Digitale Daten in den Geisteswissenschaften. Interdisziplinäre Perspektiven für semantische und strukturelle Analysen

Veranstalter: Arbeitskreis „Digital Humanities Munich“ (dhmuc) an der Bayerischen Akademie der Wissenschaften
Datum, Ort: 28.01.2016–29.01.2016, München
Bericht von: Peter Fleer, Dienst Historische Analysen, Schweizerisches Bundesarchiv, Bern

Der Arbeitskreis „Digital Humanities Munich“ (dhmuc)[1] widmete seinen ersten Workshop 2016 dem Thema „Digitale Daten in den Geisteswissenschaften. Interdisziplinäre Perspektiven für semantische und strukturelle Analysen“.[2] Insgesamt 14 Vorträge erörterten aktuelle Forschungen und Infrastrukturen im Bereich der maschinellen Textanalyse.

Folgende Institutionen zeichneten für die Organisation des Workshops verantwortlich: dhmuc – Arbeitskreis „Digital Humanities Munich“, das Institut für Computerlinguistik Universität Zürich[3], das Historische Seminar der Ludwig-Maximilians-Universität München[4], die IT-Gruppe Geisteswissenschaften (ITG) der Ludwig-Maximilians-Universität München[5] sowie die Bayerische Akademie der Wissenschaften[6].

Einführung

ECKHART ARNOLD (München), MARK HENGERER (München), NOAH BUBENHOFER (Zürich) und CHRISTIAN RIEPL (München) machten klar, dass das Vorhandensein grosser Mengen an digitalen Textkorpora den Geisteswissenschaften neue Zukunftsperspektiven eröffnet, gleichzeitig aber auch neue Herausforderungen an die Disziplinen stellt. Der Brückenschlag zwischen Philologie und Informationstechnologie erfordert von den Forschenden ein hohes Mass an Technikaffinität. Das Verständnis von Text als Daten und dessen interaktive graphische Visualisierung verändert die hermeneutischen Herangehensweisen und Methoden.

Panel 1: Korpora

Im Eröffnungsreferat stellte PETER MAKAROV (Zürich) sein PhD Projekt zum Thema „Towards automated content event analysis: Mining for protest events“ vor. Es ist Teil des POLCON Projekts unter Professor Hanspeter Kriesi.[7] Das Projekt bewegt sich zwischen Politikwissenschaft und Computerlinguistik. Ziel ist, herauszufinden, inwiefern moderne Natural Language Processing Technik die thematische Extraktion von Daten aus dem Internet unterstützen kann. Konkret geht es in der Studie um die Extraktion von News-Texten im Zusammenhang mit Öffentlichen Protesten.

Methodisch setzt das Projekt auf *Machine-Learning*-Verfahren. Hierzu müssen geeignete Klassifizierungsmodelle erstellt und Entwicklungszyklen implementiert werden, mit deren Hilfe Trainings-Korpora erstellt werden können. In einem iterativen Prozess werden diese Trainingsdaten annotiert und verbessert. Dabei hat sich gezeigt, dass linguistische Standardmodelle zur Annotation nicht geeignet waren, um die Identifikation von Textstellen, die Protest indizierten, zu verbessern. Daher werden vereinfachte, besser auf die konkrete Fragestellung zugeschnittene Annotationssysteme entwickelt.

DANIEL KNUCHEL (Zürich) referierte unter dem Titel „HIV/AIDS diskurslinguistisch – ein multimediales Korpus“ über sein Promotionsprojekt. In diesem analysiert er, welches Wissen heute zum Thema HIV/AIDS in der Öffentlichkeit zirkuliert. Dazu baute der Referent ein Korpus aus unterschiedlichen Quellen (Massenmedien, Selbsthilfe-Blogs, Social Media) auf. Er wies darauf hin, dass bei solchen Vorhaben nebst den konzeptuellen und technischen Fragen die rechtlichen Bedingungen zur Datennutzung frühzeitig geklärt werden müssen. Wichtig sei zudem, dass von Anfang an auf Nachhaltigkeit geachtet werde, um spätere Nutzungen der Daten zu ermöglichen.

[1] <http://dhmuc.hypotheses.org/uber>; alle URLs wurden überprüft am 28.04.2016.
[2] <http://dhmuc.hypotheses.org/workshop-digitale-daten#Bubenhofer>.
[3] <http://www.cl.uzh.ch/de.html>.
[4] <http://www.geschichte.uni-muenchen.de/index.html>.
[5] <http://www.itg.uni-muenchen.de/index.html>.
[6] <http://www.badw.de/>.
[7] <http://www.eui.eu/Projects/POLCON)>.

MAX HADERSBECK (München) berichtete in seinem Vortrag über die Erfahrungen mit der FinderApp „WittFind“.[8] Die webbasierte Applikation steht Forschenden nun seit vier Jahren zur Verfügung, um den Open Access zugänglichen Teil des Nachlasses von Ludwig Wittgenstein nach Wörtern, Phrasen, Sätzen und semantischen Begriffen zu durchsuchen. Sie setzt dazu regelbasierte computerlinguistische Verfahren ein. WittFind zeichnet sich dadurch aus, dass die gefundenen Textstellen zusammen mit dem Faksimileextrakt dargestellt werden und eine Überprüfung anhand des Originaltexts jederzeit möglich ist.

Unter anderem am Beispiel des englischen Neologismus „cherpumple“[9] stellte SUSANNE GRANDMONTAGNE (München), das System „NeoCrawler“[10] vor. NeoCrawler verfolgt die Entstehung und Verbreitung von Neologismen im Internet. Die Ergebnisse werden automatisiert für weitere linguistische Analysen aufbereitet. Zudem stellt das System Zeitreihenverlaufsanalysen zur Verfügung. Eine Benutzeroberfläche unterstützt die manuelle Kategorisierung der erhobenen Daten. Aus rechtlichen Gründen kann diese Möglichkeit jedoch nicht im Rahmen von *Crowd Sourcing* genutzt werden.

Die prosopographische Erforschung der Herrschaftselite der Habsburgermonarchie steht im Zentrum des Projekts „Kaiserhof“[11], das MARK HENGERER und GERHARD SCHÖN (München) vorstellten. Nebst der eindeutigen Identifikation von Personen ist die Messbarkeit qualitativer Aspekte eine der hauptsächlichen Herausforderungen, um Netzwerke und „Reichweiten von Integration“ visualisieren zu können. Dabei hat sich der Ansatz bewährt, von Begriffen mittlerer Reichweite auszugehen. Visualisierungen (etwa von Verwandtschaftsbeziehungen oder Geolokalisierungen) sind in diesem Zusammenhang von hohem heuristischem Wert.

Natural Language Processing / Suche

Zeitangaben sind eine zentrale Information in historischen Dokumenten. Das war der Ausgangspunkt für die Präsentation von NATALIA KORCHAGINA (Zürich) zu „Natural language processing for historical documents“. Doch die maschinelle Textextraktion aus den oft handschriftlichen Dokumenten ist komplex. Ziel des Forschungsvorhabens der Referentin ist es, ein Tool für die automatisierte Extraktion von Zeitangaben aus historischen Texten zu entwickeln. Als Quellengrundlage dient ein Korpus von digitalisierten Schweizer Rechtstexten zwischen dem 10. und 18. Jahrhundert.

In einer ersten Phase des Projekts wurde unter Nutzung des an der Universität Heidelberg entwickelten Zeittaggers HeidelTime[12] ein fehlerfreies, aber kleines manuell annotiertes *Gold Standard* Korpus erstellt. Auf dieser Grundlage wird sodann mit einem hybriden Vorgehen, das *machine-learning* und regelbasierte Methoden kombiniert, ein grösseres *Silver Standard* Korpus erarbeitet, das für die Extraktion von Zeitangaben herangezogen werden kann.

Zeitgenössische Rechtstexte standen im Zentrum des Forschungsprojekts von KYOKO SUGISAKI (Zürich). Sie präsentierte in ihrem Vortrag unter dem Titel „Natural language processing in speziellen Textsorten, z.B. legislative Texte“ ihre soeben abgeschlossene Doktorarbeit. Am Beispiel von online verfügbaren Schweizer Gesetzestexten erstellte sie ein qualitativ hochwertiges Korpus von fachspezifischen Texten. Im Verlauf der Arbeiten zeigte sich, dass vorhandene Referenzkorpora (etwa die Sammlung Schweizerischer Rechtsquellen) genutzt werden können, jedoch an die Spezifika des Vorhabens angepasst werden müssen. Mittels Kombination von verschiedenen Analyseverfahren und Tools (u.a. *POS-tagging*, morphosyntaktische Analyse, *Style-Checking*) konnte die Qualität der Texterkennung deutlich verbessert werden.

Visualisierung

Unter dem Titel „Visualisierungen in den Digital Humanities“ diskutierten NOAH BUBENHOFER, KLAUS ROTHENHÄUSLER und DANICA PAJOVIC (alle Zürich)

[8] <http://wittfind.cis.uni-muenchen.de/>; <http://www.cis.uni-muenchen.de/>.

[9] <https://en.wikipedia.org/wiki/Cherpumple>.

[10] <http://www.neocrawler.de/crawler/html/>.

[11] <http://www.fnz.geschichte.uni-muenchen.de/forschung/forsch_projekte/kaiser-und-hoefe/index.html>; <http://kaiserhof.geschichte.lmu.de/>.

[12] <https://github.com/HeidelTime/heideltime/>

die theoretischen Grundlagen von Visualisierungen und hinterfragten gängige Visualisierungspraktiken in den *Digital Humanities*. Ausgangspunkt ist die Feststellung, dass Visualisierungen nicht nur bei der Darstellung von Analyseergebnissen, sondern auch bei der Datenexploration eine wichtige Rolle spielen.

Besonders bei explorativen Visualisierungen sind gemäss den Referenten methodisch-technische Aspekte wichtig. Denn Diagramme sind nicht Bilder. Sie sind hoch abstrahierte Darstellungen, die Hypothesen über Sachverhalte transportieren. Visualisierungen lassen sich entlang einer Reihe von grafischen, datentypischen, diagrammatischen, semiotischen, ästhetischen, technischen und diskursiven Attributen kategorisieren und beurteilen.

Anhand der Darstellung von Geokollokationen, das heißt von sprachlichen Äußerungen über Orte zeigte Noah Bubenhofer, wie durch Sprechen eine Vorstellung von Welt konstruiert wird.[13] Die naheliegende Visualisierung von Geokollokationen auf einer Weltkarte ist dann ein voraussetzungsvoller Vorgang, der unhinterfragte Prämissen von Kartendarstellungen übernimmt. Bubenhofer plädierte daher dafür, auch nicht kartenbasierte Visualisierungen in Betracht zu ziehen. Das Beispiel illustrierte, wie Denkstile, Software und technische Möglichkeiten in Visualisierungen mit einfliessen und diese in gewisser Weise vorbestimmen.

MATTHIAS REINERT (München) präsentierte in seinem Referat „deutsche-biographie.de – ein historisch biografisches Informationssystem. Computerlinguistischer Ansatz und Visualisierung“, das aus diesem Vorhaben resultierende Internetangebot.[14] In rund 48.000 Lexikonartikeln bietet es Informationen zu rund 540.000 Personen. Für die zuverlässige Volltexterfassung und -kodierung sowie den Normdatenabgleich von Personen und Orten wurden seit 2012 computerlinguistische Verfahren eingesetzt. Hierzu wurden lokale Grammatiken und eine korpora-spezifische Wortdatenbank erstellt. Das historisch-biografische Informationssystem ermöglicht eine Geo-Visualisierung und die Darstellung von Ego-Netzwerken. In der Diskussion betonte der Referent, dass im Zusammenhang mit einem solchen interaktiven Angebot eine transparente Kommunikation über die Möglichkeiten und Grenzen des Systems und der Datenbasis unerlässlich ist, um den Nutzern die Einschätzung der Evidenz der gewonnen Resultate zu ermöglichen.

Unter dem Titel „Theatrescapes“[15] argumentierte TOBIAS ENGLMEIER (München), dass die stetig wachsenden, nun auch für die Geisteswissenschaften verfügbaren Datenbestände oft nur mit Techniken der Informationsvisualisierung erfasst werden könnten und interpretierbar seien. Das Projekt „Theatrescapes. Mapping Global Theatre Histories“ nutzt für die interaktive Kartendarstellungen WebGL (Web Graphics Library)[16] und den Google Maps API[17]. Der Referent zeigte auf, dass mittels dieses pragmatischen Ansatzes die technischen Hürden bei der Georeferenzierung von grossen Datenbeständen mit vertretbarem Aufwand überwunden und ansprechende Resultate wie etwa interaktive historisierte Kartendarstellungen erzielt werden können. Allerdings betonte er, wie schon Noah Bubenhofer vor ihm, dass Entscheidungen über den Einsatz von bestimmter Software über technische Aspekte hinausreichen und auch inhaltliche Konsequenzen haben.

EMMA MAGES (München) stellte den „Audioatlas Siebenbürgisch-Sächsischer Dialekte“ (ASD) vor, einen interaktiven Online-Atlas.[18] Er erschließt eine umfangreiche Audiodokumentation deutscher Ortsdialekte Siebenbürgens und der Marmarosch und macht diese in Transkription und Audioaufnahmen zugänglich. Nebst der eigentlichen Transkription wurde eine morphosyntaktische Etikettierung vorgenommen und eine Ontologie für die Erschliessung entworfen. Mittels Kartierung erlaubt der ASD unter anderem qualitative und quantitative Sichten auf die örtliche Verteilung der Dialekte.

In seinem Referat über die Online-Plattform „VerbaAlpina“ berichtete STE-

[13] Vgl. auch <http://www.bubenhofer.com/geocollocations/>.

[14] <http://www.deutsche-biographie.de/>.

[15] <http://www.theatrescapes.theaterwissenschaft.uni-muenchen.de/index.html>.

[16] <https://www.khronos.org/webgl/>.

[17] <https://developers.google.com/maps/>.

[18] <http://www.asd.gwi.uni-muenchen.de/>.

PHAN LÜCKE (München) über die Herausforderungen, die sich bei diesem politische Grenzen überschreitenden und mehrsprachigen Ansatz stellten.[19] Ziel des Langzeitprojekts „Verba Alpina“ ist es, den sprachlich stark fragmentierten Alpenraum in seiner kultur- und sprachgeschichtlichen Zusammengehörigkeit zu erschliessen. Das Projekt fokussiert auf die Wechselbeziehung (sowohl in onomasiologischer wie semasiologischer Perspektive) zwischen Wörtern und bezeichneten Konzepten. Die sprachlichen Zusammenhänge werden ergänzt mit ethnographischen, historischen und politischen Aspekten und in einer interaktiven Online-Karte mit *Crowd-Sourcing*-Komponente dargestellt.

Crowd

Das Referat „Text+Berg – 150 Jahre alpinistische Texte: OCR-Fehler, Crowd Correction“ von SIMON CLEMATIDE (Zürich) diskutierte die Voraussetzungen für erfolgreiches *Crowd Sourcing*.[20] Im Rahmen des Projekts Text+Berg realisierte das Institut für Computerlinguistik der Universität Zürich eine Online-Plattform zur Korrektur des OCR-Textes der digitalisierten Jahrbücher des Schweizerischen Alpenklubs (SAC) von 1864 bis 1899.[21] Entscheidend für die Motivierung von Freiwilligen und damit für den Erfolg des Vorhabens waren eine sorgfältig programmierte Benutzeroberfläche und ein einfacher Workflow. Dazu kamen begleitende Massnahmen, um die Motivation der Teilnehmenden aufrechtzuerhalten. Hierzu gehörten spielerische Elemente und Layout-Massnahmen, die den Teilnehmenden Rückmeldungen zum Datenzustand und zur geleisteten Arbeit geben. Ein Vorteil war, dass es sich beim potentiellen Teilnehmerkreis um gut organisierte und am Thema interessierte Vereinsmitglieder mit hoher Sachkenntnis handelte.

Abschliessend stellte GERHARD SCHÖN (München) das Projek „Play4Science“[22] und die in diesem Rahmen entwickelte Spiel-Plattform „Artigo“ vor. Das Projekt bringt Geisteswissenschaftler/innen, Informatiker/innen und Computerlinguist/innen zusammen, um zweckgerichtete soziale Spiel-Software („Games with a Purpose“ (GWAP)) zu entwickeln, die seit einiger Zeit auch im wissenschaftlichen Bereich erfolgreich *Crowd-Sourcing*-Ansätze unterstützen. Ziel von Play4Science ist, eine anpassbare universelle Plattform anzubieten, die von allen Fächern für verschiedenste Anwendungen sozialer Software genutzt werden kann.

Die bereits realisierte Spiel-Plattform „Artigo“[23] lädt zur Verschlagwortung von Gemälden aus einer Bilddatenbank ein. Sie schaltet zwei Mitspieler zusammen, die unabhängig voneinander relevant erscheinende Begriffe für dieselben Bilder eingeben. Die solcherart *Peer*-validierten Begriffe werden in der Datenbank gespeichert und sind für spätere Suchabfragen nutzbar.

Fazit

Der Workshop bot einen guten Überblick über den State-of-the-Art computerlinguistischer Ansätze für die digitale Aufbereitung von Textkorpora. Er wies auf die Herausforderungen hin, die sich bei der interdisziplinären Zusammenarbeit an der Schnittstelle zwischen Technik und geisteswissenschafltlicher Forschung stellen. Es wurde klar, dass eine fundierte Fragestellung entscheidend für den Erfolg von computerlinguistischen Vorhaben ist. Zugleich wurde aber unter dem Stichwort „code matters“ auch betont, dass technologische Aspekte nicht vernachlässigt werden dürfen, da sie Einfluss auf Vorgehensweisen und Resultate haben. Von den Geisteswissenschaftlern muss daher verlangt werden, dass sie wissen, was die verwendeten Algorithmen tun. Dies gilt insbesondere auch für heuristisch und explorativ eingesetzte Visualisierungen, bei denen sich die Forschenden stets zu fragen haben, ob sie den generierten Visualisierungen genug kritisch gegenüberstehen. Unter geisteswissenschaftlichen Vorzeichen kann eine in den Visual Analytics mitunter unterstellte korrelationsbasierte „ground truth“ nicht vorausgesetzt werden.

[19] <http://www.verba-alpina.gwi.uni-muenchen.de/>; <http://gepris.dfg.de/gepris/projekt/253900505>

[20] <http://textberg.ch/site/de/willkommen/>.

[21] <http://textberg.ch/site/de/korpora/>; <http://kokos.cl.uzh.ch/>.

[22] <http://www.play4science.uni-muenchen.de/index.html>.

[23] <http://www.artigo.org/> <http://www.artigo.org/karido>.

In den Diskussionen hat sich ferner die Sicherstellung der Nachhaltigkeit in computerlinguistischen Vorhaben als zentraler Aspekt herausgestellt. Dabei geht es um mehr als Datenverfügbarkeit. Entscheidend sind das Bewusstsein für die Brüchigkeit der Datengrundlagen und der Umgang mit Unschärfen und Unvollkommenheiten. In einer weiteren Perspektive identifizierte der Workshop eine Reihe zentraler Erfolgsfaktoren für *Digital Humanities*-Projekte. So gilt es die rechtlichen Bedingungen für die Datennutzung frühzeitig zu klären, eine ausbaufähige Infrastruktur aufzubauen, die Mitarbeitenden auszuwählen, auszubilden und zu begeistern sowie die langfristige Finanzierung sicherzustellen. Insgesamt bot die Tagung einen guten Überblick über die Möglichkeiten der maschinellen Analyse und Interpretation von Texten. Es herrschte Konsens darüber, dass sich die Geisteswissenschaften dadurch in den kommenden Jahren stark verändern werden.

Konferenzübersicht:

Eckhart Arnold (Bayrische Akademie der Wissenschaften) / Mark Hengerer (Ludwig-Maximilians-Universität München) / Noah Bubenhofer (Universität Zürich) Christian Riepl (Ludwig-Maximilians-Universität München): Einführung

PANEL 1: KORPORA

Peter Makarov (Institut für Computerlinguistik Universität Zürich): Towards automated content event analysis: Mining for protest events

Daniel Knuchel (Deutsches Seminar Universität Zürich): HIV/AIDS diskurslinguistisch – ein multimediales Korpus

Max Hadersbeck (Center for Information and Language Processing (CIS)): WittFind

Susanne Grandmontagne (IT-Group for the Humanities / Anglistik): NeoCrawler

Mark Hengerer, Gerhard Schön ((Historisches Seminar der Ludwig-Maximilians-Universität München / IT-Group for the Humanities): Kaiserhof

PANEL 2: NATURAL LANGUAGE PROCESSING / SUCHE

Natalia Korchagina (Institut für Computerlinguistik Universität Zürich): Natural language processing for historical documents

Kyoko Sugisaki (Institut für Computerlinguistik Universität Zürich): Natural language processing in speziellen Textsorten, z.B. legislative Texte

PANEL 3: VISUALISIERUNG

Noah Bubenhofer, Klaus Rothenhäusler, Danica Pajovic (Institut für Computerlinguistik Universität Zürich): Visualisierungen in den Digital Humanities

Matthias Reinert (Historische Kommission bei der Bayerischen Akademie der Wissenschaften): deutsche-biographie.de – ein historisch biografisches Informationssystem. Computerlinguistischer Ansatz und Visualisierung

Tobias Englmeier (Institut für Theaterwissenschaft der Ludwig-Maximilians-Universität München / IT-Group for the Humanities): Theatrescapes

Emma Mages (Institut für den Nahen und Mittleren Osten der Ludwig-Maximilians-Universität München / IT-Group for the Humanities / Romanistik): Audioatlas Siebenbürgisch-Sächsischer Dialekte

Stephan Lücke (IT-Group for the Humanities / Romanistik): VerbaAlpina

PANEL 4: CROWD

Simon Clematide (Institut für Computerlinguistik Universität Zürich): Text+Berg – 150 Jahre alpinistische Texte: OCR-Fehler, Crowd Correction

Gerhard Schön (IT-Group for the Humanities / Kunstwissenschaft): Play4Science und Artigo

Tagungsbericht *Digitale Daten in den Geisteswissenschaften. Interdisziplinäre Perspektiven für semantische und strukturelle Analysen*. 28.01.2016–29.01.2016, München, in: H-SozKult 09.05.2016.