Computationelle Textanalyse als fünfdimensionales Problem

Gius, Evelyn
gius@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Die Einschätzung und Kritik computationeller Textanalyse

In diesem Beitrag wird ein Modell vorgestellt, das zu einer Einschätzung der Komplexität von Forschungsansätzen dient, die sich Texten mit computationellen Analysen nähern. Das Modell wurde vor dem Hintergrund der (literaturwissenschaftlichen) Analyse von literarischen Texten entwickelt, es ist jedoch – ggf. mit leichten Anpassungen – für Textanalysen generell geeignet.

Die Komplexität von Digital Humanities-Projekten ist bestimmt von der Aushandlung von Vorannahmen, Methoden, der Passung zum Gegenstand, der konkreten interdisziplinären Zusammenarbeit, die fachlich, persönlich und oft auch karrierestrategisch eine große Herausforderung für die Beteiligten sein kann, bis hin zur Darstellung von Ergebnissen für eine oder mehrere Forschungscommunities. Neben Fragen der Projektplanung und -steuerung, wissenschaftspolitischen und wissenschaftskommunikativen Aspekten geht es auch um Fragen, die das eigentliche Forschungsgeschehen betreffen. Dieses wird aktuell in Bezug auf seine Relevanz und Ausrichtung diskutiert: Eine harsche Kritik von Nan Z. Da (2019a) an den Verfahren der DH initiierte eine mit dem etwas überzogenen Begriff „Digital Humanities War“ bezeichnete Auseinandersetzung.[1] Diese Debatte wird z.T. als Auseinandersetzung zwischen angeblichen Strukturalist*innen und Poststrukturalist*innen dargestellt. Zumindest von letzteren, die den Strukturalismus als solchen benennen und eine Kluft zwischen diesem und den eigenen Zugängen diagnostizieren (vgl. z.B. Dobson 2019 und Bode im Erscheinen). Hinzu kommt, dass in der Auseinandersetzung förderpolitische Aspekte zumindest als Hintergrund eine große Rolle spielen.[2]

Ein methodenunabhängiges Modell

Diese Auseinandersetzungen gehen zum Großteil an den eigentlichen Forschungszugängen vorbei. Dabei wäre es aus Sicht der Digital Humanities *und* der Literaturwissenschaft erhellend, die diskutierten Verfahren oder gar Methodenlinien detaillierter zu beschreiben und ihre Bedeutung zu reflektieren. Deshalb möchte ich ein Modell vorschlagen, das eine solche Betrachtung von computationellen Textanalyseansätzen ermöglicht und eine Grundlage bildet, auf der Textanalyse-Zugänge unabhängig von ihrer literaturtheoretischen Fundierung beschrieben, kritisiert und zu verglichen werden können.

Ausgangspunkt des Modells sind die drei Aspekte, die für jede computergestützte Textanalyse wesentlich sind: Die Phänomene, denen das Interesse gilt, die Texte, die untersucht werden, und die Art, wie Erkenntnis erzeugt wird.[3] Aus diesen Aspekten lassen sich insgesamt fünf Dimensionen ableiten, die für die Einschätzung der Komplexität eines Zugangs genutzt werden können: die Kontextualisierung von Phänomenen, die Zusammengesetztheit von Phänomenen, die Heterogenität von Texten, der Analysemodus und der Erkenntnisbeitrag computationeller Analysen.

Zusammengesetztheit von Phänomenen

Eine Einschätzung der Phänomene, die in einer computationellen Textanalyse untersucht werden, kann anhand der Phänomenbeschreibung stattfinden. Für diese kann man fragen: Wird das Phänomen als einfach, nicht weiter unterteilt, oder als aus mehreren Phänomenen zusammengesetzt betrachtet? Dabei geht es wohlgemerkt nicht um eine allgemein gültige Definition des entsprechenden Phänomens, sondern um die von den Forscher*innen genutzte Beschreibung.

Beschreibungen für dasselbe Phänomen können in unterschiedlichen Forschungsprojekten entsprechend unterschiedlich ausfallen. In Bezug auf ein aktuelles Forschungsprojekt zu Gender und Krankheit in literarischen Prosatexten[4] sind zum Beispiel folgende Unterschiede denkbar: Man könnte das Phänomen „Krankheit einer literarischen Figur“ ausschließlich daran festmachen, ob diese ärztlich behandelt wird. Man kann aber ebenso eine Reihe von Phänomenen wie körperliche Reaktionen, Aussagen der Figur etc. nutzen, um Krankheit zu bestimmen.

Kontextualisierung von Phänomenen

Neben der Bestimmung der Teile, aus denen eine Phänomenbeschreibung zusammengesetzt ist, geht es auch um die Frage, welches Wissen zur Bestimmung des Phänomens herangezogen werden muss. Dies kann zum einen Wissen sein, das der Text vermittelt. Aber es kann auch weiteres Wissen nötig werden, wie etwa spezielles Domänenwissen, zusätzliches (innerfiktionales oder außerfiktionales) Weltwissen u.ä. Die Kernfrage ist entsprechend: Braucht man über das Textwissen hinausgehendes weiteres Wissen, um ein Phänomen zu identifizieren?

Auch hier gilt: Die Einstufung der Komplexität gilt für den betrachteten Anwendungsfall, andere

Fälle haben ggf. für dieselben Phänomene andere Komplexitätsgrade. Im Projekt Gender und Krankheit wurde etwa mit Koreferenz-Auflösung experimentiert, die überwiegend auf Textphänomenen basiert. Das Krankheitskonzept wiederum wurde unter Rückgriff auf Wissen für zeitgenössische Krankheiten und Krankheitsbezeichnungen bearbeitet (etwa „Phthise“ als Bezeichnung für Tuberkulose).

Abbildung 1 stellt beispielhaft die beiden Dimensionen der Komplexität einiger Phänomene dar, die im Projekt Gender und Krankheit eine Rolle spielen.

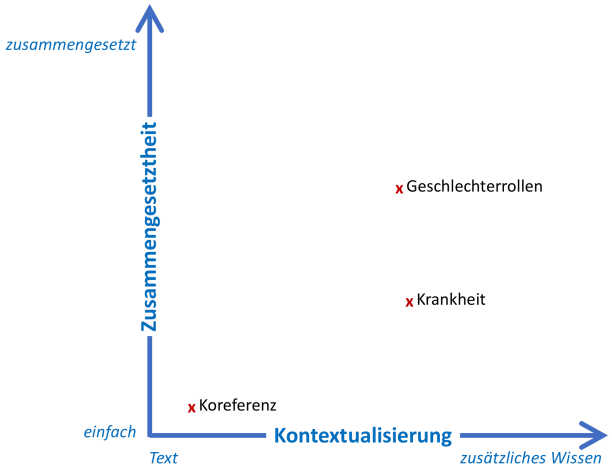

Abbildung 1: Komplexitätsdimensionen für Phänomene

Textheterogenität

Oft wird vorschnell angenommen, dass für die textorientierten Digital Humanities die nun wesentlich größere Menge an untersuchten Texten distinktiv ist. Dabei ist die Frage, ob es sich um – vermeintliche – Big Data handelt oder nicht, aus Sicht der computationellen Textanalyse nur insofern interessant, als damit die Frage zusammenhängt, ob man die Texte, die man analysiert, kennt bzw. kennen kann oder nicht. In Bezug auf die Komplexität der genutzten Texte relevanter ist hingegen die umfassendere Frage: Wie viele (wie) verschiedene Texte werden analysiert? Dabei fällt unter Heterogenität von Texten die Anzahl der Texte selbst, aber auch die Anzahl von verschiedenen Texteigenschaften, die für die Fragestellung relevant sind bzw. sein könnten. Im Fall literarischer Texte sind das typischerweise Eigenschaften wie Gattung, Genre, Epoche, Autorgender, Erscheinungsort etc.

Die Textheterogenität reicht von einem Text bis zu sehr vielen, sehr heterogenen Texten reicht[5] und ist v.a. im Vergleich zu anderen Vorhaben beurteilbar. Im Projekt Gender und Krankheit liegt eine vergleichsweise hohe Textheterogenität vor, da das Korpus aus über 2.000 deutschsprachigen Texten besteht, die verschiedene Genres, Autor*innen und Epochen zuzuordnen sind.

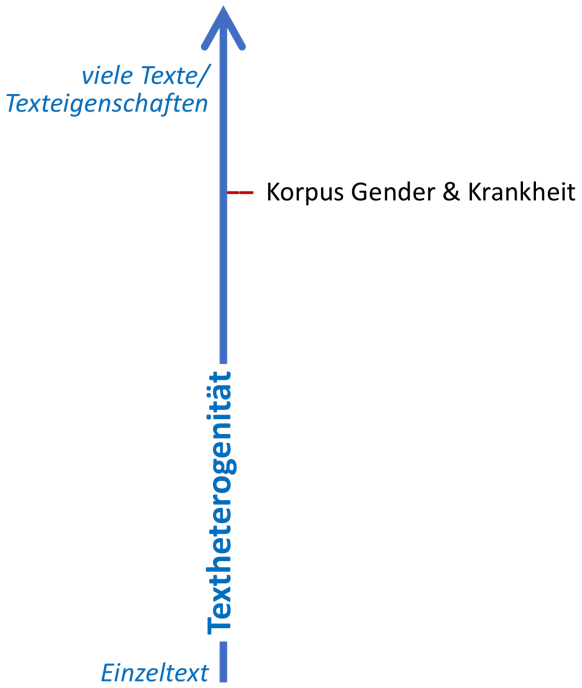

Abbildung 2: Komplexitätsdimension Textheterogenität

Analysemodus

In der Komplexitätsdimension des Analysemodus geht es darum, wer die Erkenntnisse produziert. Hier sind die beiden Möglichkeiten recht offensichtlich: Auf der einen Seite steht (menschliches) Lesen, auf der anderen Seite maschinelles Erschließen. Die Hauptfrage ist also: Wird die Textbasis durch Menschen oder durch Computer erschlossen? Dabei wird für alle Zugänge als gegeben vorausgesetzt, dass der Computer genutzt wird. Während das Lesen in Annotationen von Textstellen oder zumindest in die Ergänzung der Texte um Metainformationen resultiert, wird beim maschinellen Erschließen im Normalfall Textmining betrieben. Beide Textzugangsarten können weiter differenziert werden nach der Interpretationstheorie (etwa in text-, leser- oder autororientierte Zugänge) bzw. dem angewendeten maschinellen Verfahren (etwa in regelbasierte und Lernverfahren).

In konkreten Forschungsprojekten kommen fast immer beide Modi vor. So werden im Projekt Gender und Krankheit manuelle Annotationen von Textpassagen und halb-automatische Verfahren zur Wortfeldgenerierung für die weitere Verarbeitung oder die Methodenentwicklung mit automatischen Verfahren zur Figurenerkennung, Segmentierung und Sentimentanalyse kombiniert. Da die Zwischenschritte in der Analyse zumeist manuell überprüft und teilweise ergänzt werden, handelt es sich hier um ein Verfahren zwischen Lesen und automatischem Erschließen und damit um eine eher geringe Komplexität.

Erkenntnisbeitrag

Schließlich geht es bei der Betrachtung von computationellen Textanalysen auch darum, wie der Computer eingesetzt wird, um Erkenntnisse zu generieren. Wenn man von der literaturwissenschaftlichen Praxis der Textanalyse ausgeht, ist die komplexeste Aufgabe jene, die Textbasis insgesamt im Hinblick auf die gewählte Fragestellung zu interpretieren. Interpretation ist jedoch bislang nicht der Fokus computationeller Zugänge zu literarischen Texten. Trotzdem lohnt es sich, Interpretation als ein Extrem der Dimension der Erkenntnis zu denken. In Anlehnung an die literaturwissenschaftliche Praxis kann man die Komplexitätsdimension des Erkenntnisbeitrags computationeller Analysen als von der Analyse des Textes für ein erstes Textverständnis bis hin zur Interpretation der Textbasis als Ganzes ausgedehnt sehen.[6] Alternativ kann auch die sozialwissenschaftliche Kategorisierung von Forschungslogiken[7] in Anlehnung an Peirce (1935) in Deduktion, Induktion und Abduktion als Skala für die Erkenntnisdimension genutzt werden.

Unabhängig von der Frage, welche Systematik man für die Tätigkeiten verwendet, die mit Textverstehen befasst sind, ist die zentrale Frage in der letzten Komplexitätsdimension: Wie weit geht der Erkenntnisbeitrag der computationellen Methode? Es geht also um die Frage nach der Neuheit des computationell Erforschten. Grob kann man die Komplexitätsstufen des Erkenntnisbeitrags wie folgt erfassen: Werden in einer deduktiven bzw. einfachen Textanalyse aufgrund von bestehenden Hypothesen bzw. Regeln (also bestehenden Analysekategorien und -verfahren) durchgeführt, werden aus der Betrachtung von Texten neue Analysekategorien oder auch Taxonomien entwickelt oder handelt es sich um Hypothesen über größere Zusammenhänge in den Texten, also um ihre Interpretation?[8]

Bei der Auseinandersetzung mit der Komplexitätsdimension des Erkenntnisbeitrags ist zu beachten, dass in einer typischen literaturwissenschaftlichen Textanalyse meist alle Modi vorliegen und fließend ineinander übergehen. Für die Komplexitätseinschätzung ist relevant, welche Modi davon computationell unterstützt werden sollen. Im Fall des Projekts zu Gender und Krankheit soll etwa deduktiv die Veränderung der Figurenkonstellation anhand der Figurennennungen analysiert werden. Ein induktives Verfahren liegt vor, wenn Genderkategorien durch Clustering von Figurenrede herausgearbeitet werden (die dann wieder deduktiv in der Analyse genutzt werden). Und schließlich liegt ein abduktiver Zugang vor, wenn durch eine Gesamtbetrachtung ein neues Element entdeckt würde, das Figurenkrankheit beeinflusst.

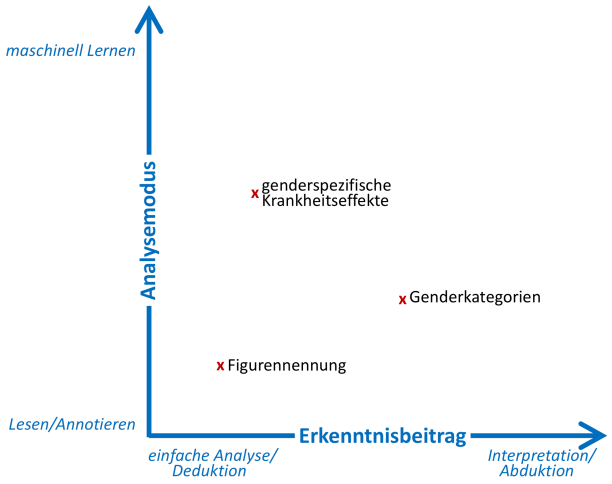

Abbildung 3: Komplexitätsdimensionen für Erkenntnis: Analysemodus und Erkenntnisbeitrag

Zur Nutzung des Modells

Wie bereits dargelegt, betrifft die Bestimmung der Komplexität in den fünf Dimensionen primär die normativen Setzungen durch die Forscher*innen. Ausschlaggebend ist weniger, wie Texte, Phänomene und Erkenntnis an sich modelliert werden *sollten*, sondern vielmehr, wie die Modellierung konkret umgesetzt wird. Die vorgeschlagenen Dimensionen sind außerdem von der mit einem Zugang verbundenen Interpretationstheorie unabhängig. Damit ist das Modell für alle literaturwissenschaftlichen Textanalyseverfahren geeignet, für jene, die in einer strukturalistischen Tradition gesehen werden können, genauso wie für solche, die eher postmoderne Zugangsweisen umsetzen – oder andere Zugänge.

Für die Betrachtung und Kritik eines Zugangs sollten alle fünf Dimensionen berücksichtigt werden. Damit vermeidet man auch vorschnelle Kritik, die sich auf eine einfache Modellierung einer Dimension beschränkt und den Zugang insgesamt als unterkomplex betrachtet, obwohl er in einer oder mehreren anderen Dimensionen Erhebliches leistet.

Darüber hinaus eignet sich das Modell als Instrument für den Entwurf eines Zugangs. Es kann in allen Phasen computationeller Textanalyse genutzt werden – vom Design des Forschungszugangs zu Beginn der Forschungsarbeit über die wiederholten Bestandsaufnahme oder Nachjustierung im Projektverlauf bis hin zur Einordnung der erzielten Ergebnisse am Ende und der Reflektion des gesamten Prozesses.

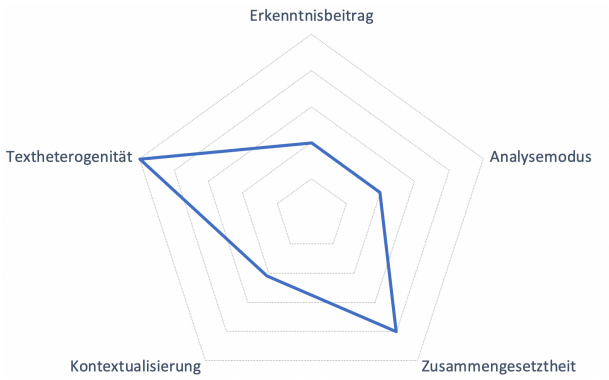

Abbildung 4: Übersichtsdarstellung: Komplexität im Projekt Gender und Krankheit

Abschließend seien noch einmal die fünf Dimensionen mit ihren Kernfragen dargestellt:

1. Komplexitätsdimension 1: die Zusammengesetztheit von Phänomenen
   - Frage *: Wird das Phänomen als einfach, nicht weiter unterteilt, oder als aus mehreren Phänomenen zusammengesetzt betrachtet?*
   - Komplexität: von einfach bis hin zu vielfach zusammengesetzten Phänomenen
2. Komplexitätsdimension 2: die Kontextualisierung von Phänomenen
   - Frage: *Braucht man über das Textwissen hinausgehendes weiteres Wissen, um ein Phänomen zu identifizieren?*
   - Komplexität: von Textwissen bis hin zu verschiedenen Arten von umfangreichem weiterem Wissen
3. Komplexitätsdimension 3: Textheterogenität
   - Frage *: Wie viele (wie) verschiedene Texte werden analysiert?*
   - Komplexität: von einem Text mit homogenen Eigenschaften bis hin zu vielen, in sich und zueinander heterogenen Texten
4. Komplexitätsdimension 4: Analysemodus
   - Frage *: Wird die Textbasis durch Menschen oder durch Computer erschlossen*?
   - Komplexität: von von Menschen annotiert bis hin zu von Maschinen durch Lernen analysiert
5. Komplexitätsdimension 5: der Erkenntnisbeitrag computationeller Analysen
   - Frage: *Wie weit geht der Erkenntnisbeitrag der computationellen Methode?*
   - Komplexität: von der Anwendung simpler Regeln auf einzelne Textelemente bis zur Interpretation der gesamten Textbasis.

Fußnoten

1. Vgl. dazu den Artikel „The Digital Humanities Debacle. Computational methods repeatedly come up short“ von Da (2019b) und als „Digital Humanities War“ zusammengefasste Reaktionen auf https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986 , sowie das „Special Forum on Responses to Nan Z. Da“ in *Cultural Analytics* auf https://culturalanalytics.org/2019/09/special-forum-on-responses-to-nan-z-da/ , sowie Jannidis (2019) und Krajewski (2019).
2. Nan Z. Da (2019b) könnte diesbezüglich so zusammengefasst werden, dass sie vorschlägt, keine Mittel mehr in die Computational Literary Studies zu stecken und damit zu verschwenden.
3. Damit ist der Zugang wesentlich spezifischer als die umfassende „TaDiRAH - Taxonomy of Digital Research Activities in the Humanities“, die alle digitalen Forschungsaktivitäten zu erfassen versucht (vgl. http://tadirah.dariah.eu/vocab/ , gesehen am 21.12.2019).
4. Vgl. dazu z.B. Gius et al. (2019), Andresen et al. (2019) und https://www.herma.uni-hamburg.de/subprojects.html , gesehen am 21.12.2019).
5. Genau genommen werden hier zwei Gegensatzpaare abgebildet: Anzahl (von Texten) und Heterogenität (von Texteigenschaften). Diese Eigenschaften werden zu einer Dimension zusammengefasst, da sie die Komplexität von Texten vergleichbar steigern.
6. „Textanalyse“ ist literaturwissenschaftlich mehrdeutig, da der Begriff sowohl eine Textanalyse meint, die das Textverständnis im Fokus hat und der anschließenden Interpretation als Voraussetzung dient, als auch den Prozess der Analyse und Interpretation insgesamt, vgl. dazu Winko (2003).
7. Vgl. dazu auch die Arbeit im Projekt hermA zu den verschiedenen Forschungslogiken im Kontext von Annotationen (Gaidys et al. 2017 bzw. www.herma.uni-hamburg.de).
8. Vgl. dazu auch Eco (1987): „[D]er Text ist ein Objekt, das die Interpretation im Verlauf ihrer zirkulären Anstrengungen um die eigene Schlüssigkeit bildet auf der Basis dessen, was sie als ihr Resultat erschafft. Ich schäme mich nicht, daß ich auf diese Weise den alten und immer noch gültigen hermeneutischen Zirkel definiere. Die Logik der Interpretation ist die Peircesche Logik der ‚Abduktion‘.“

Bibliographie

Adelmann, Benedikt / Melanie Andresen / Anke Begerow / Lina Franken / Evelyn Gius / Michael Vauth (2019): „Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics“. In *DH2019 Book of Abstracts*. Utrecht.

Bode, Katherine. Im Erscheinen. „Why you can’t model away bias“. Preprint: Modern Language Quarterly 80.3.

Da, Nan Z. (2019a): „The Computational Case against Computational Literary Studies“. *Critical Inquiry* 45 (3): 601–39. https://doi.org/10.1086/702594 .

Da, Nan Z. (2019b): „The Digital Humanities Debacle“. *The Chronicle of Higher Education*, 27. März 2019. https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986 .

Dobson, James E. (2019): *Critical Digital Humanities: The Search for a Methodology*. Topics in the digital humanities. Urbana, Illinois: University of Illinois Press.

Eco, Umberto (1987): *Lector in Fabula. Die Mitarbeit der Interpretation in erzählenden Texten*. München: Hanser.

Gaidys, Uta / Evelyn Gius / Margarete Jarchow / Gertraud Koch / Wolfgang Menzel / Dominik Orth / Heike Zinsmeister (2017) : „Project description – hermA: Automated modelling of hermeneutic processes“. *Hamburger Journal für Kulturanthropologie*. https://journals.sub.uni-hamburg.de/hjk/article/view/1213 .

Gius, Evelyn / Katharina Krüger / Carla Sökefeld (2019): „Korpuserstellung als literaturwissenschaftliche Aufgabe“. In *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts*, 164–166. Frankfurt & Mainz.

Jannidis, Fotis. (2019): „Digitale Geisteswissenschaften: Offene Fragen - schöne Aussichten“. Herausgegeben von Lorenz Engell und Bernhard Siegert. *Zeitschrift für Medien- und Kulturforschung*. https://doi.org/DOI: 10.28937/ZMK-10-1 .

Krajewski, Markus (2019): „Hilfe für die digitale Hilfswissenschaft. Eine Positionsbestimmung“. Herausgegeben von Lorenz Engell und Bernhard Siegert. *Zeitschrift für Medien- und Kulturforschung*. https://doi.org/DOI: 10.28937/ZMK-10-1 .

Peirce, Charles S (1935): *Collected Papers of Charles Sanders Peirce, Volumes V and VI: Pragmatism and Pragmaticism and Scientific Metaphysics*. Herausgegeben von Charles Hartshorne und Paul Weiss. Cambridge, Mass: Belknap Press of Harvard Univ. Press.

Winko, Simone (2003): „Textanalyse“. In *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, herausgegeben von Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, und Klaus Weimar, 3., neubearb. Aufl. Berlin: De Gruyter: 597–601.

Versch. Autoren (2019): „Special Forum on Responses to Nan Z. Da“. Journal of Cultural Analytics. 17. September 2019. https://culturalanalytics.org/2019/09/special-forum-on-responses-to-nan-z-da/ .