

Brauchen die Digital Humanities eine eigene Methodologie?

Überlegungen zur systematischen Nutzung von Text Mining Verfahren in einem politikwissenschaftlichen Projekt

Die Verfügbarkeit neuer Verfahren und Technologien in der Informatik hat in den letzten Jahren weltweit neue Forschungsansätze in den Geistes- und Sozialwissenschaften ermöglicht, die es in einer bisher nicht bekannten Weise erlauben, umfangreiche digitale oder digitalisierte Datenbestände in die geisteswissenschaftliche Forschung einzubeziehen. In der Praxis findet sich jedoch nach wie vor der Fall, dass ein Projektvorhaben entweder eine Anwendung von generischen Verfahren aus der Informatik auf geistes- oder sozialwissenschaftliche Daten darstellt, bei der die Interessen der Fachwissenschaften oft nur ungenügend berücksichtigt werden. Oder das Vorhaben stellt eine von den Fachwissenschaften „dominierte“ Adaption etablierter, nicht-computergestützter Arbeitsweisen nach den Vorgaben der Fachwissenschaftler dar, bei der die Informatik in die Rolle eines bloßen Dienstleisters gedrängt wird. In diesem Zusammenhang ist zunächst die Unterscheidung von *Datenmanagement* und *Datenauswertung* in den Digital Humanities hilfreich. Beschränken sich interdisziplinäre Kooperationen lediglich auf die Digitalisierung, Speicherung und Verwaltung von Daten, sind die Möglichkeiten für eigene Forschungsbeiträge auf Seiten der Informatik eher beschränkt. Einen wesentlichen Beitrag für die Fachwissenschaften können die Digital Humanities aber vor allem dann leisten, wenn digitale Daten nicht nur per Computer verwaltet, sondern auch ausgewertet werden sollen. In diesem Fall stellen sich besondere Anforderungen an die Kooperation, die eigene Forschungsleistungen und tiefere Verständnisse für die Arbeit des jeweils Anderen von den beteiligten Disziplinen verlangen.

Im Rahmen des gemeinsam von Politikwissenschaftlern und Informatikern betriebenen und vom BMBF geförderten Projektvorhabens „ePol - Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011“ (Wiedemann/Lemke/Niekler 2013) versuchen wir deshalb, eine Verständigungsebene zu entwickeln, die eine Klärung gegenseitiger Anforderungen und spezifisch auf diese angepasste Lösungen ermöglicht, mit denen Forschungsinteressen auf beiden Seiten angegangen werden können, und die in diesem Sinne Nachhaltigkeit erzeugt, als dass beide Seiten ihre jeweiligen Arbeitsmethoden verändern, anstatt nur mit den Ergebnissen des anderen Teilprojekts umzugehen. Eine solche Verständigungsebene muss aus unserer Perspektive vor allem Aspekte *transdisziplinärer Modellierung* umfassen – mit anderen Worten, Sozialwissenschaftler und Informatiker müssen eine gemeinsame (Definitions-)Sprache finden. Für die Informatik ist die Modellierung essentiell. Denn nur durch die Beschreibung von Sachverhalten in Form theoretischer, prozessualer und mathematischer Modelle, lassen sich Problemlösungen in Form von Computerprogrammen implementieren. Modelle sind aber immer Vereinfachungen der Realität und bedürfen der Anpassung bzw. Korrektur

¹ Abteilung Automatische Sprachverarbeitung, Institut für Informatik an der Universität Leipzig

durch die Fachwissenschaftler. Dies macht die Modellierung zum zentralen Angelpunkt zwischen den Projektbeteiligten.

Im Fall des Projektvorhabens ePol soll die These der Postdemokratie – Politik wird unter dem Einfluss des Neoliberalismus zunehmend von ökonomischen Gesichtspunkten bestimmt – erstmals umfassend empirisch analysiert werden. Grundlage bilden mehr ca. 3,5 Millionen Texte aus Tageszeitungen und Wochenzeitschriften (Zeit, FAZ, taz, SZ) im Zeitraum von 1949 bis 2011, die mit Verfahren der automatischen Sprachverarbeitung ausgewertet werden sollen.

Für die Operationalisierung der politikwissenschaftlichen Fragestellung, und damit der Bereitstellung von Kriterien für die Konzipierung eines Forschungsdesigns und die Auswahl passender Textanalyseverfahren, bieten sich zwei Zugänge an. Zum Einen können wir auf Verfahren in den Sozialwissenschaften zur Hypothesenbildung und deren empirischer Überprüfung aus dem Bereich der Inhaltsanalysen zurückgreifen. Verschiedene Methoden der qualitativen und quantitativen Textanalyse² haben sich in der Politikwissenschaft fest etabliert. Eine Anpassung dieser *Vorgehensmodelle* für (semi-)automatische Analysen großer Textmengen (z.B. vollständige Jahrgänge retrodigitalisierter Tageszeitungen) steht jedoch noch aus. Zum anderen können wir uns an Verfahren des Requirements Engineering aus der Informatik orientieren. Diese Verfahren sind zwar insbesondere in der Wirtschaftsinformatik primär auf die Betrachtung von betriebswirtschaftlich relevanten Prozessen hin orientiert (Kastens et. al. 2008), konnten jedoch auch für die Modellierung von digitalen Kulturgütern (z.B. Texte, Bilder, Musik, Objekte, Filme, Karten) als Gegenstand geistes- und sozialwissenschaftlicher Untersuchungen und für die Bewahrung kulturellen Erbes weiterentwickelt werden (vgl. Schlieder/Wullinger 2010).

Im Rahmen unseres Vortrags skizzieren wir, wie mit Hilfe eines vereinfachten Requirements Engineering die komplexen inhaltlichen Fragen des Projekts ePol in enger Absprache zwischen den Projektbeteiligten in einzelne Teilaufgaben aufgegliedert werden konnten, für die klare funktionale Anforderungen erarbeitet und deren Implementierung mit konkreten Programmpaketen spezifiziert werden konnte (vgl. Abb. 1 und 2).

Die Operationalisierung der Forschungsfrage von Seiten der politischen Theorie erfolgt über die Beobachtung der Veränderungen öffentlichen Begründens von Politikmaßnahmen in Richtung einer zunehmenden „Ökonomisierung“ (Lemke 2012). Um diese Veränderung in medialer Berichterstattung messbar zu machen, wurden im Rahmen der Anforderungsanalyse drei wesentliche Teilaufgaben spezifiziert:

1. Retrieval relevanter Dokumente: Erstellung eines Subkorpus aus den 3,5 Millionen Dokumenten unseres Gesamtkorpus nach den Kriterien einer hohen Dichte neoliberaler Sprachgebrauchsmuster und argumentativer Begründungen,
2. Identifizierung von Argumenten: Manuelle Annotation von Argumenten in unserem Subkorpus; Entwicklung und Anwendung eines Klassifikationsverfahrens, dass weitere (Kandidaten für) Argumente in den Texten automatisch findet,
3. Analyse: Validierung / Falsifizierung von Hypothesen auf den (semi-)automatisch extrahierten Daten; Erstellung geeigneter Visualisierungen zur Analyse der Zeitreihendaten.

2 Etwa die wissenssoziologische Diskursanalyse nach Keller (2007); Zur Unterscheidung von hypothesenprüfender und rekonstruktiver Sozialforschung siehe Bohnsack (2010).

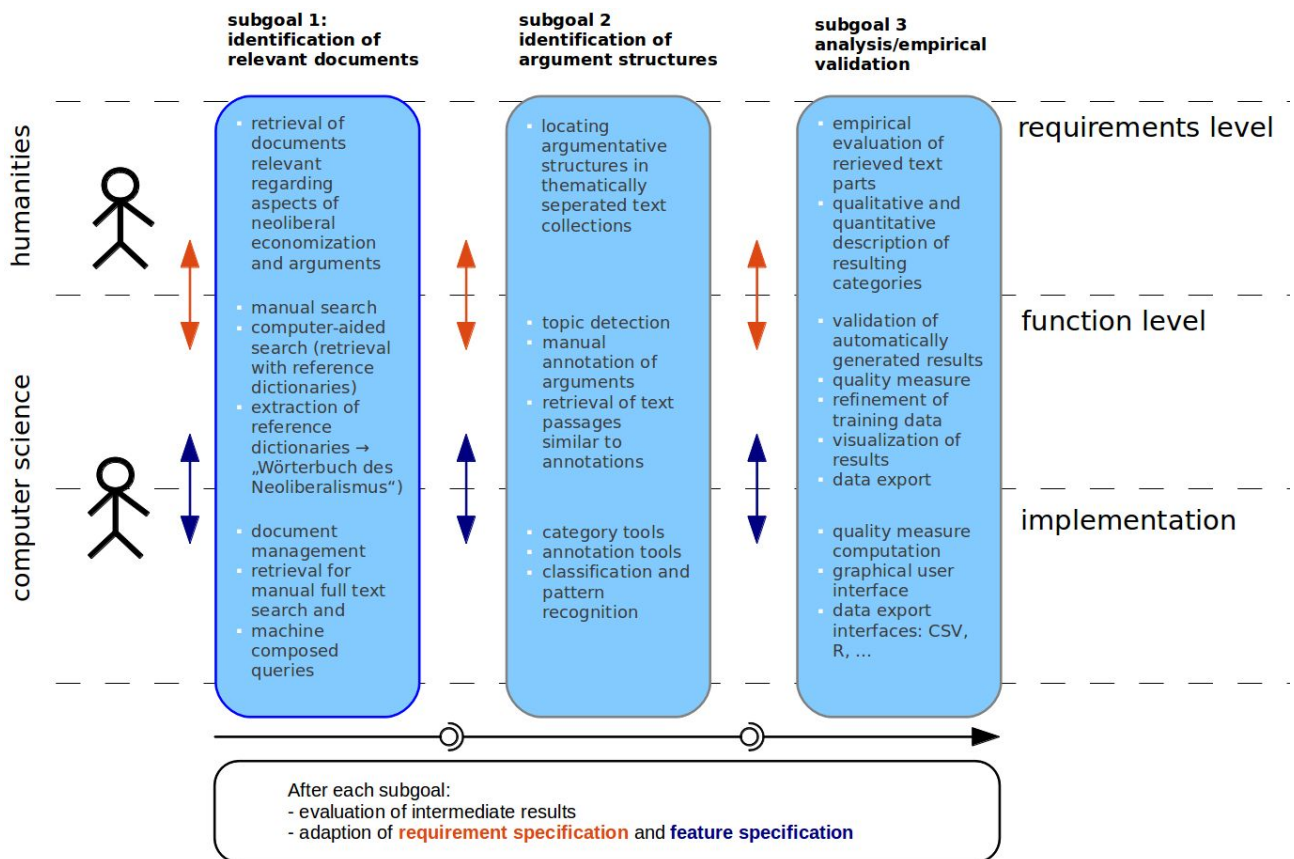


Abbildung 1 - Unterteilung der Anforderungen in Teilaufgaben und Verantwortungsebenen

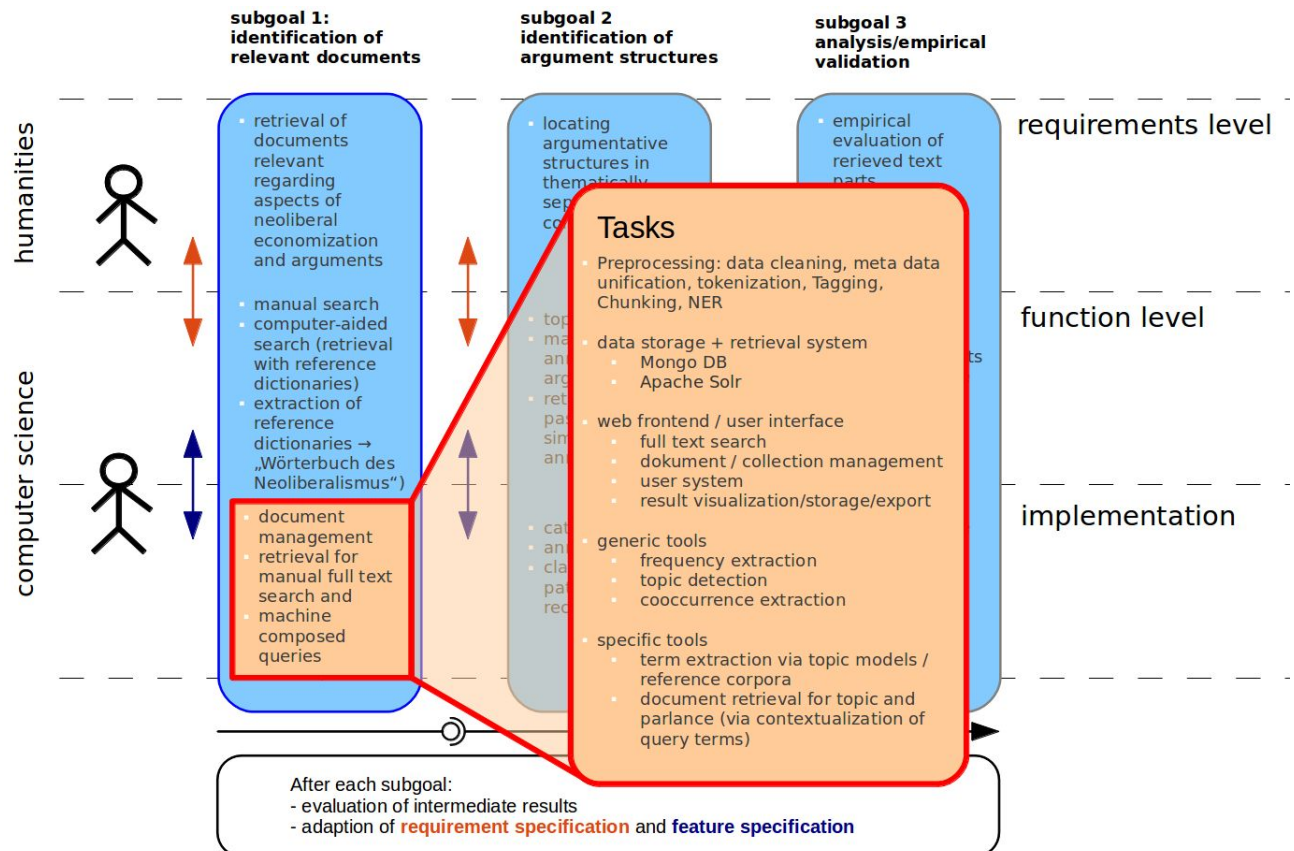


Abbildung 2 - Spezifikation konkreter Aufgaben für die Implementierung einzelner Teilaufgaben

Für diese drei Teilaufgaben werden Anforderungen von Seiten der Politikwissenschaftler formuliert und, dem „Lastenheft“ im Rahmen des Software Engineering vergleichbar, detailliert beschrieben. Auf der Implementierungsebene legen die beteiligten Informatiker fest, wie sie diese Anforderungen umzusetzen gedenken (vgl. „Pflichtenheft“). Geistes- und Sozialwissenschaftler als „informierte Anwender“ von Text Mining Verfahren und Informatiker als deren Entwickler sollten sich dabei in ihren Perspektiven auf die Umsetzung der Teilziele auf Ebene der funktionalen Spezifikation so nah wie möglich kommen. Das setzt voraus, dass die Politikwissenschaftler mögliche Analyse-Verfahren kennen und in ihren Grundlagen verstehen – zum Beispiel wenn entschieden werden muss, ob für ein induktives oder deduktives Forschungsdesign eher unüberwachte oder überwachte maschinelle Lernverfahren zum Einsatz kommen sollen. Die Informatiker wiederum benötigen ein Bewusstsein für die Anforderungen und Ziele der Politischen Theorie, um beispielsweise wie im Rahmen von ePol die Anforderung, dass es weniger um die Identifizierung klar abgegrenzter thematischer Bereiche für die Teilaufgaben 1 und 2 geht, als vielmehr um die zeitliche Veränderung eines Begründungsmodus, adäquat in ihrer Arbeit umsetzen zu können.

Für die Sicherung der Qualität eines solchen Forschungsprozesses liefert die Orientierung am Requirements Engineering ebenfalls einen nützlichen Rahmen. Die Aufteilung in wohldefinierte Teilziele ermöglicht jeweils die Evaluation von Zwischenergebnissen, bevor mit diesen die Analyse weiter fortgeführt wird. Auf Seiten der Informatik steht damit die Aufgabe, geeignete Evaluierungsverfahren für die jeweiligen Ergebnisse zu entwickeln, die wiederum mit Hilfe der Politikwissenschaftler durchgeführt werden müssen. Für Teilaufgabe 1 bedeutet das zum Beispiel die Annotation eines „Goldstandards“ von relevanten/nicht-relevanten Dokumenten in Bezug auf die Forschungsfrage mit denen die Qualität des Dokument-Retrieval bestimmt und optimiert werden kann. Systematische Evaluationen und semi-überwachtes Active Learning ermöglichen, dass Vertrauen in die computergestützten Analyseprozesse großer Datenmengen generiert und eine hohe Qualität in den oft quantifizierten Endresultaten sichergestellt wird.

Mit diesen Eigenschaften bilden die Methoden der Digital Humanities ein eigenartiges hybrides Gebilde zwischen dem positivistischen und dem interpretativen Paradigma empirischer Sozialforschung (Goldkuhl 2012). Beispielsweise dominiert bei der zunächst manuellen Identifizierung von neoliberalen Argumentationen im Teilschritt 2 des ePol-Projekts eindeutig eine hermeneutische Sicht auf die Daten. Aus dieser werden aber wiederum Sprachregelmäßigkeiten extrahiert, maschinell „gelernt“ und in einen quantifizierenden, eher positivistischen Analyseprozess überführt. Wie sehr ein solches Vorgehensmodell systematisiert und Forschungsprojekt-übergreifend festgelegt werden kann, ist noch längst nicht abschließend geklärt. Eine Debatte darüber, welche Anforderungen an die Qualitätssicherung des Forschungsprozesses, gerade im Hinblick auf die ungewöhnliche Kombination dieser oft widerstreitenden Wissenschaftsparadigmen zu stellen sind, muss noch viel intensiver geführt werden.

Wir sind jedoch der Auffassung, dass sich viele Erfahrungen aus dem ePol-Projekt zu einem Vorgehensmodell für größere Digital Humanities Projekte insgesamt generalisieren lassen. Die systematische Identifikation von (aufeinander aufbauenden) Teilzielen, ihre disziplinübergreifende Spezifikation auf verschiedenen Ebenen und ihre Untersetzung mit konkreten Aufgaben ermöglicht es, klare Erwartungen in die computergestützten Analysen zu entwickeln, den dafür notwendigen Arbeitsaufwand abzuschätzen und Kriterien für die Qualität der erreichten Ergebnisse bei den (Teil-)Zielen des Pro-

jekts zu entwickeln. Insofern viele spannende geistes- und sozialwissenschaftlichen Fragen, die mit Hilfe großer Textkollektionen beantwortet werden könnten wohl kaum mit generischen Software-Lösungen bearbeitet werden können, sind weitere methodologische Innovationen zur Ermöglichung von Transdisziplinarität dringend notwendig.

Literatur

- Bohnsack, Ralf (2010): Rekonstruktive Sozialforschung. Einführung in qualitative Methoden. Ralf Bohnsack. 8. Aufl. Opladen, Farmington Hills, Mich: Budrich.
- Goldkuhl, Göran (2012): Pragmatism vs interpretivism in qualitative information systems research. In: European Journal of Information Systems 21 (2), S. 135-146.
- Kastens, Uwe / Büning, Hans Kleine (2008): Modellierung - Grundlagen und formale Methoden, Hanser.
- Keller, Reiner (2007): Diskursforschung. Eine Einführung für SozialwissenschaftlerInnen. 3., aktualisierte Aufl. Wiesbaden: VS Verlag für Sozialwissenschaften (Qualitative Sozialforschung, 14).
- Lemke, Matthias (2012): Die Ökonomisierung des Politischen. Entdifferenzierungen in kollektiven Entscheidungsprozessen. Discussion Paper Nr. 2. Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. Hamburg; Leipzig.
- Schlieder/Wullinger (2010): Semantic Ageing of Complex Documents: A Case Study from Built Heritage Preservation, In: Informatik 2010, Service Science - Neue Perspektiven für die Informatik, Band 2, Gesellschaft für Informatik, 580-586.
- Wiedemann, Gregor; Lemke, Matthias; Niekler, Andreas (2013): Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011. Ein Werkstattbericht. In: Zeitschrift für politische Theorie 4 (1).