung (mit aufgeklappter Taxonomie-Eingabehilfe bei Design)

3. Analyse: Auswertung der Kostümdaten

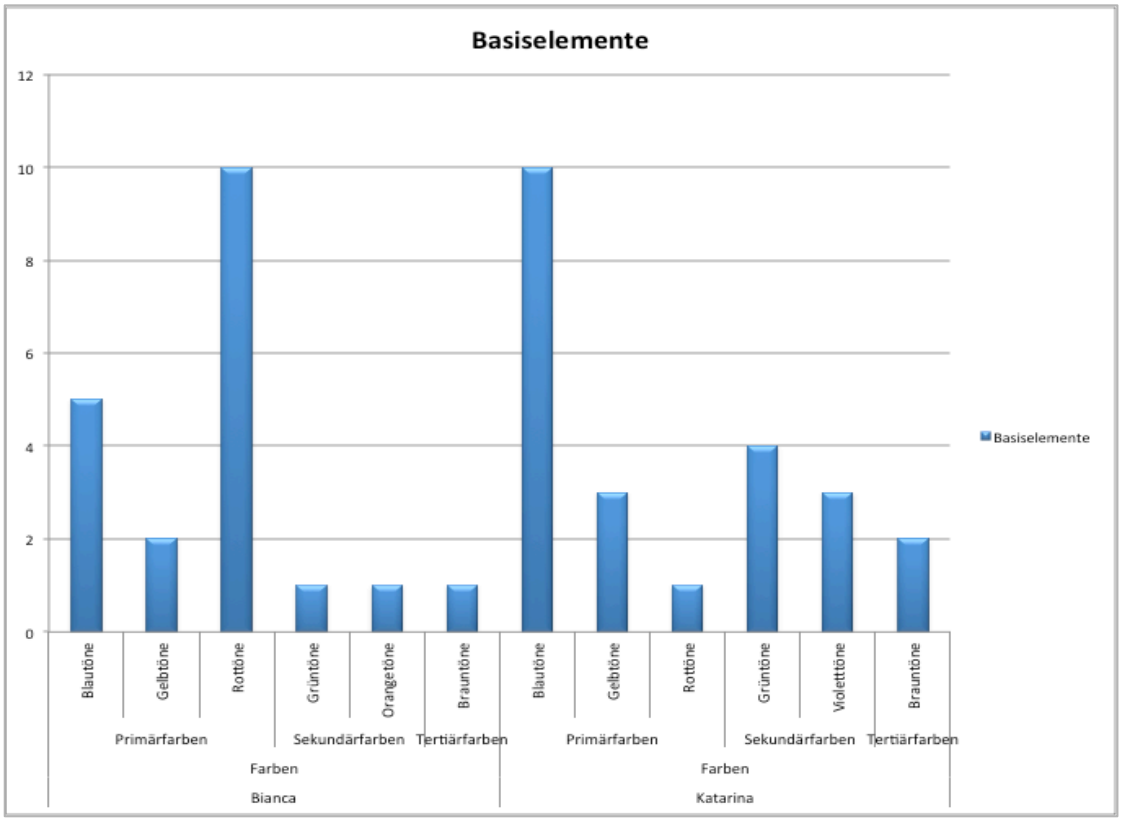
Um die mit MUSE erfassten Daten nutzerfreundlich in ihrem vollen Potential analysieren zu können, haben wir die Werkzeugumgebung so gestaltet, dass die ablaufenden informationstechnischen Auswertungsmethoden so viel Komplexität wie möglich für die Endanwender verbergen. Hierzu werden die Daten mittels eines OLAP Cubes so aufbereitet, dass mit Excel darauf zugegriffen werden kann und hier Auswertungsszenarien definiert werden können, welche die Analyse der Daten aus unterschiedlichen Blickwinkeln in all ihren Dimensionen und Verknüpfungen ermöglicht. Mittels Excel Pivot-Tabellen und als Pivot-Charts visualisiert, kann man sich nun Beispielfragestellungen wie „Welche Farbe ist am häufigsten mit welcher Charaktereigenschaft kombiniert?“, „Ist dieses kostümbildnerabhängig?“, „Werden hochgekrempelte Ärmel eher bei passiven oder aktiven Charakteren eingesetzt?“ nähern. Durch das Auftreten von Spitzen in der Häufigkeitsverteilung können dann erste Hinweise auf mögliche Muster gefunden werden.

Screenshot 4 und 5 verdeutlichen, wie man sich beispielsweise dem Einsatz von Farben im Verhältnis zu Charaktereigenschaften nähern kann. Gezeigt wird die Häufigkeitsverteilung der Farben der Kleider der beiden weiblichen Hauptrollen Katherina und Bianca aus „10 Dinge die ich an dir hasse“ (Regie: Junger, 1999).

	A	B	C
1			
2	Originaltitel	10 Things I Hate About You	
3			
4	BE Cube ID Distinct Count	Spaltenbeschriftungen	
5	Zeilenbeschriftungen	Basiselemente	Gesamtergebnis
6	**Bianca**	**15**	**15**
7	**Farben**	**15**	**15**
8	Primärfarben	14	14
9	Blautöne	5	5
10	Gelbtöne	2	2
11	Rottöne	10	10
12	Sekundärfarben	2	2
13	Grüntöne	1	1
14	Orangetöne	1	1
15	Tertiärfarben	1	1
16	Brauntöne	1	1
17	**Katarina**	**18**	**18**
18	**Farben**	**18**	**18**
19	Primärfarben	11	11
20	Blautöne	10	10
21	Gelbtöne	3	3
22	Rottöne	1	1
23	Sekundärfarben	7	7
24	Grüntöne	4	4
25	Violetttöne	3	3
26	Tertiärfarben	2	2
27	Brauntöne	2	2
28	**Gesamtergebnis**	**33**	**33**
29			

Screenshot 4: Pivot-Tabelle

Der unterschiedliche Einsatz der Rot- bzw. Blautöne bei den beiden charakterlich sehr divergierenden Schwestern, lässt bereits erste Rückschlüsse auf Konventionen in deren Einsatz zu. Dies ist allerdings nur als erster Hinweis zu verstehen, der mit weiteren Filmen, Rollen, Charaktereigenschaften, etc. zu überprüfen ist. Genau dabei unterstütz der OLAP Cube.



Screenshot 5: Pivot-Chart zu der Tabelle aus Screenshot 5

4. Ausblick

Zwar ist MUSE, als spezialisiertes Tool zur Kostümerfassung domänenabhängig, die dahinterstehende Methode und das Konzept des Musters zur Wissenserfassung und -repräsentation sind aber auch für die Anwendung in anderen Bereichen der Geisteswissenschaften ein vielversprechender Ansatz und gehen weit über die Kostümforschung hinaus [BL14].
Angedacht ist der Einsatz zur Extraktion von musikalischen Mustern, um Charakteristika und Topoi musikalischer Artefakte, die sich mit den herkömmlichen musikwissenschaftlichen Konzepten wie „Thema“, „Motiv“ oder „Stil“ nicht erfassen lassen, herauszuarbeiten und eventuell im Hinblick auf ihre semantische oder expressive Funktion deuten zu können.

5. Referenzen

[AIS85] Alexander, C.; Ishikawa, S.; Silverstein, M.; Jacobson, M.; Fiksdahl-King, I.; Angel, S.: A Pattern Language: Towns, Buildings, Constructions. Oxford University Press, 1977.

[Ba13] Barzen, J.: Taxonomien kostümrelevanter Parameter: Annäherung an eine Ontologisierung der Domäne des Filmkostüms, Universität Stuttgart, Technischer Bericht Nr. 2013/04, 2013.

[BL14] Barzen, Johanna; Leymann, Frank: Kostümsprache als Mustersprache: Vom analytischen Wert Formaler Sprachen und Muster in den Filmwissenschaften, In: DHd 2014.

[FBB14] Falkenthal, M.; Barzen, J.; Breitenbücher, B.; Fehling, C.; Leymann, F.: From Pattern Languages to Solution Implementations. In: Proceedings of the 6th International Conference on Pervasive Patterns and Applications, Venice, 2014.

[SBL12] Schumm, D.; Barzen, J. ; Leymann, F.; Ellrich, L.: A Pattern Language for Costumes in Films. In: Proceedings of the 17th European Conference on Pattern Languages of Programs (EuroPLoP), Irsee, 2012. ACM Press, New York, 2012.

Anforderungen und Bedürfnisse von Geisteswissenschaftlern an einen digital gestützten Forschungsprozess

Oona Leganovic, Viola Schmitt, Juliane Stiller, Klaus Thoden & Dirk Wintergrün
Max-Planck-Institut für Wissenschaftsgeschichte

Cluster 1 von DARIAH-DE[1] (Wissenschaftliche Begleitforschung) hat zum Ziel den geisteswissenschaftlichen Forschungsprozess zu analysieren um Bedürfnisse von Fachwissenschaftlern im Hinblick auf virtuelle Forschungsinfrastrukturen besser zu verstehen. Durch diese Arbeit sollen die innerhalb von DARIAH-DE entwickelten Dienstleistungen an die fachwissenschaftlichen Anforderungen angepasst werden. Um dies zu verwirklichen hat sich Cluster 1 die folgende drei Schritte vorgenommen:

1. Analyse der Beziehung zwischen geisteswissenschaftlichen Forschungsprozessen und den von digitalen Tools abgedeckten Prozessen,
2. Kartierung bisher genutzter digitaler Tools und Methoden um eventuelle Lücken aufzudecken,
3. Formulierung von Anforderungen an virtuelle Forschungsumgebungen zur Unterstützung des geisteswissenschaftlichen Forschungsprozesses.

Dieses Poster wird die Ergebnisse des ersten Arbeitsschrittes darstellen.
Es wurden vorhandene Modelle, die den geisteswissenschaftlichen Forschungsprozess konzeptionell erfassen auf ihre Gemeinsamkeiten und Unterschiede hin untersucht. Darauf aufbauend wurde ein auf unsere Bedürfnisse zugeschnittener Forschungskreislauf modelliert, der sowohl digitale als auch klassische Forschungsprozesse einbeziehen soll.

Es gibt eine Vielzahl von Modellen, die den Forschungsprozess vereinfacht darstellen und ihn auf Konzepte oder Aktivitäten reduzieren. Die Grundlage all dieser Überlegungen hat Unsworth mit seinen Primitiven gelegt (Unsworth, 2000). In eine ähnliche Richtung geht TADiRAH (Taxonomy of Digital Research Activities in the Humanities) - eine Taxonomie geisteswissenschaftlicher Forschungsmethoden und -ziele (Borek et al., 2014). Auch Bernardou und andere (2010) haben ein Modell entwickelt, das den geisteswissenschaftlichen Forschungsprozess abbildet mit besonderem Augenmerk auf die Ziele der beschriebenen Aktivitäten. Innerhalb des EU-geförderten Projektes DM2E[2] (Digital Manuscripts to Europeana) wurde in einem der Meilensteine das Scholarly Domain Model (SDM) beschrieben (Gradmann & Hennicke, 2012).

Wir haben die Modelle Unsworth's Primitives, TaDiRAH und das SDM aufeinander abgebildet um Gemeinsamkeiten und Unterschiede festzustellen. Abbildung 1 zeigt die Primitiven von Unsworth (2000), deren Verhältnis zu den Primitiven und Aktivitäten des SDM und der TaDiRAH Taxonomie. Man kann gut erkennen, dass es viele Überschneidungen, auch in der Terminologie, gibt. Weiterhin herrscht Einigkeit über die Aktivitäten, die während des geisteswissenschaftlichen Forschungsprozesses stattfinden. Auffällig ist, dass die Aktivitäten sich natürlich unterscheiden im Hinblick auf den Teil des Prozesses, den sie abbilden. So ist

[1] https://de.dariah.eu/
[2] http://dm2e.eu/

TaDiRAH sehr auf die Abbildung digitaler Arbeitsprozesse, die mit Software erledigt werden fokussiert und hat deswegen eine Aktivität “Storage”. Dies spielt in den anderen Taxonomien eine untergeordnete Rolle und ist oft Teil von anderen Aktivitäten, wie “Aggregation” beim SDM und “Sampling” bei Unsworth.

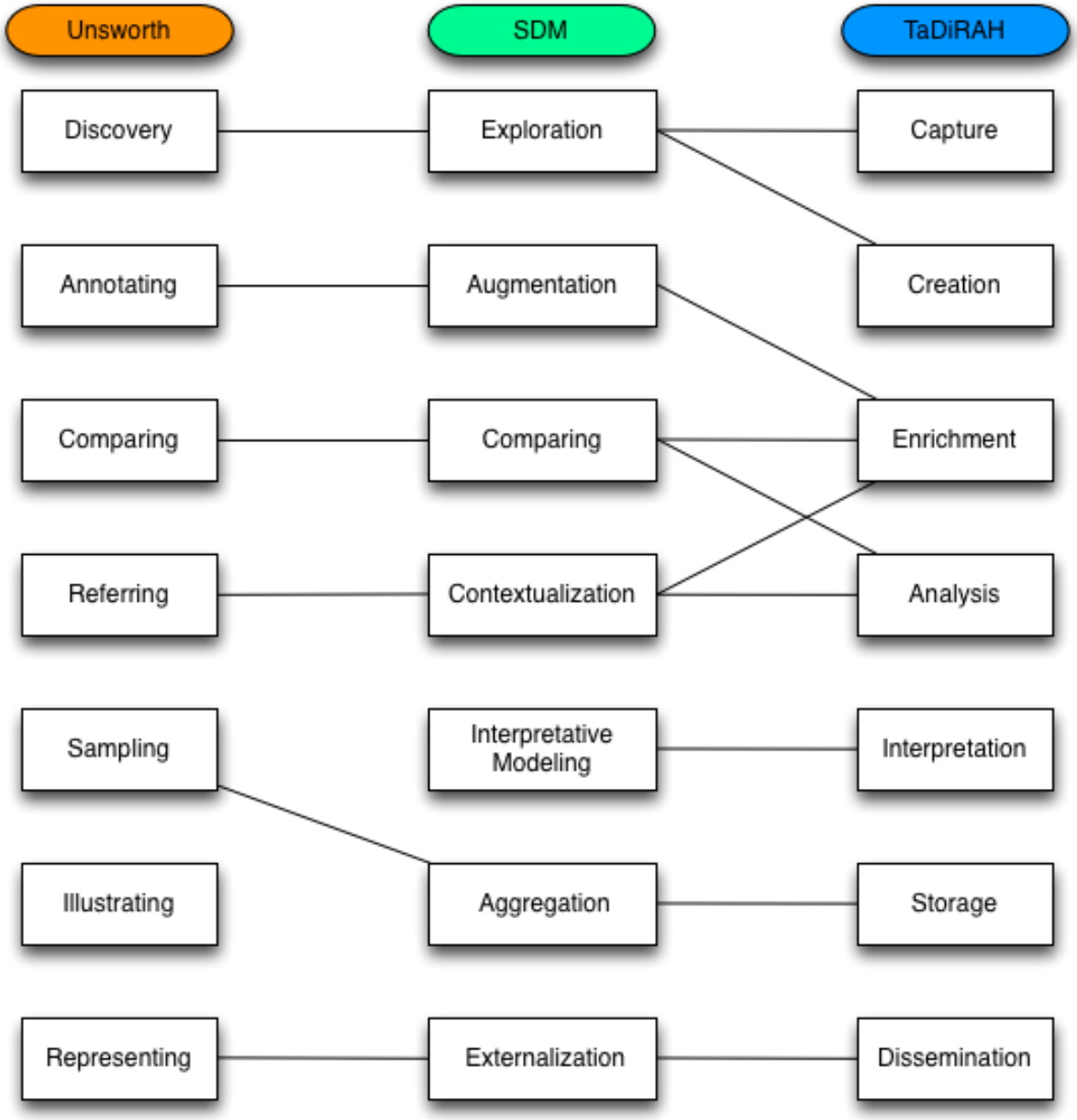


Abbildung 1: Abbilden der Unsworth’schen Primitiven, des Scholarly Domain Models und TaDiRAH

Die vorangegangen Modelle versuchen auf Basis der vorhandenen Tools Kategorisierungen zu erzielen oder auf Basis des Forschungsprozesses Arbeitsabläufe zu konzeptualisieren. Wir wollen anhand des geisteswissenschaftlichen Forschungsprozesses darstellen, was Tools

leisten müssen um diesen zu unterstützen. Ziel ist es Lücken aufzudecken und zu verstehen, wo digitale Dienstleistungen den Forschungsprozess besser unterstützen können und müssen.

Als ein Schritt zu diesem Ziel untersuchen wir, wie dieser prototypische Forschungsprozess sich in einem digitalen Arbeitsablauf abbilden lässt und wo sich Lücken befinden und wodurch diese entstehen. Dafür haben wir uns auf Grundlage der oben beschriebenen Modelle vor allem auf die Ergebnisse (und den Output) der verschiedenen Aktivitäten konzentriert (Abbildung 2).

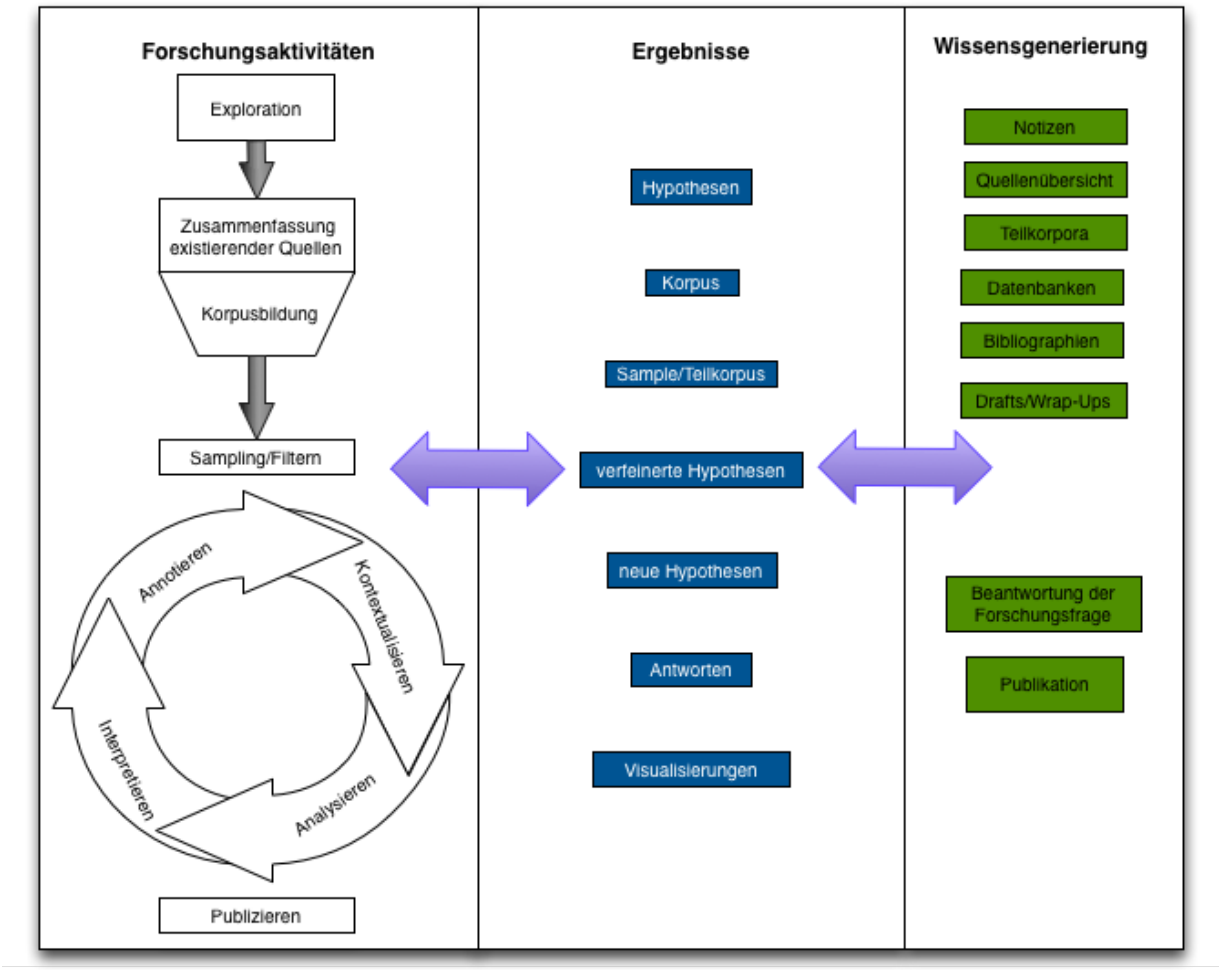


Abbildung 2: Forschungsaktivitäten, deren Ergebnisse und Output als Wissensgenerierung.

Es wurden die wichtigsten Ergebnisse, die in einem digitalen Prozess gespeichert und weiterverarbeitet werden, gelistet. Dabei ist zwischen den Zwischenergebnissen jedes einzelnen Schritts (Spalte 2 Abbildung 2) und dem Output, der in seiner gegenständlichen Form in der nächsten Aktivität verarbeitet wird (Spalte 3 Abbildung 2), zu unterscheiden. Dieses generierte Wissen kann mit anderen einzelnen Forschern aber auch der Öffentlichkeit geteilt werden. Dies kann die Publikation sein, die klassischerweise am Ende des Forschungsprozesses steht, aber auch Quellenübersichten, Datenbanken oder Bibliographien, die vor der Veröffentlichung angelegt werden. Häufig kann der Forschungsprozess nicht eindeutig modelliert und mit Aussagen versehen werden, die starr vorgeben, dass bestimmte Aktivitäten immer zu einer bestimmten Form des Ergebnisses und der Wissensgenerierung führen. Trotzdem ist es durchaus sinnvoll sich im Bezug auf die digitale Unterstützung des

Forschungsprozesses deutlich zu machen, dass jede Aktivität eine Form des Outputs produziert, auf dem der Forscher idealerweise im nächsten Schritt seines Denkprozesses aufbauen möchte. Dies gilt vor allem für den digitalen Arbeitsverlauf; lästiges hin- und her Kopieren und Konvertieren in verschiedene Datenformate beim Wechsel von Tools ist eines der Probleme, die Brüche im digitalen Forschungsprozess hervorrufen.

Mit diesem Poster möchten wir unsere Überlegungen vorstellen und zur Diskussion einladen, wie die Bedürfnisse von Fachwissenschaftlern in virtuellen Forschungsumgebungen Berücksichtigung finden und digitale Dienstleistungen aufgebaut werden können, die Geisteswissenschaftler in ihrer Arbeit unterstützen.

Literatur

Benardou, Agiatis; Constantopoulos, Panos; Dallas, Costis; Gavrilis, Dimitris (2010): A Conceptual Model for Scholarly Research Activity. iConference 2010. Online: https://www.ideals.illinois.edu/handle/2142/14945

Borek, Luise; Quinn Dombrowski; Matthew Munson; Jody Perkins; Christof Schöch (2014): Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects, Digital Humanities Conference 2014, Lausanne, Switzerland, July 7-12, 2014

Gradmann, Stefan; Hennicke, Steffen (2012): Intermediary Research Report on DH Scholarly Primitives (MS 3). Project DM2E.

Unsworth, J. (2000, May). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? Symposium on Humanities computing: Formal methods, experimental practice, King's College, London, Online: http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html

‚Was heißt und zu welchem Ende produziert man ein geisteswissenschaftliches E-Journal?'

Innovationspotentiale des digitalen Publizierens am Beispiel der *Zeitschrift für Digital Humanities* (ZfDH)

Constanze Baum (Wolfenbüttel), Timo Steyer (Wolfenbüttel)

Die *Digital Humanities* sind dabei, die geistes- und kulturwissenschaftliche Forschung grundlegend zu verändern. Durch die Inhalte und Methoden der *Digital Humanities* werden aber nicht nur neue Zugangswege, Fragen und Auswertungsmöglichkeiten zu bzw. an Primärquellen ermöglicht, sondern es eröffnen sich auch für die Präsentation und Publikation von Forschungsdaten und -ergebnissen innovative Alternativen zu traditionellen Printmedien. Die *Zeitschrift für Digital Humanities* (ZfDH) wird diese beiden Felder miteinander kombinieren, indem sie als dezidiertes Organ für die *Digital Humanities* im deutschsprachigen Raum nicht nur Themen der *Digital Humanities* veröffentlicht, sondern selbst ein Produkt der *Digital Humanities* ist: Hier werden neue Verfahren und Methoden digitalen Publizierens im Sinne einer Prototypentwicklung eines E-Journals für die Geisteswissenschaften ausgelotet. Das Poster wird sowohl die Innovationspotentiale des E-Journals zur Diskussion stellen, als auch über den gegenwärtigen Stand des Projektes informieren. Insofern scheint es gerechtfertigt, in Anlehnung an Schillers Antrittsrede vor 125 Jahren – „Was heißt und zu welchem Ende studiert man Universalgeschichte?" (1789) – programmatisch und grundsätzlich über Wege und Potentiale eines E-Journals im Bereich der Geisteswissenschaften nachzudenken und zu fragen: Was heißt und zu welchem Ende produziert man ein geisteswissenschaftliches E-Journal?

Es werden dabei Felder des digitalen Publizierens aufgezeigt, die die Bereiche der Beitragsakquise ebenso wie einen digital grundierten Workflow, ein offeneres Review-Verfahren und die vielversprechenden Möglichkeiten der E-Distribution der Zeitschrift betreffen. In den Geisteswissenschaften fehlen hier auf vielen Feldern noch Standards und Normen für E-Journale. Insofern versteht sich die ZfDH als Pilotprojekt und Innovationsgeber. In Anlehnung und Abgrenzung zu Projekten aus den Natur- und Technikwissenschaften sollen daher die Potentiale herausgearbeitet werden, die das digitale Publizieren in den Geisteswissenschaften haben kann. Denn im Bereich der Softwareentwicklung ist trotz verschiedener vorhandener Programme (*Open Journal System*) der Innovationsgrad längst nicht auf dem Niveau, wie er auf anderen Feldern der *Digital Humanities* bereits erreicht worden ist. Die Entwicklung der *Zeitschrift für Digital Humanities* beinhaltet daher sowohl die Ausarbeitung eines innovativen Workflows als auch dessen konkrete technische Umsetzung. Ausgegangen wird hierbei nicht von einer fertigen Softwarelösung, vielmehr wird die Software entsprechend den formulierten Anforderungen an das E-Journal modular entwickelt und aufgebaut.

Innovation im Bereich von wissenschaftlich orientierten E-Journals besteht vor allem in der freien Verfügbarkeit der Inhalte (OA) und der wesentlich schnelleren Publikation der Artikel, ohne dabei auf eine umfangreiche Qualitätskontrolle zu verzichten. Gedacht ist zurzeit an ein transparentes Review-

Verfahren, das es dem Autor ermöglichen wird, das jeweilige Gutachten einzusehen und darauf zu reagieren und bei Bedarf eine revidierte Fassung einzureichen sowie eine Gesamtbeurteilung der Gutachter für alle Nutzer öffentlich zu machen. Alle Fassungen bleiben mittels eindeutiger DOI-Nummern recherchier- und archivierbar. Vorab steht die Entscheidung, die redaktionsgeprüfte Erstfassung eines Artikels bereits nach einer ersten Routine online zu stellen. Die Qualitätskontrolle ist demzufolge im Sinne einer Liberalisierung von Wissensdiskursen (*Open Science*) als moderiertes *post-publication-peer-review*-Verfahren angedacht.

Um die Nachnutzung der Artikel zu gewährleisten, werden die Artikel unter einer freien Lizenz veröffentlicht und in XML bereitgestellt. Ob TEI sich auch als Standard für wissenschaftliche Sekundärliteratur im E-Journalbereich eignet, ist eine der zentrale Forschungsfragen des Projektes. Innovationspotentiale bestehen auch im Bereich weiterer Serviceleistungen, die Printmedien nicht bieten können, dazu zählen semantische Anreicherungen wie ein weitreichendes Verlinkungssystem, die Einbindung bestehender Normdaten und die Distribution der Zeitschrift über standardisierte Schnittstellen (Katalogisierung, Indexierung). Ein implementiertes Metriksystem liefert Angaben über die wissenschaftliche Nachnutzung einzelner Artikel. Die Artikel des E-Journals werden periodisch in Ausgaben zusammengefasst, dies dient vor allem der Erschließung und Distribution. Umfangreiche Suchfunktionen, Verschlagwortungen, Metadaten und Rubrizierungen bieten weitere digitale Möglichkeiten der Erschließung der Artikel, die parallel dazu zur Verfügung gestellt werden.

Darüberhinaus eröffnen sich für digitale Publikationen über den Text hinaus weitreichende Optionen für die Einbettung digitaler Medien, seien es Bilder, Videos, Tondokumente, Blogbeiträge oder Twitterfeeds. Die (dynamische) Aggregation unterschiedlicher Ressourcen bringt die Frage auf, wie eine persistente Identifizierung der einzelnen Bestandteile möglich sein wird und welche wissenschaftliche Relevanz diesem Quellenmaterial in der Forschung künftig zugewiesen wird. Es stellt sich demnach auch die Frage, inwieweit durch solche Formen digitalen Publizierens neue Inhalte für die wissenschaftliche Beschäftigung erschlossen werden können.

Wittgensteins Nachlass: Aufbau und Demonstration der FinderApp WiTTFind und ihrer Komponenten

Yuliya Kalasouskaya, Matthias Lindinger, Stefan Schweter, Roman Capsamun
Y.Kalasouskaya1@campus.lmu.de, matthias.lindinger@campus.lmu.de,
Stefan.Schweter@campus.lmu.de, r.capsamun@campus.lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

1 Einleitung

Das von uns erstellte Poster soll den Aufbau und Einsatz der FinderApp WiTTFind mit den zugehörigen WAST-Tools[1] als *open source* Tool vorstellen. Im Mittelpunkt stehen die optimierte Browseroberfläche, zugrunde liegende Texte der FinderApp, Faksimile mit OCR, Faksimile Reader und den Einsatz des Finders als open source Programm. Für Interessierte werden wir die FinderApp vorführen.

2 Neuerungen der FinderApp

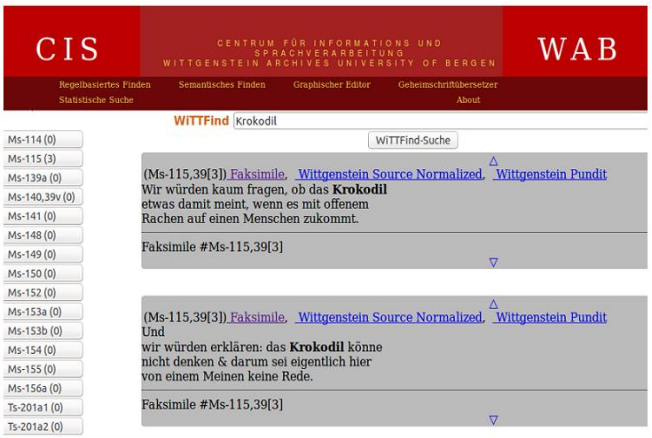
Seit 2 Jahren wird mit unserem Finder in der Nachlassforschung von Ludwig Wittgenstein gearbeitet und die von uns entwickelten Programme werden stetig optimiert. Eine zusätzliche Motivation für die Weiterentwicklung und Erweiterung war auch die Verleihung des EU-AWARDs 2014, der vom EU Projekt Digitised Manuscripts to Europeana (DM2E) ausgeschrieben wurde. Neuerungen des Finders bestehen darin, dass die Weboberfläche optimiert wurde, mehrere Dokumente parallel durchsucht werden können und eine lemmatisierte symmetrische Vorschlagssuche sowie ein neuer Faksimile-Reader integriert wurden. Die wichtigste Neuerung bei unserem Finder ist jedoch, dass WiTTFind für andere Forschungsprojekte geöffnet wurde und als open source für andere Projekte der Digital Humanities einsetzbar sein wird. Die Applikation ist unter dem folgenden Permalink zu erreichen:

http://wittfind.cis.uni-muenchen.de:

3 Die Komponenten des Finders

3.1 Benutzeroberfläche

Als erstes wollen auf dem Poster wir die Gestaltung der neuen benutzerfreundlichen und interaktiven Hauptseite der FinderApp vorstellen. Um die Bedienung der Anwendung übersichtlich zu gestalten, werden den Nutzern verschiedene Suchumgebungen angeboten. In der neuen Version können mehrere Text-Ressourcen parallel durchsucht werden, weshalb die Applikation um eine *multidoc*-Struktur erweitert wurde. Die Darstellung der Treffer wird auf die ausgewählten Dokumente beschränkt. In der nächsten Abbildung ein Beispiel zur *multidoc* Oberfläche:



Die Antwort auf die Benutzeranfrage wird mit Hilfe des *LocalStorage* Konzepts im Browser gespeichert. Bearbeitet werden die Daten und die *multidoc* Funktionen nur auf der Client Seite (Web-Browser) ohne Server-Zugriff. Zur Interaktivität und Lebendigkeit der Seite tragen die modernen Techniken von *JQuery* und *HTML5* bei.

[1] Wittgenstein Advanced Search Tools

3.2 Faksimile-Reader

Das zweite Thema des Posters stellt den Faksimile-Reader (siehe Bild:Faksimile-Reader) vor, der es erlaubt komplementär durch die Faksimile der Edition zu blättern und gleichzeitig die gefundenen Textstellen im Bild hervorzuheben. Dieser ist in Javascript sowie den Bibliotheken *jQuery* und *turn.js* programmiert. Zum *Highlighting* der einzelnen Treffer wird eine Liste von Koordinaten verwendet, die im Javascript-eigenen JSON-Format vorliegt. Zur schnellen Darstellung der Faksimile werden immer nur die Seiten geladen, die der Anfrage des Benutzers entsprechen. Somit kann der Anwendungsnutzer das komplette Dokument in Faksimile-Form durchblättern. Weitere Features des Readers sind eine dynamische Anpassung der Ansicht an das Browserfenster und eine kurze Bedienungsanleitung, die beim Start angezeigt wird. Damit die Faksimile zusammen mit den gefundenen, farblich hervorgehobenen Treffern dargestellt werden können, müssen die Faksimile mit der open source OCR Software *tesseract* verarbeitet werden. Je nach Qualität der Faksimile müssen die extrahierten OCR-Ergebnisse manuell nachbearbeitet werden. Dazu haben wir spezielle Tools entwickelt.

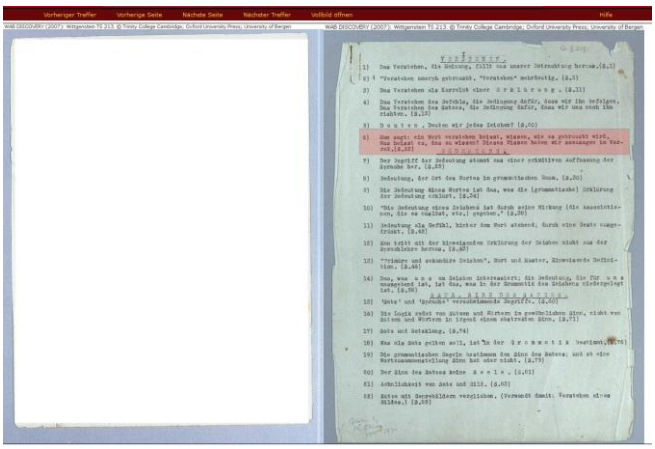


Bild:Faksimile-Reader

4 Unsere FinderApp für andere Digital Humanities Projekte

In einem weiteren Thema des Posters geht es um eins der wichtigsten Ziele unseres Projekts: Die FinderApp und die WAST-Tools sollen plattformunabhängig einem breiten Forschungskreis zur Verfügung stehen.

4.1 Die Texte der Edition

Die FinderApp findet Wörter, semantische Begriffe und Satzphrasen über mehrere Dokumente hinweg, sofern die Dokumente in dem XML-TEI-P5 Format vorliegen. Dieses XML-Format wird von uns CISWAB genannt und in einer eigenen *Document Type Definition* (DTD) beschrieben.

4.2 Elektronisches Vollformenlexikon

Zu den Texten einer Edition benötigt die FinderApp ein elektronisches Lexikon im DELA Format[2] Das CIS verfügt über das größte deutsche Vollformenlexikon, das bei der Entwicklung eines eigenen „Editionslexikons“ herangezogen werden kann.

4.3 Software-Kompatibilität unserer FinderApp

Da bei unseren Programmen eine große Anzahl unterschiedlicher Programmiersprachen und Libraries im Einsatz sind, die von Standarddistributionen abweichen, setzen wir die quelloffene Containervirtualisierungssoftware namens *docker* ein. Bei dieser Technologie werden alle von WiTTFind benötigten Programme, Module und Libraries in einem Softwarecontainer zusammengefasst. Jeder Anwender, der auf seinem Rechner die *docker* Software.[3] installiert hat, kann unseren Finder mit WAST-Tools lokal auf seinem Rechner einsetzen. Zur Entwicklungsverwaltung verwenden wir das

[2] Laboratoire d'Automatique Documentaire et Linguistique, Paris

[3] https://www.docker.com/

Versionsverwaltungsprogramms Git und das web-basierte Versionsverwaltungs-Management-Werkzeug GitLab.

Am Posterstand werden wir auf Laptops mit verschiedenen Betriebssystemen unsere FinderApp WiTTFind und WAST-Tools vorstellen.

5 PUBLIKATION UND AWARD

EU AWARD: http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal: Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST). Madrid, DATeCH 2014: 91-96 http://wast.cis.uni-muenchen.de/tutorial

Computerlinguistische Verfahren zur Aufdeckung struktureller Ähnlichkeiten in Narrativen

Einführung

In diesem Beitrag stellen wir eine Methode zur automatischen Erkennung von strukturellen Ähnlichkeiten narrativer Texte auf der Handlungsebene vor. Dafür operationalisieren wir strukturelle Ähnlichkeiten als (intertextuelle) Verbindungen (*Alignments*) zwischen Ereignissen. Die verwendeten Alignierungsalgorithmen bauen auf automatisch erzeugten linguistischen Analysen der Texte auf und verwenden als Kriterien Eigenschaften verschiedener linguistischer Ebenen. Ziel unseres Ansatzes ist es, materiell in Texten vorliegende Ähnlichkeiten auffindbar zu machen und hervorzuheben, so dass sie von Wissenschaftlerinnen und Wissenschaftlern zielgerichtet analysiert und interpretiert werden können.

Anwendungsszenarien

Die Untersuchung struktureller Ähnlichkeiten zwischen Narrativen spielt in vielen geisteswissenschaftlichen Disziplinen eine Rolle. Als Beispielszenarien verwenden wir die Märchen- und Ritualforschung.

Ähnlichkeiten zwischen **Märchen** sind auf verschiedenen Granularitätsebenen untersucht worden. Propp (1958) veröffentlichte eine Analyse, in der in russischen Märchen prototypische Handlungen und Charaktere identifiziert werden. Regelmäßigkeiten im Auftreten von Handlungen und Charakteren werden in einer sog. „Morphology of the Folktale“ erfasst. Damit sollen typische Handlungsmuster (Ereignis X folgt auf Ereignis Y) beschrieben werden. Am anderen Ende der Granularitätsskala existieren Sammlungen wie der ATU-Index (Uther, 2014), in dem Märchen mit gleichen Handlungselementen (Aussetzen von Kindern) oder Charakteren (Lebkuchenhaus) in Klassen zusammengefasst werden.

Im Bereich der **Ritualforschung** werden Rituale aus diversen religiösen, kulturellen oder politischen Kontexten untersucht. Unter dem Stichwort „Ritualgrammatik“ (vgl. Hellwig und Michaels, 2013) wird diskutiert, dass in verschiedenen Ritualen ähnliche Handlungen vorkommen und Teilnehmer ähnliche Rollen übernehmen. Verschiedene Forscher vertreten die Auffassung, dass die Zusammensetzung wiederkehrender Ereignisse zu Ritualen Regeln folgt. Existierende Überlegungen zur Ritualgrammatik sind nicht formalisiert und daher für eine automatische Analyse nur begrenzt nutzbar.

Um unsere Methode entwickeln und testen zu können, haben wir für diese beiden Szenarien ein englischsprachiges Korpus zusammengestellt, das mehrere Beschreibungen des gleichen Typs enthält (ATU-Märchenklasse bzw. Ritualtyp).

Computerlinguistische Verarbeitung

Wir wenden die gleichen computerlinguistischen Komponenten auf beide Korpora an. Damit werden linguistische Repräsentationen für Wortarten, (syntaktische) Dependenzrelationen, semantische Rollen, Wortbedeutungen und Koreferenzketten erstellt. Verknüpft ergeben diese Annotationen eine Diskursrepräsentation, die als Basis für die Alignierungsverfahren verwendet wird. Da Ritualbeschreibungen untypische linguistische Phänomene enthalten, wurden sämtliche Komponenten auf die Domäne angepasst (*Domain Adaptation*). Dadurch konnten deutliche Qualitätssteigerungen der computerlinguistischen Analyse erreicht werden.

Alignierungsexperimente

Drei Alignierungsalgorithmen mit unterschiedlicher Mächtigkeit wurden verglichen: *Sequence alignment* (Needleman-Wunsch, 1970) ist der einfachste Algorithmus, der ausschließlich paarweise und nicht-kreuzende Alignierungen erzeugen kann. *Graph-based predicate alignment* (GPA; Roth, 2014, Roth & Frank, 2012) kann paarweise und kreuzende Alignierungen erzeugen. *Bayesian model merging* (BMM; Stolcke & Omohundro, 1993) ist der mächtigste Algorithmus, der Alignierungen beliebiger Länge mit Überkreuzungen erzeugen kann. Diese drei Algorithmen wurden in zwei Experimenten evaluiert: In einer intrinsischen Evaluation wurden die Ergebnisse mit einem von zwei Ritualwissenschaftlern parallel erzeugten Goldstandard verglichen (κ=0.61). Dabei erzielte BMM die besten Ergebnisse insgesamt und GPA die besten Ergebnisse auf einem Einzeldokumentpaar.
Im zweiten Experiment wurde aus den automatisch erzeugten Alignierungen ein Maß für Dokumentenähnlichkeit berechnet und in einem Clustering-Verfahren eingesetzt. Das Ergebnis des Clusterings – eine Einteilung der Dokumente auf Basis der errechneten strukturellen Ähnlichkeit – konnte dann mit der Gruppierung verglichen werden, die „natürlicherweise“ in den Korpora vorkommt (Ritualtypen bzw. ATU-Klassen). Dabei zeigten sich wieder GPA und BMM als die leistungsstärksten Algorithmen.

Visualisierung und Nutzung

Um es Wissenschaftlerinnen und Wissenschaftlern aus der Ritual- bzw. Märchenforschung zu ermöglichen die Analysen zu nutzen, haben wir Visualisierungen entwickelt, die eine systematische Untersuchung der gefundenen Ähnlichkeiten ermöglichen. Auf einer Vogelperspektive stellen wir die Dokumentenähnlichkeit in einer Heatmap dar. Auf interessante, dicht verknüpfte Stellen können wir hinweisen, indem für jedes Ereignis ein *connectivity score* in einem Diagramm anzeigt wird. Eine detaillierte Darstellung der Einzelereignisse (mit Teilnehmern und Kontext-Ereignissen) ist ebenfalls möglich. Direkt aus der Diskursrepräsentation können wir außerdem eine Visualisierung des sozialen Netzwerks erzeugen, in der wichtige Entitäten (Charaktere, Gegenstände und Materialien) in einem Netzwerk angezeigt und gemeinsam auftretende Figuren verknüpft werden.

Konklusion

Der Posterbeitrag präsentiert eine Methode zur Erkennung struktureller Ähnlichkeiten zwischen narrativen Texten. Die Ähnlichkeiten werden basierend auf computerlinguistischen Analysen vollautomatisch identifiziert und können zielgerichtet auf unterschiedlichen Granularitätsebenen dargestellt und manuell inspiziert werden. Damit eignet sich die Methode auch zur Analyse von größeren Datenmengen, ohne bestimmte Interpretationen vorwegzunehmen. Eine ausführliche Darstellung des Verfahrens sowie des geisteswissenschaftlichen Anwendungskontexts findet sich in Reiter (2014) und Reiter et al. (2014).
Auf einer methodischen Ebene zeigt sich in diesem Projekt, dass komplexe linguistische Analysen auch für nicht-kanonische Textsorten erstellt werden können und eine vielversprechende Ausgangsbasis für Analysen darstellen. Die Besonderheiten natürlicher Sprache (z.B. Ambiguität, Vielseitigkeit) stellen für automatische Verarbeitung eine große Herausforderung dar, werden aber in der Computerlinguistik bereits untersucht. Auf (computer-)linguistische Analysen aufzubauen erlaubt die Untersuchung komplexer semantischer Phänomene, die vergleichsweise eng mit den Zielkategorien vieler Geisteswissenschaften verwandt sind.

Bibliographie

Oliver Hellwig and Axel Michaels. Ritualgrammatik. In Christiane Brosius, Axel Michaels, and Paula Schrode, Hrsg., *Ritual und Ritualdynamik*, S. 144–150. Vandenhoeck & Ruprecht, Göttingen, Germany, 2013.

Saul B. Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, March 1970.

Vladimir Yakovlevich Propp. *Morphology of the Folktale*. University of Texas Press, Austin, TX, 2nd edition, 1958. Translated by Laurence Scott (Original work published 1928).

Nils Reiter. *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. PhD thesis, Heidelberg University, June 2014.

Nils Reiter, Anette Frank, and Oliver Hellwig. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605, 2014.

Michael Roth. *Inducing Implicit Arguments via Cross-document Alignment – A Framework and its Applications*. PhD thesis, Heidelberg University, 2014.

Michael Roth and Anette Frank. Aligning predicates across monolingual comparable texts using graph-based clustering. In Jun'ichi Tsujii, James Henderson, and Marius

Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 171–182, Jeju Island, Korea, July 2012.

Andreas Stolcke and Stephen Omohundro. Hidden markov model induction by bayesian model merging. In Steve J. Hanson, J. D. Jack D. Cowan, and C. Lee Giles, Hrsg., *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, California, 1993.

Hans-Jörg Uther. *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. Number 284–286 in FF Communications. Suomalainen Tiedeakatemia, Helsinki, 2004.

Netzwerke sehen

Matej Ďurčo, ACDH-ÖAW

In diesem Beitrag stellen wir eine Webapplikation zur Visualisierung und interaktiven Erkundung von Graphen und Netzwerken vor. Die Applikation ist ursprünglich im Kontext der Forschungsinfrastruktur CLARIN entstanden, mit dem Ziel die komplexe Datendomäne der Metadaten-Profile der Component Metadata Infrastructure[1] (CMDI) (Broeder et al., 2010) besser fassbar zu machen. (Ďurčo, 2013). Im Laufe der Entwicklung hat sich diese Applikation zu einem generischen Viewer für jede Art von graph-basierten Daten weiterentwickelt.

Die Applikation kann auch im Vergleich mit alternativen weit verbreiteten Tools bestehen. Gephi[2] bietet zwar wesentlich mehr Funktionalität zum automatischen Analysieren von Graphen, ist aber eine Client-Applikation, die lokal installiert wird, und die Möglichkeiten der dynamischen Navigation im Graphen sind auch nicht so reichhaltig, wie in der vorgestellten Applikation. Die traditionelle command-line Applikation GraphViz[3] ist zwar sehr stark im eleganten Layoutieren der Graphen, ist aber eine rein statische Anwendung ohne graphisches User Interface, bietet also keine Möglichkeit interaktiv zu arbeiten.

Die Applikation basiert auf der open-source javascript Bibliothek d3[4] und läuft nach dem anfänglichen Laden vollständig client-seitig. Es bietet mehrere miteinander verknüpfte Ansichten und eine Reihe von Optionen zum Manipulieren der dargestellten Graphen. So ist es möglich mehrere Knoten auszuwählen und sich aus dem zugrundeliegenden geordneten Graphen beliebig viele Ebenen von Vorgänger- bzw. Nachfolgerknoten anzeigen zu lassen. Ebenfalls werden mehrere vordefinierte Layout-Algorithmen angeboten. Das Layout kann die Stärke der Verbindungen reflektieren, ebenso kann die Größe und Farbe der Knoten verwendet werden, um weitere Dimensionen visuell zu kodieren. Der aktuell angezeigte Graph, kann entweder als Link verschickt oder als SVG-Grafik exportiert und weiter verarbeitet werden.

Daten

Neben den CMDI Metadaten, für welche die Applikation ursprünglich vorgesehen war, wurde die Applikation bereits erfolgreich mit ganz anderen Datensätzen erprobt. So wurde zum Beispiel das „Philosophen-Influenz-Netzwerk“ visualisiert. Dafür wurden die Daten aus der dbpedia über den SPARQL-Endpoint abgefragt (Philosophen und ihre *influenced/influencedBy* Beziehungen), diese wurden durch eine einfache XSLT-Transformation in das Eingabeformat umgewandelt und in die Applikation importiert. Mit minimalem Aufwand konnte dergestalt eine große Datenmenge wesentlich besser erfasst und erkundet werden als dies mit konventionellen Instrumenten möglich wäre. (sieh Abb. 1)

Es ist vorgesehen, weitere Datensätze aufzubereiten und sie über die Applikation verfügbar zu machen. Dies wären zum einen Taxonomien, die in unterschiedlichen DH-Disziplinen Verwendung findet. Hier bietet sich SKOS als primäres Input-Format an, da es weite verbreitet ist und entsprechende Transformationen von SKOS in das interne Graph-Format eine ganze Klasse von Datensätzen für die visuelle Erkundung in der Applikation zugänglich machen würde. Eine weitere Klasse potentieller Daten sind prosopographische Annotationen, durch deren Auswertung sogenannte Kookurenznetzwerke visualisiert werden können. Hierfür wurden schon Experimente mit Daten aus dem Schnitzler Tagebuch (8.500 Personen mit ca. 77.000 Nennungen) und aus der Zeitschrift ‚Die Fackel‘ (15.000 Personen mit über 123.000 Nennungen) durchgeführt.

[1] http://clarin.eu/cmdi
[2] https://gephi.github.io/
[3] http://www.graphviz.org/
[4] https://github.com/mbostock/d3/

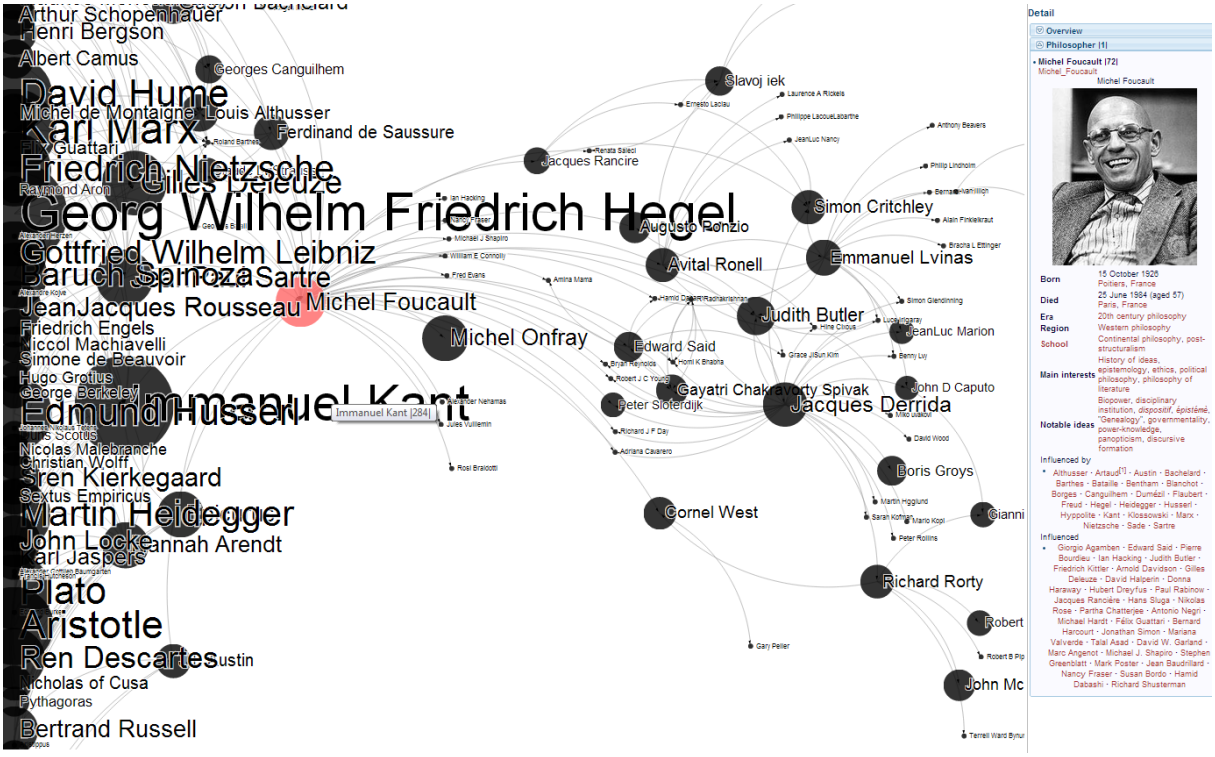


Abb 1 Screenshot der Visualisierung des "Philosophen-Influenz-Netzwerks". Es werden die VorgängerInnen und NachfolgerInnen von Michel Foucault dargestellt.

Nächste Schritte

Die Applikation ist bereits lauffähig und wird in ihrem ursprünglichen Kontext produktiv eingesetzt[5], sie wird aber auch laufend weiterentwickelt. So ist eine Erweiterung geplant, um verbreitete standardisierte Graph-Formate (wie GraphML, GDF, GML) als Input und Output Formate- zu unterstützen. Ebenso ist ein Refactoring des Codes notwendig, um die domänenspezifischen Aspekte von der generischen Applikation zu trennen und diese als ein sauberes wiederverwendbares konfigurierbares javascript-Modul anzubieten, das leicht in komplexere Applikationen eingebaut werden kann. Im Hinblick auf die eigentliche Visualisierung ist es wünschenswert, andere Darstellungsformen für die Knoten (momentan nur Kreise) anzubieten. Für tiefgreifende Analysen sind Graph-Operationen erforderlich, z.B. Vergleich von zwei Graphen und das Berechnen des gemeinsamen Subgraphen, Clustering u.ä.

Die Applikation wird als eigener Visualisierungsservice, der auch externe Daten annehmen, verarbeiten und visualisieren kann, als einer der Dienste des ACDH-ÖAW angeboten werden. Der Code wird open-source frei zur Verfügung gestellt.

Referenzen

Broeder, D.; Kemps-Snijders, M.; Uytvanck, D. V.; Windhouwer, M.; Withers, P.; Wittenburg, P. & Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In Calzolari, N.; Choukri, K. & others (Eds.). Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA).

Ďurčo, M. (2013). SMC4LRT - Semantic Mapping Component for Language Resources and Technology. *Technical University, Vienna*.

[5] http://clarin.oeaw.ac.at/smc-browser

Ontologiegestütze geisteswissenschaftliche Annotationen mit dem OWLnotator

Giuseppe Abrami Alexander Mehler Susanne Zeunert

Begriffe wie Annotationen, Relationen, Ontologien und Inferenz begegnen uns in allen Projekten, die sich mit der (geistes)wissenschaftlichen Erschließung von Korpora beschäftigen. Hierbei werden die Korpora mit den entsprechenden Annotationen des Anwendungsgebiets versehen, um auf dieser Grundlage Forschungsfragen zu beantworten. Annotationen sind ein wichtiges Mittel in der Analyse von Korpora; allerdings entwickeln die meisten Projekte ihre eigenen Strukturen und Formen der Annotations-Abbildung und -Verwaltung. Am Anfang der *Digital Humanities* wurden Annotations-Schemata teilweise fest codiert. Nunmehr werden vermehrt Beschreibungssprachen wie RDF Schema (RDFS) und die Web Ontologie Language (OWL)[1] eingesetzt. Da uns Ontologien flexible Annotationsmöglichkeiten bieten, jedoch die permanente Wartung und Anpassung von Software durch Informatiker auf lange Sicht keine effiziente Lösung ist, wurde in unserem interdisziplinären Projekt, gefördert durch LOEWE[2], zur inhaltlichen Erschließung der *Illustrationen zu Goethes Faust* der *OWLnotator*, ein ontologiebasiertes Annotationswerkzeug zur Erstellung und Analyse von Intra- und Intermedialen Relationen entwickelt. Dank der flexiblen Annotationsmöglichkeiten des *OWLnotators* können Geisteswissenschaftler durch das Erstellen von eigenen Ontologien sehr schnell und einfach ontologiegestützt Annotationen im Einzel- oder Batch-Betrieb erstellen, ändern oder löschen.

Ein digitalisiertes Korpus von 2 500 Faustillustrationen bildet die Grundlage für die semantische Erschließung durch eine kunsthistorische Ontologie. Auf dieser Basis demonstrieren wir im Full-Paper die inter- und intramedialen Relationen zwischen dem Faust-Text und den dazugehörigen Bildern im *OWLnotator*. Das Korpus der *Illustrationen zu Goethes Faust* ist hierbei für diese Untersuchung in besonderer Weise geeignet, da einige Illustrationen Bildinhalte haben, welche im Text nicht erwähnt oder beschrieben wurden. Für eine hinreichend aussagekräftige inhaltliche Erschließung der Bildbestände ist es notwenig, die Bilder detailliert zu beschreiben. Dafür werden die Bilder *segmentiert* (cf. Abrami, Freiberg und Warner 2012) und detailliert annotiert. Zur Korpusverwaltung wird die ImageDB, ein Tool des *eHumanities Desktop* (Gleim, Mehler und Ernst 2012), verwendet, welche die Bildsegmentierung durchführt und als Annotation mittels des *OWLnotators* speichert. Der *eHumanities Desktop* ist eine plattformunabhängige, browserbasierte, flexible und skalierbare virtuelle Forschungsumgebung für Geisteswissenschaftler welche neben den genannten Tools weitere Werkzeuge zur Verwaltung, Analyse und Aufbereitung von Text- und Bild-Korpora wie auch von Lexika umfasst.

An den Korpora und den Annotationen können mehrere Forscher, in einer Arbeitsgruppe oder darüber hinaus, gleichzeitig arbeiten und je nach Forschungsschwerpunkt und Fragestellungen die annotierten Elemente entsprechend der gewünschten ontologischen Betrachtungsweise auswerten. Hierzu müssen die Forscher nur eine eigene Ontologie erstellen und diese im *OWLnotator* anwenden. Der *OWLnotator* kann jede syntaktisch gültige Ontologie nutzen und dank der in der Web Ontology Languge spezifizierten Möglichkeiten der Klassen- und Relationsvererbung sowie der darin enthaltenen Inferenz-Methoden zur inhaltlichen Analyse des Korpus eingesetzt werden. Dank dieses ausdrucksmächtigen Werkzeugs sind wir künftig in der Lage, semantische Wissensnetzwerke sehr schnell und einfach aufzubauen und diese mit der Forschungsgemeinschaft zu teilen.

Unser Ziel ist es, durch den Einsatz von ontologischen Annotationen auf der Grundlage von OWL ein austauschbares Wissensnetzwerk zu generieren, welches unabhängig von der Software eingesetzt werden kann. Voraussetzung ist natürlich, dass die Software Ontologien lesen, interpretieren und verwalten kann, wie dies

[1] `http://www.w3.org/TR/owl-features/`

[2] Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, `www.proloewe.de`

durch den *OWLnotator* dynamisch und effizient geschieht. Die kunsthistorischen Ontologien unseres Projektes beinhalten die Annotationen der abgebildeten Personen auf den *Illustrationen zu Goethes Faust* sowie die Annotation ihrer Proxemik und Gesten. Durch atomare[3] Annotationen schaffen wir die Grundlage zur Interpretation der annotierten Bildinhalte. Folgeprojekte oder ähnliche Fragestellungen können an den im Projekt erstellten Ontologien anknüpfen und darauf aufbauen sowie die Ontologien selbstverständlich erweitern. Die Ergebnisse sowie die Verknüpfung zwischen den Bildinhalten mit dem dazugehörigen Text sowie die weiter- und tiefergehende ontologische Annotation auf Text- und Bild-Ebene bieten für die Forschung, für die Lehre sowie für die Präsentation von Beständen und Korpora eine breite und stabile Grundlage und sind gleichzeitig jederzeit austauschbar und weiterverwendbar.

Durch unsere Arbeiten möchten wir Geisteswissenschaftlern die *Scheu* vor dem Einsatz digitaler Werkzeuge auch im Bereich komplexester Ontologien nehmen. Es geht darum, sehr große Korpora überhaupt erst auf der Basis dynamisch, im Wissenschaftsprozess wachsender Ontologien erschließbar zu machen. Mit dem *OWLnotator* bieten wir universell einsatzfähiges Annotationswerkzeug an welches Open Source zur Verfügung steht und allen Nutzern die Möglichkeit gibt, Annotationen ontologiebezogen durchzuführen, ohne über das *Wie* der Umsetzung nachdenken zu müssen. Insbesondere sollen Geisteswissenschaftler die Gelegenheit erhalten, nun über das *Was* des Annotationsinhaltes nachzudenken – dazu befähigt sie der *OWLnotator* in den Anwendungsgebieten der Geisteswissenschaft.

Literatur

Abrami, Giuseppe, Michael Freiberg und Paul Warner (2012). „Managing and Annotating Historical Multimodal Corpora with the eHumanities Desktop - An outline of the current state of the LOEWE project Illustrations of Goethe's Faust ". In: *Proceedings of the Historical Corpora Conference, 6-9 December 2012, Frankfurt*.

Gleim, Rüdiger, Alexander Mehler und Alexandra Ernst (2012). „SOA implementation of the eHumanities Desktop". In: *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany*.

[3] Nicht mehr weiter teilbare

Sabine Seifert
Humboldt-Universität zu Berlin
Institut für deutsche Literatur
Nachwuchsgruppe „Berliner Intellektuelle 1800–1830“
sabine.seifert@hu-berlin.de

Poster Abstract

Gelehrsamkeit will verlinkt werden. Zur digitalen Erschließung von August Boeckhs Nachlass und Bibliothek

Das Wirken August Boeckhs[1] (1785–1867), einer zentralen Figur im wissenschaftlichen Preußen des 19. Jahrhunderts, ist einer der Forschungsschwerpunkte der Nachwuchsgruppe „Berliner Intellektuelle 1800–1830“, geleitet von Dr. Anne Baillot an der Humboldt-Universität zu Berlin. Boeckhs bisher unedierte Handschriften bilden die Grundlage eines dreiteiligen Erschließungs- und Datenaufbereitungsvorhabens und werden auf folgende Weise zugänglich gemacht: 1) mittels einer digitalen Auswahledition, 2) durch die Rekonstruktion von Boeckhs Büchersammlung und 3) durch die Erschließung des handschriftlichen Nachlasses. Diese drei Bereiche sind eigenständige Forschungsunternehmen mit jeweils eigenen Ansprüchen und Forschungsfragen, die parallel ablaufen, sich aber durch die Verbindung auf digitaler Ebene gegenseitig befruchten und ergänzen.

Zu 1) Den Ausgangspunkt für das Poster soll die Edition ausgewählter Briefe und Berichte von und an Boeckh bilden, die im Rahmen der digitalen Edition „Briefe und Texte aus dem intellektuellen Berlin um 1800“[2] erfolgt. Diese führt nicht nur verschiedene Schriftsteller/-innen, Wissenschaftler und Intellektuelle zusammen, sondern auch verschiedene Textsorten und Themenschwerpunkte. Schon dadurch werden die edierten Handschriften Boeckhs in einem größeren, sie selbst übersteigenden Kontext verortet. Dieser Breite im Ansatz muss die zugrunde liegende Datenstruktur gerecht werden. Es wurden projekteigene Kodierungsrichtlinien[3] nach den TEI P5 Guidelines[4] entwickelt, die nur in Ausnahmefällen wie bei briefspezifischen Metadaten abgewandelt werden. Die Handschriften werden als Digitalisate zur Verfügung gestellt und die Transkriptionen in einer diplomatischen Umschrift sowie einer Lesefassung angeboten – ein Aspekt, der aufgrund der technischen Möglichkeiten relativ leicht zu realisieren ist, aber von kaum einer digitalen Edition tatsächlich umgesetzt wird. Die Auszeichnung von Personen, Orten, Werken und Organisationen und deren Erfassung in projektinternen Indizes ermöglichen eine umfassende Verknüpfung. Beides bildet die Grundlage für die Rekonstruktion (und geplante Visualisierung) der Netzwerke der Berliner Intellektuellen und für die Beantwortung der Forschungsfrage, wie sich diese Netzwerke entwickelt haben. Die Verwendung von Stan-

[1] Für neueste Forschungen siehe u.a.: Christiane Hackel, Sabine Seifert (Hrsg.), *August Boeckh. Philologie, Hermeneutik und Wissenschaftsorganisation*, Berlin 2013; Werther, Romy (Hrsg.), *Alexander von Humboldt. August Böckh. Briefwechsel*. Unter Mitarb. v. Eberhard Knobloch, Berlin 2011; Poiss, Thomas, „August Boeckh als Universitätspolitiker“, in: Anne Baillot (Hrsg.), *Netzwerke des Wissens*, Berlin 2011. S. 85–112.
[2] http://tei.ibi.hu-berlin.de/berliner-intellektuelle/?de
[3] http://tei.ibi.hu-berlin.de/berliner-intellektuelle/encoding-guidelines.pdf
[4] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

dards für webbasierte Textpräsentationen (XML/TEI, CC-BY-Lizenz, Normdaten, persistente URLs, ISO-Codes, etc.) bieten die Möglichkeit zu Kollaborationen mit anderen Projekten.

Zu 2) Die in den veröffentlichten Handschriften erwähnten Personen und Schriften finden sich häufig als Autoren und Werke in Boeckhs Büchersammlung wieder, die ca. 6000 Bände umfasste. Diese wird anhand einer von Boeckh selbst angefertigten und ebenfalls edierten Bücherliste virtuell rekonstruiert. Mit diesen Ergebnissen wird es möglich, Boeckhs Wissenshorizont und die Verbindungen zwischen Boeckhs eigenem wissenschaftlichen Arbeiten und dem seiner Fachkollegen und anderer zeitgenössischer Geisteswissenschaftler, ehemaliger Schüler und wissenschaftlicher Institutionen nachzuzeichnen. Darüber hinaus werden Boeckhs eigene Exemplare, aufbewahrt in der Universitätsbibliothek der Humboldt-Universität zu Berlin, auf Marginalien überprüft, die dann gegebenenfalls als Digitalisat zur Verfügung gestellt werden können.

Zu 3) Der bisher nicht an einer zentralen Stelle recherchierbare handschriftliche Nachlass Boeckhs soll, beginnend bei den verschiedenen Berliner Archiven und Bibliotheken, möglichst vollständig erschlossen werden. Für eine systematische und erstmalig institutionenübergreifende Darstellung dieser Daten wurde eine Plattform[5] eingerichtet, die in ihrer technischen Struktur mit der Edition in Verbindung steht. Durch die Kooperation mit der Staatsbibliothek zu Berlin–PK konnten ca. 900 Kalliope-Einträge importiert werden.[6] Da diese bibliothekarischen Einträge gerade in Bezug auf Boeckhs Korrespondenz teilweise eine sehr grobe Struktur aufweisen, wurden sie nach Autopsie und aufgrund von Forschungsergebnissen mit detaillierteren Metadaten zu jedem einzelnen Brief angereichert. Zusätzlich wurden, in Übereinstimmung mit der Edition, die in den Briefen genannten Personen, Werke etc. verzeichnet, um auch einen entitätenbezogenen Zugriff zu ermöglichen.

Nur durch die digitale Erfassung und Präsentation wird es möglich, dass sich diese drei Forschungsunternehmen in ihren wissenschaftlichen Ansätzen und Ergebnissen gegenseitig ergänzen und einen übersichtlichen und systematischen Zugriff für die Forschung bieten können. Durch die Indizes ist eine gleichzeitige Suche in den Handschriften der Edition, der Bücherliste und der mit ihr verbundenen bibliographischen Angaben und den Metadaten der Nachlassdokumente möglich. So können verschiedene Diskurse verfolgt und Tendenzen in der Philologie, in der Wissenschaftsorganisation in und über Preußen hinaus sichtbar gemacht werden. Die digitale Umgebung allgemein und konkret die Recherche und Nutzung derselben Datenbasis für gleichzeitig drei Unternehmen, die denselben Bezugspunkt – die Person Boeckh – haben, aber doch in sich eigenständig sind, können somit zur Forschung beitragen und Forschungsdaten zur Verfügung stellen, die von wissenschaftlichen Institutionen, Bibliotheken und Archiven genutzt werden können. Mit dem Poster möchte ich die digitale Edition in Verbindung mit der Rekonstruktion der Bibliothek und der Nachlasserschließung vorstellen und zeigen, welche Möglichkeiten digitale Methoden für die Forschung bieten und wie (Meta-)Daten als Schnittstelle zwischen Literaturwissenschaft, Bibliotheken und Archiven fungieren können.

[5] http://tei.ibi.hu-berlin.de/boeckh/
[6] http://kalliope.staatsbibliothek-berlin.de/

DFG-Projekt „Entwicklung eines MEI- und TEI-basierten Modells kontextueller Tiefenerschließung von Musikalienbeständen am Beispiel des Detmolder Hoftheaters im 19. Jahrhundert (1825–1875)“

Dr. Irmlind Capelle | Kristina Richts M.A., MA LIS

Die Kooperation von Bibliotheken und Wissenschaft erhält gegenwärtig vor dem Hintergrund des digitalen Wandels und der Entwicklung virtueller Forschungsumgebungen eine immer stärkere Bedeutung. Als Grundlage für eine solche Kooperation und die Zusammenführung der in unterschiedlichen Formaten vorliegenden Datenbestände ist die Entwicklung geeigneter Datenstandards unverzichtbar. Für den Bereich der Musikwissenschaft bringt der relativ junge Standard der Music Encoding Initiative (MEI) die für eine solche Zusammenführung notwendigen Anforderungen mit. Durch die Implementierung des Modells der Functional Requirements for Bibliographic Records (FRBR) im Jahr 2013 haben die Entwickler des Formats bereits einen entscheidenden Schritt in Richtung einer Zusammenführung mit in Bibliotheken vorliegenden Daten vollzogen. Das FRBR-Modell bildet dabei zum einen die Grundlage für das neue Katalogisierungsformat Resource Description and Access (RDA), zum anderen eignet es sich aber auch in besonderem Maße für die Beschreibung und Speicherung musikwissenschaftlicher Quellen. So sind erste Werkverzeichnisse dabei, die Vorteile des Modells zu nutzen – ein prominentes Beispiel ist der jüngst vom Danish Centre for Music Publication der Königlichen Bibliothek in Kopenhagen in digitaler Form veröffentlichte Carl Nielsen Works Catalogue (CNW).
Im Rahmen des hier vorzustellenden, von der Deutschen Forschungsgemeinschaft (DFG) ab September 2014 über den Zeitraum von zunächst zwei Jahren geförderten Projekts steht die Entwicklung eines Modells zur kontextuellen Tiefenerschließung von Musikalienbeständen im Fokus. Vor dem Hintergrund der engen Kooperation von Bibliothek und Wissenschaft, die in Detmold ab Mitte 2015 auch räumlich und institutionell durch die Entstehung des neuen Zentrums „Wissenschaft | Bibliothek | Musik“ umgesetzt wird, beleuchtet das Projekt den Gegenstand sowohl von der wissenschaftlichen als auch von der bibliothekarischen Seite. So besteht das technische Ziel des Projekts darin, ein von anderen Bibliotheken mit vergleichbaren Beständen nachnutzbares Modell auf der Grundlage der XML-basierten Codierungsstandards der Music Encoding Initiative (MEI) sowie der Text Encoding Initiative (TEI) zu entwickeln, die beide sowohl eine bibliothekarische als auch eine wissenschaftliche Erfassung der Dokumente unterstützen und durch ihre Anbindung an internationale Datenstandards die Möglichkeit eines gezielten Mappings zu den Datenbeständen anderer Bibliotheken oder Forschungseinrichtungen mit sich bringen. Die bereits vorhandenen Vorteile speziell von MEI werden dabei gezielt im Hinblick auf ihre Anbindung an bibliothekarische Datenbestände weiterentwickelt und eine Anwendung erprobt.
Auf inhaltlicher Ebene sollen auf der Basis des entwickelten Modells die Vorteile einer kontextuellen Erschließung anhand des außergewöhnlich reichhaltig dokumentierten Musikalien- und Aktenbestands aus der Blütezeit des Detmolder Hoftheaters von 1825 bis 1875 demonstriert werden. Diese in der Lippischen Landesbibliothek Detmold erhaltenen musikalischen und archivalischen Quellen sind bislang entweder nur standardmäßig z. B. im Internationalen Quellenlexikon der Musik (RISM) (Musikalien) erfasst oder sogar lediglich durch

maschinenschriftliche Regesten (Theaterakten) sowie z. T. handschriftliche Zettelkataloge ausgewertet. Ergänzt werden sie durch Materialien aus dem Landesarchiv Detmold (Personalakten etc.) und dem Staatsarchiv Osnabrück (Theaterzettel).

In der ersten Projektphase geht es darum, die überlieferten musikalischen Quellen, die sowohl Partituren, Stimmen und Partien als auch Libretti und Rollenhefte umfassen, einerseits detailliert zu beschreiben (inkl. enthaltener Einlagen bzw. Striche sowie handschriftlicher Einträge zu Personen und Aufführungen) und andererseits die archivalischen Quellen im Volltext oder als Regesten zu erfassen. Im Rahmen der kontextuellen Tiefenerschließung sollen dann den erschlossenen Musikalien z. B. Informationen aus den Einnahme-Journalen oder den Regiebüchern des Theaters zugeordnet werden. So könnte in der Folge etwa ein mit Normdaten angereichertes Rollenverzeichnis aller mitwirkenden Schauspieler oder Sänger erstellt werden.

Um die Erkenntnisse, die aus dieser Erschließung der Daten entstehen, anschaulich zu visualisieren und möglichst offene Schnittstellen zur Weiternutzung zu bieten, werden die Projektergebnisse in einem Portal zusammengeführt, in dem Digitalisate der Materialien (in Auswahl) mit den XML-basierten Erschließungsdokumenten unter Rückgriff auf die in Detmold entwickelte Software Edirom Online verknüpft werden. Damit wird nicht nur für Forscher oder interessierte Laien eine Möglichkeit geschaffen, sich ein sehr viel präziseres Bild vom Wirken des Detmolder Hoftheaters in all seinen Facetten zu machen, sondern ein Repositorium geboten, das vielfältige Anknüpfungspunkte für weitere kulturwissenschaftliche Fragestellungen im Umkreis dieser wichtigen Institution des Hofes bietet.

ediarum – Eine digitale Arbeitsumgebung für Editionsvorhaben

Stefan Dumont (dumont@bbaw.de), Martin Fechner (fechner@bbaw.de)

An der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) sind zahlreiche geisteswissenschaftliche Forschungsvorhaben unterschiedlichster Fachrichtungen angesiedelt. Die TELOTA-Arbeitsgruppe (»The Electronic Life of the Academy«) unterstützt diese Vorhaben in allen digitalen Belangen und entwickelt Softwarelösungen für die tägliche Forschungsarbeit der Wissenschaftler/-innen.

Die Erfahrung hat gezeigt, dass die Bereitschaft, TEI-Kodierung in Editionsvorhaben zu verwenden, von der Benutzerfreundlichkeit der Eingabeoberfläche abhängt. Aus der Perspektive der Wissenschaftler/-innen erscheint es als ein Rückschritt, direkt im XML-Code zu arbeiten, wenn man vorher in Programmen wie MS Word gearbeitet hat. Eine neue Softwarelösung muss daher mindestens den gleichen Komfort bieten wie das zuvor benutzte Programm. Idealerweise würde sie sogar den gesamten Lebenszyklus einer Edition abdecken: von der ersten Phase der Transkription bis hin zur Publikation in Web und Druck.

TELOTA hat mit »ediarum« eine solche digitale Arbeitsumgebung entwickelt. Diese Lösung besteht aus mehreren Softwarekomponenten, die es den Wissenschaftler(inne)n erlauben, Transkriptionen von Manuskripten in TEI-XML anzufertigen, zu bearbeiten und zu veröffentlichen.

Als zentrale Softwarekomponente der neuen Arbeitsumgebung wird »oXygen XML Author« eingesetzt. Die Bearbeiter arbeiten in oXygen XML Author nicht in einer Codeansicht, sondern in der benutzerfreundlichen »Autorenansicht«, die über Cascading Stylesheets (CSS) gestaltet wird. Außerdem kann der Endanwender über eine eigene Werkzeugleiste per Knopfdruck Auszeichnungen vornehmen. So können z.B. in Manuskripten Streichungen markiert oder Sachanmerkungen eingegeben werden. Auch können Textstellen ausgezeichnet und gleichzeitig über eine komfortable Auswahlliste mit dem jeweiligen Eintrag eines zentralen Registers (Personen-, Ortsregister etc.) verknüpft werden. Der gesamte Text kann dadurch einfach und schnell mit TEI-konformen XML ausgezeichnet werden.

Die digitale Arbeitsumgebung nutzt die native XML-Datenbank »exist-db« als zentrales Repositorium für die XML-Dokumente. Die Datenbank ist auf einem Server installiert und online zugänglich. Dadurch können alle Projektmitarbeiter auf ein und denselben Datenbestand zugreifen und zusammenarbeiten.

Neben der eigentlichen Arbeitsumgebung in oXygen XML Author, wird für die Forschungsvorhaben auch jeweils eine Website auf Basis von eXist, XQuery und XSLT erstellt. In ihr kann von den Wissenschaftler(inne)n der aktuelle Datenbestand leicht durchblättert bzw. durchsucht werden. Die Website kann - je nach Bedarf - nur den Bearbeitern oder der gesamten Öffentlichkeit gemacht werden.

Als weitere Ausgabemöglichkeit wird mit Hilfe von ConTeXt eine Druckausgabe implementiert, die automatisch aus den aktuellen TEI-XML-Dokumenten ein PDF erstellt. Die Gestaltung und Formatierung kann - nach entsprechender Konfiguration - dabei gedruckten Bänden der jeweiligen Edition entsprechen. Jedem TEI-Element wird über eine Konfigurationsdatei eine entsprechende Formatierungsanweisung für den Druck übergeben. So können z.B. Text- und Sachapparat als Fußnoten dargestellt werden, die mit Hilfe von

Zeilennummerierung und Lemmata auf den Fließtext verweisen. Die Druckausgabe erstellt bei Bedarf auch das passende Register zu den jeweiligen Transkriptionen und löst Querverweise zwischen Texten auf.

Die Arbeitsumgebung wird seit 2012 von Wissenschaftler(inne)n verschiedener Forschungsvorhaben bei ihrer täglichen Arbeit benutzt. Nach ihrer Meinung befragt, waren sich die Nutzer darin einig, dass durch die neue Arbeitsumgebung die Editionsarbeit erleichtert und viel Zeit gespart wird. Auch die Möglichkeit, die Ergebnisse der Arbeit direkt in einer Webpräsentation oder Druckausgabe zu kontrollieren, wurde positiv gesehen. Sehr erleichtert äußerten sich die Mitarbeiter/-innen darüber, dass ihnen keine Arbeit im XML-Code selbst zugemutet wird, sondern alle Texte in einer grafischen und einfach zu bedienenden Programmoberfläche mit XML ausgezeichnet werden können.

Nach der erfolgreichen Pilotumsetzung im Akademievorhaben »Friedrich Schleiermacher in Berlin 1808-1834. Briefe, Vorlesungen, Tageskalender« wurde »ediarum« in zwei weiteren Akademienvorhaben eingesetzt: »Commentaria in Aristotelem Graeca et Byzantina« und »Regesta Imperii - Friedrich III.« (letzteres in Kooperation mit der AdW Mainz) Für jedes Projekt wurden die TEI-XML-Schemata sowie die Funktionen an die verschiedenen Manuskripttypen und Forschungsanforderungen angepasst. Derzeit wird »ediarum« für die Historisch-kritische Edition der Schriften Jeremias Gotthelf zur Verfügung gestellt (in Kooperation mit der Universität Bern). Weitere Implementierungen befinden sich derzeit in Planung..

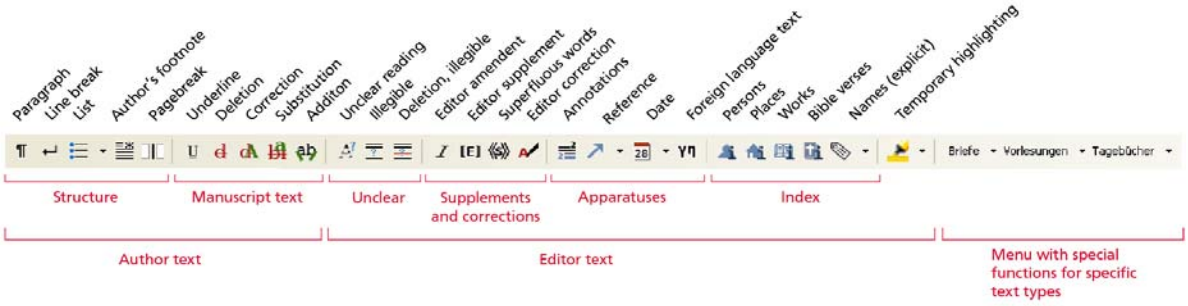
Weitere Informationen

- Projektwebsite: http://www.bbaw.de/telota/software/ediarum

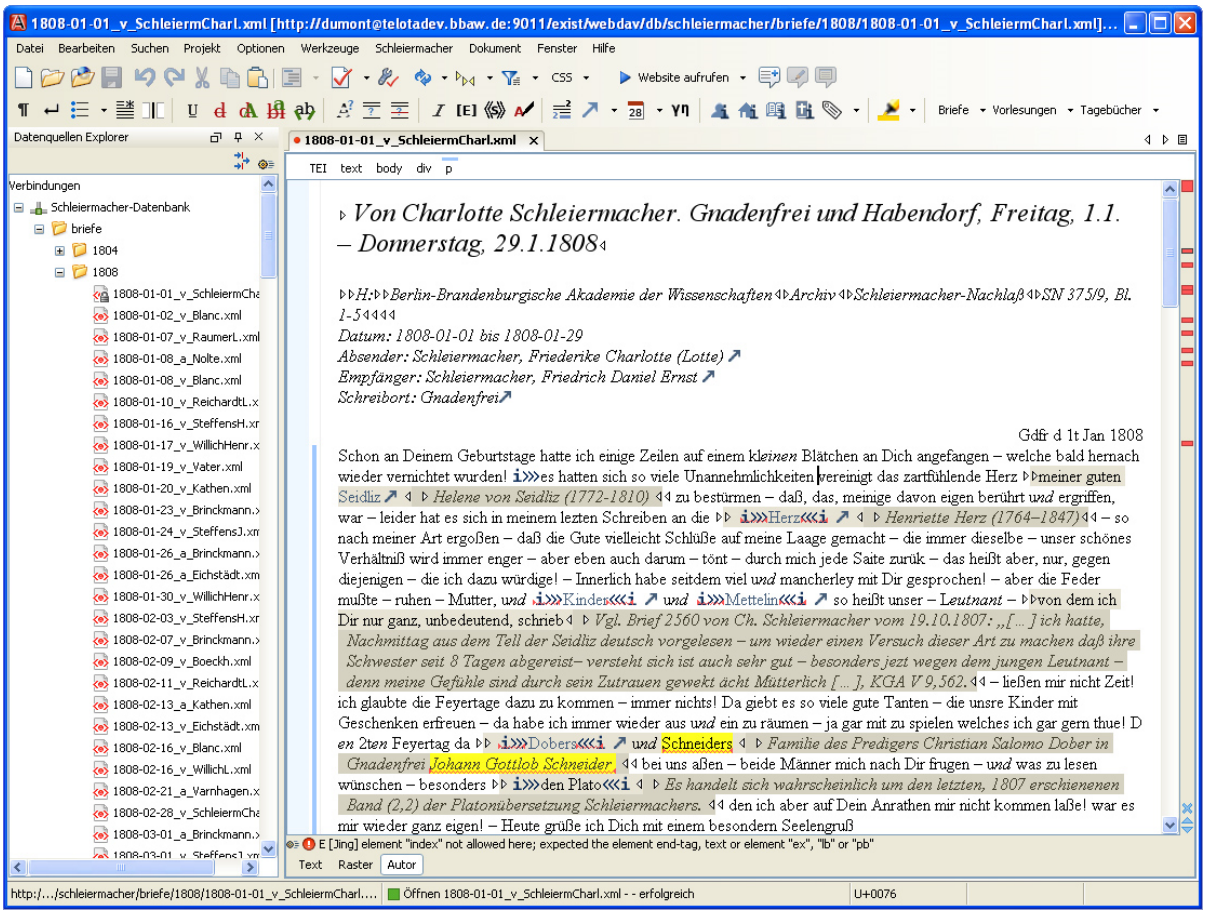
Literatur

- Dumont, Stefan; Fechner, Martin: Digitale Arbeitsumgebung für das Editionsvorhaben »Schleiermacher in Berlin 1808—1834« In: digiversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften. URL: <http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834/>
- Burnard, Lou; Bauman, Syd (Hg.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Charlottesville, Virginia, USA 2014. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>
- User Manual oXygen XML Author 14. URL: <http://www.oxygenxml.com/doc/ug-editor/>
- eXist Main Documentation. URL: <http://www.exist-db.org/exist/documentation.xml>
- ConTeXt Dokumentation. URL: <http://wiki.contextgarden.net/Main_Page>

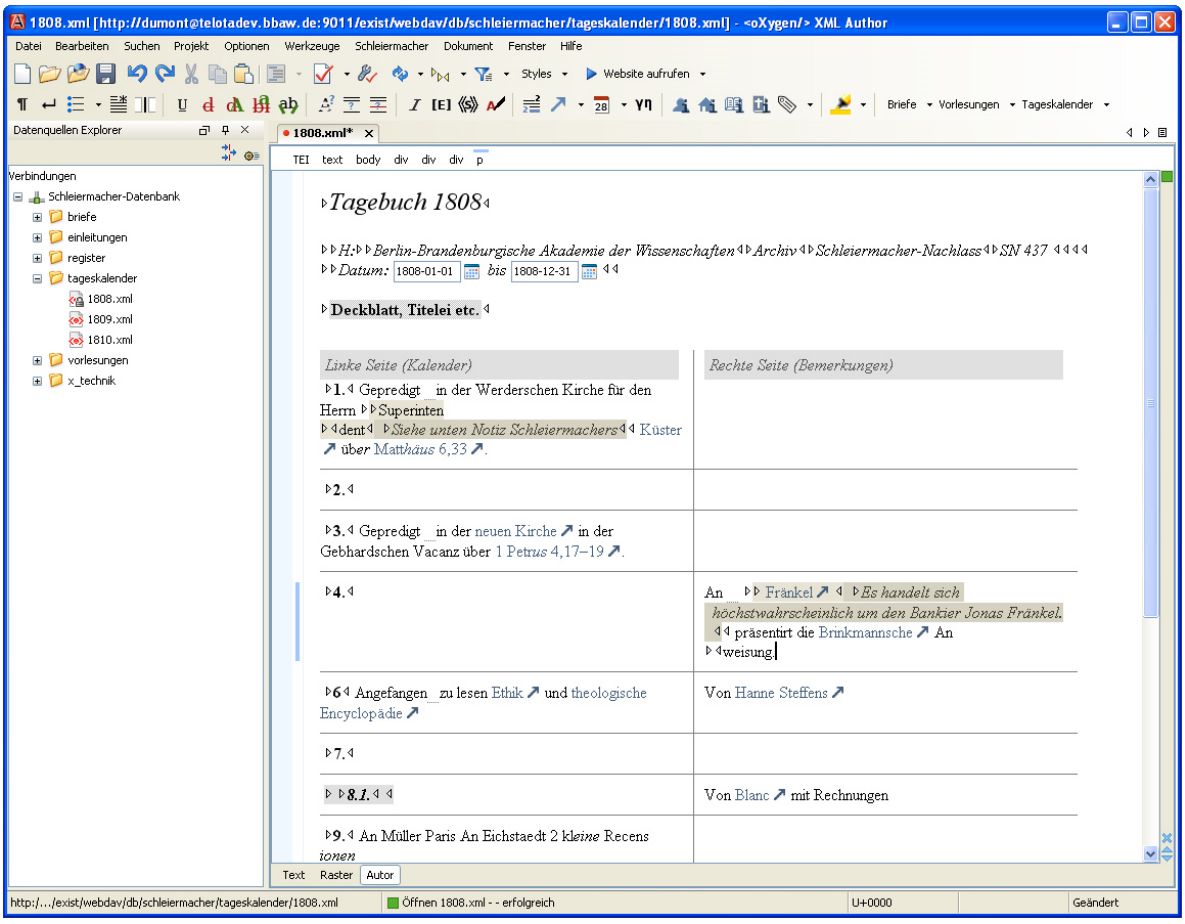
Screenshots



Eigene Werkzeugleiste für die Schleiermacher-Edition in oXygen XML Author



Transkription eines Briefes in oXygen XML Author



Transkription eines Tageskalenders in oXygen XML Author

Kombinierte Text- und Geo-Suche zum Durchsuchen einer Georeferenzierten Online-Bibliographie

Bastian Entrup[1], Vera Ermakova[2], Ines Schiller[2] und Henning Lobin[2]

[1]Angewandte Sprachwissenschaft und Computerlinguistik
bastian.entrup@germanistik.uni-giessen.de
[2]Zentrum für Medien und Interaktivität
{vera.ermakova|ines.schiller|henning.lobin}@zmi.uni-giessen.de
Justus-Liebig Universität Gießen
Germany

1 Einleitung und Motivation

Das GeoBib Projekt entwickelt eine georeferenzierte Online-Bibliographie der frühen Holocaust- und Lagerliteratur zwischen 1933 und 1949 mit über 700 Werken und ca. 850 Autoren und Herausgebern. Anders als eine klassische Bibliographie werden auch handlungsrelevante Orte sowie biographische Daten inklusive einer schriftlichen Biographie zu Autoren und Herausgebern erfasst. Die bibliographischen Daten umfassen zusätzlich Informationen wie sie für ein Literaturlexikon nicht unüblich sind, z.B. Rezeptionen und Werksgeschichten. Die Kombination und Verlinkung dieser Entitäten, Personen, Werke und Orte, macht das Besondere und den Mehrwert der Online-Bibliographie aus.

2 Funktionen und Implementation

2.1 Funktionen und Implementation der Text-Suche

Um die entstehende Bibliographie und die dafür erstellten Texte (z.B. die Inhaltszusammenfassungen oder die Autorenbiographien) sowie die bibliographischen Daten (Autoren, Herausgeber, Verlag usw.) durchsuchbar zu machen wird Apache Solr [1] und das Open Source Projekt *glp4lucene*[2] zur Verarbeitung von Suchanfragen und Erstellung des Indexes verwendet.

Die natürlich Variabilität und Ambiguität einer Sprache machen Verarbeitungsschritte aus dem Bereich des Natural Language Processings (NLP) notwendig. Im Bereich des Information Retrievals (IR) hat sich das Stemming als einfaches, regelbasiertes Verfahren zur Vereinheitlichung unterschiedlicher Wortformen auf einen gemeinsamen Stamm durchgesetzt (vgl. [3,5]). Aus linguistischer

[1] https://lucene.apache.org/solr/.

[2] Zu finden unter `https://sourceforge.net/projects/glpforlucene/`. Das Paket umfasst die Ergänzung von Synonymen, eine Lemmatisierungsfunktion sowie eine Termgewichtungsmethode, die auf der Wortart der Terme basiert.

Sicht ist Stemming jedoch nicht so erstrebenswert wie eine Lemmatisierung, die Reduktion von verschiedenen Wortformen auf ein gemeinsames Lemma, da beim Stemming die Ambiguität einer Sprache erhöht wird. Das hier genutzte Verfahren basiert auf dem MATE Tool [2] und verwendet das in [9] beschriebene deutsche Modell.

Basierend auf einem Lemma können Synonyme in GermaNet [4] nachgeschlagen und dem Suchindex hinzugefügt werden. Das Hinzufügen der Synonyme geschieht schon während der Indexierung der Daten[3].

Die einfache textbasierte Suche durchsucht die wahrscheinlichsten Suchfelder nach einem Suchbegriff und nutzt dabei die Lemmatisierung des Indexes, um deklinierte oder konjugierte Formen zu finden. Zusätzlich sind im Index Synonyme vorhanden, so dass eine Suche nach *Gefängnis* auch Vorkommnisse von z.B. *Zuchthaus* findet. Die Suchergebnisse sind nach verschiedenen Personen-, Text- und Ortskategorien facettiert (s. Abb. 1).

Die Erweiterte-Suche liefert entweder Texte, Autoren/Herausgeber oder Orte als Ergebnis zurück. Wenn nach Texten gesucht wird, kann die Suche nach biographischen Daten der Autoren/Herausgeber (z.B. Name, Geburtsjahr oder Sterbeort), aber auch nach bibliographischen Daten (z.B. nach dem Verlag, dem Erscheinungsjahr oder -ort) gefiltert werden. Eine mögliche Suchanfrage wäre z.B. *Texte, deren Autoren weiblich sind.* Ähnliche Einschränkungen sind auch für Personensuchen möglich; z.B.: *Eine Autorin, die im Jahr 1939 einen oder mehrere Texte bei einem bestimmten Verlag veröffentlicht hat.*

2.2 Funktionen der Geo-Suche

Die Geo-Suche basiert auf einem Kartensatz Europas zur Zeit zwischen 1939 und 1945, der speziell für dieses Projekt aus verschiedenen Datensätzen kompiliert wurde (vgl. [6,7]). Für jedes Jahr wurde versucht, eine vollständige Karte mit den Grenzen Europas zu erstellen [8]. Die Jahre können über einen Slider unter der Karte ausgewählt werden. Auf der Karte dargestellte Datensätze können durch Klicken, Zoomen oder andere Werkzeuge ausgewählt werden.

Orte können in ein Suchfeld eingeben werden. Auf Grund der hohen Ambiguität von Toponymen wird dem User bei der Eingabe eines Ortsnamens eine Liste mit Vorschlägen angezeigt. Auf diese Weise gefundene Orte können dann mit einer Umkreissuche erweitert werden. Das ermöglicht zielgenaue regionale Recherchen, die für Heimatforscher und pädagogische Zwecke sinnvoll sind.

Ein Graph unterhalb der Karte (s. Abb 2) zeigt die Häufigkeit von Ereignissen (z.B. Anzahl von Handlungsorten) für jedes Jahr an. So können auf einen Blick Schwerpunkte ausgemacht werden. Ein Slider unter diesem Graph macht

[3] Das ist bei dem relativ kleinen Datensatz vertretbar. Im Vergleich zu einer Verarbeitung während des Suchvorgangs, sorgt dies für eine geringere Auslastung und weniger Wartungsbedarf des resultierenden Systems. Das verwendete Software Paket erlaubt allerdings beide Möglichkeiten.

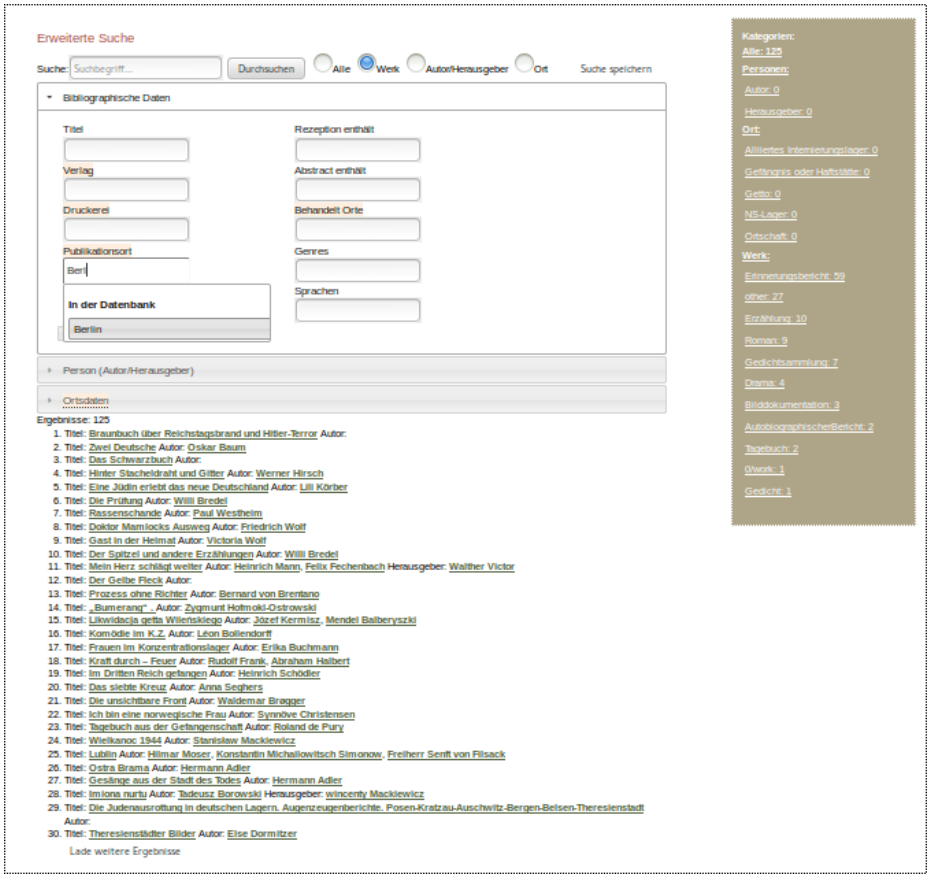


Abb. 1. Screenshot eines aktuellen Prototypen: Darstellung der Suchergebnisse und Vorschau der Eingabemaske.

es möglich sich nur Handlungsorte eines bestimmten Zeitraumes anzeigen zu lassen.

2.3 Verbindung von Text- und Geo-Suche

Die Verbindung der verschiedenen Entitäten in der Datenbank macht die Kombination der beiden Systeme möglich. Autoren/Herausgeber sind mit ihren Geburts- und Sterbeorten verbunden, außerdem mit den Orten in den von ihnen geschriebenen Texten. Werke sind mit ihren Erscheinungs- und Handlungsorten verbunden.

Die Beispiel-Suchanfragen können wie folgt ergänzt werden: *Texte, deren Autoren weiblich sind und in Berlin geboren wurden* und *Eine Autorin, die im Jahr*

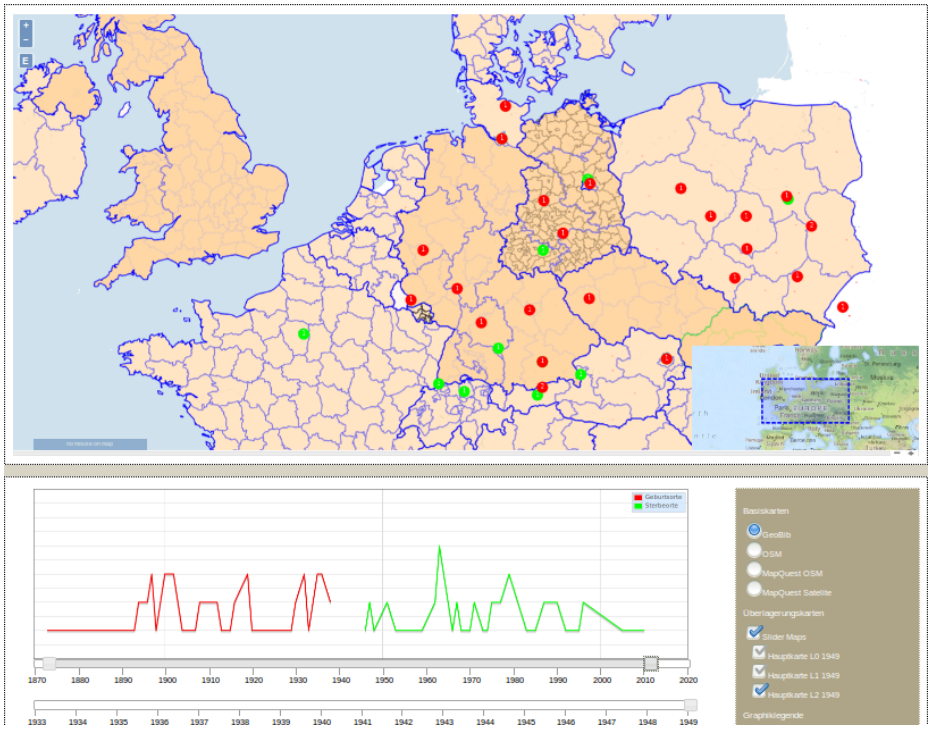


Abb. 2. Screenshot eines aktuellen Prototypen: Karte mit Angezeigten Geburts- (rot) und Sterbeorten (grün) basierend auf 125 Beispieltexte in den Grenzen von 1949.

1939 einen Text über Geschehnisse in Auschwitz bei einem bestimmten Verlag veröffentlicht hat. Auch Texte von Autoren, die in einer bestimmten Region geboren wurden oder Texte, die von einem bestimmten Lager handeln, sind so auffindbar.

Umgekehrt sind aber auch Orte auffindbar, die als Handlungsort zu bestimmten Zeiten eine Rolle spielen oder die Publikationsorte von bestimmten Werken sind. So lassen sich alle Orte finden, die in Werken eines bestimmten Autoren vorkommen.

3 Aussicht

Die Verbindung von Texten mit Geo-Daten ist nicht nur eine besondere Herausforderung an die Darstellung, die Organisation und die Suche nach Informationen, sondern bietet viele Möglichkeiten: Die Verteilung von Handlungsorten der Texte auf einer Karte bietet ein räumliches Verständnis eines Textes oder einer Sammlung von Texten. Besondere lokale Schwerpunkte können auf einen Blick erfasst werden.

Viele der im Projekt erfassten Texte gelten heute als vergessen. Sie werden nun das erste Mal systematisch durchsuchbar gemacht. Die Kombination von bibliographischen, biographischen, geographischen und inhaltlichen Daten ermöglicht einen völlig neuen (räumlichen) Zugang zu den Texten und den Ereignissen des Holocaust. So sind das Stellen und die Beantwortung neuer Forschungsfragen auf Grundlage einer breiten Textbasis und unter Berücksichtigung der geographischen Verteilung möglich.

Literatur

1. Binder, F., Entrup, B., Schiller, I., Lobin, H.: Uncertain about Uncertainty: Different Ways of Processing Fuzziness in Digital Humanities Data. In: Digital Humanities 2014, Book of Abstracts, pp. 97-100. Ecole Polytechnique Fédérale de Lausanne (EPFL) and The University of Lausanne (UNIL), Switzerland, 7-12 July 2014 (2014), http://dharchive.org/paper/DH2014/Paper-874.xml
2. Bohnet, B.: Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 89–97. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
3. Braschler, M., Ripplinger, B.: How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7(3-4), 291–316 (2004)
4. Hamp, B., Feldweg, H.: GermaNet - a Lexical-Semantic Net for German. In: Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. pp. 9–15 (1997)
5. Kraaij, W., Pohlmann, R.E.: Viewing Stemming as Recall Enhancement. In: In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 40–48 (1996)
6. Schaarschmidt, S.: Bestandserhebung zu verfügbaren digitalen geographischen Grundlagenkarten (2013), http://geb.uni-giessen.de/geb/volltexte/2014/10572
7. Schaarschmidt, S.: Bedarfsanalyse zu weiterem Kartenmaterial (2014), http://geb.uni-giessen.de/geb/volltexte/2014/11102
8. Schiller, I., Entrup, B., Binder, F., Schaarschmidt, S., Lobin, H.: Using a GIS for Search and Visualization of Literary Works in the Digital Humanities. In: gis.SCIENCE - Die Zeitschrift für Geoinformatik 4 (to appear) (2014)
9. Seeker, W., Kuhn, J.: Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: LREC. pp. 3132–3139 (2012)

Digitalisierung der Universitätssammlungen der FAU Erlangen-Nürnberg

Auf dem Weg zum Semantic Web in Eigenregie

Martin Scholz und Udo Andraschke

(vorname.nachname@fau.de)

Die Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) besitzt über 20 Sammlungen aus den verschiedensten Fachbereichen. [1] Nicht minder verschieden gestalten sich Grad und Umfang ihrer Erfassung und Digitalisierung. Mit der Einrichtung einer Zentralkustodie im Jahr 2011, die die Aktivitäten und Ausrichtung der Sammlungen bündeln und sie als wichtige wissenschaftliche Infrastrukturen weiter ausbauen sollte, wurde auch das Ziel formuliert, die digitale Datenerfassung und Präsentation der Sammlungen voranzutreiben. Von zentraler Bedeutung sind dabei gemeinsame Erfassungsstandards und -formate sowie eine gemeinsame Software-Lösung und Webpräsenz.

Die Wahl der geeigneten Software-Infrastruktur fiel bewusst auf die Virtuelle Forschungsumgebung WissKI (wiss-ki.eu) [4], da sie

a) unter einer Open Source-Lizenz verfügbar ist (GPL),
b) konsequent auf offenen Standards und Formaten aufbaut und
c) an der FAU mitentwickelt wird.

WissKI wird seit 2009 von der Arbeitsgruppe Digital Humanities des Departments für Informatik der FAU in Kooperation mit dem Germanischen Nationalmuseum in Nürnberg sowie dem Zoologischen Forschungsmuseum Alexander Koenig in Bonn als web-basiertes Content Management System für die Dokumentation von Kulturerbe im musealen und wissenschaftlichen Kontext entwickelt und befindet sich in den genannten Institutionen bereits im Einsatz. WissKI bietet den Nutzern gewohnte Eingabe- und Präsentationsschnittstellen, wie etwa feldbasierte Eingabemasken oder die Möglichkeit zum Freitext. Die Daten werden jedoch im Hintergrund nativ auf Basis von Semantic Web-Technologien (Ontologien, RDF [2]) erfasst. Dies ermöglicht auch technisch ungeschulten Nutzern das Einpflegen hoch vernetzter Datenbestände – sowohl lokal als auch global – und gleichzeitig das Erfassen der Bedeutung der Daten, um deren Interpretierbarkeit auf lange Zeit zu sichern. Dabei schreibt das System keine ontologischen Kategorien vor, sondern kann innerhalb des jeweiligen Anwendungsbereichs frei angepasst werden. WissKI ist nicht als zentraler Webdienst konzipiert, vielmehr kann die Software kostenlos heruntergeladen und auf einem Server als an die eigenen Bedürfnisse angepasste WissKI-Instanz eingesetzt werden.

Als fachübergreifende, verbindende Ontologie – eine sog. Referenzontologie – wurde der offene Standard CIDOC CRM [3] (ISO 21127) bzw. die OWL DL-Implementation Erlangen CRM (erlangen-crm.org) [5] gewählt, da das CIDOC CRM

a) speziell auf die Dokumentation von Kulturerbe ausgerichteter ist und
b) als international anerkannter Standard Sicherheit in Langzeitfragen gibt.

Die Referenzontologie garantiert zum einen ein Mindestmaß an fachübergreifender Interpretierbarkeit der Daten durch die Definition grundlegender Klassifikationsstrukturen und ermöglicht zum anderen die modulare Erweiterung um

fachspezifische Begrifflichkeiten.

Zur Umsetzung des angezeigten Vorhabens wurde das Pilotprojekt *WissKI@Sammlungen der FAU* [6] ins Leben gerufen. Partner sind neben der Zentralkustodie und der AG Digital Humanities drei ausgesuchte Universitätssammlungen: das Herbarium Erlangense, die Informatiksammlung Erlangen sowie die Schulgeschichtliche Sammlung. Die beteiligten Sammlungen spiegeln die oben genannte Heterogenität in hohem Maße wieder, so dass die unterschiedlichen Eigenarten und Bedürfnisse der Sammlungen der FAU weitgehend repräsentiert sind.
Das Pilotprojekt hat experimentellen Charakter. Es soll WissKI für den Einsatz in den Sammlungen erproben und einen Migrationspfad für die gesamten Universitätssammlungen entwickeln. Ein wichtiger Teilaspekt ist dabei der Transfer der Daten aus den bestehenden Datenbanksystemen in zuvor eingerichtete WissKI-Instanzen. Dennoch versteht sich das Vorhaben nicht als rein technikgetrieben, sondern sieht die Software als ein Instrument zum Ausbau der Sammlungen zu Forschungsinfrastrukturen.

Das Projekt startete im November 2013 ohne größere finanzielle Ausstattung. Treibende Kraft war und ist das Eigeninteresse der beteiligten Partner. Seit Mitte 2014 wird das Projekt durch eine studentische Hilfskraft unterstützt.
Aufgrund der räumlichen Nähe aller Projektbeteiligten haben sich Workshops in regelmäßigen Abständen als Arbeitsmodus bewährt. Von großer Bedeutung ist dabei der gegenseitige Austausch, sowohl zwischen den Sammlungen untereinander als auch zwischen den Sammlungen und der Informatik, repräsentiert durch die AG Digital Humanities. Parallel dazu wurde eine WissKI-Instanz als Sandbox zum Üben eingerichtet und mit einem Forum und Wiki ausgestattet, so dass bspw. auch Tutorien erstellt und gemeinsam bearbeitet werden können.
In der ersten Phase des Projekts (von Ende 2013 bis Mitte 2014) wurden monatliche Arbeitstreffen mit allen Projektteilnehmern anberaumt, um Themen zu behandeln, die alle Sammlungen angehen und um eine gemeinsame Wissensbasis zu schaffen. Nach ca. 6 Monaten wurden ergänzend Treffen zwischen der AG Digital Humanities (IT) und je einer Sammlung (Anwender) eingeführt, um den Spezifika der einzelnen Sammlungen besser Rechnung zu tragen.

Das Projekt wurde aufgrund der inhaltlichen Aufgaben in zwei Phasen unterteilt:
Die bereits abgeschlossene Phase 1 beinhaltete alle Maßnahmen bis zum Transfer der Daten nach WissKI, die sich in fünf Schritte gliedern lassen:

1. Vertrautmachen mit der (Semantic) Web-Technologie und der WissKI-Infrastruktur
2. Identifikation von Gemeinsamkeiten und Unterschieden der bereits vorhandenen Daten und Datenbankschemata
3. Erstellen bzw. Erweiterung der Domänenontologie auf Basis des CIDOC CRM
4. Definition der Eingabemasken und Datenfelder und entsprechende Konfiguration der WissKI-Software
5. Iteration der Punkte 2-4

Phase 2 widmet sich dem Datentransfer und dem Aufbau eines gemeinsamen Portals in folgenden Schritten:

6. Definition der Abbildungsvorschriften zwischen bestehenden Datenbanken und Domänenontologie
7. Import von Testdaten aus den bestehenden Datenbanken

8. Test und Korrektur des vorgenommenen Imports
9. Iteration der Punkte 6-8
10. Einbinden bzw. Erstellen von Normdaten
11. Einbinden des Datenbestandes in ein gemeinsames Präsentationsportal

Alle Schritte wurden und werden begleitend in Form von Tutorien dokumentiert. Sie spiegeln Erfahrungen, Diskussionen und Best Practices wieder und bilden einen wichtigen Eckpfeiler für die Migration weiterer Sammlungen der FAU. Sie stellen in ihrer Gesamtheit eine stetig wachsende Gebrauchsanweisung für den Einsatz von WissKI auf der einen und Leitfaden zur semantische Modellierung von Sammlungen auf der anderen Seite dar. Die Tutorien sind öffentlich zugänglich und können auch Dritten außerhalb der FAU als Leitfäden dienen. [7]

Zum jetzigen Zeitpunkt (November 2014) ist Phase 1 abgeschlossen, Phase 2 befindet sich noch in der Umsetzung.

Im Vortrag soll daher auf Phase 2 nicht näher eingegangen werden. Vielmehr sollen Phase 1 analysiert und anhand von Beispielen einige der Hindernisse und Chancen des Vorgehens hervorgehoben werden:

1. Das Projektformat mit regelmäßigen Workshops und die aktive Einbindung von Sammlungsmitarbeitern setzt deren Bereitschaft voraus, sich eingehend mit aus Sammlungssicht meist fachfremden Methoden und Techniken auseinanderzusetzen. Im Bereich des Semantic Web handelt es sich zudem um ein relativ neues und dynamisches Gebiet der Informatik, das nicht mit jahrzehntelanger Erfahrung und entsprechend ausgereiften Werkzeugen aufwarten kann wie etwa relationale Datenbankensysteme.
   Naturgemäß bildet auch der Zeitaufwand (und damit indirekt die personellen Kapazitäten einer Sammlung) eine Hürde.
2. Im Gegenzug festigt das Format bei den Sammlungsmitarbeitern das Verständnis und die Akzeptanz für die eingesetzten Technologien und Methoden. Das Eigeninteresse der Sammlungen wird klar erkennbar, was in den Augen der Autoren die erfolgreiche Umsetzung des Vorhabens trotz begrenzter Mittel entscheidend begünstigt.
3. Nicht zuletzt bauen die Sammlungen über die beteiligten Mitarbeiter Kompetenz im Bereich IT und semantischer Modellierung auf. Die Sammlungen können sich im Idealfall untereinander austauschen und mithin gegenseitig helfen und unterstützen. Realistischerweise ist dies im Pilotprojekt bei einfacheren Fragestellungen der Bedienung und Modellierung gegeben.
4. Die ontologische Modellierung und der fächerübergreifende Charakter der Workshops führen zu einer vertieften Reflexion über die eigene Sammlung und den sammlungs- bzw. fächerübergreifenden Kontext. So wurde bspw. das Bewusstsein für fachspezifische Termini bei gleichzeitiger Notwendigkeit für gemeinsame Terminologien gestärkt. Das Auftreten teils unerwarteter Überschneidungen in den einzelnen Disziplinen ermöglichte u.a. auch eine genauere Definition sonst kaum weiter hinterfragter Begrifflichkeiten.
5. Das CIDOC CRM als Referenzontologie bietet hier einen guten Ausgangspunkt, um gemeinsame Strukturen und Prozesse trotz unterschiedlicher Fachbegriffe herauszuarbeiten und deren Bedeutung klar zu formulieren. Andererseits können die in der Referenzontologie vorgegebenen Strukturen zu zunächst eigenwilligen Ergebnissen führen. Die starke Betonung von Ereignissen im CIDOC CRM steht zum Beispiel im scheinbaren Widerspruch zur objektzentrierten Dokumentation vieler Sammlungen und erfordert teilweise ein Überdenken tradierter Muster und eine Korrektur institutionalisierter Denkgewohnheiten. Dies kann wiederum die oben genannte Reflexion anregen.

6. Durch die Analyse der Datenstrukturen werden darüber hinausgehende Gemeinsamkeiten sichtbar. Grundlegende Herausforderungen wie die Einbindung und Verwendung gemeinsamer Normdaten und die Sicherung der Datenqualität können benannt, diskutiert und einheitlich angegangen werden.

Nach Abschluss der ersten Projektphase kann somit resümiert werden, dass die Umsetzung des vorliegenden Vorhabens – der einheitlichen Digitalisierungen der Sammlungen der FAU in Eigenregie – zwar einen deutlichen Einsatz von den Sammlungen selbst einfordert, sich aber Mehrwerte ergeben haben, die für sie mittel- und langfristig von Vorteil sind.

Literatur & Internetseiten:

[1] U. Andraschke und Marion Ruisinger, Die Sammlungen der Universität Erlangen-Nürnberg, 2007.
[2] G. Antoniou, P. Groth und F. van Harmelen, A Semantic Web Primer, MIT Press, 2012.
[3] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff (Hrsg.), Definition of the CIDOC Conceptual Reference Model, 2011.
[4] M. Scholz und G. Goerz, WissKI: A Virtual Research Environment for Cultural Heritage. In Proceedings of ECAI. 2012, 1017-1018.
[5] http://erlangen-crm.org (aufgerufen am 09.11.2014)
[6] http://wisski.cs.fau.de/sammlungen (aufgerufen am 09.11.2014)
[7] http://wisski.cs.fau.de/sammlungen/tutorials (aufgerufen am 09.11.2014)

Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen

André Blessing
Universität Stuttgart
andre.blessing@ims.uni-stuttgart.de

Melanie Dick
Universität Hildesheim
melaniedick@gmx.net

Ulrich Heid
Universität Hildesheim
heid@uni-hildesheim.de

Abstract

In vielen Projekten der Digital Humanities werden große Textmengen im Hinblick auf eine Forschungsfrage ausgewertet. Das interdisziplinäre Projekt *eIdentity* (BMBF, FKZ. 01UG1234) widmet sich beispielsweise der Frage nach multiplen kollektiven Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Damit sprachtechnologische Werkzeuge überhaupt auf das dort verwendete mehrsprachige Zeitungskorpus angewendet werden können, muss dieses zunächst von nicht für die Forschungsfrage relevanten Artikeln („off-topic-Artikel“) bereinigt werden. Nur so kann sichergestellt werden, dass nur Texte in die Auswertung einfließen, die Gegenstand der Forschungsfrage sind.

Viele Digital Humanities-Studien verwenden zur off-topic-Filterung lediglich Metadaten, wie zum Beispiel den Namen der Quelle, das Veröffentlichungsdatum oder den Autor. Für die Bereinigung des eIdentity-Korpus genügen diese Informationen aber nicht: es muss zusätzlich eine inhaltliche Filterung vorgenommen werden. Kantner et al. (2011) erstellten dazu manuell Schlagwortlisten um relevante und nicht relevante Artikel zu identifizieren. Die Erstellung solcher Listen ist allerdings zeitaufwendig und in der Praxis stellten sich diese als nicht vollständig heraus. Die Aufgabe der off-topic-Filterung ist ähnlich wie sogenannte „Spam-Filter“ für e-Mail; allerdings kann ein Spam-Filter anhand großer Mengen von Daten trainiert werden, weil der Nutzer in der Regel alle Nachrichten manuell nach Relevanz bewertet. In Digital Humanitites-Projekten ist die Anzahl der zu klassifizierenden Texte dafür zu groß; es braucht also Klassifikationsverfahren, die schon auf kleinen Mengen annotierter Texte gute Ergebnisse liefern.

In unserer Arbeit stellen wir einen neuen Ansatz vor; er geht aus von einer manuell annotierten Grundmenge von für die Forschungsfrage als relevant beziehungsweise irrelevant annotierten Artikeln. Ein Problem bei der Annotation ist die Auswahl der für das Training des Klassifikators nützlichen Artikel. Wenn zufällig ausgewählt wird, kann es sein, dass die Auswahl nicht repräsentativ für die zu klassifizierende Textmenge ist: wenn z.B. die Mehrheit aller Texte für die Forschungsfrage relevant ist, würden zu viele relevante und zu wenig irrelevante Texte annotiert. Hier kann durch die Nutzung von „Topic Modelling“ mit Latent Dirichlet Allocation (LDA; vgl. Blei et al. 2003) sichergestellt werden, dass eine besser nutzbare Auswahl getroffen wird.

Im ersten Schritt, der Feature-Extraktion, werden zunächst Merkmale aus Textdokumenten extrahiert. Diese extrahierten Merkmale werden in einem zweiten Schritt an einen Klassifikator übergeben, welcher die Artikel als relevant oder irrelevant kategorisiert (vgl. Abbildung 1).

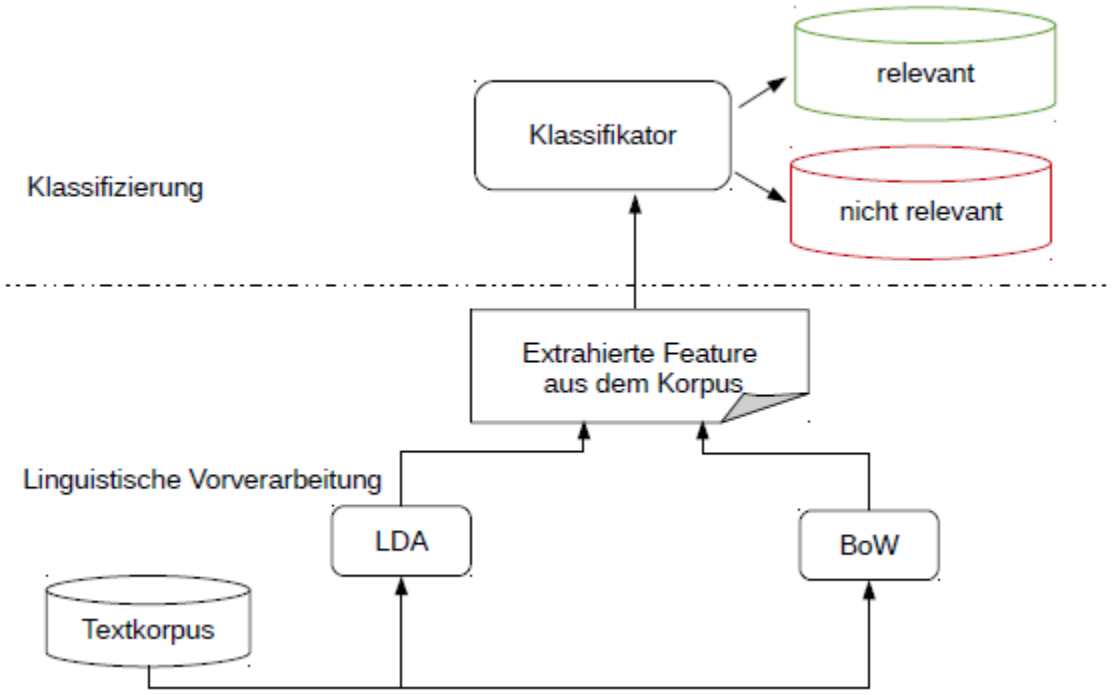


Abbildung 1: Zweistufiges Klassifikator-Modell

Für die Featureextraktion wird in unserem Experiment neben dem gängigen Bag-of-Words (BOW)-Modell, Topic Modelling durch LDA, ein generatives Wahrscheinlichkeitsmodell, eingesetzt und die Ergebnisse werden verglichen. Mit LDA können die Artikel entsprechend den vom System bestimmten „Topics“ vorsortiert werden. Ein „Topic“ soll mittels eines Wortclusters im abstrakten Sinne ein Themengebiet beschreiben. Der Klassifikator kann nun diese Topics explorieren (vgl. Abbildung 2) und anhand ihrer „Schlagwörter“ eine für den Forschungsgegenstand relevantere Auswahl kodieren. In unserem Forschungsprojekt konnten so sehr gut irrelevante Artikel zum Thema Sport, historische Konflikte, Buch- oder Filmkritik aufgespürt und als nicht relevant annotiert werden.

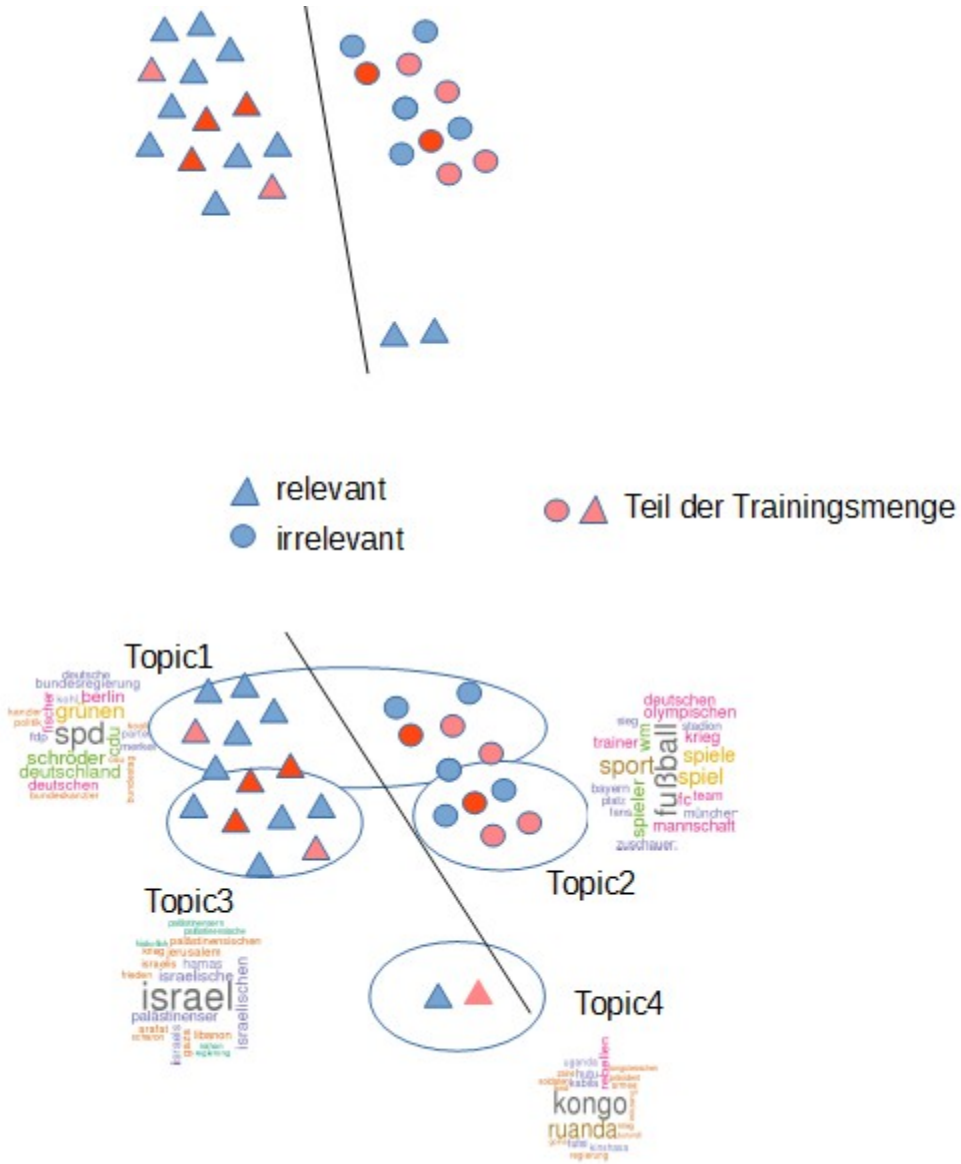


Abbildung 2: Exploration der Topics: oben: Zufallsauswahl; unten: LDA Topic-Modellierung

Im Rahmen der Vorverarbeitung wird zunächst eine Stopwortliste auf das Korpus angewendet. Häufig auftretende Wörter der geschlossenen Wortklassen werden damit als Merkmal ausgeschlossen. Als Baseline wird ein BoW-Modell verwendet, welches die Häufigkeit aller Wörter eines Dokuments als Hauptmerkmal für den Klassifikator aufbereitet. LDA hingegen lässt auf Wortebene für jedes Wort eine anteilige Zuordnung zu mehreren Topics zu. Die Information über die Verteilung über eine zuvor definierte Anzahl von Topics (beispielsweise 150) dient als Grundlage für den Klassifikator. LDA übergibt im Vergleich zum BoW-Modell Informationen an den Klassifikator, welche dieser wesentlich effizienter verarbeiten kann. Im zweiten Schritt folgt die Klassifikation in relevante und irrelevante Artikel (vgl. Abbildung 1).

Erste Ergebnisse haben gezeigt, dass das Topic Modelling für die Textklassifikation geeignet ist. Die manuell bewerteten Artikel (ca. 70 Dokumente – mehrfach bewertet) zeigten Accuracy-Werte von durchschnittlich 0,89. Erste Accuracy-Werte aus der automatischen Klassifikation (ca. 2.500 Dokumente – meist einfach

bewertet) mit LDA sowie BoW in der Vorverarbeitung, bewegen sich auch zwischen 0,8 und 0,9: die automatische Annotation liefert also die gleiche Qualität wie die menschlichen Annotatoren. Der Erfolg der Klassifizierung (Accuracy) beim LDA ist allerdings stark von der Anzahl der manuell ausgewählten Topics abhängig.

Weitere Variationen der beiden Verfahren wie beispielsweise eine Wortselektion mittels Part-of-Speech (POS)-Tagging sowie die Verwendung eines Stemmers, werden jeweils zusammen mit der linguistischen Vorverarbeitung eingesetzt: so kann deren Einfluss auf die Qualität der Ergebnisse des Klassifikators analysiert werden und die bestmögliche Merkmalsextraktion für die Entscheidung über die Artikelrelevanz kann bestimmt werden.

Literatur

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003): *Latent dirichlet allocation.* IN: The Journal of Machine Learning Research 3, S. 993-1022.

Cathleen Kantner, Amelie Kutter, Andreas Hildebrandt, Mark Püttcher (2011): *How to get rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts, International Relation Online Working Paper.* 2011/2, Juli 2011, Stuttgart: Universität Stuttgart.

eIdentity (2014): *Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften (eIdentity).* URL: http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html Stand: 08.10.2014.

Vorschlag für ein POSTER:

ADHO Special Interest Group for Libraries and Digital Humanities

Special Interest Group (SIG) Organisatoren und Autoren:

Zoe Borovsky, UCLA Libraries Libraries, U.S.A.

Angela Courtney, Indiana University Libraries, U.S.A.

Isabel Galina, Universidad Nacional Autónoma de México

Stefanie Gehrke, Biblissima, France

Hege Stensrud Høsøien, National Library, Norway

Sarah Potvin, Texas A&M University Libraries, U.S.A.

Thomas Stäcker, Herzog August Library, Germany

Glen Worthey, Stanford University Libraries, U.S.A.

Das Poster hat zum Ziel, den Vorschlag einer Etablierung einer *ADHO Special Interest Group for Libraries and Digital Humanities* vorzustellen und im Rahmen der DHd 2015 zu diskutieren. Es wird die Gegenstände näher erläutern, die zu dem Antrag geführt haben, seine Genese als internationale Unternehmung darstellen und den Organisatoren dieser Gruppe die Möglichkeit geben, den Vorschlag im persönlichen Gespräch möglichen Unterstützern und an der Arbeit in der Gruppe Interessierten bekannt zu machen und zu erläutern. Konferenzteilnehmer, die nicht unmittelbar im Bibliotheksbereich arbeiten, soll das Poster verdeutlichen, von welcher Bedeutung DH heute in Bibliotheken ist.

Ziele

ADHO *Libraries and DH SIG* zielt darauf, die Zusammenarbeit und Kommunikation zwischen BibliothekarInnen und WissenschaftlerInnen zu fördern. Durch Einrichtung dieser SIG wird ADHO seinem Ziel gerecht, den Austausch zwischen den ADHO Organisationen und neuen DH Initiativen, die sich von bibliothekarischer Seite aus entwickeln, zu etablieren. Wir sind der Überzeugung, dass diese Verbindung zu einer sich intellektuell befruchtenden “doppelten Staatsbürgerschaft” führt, wo BibliothekarInnen und DH WissenschaftlerInnen gleichermaßen

in beiden Bereichen zu Hause sind. Durch die Förderung einer solchen “doppelten Staatsbürgerschaft” werden Bibliotheken und BibliothekarInnen in die Lage versetzt, Möglichkeiten besser zu erkennen, wie sie sich in DH Projekte und Forschungsarbeiten einbringen können sowie insgesamt die Herausforderungen, vor denen sie stehen, besser zu bewältigen. Diese Herausforderungen schließen z.B. ein a) Finanzierungsmöglichkeiten zu ermitteln, Freistellungen und Schulungen zu ermöglichen, technische Infrastruktur bereitzustellen, um DH Projekte durchzuführen, b) die wechselnden Begriffe von „Dienstleistung“ und „Forschung“ mit Blick auf die meist kooperative Natur von DH Projekten zu hinterfragen und c) eine Kultur der digitalen Forschung in der Bibliothek zu etablieren.

Das Ziel der ADHO *Libraries and DH SIG* wird sein:

- Rat und Unterstützung anzubieten für die neu sich herausbildende Gruppe von BibliothekarInnen, die entweder eigene oder DH Projekte mit nicht der Bibliothek angehörigen digitalen Geisteswissenschaftlern verfolgen,
- sich einzusetzen für Initiativen, die sowohl für Bibliotheken als auch DH von Interesse und von Vorteil sind (z.B. “Best Practices for TEI in Libraries” und andere Richtlinien oder *best practice* Beispiele mit Bezug auf DH, die sich auf die Bibliothek beziehen)
- zu dokumentieren, wie sich BibliothekarInnen und Bibliotheken diesen Herausforderungen stellen
- Informationen zu liefern über verfügbare Ressourcen und Möglichkeiten (z.B. Schulungen, Drittmittel), die die Zusammenarbeit von verschiedenen, im Bereich der DH Forschenden, insbesondere in der Bibliothek, befördern,
- beispielhafte Projektergebnisse von BibliothekarInnen zu zeigen, die im Bereich der DH arbeiten,
- bibliothekarische Sichtweisen und Kompetenzen der gesamten DH community zu vermitteln.

Ein erstes Ziel der SIG wird sich darauf konzentrieren, Arbeitsbeziehungen zwischen internationalen Organisationen mit Bibliotheksbezug zu entwickeln, wie z.B. der ACRL DH Interest Group, der Digital Library Federation, der TEI in Libraries Special Interest Group, der

Society for American Archivists, der Association for Information Science and Technology oder der International Federation of Library Associations and Institutions.

Tätigkeiten

Mit Blick auf konkrete Aktivitäten würde die SIG sich einsetzen, um in Zusammenarbeit mit bibliothekarischen Organisationen Folgendes zu erreichen:

- Ermittlung und Nachweis von Bibliotheken, die DH Projekte durchführen und DH Organisationen, in denen Bibliotheken aktive Partner sind (z.B. das TEI Consortium SIG on Libraries)
- Konferenz-Sessions zu organisieren, einerseits für BibliothekarInnen auf DH Konferenzen, andererseits für andere DH Interessierte auf Tagungen, die sich in erster Linie an BibliothekarInnen richten (wie ALA, ACRL, ARLIS, DLF, Bibliothekartag, etc.)
- Workshops, Schulungen und Konferenz-Sessions zu organisieren, die dazu dienen, BibliothekarInnen stärker in die allgemeine DH Community zu integrieren und DH bezogene Bibliotheksprojekte vorzustellen.

Teilnehmen kann jeder, der Interesse an der Sache hat. Derzeit haben 130 Personen ihr Interesse bekundet, bei der SIG mitzuwirken. Wir denken jedoch, dass das potentielle Interesse weltweit weit höher liegt. Die Hoffnung besteht, dass durch die Posterpräsentation die SIG auch in der deutschsprachigen Community bekannter gemacht wird und neue Mitglieder geworben werden können.

Hinweise

Eine öffentliche Zotero Group zum Thema DH in libraries findet sich hier: https://www.zotero.org/groups/adho_library_sig

Poster proposal DHd 2015

SMuFL-Browser und oXygen GlyphPicker Plugin

Werkzeuge zur Integration musikalischer Symbole in TEI

Alexander Erhard* Peter Stadler†

Die digitale Edition von musikalischen Texten und Texten über Musik bedarf an vielen Stellen der Darstellung musikalischer Zeichen und Symbole. Im Bereich 1D100–1D1FF des aktuellen Unicode-Standards sind zwar „Musical Symbols“[1] definiert, diese insgesamt 220 Zeichen decken aber nur einen Bruchteil des in der Praxis benötigten Repertoires ab. Eine breiter angelegte Systematik musikalischer Zeichen liegt in den Spezifikationen des *Standard Music Font Layout* (SMuFL)[2] vor, welche musikalischen Symbolen – ähnlich der *Medieval Unicode Font Initiative* (MUFI)[3] im Bereich mittelalterlicher Zeichen – die Codepoints der Unicode Private Use Area zuordnen. Obwohl gegenwärtig kein Versuch unternommen wird, diese Zeichen in den offiziellen Unicode-Standard einzubringen, so stellt SMuFL doch für den Bereich musikalischer Symbole aufgrund seiner breiten Abdeckung einen de-facto-Standard dar.

Der Gebrauch von Unicode-Zeichen ist (neben dem Auszeichnen nach MEI oder MusicXML, dem Einbinden von Grafiken etc.) eine der von der TEI Music SIG in ihren Empfehlungen zu „TEI with Music Notation“[4] diskutierten Möglichkeiten, musikalische Zeichen in TEI-Dokumenten zu repräsentieren. Einen

*Richard Strauss: Werke. Kritische Ausgabe, Universität München
†Carl-Maria-von-Weber-Gesamtausgabe, Universität Paderborn

[1] Vgl. Perry Roland, *Proposal for Encoding Western Music Symbols in ISO/IEC 10646*, revised February 19, 1998, online verfügbar unter `https://archive.today/PzkaT`

[2] `http://www.smufl.org`

[3] `http://folk.uib.no/hnooh/mufi/`

[4] `http://www.tei-c.org/SIG/Music/twm/`

besonders geeigneten Ort hat die Verwendung von Musiksymbolen unseres Erachtens dort, wo einzelne musikalische Zeichen (oder kurze Sequenzen) losgelöst von einem größeren musikalischen Kontext in Worttext eingeflochten sind. Um das Finden und Einfügen dieser Symbole nach dem SMuFL-Standard in TEI Dokumenten zu vereinfachen, haben wir den Webservice „SMuFL-Browser“[5] sowie das oXygen-Plugin „GlyphPicker“[6] entwickelt.

Grundlage des Webservices sind Definitionen der mehr als 2000 Zeichen und Symbole im TEI-Format (mittels `<charDecl>` und `<char>`), die auf den SMuFL-Spezifikationen beruhen und als standardisierte Zielpunkte bei der Codierung von Musiksymbolen in TEI-Dokumenten (z. B. in der Form `<tei:g ref="http://mywebservice/smufl-browser/restQuarter"/>`) dienen können. Die Web-Oberfläche des SMuFL-Browsers erlaubt das bequeme Durchsuchen der Definitionen und stellt für jedes Musiksymbol neben Beispiel-Graphiken auch Code-Fragmente zum Einfügen in TEI-Dokumente bereit. Die Funktionalität orientiert sich an der ENRICH gBank application,[7] geht aber durch die Bereitstellung einer REST-Schnittstelle für maschinelle Abfragen darüber hinaus. Via Content Negotiation werden Anfragen neben HTML auch in den Formaten TEI-XML oder JSON beantwortet, wodurch der Webservice auch als flexible Datengrundlage externer Tools dienen kann.

Die oXygen-Erweiterung „GlyphPicker“ versteht sich als Ergänzung zur regulären Zeichentabelle von oXygen und nutzt dafür den vorgenannten Webservice SMuFL-Browser. Sie unterstützt das Auffinden und Einfügen von Unicode-fremden Zeichen, für die Definitionen in TEI mittels der Elemente `<char>` und `<charDecl>` vorliegen. Das Plugin bereitet entsprechende Definitionen zu Zeichentabellen in oXygen auf und stellt Mittel bereit, Verweise auf diese Definitionen (in Form von `<g>`-Elementen) in Dokumente einzufügen. Die Datenquellen der Zeichen-Definitionen sind im Plugin frei bestimmbar: Webservices wie der SMuFL-Browser werden ebenso unterstützt wie lokal abgelegte TEI-Dateien mit projektspezifischen Vorgaben. Das Plugin ist in der Texteditor- und Autor-Ansicht von oXygen nutzbar und kann sowohl selbständig als auch in Kombination mit einem Autor-Framework eingesetzt werden.

[5] `https://github.com/Edirom/SMuFL-Browser`
[6] `https://github.com/aerhard/glyphpicker`
[7] `http://www.manuscriptorium.com/apps/gbank/`

Dingler Dissemination
Highlights aus 6 Jahren »Digitalisierung des Polytechnischen Journals«

Marius Hug, M. A.; Martina Gödel, M. A.;
Timo Arndt, B.A.; Una Schäfer, B.A.

Einreichung zum Call for Posters
DHd-Tagung 2015
Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation
7. November 2014

Abstract

Am 28. Februar 2015 endet die Laufzeit des von der DFG geförderten Projekts »Digitalisierung des Polytechnischen Journals« am Institut für Kulturwissenschaft der Humboldt Universität zu Berlin. Damit ist die DHd 2015 genau der richtige Zeitpunkt, um Bilanz zu ziehen. Mit unserem Posterbeitrag möchten wir die wichtigsten Ergebnisse aus der kooperativ angelegten Projektarbeit präsentieren. Highlights sind dabei: Nachhaltige Datenspeicherung und -präsentation, Ausführliche Tagging- und Projektdokumentation per ODD, Zurverfügungstellen der Daten über geeignete Schnittstellen zu Analysezwecken bspw. für Computerlinguisten, Aufbereitung der im TEI P5-Format vorliegenden Daten zur wissenschaftlichen Weiterverarbeitung (bspw. Umwandlung historischer Währungen, Visualisierung auf einer Timemap), Bearbeitung des sehr umfangreichen Bildmateriels mittels Image-Markup-Tool, sowie ein vollkommen neuartiger Transfer unserer Daten — aus der virtuellen Welt in die Welt der Objekte — als Grundlage für die Kooperation mit einem Museum.

1 Dinglers Polytechnisches Journal (DPJ)

Das »Polytechnische Journal« wurde 1820 vom Augsburger Fabrikanten und Chemiker Johann Gottfried Dingler begründet. Dingler studierte wichtige Zeitschriften (die meisten davon aus England, Frankreich, später aber auch den USA), wählte relevante

Artikel aus, übersetzte und publizierte sie in seinem Journal. Mit einer Laufzeit von 111 Jahren ist diese Zeitschrift ein beispielloses, europaweites Archiv der Technik-, Wissens- und Kulturgeschichte. Besonders bemerkenswert ist die Aktualität der Publikation: So verging kaum Zeit zwischen Erstveröffentlichung der Artikel und Erscheinen der übersetzten Version im DPJ.

2 Das Digitalisierungsprojekt

Im von der DFG geförderten Projekt am Institut für Kulturwissenschaft der Humboldt-Universität zu Berlin wurde der komplette Bestand vom DPJ digitalisiert. Die Bilddigitalisierung wurde an der SLUB-Dresden durchgeführt. Für die Textdigitalisierung und Basisauszeichnung der über 200.000 Seiten war der Dienstleister Editura GmbH zuständig. Alle Bände sind per TEI-P5 kodiert. Das Journal ist online (CC by-nc-sa 3.0) unter `www.polytechnischesjournal.de` verfügbar.

Nachhaltigkeit, Clarin-D

Ein großes Problem für Digitalisierungsprojekte ist die nachhaltige Verfügbarmachung der Daten, Stichwort: Langzeitarchivierung. Für das Projekt Dingler-Online bedeutet die Zusammenarbeit mit dem BMBF geförderten Verbundprojekt **CLARIN-D** einerseits die Möglichkeit der Dissemination der Projektdaten, andererseits ist dadurch eine langfristige Sichtbarkeit des Projekts garantiert.

Durch die Kooperation mit der Berlin-Brandenburgischen Akademie der Wissenschaften, konkret dem DFG-Projekt **Deutsches Textarchiv** (DTA), profitiert Dingler-Online sehr direkt von deren technischem Know-how. Konkret zu nennen wären hier bspw. die orthographische Normalisierung und linguistische Analyse (POS, Tokenisierung, Lemmatisierung, ...) sowie die Nutzung der elaborierten Rechercheschnittstelle.

Ausführliche Tagging- und Projektdokumentation per ODD

Alle editorischen Entscheidungen, die verwendeten Elemente und Attribute wurden in der ODD-Datei ausführlich beschrieben bzw. festgelegt. Für die eingesetzten Attribute wurden geschlossene Listen mit projektspezifischen Werten definiert und ihr Einsatz erklärt. Konkrete Quelltextbeispiele aus dem Projekt veranschaulichen das Vorgehen. Eine Transformation in das HTML-Format mittels des TEI-Tools OxGarage ermöglicht die Sichtbarmachung dieser Dokumentation im Look and Feel der TEI Guidelines selbst.

Das restriktiv formulierte Datenmodell und seine transparente Dokumentation unterstützen die möglichst schwellenarme automatisierte interne und externe Weiterverarbeitung. Die Dokumentation ist online unter `http://dingler.culture.hu-berlin.de/Schema/dingler.html` verfügbar.

Historische Daten

Ein wichtiges Thema – nicht zuletzt aufgrund der stets größer werdenen digital vorliegenden Datenmengen – ist die Visualisierung. Exemplarisch wurde in unserem Projekt ein solches Verfahren anhand von Patentdaten durchgeführt (s. Abb. 1). Diese eignen sich in besonderem Maße, da die Einheit Patentschrift sehr überschaubar ist, aber dennoch die für ein tief granuliertes TEI-Tagging benötigten Elemente enthält: Names, Dates, People, und Places (TEI P5 Guidelines, ch. 13). Der Workflow, mit Hilfe dessen die in Patentlisten vorliegenden Einträge in das zur Darstellung auf einer Timemap benötigte KML (Keyhole Markup Language) transformiert wurde, ist gut dokumentiert.

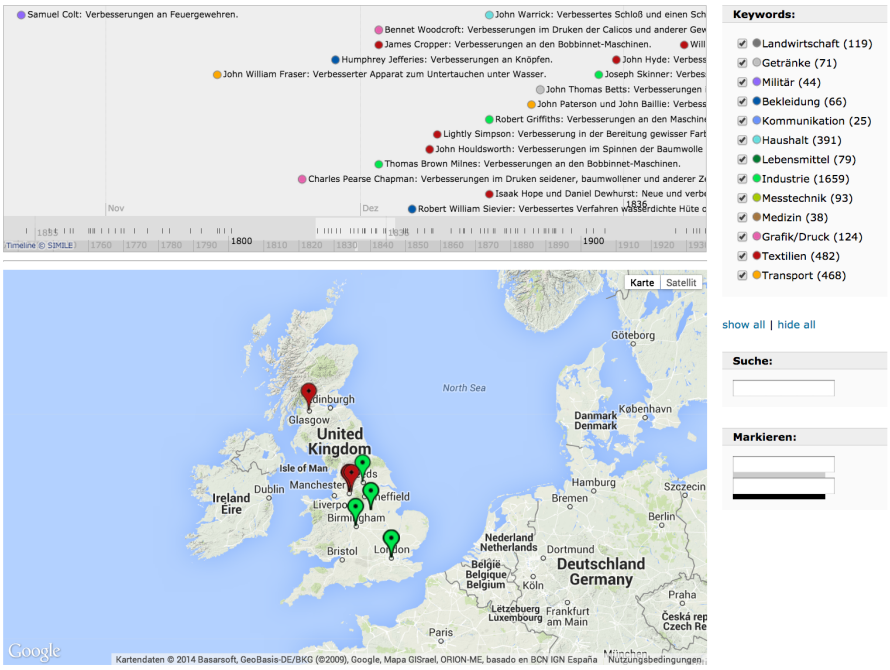


Abbildung 1: Visualisierung der TEI-Daten auf einer Timemap.

Historische Varianz stellt eine weitere Herausforderung für unsere Daten dar. Dies betrifft nicht nur die Texte, sondern auch Zahlen und Einheiten. Hier wurden die in den digitalisierten Daten vorkommenden Währungen und Einheiten gesammelt und über Listen entsprechende Umrechnungen zugewiesen, wobei sich bspw. für den Wiener Fuß folgendes Bild ergibt:

Image Markup Tool

Neben dem Textbestand (etwa 420 Mio. Zeichen) ist v. a. die große Menge an Bildern bzw. Zeichnungen hervorzuheben, wobei neben rund 3500 Falttafeln – eine Tafel enthält bis zu 114 Einzelfiguren – auch zahlreiche Figuren von Text umflossen gedruckt wurden. Um der Bedeutung des Bildmaterials im technischen Kontext gerecht zu werden, wurde

```
wienerfuss:
    name: Wiener Fuß
    unit: Wiener Fuß
    wp: http://de.wikipedia.org/wiki/Fuß_(Einheit)
    conversions:
        zentimeter: x * 31.608
        millimeter: x * 316.08
```

Abbildung 2: Syntax des projekteigenen Einheitenrechners.

hier bei der Erschließung besonders großer Aufwand betrieben.

Die Tafelwerke des »Polytechnischen Journals« wurden auf der Ebene der Einzelfiguren mit Koordinaten versehen. Der Mehrwert dieser Auszeichnungsstrategie besteht für den Nutzer zum einen in der konkreten Referentialisierung von Textpassage und Einzelfigur sowie individueller Anordnungs-, Betrachtungs- und Ausgabemöglichkeiten.

3 Museum

Hintergrund der von uns auf Grundlage unserer im Projekt aufbereiteten Daten angeregten Kooperation mit einem Museum – erste Prototypen sind gerade im Einsatz – ist folgende These: Weder ist Sammlung jenseits von Wissenschaft noch Forschung jenseits der Dinge möglich. Von einer Zusammenarbeit profitieren demnach beide Seiten. Ziel ist es, ausgewählte Objekte eines Museums multimedial erlebbar zu machen und damit den Besuchern ein neuartiges, interessantes Objekterlebnis zu ermöglichen. Andererseits erzeugt diese *Nachnutzung* unserer Daten eine längerfristige Sichtbarkeit der in den Projektdaten gespeicherten Informationen.

Mit Hilfe einer kostenlos zur Verfügung stehenden und speziell für das Projekt entwickelten App, werden dem Museumsbesucher zu ausgewählten Objekten vertiefende Informationen angeboten.

Denkbar sind hier bspw.:

- weiterführende Informationen zum Ausstellungsgegenstand und seiner Geschichte
- Informationen zu beteiligten Akteuren (Erfinder, Produzenten, Firmen...)
- veranschaulichende Bilder/Figuren
- verwandte Themenfelder, bspw. per Schlagwortwolke

Die App-Entwicklung wird bis zum Februar soweit sein, dass wir diese neuartige Nutzung unserer Daten tatsächlich hands-on vorführen können.[1]

[1] An dieser Stelle möchten wir darauf hinweisen, dass die Einreichung von Christian Kassung für einen Vortrag mit dem Titel »Making Things Chatter« eine weitere Technologie zur Verlinkung von Daten und Objekten präsentiert. Im Unterschied zu unserem Poster liegt der Fokus dort auf dem Museum.

4 Fazit

Zusammengefasst verfolgt unsere Posterpräsentation eine doppelte Strategie: 1) Das in der Community teilweise schon bekannte Projekt kann in verschiedenster Hinsicht Erfahrungen der letzten Jahre weitergeben und so für andere (kleinere) Digitalisierungsprojekte inspirierend sein. 2) Wir würden uns freuen, mit unserer neuartigen Idee einer Verknüpfung von Text- und Objektdaten (aus dem Museum) mit der Community in Diskussion zu treten und für das weitere Vorgehen von zu erwartenden Synergieeffekten zu profitieren.

5 Webressourcen

- CLARIN-D: http://de.clarin.eu/de/
- DinglerOnline: http://www.polytechnischesjournal.de
- Dingler ODD: http://dingler.culture.hu-berlin.de/download
- DTA: http://www.deutschestextarchiv.de
- Google Timemap: https://code.google.com/p/timemap/
- Image Markup Tool: http://tapor.uvic.ca/~mholmes/image_markup/
- KML: https://developers.google.com/kml/

Erweiterte Publikationen in den Geisteswissenschaften
Zwischenergebnisse des DFG-Projektes Fu-PusH

Ben Kaden und Michael Kleineberg
Universitätsbibliothek der Humboldt-Universität zu Berlin, Deutschland

Das DFG-Projekt *Future Publications in den Humanities* (Fu-PusH), angesiedelt am Jacob-und-Wilhelm-Grimm-Zentrum der Humboldt-Universität zu Berlin, untersucht die Potentiale des digitalen Publizierens in den Geisteswissenschaften und erarbeitet szenarienbasiert Handlungsempfehlungen für akademische Infrastruktureinrichtungen wie insbesondere Universitätsbibliotheken und Rechenzentren, um den funktionalen Anforderungen unterschiedlicher geisteswissenschaftlicher Fachrichtungen gerecht zu werden.

Für Publikationsformen, die sich nicht mehr primär an der Druckkultur orientieren mit dem Versuch Printmedien etwa in Form von Monographien, Fachartikeln oder Sammelbandbeiträgen lediglich digital nachzubilden, sondern die genuinen Eigenschaften des Digitalen in den Mittelpunkt stellen, bietet sich die Bezeichnung *enhanced publications* bzw. „erweiterte Publikationen“ an. Solche Publikationsformen werden häufig als komplexe digitale Dokumente bzw. Dokumentensysteme charakterisiert, die sich unter anderem durch nicht-lineare Hypertextstrukturen, multimediale Zusatzmaterialien, integrierte Forschungsdaten, adaptive Darstellungsvarianten, dynamische Versionierung, kontextuelle Anreicherung sowie maschinenlesbare semantische Strukturierung auszeichnen. Ihre Vorteile liegen in einer engen Verknüpfbarkeit heterogener Elemente wie beispielsweise Digitalisate, Textkorpora, Datenbanken, Annotationen, Normdateien, Geoinformationen und narrativ-interpretativen Auseinandersetzungen mit diesen Objekten.

Auf diese Weise bieten erweiterte Publikationsformen die Möglichkeit nicht nur die Forschungsergebnisse, sondern auch die zu Grunde liegenden Forschungsdaten bzw. Forschungsprozesse in einem gemeinsamen Kontext zur Verfügung zu stellen, wobei die Grenzen zwischen Bearbeitungsraum, Kommunikationsraum und Veröffentlichungsraum sehr durchlässig werden.

Erweiterte Publikationen lassen sich demnach vor allem dadurch kennzeichnen, dass sie die in den Geisteswissenschaften etablierte Grundform der narrativen Auseinandersetzung mit einem Forschungsgegenstand an mindestens drei Stellen öffnen: Erstens kann ein direkter Bezug zu den Forschungsgrundlagen hergestellt werden, etwa durch eine Einbindung von bzw. Verlinkung zu digital vorliegenden Forschungsquellen wie Referenztexten, Abbildungen, Tondokumenten oder Filmsequenzen. Zweitens kann das narrative Element selbst über entsprechende semantische Tiefenauszeichnung durch Annotationen und Metadaten zu einem vielfältig vernetzbaren und maschinell prozessierbaren Datum werden. Drittens werden Interaktions- und Vernetzungsspuren solcher Dokumente wie beispielsweise Zitationen, Verlinkungen, Rezensionen, Verschlagwortungen oder Nutzungstatistiken darstell- und auswertbar.

Ob und inwieweit sich derartige Publikationskonzepte tatsächlich in der Praxis der Wissenschaften durchsetzen werden, hängt freilich vom Bedarf und auch der Bereitschaft der jeweiligen Fachgemeinschaften ab. Um auf diese Fragestellung einen substantiellen Zugriff zu erhalten werden im Fu-PusH-Projekt die Bedarfe, funktionale Anforderungen und Einstellungen systematisch in Interviews mit ExpertInnen aus dem Bereich der Geisteswissenschaften, aber auch mit Vertretern von Infrastruktureinrichtungen sowie Intermediären wie Verlagen und Anbietern alternativer Publikations-plattformen ermittelt.

Bei den zielgruppenorientierten Befragungen handelt es sich um qualitative und offene Leitfaden-interviews, die ein möglichst breites Spektrum an Perspektiven und thematischen Facetten abdecken sollen. Das Erhebungsinteresse schließt dabei neben technologischen Desiderata hinsichtlich digitaler Arbeits- und Publikationsumgebungen auch wissenschaftskulturelle, wissenschaftsstrukturelle sowie wissenschaftspolitische Anforderungen und Spielräume ausdrücklich ein.

In der Präsentation arbeiten wir zunächst den definitorischen Rahmen für erweiterte Publikationen heraus und spezifizieren funktionale Anforderungen an wissenschaftliche Veröffentlichungsverfahren. Im Anschluss setzen wir dies in Relation zu den Ergebnissen der Befragungen. Dabei differenzieren wir einen Ist-Zustand und einen auf einer Desiderats-Analyse basierenden Perspektiv-Zustand hinsichtlich der Publikationskulturen in verschiedenen geisteswissenschaftlichen Fachrichtungen. Auf diese Weise sollen aktuelle Transformationsprozesse in den Geisteswissenschaften sichtbar gemacht werden. Im Fokus stehen dabei insbesondere Einstellungs- und Handlungsmuster in Bezug auf:

- das wissenschaftliche Publizieren generell,
- die Erhebung, den Umgang sowie die Nachnutzung von Forschungsdaten,
- mögliche methodologischen Veränderungen unter dem Einfluss der Digital Humanities,
- das Publikationsverhalten insbesondere vor dem Hintergrund von Open Access,
- das Forschungsverhalten im Kontext von Open Science bzw. Open Scholarship,
- das Qualitätssicherungsverfahren des wissenschaftlichen Publizierens (Peer Review, etc.),
- die Dienstleistungen von Infrastruktureinrichtungen (z.B. Rechenzentren, Bibliotheken, Archive),
- die von Wissenschaftspolitik und Förderinstitutionen gesetzten Rahmenbedingungen,
- sowie mögliche Risiken im Zuge der digitalen Transformation.

Die Zwischenergebnisse des Fu-PusH-Projektes zeigen bereits sehr deutlich die Unterschiede im Forschungs- und Publikationsverhalten sowohl zwischen den Geisteswissenschaften und den so genannten MINT-Disziplinen als auch innerhalb des disziplinären Spektrums der Geisteswissenschaften selbst.

In diesem Zusammenhang soll die Frage verfolgt werden, inwieweit fachspezifische Publikationskulturen auch unterschiedliche technische und konzeptionelle Lösungen im Bereich der erweiterten Publikationen erfordern. Dies ist von besonderer Bedeutung, wenn man im Gegenzug die Herausforderung technischer Standardisierung zur Gewährleistung von Interoperabilität berücksichtigt. An dieser Stelle werden die Risiken deutlich, die generell von Technologien im Kontext der Digital Humanities ausgehen. Zum einen liegen bisher kaum Erfahrungswerte vor, mit denen sich eine tatsächliche Relevanzbewertung von Informationsinfrastrukturen bzw. Publikationsszenarien vornehmen lässt. Zum anderen besteht die Gefahr, dass neue technische Dispositive bestimmte Forschungs- und Erkenntnispraxen begünstigen und dafür andere weniger angemessen berücksichtigen.

Dies unterstreicht zusätzlich die Bedeutung der Modellierung komplexer Szenarien bevor Innovationsschritte angestoßen werden, da naturgemäß der Erfolg derartiger technischer Entwicklungen maßgeblich von der Passung mit dem tatsächlich Bedarf und den Erwartungen – auch perspektivisch – der jeweiligen Zielgruppen abhängt. Insofern, und dies ist eine zentrale Erkenntnis auch dieses Projektes, müssen Schritte von Seiten der Infrastruktureinrichtungen, die die Forschungsrealität der Wissenschaftsgemeinschaften betreffen, im Dialog mit diesen erarbeitet werden.

Europeana Sounds – Ein Portal zu Europas klingendem Kulturerbe

Ute Sondergeld, Max Kaiser (Österreichische Nationalbibliothek, Wien)

Abstract

Die Massendigitalisierungsprojekte der vergangenen Jahre und die in diesem Zusammenhang entstandenen Portale und Repositorien haben zwar dazu beigetragen, den Zugang zu Primär- und Sekundärquellen des Kulturerbes zu erleichtern, unterliegen oftmals aber noch immer institutionellen oder regionalen Begrenzungen oder sind fokussiert auf bestimmte Dokumenttypen. Die Zusammenführung heterogener internationaler Datenbestände und verschiedener Informationstypen in einem zentralen Verweissystem sowie die Bereitstellung von Werkzeugen zu ihrer Bearbeitung und Weiterverarbeitung bietet die Chance, die Voraussetzung wissenschaftlichen Arbeitens – Sichtung, Auswahl und Kontextualisierung geeigneten Quellenmaterials zur Entwicklung von Forschungsfragen – zu stärken.

Das Projekt *Europeana Sounds* zielt darauf ab, die Datenbasis für den bisher weniger beachteten Themenbereich der Audioinhalte und der damit verwandten Dokumente innerhalb der digitalen Bibliothek *Europeana* zu stärken und so neben den bereits bestehenden Aggregatoren *APEX* (Archive), *EUscreen* (Fernsehen), *European Film Gateway* (Film) und *TEL* (Bibliotheken) die Infrastruktur für eine weitere Domäne der europäischen digitalen Bibliothek aufzubauen. Die Spannbreite der im Projektrahmen zu referenzierenden Objekte reicht von Musik aller Sparten über Radiosendungen und Sprachaufnahmen bis hin zu Klanglandschaften, Natur- und Umweltgeräuschen. Der Einschluss verwandter Materialien wie Fotografien, Korrespondenzen, Textbücher, Musikdrucke und -handschriften trägt dazu bei, den Bestand audiobezogener Inhalte innerhalb der *Europeana* um mehr als das Doppelte auf insgesamt über eine Million Referenzen zu steigern und für Europa kulturell und historisch bedeutsame Objekte zentral zugänglich zu machen.

Grundlage der Datenaggregation bildet ein spezifisches, den Anforderungen von Audioobjekten entsprechendes *European Data Model Profile for Sound* sowie eigens entwickelte kontrollierte Vokabulare, die verschiedene Ebenen der referenzierten Objekte beschreiben. Basierend auf internationalen Normdaten tragen diese Vokabulare dazu bei, Metadatenqualität und das Retrieval multilingualer Daten, einer der großen Herausforderungen internationaler Datenbanken, sicher zu stellen.

Die Entwicklung von Tools zur Bearbeitung von Metadaten und Anwendungen zur Weiterverarbeitung von digitalen Objekten eröffnen Interaktionsmöglichkeiten mit dem Datenbestand. Zum Teil auf Anwendungen beruhend, die in anderen *Europeana*-Projekten entwickelt wurden, sollen die Werkzeuge zum Beispiel eine Korrektur und Transkription digitaler Objekte sowie ihre Klassifikation durch *social tagging* ermöglichen. Eine Kontextualisierung von Inhalten ist auf objektiver Ebene durch die Verlinkung zu ähnlichen Ressourcen oder Hintergrundinformationen sowie auf subjektiver Ebene durch persönliche Kommentare und Diskussionen vorgesehen. Zusammen mit der Möglichkeit einer individuellen Zusammenstellung von Objekten (Kuratierung) und der Einrichtung eines persönlichen Bereiches innerhalb des Portals wird eine Infrastruktur zur Verfügung gestellt, die sowohl dem allgemeinen Publikum, Experten wie auch Forschenden Möglichkeiten der Datenbearbeitung und –generierung bietet (Oomen & Aroyo, 2011; Chen, 2014).

Die Verbreiterung der Datenbasis durch die Zusammenführung verschiedener Bestände und die Bereitstellung einer Infrastruktur zu deren Weiterverarbeitung kann auf der einen Seite zu einer qualitativen Verbesserung des Informationssystems *Europeana* führen, eröffnet andererseits Möglichkeiten für die geisteswissenschaftliche Forschung und der Generierung neuen Wissens über das europäische Kulturerbe.

Das Projekt *Europeana Sounds* wird im Zeitraum von Februar 2014 bis Jänner 2017 von insgesamt 24 Institutionen aus 12 Ländern durchgeführt und von der Europäischen Kommission im Rahmen des ICT Policy Support Programme ko-finanziert. Die Österreichische Nationalbibliothek stellt im Rahmen des Projekts ihre wertvollsten Musikhandschriften von Komponisten des 17. bis 19. Jahrhunderts, die ihren Ruf als eine der bedeutendsten historischen Musiksammlungen weltweit begründen, zur Verfügung.

Literatur

Chen, C. (2014): Design for User Engagement on Europeana Channels. Master thesis, Delft University of Technology, Faculty of Industrial Design Enginieering

Oomen, Johan & Aroyo, Lora: Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges.

Annotation und Analyse des literaturtheoretischen und -kritischen Diskurses in deutschsprachigen Poetiken (1770 bis 1960)

Die Poetik bildet als Wissensgebiet die theoretische Basis der Literatur- und Sprachwissenschaft von der Antike bis ins 20. Jahrhundert hinein. Als Poetik wird zugleich die Textsorte bezeichnet, die diese Theoriegrundlagen enthält und in der diese diskutiert und literaturkritisch geprüft werden. Im Zuge dieser diskursiven Verhandlung werden Verweise auf jeweils andere Poetik-Autoren und literarische Beispiele benutzt und teilweise kritisch bewertet. Die Analyse der quantitativen und qualitativen Aspekte dieser diskursiven Verweisungsstrukturen ist ein zentrales Ziel des Projekts ePoetics, einem BMBF-geförderten Kooperationsvorhaben der Universität Stuttgart und der Technischen Universität Darmstadt, und trägt zur Erforschung der Entwicklung grundlegender literaturtheoretischer Begriffe und Konzepte bei. Im Rahmen dieser Untersuchung wurden zwanzig deutschsprachige, für die Zeit von 1770 bis 1960 repräsentative Poetiken ausgewählt. Diese werden als TEI-konformes Corpus (inklusive der im Folgenden dargestellten Annotationsebenen) im Repositorium der virtuellen Forschungsumgebung TextGrid publiziert und für die weitere Erforschung nachnutzbar zur Verfügung gestellt. Darüber hinaus sollen auch im Projekt entwickelte Tools nachgenutzt werden können.

Um das Auftreten der zu untersuchenden vernetzten Verweisstrukturen in ihren unterschiedlichen Ausprägungen zu zeigen, aber auch das Vorgehen bei der Annotation und Analyse einzelner Begriffe und Konzepte zu erläutern, werden im geplanten Vortrag ‚Das Erhabene' und die ‚Metapher' als Beispiele hinzugezogen. Diese Begriffe eignen sich für Analysen von diskursiven Verweisungsstrukturen besonders gut, weil sie ihrer Herkunft nach aus der Ästhetik bzw. Rhetorik stammen und in der Poetik mit Rückverweis auf ihre Ursprünge und Urheber aufgeführt werden.

So geht das Konzept des Erhabenen als ambivalentes Gefühl der Überwältigung (bspw. bei der Betrachtung von Kunstwerken) zurück auf (Pseudo-) Longin, in dessen Nachfolge vor allem Kant und Burke den Begriff als ästhetische Kategorie definiert und umfassend diskutiert haben. Verweise auf diese beiden finden sich daher auch im Untersuchungscorpus wieder, allerdings mit entscheidenden Unterschieden. Denn das ästhetische Konzept des Erhabenen mit seinen zugehörigen Ersatz- und Ergänzungsbegriffen (Sublimes, Schreckliches usf.) erleidet im Untersuchungszeitraum einen Bedeutungsverlust, so wie die Ästhetik im Allgemeinen ihre poetologische Relevanz einbüßt, und wird entweder gar nicht mehr in seinem ursprünglichen Kontext behandelt, sondern nur noch im Zusammenhang mit erhabenem Stil innerhalb der Genera dicendi, oder erfährt ablehnende Beurteilung. Bei der Betrachtung des Begriffs und der

dazugehörigen Verweise muss insofern zwischen positiven und negativen Bewertungen unterschieden werden. Dies lässt sich durch das folgende Beispiel verdeutlichen: Bei Beyer (1882-84), Scherer (1888) und Wolff (1899) sind nicht nur zahlreiche Textstellen zu finden, in denen das Erhabene in seinem ursprünglichen Kontext behandelt wird, sondern auch jeweils Verweise auf Kant und/oder Burke als Urheber dieses Konzepts. Bei Beyer und Wolff findet eine produktive Auseinandersetzung mit dem Erhabenen statt, während Scherer den Begriff zwar in seiner ästhetischen wie poetologischen Auslegung nachverfolgt, ihn in seiner eigenen, empirisch ausgerichteten Poetik aber nicht reaktiviert, sondern rückblickend auf seinen Ursprung und seine Rezeption verwirft. D. h. Scherer nimmt zwar Bezug auf Kant, Burke und Autoren anderer literaturtheoretischer Werke, die über einen Diskurs des Erhabenen vernetzt sind, schließt sich diesem aber nicht an, sondern lehnt dessen Weiterführung ab.

Dieses erste Beispiel zeigt, dass die rein quantitative Identifikation, Annotation und Analyse relevanter Textstellen nicht ausreichen, um den Diskurs über das literaturtheoretische Konzept und dessen Entwicklung zu untersuchen. Eine komplexere Mehrebenen-Annotation ist erforderlich, die die Auszeichnung von Explikations- und Beschreibungskomponenten und Verweisstrukturen mit einer Bewertungsebene verbindet. Das für diese Anforderungen erstellte Annotationsschema umfasst daher konkrete Kategorien, die die Repräsentation des Begriffes und die Verweisungsstruktur im Text erfassen, und abstrakte Kategorien, die einerseits die Bewertungsebene abdecken, andererseits aber auch zur Überprüfung von Hypothesen dienen, die in einer vorhergehenden hermeneutischen Studie über das Corpus formuliert wurden (vgl. Sandra Richter: „A History of Poetics“). Im genannten Beispiel wäre die Hypothese, dass die Poetik im Ausgang aus dem 18. Jahrhundert noch expliziten Bezug auf die ästhetische Kategorie des Erhabenen nimmt und es unter Verweis auf Burke und/oder Kant diskutiert, während sich mit zunehmender Empirisierung der Poetik im Verlauf des 19. Jahrhunderts ein Bedeutungsverlust vollzieht und das Erhabene – wenn überhaupt – nur noch als stilistischer Aspekt thematisiert wird. Diesbezüglich wird ausgezeichnet, ob ein Bezug zur Ästhetik oder zur Stilistik besteht und inwiefern eine Bewertung erfolgt. So lässt sich nachvollziehen, dass eine derartige Entwicklung der Bedeutung des Konzepts Ergebnis eines wechselseitigen Diskurses ist.

Diese Diskursstruktur wird auch auf der Annotations-Ebene der konkreteren Repräsentation des Begriffs im Text erfasst. Dazu werden zusätzlich Verweisungen auf Personen bzw. Autoren und Werke ausgezeichnet (ebenfalls in Verbindung mit einer Bewertungsebene). Unterschieden wird dabei zwischen drei Textebenen: dem eigentlichen Poetikentext (Aussagen des Autors der jeweiligen Poetik), der Sekundärliteratur (Aussagen aus anderen literaturtheoretischen Texten, aber auch aus anderen Poetiken unseres Corpus) und der Primärliteratur (zur

Veranschaulichung herangezogene Beispiele aus literarischen Werken). Darüber hinaus wird bei gegebener Referenz auch zwischen den Verweisungsformen Zitat, expliziter und impliziter Paraphrase unterschieden. Vor allem letztere ist interessant, wenn sich nachweisen lässt, dass ein Autor einem anderen in seinen Ausführungen folgt, ohne dies anzugeben. Dadurch lassen sich über angegebene Verweisungsstrukturen hinweg auch unausgesprochene Übernahmen zurückverfolgen und ein diskursives Beziehungsgeflecht innerhalb des Corpus und darüber hinaus erfassen und sichtbar machen – auch auf der Ebene der literarischen Primärliteratur.

Dies lässt sich an einem zweiten Beispiel verdeutlichen. Bei der Definition der Metapher wird meist auf Aristoteles zurückgegriffen. Dieser versteht sie als „Übertragung“ zwischen einem eigentlichen und einem uneigentlichen Begriff und differenziert verschiedene Formen (vgl. Aristoteles: Poetik, Kap. 21). Diese grundlegende Definition lässt sich im Corpus „verfolgen“. Markant ist, dass einzelne Poetiken den Metaphern-Begriff als direkte Paraphrase von Aristoteles definieren und ihn gleichzeitig mit dessen teils literarischen Beispielen beschreibend darlegen. So finden sich in der Poetik von Borinski (1895) exakt dieselben Primärtext-Beispiele von Homer, die auch Aristoteles in seiner Poetik nennt. In einigen Poetiken lassen sich jedoch auch nur Ähnlichkeiten bei der Explikationsformulierung oder in der Unterscheidung verschiedener Metaphern-Formen bzw. Unterkategorien erkennen. Dies ist bspw. bei Beyer der Fall. Er definiert die Metapher als verkürzten Vergleich, bei dem der Vergleichspartikel wegfällt. Damit folgt er der Definition von Quintilian (neben Aristoteles die zweite grundlegende Begriffsbestimmung), ohne jedoch explizit die Quelle zu nennen. Darüber hinaus verweist er auf weitere Poetiken unseres Corpus, die den Begriff ebenfalls nach Quintilian bestimmen (Wackernagel 1873, Vischer 1846-57, Gottschall 1858). Das Beispiel zeigt, wie die Auslegung eines theoretischen Begriffs und dessen Entwicklung durch das Corpus hindurch mittels qualitativer Vergleiche und Analysen der (auch impliziten) Verweisstruktur nachweisbar ist.

Darüber hinaus wird die theoretische Definition eines griffigen Begriffs wie der Metapher häufig durch literarische Beispiele veranschaulicht, die sich für die Analyse auf der Ebene der Primärliteratur eignen. Zur Unterscheidung von Vergleich und Metapher verweist Beyer nicht allein auf Gottschall, sondern darüber hinaus auf ein bei diesem angeführtes Shakespeare-Zitat. An dieser Stelle verschränkt sich also die Analyse des Beziehungsgeflechts der Poetiken untereinander mit der Analyse der zitierten Primärliteratur. Literarische Beispiele werden in den Poetiken zur Veranschaulichung beschriebener Konzepte und für die (literatur-)kritische Stellungnahme im Hinblick auf die theoretischen Aspekte verwendet, sodass auch hier eine vergleichende Analyse möglich ist. Bestimmte Autoren und deren Werke tauchen in ähnlichen Zusammenhängen in einem großen Teil der Poetiken auf. Für die Metapher ist dies

Shakespeare. Clodius (1804), Gottschall und Dilthey (1887) nennen ihn übereinstimmend als einen der metaphernreichsten Dichter und damit als Vorbild für den richtigen Gebrauch von Metaphern. Ihm werden aber auch Negativbeispiele gegenübergestellt. Dies sind vor allem die antiken Autoren Sophokles und Aischylos, aber auch Goethe taucht in diesem unrühmlichen Zusammenhang immer wieder auf, was mit dem Unterschied von dessen epischem Stil zu Shakespeares dramatischem Stil begründet wird. Autorenbezogene Zuschreibungen wie diese lassen sich über das gesamte Corpus nachvollziehen und (auch diachron) vergleichend analysieren. Beispielsweise ist nach der jüngeren Poetik von Staiger nicht mehr Shakespeare der Prototyp des dramatischen Autors sondern Schiller, während Goethe diesem als Muster des lyrischen Dichters gegenübersteht. Durch solche Analysen ist es möglich, Prozesse der Kanonisierung und Ent-Kanonisierung einzelner Autoren und Werke nachzuvollziehen.

Sie helfen aber auch dabei, die Denkwelt einzelner Poetiken abzubilden. Bspw. erschien Wackernagels Werk zwar erst 1873 postum, es geht jedoch zurück auf eine akademische Vorlesungsreihe von 1836/7, was sich anhand der Auswahl der zitierten Primärliteratur einwandfrei nachvollziehen lässt. Ähnliches gilt für Autoren wie Staiger (1946) und Wehrli (1951), deren Schweizer Herkunft eine andere Textauswahl zumindest vermuten ließe. Während sich dies für Wehrli etwa im Hinblick auf Goethe bestätigen lässt, wird dieser für Staiger jedoch zum kanonischen Autor schlechthin.

Die computergestützte Auswertung all dieser Aspekte ermöglicht das Erkennen von Mustern und die Formulierung neuer Hypothesen bzw. Fragestellungen. Gleichzeitig bietet das Nebeneinander von abstrakter Interpretationsebene und konkreter Textebene Möglichkeiten des Abgleichs anhand der verschiedenen Kategorien. Das Annotationstool (UAM Corpus Tool) erlaubt eine Erweiterung des Schemas, sodass auch Aspekte, die im Verlauf der Untersuchung zu neuen Hypothesen führen, abgedeckt werden können. Auf der Basis der Auszeichnung der mit diesen Methoden selektierten Fundstellen nach dem beschriebenen Annotationsschema werden computergestützte Analysen und Visualisierungen durchgeführt. Auf diese Weise werden hermeneutische und algorithmische Verfahren im Sinne des ‚Algorithmic Criticism' verbunden. Dieses Vorgehen wird ausgeweitet auf weitere literaturtheoretische Kategorien auf verschiedenen Ebenen (z. B. Figur und Drama), sodass letztlich durch die kontextualisierende Aufbereitung, Vernetzung, Visualisierung und Analyse der enthaltenen Daten neue Erkenntnisse im Hinblick auf die Entwicklung der bedeutendsten literaturtheoretischen und -kritischen Diskurse und Konzepte in ihrem Zusammenhang ermöglicht werden.

Germania Sacra Online – Das Forschungsportal für kirchliche Personen und Institutionen bis 1810

Bärbel Kröger und Dr. Christian Popp, Germania Sacra

Wer im Netz nach wissenschaftlichen Informationen beispielsweise zu einem mittelalterlichen Kloster recherchiert, kann schon heute im Idealfall auf dem Forschungsportal der Germania Sacra ein ganzes Bündel von Informationen erhalten: Basisdaten zur Geschichte der Institution, kartographisch visualisierte Standortinformationen, Normdaten (GND, DBPedia, GeoNames), Links zu weiterführenden regionalen Online-Angeboten und bibliographische Informationen, Verknüpfungen zu dem in der Personendatenbank der Germania Sacra erfassten Klosterpersonal, die weitere fachübergreifende Verweise enthalten und damit den Weg zu neuen Erkenntnissen ermöglichen.

Die Germania Sacra hat in den vergangenen Jahren ein breites Portfolio digitaler Angebote zur Kirche des Alten Reiches erstellt. Hauptsäulen sind die digitalisierten Handbücher zur Geschichte kirchlicher Institutionen, die im Rahmen des Langzeitprojektes seit 1917 erarbeitet worden sind, ein umfangreiches digitales Personenregister zum kirchlichen Personal sowie die Datenbank zu Klöstern und Stiften des Alten Reiches. Alle digitalen Angebote der Germania Sacra sind work-in-progress: neue Bände werden nach 3 Jahren digital zur Verfügung gestellt, das Personenregister wird laufend erweitert (ca. 26.000 Einträge, Stand November 2014), die Klosterdatenbank befindet sich in der Aufbauphase (ca. 900 Einträge, Stand November 2014).

Die projektinterne Vernetzung der Daten wurde im Zuge der Integration in das Digitale Portal der Akademie der Wissenschaften zu Göttingen sichergestellt. Die hierfür verwendeten Technologien und Funktionalitäten werden im Rahmen des Vortrags präsentiert. Der Schwerpunkt des Vortrages wird jedoch auf der projektübergreifenden Vernetzung von Daten liegen. Ausführlich skizziert werden die hierfür bereits entwickelten Lösungsansätze sowie die zukunftsweisenden Kooperationen, die zu einer neuartigen digitalen Wissenslandschaft über die Kirche des Alten Reiches führen sollen.

Ein wichtiger Baustein für die Vernetzung ist die systematische Anreicherung der Datenbestände mit Normdaten. Für viele der durch die Forschung der Germania Sacra generierten Informationen kann auf bereits vorhandene Normdaten zurückgegriffen werden. Besonders relevant für unser Projekt ist der Datenbestand der Deutschen Nationalbibliothek mit den dort verwendeten Datensatznummern der Gemeinsamen Normdatei (GND). Für Personendaten wird üblicherweise das Beacon-Format verwendet, das das automatische Generieren von Links zu externen Datenquellen ermöglicht. Für andere Daten als Personen, etwa für Körperschaften, wird das Beacon-Format bisher kaum genutzt. Mit der Klosterdatenbank der Germania Sacra soll die Verwendung dieser Technik für Klöster und Stifte erprobt und eingeführt werden. Die Identifizierung der einzelnen Klöster und Stifte in der Gemeinsamen Normdatei der Deutschen Nationalbibliothek ist bereits vielfach erfolgt, fehlende Einträge in der GND werden durch die Germania Sacra ergänzt. So können in der Datenbank automatisiert direkte Links nicht nur zu externen Datenbanken, sondern auch zu relevanten Datensätzen in Bibliothekskatalogen, Bestandsübersichten von Archiven, Quelleneditionen, Bibliographien, Porträtsammlungen und weiteren Informationsangeboten bereitgehalten werden.

Um den Möglichkeiten zur semantischen Recherche einen Weg zu bereiten, werden die Inhalte der Datenbanken auf der Basis von Linked Data angereichert und im RDF-Format ausgegeben, dabei wird auf etablierte existierende Vokabulare zurückgegriffen. Für die Ausgabe der Datensätze im RDF-Format werden Normdaten für Orden, Bistümer, Personen wie auch Normdaten für Geografika (GeoNames) verwendet. Vorhandene Einträge in der Wikipedia werden referenziert. Das modellierte Schema bietet hohes Potential, das Informationsnetz zu den Beziehungen von Personen und geistlichen Institutionen für den Zeitraum des Mittelalters und der Frühen Neuzeit zu verdichten.

Während bei der Verknüpfung von Daten zu kirchlichen Institutionen aufgrund eindeutiger Identifikatoren (z.B. Name, Orden, Standortinformation) vergleichsweise gut automatisierte Lösungsansätze zu finden sind, ist die automatisierte Vernetzung von personengeschichtlichen Datenbanken aus dem Bereich der Mittelalter- und Frühneuzeitforschung nach wie vor ein ungelöstes Problem im Bereich der Digital Humanities. Die Zuordnung von Informationen aus unterschiedlichen Datenbanken zu einer bestimmten Person ist schwierig. Häufig liegen nicht genügend Daten vor, die eine sichere Identifizierung ermöglichen (Geburts- und Sterbedatum, Herkunftsort, Ämter und Amtsdaten). Erschwerend kommen die zum Teil erheblich abweichenden Namensvarianten, Übersetzungs- und Transkriptionsfehler, latinisierte Formen und der spät einsetzende Gebrauch von Zweitnamen hinzu.

Daher entwickelt die Germania Sacra in Kooperation mit dem Deutschen Historischen Institut in Rom (DHI) und dem Repertorium Academicum Germanicum (RAG) eine projektübergreifende Personenrecherche. Dabei gilt es geeignete technische Lösungen zu finden. Hier können beispielsweise Algorithmen, die phonetische und orthographische Varianten auffindbar machen, oder die Verwendung von Thesauri zur Erkennung latinisierter Namensformen hilfreich sein. Diese Metasuche soll nicht nur als ein datenbankübergreifendes Recherchetool fungieren, sondern zugleich unter Verwendung von Technologien des Crowdsourcing interaktive Verknüpfungsmöglichkeiten für wissenschaftliche Nutzer bieten, die so ihre Identifikationsvorschläge einzelner Personendatensätze in die Datenbanken zurückmelden können.

Gerade die Verknüpfung von Personendaten aus unterschiedlichen Forschungsprojekten und unterschiedlichen Quellenbeständen lässt eine Generierung neuen Wissens erwarten, die in dieser Form nur durch den Einsatz digitaler Werkzeuge möglich ist.

Die nachhaltige Nutzung der Forschungsdaten wird durch die Integration von Germania Sacra Online in das Digitale Portal der Akademie der Wissenschaften zu Göttingen gewährleistet, die die hierfür erforderliche Infrastruktur zur Verfügung stellt.

Semantische Anreicherung von Bildnetzwerken mit HyperImage und Yenda – Das Hachiman Digital Handscrolls Projekt

Dr. Jens-Martin Loebel, Dipl.-Inf. Heinz-Günter Kuper

Abstract für ein Poster auf der DHd 2015 – Von Daten zu Erkenntnissen
23.-27. Februar 2015, Graz

Das Poster wird die Erweiterungen der virtuellen Forschungs- und Publikationsumgebung *HyperImage*[1] sowie die semantische Annotationsplattform *[Yenda]*[2] im Rahmen des Hachiman-Projekts vorstellen.

Das wesentliche Ziel des *Hachiman Digital Handscolls* Projektes ist es, monumentale oder bewegliche Bildformate einer Forschungsgemeinschaft digital so vorzustellen, dass damit disziplinäre, sprachliche und regionalspezifische Grenzen aufgehoben werden. Inhaltlicher Gegenstand dieses Pilotprojekts ist ein Konvolut von sieben illuminierten japanischen Querrollen des 14. – 17. Jahrhunderts, die in leicht variierenden Versionen die Hachiman-Legende wiedergeben.

Die Materialität der Querrollen setzt einen physischen Kontakt und direkte Interaktion voraus: Sie müssen beim Ansehen mit beiden Händen entrollt werden und der genaue Bildabschnitt kann selbst bestimmt werden. Dieser Umstand und ihre Maße von bis zu 18 Metern pro Rolle müssen im Rahmen einer Untersuchung und Präsentation in einer Printpublikation oder statischen Datenbank zwangsläufig zu einem unbefriedigenden Ergebnis führen. Eine statische, auf Text- oder Bildabschnitte fixierte digitale Darstellung dieser Artefakte ist problematisch und in diesem Zusammenhang wenig zielführend.

Möglicherweise ist darin einer der Gründe zu suchen, weshalb bisher keine vergleichbare wissenschaftlich-digitale Aufbereitung solcher Querrollen mit gleichem Sujet erfolgt ist. Das Hachiman Digital Handscolls Projekt möchte durch den Einsatz und gezielten Ausbau der virtuellen Forschungs- und Publikationsumgebung *HyperImage* und dessen Nachfolger Yenda zur semantischen Annotation genau diese Lücke schließen.

Durch Transkriptionen, Übersetzungen und visuelle sowie textuelle Annotationen, die gleichzeitig mit den dazugehörigen Textpassagen angezeigt werden können, wird den Betrachtern die inhaltliche Bedeutung vermittelt und der Text somit entmystifiziert und einem breiteren Publikum zugänglich gemacht.

HyperImage stellt dabei als technologische Basis Mittel und Werkzeuge bereit, die Rollen digital zu erschließen, zu annotieren, und – mittels des neuen Werkzeugs Yenda – semantische Verbindungen zwischen einzelnen Rollenabschnitten und Bild- und Textdetails zu visualisieren und diese mit Normdaten und Link-Open-Data-Beständen zu verknüpfen.

---

[1] siehe Website des Open-Source-Projekts unter http://hyperimage.ws/

[2] Die semantische Annotationsplattform [Yenda] wird ab Frühjahr 2015 als Open-Source zur Verfügung stehen und u. a. HyperImage als Web-basierte Arbeitsumgebung integrieren. Weitere Informationen finden sich unter https://yenda.tools/

HyperImage ist als Werkzeug zur Unterstützung des Bilddiskurses in den Digitalen Geisteswissenschaften seit vielen Jahren etabliert und wird in Forschung und Lehre an Forschungseinrichtungen in Deutschland und Europa eingesetzt. Mit HyperImage können beliebig viele Details innerhalb eines Bildes präzise markiert und beschrieben sowie Annotationen des Corpus untereinander verlinkt und über Indizes erschlossen werden können. Einfach gesagt: Was Hypertext für Text ist, ist HyperImage für Bilder (siehe Abb. 1 und Abb. 2).

Zwischenergebnisse wie endgültige Fassungen können jederzeit als hypermediale online- oder offline-Publikation erstellt werden. Verschiedene einzeln eingeführte und erprobte Verfahren und Datenrepositorien (u. a. das *prometheus Bildarchiv*)[3] sind in HyperImage komfortabel zu einer einzeln oder kollaborativ nutzbaren Forschungs- und Publikationsumgebung zusammengeführt.

Ziel ist es, die digitale Darstellung von beweglichen Bild-und-Text-Formaten zu verbessern. Das innovative Open-Source-System HyperImage dient als zentrales Werkzeug, um sieben der japanischen Querrollen im Web zu präsentieren und somit historische Artefakte für zeitgenössische Sehgewohnheiten zu vermitteln und den Zugang zu historischem, literarischem und visuellem Wissen zu erleichtern.

Die Studie ist ein Kooperationsprojekt zwischen dem Institut für Kunstgeschichte Ostasiens, der Heidelberg Research Architecture (HRA) des Exzellenzclusters Asien und Europa, dem SFB 933 „Materiale Textkulturen“ und der Firma bitGilde IT Solutions UG.[4] Die Digitalisate der Schriftrollen wurden von einer Reihe von Schreinen und Museen in Japan, den USA und Deutschland bereitgestellt (u. a. Hakozaki-gu, Kotozaki Hachiman-gu, Umi-Mori Art Museum, Yura Minato Jinja, Asian Art Museum San Francisco und Staatsbibliothek Berlin).

Die Berliner Firma bitGilde IT Solutions UG (eine Ausgründung der beiden Hauptentwickler von HyperImage) übernimmt die universitäts- und institutsübergreifende Koordinierung und nachhaltige Weiterentwicklung des Systems als Open Source. Mit der Ausgründung wird ein innovatives Konzept zur Verstetigung und Langzeitsicherung der Forschungsergebnisse aus Drittmittelprojekten verfolgt.

Durch das Projekt wird es erstmals möglich, diese Querrollen in digitaler Form einer breiten Öffentlichkeit über das Internet zugänglich und erfahrbar zu machen. Die Online-Veröffentlichung ist für Anfang 2015 geplant. Die Projektwebsite[5] befindet sich derzeit im Aufbau.

Das Poster wird die zentralen Funktionen und Erweiterungen von HyperImage im Zusammenspiel mit der Annotationsplattform Yenda anhand der Ergebnisse des Hachiman-Projektes vorstellen.

---

[3] http://prometheus-bildarchiv.de

[4] siehe http://bitgilde.de

[5] http://www.zo.uni-heidelberg.de/iko/hdh/

Abbildungen

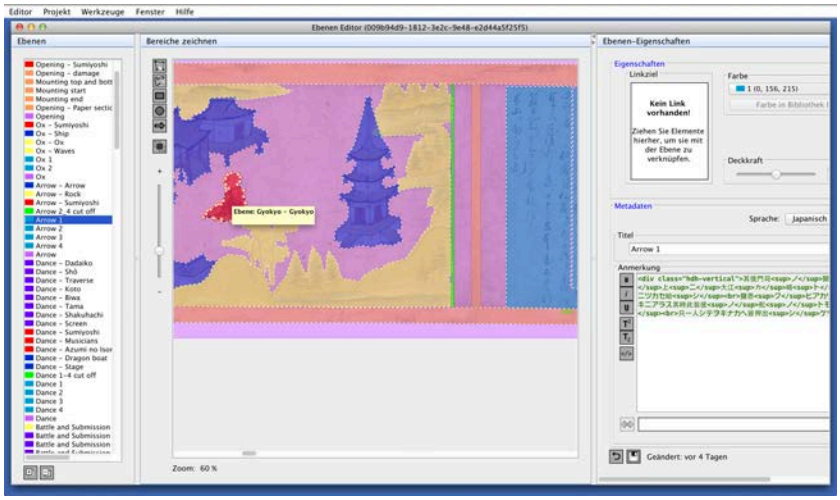


Abb. 1: *Visuelle Annotation und Transliteration der Rollen mit HyperImage.*

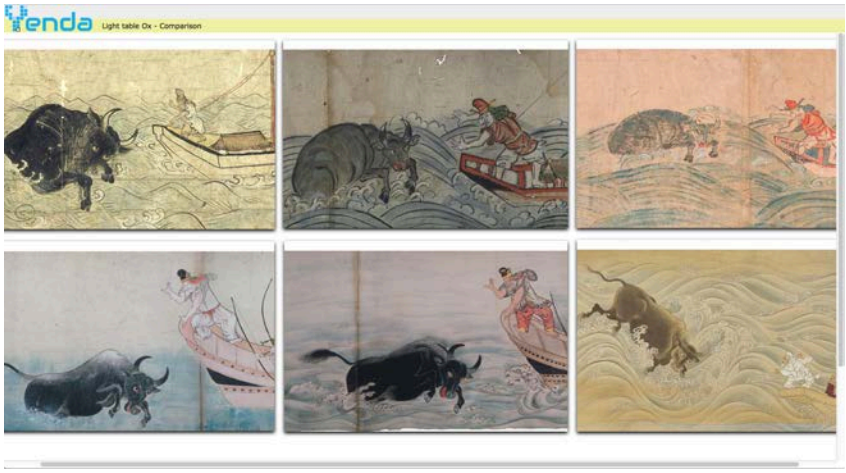


Abb. 2: *Living Scrolls-Technologie von Yenda. Gleichzeitige Visualisierung ein und desselben Motivs von sechs verschiedenen Rollen. Jede Rolle ist einzeln in der Webpublikation im Browser navigierbar.*

Weiterführende Literatur

Loebel, J.-M., Kuper, H.-G.; et al.: Hachiman Digital Handscrolls – Semantische Anreicherung mit HyperImage und Yenda. In: Bienert, A.; Hemsley, J.; Santos, P. (Hrsg.): *EVA Berlin 2014 – Elektronische Medien & Kunst, Kultur und Historie*. Berlin: Konferenzband, ISBN: 978-3-88609-755-5, 2014, S. 262-267.

Kuper, H.-G.; Loebel, J.-M.: HyperImage: Of Layers, Labels and Links. In: *Proceedings of RENEW – the 5th edition of the International Conference on the Histories of Media Art, Science and Technology*, Riga, 2014.

Humanities Data Centre – grundlegende Überlegungen in der Designphase eines geisteswissenschaftlichen Forschungsdatenzentrums

Stefan Buddenbohm[1], Claudia Engelhardt[2], Ulrike Wuttke[3]
[1]Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften, [2]Staats- und Universitätsbibliothek Göttingen, [3]Akademie der Wissenschaften zu Göttingen
buddenbohm@mmg.mpg.de, claudia.engelhardt@sub.uni-goettingen.de, uwuttke@gwdg.de

Schlagwörter: Forschungsdatenzentrum, Langzeitarchivierung, Forschungsdatenmanagement

Zusammenfassung: Forschungsdaten sind sowohl Ergebnis von Forschung als auch Grundlage für neue Forschungsfragen. Die zunehmende Nutzung digitaler Ressourcen und Methoden in der Forschung widerspiegelt sich sowohl im wachsenden Umfang als auch in der zunehmenden Komplexität von digitalen Forschungsdaten, sowohl in den Geisteswissenschaften wie auch in anderen Disziplinen. Aus verschiedenen Gründen ist die Erhaltung dieser Forschungsdaten notwendig: Dokumentationszwecke beispielsweise für Förderer oder aufgrund rechtlicher Bestimmungen, Nachvollziehbarkeit und Reproduzierbarkeit von Forschungsergebnissen, aber vor allem auch die Möglichkeit der Nachnutzung für neue Forschungsvorhaben. Die Herausforderungen hinsichtlich des Forschungsdatenmanagement und der Langzeitarchivierung können jedoch nur mit einem umfassenden Verständnis ihrer Entstehungs- und Nutzungsbedingungen gemeistert werden. Da diese von Infrastrukturanbietern nur im engen Austausch mit den Fachdisziplinen eruiert werden können, scheinen disziplinspezifische Forschungsdatenzentren am besten geeignet, die damit verbundenen Aufgaben zu übernehmen.
Während der Designphase des Humanities Data Centres (HDC, 2014-2016) werden daher im Dialog mit der Wissenschaft und Infrastruktureinrichtungen die Grundlagen für den Aufbau eines Forschungsdatenzentrums für die Geisteswissenschaften geschaffen. Das Projektkonsortium besteht neben geisteswissenschaftlichen Forschungseinrichtungen aus Rechenzentren und einer Universitätsbibliothek.
Grundsätzlich lässt sich die Langzeitarchivierung von Forschungsdaten entlang von drei aufeinander aufbauenden Ebenen strukturieren:

- Bitstream Preservation: Der physische Erhalt des gespeicherten Datenobjekts (Bitstream) auf einem entsprechenden Speichermedium,
- Technische Nachnutzbarkeit: Sicherstellung der Zugänglichkeit der Forschungsdaten auch bei veränderten technischen Bedingungen,
- Intellektuelle Nachnutzbarkeit: Sicherstellung der vollständigen Nutzbarkeit und Interpretierbarkeit des intellektuellen Gehalts der Forschungsdaten, beispielsweise durch Metadaten und die Dokumentation von Kontextinformationen, die das ursprüngliche Forschungsszenario nachvollziehbar machen.

Darüber hinaus hängt die Nachhaltigkeit von Forschungsdaten stark von einem stabilen, organisatorischen Rahmen ab, innerhalb dessen die entsprechenden Umgebungen und Werkzeuge bereitgestellt werden können. Nicht zuletzt ist aber der beständige Austausch mit den wissenschaftlichen Nutzern von großer Bedeutung, um mit dem Angebot (dem Forschungsdatenzentrum) den Anforderungen der Wissenschaft zu entsprechen beziehungsweise dieses neuen Entwicklungen und Bedürfnissen anzupassen.
Vor diesem Hintergrund stellen sich bei der Konzeption eines geisteswissenschaftlichen Forschungsdatenzentrums, das sowohl die Langzeitarchivierung als auch die Bereitstellung der Forschungsdaten für die Nachnutzung sicherstellen soll, verschiedene Fragen:

- Was sind Forschungsdaten in den Geisteswissenschaften und welche Forschungsdatentypen sollen vom Angebot des Datenzentrums berücksichtigt werden? Wie können geeignete Objektmodelle für die Bereitstellung und Archivierung dieser Forschungsdaten aussehen? Wie kann mit Forschungsdaten umgegangen werden, die nicht dokumentenbasiert sind, sondern beispielsweise aus Datenbanken bestehen?

- Wie kann die Zusammenarbeit zwischen Wissenschaft und Forschungsdatenzentrum erfolgreich sein? Welche Angebote hinsichtlich Beratung und Schulung sind besonders geeignet, um der Bedeutung des Forschungsdatenmanagements gerecht zu werden? Welche Implikationen hat das für mögliche Organisationsformen (-einheiten) eines Forschungsdatenzentrums?
- Die langfristige Nachhaltigkeit und Nachnutzbarkeit von Forschungsdaten ist nicht nur ein technisches, sondern vor allem ein organisatorisches Thema. Bestimmte geisteswissenschaftliche Forschungsdaten (zum Beispiel Editionen, Korpora, Wörterbücher) behalten über einen längeren Zeitraum ihre Forschungsrelevanz. Wie lässt sich diese Anforderung in organisatorischer und infrastruktureller Hinsicht umsetzen?
- Welche bestehenden und zukünftigen Standards für ein Forschungsdatenzentrum sind zu beachten, um Interoperabilität und Kooperation zwischen Forschungsdatenzentren zu fördern? Wie kann dies in Einklang mit der Anforderung der Skalierbarkeit gebracht werden?
- Ein Forschungsdatenzentrum muss über einen längeren Zeitraum lernen und sein Angebot anpassen. Gleichzeitig ist aber die Stabilität der konkreten (technischen) Angebote wichtig: für die technische Infrastruktur zum stabilen Aufbau der technischen Dienste; für Nutzer des Forschungsdatenzentrums um bereits zu Projektbeginn die Angebote in ihr Datenmanagement einplanen zu können und auch zu Projektende noch darauf vertrauen zu können. Stabilität und Innovation sind dabei zwar keine Gegensätze, müssen aber gegeneinander abgewogen werden. Wie leistet ein Forschungsdatenzentrum diesen Ausgleich zwischen Erneuerung und Stabilität des Angebotes?

Ein Wizard für die Erschließung strukturierter Textdaten

Fritz Kliche[1], **Nicolas Schmidt**[2], **Ulrich Heid**[1]

[1]Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
[2]Institut für Betriebswirtschaft und Wirtschaftsinformatik, Universität Hildesheim
{kliche,schmi032,heid}@uni-hildesheim.de

Wir stellen einen *Wizard* vor, mit dem strukturierte Textdaten in einer Browser-Anwendung erschlossen werden können, um die textlichen Inhalte und Metadaten für textwissenschaftliche Analysen nutzen zu können. Der *Wizard* ist ein interaktives Werkzeug, das es dem Benutzer erlaubt, von Beispielfällen auf größere Mengen von Daten zu generalisieren, ohne dass er dazu zu programmieren braucht.

Die Voraussetzung sind Textdaten, deren Textstruktur nicht für jeden Text unterschiedlich ist, sondern wo sich ähnliche Textstrukturen über größere Mengen von Einzeltexten hinweg beobachten lassen. Beispiele sind Sammlungen von Zeitungsartikeln, Sammlungen von Blogs und user-generated content oder Protokolle von Parlamentsdebatten. Solche Texte sind einerseits nicht standardisiert, andererseits doch relativ homogen repräsentiert, mindestens innerhalb jeder Kollektion, jedes Zeitungs- oder Blog-Archivs. Die extrahierten Inhalte werden als *Textobjekte* in einer Datenbank abgelegt und mit Labels versehen, die den Zugriff ermöglichen. Über diesen Zugriff können die Textobjekte in eine neue Datenstruktur überführt werden. Der *Wizard* entsteht innerhalb des DH-Projekts *e-Identity* (Blessing et al., 2013), in dem ein umfangreiches Sample von Zeitungsartikeln (>800.000 Artikel) aus 5 digitalen Medienportalen erschlossen und ein Korpus erstellt wurde, in dem die textlichen Inhalte der Artikel und die begleitenden Metadaten kategorisiert vorliegen.

Der *Wizard* führt den Anwender durch die Funktionen, die über eine Browser-GUI gesteuert werden. Zunächst können Textdaten in unterschiedlichen Formaten (RTF, DOCX, ODT, TXT, HTML) und Zeichencodierungen (UTF-8, ISO-8859-1) importiert werden. Die anschließende Erschließung erfolgt in zwei Schritten: (1) Die Textdaten werden zunächst in strukturelle Einheiten (z. B. in *e-Identity*: Zeitungsartikel) segmentiert; (2) in diesen Einheiten werden anschließend textliche Inhalte und Metadaten erkannt und klassifiziert, indem ihre Anfangs- und Endpunkte im fortlaufenden Textmaterial identifiziert werden. Dafür werden in einem Vorschau-Fenster Ausschnitte der importierten Textdaten angezeigt. Der Anwender erstellt anhand solcher Beispiele *Extraktionsregeln*, d. h. Muster, nach denen Textobjekte identifiziert werden. Mit der Erstellung mehrerer Extraktionsregeln entsteht ein Regelset als eine Schablone, mit der in den importierten Daten Inhalte erkannt und extrahiert werden. Für die Extraktionsregeln wurden Elemente einer Regelsprache implementiert, über die der *Wizard* computerlinguistische Konzepte (reguläre Ausdrücke, Text Mining, computerlinguistische Prozessierung) textwissenschaftlichen Anwendern möglichst intuitiv zugänglich macht. Die im Folgenden dargestellten Konzepte wurden umgesetzt:

Integrierte computerlinguistische Werkzeuge
Der *Wizard* integriert computerlinguistische Verarbeitungsschritte zur Tokenisierung, Lemmatisierung, Wortartenerkennung und zur Erkennung von Eigennamen. Weiter werden Tokens verschiedenen „Tokentypes“ zugeordnet (Versalien, groß- oder kleingeschriebene Wörter, Zahlwörter usw.). Die entsprechende computerlinguistische Verarbeitung soll einerseits im Hintergrund stattfinden; andererseits soll sie von den Anwendern gesteuert werden können. Wir trennen dazu die Erstellung der Extraktionsregeln von ihrer Anwendung. Nach der Erstellung einer Schablone validiert der *Wizard* deren Regeln und prüft, welche computerlinguistischen Vorverarbeitungsschritte sie verlangen. Der Anwender muss also nicht entscheiden, welches computerlinguistische Werkzeug an welcher Stelle zum Einsatz kommen soll, sondern welches Prozessierungsergebnis angestrebt wird. Wo nötig, wird dazu ein computerlinguistischer Verarbeitungsschritt vom Werkzeug vorgeschlagen und eingeschoben.

Unterschiedliche Textobjekte
In den Daten können unterschiedliche Textobjekte definiert werden. Als Textobjekte sind Tokens, Segmente (d. h. einzelne Zeilen) und mehrzeilige Objekte möglich.

Merkmale zur Identifikation
Zur Erstellung der Extraktionsregeln können unterschiedliche textliche Merkmale berücksichtigt werden. Als Indikator eines Textobjekts können (1) ein Ankerwort oder (2) ein regulärer Ausdruck definiert werden; (3) die maximale und die minimale Länge eines Segments können festgelegt werden; (4) um das Segment im Kontext zu definieren, können Ankerwörter und reguläre Ausdrücke zum Vorgänger- oder Nachfolgersegment des zu bestimmenden Segments definiert werden. (5) Schließlich kann die Abfolge unterschiedlicher Typen von Tokens definiert werden, die über die Wortart, einen Abgleich mit Terminologielisten (z. B. eine Liste von Monatsnamen), „Tokentypes“ oder über eine feste Zeichenkette charakterisiert werden.

Funktionen der Extraktionsregeln
Die Extraktionsregeln dienen zunächst der Identifikation von Textobjekten; sie können auch als Blocker fungieren, die die Identifizierung von Objekten durch andere Regeln verhindern.

Der Aufbau einer eigenen Datenstruktur
Die erkannten Objekte werden in einer Datenbank mit ihrem Label abgelegt. Durch das Label kann auf die Daten zugegriffen werden. Damit sind die Daten für den Aufbau einer neuen Datenstruktur zugänglich. Die Daten können in ein XML-Format konvertiert werden. Wir planen die Möglichkeit zur Konvertierung in gängige Formate: TEI, CMDI (Broeder et al., 2012), etc.

Anwendung für Textwissenschaftler in DH-Projekten
Das Poster richtet sich besonders an textwissenschaftliche Anwender. Screenshots stellen die Arbeitsschritte zur Erschließung strukturierter Textdaten dar. Aus computerlinguistischer Sicht zeigt das Poster ein Beispiel, wie linguistische Annotationen in ein Softwareprojekt für die Digital Humanities eingebunden werden. In einer Demonstration können die Benutzerschnittstellen des Systems vorgeführt werden.

Literatur

Blessing, André; Sonntag, Jonathan; Kliche, Fritz; Heid, Ulrich; Kuhn, Jonas; Stede, Manfred (2013). Towards a tool for interactive concept building for large scale analysis in the humanities. In: *Proceedings des 7. Workshops „Language Technology for Cultural Heritage, Social Sciences, and Humanities“*. Association for Computational Linguistics, Sofia, Bulgarien.

Broeder, Daan; Windhouwer, Menzo; van Uytvanck, Dieter; Goosen, Twan; Trippel, Thorsten (2012). CMDI: a component metadata infrastructure. In: *Proceedings des Workshops „Describing Language Resources with Metadata“*. LREC 2012, Istanbul, Türkei.

Database of historical places, persons and lemmas

Natalia Korchagina [1, 2]

[1] *Schweizerische Rechtsquellenstiftung / Zürich, Switzerland*
[2] *Institut für Computerlinguistik / Universität Zürich, Switzerland*

Proposed session - **Poster session**

Keywords - **digital humanities, database, RDF triple store, multilinguality, NLP for historical texts**

The importance of the representation of humanities material as structured, interconnected objects has grown with the recent emergence of the ideas of Linked Data and Semantic Web. Efficient cataloging and storing of humanities data can facilitate the research and knowledge exchange within the field. In this respect, the use of modern technologies of data storage, i.e. databases, is a crucial point for a digital humanities project.

The Swiss Law Sources Foundation has been handling the critical edition and publishing of Swiss historical legal manuscripts for over a hundred years. By today, over 100 volumes of texts have been published, about 30 of them are available as digital editions. This collection contains texts in German, French, Italian, Romansh and Latin languages. The texts' creation time ranges from the 10th to the 18th centuries representing, for this reason, a rich source not only of historical, but also of linguistic information on language evolution. Each of these volumes contains a back-of-the-book index of persons, places and lemmas mentioned.

The creation of the database should facilitate the edition of the index of future volumes, as well as to be a starting point for users looking for information on a specific personality/place which could have been mentioned in several Foundation's volumes. The database will have the four CRUD (create, read, update, delete) basic functions of persistent storage, and will provide its users with an intuitive GUI. This database is intended for use in two directions: first, for the edition of indexes of the upcoming volumes where editors will update/create database entries via GUI instead of working with Excel files; and second, for large public browsing the database via GUI in read-only mode. To give some more details on a historical person/place the database entries will be linked (when possible) with the corresponding GND (Integrated authority file of the German National Library) and HLS (Historische Lexikon der Schweiz) entries.

The technology to be used is RDF triple store. This is a NoSQL (non-relational) mechanism for storing and retrieval of data. In a triple store each data entity is composed of subject-predicated-object (triple), like "John knows Mary". An RDF triple store can be viewed as a graph, where an object of one entity is a subject of another, and so on. Graph data representation is particularly pertinent for Digital Humanities where the data is highly interconnected. On practice this kind of data representation guaranties fast path-walking for complex queries enabling knowledge discovery. Furthermore, an RDF triple store has a simple and uniform standard data model, and is governed by a powerful standard query language SPARQL. An RDF triple store also provides a standardized interchange format (e.g. N-triples) for import/export which is important for data transfer/exchange. Thus, RDF triple store is a mature, stable technology convenient for persistent data storage. Moreover, RDF is a standard model for data interchange over the emerging Semantic Web and Linked Open Data cloud. As a future goal, we aim to integrate our RDF data into other international projects (e.g. DBpedia, Europeana) for a higher visibility. Another future direction would be participation in such "meta"-projects as, for example, Bibliographie-Portal (http://www.biographie-portal.eu).

Mit den Informationswissenschaften von Daten zu Erkenntnissen

Sandra Balck, Prof. Dr. Stephan Büttner, Denise Ducks, Ann-Sophie Lehfeld, Eva Schneider, Evelyn Vietze

Fachhochschule Potsdam, Fachbereich Informationswissenschaften

1. Transformationsprozess Wissen- Information

Der Transformationsprozess von Wissen zu Informationen ist ein originär informationswissenschaftliches Problem. In der informationswissenschaftlichen Betrachtung ist Wissen der Ausgangspunkt von Daten und Informationen. Informationen gehen demnach nicht, wie in der klassischen DIKW-Pyramide (Data-Information-Knowledge-Wisdom) angenommen, aus Daten hervor sondern werden durch einen doppelten Transformationsprozess aus Wissen generiert. Anstelle eines hierarchischen Modells wird eine funktionale Unterscheidung zwischen formal-syntaktischen, semantischen und pragmatischen Ebenen von Information vertreten[1]. (*Dieser Transformationsprozess wird im Poster durch eine Grafik visualisiert.)*

2. Beitrag der Informationswissenschaften

Die Informationswissenschaften (IW) verfügen über Methoden welche es ermöglichen vorhandenes Wissen aus Informationsbeständen zu extrahieren. Auch in Digital Humanities (DH)-Projekten werden neue Daten u.a. mit Hilfe der Methoden der Informationswissenschaften generiert und neue Erkenntnisse gewonnen. Dies betrifft alle die Informationswissenschaften tangierenden Disziplinen (Linguistik, Informatik u.a.). Vorhandenes Wissen ist auf Grund interdisziplinärer Zusammenarbeit nicht mehr klar voneinander zu trennen und sollte es auch nicht sein. Dies erfordert eine Optimierung des Transformationsprozesses. Die Methoden der Informationswissenschaften bieten dazu ein geeignetes Toolset.

IW ist, nicht nur, wie z.B. im Drei-Sphären-Modell von DARIAH[2] angenommen, ein geisteswissenschaftliches Einzelfach, sondern bietet eine gemeinsame, disziplinübergreifende theoretische Grundlage für DH. Bereits Roberto Busa wies im Companion to Digital Humanities[3] explizit darauf hin, dass der größte der drei Stränge der DH als "documentaristic" or "documentary", zu bezeichnen sei.

Bei den Bestrebungen für ein Referenzcurriculum im Rahmen der DARIAH-Initiative wurde die informationswissenschaftliche Ausbildung bisher kaum oder gar nicht wahrgenommen.

---

[1] vgl. Kuhlen (2013)
[2] vgl. ARIAH Working Papers (2013) Schreibmann, et al (2004)
[3] vgl. Schreibmann, et al (2004)

Ein Vergleich der zentralen Themen der DH mit denen der IW zeigen jedoch große Übereinstimmungen:

Suchverfahren
Text Mining und Sprachverarbeitung
(Forschungs-)Datenmanagement
Fachspezifische Datenbanken
Fachinformation
Geographische Informationssysteme
digitale Bildverarbeitung
User studies
Hermeneutik (third current)
Digitale Edition und
Langzeitarchivierung

3. Module Studiengang Informationswissenschaften der Fachhochschule Potsdam

In der informationswissenschaftlichen Ausbildung der Fachhochschule Potsdam spielen die Kernkompetenzen der DH eine wesentliche Rolle (siehe Tabelle).

Module	**Credit Points**
Erschließung	15-25
Datenbanken	5-25
Information Retrieval	5-20
Digitale Editionen	7
Dokument	10
Informationsvisualisierung	6
Modellierung / XML	10
Linguistik / Textmining	5
Datenmining / Semantic Retrieval	14
Wissenschaftsmethodik	5

Die Ausprägung der einzelnen Module unterscheidet sich hierbei innerhalb der drei angebotenen Studiengänge Archiv, Bibliothekswissenschaft sowie Information und Datenmanagement.[4]

4. DH als Anwendung informationswissenschaftlicher Methoden

Der drei-semestrige Masterstudiengang Informationswissenschaften bietet eine fachwissenschaftliche Weiterführung informationswissenschaftlicher Grundlagen mit zwei vertiefenden Profilierungsmöglichkeiten und baut auf einem informationswissenschaftlichen Bachelorstudium auf. Inhaltlich ist eine

[4] Vgl. Studien- und Prüfungsordnung (2014)

Spezialisierung durch die Wahl einer von zwei Profillinien im zweiten Semester möglich.

Profil 1: Records Management und Digitale Archivierung
Profil2: Wissenstransfer und Projektkoordination

Die Profillinie "Wissenstransfer und Projektkoordination" vermittelt dabei Kompetenzen, die sich mit den Kerninhalten der DH überschneiden.

Im Track Wissenstransfer ist es demzufolge notwendig und folgerichtig DH als Anwendung informationswissenschaftlicher Methoden zu integrieren. Die Integration der Digital Humanities soll dabei nicht durch die Unterbringung neuer Inhalte, sondern die namentliche Verankerung (Implizites explizit machen) und somit Sichtbarmachung der bereits als IW-Methoden im Curriculum verankerten Kompetenzen erfolgen. IW fungiert dabei als Mittler zwischen D (Informatik) und H (Geisteswissenschaften).

Im Bestreben eine gemeinsame Lobby für den geisteswissenschaftlichen Transfer zu etablieren, wird keine Fusion sondern die Kooperation beider Disziplinen angestrebt. Diese könnte sich unter anderem in einer gemeinschaftlichen Ausbildung niederschlagen.

Literatur

DARIAH-DE Working Papers 2013-1
Sahle, P.: Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities
nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2013-1-5

Kuhlen, R.: A1 Information – Informationswissenschaft
in: Kuhlen, R.; Semar, W.; Strauch, D. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. 6. Ausgabe. Berlin 2013: Walter de Gruyter

Schreibmann, S.; Siemens, E.; Unsworth, J. (Ed): A Companion to Digital Humanities Oxford, Blackwell (2004)

Studien- und Prüfungsordnung für die Bachelorstudiengänge
Archiv, Bibliothekswissenschaft, Information- und Datenmanagement des Fachbereichs Informationswissenschaften der Fachhochschule Potsdam - Besondere Bestimmungen (2014 intern)

Carolin Hahn ▪ ✉ Wollankstaße 64A, 13359 Berlin ▪ ☎ 0152 2251 6963 ▪ 💻 carolin.hahn@ymail.com Abstract [Poster]

Poster-Bewerbung. DHd 2015, Graz – „Von Daten zu Erkenntnissen"

Themenbereich: (a) Mehrwert digitaler Methoden und Technologien für Erkenntnisprozesse

Arbeitstitel: Auch ich in Rom! Die literarische Inszenierung sozialer Netzwerke und Wissenstransfers in deutschsprachigen Reiseberichten (1816-1833)

Fachbereich, Universität: Neuere Deutsche Literaturwissenschaft, Humboldt-Universität zu Berlin

Betreuer: Prof. Dr. Steffen Martus, Prof. Dr. Anne Baillot

Zusammenfassung: Das Rom des 19. Jahrhunderts ist auch die Stadt der Bildungsreisenden. Villen, Cafés und Galerien fungieren als Treffpunkte; Erlebnisse und Fachwissen aus Kunstgeschichte, Mineralogie, Philosophie etc. werden ausgetauscht. Das nachfolgend beschriebene Forschungsprojekt analysiert die literarische Inszenierung sozialer Netzwerke und Wissenstransfers in Rom, dargestellt in deutschsprachigen Reiseberichten (1816-1833). Zur Anwendung kommen computergestützte Methoden: Die Berichte werden in XML semantisch ausgezeichnet, die gewonnenen Daten sollen daraufhin in eine Datenbank eingespeist und visualisiert werden. Im zweiten Teil der Untersuchung werden diese unter literatursoziologischen und wissenspoetologischen Gesichtspunkten interpretiert.

Untersuchungsgegenstand und Zielsetzung: Im deutschsprachigen Raum kommt es seit den aufgeklärten Lesegesellschaften und der späteren Salon-Kultur zu zahlreichen Vereinsgründungen. Der Habitus des geselligen Gelehrten wird auch von Reisenden in Rom gepflegt: Intellektuellenzirkel, wie der Deutsche Künstlerverein oder die Zusammenkünfte im Antico Caffè Greco, entstehen. Fachwissenschaftliches Wissen wird gemeinsam erarbeitet und diskutiert – ‚Netzwerken' ist wesentlicher Bestandteil der Aufenthalte. Dieses Phänomen spiegelt sich in seiner Dynamik besonders in Reiseberichten wider: Goethe fährt nach Rom, beschreibt abendliche Malzirkel (Goethe [HA] 1974: 134-136) etc. Die dänisch-deutsche Schriftstellerin Friederike Brun wohnt mit dem Ehepaar von Humboldt zusammen und berichtet Begegnungen und philosophische Gespräche (Brun 1833: 171-179). Diese und andere literarisierte Begegnungen **sollen im dargelegten Projekt erstmalig umfassend analysiert und visualisiert werden.** Die Forschungsfrage lautet: **Welche sozialen Netzwerke werden in den Reiseberichten inszeniert und welche gemeinsam erörterten Diskurse prägen die Zusammenkünfte?**

Forschungskontext: Das Projekt ordnet sich in einen äußerst aktuellen Forschungskontext über Reisenetzwerke im 19. Jahrhundert ein: **Geschichte**: DHI Rom: *Künstler, Agenten und Sammler in Rom 1750-1850*; **Kunstgeschichte**: Karl S. Rehberg: SFB 804 *Transzendenz und Gemeinsinn* (Netzwerke deutscher und französischer Künstler in Rom); **Musikwissenschaften**: *Europäische Musiker in Venedig, Rom und Neapel (1650-1750).*[1] Eine Analyse der Reiseberichte unter literaturwissenschaftliche Perspektive wurde bisher noch nicht vorgenommen. Diese Lücke soll nun geschlossen werden. Methodisch orientiert sich das Dissertationsvorhaben dabei insbesondere an dem oben genannten musikwissenschaftlichen Projekt. Damit wird der inhaltliche sowie methodische Anschluss an das Forschungsumfeld gewährleistet beziehungsweise wird dieses durch die literaturwissenschaftliche Perspektive angereichert. Die Einbindung der Forschungsergebnisse in eine Onlineplattform soll die

[1] Projekt-URLs siehe Literaturverzeichnis.

öffentliche Nutzung und Langzeitverfügbarkeit der Daten gewährleisten (Shillingsburg 2006: 12). Eine CC-BY-Lizenz wird bevorzugt, um Verbreitung zu ermöglichen.

Methodik und Vorgehen: Digitale Methoden kommen zur Anwendung, die einen **neuen Blick auf die Vielfalt der Reisebeschreibungen** eröffnen sollen. Die Strukturierung und Visualisierung können Interferenzen sichtbar machen, die durch eine klassische Textanalyse in diesem Umfang kaum sichtbar wären. Das zugrundeliegende **Korpus** wird aus ca. 30 Texten bestehen, eingeleitet durch Goethes *Italienische Reise* (1816) und beschlossen durch Friederike Bruns *Römisches Leben* (1833). Dieser Zeitraum ist ein Höhepunkt in der Italien-Reiseliteratur, es werden besonders viele Berichte verlegt.

Die meisten relevanten Berichte liegen bereits als PDF vor. Diese werden mit Hilfe der OCR-Erkennungssoftware ABBYY FineReader maschinenlesbar gemacht und in **XML** ausgezeichnet: Den Personen sind Attribute wie Beruf, regionale Herkunft, persönliches Verhältnis, Gesprächsthemen etc. zuzuordnen. Wo vorhanden, kommen **Normdaten-Identifizierungen** (GND-Referenzen) zum Einsatz. In einem nächsten Schritt werden die jeweiligen Personennetzwerke mit Gephi **visualisiert**,[2] die Datenmenge wird damit leichter interpretierbar. Zudem wird angestrebt, eine zeitliche Dimension mit Hilfe einer Timeline einzubauen, sodass eine dynamische Darstellung der Daten gewährleistet wird. Hierauf aufbauend sind die erhobenen **Daten auszuwerten**. Dazu werden Theorien aus dem Bereich der Literatursoziologie sowie der Wissenspoetologie herangezogen.[3]

Relevante Informationen auf dem Poster:

- Untersuchungsgegenstand, Forschungsfrage, exemplarisches Textbeispiel
- Methodik, Vorstellung verwendeter Tools
- Eigener Forschungsstand zum Zeitpunkt der Tagung (veranschaulicht in einem Projekt-Zeitstrahl)

Zitierte Literatur und Projekte:

- **Baßler**, Moritz (Hrsg.): New Historicism. Tübingen ^{2}2001.
- **Brun**, Friederike: Römisches Leben. Band 1, Leipzig 1833.
- **Europäische Musiker in Venedig, Rom und Neapel (1650-1750):** www.musici.eu [29.10.2014].
- **Goethe**, Johann W. von: Italienische Reise, in: Goethes Werke. Hrsg. von Erich Trunz. Hamburger Ausgabe. Band 11: Autobiographische Schriften III. Hamburg ^{8}1974.
- **Hogrebe**, Wolfram: Societas Teutonica. Erlangen/ Jena 1996.
- **Klausnitzer**, Ralf: Literatur und Wissen. Berlin 2008.
- **Künstler, Agenten und Sammler in Rom 1750-1850**: http://dhi-roma.it/projekte-aktuell+M5a1e59c05e9.html [29.10.2014].
- **Rehberg**, Karl-Siegbert: SFB 804 Transzendenz und Gemeinsinn. www.sfb804.de [29.10.2014].
- **Shillingsburg**, Peter L.: From Gutenberg to Google. Cambridge 2006.
- **Vogl**, Joseph (Hrsg.): Poetologien des Wissens um 1800. München 1999.

[2] Siehe http://gephi.github.io/ [29.10.2014].

[3] Siehe bspw. Hogrebe 1996. Zum New Historicism siehe Baßler 2001. Zur Theorie einer „Poetologie des Wissens" Klausnitzer 2008: 169-183 und Vogl 1999.

Digitalisierung eines NS-Bildarchivs – Konstruktion von NS-Lebenswelt
Posterpräsentation

Unser Projekt repräsentiert das Erste dieser Art im Feld der Theaterwissenschaft und unsere Intention ist es weitere Initiativen für unsere Disziplin anzuregen. Das sogenannte Bildarchiv ist Teil des Archivs und der historischen Theatersammlung des Instituts für Theater-, Film- und Medienwissenschaft der Universität Wien (TFMA). Im April 2011 wurde dieses verschollen geglaubte, umfangreiche historische Bildarchiv des ehemaligen „Zentralinstituts für Theaterwissenschaft" wiederaufgefunden. Der Bestand umfasst vorwiegend Fotografien (Schauspielerporträts und Theaterfotografien zwischen 1880–1945), Stiche und Grafiken aus dem 19. Jahrhundert, insgesamt ca. 2000 Einzelstücke. Den Hauptteil bildet die visuelle Dokumentation von NS-Theatern, die in solcher Vollständigkeit keine österreichische oder deutsche Sammlungsinstitution aufzuweisen hat. Es handelt sich hierbei um Fotografien sämtlicher Produktionen an Wiener Bühnen im Zeitraum von 1938 bis zur Theatersperre im Juli 1944, von Prager Bühnen und Produktionen sogenannter Grenzlandtheater. Dieses Fotomaterial wurde dem Institut von verschiedenen Pressefotographen zur Verfügung gestellt. Weitere historische Fotos sind frühe und äußerst rare Schauspielerporträts, sehr häufig mit handschriftlichen Widmungen versehen. Der Aufbau dieses Bildarchivs war eine der ersten Maßnahmen des 1943 an der Universität Wien gegründeten „Zentralinstituts für Theaterwissenschaft". In Kontext gesetzt zu den Zielvorgaben seitens des Reichserziehungsministeriums, nämlich nach dem Krieg als Reichsinstitut die Bedeutung von Theater und Film für das großdeutsche Reich zu definieren, hat dieser Fotobestand große Brisanz.

Dieses Bildarchiv bietet eine exzellente Möglichkeit zur Entwicklung einer Digital Humanities-Strategie für unser Fach und erleichtert fächer- und institutionenübergreifende Zusammenarbeit und Vernetzung. Unser Projekt zielt auf einen intensiven Austausch zwischen verwandten wissenschaftlichen Feldern ab. Die Interdisziplinarität ist eine wichtige Säule des Projekts, sowohl technisch als auch wissenschaftlich motiviert.

Dabei soll vorrangig auf bereits bestehende Tools und Standards zurückgegriffen werden, wobei an manchen Stellen auch Eigenentwicklungen notwendig sein werden. Diese werden idealerweise in bereits aktiv genutzte Strukturen einfließen. Zudem wird mit der Digitalisierung des Bildarchivs die strukturelle Grundlage geschaffen, um den gesamten Bestand des TFMA digital aufzubereiten. Es wird dafür ein für die Theaterwissenschaft prototypischer Workflow entwickelt werden.

Für die Präsentation und Bearbeitung der digitalisierten Objekte wird eine Webplattform erstellt, die die wissenschaftlichen Ergebnisse unseres Projekts zu Beginn in den Mittelpunkt stellt, dabei aber für zukünftige Forschungsarbeiten offen und jederzeit andockbar bleibt.

Die digitalisierten Objekte werden der Forschungsgemeinschaft und Öffentlichkeit in der freien Lizenz CC BY 4.0 zur Verfügung gestellt und in einem Open Access-Format angeboten. Durch die Wahl dieser Lizenzierung ist die Basis für eine breite (Nach-)Nutzung gelegt. Zudem wird auf Langzeitarchivierung Rücksicht genommen, indem die Digitalisate im Repositorium von PHAIDRA (Digital Asset Management System mit Langzeitarchivierungsfunktionen der Universität Wien, http://phaidra.univie.ac.at) eingebunden werden. Zudem werden mit Tools auf der Webplattform neuartige Zugriffe auf den Bestand möglich sein, der ein innovatives, zeitgemäßes wissenschaftliches Arbeiten unterstützt.

Mit den Bedeutungseinschreibungen zu Theater und Film lassen sich nicht alleine ästhetische Fragekomplexe aufwerfen. Noch dringlicher stellen sich dabei Fragen nach NS-Menschenbildern. Konstruktionsvorgänge dieser Menschenbilder werden über Theater- und Filmproduktion erkennbar. Die Digitalisierung dieses NS-Bildarchivs birgt für die internationale Forschung unterschiedlicher Disziplinen großes Innovationspotential, da sich über diese Materialien nicht allein Repräsentationsformen untersuchen lassen, sondern auch sichtbar wird, wie parallel zur Vernichtung der als „nichtarisch“ gekennzeichneten Personen ein neues NS-Menschenbild konstruiert wird.

Auf unserem Poster skizzieren wir die eben beschriebenen Abläufe exemplarisch anhand von ausgewählten Theaterfotografien aus diesem NS-Bildarchiv, um aufzuzeigen, wie eine digitalisierte Aufbereitung ideologische Sammlungsstrategien und Wissenschaftspolitik sichtbar macht. Die einzelnen Schritte werden sowohl auf technischer als auch inhaltlicher Ebene erläutert.

Projektteam

PD Mag. Dr. Birgit Peter
Mag. Klaus Illmayer
Mag. Johannes A. Löcker

Archiv und theaterhistorische Sammlung (TFMA)
tfm | Institut für Theater-, Film- und Medienwissenschaft
Universität Wien
Hofburg, Batthyanystiege
1010 Wien

Kontakt: birgit.peter@univie.ac.at

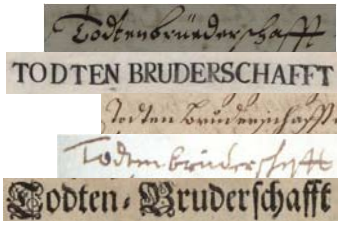
Eva Wohlfarter
Österreichische Akademie der Wissenschaften
Austrian Centre for Digital Humanities

JENSITES – Schriften einer Totenbruderschaft digital

Das Poster **JENSITES – Schriften einer Totenbruderschaft digital** informiert über ein kleines Forschungsprojekt (www.oeaw.ac.at/icltt/bruderschaftsdrucke), das derzeit – von der Stadt Wien gefördert – am Institut für Corpuslinguistik und Texttechnologie durchgeführt wird.

Zum Hintergrund: Bruderschaften waren christliche Vereinigungen, die im Zuge der Gegenreformation einen neuen Aufschwung erfahren und die religiöse Alltagskultur in ganz Europa stark geprägt haben. Während den barocken Bruderschaften als Forschungsgegenstand im internationalen Vergleich ein wachsendes Interesse entgegengebracht wird („Society for Confraternity Studies"), steht ihre Erforschung im deutschsprachigen Raum, von wenigen Einzelstudien abgesehen, erst am Anfang.

Das genannte Projekt hat daher den Anspruch, anhand einer in Wien gegründeten, sogenannten **Totenbruderschaft** zu zeigen, wie man sich diesem vernachlässigten, in höchstem Maße interdisziplinären Forschungsgegenstand mit digitalen Methoden nähern kann:



1) Sichtung und Digitalisierung der Quellen

Die Quellen der kaiserlich-königlichen Totenbruderschaft befinden sich in verschiedenen kirchlichen und öffentlichen Archiven und Bibliotheken. Die Bruderschaft hat einerseits Handschriftliches wie etwa Gründungsdokumente, Urkunden, Verträge und interne Aufzeichnungen hinterlassen und andererseits Gedrucktes wie etwa Statuten, Gebetsbücher für Mitglieder, Neujahrskalender, Predigten und Memento mori-Dichtung publiziert. Durch den Digitalisierungsprozess wird dieses dislozierte Quellenmaterial virtuell zusammengeführt.

2) Aufbereitung und Annotation der Quellen

Um das reichhaltige Quellenmaterial zu erschließen, werden die gescannten Image-Digitalisate der Originale in maschinenlesbare Volltextversionen umgewandelt und in ein XML-Format (Version P5) überführt. In einem weiteren Schritt werden die Textdaten in einem semi-automatischen Verfahren mit linguistischen Informationen versehen (Tokenisierung, Wortklassenzuordnung und Lemmatisierung). Das Projekt profitiert hier von bereits abgeschlossenen Projekten, in denen am Insititut für Corpuslinguistik und Texttechnologie daran gearbeitet worden ist, das automatische Tagging (Treetagger) durch die Verwendung bereits korrigierter Daten aus dieser Zeit zu verbessern und das standardisierte *Stuttgart Tübingen TagSet* für die Sprachstufe des Älteren Neuhochdeutsch zu adaptieren. Das

gewonnene Textmaterial, das sorgfältig kollationiert und korrigiert wird, erweitert einerseits die Datenbasis von historischen Texten aus dieser Zeit und ermöglicht andererseits die kontinuierliche Weiterentwicklung eines stabilen Methodeninventars für nicht-kanonische Varietäten.

3) Virtuelle Kontextualisierung der Quellen

Da die Totenbruderschaft eine kaiserliche Gründung war und ihr Vorstand und ihre Mitglieder zum Teil dem hohen Adel angehörten, sollen die in den Texten vorkommenden Personennennungen mit den vorhandenen RDF-Datensätzen der "Deutschen Biographie" http://www.deutsche-biographie.de/ verlinkt werden. In Kooperation mit der Wiener Stadt- und Landesbibliothek wird derzeit an Referenzierungsmöglichkeiten von Orten, Bauwerken und Institutionen mit der kürzlich präsentierten historischen Wissensplattform "Wien Geschichte Wiki" https://www.wien.gv.at/wiki/ gearbeitet. Mit dem Einsatz verschiedener Textanalysetools werden sich sprachliche und inhaltliche Bezüge innerhalb der Quellengruppe nachweisen und visualisieren lassen, was eine wesentliche Voraussetzung zur Interpretation und Funktionsbeschreibung dieser Texte sein wird.

Das Projektvorhaben – die Erforschung der Quellen der Totenbruderschaft – verbindet philologische Expertise mit moderner Informations- und Kommunikationstechnologie. Es hat daher zwar eine germanistisch-philologisch-kulturwissenschaftliche Ausrichtung, ist aber allein durch die gewählte Methodik der digitalen Quellenaufbereitung **interdisziplinär** angelegt: Die Quellen, die darin beispielhaft erschlossen werden, sind nicht nur ein wesentlicher Beitrag zu sozial-, kultur- und alltagsgeschichtlichen Aspekten der Stadt Wien, sondern auch Forschungsgegenstand der Sprachgeschichte und der Theologiewissenschaft und ermöglichen prosopographische Studien und historische Netzwerkforschungen.

Die Totenbruderschaft ist nur eine von vielen barocken Bruderschaften, doch eignet sich die günstige, bislang kaum erforschte Quellenlage in besonderer Weise dazu, beispielhaft aufbereitet und analysiert zu werden. **Es ist daher das Ziel des Projekts, die erhaltenen Quellen der Bruderschaft mit zeitgemäßen Methoden zu erschließen und in einem breiteren kulturwissenschaftlichen Kontext auszuwerten.** Auf Basis des digital aufbereiteten Materials sollen erstmals fundierte Aussagen über Geschichte, Mitglieder, Tätigkeit, Funktion und kulturelle Bedeutung dieser religiös begründeten Sozietät formuliert werden.

Bei der Quellenaufbereitung wird die Archivierung der Texte bereits mitbedacht – nach Abschluss des Projekts sollen die Daten Teil der Sammlung ABaC:us - Austrian Baroque Corpus werden.

TTLab Preprocessor – Eine generische Web-Anwendung für die Vorverarbeitung von Texten und deren Evaluation

Rüdiger Gleim und Alexander Mehler

Goethe-Universität Frankfurt

1 Einführung und Motivation

Dieser Beitrag stellt den *TTLab Preprocessor* (kurz: *TTLab PrePro*) als generische Web-Anwendung für die Vorverarbeitung von Texten in den *Digital Humanities* vor. Er erörtert die Architektur des *TTLab PrePro*, exemplifiziert das von ihm anvisierte Nutzungsszenario und fasst seinen aktuellen Entwicklungsstand zusammen.

Die linguistische Vorverarbeitung von Texten ist ein integraler Bestandteil jeder automatischen Textanalyse. Dies beinhaltet unter anderem die Erkennung der dem jeweiligen Text zugrundeliegenden Sprache(n), die Erkennung seiner logischen Dokumentstruktur, die Tokenisierung und Lemmatisierung seiner lexikalischen Konstituenten und die Annotation ihrer Wortarten (*PoS-Tagging*). Es existiert eine Reihe von Software-Systemen und -Komponenten, welche die Vorverarbeitung für verschiedene Sprachen umsetzen. In der Literatur werden dabei etwa für das PoS-Tagging Erkennungsraten von über 95% dokumentiert.[1] Für viele Fragestellungen, wie z.B. die Textklassifikation, fällt eine entsprechende Fehlerquote von ca. 5% kaum ins Gewicht. Im Bereich der *Digital Humanities*, bei der es etwa um die qualitative Analyse einzelner Wortbedeutungen geht, sind jedoch bereits Fehlerquoten von 1% oftmals inakzeptabel.[2] Gerade in diesem Bereich ist die automatische Vorverarbeitung zumeist der Ausgangspunkt für die nachfolgende unabdingbare manuelle Korrektur der Annotationen.

So stellt sich die Frage etwa zu Beginn eines Forschungsprojekts, wie hoch die erwartete Fehlerquote für Texte der untersuchten Sprache beim Einsatz eines bestimmten Präprozessierers ist. Zur Beantwortung dieser Frage kann eine Sammlung von Texten manuell vorverarbeitet und als so genannter *Gold-Standard* zur Bewertung der automatischen Vorverarbeitung herangezogen werden. Vergleicht man die Annotationsergebnisse verschiedener Systeme mit einem solchen Goldstandard, so können Kennzahlen zur Ermittlung der erwarteten Fehlerrate gewonnen werden, um schließlich den Aufwand für entsprechende manuelle Korrekturen zu schätzen. Da die Parametrisierung sowie die Ein- und Ausgabeformate verschiedener Systeme zur Vorverarbeitung variieren, ist die Durchführung einer solchen Evaluation aufwendig und ihrerseits fehleranfällig. Die Funktion, verschiedene Systeme über eine generische Schnittstelle nicht nur verwendbar, sondern auch evaluierbar zu machen, bildet folglich den funktionalen Kern des *TTLab PrePro*.

[1] Diese Rate schwankt erwartungsgemäß je nach Sprache und Genre der untersuchten Texte [Giesbrecht and Evert, 2009].

[2] Anne Bohnenkamp-Renken (2013); *persönliche Kommunikation*.

2 TTLab Preprocessor Web-Anwendung

Der *TTLab PrePro* ermöglicht die Vorverarbeitung von Texten, die automatische Evaluation auf der Basis von Goldstandards und die einzelfallbezogene Fehleranalyse. Die Eingabe in das System kann direkt über den Browser in Form einer Texteingabe, die Angabe einer Webressource oder den Upload von Dateien erfolgen. Die Upload-Funktion ermöglicht nicht nur das Hochladen mehrerer Dateien auf einmal, sondern auch die Verwendung von komprimierten Archiven. An Dateiformaten werden unter anderem HTML, PDF, RTF und DOC unterstützt. In der Voreinstellung wird die Sprache der Inputtexte automatisch erkannt und der für die jeweilige Zielsprache voreingestellte Präprozessierer verwendet. Es ist auch möglich, diese Parameter explizit zu setzen. Die Ausgabe erfolgt mittels *TEI P5* [TEI, 2014]. Die Ergebnisse können direkt im Browser in verschiedenen Sichten betrachtet und frei heruntergeladen werden.

Werden TEI-P5-Dokumente als Eingabe verwendet, so werden diese vom System – wie bei jedem anderen Eingabeformat – auf den unstrukturierten Text heruntergebrochen. Anschließend werden sie durch den Präprozessierer vorverarbeitet und in TEI P5 repräsentiert. Bilden annotierte TEI-P5-Dokumente den Input, so können diese als Goldstandard interpretiert werden. Das System evaluiert in diesem Falle den jeweils ausgewählten Präprozessierer auf der Basis dieses Goldstandards. Da die Tokenisierung zwischen den zu vergleichenden Dokumenten variieren kann, wird zunächst mittels dynamischer Programmierung ein Alignment der Token durchgeführt. Anschließend wird das Ergebnis der Lemmatisierung sowie des Taggings mit dem Goldstandard verglichen. Auf diese Weise können die aus dem *Machine Learning* bekannten Maße *Precision*, *Recall* und *F-Score* berechnet werden. Die Ergebnisse werden direkt im Browser angezeigt – sowohl für die einzelnen Dokumente, als auch für das Eingabekorpus insgesamt. Analog wird eine Rangverteilung der häufigsten Tagging- und Lemmatisierungsfehler (nach abnehmender Häufigkeit) visualisiert. Schließlich können die Tagging- und Lemmatisierungsfehler in einer tabellarischen Ansicht im jeweiligen Satzkontext untersucht werden. Abbildung 1 exemplifiziert eine solche Ansicht von Evaluationsergebnissen. Die obere Tabelle beinhaltet eine Liste aller evaluierten Dokumente mit den Gesamtergebnissen. Für ein ausgewähltes Dokument können, wie in diesem Beispiel gezeigt, Belegstellen von Tagging-Fehlern im Satzkontext aufzeigt werden. Dies erlaubt das gezielte Nachverfolgen und Beheben von Fehlern.

Der *TTLab PrePro* ist als Java- und JavaScript-basierte Client-Server-Architektur implementiert. Die Benutzeroberfläche ist mithilfe des JavaScript-Frameworks ExtJS realisiert. Das in *Apache Tomcat* laufende *Java Servlet* bearbeitet die Nutzeranfragen, ruft externe Systeme zur Vorverarbeitung auf, führt ggf. Evaluationen durch und bereitet die Ergebnisse für die Darstellung im Browser auf. In der aktuellen Version sind zwei Systeme des *TTLab Preprocessor* [Mehler et al., 2015, Waltinger, 2010] integriert sowie das System namens *Stanford CoreNLP* [Manning et al., 2014].

3 Zusammenfassung und Ausblick

Der vorliegende Beitrag stellt den *TTLab PrePro* als System zur Vorverarbeitung von Texten und darauf basierenden Evaluationen vor. Das mit der geplanten Publikation veröffentlichte System ist frei ver-

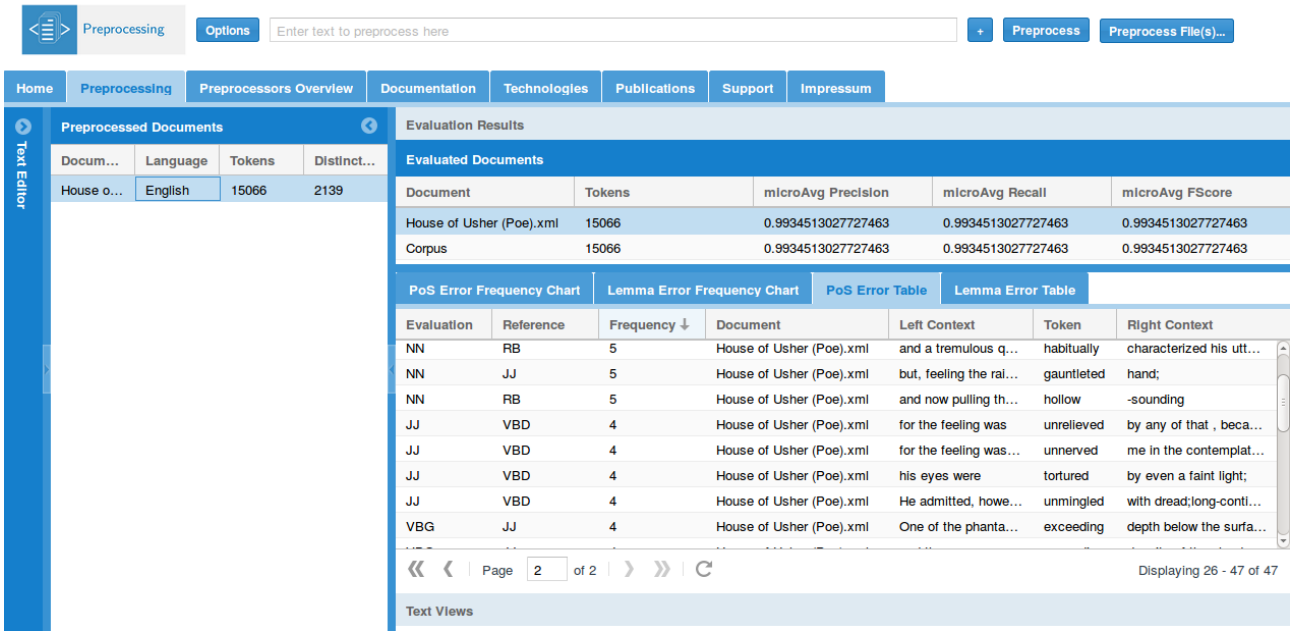


Abbildung 1: Ansicht von Evaluationsergebnissen, welche Tagging-Fehler mit Belegstellen im Satzkontext aufzeigt.

wendbar (*open access*). Die Weiterentwicklung zielt auf die Nutzbarmachung des UIMA-Frameworks[3]. Zum einen, um den Pool der verfügbaren Systeme zur Vorverarbeitung zu vergrößern, zum anderem, um umfangreiche Parameterstudien über die einzelnen Komponenten durchführen zu können. Ferner soll eine Normalisierung von PoS-Tagsets für die Evaluation entwickelt werden. Der *TTLab PrePro* zielt vor allem darauf, von Geisteswissenschaftlerinnen und -wissenschaftlern auch ohne Informatik-Vorkenntnisse genutzt werden zu können. Unterstützt werden derzeit die Sprachen Latein [Mehler et al., 2015], Englisch und Deutsch.

Der *TTLab PrePro* kann unter der URL `http://prepro.hucompute.org` getestet werden. Ein TEI-P5-Dokument zum Testen der Evaluation steht unter der Adresse `http://prepro.hucompute.org/examples/poe.tei` bereit.

Danksagung

Diese Arbeit ist im Rahmen des BMBF-Projekts *Computational Historical Semantics* (`www.comphistsem.org`) entstanden, für dessen Unterstützung wir uns herzlich bedanken.

Literatur

TEI P5: Guidelines for electronic text encoding and interchange, 2014. URL `\url{http://www.tei-c.org/Guidelines/P5/}`.

[3] `https://uima.apache.org/`

Eugenie Giesbrecht and Stefan Evert. An evaluation of part-of-speech taggers for the web as corpus. In *Proceedings of DGfS-CL Postersession 2009*, 2009.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Alexander Mehler, Tim vor der Brück, Rüdiger Gleim, and Tim Geelhaar. *Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger*. Theory and Applications of Natural Language Processing. Springer, Berlin/New York, 2015.

Ulli Waltinger. *On Social Semantics in Information Retrieval*. Phd thesis, Bielfeld University, Germany, 2010.

Projekt Altägyptische Wörterbücher im Verbund: Digital unterstütze Analyse der Entwicklung ägyptischer Wörterbücher

In der Frühzeit der wissenschaftlichen Untersuchung des Ägyptischen im 19. und frühen 20. Jahrhundert ist eine Vielzahl von Wörterbüchern, Wortlisten und Glossaren entstanden. Da die Ägyptische Sprache erst allmählich entschlüsselt und verstanden wurde, entwickelte sich auch die Sicht auf die Erfassung der ägyptischen Wörter weiter. Alle aktuell bekannten Wörter des Ägyptischen sind im Thesaurus Linguae Aegyptiae (http://aaew.bbaw.de/tla, auch Berliner Wortliste, kurz BWL) erfasst. Um nachzuvollziehen, wie sich das Verständnis des Ägyptischen über die Zeit verändert hat, schafft das vorliegende Projekt eine Infrastruktur dafür, das Vorkommen von Wörtern in ägyptischen Wörterbüchern und anderen lexikographisch relevanten Publikationen mit den Einträgen der BWL zu verknüpfen. Mit dem entwickelten Werkzeug sind automatisierte Auswertungen der Entwicklung des Verständnisses der ägyptischen Lexik – und damit der Lexikographie – möglich, die ohne IT-Unterstützung nicht denkbar wären.

Das Projekt „Altägyptische Wörterbücher im Verbund“ orientiert sich an Wörterbuchportalen wie "Wörterbuchnetz" (http://woerterbuchnetz.de/), "OWID" (http://www.owid.de/), "Etymologiebank" (http://www.etymologiebank.nl) und Infolux (http://infolux.uni.lu/worterbucher/). Spezifisch für das Ägyptische ist, dass die meisten Wörterbücher handgeschrieben sind und neben deutschen, französischen, englischen oder italienischen Textteilen auch Hieroglyphen und Partien in Demotisch, Koptisch, Griechisch, Lateinisch, Hebräisch und Arabisch enthalten (können), ganz abgesehen von Transliterationen mit Zeichen, die nicht alle im Unicode-Format existieren. Unter diesen Umständen ist eine Texterfassung per OCR und eine anschließende Auszeichnung in TEI/XML nicht möglich, so dass Verknüpfungen von Bilddaten mit einer Lemmaliste vorgenommen und Metadaten an diese Verknüpfungen angehängt werden müssen.

Die unterschiedlich strukturierten (teils nach hieroglyphischer Orthographie, teils nach Transliteration, teils nach konventionalisierter hieroglyphischer Zeichenliste) und in der Methodik der Worttransliteration individuell verfahrenden Wörterbücher machen es unmöglich und unnötig, für jedes Wörterbuch eine eigene Lemma-Liste zu erstellen und diese mit einer Hyper-Lemma-Liste zu verknüpfen. Stattdessen wird die BWL als standardisierte Lemma-Liste eingeschaltet, was zudem eine zukünftige Verknüpfung mit der ägyptischen Textdatenbank ermöglicht. So werden die alten Wörterbücher langfristig gesehen mit einem System, das ägyptische Volltexte erfasst, verschränkt.

Zur Erfassung und Aufbereitung der Informationen über die Entwicklung ägyptischer Wörterbücher wurde ein webbasiertes Werkzeug erstellt. Dieses bietet eine für mehrere Forscher mit differenzierenden Fragestellungen gleichzeitig zugängliche Oberfläche, um die Verknüpfung von Wörterbuch- und anderen Publikationseinträgen mit Wörtern der BWL zu ermöglichen. Die BWL wurde in einem Vorverarbeitungsschritt extrahiert, so dass die Wörter mit Metadaten als Graphiken vorlagen. Die einzelnen Seiten der oft handgeschriebenen Publikation lagen ebenfalls als Graphiken vor. Metadaten zu den Graphiken und zur Verknüpfung wurden in einer MySQL-Datenbank gespeichert, auf die anschließend mit Hibernate zugegriffen wurde. Hierbei wurde bspw. die Möglichkeit geschaffen, zu jeder einzelnen Verknüpfung zu annotieren, wie zutreffend die damalige Transliteration und Übersetzung aus heutiger Sicht waren.

Um die Verknüpfung herzustellen, musste die Möglichkeit geschaffen werden, in einer Eingabeoberfläche zu erfassen, welcher Bereich einer Seite mit welchem Eintrag der BWL zusammenhängt. JSF und PrimeFaces enthalten bereits verschiedene Funktionalitäten, um diese Verknüpfung herzustellen, bspw. Datentabellen mit Suchfunktion für die Wörter der BWL und ein Verfahren, um Bildbereiche in den Publikationsseiten zu extrahieren. Aus diesem Grund wurden diese zur Umsetzung der Weboberfläche eingesetzt und so ein Werkzeug geschaffen, mit dem sich Wörter der BWL mit Publikationsseitenbereichen verknüpfen lassen.

Weiterhin wurde eine Möglichkeit geschaffen, die Publikationen nach festgelegten Kriterien zu analysieren. Hierbei wurden Anfrageoptionen nach Metadaten der einzelnen Verknüpfungen und statistischen Daten der Metadaten mit MySQL implementiert. Diese lassen sich bei Bedarf problemlos erweitern. Auf diese Weise lassen sich die großen Datenmengen der Wörterbücher schnell und effizient auswerten.

Mit dem beschriebenen Werkzeug wurden seitdem verschiedene ägyptische Wörterbücher des 19. Jahrhunderts erfasst und analysiert. Hierbei wurden in insgesamt zehn Publikationen 30371 Verknüpfungen erstellt. Die sukzessive erweiterbare Datenbank und die installierten Funktionen werden den Nutzer in die Lage versetzen, die einzelnen Etappen der Erforschung der ägyptischen Lexik und die jeweils zu Grunde gelegte Konzeption und Methodik besser nachvollziehen zu können. Im idealen Falle soll es möglich sein, die Forschungsgeschichte eines Lexems von den Anfängen der Ägyptologie bis zum Erscheinen des Wörterbuchs der Ägyptischen Sprache (1926-31), das die Basis der BWL darstellt, und darüber hinaus skizzieren zu können. Je nach Suchkriterium können dabei exemplarische bzw. statistische Daten ausgewertet werden, wie z.B. Qualität und Quantität der verwendeten Primär- und Sekundärquellen, Häufigkeit von lexikographischen Fehlern und Missdeutungen, die entweder mangels besserer Kenntnis und auch gelegentlich aufgrund fehlender Sorgfalt vorgekommen sind.

Insgesamt ermöglicht die so geschaffene IT-Infrastruktur eine Analyse der frühen ägyptischen Lexikographie, die sonst nicht bzw. nur durch sehr zeitaufwändige Recherchen möglich wäre. Die Präsentation wird sowohl die zugrundeliegenden Fragestellungen der Ägyptologie und bereits gefundene Lösungen als auch die technische Infrastruktur, die zur Unterstützung der Beantwortung der Fragestellungen geschaffen wurde, darstellen.

Ingo Börner, Angelika Hechtl

Quantitative Aufführungsanalysen zu Stücken Johann Nestroys

Die Posterpräsentation stellt einen Ansatz einer computergestützten quantitativen Aufführungsanalyse vor.

Basierend auf dem von Solomon Marcus (1970) vorgelegten mathematischen Dramenmodell, das durch Manfred Pfisters (2001) seine theoretische Fundierung sowie von Hartmud Ilseman (1890) eine praktische Umsetzung in der Analyse der Dramen von Shakespeare erfahren hat, werden im projektierten Vorhaben die Möglichkeiten zur Anwendung quantitativer Verfahren zur Analyse von Inszenierungen ausgelotet. In der bisherigen Anwendung quantitativer Verfahren auf Dramen stand der Dramentext alleine im Zentrum des Erkenntnisinteresses. Die konkrete Umsetzung als Inszenierung ist im Unterschied zum Film bisher nicht unter Rückgriff auf quantitative Methoden untersucht worden.

Die quantitative Aufführungsanalyse ermöglicht es, unterschiedliche Inszenierungen der Stücke Johann Nepomuk Nestroys anhand des Merkmals Bühnenpräsenz untereinander und mit dem Dramentexten zu vergleichen. Die entwickelte Methode wird exemplarisch anhand einiger ausgewählter Dramentexte („Der Talisman“, „Der böse Geist Lumpazivagabundus“ und „Der Zerrissene“) erprobt.

Untersuchungsgegenstand bilden sowohl aktuelle Aufführungen an Wiener Bühnen, als auch Aufzeichnungen von älteren Aufführungen (wie etwa Salzburger Festspielinszenierungen). Mit dem Merkmal „Bühnenpräsenz“ wurde in Anlehnung an das von Erika Fischer-Lichte (2007) beschriebene System theatralischer Zeichen als jenes Charakteristikum identifiziert, welches einen Vergleich von Inszenierungen untereinander und mit dem Dramentext ermöglicht. Es wird ermittelt, welche SchauspielerInnen zu welchem Zeitpunkt auf der Bühne anwesend sind. Die erhobenen Daten lassen sich mit dem Dramentext in Beziehung setzen.

Das erhobene Datenmaterial wird mittels R aufgearbeitet und visualisiert. So sollen im Text vorhandene Strukturen und ihre konkrete Realisierung in den unterschiedlichen Aufführungen nachvollziehbar gemacht werden.

Fischer-Lichte, E. (2007): *Semiotik des Theaters. Bd. 1. Das System der theatralischen Zeichen*. Tübingen.

Ilsemann, H. (1998): *Shakespeare Disassembled. Eine quantitative Analyse der Dramen Shakespeares*. Frankfurt a. M.

Marcus, S. (1970): „Ein mathematisch-linguistisches Dramenmodell“. In: *Zeitschrift für Literaturwissenschaft und Linguistik*, S. 139–152.

Pfister, M. (2001): *Das Drama. Theorie und Analyse*. München.

Dr. Jakub Šimek

Universität Heidelberg
Germanistisches Seminar
Hauptstraße 207–209
D-69117 Heidelberg
+49-(0)6221-543217

jakub.simek@gs.uni-heidelberg.de

Christoph Forster

datalino. Forster, Fabian, Krumnow PartG
Martin-Luther-Straße 120
D-10825 Berlin
+49-(0)30-78893232

forster@datalino.de

DHd-Tagung 2015
Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation

Kodierung, Analyse und Visualisierung mittelalterlicher Kodexstrukturen im editorischen Kontext

Posterpräsentation – Abstract

Für die Beschreibung der Lagen- und Blattstruktur mittelalterlicher Kodizes wird in gängigen Handschriftenbeschreibungen meist die Chroust'sche Lagenformel verwendet. Diese zeigt durch römische Buchstaben die Anzahl der Doppelblätter in einer Lage an, während hochgestellte Ziffern auf die Blattzählung verweisen. Bei Wiederholung gleichartiger Lagen in einem Kodex wird die Art dieser Lagen zusammen mit deren Anzahl nur einmal angegeben. Fehlende Blätter werden nur durch Minuszeichen, eingefügte durch Pluszeichen und deren Anzahl bei einer Lage angedeutet, ohne dass normalerweise eine genaue Zuordnung zum konkreten Doppelblatt möglich wäre. Die TEI sieht für den Inhalt des Elements `<collation>` keine genauere Spezifizierung vor.

Eine Schwäche derartiger Beschreibungen ist, dass sie einerseits nicht ohne Weiteres maschinenlesbar und eindeutig genug sind, um mit digitalen Faksimiles verknüpft zu werden, und dass sie andererseits von den meisten Benutzern gedruckter und digitaler Ausgaben kaum wahrgenommen werden. Dabei ist die physische Lagen- und Blattstruktur gerade bei individuell hergestellten Buchartefakten wie mittelalterlichen Kodizes häufig wesentlich für das Verständnis der Textgestaltung und -überlieferung sowie der Seitenarrangements. So sind etwa Textlücken in einer Abschrift potenziell auf fehlende Blätter in der Lagenmitte der Vorlage zurückzuführen oder Texterweiterungen mit verfügbarem Freiraum am Ende einer Lage erklärbar.

Die Perzeption der Zusammenhänge zwischen Lagen- und Blattstrukturen einerseits und dem Text andererseits ist für den Benutzer herkömmlicher Ausgaben kaum möglich, selbst wenn einer Edition eine Beschreibung des texttragenden Artefakts beigefügt ist. Bei der Textlektüre sind Informationen dieser Art in herkömmlichen Ausgaben nicht direkt verfügbar. Selbst dort, wo bisher versucht wurde, Lagenstrukturen als Begleitfunktion eines digitalen Faksimiles zu visualisieren (›Canterbury Tales Project‹, ›Parzival-Projekt‹),

wurden lediglich statische Lagenskizzen mit Einzelseiten verknüpft, sodass weder ein analytischer Zugriff noch eine dynamische Navigation und Visualisierung möglich waren.

Unser Ansatz, der im Zusammenhang mit der zur Zeit entstehenden Plattform ›Welscher Gast digital‹ (einem Kooperationsprojekt des Sonderforschungsbereichs 933 ›Materiale Textkulturen‹ und der Universitätsbibliothek Heidelberg) entwickelt wird, setzt bei der TEI-konformen Kodierung der physischen Lagen- und Blattstrukturen mittelalterlicher Handschriften des ›Welschen Gastes‹ an, die direkt im Code der Texttranskription notiert werden. Dadurch werden Abfragen möglich über die Zusammenhänge zwischen hierarchischen Strukturen des Werkes (Bücher, Kapitel, Verspaare, Verse) und materiellen Strukturen des texttragenden Artefakts. Bei der Kodierung arbeiten wir mit mehreren Typen der `<surfaceGrp>`-Elemente (`binding`, `gathering`, `bifolium`, `leaf`), die durch ihre Schachtelung die physische Zusammensetzung der Kodizes abbilden. Eventuelle Defekte (fehlende Blätter) und Ergänzungen (eingeklebte oder eingenähte Zusatzblätter) werden an entsprechenden Elementen durch Attribute und fehlende bzw. zusätzliche Knoten direkt realisiert, womit die Lagenzusammensetzung präzise beschrieben ist.

Auf der Basis dieser Kodierung (und einer aus Performanzgründen daraus generierten relationalen Struktur) entwickeln wir eine visuelle SVG-Schnittstelle, die dem Benutzer des digitalen Faksimiles eine Navigation durch die physischen Kodexstrukturen und einen davon ausgehenden Zugang zu Text und Bild ermöglicht. Der Benutzer kann dadurch eine konkrete Lage ansteuern und darin schematisch blättern.

An den Eckpunkten der physischen Struktur (Seitenumbrüche, Blattwechsel, Lagengrenzen) wird zudem das Zusammenfallen oder die Überlappung mit feinkörnigen hierarchischen Strukturen des Werkes (Vers- und Doppelversgrenzen) durch farbig differenzierte Symbole angezeigt. Schließlich visualisieren parallel mit der Lagenanzeige verlaufende Farbleisten Übereinstimmungen und Unterschiede der materiellen Einheiten des Buches und der ideellen Makrostrukturen des Werkes.

Anstelle einer separaten Beschreibung befolgt unser Ansatz die Maxime einer in die digitale Edition integrierten Veranschaulichung. Damit stehen nicht nur die abgelegten Daten der wissenschaftlichen Analyse zur Verfügung, sondern die Visualisierungen an sich erschließen neue Perspektiven auf das physische ›Gewordensein‹ und die zugrundeliegende Planung der Kodizes. Auf diese Weise vermitteln die Darstellungen nicht nur die Datenbasis, sondern sind – ganz im Sinne der digitalen Geisteswissenschaften – ihrerseits eigenständige Impulsgeber für neue Interpretationen, die wiederum zum Ausgangspunkt neuer Fragestellungen werden können.

Sonic materialization of linguistic data

The problem of sonification

Kramer Gregory (1994) in his book "*Auditory Display: Sonification, Audification, and Auditory Interfaces*" defines sonification as "use of non-speech audio to convey information or perceptualize data".
In our digital age we can store, edit and examine almost all qualities and quantities as data. Sound itself can be considered as a pure stream of information able to be modulated, transformed and analyzed in a lot of different ways.
The success of sonification occurs when the sound reveals one or more qualities of data or data reveals one or more qualities of sound. Thus, this kind of materialization of data is an interdisciplinary act which involves both the proper analysis of data and the structure of sound.
While technology provides us with a wide variety of tools, the core of the problem still exists. As this kind of interdisciplinary knowledge is hard to be combined, there aren't enough available tools which help artists to escape from an arbitrary mapping of data to sound qualities. This leads to arbitrary results both for the artist and the listener as the sonification process doesn't take advantage of neither the auditory perception properties nor sound's advantages in temporal, amplitude and frequency resolution. As a result, in most cases, sonification fails its purpose which "is to encode and convey information about an entire data set or relevant aspects of the data set." (The Sonification Handbook 2011)

Sound and linguistics

"Sonic Materialization of Linguistic Data" is a series of work and a research project aiming to provide sound artists with the tools for the proper linguistic analysis of the mined data.
In our age of constant connectivity, social media - and especially the twitter text-based platform- can be considered as a monitor corpus which evolves perpetually and it is in a process of constant change. In order to create new structures and transform this chaotic

stream of data into new material - in our case sound, it needs to be organized according to its different kind of properties- here its linguistic aspects**.** With our work “Sonic Materialization of Linguistic Data“ we provide different software modules that can perform real time linguistic analysis of data and output the result for sonification purposes.

Our software consists of different kind of modules, from which the user can choose only one or a combination of more. Here we present the *Stress Module*. The program enables the user to aggregate data from different hashtag [#] feeds on twitter in real time. The incoming data is being processed according to their linguistic features and in particular stress. The algorithm performs a series of tasks and extracts the stressed syllables of the aforementioned data. The output is a phonetic transcription code which represents each phoneme of the input twitter feed. The encoded outputted list of data also includes suggestions for the sonic mapping that occurs from data’s linguistic features and the sound’s nature. For instance, the strong syllables are a numerical output which represents a longer sound event (time envelope), whereas the weaker syllables are a numerical representation of a briefer sound event. Similar kind of optional mapping can also affect other sound features such as pitch, timbre, ADSR envelopes, modulation etc.

Stress, which can be considered as a prosodic feature, manifests itself in the speech stream in several ways. Stress patterns seem to be highly language dependent, considering that there is a dichotomy between stress timed and syllable timed languages. In stress timed languages primary stress occurs at regular intervals, regardless of the number of unstressed syllables in between, whereas in syllable timed languages syllables tend to be equal in duration and therefore are inclined to follow each other at regular intervals of time. According Halliday(1985: 272), “salient syllables occur in stress timed languages at regular intervals”. Strong syllables bear primary or secondary stress and contain full vowels, whereas weak syllables are unstressed and contain short, central vowels.

Particularly in English, which is a stress language, speech rhythm has a characteristic pattern which is expressed in the opposition of strong versus weak syllables. Stressed syllables in English are louder, but they also tend to be longer and have a higher pitch. Despite the fact that stress can be also influenced by pragmatic factors such as emphasis, our project aims to capture the natural stress pattern of English in order to

extract meaning from sound patterns too, as they will be delineated by the phonetic structure of natural language.

Presentation

For the presentation of the project we are proposing a poster with the description of how exactly the software works and what its aim is. We also would like to include a pair of headphones and a small screen (or projector) in order to have the data analysis and the sonification process in real time for the audience to experience.

References

Kramer, Gregory 1994. Auditory Display: Sonification, Audification, and Auditory Interfaces. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Vol. XVIII. Addison Wesley, Reading, Mass.

Halliday, M. A .K. 1985. An Introduction to Functional Grammar. London: Arnold.

The Sonification Handbook. Edited by Thomas Hermann, Andy Hunt, John G. Neuhoff (Eds.). Berlin: Logos Verlag Berlin 2011

Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten

Uwe Springmann (LMU München und Humboldt-Universität zu Berlin)
& Anke Lüdeling (Humboldt-Universität zu Berlin)

Unser Vortrag stellt eine Methode zur optischen Zeichenerkennung (OCR) von frühen Drucken vor, die deutlich bessere Resultate zeigt als vorherige Methoden. Mithilfe des Verfahrens können leichter und schneller Korpora mit frühen Texten erstellt werden, die dann nur noch nachkorrigiert werden müssen. Mit dem Aufbau solcher Korpora aus frühneuzeitlichen Drucken werden Basisressourcen für alle darauf aufbauenden Forschungen sprachlicher, historischer und kulturgeschichtlicher Art in den Digital Humanities bereitgestellt. Wir exemplifizieren unsere Methode mit Daten aus dem RIDGES-Korpus[1], einem diachronen Korpus, das deutschsprachige Kräutertexte enthält, die zwischen 1487 und 1870 entstanden sind.

Mangels maschineller Unterstützung ist die Erstellung eines solchen Korpus aufwändig und vom Finden korpusrelevanter gut lesbarer Vorlagen über die Einweisung von Hilfskräften in die Transkription ungewohnter Zeichen und paläographischer Konventionen und einer für die Korrektur der Transkription notwendigen breiten Sprach- und Sachkenntnis geprägt. Eine historische Orthographie und der unvermittelte Wechsel von deutschem Fraktur-Text zu lateinischen Zitaten in Antiqua sowie griechischen Wörtern erschweren die Erstellung der Transkription zusätzlich. Insbesondere die frühen Drucke (Wiegendrucke, aber auch noch Drucke aus dem 17. Jahrhundert) sind hier schwierig.

Der Traum von einer automatischen Unterstützung bei der Konvertierung früher Drucke durch allgemein zugängliche Methoden einer OCR, die ein entsprechendes Training der Erkennungsroutinen auf die verwendeten Schriften sowie die Eigentümlichkeiten des Druckbildes voraussetzen, ließ sich bisher angesichts proprietärer, einem umfangreichen Training für Außenstehende nicht zugänglicher Industrieprodukte (z.B. Abbyy Finereader[2]) bzw. zwar quelloffener und grundsätzlich trainierbarer, aber an der gestellten Aufgabe scheiternder Software (z.B. Tesseract[3]) nicht verwirklichen. Neben diesen grundsätzlichen Mängeln stand einem solchen Ansatz bisher auch der Umstand entgegen, dass ein Training eine systematisch erstellte diplomatische, d.h. am Druckbild orientierte und nicht-normalisierende Transkription von Texten voraussetzt.

Im Jahr 2013 wurden die bei Mustererkennungsaufgaben sehr erfolgreichen rekurrenten neuronalen Netzwerke mit langem Kurzzeitgedächtnis (LSTM: long short-term memory; Hochreiter & Schmidhuber 1997) durch Thomas Breuel erstmals in die OCR eingeführt und in das quelloffene, schon länger bestehende System OCRopus (Version 0.7)[4] integriert (Breuel et al. 2013). Das RIDGES-Korpus enthält Textausschnitte aus vielen Kräuterbüchern. Diese Ausschnitte (meist ca. 30 Textseiten) wurden eng diplomatisch transkribiert.[5] Das Training dieses Systems mit Hilfe der diplomatischen Transkription ("ground truth") zeigt Ergebnisse, die bei jedem der vorliegenden Texte eine Rate korrekt erkannter Zeichen (einschließlich Ligaturen, Diakritika

---

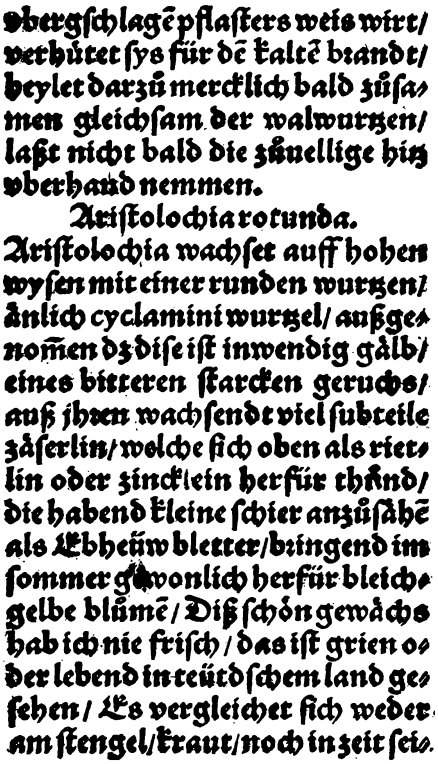
1 Ridges steht für Register in Diachronic German Science; Ziel des Ridges-Projekts ist die qualitative und quantitative Analyse der Entstehung eines deutschsprachigen wissenschaftlichen Registers. Dazu gibt es viel Literatur (so z. B. Klein 2011 oder Habermann 2003), die sich bisher aber auf nicht digital vorliegende Texte stützen musste und daher kaum für statistische Registeranalysen ausgewertet werden konnte. Das Korpus ist unter der CC-BY-Lizenz verfügbar unter http://korpling.german.hu-berlin.de/ridges/index_de.html. Das Korpus ist tief annotiert und wächst ständig.

2 http://www.abbyy.de/

3 http://code.google.com/p/tesseract-ocr/

4 http://www.ocropus.com

und Leerzeichen) von über 96% selbst ohne Verwendung von Sprachmodellen und Nachkorrekturen ergibt, während bisherige Versuche mit kommerziell erhältlicher Software bzw. Tesseract, an denen ein Autor seit Jahren beteiligt ist, kaum an die Grenze von 90% heranreichen (Springmann et al. 2013).[6]



Vbergſchlagēpflaſters weis wirt / verhůtet ſys für dē kaltē brandt / heylet darzů mercklich bald zůſa= inen gleichſam der walturtzen / laſzt nicht bald die zůuellige hitz tberhaud nemmen.
Ariſtolochia rotunda.
Ariſtolochia wachſet auff hohen wyſen mit einer runden wurtzen / ånlich cyclamini wurtzel / auſzge= nomēn dz diſe iſt inwendig gålb / eines bitteren ſtarcken geruchs / auſz jhren wachſendt viel ſubteile zåſerlin / welche ſich oben als riet= lin oder zincklein herfür thůnd / die habend kleine ſchier anzůſåhē als Ebheüw bletter / bringend im ſommer gvonlich herfür bleich= gelbe blůmē / Diſz ſchön gewåchs hab ich nie friſch / das iſt grien o= der lebend in teütdſchem land ge= ſehen / Es vergleichet ſich weder am ſtengel / kraut / noch in zeit ſei=

Adam von Bodenstein (1557): Wie sich meniglich … . Unkorrigierter OCR–Output einer vorher nicht gesehenen Seite nach Training auf 49.000 zufällig ausgewählten Textzeilen (Bild + zugeordnete ground truth) aus einer Traningsmenge von 34 diplomatisch transkribierten Seiten. Die OCR zeigt 7 verbleibende Fehler auf dieser Seite (das entspricht der durchschnittlichen Zeichenerkennungsrate auf einer Testmenge von Seiten von 99,0%).

Der Grund für diese hochgenaue Erkennungsrate liegt darin, dass OCRopus im Gegensatz zu bisherigen Methoden keine Erkennung auf Zeichenbasis über ein Template-Matching-Verfahren durchführt, bei dem ein errechnetes „mittleres“ Zeichen (das Template) auf Übereinstimmung mit einem zu erkennenden Zeichen überprüft wird, sondern jede Druckzeile durch Zerlegung in bis zu 1000 vertikale

5 Die Transkription wurde von Studierenden in verschiedenen Seminaren begonnen und später korrigiert. Daneben gibt es zwei Normalisierungsebenen und verschiedene Annotationsebenen.

6 Lediglich für die kommerzielle Software B.I.T. Alpha wurden ähnlich gute Ergebnisse für Drucke des 16. und 17. Jahrhunderts berichtet, wobei die erreichbare Genauigkeit von einem für Außenstehende kaum nachzuvollziehenden In-House-Training des kommerziellen Anbieter abzuhängen scheint (Stäcker in Federbusch et al. 2013).

Streifen schneidet, so dass jeder Buchstabe und jeder Wortzwischenraum in bis zu 30 Streifen zerlegt wird. Für jeden Streifen werden im Laufe des Trainings über den Vergleich von gedruckter Zeile mit ihrer zugeordneten Transkription die Parameter des neuronalen Netzes so eingestellt, dass mit hoher Wahrscheinlichkeit der richtige Buchstabe ausgegeben wird. Die Übersegmentierung der Buchstaben führt zu einer höheren Auflösung bei der Erkennung, so dass auch zwischen ähnlichen Zeichen wie langem s (ſ) und f problemlos unterschieden werden kann.

Die Aussicht, dass sich nunmehr jeder Interessierte Texte in hoher Genauigkeit in elektronischer Form verschaffen kann, selbst wenn die zugrundeliegenden Drucke aus früheren Jahrhunderten stammen, wird anhand unserer Erfahrungen mit dem RIDGES-Korpus hinsichtlich ihrer Voraussetzungen und des damit verbundenen Aufwandes kritisch beleuchtet. Dabei werden sowohl die Rolle des Trainings als auch der Nachkorrektur sowie die Stellung der OCR im gesamten Prozess der Korpuserstellung diskutiert. Die verwendeten Werkzeuge sowie Trainings- und Testdaten samt einer Anleitung zur Nutzung des Systems werden unter einer Open-Source-Lizenz veröffentlicht und stehen der Allgemeinheit in Kürze zur Verfügung.

Referenzen

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013). High-performance OCR for printed English and Fraktur using LSTM networks. In *Document Analysis and Recognition (ICDAR), 2013,* 683-687.

Federbusch, M., Polzin, C., & Stäcker, T. (2013). *Volltext via OCR - Möglichkeiten und Grenzen: Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Erfahrungsbericht aus dem Projekt "Helmstedter Drucke Online" der Herzog August Bibliothek Wolfenbüttel/von Thomas Stäcker*. Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Berlin.

Habermann, M. (2003). Der Sprachenwechsel und seine Folgen. Zur Wissensvermittlung in lateinischen und deutschen Kräuterbüchern des 16. Jahrhunderts. In: *Sprachwissenschaft 28*, 325–354.

Klein, W.-P. (2011). Die deutsche Sprache in der Gelehrsamkeit der frühen Neuzeit. Von der *lingua barbarica* zur *HaubtSprache*. In: Jaumann, Herbert (Hg.) *Diskurse der Gelehrtenkultur in der Frühen Neuzeit. Ein Handbuch*. de Gruyter, Berlin/New York, 465–516

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). OCR of historical printings of Latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* 71-75.

StreetartFinder – Eine Datenbank zur Dokumentation von Kunst im urbanen Raum

1. Einleitung

Streetart ist ein Sammelbegriff für Graffitis, schablonen- oder handgezeichnete Bilder, Poster, Aufkleber aber auch Installationen und Skulpturen im öffentlichen Raum (Reinecke, 2012, S. 17). Eine Vielzahl bestehender Publikationen zeigt, dass Streetart auch als wissenschaftliches Forschungsobjekt zunehmend an Relevanz gewinnt (vgl. etwa Klitzke & Schmidt, 2009; Philipps & Barlösius, 2014; Reinecke 2012; Wacławek, 2012; u.v.a.). Mit dem *StreetartFinder*[1] wurde ein Tool geschaffen, das erlaubt, diese Kunstwerke im urbanen Raum in digitaler Form zu dokumentieren, und so eine Datenbank für weitere Forschung in diesem Feld zur Verfügung zu stellen.

2. Konzeption und wesentliche Funktionen des StreetartFinder

StreetartFinder wurde mit gängigen Web-Technologien umgesetzt, und steht sowohl als *Desktop*- als auch als *Mobile*-Variante zur Verfügung. Nutzer können Fotos von Streetart-Objekten auf die Webseite laden, und dabei Metadaten wie „Name des Uploaders", „Tags / Schlagworte" sowie einen optionalen „Beschreibungstext" angeben. Zusätzlich sind die Uploader angehalten, ihr jeweiliges Objekt zu klassifizieren, wobei derzeit folgende Optionen zur Auswahl stehen: *Graffiti*, *Stencil*, *Painting*, *Paste-Up*, *Installation*, *Sonstiges*. Zuletzt können die Nutzer optional den Standort der Streetart durch Markierung in einem *GoogleMaps*-Ausschnitt vornehmen.

Streetart kann gefiltert nach Städten, Kategorie, Bewertung oder Anzahl der Views dargestellt werden (vgl. Abb. 1). Die Besucher der Seite können die bestehenden Streetart-Fotos bewerten oder aber ein Objekt als „nicht länger vorhanden" melden.

[1] http://streetartfinder.de/, zuletzt aufgerufen am 23.10.2014

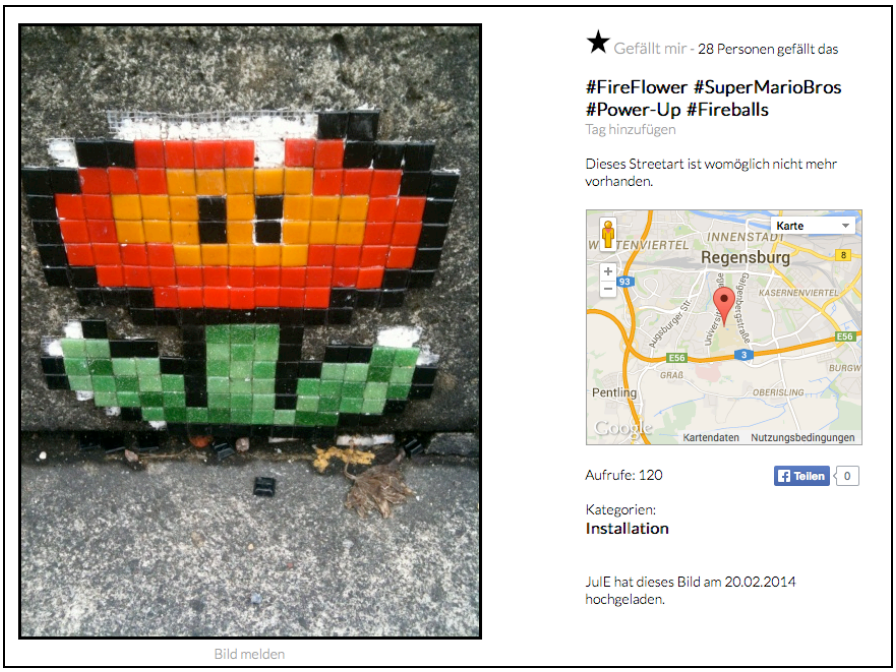


Abbildung 1: Darstellung eines Streetart-Objekts mit verschiedenen Metadaten auf der Webseite.

Eine wesentliche Funktion stellt außerdem die Visualisierung verschiedener Streetart-Objekte auf einer interaktiven Karte dar, die mit Hilfe der *GoogleMaps API*[2] realisiert wurde (vgl. Abb. 2). Auf dieser Karte kann etwa dargestellt werden wo sich welche Art von Streetart befindet.

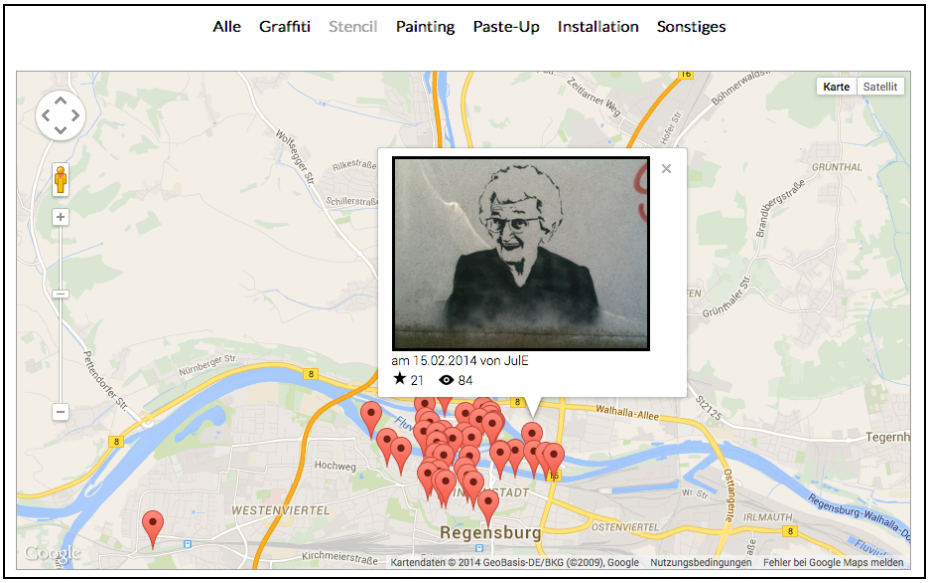


Abbildung 2: Interaktive GoogleMaps-Visualisierung der Streetart-Objekte.

[2] https://developers.google.com/maps/, zuletzt aufgerufen am 23.10.2014

3. Aktueller Stand der Datenbank

Dieser Abschnitt beschreibt den aktuellen Stand (22. Oktober 2014) der Streetart-Datenbank seit der Live-Schaltung am 15. Februar 2014. Seither haben 1.845 unterschiedliche Nutzer (davon 38,5% wiederkehrende Besucher) die Seite besucht. Es befinden sich momentan 10 deutsche Städte in der Datenbank, wobei für Köln (176) und Regensburg (149) mit Abstand am meisten Streetart-Objekte zu verzeichnen sind. Insgesamt gibt es aktuell 475 Objekte in der Datenbank, von denen 442 einer eindeutigen Streetart-Kategorie zugeordnet wurden. Für die restlichen 33 wurden jeweils zwei Kategorien von den Uploadern vergeben. Tabelle 1 zeigt die Häufigkeitsverteilung der einzelnen Kategorien (inklusive der 33 Doppelkategorisierungen). Das Material wurde von insgesamt 71 unterschiedlichen Nutzern hochgeladen.

Kategorie	Anzahl	Prozentualer Anteil
Graffiti	230	45%
Paste-Up	102	20%
Stencil	80	16%
Installation	36	7%
Sonstiges	33	7%
Painting	27	5%

Tabelle 1: Häufigkeitsverteilung der einzelnen Kategorien.

Von den 475 Bildern wurden in den letzten 7 Monaten 35 von Nutzern als „nicht mehr vorhanden“ gemeldet. Alle Bilder zusammen wurden insgesamt 23.330 mal betrachtet. Außerdem wurden insgesamt 892 Tags (522 *unique tags*) vergeben, deren Hauptfunktion die Beschreibung des *Orts* (Köln, Ehrenfeld, Jugendzentrun, etc.), des *Datums*, des *Künstlers* (stencilove, riq, unbekannt, etc.), der *Streetart-Kategorie* (paste-ups, graffiti, etc.) oder aber des *Inhalts* bzw. des *Motivs* (Katze, Banane, Kurt Cobain, etc.) ist (vgl. Abb. 4).

Abbildung 4: Wordcloud der vergebenen Hashtags. Die 5 häufigsten Tags sind dabei *ehrenfeld* (66), *köln* (66), *unbekannt* (19), *mann* (11) und *bürgerzentrum* (10).

4. Fazit

Die bisherigen Nutzerzahlen zeigen, dass *StreetartFinder* von den Benutzern als Tool zur Dokumentation von Kunst im urbanen Raum gut angenommen wird. Es entsteht auf diese Weise eine einzigartige Datenbank, in der neben Fotografien der jeweiligen Streetart auch Metadaten mitgespeichert werden, die verschiedene soziologische, kultur- und medienwissenschaftliche Fragestellungen erlauben, z.B.:

- Welcher Typ von Streetart kommt am häufigsten vor?
- Gibt es im Laufe der Zeit Trends für bestimmte Typen bzw. gibt es Ballungsgebiete, in denen vor allem ein bestimmter Typ von Streetart vorherrscht?
- Wie lange ist die durchschnittliche Lebensdauer von Streetart, und gibt es einen Zusammenhang mit dem Ort / Typ?
- Was sind die Hauptfunktionen von Streetart?

Neben Überlegungen zur weiteren Verbreitung des Tools, vor allem auch in anderen Städten, planen wir zusätzlich einen Web-Zugang zu allen relevanten Metadaten für interessierte Wissenschaftler.

5. Literaturverzeichnis

Klitzke, K. & Schmidt, C. (2009). Street Art: Legenden zur Straße. Berlin: Archiv der Jugendkulturen.

Philipps, A. & Barlösius, E. (2014). Zur Sichtbarkeit von Street Art in Flickr. Methodische Reflexionen zur Zusammenarbeit von Soziologie und Informatik. In Abstracts of the Dhd 2014, Passau.

Reinecke, J. (2012). Street-Art: eine Subkultur zwischen Kunst und Kommerz. Bielefeld: Transcript Verlag.

Wacławek, A. (2012). Graffiti und Street Art. Berlin: Deutscher Kunstverlag.

Virtuelle Rekonstruktion des Regensburger Ballhauses

1. Projektkontext und wesentliche Ziele

Im Rahmen einer Vortragsreihe zum 350-jährigen Reichstagsjubiläum in der Stadt Regensburg wurde in Ergänzung zum Thema „Das Jahrhundert des Dramas und der Komödien: Blüte des Regensburger Theaterlebens“[1] eine virtuelle 3D-Rekonstruktion des heute nicht mehr vorhandenen Regensburger Ballhauses am Ägidienplatz erstellt. Die 3D-Rekonstruktion stellt einerseits das Innenleben des Ballhauses dar und liefert andererseits textuelle Informationen zu interessanten Objekten. Die Rekonstruktion kann mit Hilfe der Virtual Reality-Brille *Oculus Rift* interaktiv exploriert werden. Das Projekt ist damit im Kontext der Museumspädagogik anzusiedeln (vgl. Flügel 2009; Wagner 2007; Waidacher & Raffler 2005).

Umfangreiche Informationen zur Geschichte des Regensburger Ballhauses am Ägidienplatz finden sich in Meixner (2008): Die Baugeschichte des Ballhauses beginnt bereits im Jahre 1603, als zunächst hölzerner Bau, der vornehmlich für Sportereignisse genutzt wurde. Dieses Gebäude wurde schließlich im Jahre 1736 durch einen Neubau ersetzt, der dann auch stärker für Theateraufführungen genutzt wurde. In seiner Hochzeit war das Ballhaus am Ägidienplatz das kulturelle Zentrum des Immerwährenden Reichstags in Regensburg. Durch die Eröffnung des Theaters am Bismarckplatz im Jahre 1804 verlor das Ballhaus langsam an Bedeutung. In der Folge verfällt das Gebäude zunehmend und wird schließlich im Jahre 1922 abgerissen.

Hauptziele des Projekts

- Rekonstruktion des Innenraums des Ballhauses (1736-1922) mit der Präsentation einer barocken Kulissenbühne
- Interaktion durch *Virtual Reality*-Umsetzung statt statische Präsentation eines 3D-Modells
- Umsetzung einer zusätzlichen pädagogische Komponente durch das Augmentieren weiterführender Information über das Ballhaus im virtuellen 3D-Raum

2. Unvollständigkeit der Quellenlage als wesentliche Herausforderung

Die exakte Gestaltung des Innenraums ist sehr schwer zu rekonstruieren, da nur wenige Quellen aus dieser Zeit überliefert wurden. Soweit Quellen vorliegen, sind diese zumeist Skizzen von Zeitzeugen und Dokumente aus dem Hofarchiv Thurn und Taxis (vgl. Abb. 1).

[1] Referentin: Hannah Ripperger; weitere Informationen zum Vortrag im Programmheft zur Vortragsreihe (S. 17), online verfügbar unter: https://www.regensburg.de/sixcms/media.php/121/der-reichstag-in-45-minuten.pdf, zuletzt abgerufen am 27.10.2014.

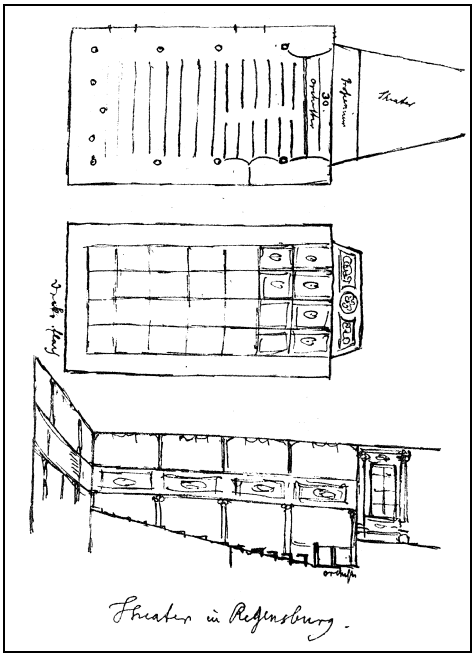


Abbildung 1: Skizzen von Friedrich Gilly, 1798 (Bildquelle: Meixner, 2008, S. 126).

Weitere Herausforderungen ergeben sich durch historische Maßeinheiten (z.B. „Regensburger Schuh“ statt Meter), oder durch Skizzen ohne Maßstab und Maßangaben. Zudem gibt es oftmals keine Abgrenzung zwischen verschiedenen Bauphasen des Hauses. Um diese unvollständigen oder fehlenden Informationen zu ergänzen, wurden schließlich Vergleiche zu anderen Theaterräumen in Deutschland angestellt (etwa Gotha und Passau), und allgemeine Stilmerkmale aus der Kunstgeschichte umgesetzt (vgl. Meixner, S.128 f). Diese heterogene Quellenlage ist in Abbildung 2 zusammengefasst dargestellt:

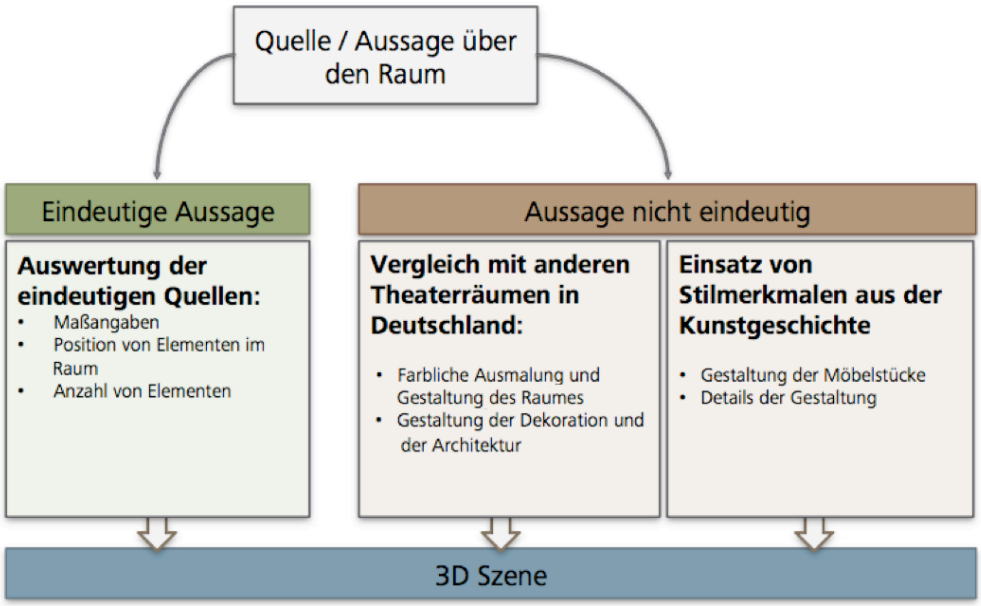


Abbildung 2: Überblick zur heterogenen Quellenlage, die als Grundlage für die Rekonstruktion verwendet wurde.

3. Technische Umsetzung

Die Rekonstruktion wurde mit Hilfe des 3D-Modellierungstools *Blender*[2] umgesetzt. Zu Beginn wurde die Geometrie des Theatersaals sowie dessen Möblierung modelliert. Zusätzlich wurde mit Hilfe der Bildbearbeitungssoftware Photoshop und einigen Referenzbildern die farbliche Gestaltung des Raums nachgebildet (vgl. Abb. 3).

Abbildung 3: Ausschnitt aus der virtuellen Rekonstruktion des Regensburger Ballhauses am Ägidienplatz.

Nach der Rekonstruktion des Raums folgte der Export in die Game-Engine *Unity3D*[3]. Dort wurde die Geometrie mit Farbinformationen und mit Oberflächenstrukturen ausgestattet. Danach wurden alle weiteren Interaktionsmöglichkeiten implementiert. Dabei wurde auf das *Oculus Rift SDK*[4] zurückgegriffen: Zwei Kameras rendern die Szene und verkrümmen das gerenderte Bild, um es gemäß der Linsenkrümmung im *Head Mounted Display* (HMD) korrekt anzeigen zu können (vgl. Abb. 4).

[2] http://www.blender.org/, zuletzt abgerufen am 27.10.2014
[3] http://unity3d.com/, zuletzt abgerufen am 27.10.2014
[4] http://www.oculus.com/, zuletzt abgerufen am 27.10.2014

Abbildung 4: 3D-Szene aus Perspektive der Virtual-Reality-Brille *Oculus Rift*.

Mithilfe eines einfachen Game-Controllers kann sich der Nutzer im Raum bewegen. Ferner kann über die Bewegung mit dem Kopf die Rotation der Kamera bestimmt werden. Über den Mittelpunkt des Bildschirms und entsprechendes *Raycasting*[5] in den Raum wird überprüft, welches Objekt der Nutzer gerade anschaut. Je nachdem können verschiedene Informationen über die Elemente im Raum, etwa die Kulissenbühne (vgl. Abb. 5), mit einem Tastendruck über den Game-Controller abgerufen werden.

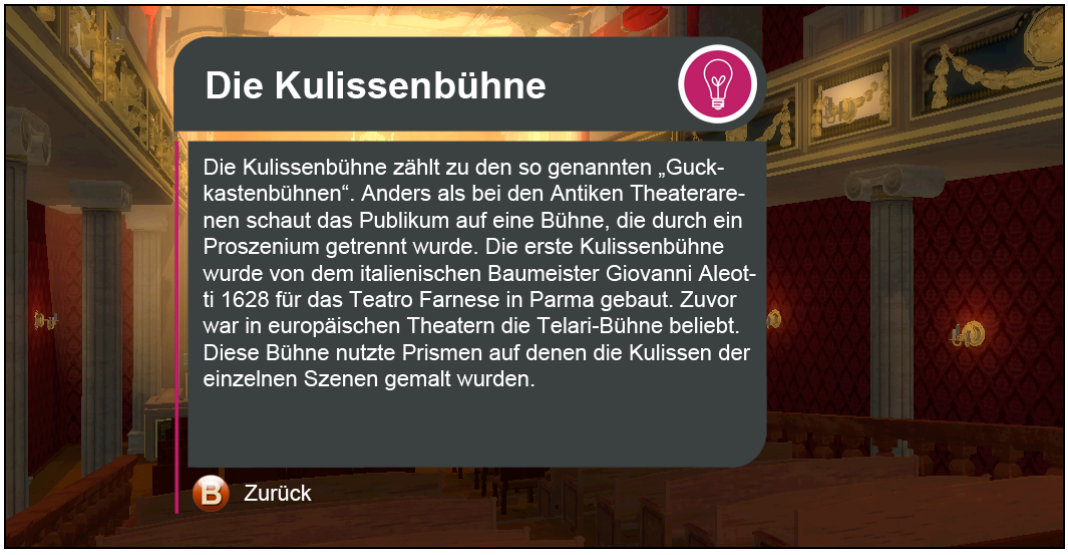


Abbildung 5: Informationsanzeige zur Funktionsweise der Kulissenbühne.

[5] *Raycasting* ist ein Begriff aus der Computergrafik. Vereinfacht gesagt „tastet“ dabei ein virtueller Strahl den dreidimensionalen Raum nach Objekten ab, die dann bei Bedarf aktiviert werden können.

4. Demonstration

Ein Demo-Video der virtuellen Rekonstruktion ist verfügbar unter:

- http://dhregensburg.wordpress.com/2014/07/25/virtuelle-rekonstruktion-regensburger-ballhaus/

Im Falle der Annahme des Abstracts ist geplant, die Rekonstruktion im Rahmen der Poster-Präsentation live mithilfe einer Virtual-Reality-Brille (*Oculus Rift*) zu demonstrieren.

5. Literaturverzeichnis

Flügel, K. (2009). Einführung in die Museologie. 2., überarb. Aufl. Darmstadt: WBG.

Meixner, C. (2008). Musiktheater in Regensburg im Zeitalter des Immerwährenden Reichstages. Sinzig: Studio Verlag.

Wagner, E. (2007). Museum, Schule, Bildung. Aktuelle Diskurse, innovative Modelle, erprobte Methoden. München: kopaed.

Waidacher, F. & Raffler, M. (2005). Museologie - knapp gefasst. 1. Aufl. Stuttgart: UTB.

Poster | Abstract

Statistisch gestützte Visualisierung von Informationsgliederungen in den Bundestagsreden

Dr. Zakharia Pourtskhvanidze, *pourtskhvanidze@em.uni-frankfurt.de*
Institut für Empirische Sprachwissenschaft. Goethe-Universität Frankfurt/M

Der handlungsbezogene Aspekt der Sprache wird besonders deutlich in den öffentlichen Auftritten der Politiker. Neben der für die gesprochene Sprache spezifischen Verwendung von lexikalischen Mitteln erscheinen in den Reden zuhörerorientierte Einsetzung von syntaktischen Konstruktionen (Z.B. Anrede, Rhetorische Fragen) und auf die Interaktion abgestimmte Realisierung von pragmatischen kommunikativen Strategien (Z.B. Gesichtserhaltung).

Der linguistisch interpretierte Begriff *Hervorhebung* bezüglich der Gestaltung des informationellen Gehalts einer Äußerung eignet sich besonders als ein Ausgangspunkt für die Verbildlichung (Visualisierung) besonders prominenter Informationseinheit im Vergleich zur informationell neutralen Einheit eines gesprochenen Diskurses.

Die **empirische Basis** der auf dem Poster beschriebenen Analyse stellt das Plenarprotokoll (Stenographischer Bericht) der 2. Sitzung des Deutschen Bundestages am 18. November 2013 zum Thema *die Abhöraktivitäten der NSA und die Auswirkungen auf Deutschland und die transatlantische Beziehungen* dar (ca. 20.000 Token bei 17 Sprechern aus 5 Parteien).

Aufgrund der linguistischen Analyse von Texten wird eine **Ranking-Tabelle** der grammatischen Instrumente der Informationsgliederung erzeugt. In der Tabelle bekommen die einzelnen Instrumente (Partikeln, Vorfeldbesetzung, Spalt-Satzstruktur, Satzarten, Referentielle Bezüge (Anapher bzw. Katapher) etc.) in den Zahlen ausgedruckte Werte (-15 min. … 0 … 75 max.). Die fokussierende (rhematisierende) Instrumente bekommen tendenziell hohe Werte (daher stärker hervorgehoben), wogegen die topikalisierte (thematisierende) Instrumente bekommen tendenziell niedrige Werte (daher flachere Visualisierung).

Der *erste Schritt* des Visualisierungsvorgangs sieht die automatische **Ersetzung** des jeden in der Ranking-Tabelle notierten Tokens resp. Skopus im Protokoll-Text mit der entsprechenden Zahl, während der Rest von Tokens im Text automatisch mit dem Wert „0“ ersetzt wird. Die Texte werden in CSV-Dateiformat abgebildet und damit für unterschiedliche Statistik-Analysen endgültig aufbereitet.

Die Visualisierung der Daten erfolgt im *zweiten Schritt* durch eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken **R**.

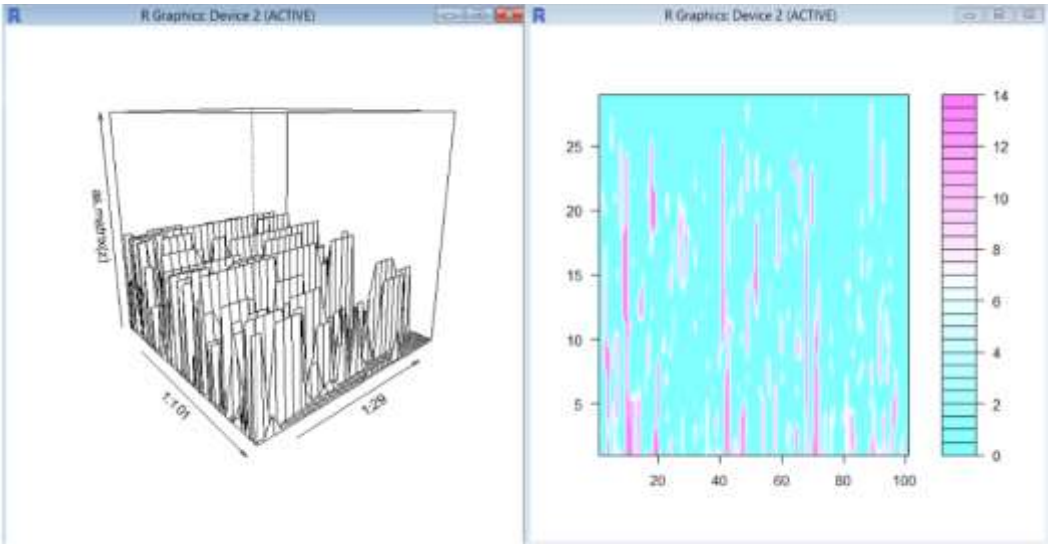


Abb. 1. Visualisierung der Rede des Abgeordneten G. Gysi (Die Linke) mit R-Funktionen: *persp* (links) und *filled.contour* (rechts)

Im Kontext der Auffassung *Make your data tell a story* eröffnet die Visualisierung von Informationsgliederungen die Möglichkeiten die generierten Bilder auf die Zeitschiene zu ordnen und zwar limitiert auf eine Person oder einem Themas.

- Gibt es mögliche Kausalitätstendenzen zwischen den „Informationslandschaften“ der Reden und dem Thema desselben?
- Lässt sich ein Dominanz-Muster für die bestimmten sprachlichen Instrumente der Informationsgliederung abhängig vom Thema etablieren?
- Gibt es individuelle Muster von informationellen Hervorhebungen?
- Lässt sich die Entwicklung des Redestils eines Sprechers in den Visualisierungen von Informationseinheiten über die Zeit dokumentieren?

Eine disziplinübergreifende Analyse mit den Bezügen auf die Gender- und Meinungs-Forschungen ist denkbar.

Vorgesehen ist die Erweiterung der empirischen Daten in Richtung der Vielfalt der Genres und die Einsetzung von alternativen Visualisierungssoftware (z.B. RStudio).

Das weitere Forschungsvorhaben wird gegenwärtig an der Goethe-Universität für die Aufbau einer integrativen Kooperationsplattform zwischen den Fachbereichen „Sprach- und Kulturwissenschaften“ und „Informatik und Mathematik“ bedacht.

Vom Luftbild zum 3D-Modell

zum Einsatz von unbemannten Luftfahrzeugen in der Archäologie des Vorderen Orients

Benjamin Glissmann, Jason Herrmann[1], Matthias Lang[2]

[1] Institut für die Kulturen des Alten Orients der Universität Tübingen, [2] eScience-Center der Universität Tübingen

Unbemannte Luftfahrzeuge – meist als Drohnen bezeichnet – entwickeln sich immer mehr zu einem wichtigen Dokumentationswerkzeug in der archäologischen Feldforschung. Waren die Geräte noch vor wenigen Jahren äußerst kostspielig sowie schwierig zu Fliegen und zu Warten, lassen sich heute bereits für niedrige dreistellige Beträge Drohnen erwerben, deren Betrieb auch durch einen Laien schnell zu erlernen ist.

In einem Großteil der Projekte dienen die UAVs (unmanned aerial vehicle) meist zur Aufnahme von Luftbildern und Filmen zu reinen Visualisierungs- und Präsentationszwecken, eine Integration der Geräte in den eigentlichen Dokumentations- und Forschungsprozess findet meist nicht statt.

Integration von unbemannten Luftfahrzeugen in die archäologische Dokumentation

In diesem Beitrag soll diskutiert werden, wie diese Systeme sinnvoll in den Workflow der Dokumentation integriert werden können und welchen Mehrwert sie gegenüber herkömmlichen Methoden besitzen und diese in Teilen obsolet machen.

Dies soll am Beispiel eines Surveys im kurdischen Teil des Irak aufgezeigt werden, der seit 2013 an der Universität Tübingen durch Peter Pfälzner durchgeführt wird. Ziel des Projektes ist die Identifikation und die Untersuchung von künstlichen Siedlungshügeln, sogenannten Tells, die meterhoch aus der Landschaft ragen und durch eine stetige Besiedlung desselben Ortes teils über mehrere Jahrtausende entstehen.

Besonders Beachtung sollen in diesem Beitrag die spezifischen Anforderungen der Archäologie des Vorderen Orients sowie die Auswirkungen der extremen äußeren Bedingungen auf die Arbeiten finden.

Dokumentation der Fundstellen

Eine der wichtigsten Aufgaben bei der Dokumentation der Tells liegt in einer genauen geographischen Verortung und einer möglichst präzisen Kartierung des gesamten Befundes. Diese Informationen können dann in einem nächsten Schritt in elektronischen Geoinformationssystemen verwaltet und analysiert werden, um beispielweise Rückschlüsse zur Siedlungsstruktur einer ganzen Region in einer spezifischen Epoche zu ermöglichen.

Aufgrund der großen Anzahl von Fundstellen im Untersuchungsareal muss diese Erfassung schnell und effizient erfolgen. Zusätzlich erschwert werden die Arbeiten durch die Teilweise große Entfernung der einzelnen Sites, durch das häufig unwegsame Gelände und durch das vollständige Fehlen von bekannten Vermessungspunkten. Um diesen Problemen zu begegnen wurden verschiedene Ansätze evaluiert.

Herkömmliche Vermessungsmethoden

Eine tachymetrische Erfassung der Befunde ist aufgrund des Fehlens korrespondierender Festpunkte mit bekannter Koordinate nur in einem lokal beschränkten Vermessungsnetz möglich. Zudem ist bei dieser Methode aufgrund der Morphologie der Tells ein häufiges Umstationieren des Tachymeters

erforderlich, da stets eine Sichtverbindung zwischen Gerät und Winkelprisma notwendig ist. Darüber hinaus muss der Vermesser jeden Punkt der zu vermessenden Struktur anlaufen, um punktgenaue Daten zu erheben. Aufgrund der Geländestruktur war dies bei einem Großteil der zu vermessenden Sites nicht zu gewährleisten.

Somit muss diese Methode als ineffizient und aufgrund der fehlenden absoluten Verortung in einem standardisierten Koordinatensystem als ungeeignet angesehen werden.

Als zweite Methode wurde die Vermessung der Sites mittels GPS angedacht. Dies erlaubt zwar eine Vermessung in einem absoluten Koordinatensystem, nach wie vor muss jedoch jeder Punkt zeitaufwendig angelaufen werden, um eine präzise Kartierung zu gewährleisten.

Als weiterer Nachteil beider Methoden muss zudem angeführt werden, dass sie lediglich punktförmige Daten produzieren, die in einem zweiten Schritt erst aufwendig zu fertigen Karten zusammengeführt werden müssen.

Einsatz von luftgestützter Photogrammmetrie

Aus diesen Gründen haben wir uns dazu entschieden, ausschließlich luftgestützte Photogrammmetrie zur Vermessung der Siedlungshügel zu verwenden, die eine effiziente und präzise Erfassung erlaubt.

Setup und Einsatz der Drohne

Grundlage dieser Methoden sind Luftbilder sowie mittels GPS eingemessene Passpunkte. Zur Aufnahme der Bilder kam ein Quadrokopter des Typs *DJI Phantom Vision +* zum Einsatz, der sich per GPS positioniert und mit einer Funkfernsteuerung sowie einer Smartphone-App gesteuert wird. In dieser App sind ein Livebild der eingebauten Kamera sowie Telemetriedaten wie Flughöhe und Akkustand verfügbar. Aufgrund der hellen Umgebung waren diese Daten jedoch nur bedingt ablesbar und eine Positionierung der Drohne musste weitestgehend über einen Sichtkontakt zum Gerät selber erfolgen. Hier sollen in Zukunft verschiedene Schutzvorrichtungen vor zu starker Sonneneinstrahlung evaluiert werden. Die Aufnahme der Bilder erfolgt intervallgesteuert, so dass sich der Pilot lediglich um die Positionierung der Drohne über dem Grund zu kümmern hat. Als geeignete Flughöhe haben sich ca. 50 Meter erwiesen, die sowohl eine große Überlappung der Bilder als auch eine entsprechende Auflösung gewährleistet. Aufgrund der GPS-Positionierung der Drohne kann diese Flughöhe ohne Mühe konstant gehalten werden.

Um ein späteres Zusammenfügen der Bilder positionsgenau zu ermöglichen sind auf dem Grund Passpunkte verteilt, die mittels eines GPS-Gerätes genau eingemessen werden. Die Bilder der Drohne tragen zwar ebenfalls GPS-Informationen, diese zeigen jedoch die genaue Position der Drohne bei der Aufnahme an und nicht die Position des aufgenommenen Areals.

Als problematisch hat sich die Auswirkung großer Hitze auf die Elektronik der Drohne erwiesen, Abbrüche des Funkkontaktes zwischen Fernsteuerung und Drohne waren die Folge. Ein Schutzmechanismus lässt in diesem Fall den Kopter jedoch zu seinem Startpunkt zurückkehren, so dass hier ein sicherer Betrieb stets gewährleistet ist.

Erstellung von kartographischen Informationen aus Luftbildern mittels SFM

In einem zweiten Schritt müssen nun die Luftbilder prozessiert und mit den GPS-Koordinaten verbunden werden. Hierzu wird die Software Agisoft Photoscan Pro verwendet, die eine weitestgehend automatisierte Verarbeitung der Bildinformationen mittels Structure from Motion (SFM) zu einem fertigen 3D-Modell erlaubt. Hierzu werden in den Bildern gemeinsame Strukturen wie Eckpunkte oder Linien durch die Software erkannt und in einem dreidimensionalem Raum verortet. Hierzu ist eine Überlappung der verwendeten Bilder notwendig. Der Methode liegen dieselben

Prozesse zu Grunde, die das menschliche Gehirn zur Konstruktion dreidimensionaler Informationen verwendet.

In einem nächsten Schritt werden diese Passpunkte trianguliert und vernetzt. Dieses sogenannte Mesh kann nun wiederum mit einer photorealistischen Textur versehen werden, die ebenfalls aus den Luftbildern abgeleitet wird. Die auf den Bildern zu identifizierenden Passpunkte müssen mit den GPS-Koordinaten verbunden werden, um das Modell mit absoluten Größeninformationen zu verbinden. Ohne diese Zusatzinformationen ist das Modell maßstabslos und kann kaum sinnvoll genutzt werden.

Um nun aus diesem Modell eine Karte zu erzeugen, ist ein Export in ein Digital Elevation Modell (DEM) notwendig. Dieses besteht aus einem Rasterbild, in dem die Höheninformationen der einzelnen Pixel durch unterschiedliche Grautöne repräsentiert werden, die mit absoluten Höhenmetern verbunden sind. Die Ausdehnung sowie die Lage des DEMs sind durch die absolute Koordinaten in einem vorher zu bestimmenden Koordinatensystem definiert, die in einer Zusatzdatei abgelegt werden. Das DEM und diese Zusatzdatei können nun mit nahezu jedem Geoinformationssystem eingelesen und weiterverarbeitet werden. Zur Erstellung einer topographische Karte dient nun eine farbliche Repräsentation des DEM sowie hieraus abgeleitete Konturlinien.

Neben einem Export als DEM bedient Photoscan auch weitere Datenformate wie OBJ oder Collada, die eine Weiterverarbeitung des Modells in beliebigen 3D-Umgebungen gestattet.

Als Nachteil der hier präsentierten Methode muss das Fehlen jeglicher Features gelten, die sich im 3D-Modell nicht erkennen lassen. So sind Straßen, Wege sowie unterschiedliche Landnutzungen im DEM nahezu unsichtbar. Photoscan erlaubt jedoch neben dem Export als DEM auch gleichzeitig die Erstellung eines Orthofotos, das eine koordinatenrichtige, rechtwinklige Aufsicht auf das Modell darstellt. Dieses Orthofoto kann ebenfalls im Geoinformationssystem verarbeitet werden und die im Geländemodell unsichtbaren Informationen lassen sich dort nun digitalisieren und gemeinsam mit der topographischen Karte visualisieren.

Erfahrungen aus Ausblick

Die hier vorgestellte Methode hat sich bei der Dokumentation der Siedlungshügel überaus gut bewährt und muss als deutlich effizienter als die herkömmlichen Vermessungsmethoden gelten. Diese Aussage lässt sich jedoch nicht verallgemeinern. So ist eine derartige Vorgehensweise in einer dichten Vegetation deutlich schwieriger einzusetzen, als in der wüstenähnlichen Landschaft des Nordirak. Ebenso hat sich die dünne Besiedlung des Areals als großer Vorteil erwiesen, in einem dichtbesiedelten Gebiet ist der Einsatz solcher Systeme aufgrund von Aspekten der Sicherheit durchaus fragwürdig.

Als weiterer großer Vorteil der Methode kann die leichte Weiterverarbeitung der Daten in Geoinformationssysteme sowie 3D-Anwendungen angesehen werden, wodurch in einem einzelnen Arbeitsschritt vielfältige Anforderungen erfüllt werden können.

Die erreichbare Präzision hängt sehr stark von der Auflösung der Kamera, der Flughöhe sowie der Genauigkeit der Vermessung der Passpunkte ab. Für die großflächige Vermessung der Tells ist die erreichte Genauigkeit jedoch absolut ausreichend. In der kommenden Kampagne soll nach Möglichkeit einer zurzeit im Test befindliche, deutlich größere Drohne mit einem hochauflösenden Kamerasetup verwendet werden, das eine Bodenauflösung von 1-2 Zentimetern erlaubt und auch eine Kartierung kleinteiliger Befunde erlaubt.