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).

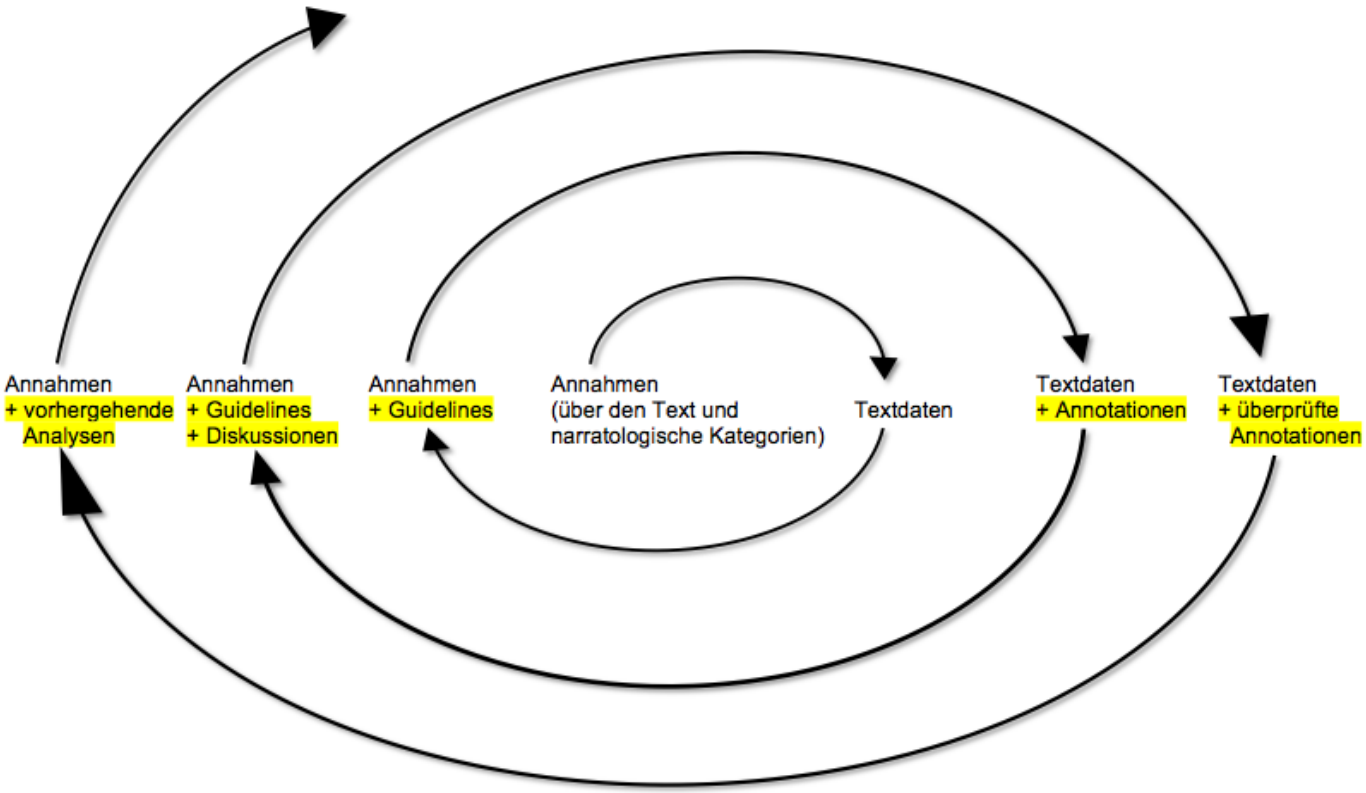

Abbildung 2: Der erweiterte hermeneutische Zirkel

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.

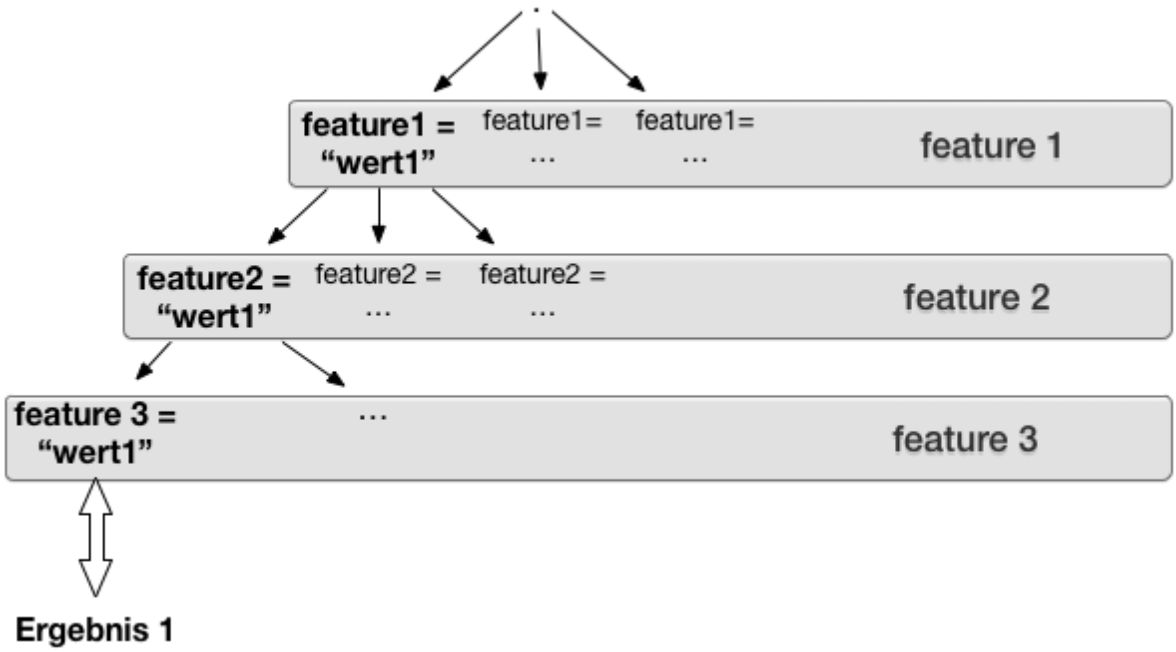

Abbildung 3: Schematische Visualisierung eines Decision Trees.

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.