DE GRUYTER

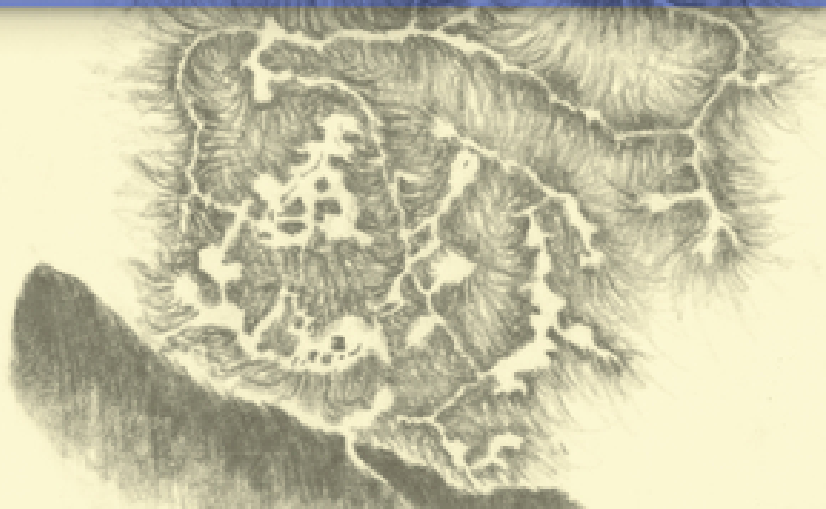
Nils Reiter, Axel Pichler, Jonas Kuhn (Hrsg.)

REFLEKTIERTE ALGORITHMISCHE TEXTANALYSE

INTERDISZIPLINÄRE(S) ARBEITEN IN DER CRETA-WERKSTATT

CRETA

CENTER FOR REFLECTED TEXT ANALYTICS



DE G

Reflektierte Algorithmische Textanalyse

Reflektierte Algorithmische Textanalyse

Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt

Herausgegeben von
Nils Reiter, Axel Pichler und Jonas Kuhn

DE GRUYTER

Gefördert durch das Bundesministerium für Bildung und Forschung.
Förderkennzeichen: 01UG1901.

ISBN 978-3-11-069385-0
e-ISBN (PDF) 978-3-11-069397-3
e-ISBN (EPUB) 978-3-11-069402-4
DOI https://doi.org/10.1515/9783110693973

rt unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz. Weitere Informationen finden Sie unter http://creativecommons.org/licenses/by-nc-nd/4.0/.

Library of Congress Control Number: 2020940471

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

© 2020 Nils Reiter, Axel Pichler und Jonas Kuhn
publiziert von Walter de Gruyter GmbH, Berlin/Boston
Dieses Buch ist als Open-Access-Publikation verfügbar über www.degruyter.com.

Umschlagabbildung: „Plan des Höhlensystems auf Kreta“ von Franz Sieber (1821),
nachgezeichnet von Eliana Heredia
Satz: Nils Reiter
Druck und Bindung: CPI books GmbH, Leck

www.degruyter.com

Inhalt

Teil II: **Methodenentwicklung**

Teil III: **Fallstudien**

Teil IV: **Erzählebenen annotieren: Ein Shared Task**

Teil V: **Ausblick**

Jan Christoph Meister

Geleitwort

Der vorliegende Band versammelt achtzehn Aufsätze, die – im ersten Abschnitt theoretische Aspekte behandelnd, im zweiten auf Grundsatzfragen der Methodenentwicklung fokussierend und im dritten anhand von Fallanalysen und Beispielen – um die Frage kreisen, welche Voraussetzungen eine theoretisch-methodologisch reflektierte Form der algorithmischen Textanalyse zu erfüllen hat. Das Buch präsentiert damit Ergebnisse der Forschungsarbeit eines interdisziplinären Forschungszentrums, das sich seit 2016 systematisch und im Modus eines stringent organisierten interdisziplinären Austauschs zwischen hermeneutisch-geisteswissenschaftlich orientierten Textdisziplinen, formalisierungsaffiner Computerlinguistik und datenmodellierender Informatik mit der Frage beschäftigt, wie praxeologische, methodologische und theoretisch-konzeptionelle Perspektiven in der Methodenentwicklung der Digital Humanities im Teilbereich der Textanalyse fruchtbar kombiniert werden können. Mit diesem ebenso ambitionierten wie klar fokussierten Ansatz hat CRETA, das Centrum für reflektierte Textanalyse an der Universität Stuttgart, mittlerweile national wie international ein herausragendes Profil gewonnen.

Die DH zählen heute, keine zehn Jahre nach der Gründung der Digital Humanities im deutschsprachigen Raum (DHd) im Jahr 2012, zum pluralistischen Methodenrepertoire der Geisteswissenschaften. Selbst von bekennenden Traditionalisten werden sie zunehmend als ein Methodenparadigma wahrgenommen, das den Geisteswissenschaften eine Option auf die Teilhabe am digitalen Zeitalter eröffnet, wenn nicht gar interessante methodische Impulse liefern könnte.

Was dabei allerdings in der Binnen- wie der Außenperspektive leicht in Vergessenheit gerät, ist, dass das neue Methodenparadigma seine gegenwärtige Kontur nicht erst Dank der strategischen Förderinitiativen der letzten zehn Jahre gewonnen hat, mit denen öffentliche und private Drittmittelgeber die Digital Humanities insbesondere im deutschsprachigen Raum entscheidend gefördert haben.[1]

1 Hervorzuheben sind hier im deutschen Kontext neben diversen föderalen und privaten Initiativen insbesondere die einschlägigen Förderlinien der Volkswagenstiftung, der Alexander-von-Humboldt-Stiftung und der Deutschen Forschungsgemeinschaft (DFG). Das Projekt CRETA verdankt seine Existenz der impulsgebenden eHumanities-Förderinitiative des Bundesministeriums für Bildung und Forschung (BMBF).

Jan Christoph Meister, Universität Hamburg

Open Access. © 2020 Jan Christoph Meister, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-001

Der Erfolg der DH ist auch und nicht zuletzt der einer Forschungsidee, die im Verlauf von gut fünfzig Jahren Forschungsarbeit ihren Gegenstandsbereich wie ihre methodisch-theoretische Ausprägung einerseits sukzessive erweitert, dabei jedoch andererseits ihr interdisziplinäres Ethos bewahrt hat. In der Außenwahrnehmung der DH stand dabei oft die technologische Komponente – ‚der Computer' – im Vordergrund. Aber das ist eine probate Verdinglichung, die den Blick auf den konzeptionellen Kern und die historische Dimension eines philosophischen Anliegens verstellt, das bereits bei Raimundus Lullus und Gottfried Wilhelm Leibniz Konturen annimmt und auch in späteren Epochen immer wieder virulent wurde. Dieses Anliegen gilt der Frage, inwieweit nicht nur Natur-, sondern auch Geistesphänomene formal beschrieben, analysiert und rekonstruiert werden können.

Dass so ein Unternehmen irgendwann darauf hinauslaufen müsste, nicht nur das Denken und Schlussfolgern an sich, sondern auch die kulturellen Produkte dieses Denkens und Folgerns formal zu behandeln, hat pikanterweise ausgerechnet der Satiriker Jonathan Swift 1726 als einer der ersten prognostiziert (Swift 1986, Teil III, Kapitel 7). Sein Held *Gulliver* wird Zeuge, wie ein Professor der *Akademie von Lagado* Studenten an einer Maschine kurbelnd nach dem Zufallsprinzip Wortkombinationen generieren lässt. Dabei kommt natürlich viel Unsinn heraus, aber manchmal fügen sich die Worte auch zu sinnvollen Folgen und Sätzen und die werden dann manuell selektiert, um sie wiederum zu Texten zusammenzufügen. Das Ganze ist wie gesagt ein Akademieprojekt: Als Langzeitvorhaben angelegt zielt es deshalb, wie Swift seinen *Gulliver* berichten lässt, darauf, die Universalbibliothek aller nur möglichen Bücher überhaupt zu generieren.

Wer sich um die statistischen und mathematischen Finessen aktueller Verfahren wie *topic modeling* oder *word2vec* nicht weiter bekümmern will, kann Swifts Persiflage der Leibnizschen Kombinatorik natürlich umstandslos als ein Menetekel lesen, das bereits die Kapitulation des geisteswissenschaftlichen Forschungsethos vor der ‚anderen', dem Geistigen inkompatiblen Kultur mathematisch-naturwissenschaftlicher Disziplinen voraussagt. Aber das ist selbst unter dieser problematischen Voraussetzung schlicht zu kurz gedacht: denn der Topos von den ‚zwei Kulturen' des Wissens war schon 1959, als C. P. Snow ihn formulierte, genau das – ein Topos, ein rhetorisches Stilmittel. Als Snow das Bild in seiner berühmten *Rede Lecture* verwendete, um auf die Problematik zweier sich verfestigender, voneinander zunehmend abgeschotteter ‚Kulturen' des Wissens hinzuweisen, tat er das in polemischer Absicht. Bewirken wollte der Bildungspolitiker Snow hingegen die Wiederaufnahme des intellektuellen Diskurses zwischen den *scientists* und den *men of letters*. In der Pflicht sah er dabei in erster Linie die Vertreter der *liberal arts* und die Kulturschaffenden, die Snow einer bornierten, wissenschaftsfeindlichen Gesprächsverweigerung bezichtigte (Snow [1959] 2001).

Dieser kleine Exkurs zeigt: Um ein Forschungsparadigma in seinem Anspruch zu verstehen, sollten wir es nicht nur pragmatisch und methodologisch-theoretisch, sondern auch wissenschafts- und ideengeschichtlich kontextualisieren. Wenn sich heute die Digital Humanities darum bemühen, die der Domäne der Geisteswissenschaften zugerechneten Phänomene der menschlichen Kultur, des Sinnverstehens, der Konstituierung, Kommunikation, Tradierung und Analyse von Bedeutung bis hin zur kritischen Reflexion der Verstehensprozesse selbst auf dem Wege einer formalen Modellierung und in Zusammenarbeit mit Informatikern zu erforschen, dann tragen sie insofern seit einem halben Jahrhundert zur Verwirklichung der im Laufe der Moderne immer wieder eingeforderten domänenübergreifenden Gesprächskultur bei. Und umgekehrt setzt nun die Frage, die im Fokus des vorliegenden Bandes steht – nämlich wie man den grundsätzlich interdisziplinären Forschungsansatz der DH im spezifischen Bezug auf den engeren Anwendungsbereich der algorithmischen Textanalyse modular organisieren, schrittweise ausbauen, systematisch validieren und zyklisch optimieren kann –, an genau der Problemkonstellation an, die für die Entwicklung der digitalen Geisteswissenschaften insgesamt die frühesten Impulse geliefert hat. Diese Konstellation lässt sich vielleicht am besten so umreißen: Wie lassen sich das komplexe Phänomen Text, das semiotische System der natürlichen Sprache und die Methode der formalen Modellierung und Analyse von Objekten und Prozessen mit rechnergestützten Verfahren so zueinander in Beziehung setzen, dass daraus beispielhaft eine neue epistemische Praxis für die Geisteswissenschaften resultiert?

Diese Frage begleitet uns seit den Anfängen der disziplinären Formierung unseres Forschungsfeldes. *Nomen est omen*: Schon 1964 gründete der Germanist und Mediävist Roy Wisbey an der University of Cambridge das erste *Literary and Linguistic Computing Centre* und 1970 veranstaltete er dort zusammen mit Michael Farringdon auch das erste internationale Symposium „The Computer in Literary and Linguistic Research". Aus dem Veranstaltungstitel leitete sich dann der Name für den 1973 am King's College London gegründeten ersten Fachverband ab: *Association for Literary and Linguistic Computing* (ALLC). Einen programmatischen Fokus auf die Trias von Literatur, Linguistik und Computing signalisierte in gleicher Weise die von 1973 bis 1988 jährlich ausgerichtete *ALLC Conference* und ebenso die erste einschlägige Fachzeitschrift, das *Journal of Literary and Linguistic Computing*. Erst als in den 1990er Jahren das Spektrum digitalisierter Medienformate breiter wurde, setzte sich allmählich die Bezeichnung *Humanities Computing* durch. Die heute international geläufige Bezeichnung *Digital Humanities* kam hingegen erst nach der Jahrtausendwende auf; wir verdanken sie, wie sich John Unsworth erinnert hat, einer Entscheidung, die einen gelungenen Kompromiss zwischen den Interessen von Verlagsmarketing und wissenschaftlicher Profilbildung darstellt:

> The real origin of that term [digital humanities] was in conversation with Andrew McNeillie, the original acquiring editor for the Blackwell *Companion to Digital Humanities*. We started talking with him about that book project in 2001, in April, and by the end of November we'd lined up contributors and were discussing the title, for the contract. Ray [Siemens] wanted "A Companion to Humanities Computing" as that was the term commonly used at that point; the editorial and marketing folks at Blackwell wanted "Companion to Digitized Humanities." I suggested "Companion to Digital Humanities" to shift the emphasis away from simple digitization. (Kirschenbaum 2010)

Wir dürfen die Nuancierung der wechselnden Bezeichnungen *Literary and Linguistic Computing* über *Humanities Computing* bis hin *zu Digital Humanities* also nicht überinterpretieren; Namensgebungen können auch kontingenten Faktoren geschuldet sein. Aber die Abfolge der Bezeichnungen macht immerhin deutlich, dass neben der schrittweisen Erweiterung des Skopus vom engeren Gegenstandsbereich der Textwissenschaften hin zur Gesamtdomäne der Geisteswissenschaften sich zugleich das Bewusstsein für den methodologischen Effekt einer Kombination von verstehensorientiert-interpretierenden mit formal-modellierenden Verfahren geschärft hat. Statt um eine ‚simple digitization' von Objekten und Verfahren der traditionellen Geisteswissenschaften mit den Techniken der digitalen Datenerfassung und Datenmanipulation ging und geht es in den DH dabei um eine methodologische Erweiterung und Bereicherung der Geisteswissenschaften.

Dazu ist nun allerdings ein Brückenschlag erforderlich, denn insbesondere der hermeneutisch orientierte geisteswissenschaftliche Forschungsansatz ist seiner Natur nach zunächst phänomenologisch motiviert: Die traditionellen, idealtypisch im 19. Jahrhundert entstandenen Geisteswissenschaften fragen nach der Bedeutung und lebensweltlichen Relevanz symbolischer Artefakte für den Menschen; sie begnügen sich nicht mit der abstrakten und interessefreien analytischen Bestimmung von Phänomenen als Objekten an sich. Erst im 20. Jahrhundert kommen mit der strukturalen Linguistik und der formalistischen Erzählforschung die ersten Disziplinen auf, für die nicht mehr diese phänomenologisch-historische Perspektive auf die Wirkung eines Zeichens auf und für den Menschen und die Kultur im Vordergrund steht, sondern die analytische auf das immanente, dynamische Funktionsgefüge des Zeichenkomplexes und dessen Prozesslogik. Diese zweite Fragehaltung setzt allerdings mindestens zweierlei voraus: erstens die Bereitschaft, im Interesse des Erkenntnisfortschritts vorübergehend von der Problematik historischer Kontingenz und der Kontextualisierung des Beobachtungsfeldes zu abstrahieren, es also gewissermaßen bewusst ‚stillzustellen'; zweitens die Bereitschaft, das lebensweltlich als ein auf uns wirkendes Ganzes erfahrene – die Sprache, die Erzählung, das Kunstwerk, die Gesellschaft etc. – als ein aus diskreten Elementen zusammengesetztes zu konzeptualisieren. Denn eben dies sind die beiden notwendigen Voraussetzungen einer digitalen Konzeptuali-

sierung: Abstraktion von der je konkreten Kontextgebundenheit, Erfassung und Modellierung der Objektdomäne in Form diskreter Beobachtungsinstanzen.

Schon Willard McCarty hat in seiner 2005 vorgelegten Monographie *Humanities Computing* allerdings darauf aufmerksam gemacht, dass das Projekt der Digital Humanities sich nicht auf ein Pipelinemodell reduzieren lässt, in dem geisteswissenschaftliche ‚Probleme' isoliert, formal beschrieben, dann an die Informatik weitergereicht, dort operationalisiert und mit teils generischen, teils spezialisierten Tools bearbeitet und als ‚gelöst' an die Geisteswissenschaften rückgemeldet werden, um zum nächsten ‚Problem' voranzuschreiten. Ans Ende seiner Überlegungen stellt McCarty *A preliminary Agenda*. Sie umfasst zehn Punkte, von denen heute, fünfzehn Jahre später, die letzten neun weniger dringlich und aktuell erscheinen mögen – nicht aber der erste Punkt ‚Analysis'. Dazu schreibt McCarty:

> The first item is to understand how computing affects analysis in the humanities beyond simply fetching, counting and formatting data – how, in other words, it affects analysis itself rather than its scope, speed or convenience. [...] I have placed this question first on the agenda because otherwise we elide the profound difference between artifact and data and so are in no position to understand the significance and reception of the results we obtain from these data. We need to know what that difference is, and how an analysis based on it might differ from what we have been doing all along. (McCarty 2005, S. 206)

Wie aber können wir diese Fragestellung in der konkreten interdisziplinären Forschungsarbeit nicht nur abstrakt bewusst halten, sondern sie konzeptionell so in den Forschungsprozess einfließen lassen, dass daraus mehr wird als ein fruchtloses philosophisches *ceterum censeo*?

CRETA hat dieses Problem mit einem Ansatz zur reflektierten Erarbeitung textanalytischer DH-Workflows gelöst, der nach meiner Einschätzung vorbildhaft ist. Dazu trägt erstens die ‚Werkstatt-Idee' bei, die als organisatorisches Prinzip einen explorativen, bottom-up Austausch zwischen den beteiligten Disziplinen ermöglicht. Zweitens berücksichtigt der Ansatz von CRETA die geisteswissenschaftliche Erfordernis, nach der Abstraktion vom Kontext, die für den formal-modellierenden Zugang Voraussetzung ist, auch wieder zum Kontext zurückzukehren – und dann womöglich in einen neuen Zyklus einzutreten, der ein komplexeres Gegenstandsmodell zur Folge hat. Die interdisziplinäre Forschung ist damit konzeptionell als ein iterativer Prozess organisiert[2], der wechselseitig befruchtend sein kann, statt eine modulare „Entkoppelung der disziplinären Fel-

2 Siehe hierzu Abb. 1, „Prototypischer Arbeitsablauf" in Pichler und Reiter 2020, S. 44.

der“ (Kuhn 2020b, S. 19) zu erzwingen.[3] Drittens lernen die an dem CRETA-Projekt beteiligten Disziplinen voneinander nicht nur in Bezug auf die Konzepte, Modelle und Verfahren, die bei der Erarbeitung, Validierung und Anwendung von Prozessen und Werkzeugen der algorithmischen Textanalyse im Vordergrund stehen. Sie begegnen darüber hinaus auch einer fremden disziplinären Forschungspraxis, reflektieren diese und erhalten so Gelegenheit, etablierte Routinen in die eigene Praxis zu übernehmen, wo dies sinnvoll erscheint. Ein Beispiel für den letzten Punkt und damit zugleich für die programmatische Dimension von CRETA diskutiert Kuhn (2020a) ab Seite 84; hier wird wegweisend die Möglichkeit einer Integration des in der Informatik etablierten *rapid probing of analysis models* mit einem hermeneutischen Ansatz erörtert.

Die Beiträge des vorliegenden Bandes dokumentieren so einerseits auf beeindruckende Weise, wie und mit welchen Ergebnissen die Kooperationspartner in CRETA die praxeologische Aufgabe, „technische Methoden und Arbeitsablaufpraktiken zur Textanalyse im Forschungsbereich der DH zu entwickeln“ (Pichler und Reiter 2020, S. 43), angegangen sind. Aber der Band als ganzer ist zugleich mehr als die Summe seiner Teile: Er demonstriert auch das ambitionierte methodologische Niveau, auf dem sich die Digital Humanities als interdisziplinäres Forschungsprojekt im Schnittfeld von Geisteswissenschaften und Informatik mittlerweile bewegen können. Die Voraussetzung dafür signalisiert CRETA, das „Centrum für reflektierte Textanalyse“, mit der Wahl des Adjektivs in seinem Namen. Es sei hiermit doppelt unterstrichen.

Literatur

Kirschenbaum, Matthew G. (2010). „What Is Digital Humanities and What’s It Doing in English Departments?“ In: *ADE Bulletin* 150. URL: https://mkirschenbaum.files.wordpress.com/2011/03/ade-final.pdf (besucht am 1. Juni 2020).

Kuhn, Jonas (2020a). „Computational Text Analysis within the Humanities“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 63–106.

Kuhn, Jonas (2020b). „Einleitung“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 9–40.

McCarty, Willard (2005). *Humanities Computing*. Houndmills/New York: Palgrave Macmillan.

3 Das schließt nach der CRETA-Philosophie jedoch umgekehrt eine Wahl des Domänenexperten/Entwickler-Modells nicht aus, wo diese pragmatisch begründet werden kann. Siehe Kuhn 2020b, S. 23, Fußnote 13

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Snow, Charles Percy [1959] (2001). *The Two Cultures*. London: Cambridge University Press.

Swift, Jonathan (1986). *Gulliver's Travels*. Wiederauflage 2008. Oxford: Oxford University Press.

Jonas Kuhn

Einleitung

Das vorliegende Buch gibt mit einer Reihe von Aufsätzen Einblicke in die ‚Werkstatt' des Stuttgarter *Center for Reflected Text Analytics* (CRETA), das 2016 mit Förderung durch das Bundesministerium für Bildung und Forschung (BMBF) seine Arbeit als interdisziplinäres Zentrum für die Methodenentwicklung in den digitalen Geisteswissenschaften oder Digital Humanities (DH) aufnahm. Die Kooperation in CRETA[1] zur algorithmischen Textanalyse setzt den Gedanken der Interdisziplinarität in zwei Dimensionen gleichzeitig um: Erstens kombiniert das Zentrum Modellierungsansätze, Arbeitspraktiken und methodisches Wissen aus informatiknahen Fächern einerseits und textwissenschaftlichen Disziplinen der Geistes- und Sozialwissenschaften andererseits – also entlang der Hauptachse der DH, die man sich als Horizontale denken kann. Zugleich führt CRETA orthogonal dazu – gleichsam entlang einer vertikalen Achse – *innerhalb* von Clustern methodisch verwandter Fächer die jeweiligen disziplinspezifischen Blickwinkel zusammen. So werden komplementäre textwissenschaftliche Herangehensweisen in Literaturwissenschaften, Philosophie, Politikwissenschaft und Linguistik beleuchtet und Techniken aus Computerlinguistik, maschinellem Lernen und *Visualisierung* miteinander verbunden, um jeweils die einzusetzenden Analyseschritte optimal an Gegenstand und Fragekontext anzupassen. (Eine ausführlichere Darstellung des disziplinübergreifenden Selbstverständnis und der Zielsetzung für CRETA folgt in Abschnitt 2.2.)

Ein interdisziplinäres Vorgehen erfordert immer einen Abstimmungsprozess und dieser muss umso intensiver geführt werden, je stärker sich die methodischen Grundansätze unterscheiden und je mehr unterschiedliche Fachdisziplinen beteiligt sind (Reiter, Kremer et al. 2020 dokumentiert ausführlich, mit welchen Maßnahmen wir in CRETA diesen Abstimmungsprozess gestaltet haben). Die Erfahrung in CRETA zeigt jedoch, dass sich eine Abstimmung rasch bei der Entwicklung von effektiven algorithmischen Ansätzen für die Textanalyse auszahlt – etwa für die Extraktion von Relationen zwischen zentralen Akteuren oder Konzepten in einem Text, die den systematischen Vergleich von Texten in einem Korpus

1 Wir folgen der Sprechweise, die sich im Umfeld des Zentrums rasch etabliert hat, und verwenden das Akronym wie einen Eigennamen für einen abstrakten Ort, der für die zugrundeliegende disziplinübergreifende Arbeitspraxis steht.

Jonas Kuhn, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Jonas Kuhn, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-002

unterstützt. Gewinnbringend ist die Abstimmung nicht nur in der horizontalen Dimension mit ihren offensichtlichen Schnittstellen zwischen technischen Analyseverfahren einerseits und inhaltlich-konzeptuellem Wissen andererseits, sondern gerade auch in der vertikalen Dimension: in unterschiedlichen textwissenschaftlichen Disziplinen sind es häufig unterschiedliche Facetten der Versprachlichung von relevanten Relationen, die im Zentrum der jeweils etablierten Analyseansätze stehen; ein Einbeziehen der Sichtweise aus Nachbardisziplinen erleichtert jedoch nicht selten das praktische Vorgehen bei der Korpusanalyse – greifen doch beim Textverständnis sämtliche Facetten ineinander. Das Nebeneinanderstellen von disziplinspezifischen Analyseansätzen schärft den Blick auf deren Eigenheiten und kann so erheblich zu einer fundierten textbasierten Charakterisierung der Konzepte und ihrer Operationalisierung beitragen. So erweist sich – entgegen einer denkbaren Befürchtung, es könnte sich als problematisch erweisen, die textanalytischen Traditionen unterschiedlicher Disziplinen zu ‚vermischen': Gerade wenn sich die Akzentsetzung in der Textbetrachtung über textwissenschaftliche Fachdisziplinen hinweg stark unterscheidet, kann die vergleichende Betrachtung einen spürbaren Mehrwert erzeugen – vorausgesetzt, die Beteiligten bringen die nötige Offenheit mit, den etablierten Ansatz des eigenen Fachs um zusätzliche Facetten zu ergänzen.

Jenseits des Zusammenfügens von textanalytischen Facetten für die Behandlung eines konkreten Textgegenstands liegt die Zielsetzung für CRETA darin, wirksame Leitlinien für eine Praxis der reflektierten algorithmischen Textanalyse zu identifizieren – im Rahmen eines Vorgehens, das beide genannten Dimensionen der interdisziplinären Kooperation berücksichtigt. Erhebt CRETA damit den Anspruch, *das* umfassende Methodengebäude für die DH schlechthin zu entwickeln oder vorzubereiten? Keineswegs. Dies würde auch nicht zum Selbstverständnis passen, das sich in der internationalen Forschungscommunity entwickelt hat: Die Exponenten der DH oder der digitalen Geisteswissenschaften gehen nicht von einem fixen und übergreifenden Kanon von Methoden und Arbeitspraktiken aus. Vielmehr charakterisiert sich das Feld gerne offen als die Gesamtheit aller wissenschaftlichen Aktivitäten, in denen die Möglichkeiten des Computers bei der Bearbeitung von Forschungsfragen und -gegenständen aus den Geisteswissenschaften ausgenutzt werden.[2] Die meisten DH-Tagungen und -Workshops zeichnen sich in der Tat durch einen Methodenpluralismus aus und die Community begegnet Versuchen, neue Ansätze einzubringen, generell mit Offenheit und Wohlwollen. So erlebten es auch die Forschungsteams, in denen einige der späteren CRETA-Beteiligten Anfang der 2010er-Jahre Strategien zur Integration von Verfah-

2 Siehe z. B. Kirschenbaum 2012.

ren aus der Computerlinguistik und den *visual analytics* in die Forschung zu Untersuchungsgegenständen aus den Geistes- und Sozialwissenschaften erkundeten[3] – mit dem erklärten Ziel, die eingesetzten komputationellen Analyseverfahren sorgfältig an die jeweilige Forschungsfrage anzupassen und in *workflows* einzubinden, die eine fundierte empirische Evaluation und theoretische Reflexion ermöglichen. Die Akzentuierung der Adaptierbarkeit und einer modularen Kombination von kontrollierbaren Komponenten, die diesem Zugang zugrundeliegt, wurde vielfach als Ergänzung der Modellierungspraxis in dem schon wesentlich länger etablierten Bereich des *computing in the humanities* bzw. *Humanities Computing*[4] begrüßt.

Gerade aufgrund der praktizierten Methodenvielfalt in der DH-Forschungslandschaft ist es allerdings ratsam für ein einzelnes Projekt oder ein systematisch vorgehendes Zentrum wie CRETA, das arbeitspraktische Selbstverständnis klar zu bestimmen. Und auch für die jeweilige Forschungscommunity in DH-Teilfeldern ist eine Verständigung auf methodische Standards und ein geteiltes Problemverständnis ein wichtiger Schritt bei der Etablierung als wissenschaftliche Disziplin, etwa um Konsens zu den Spielregeln in *peer-reviewing*-Verfahren herzustellen.

In der Tat verfolgt CRETA trotz der dargestellten interdisziplinären Breite methodisch-arbeitspraktisch eine sehr spezifische Agenda, die in den folgenden Abschnitten und ab Seite 43 im Beitrag von Pichler und Reiter (2020) weiter ausgeführt wird. Die vorliegende Einleitung will vier Dinge leisten: Abschnitt 1 umreißt den Modellierungsbegriff, auf den sich die fachübergreifende Arbeit in CRETA stützt. Anschließend soll in Abschnitt 2 knapp nachgezeichnet werden, aus welcher Motivation heraus der Fokus auf die reflektierte algorithmische Textanalyse in CRETA entwickelt wurde (2.1) und welche konkrete Zielsetzung CRETA verfolgt (2.2). Abschnitt 3 geht auf strukturelle Herausforderungen ein, die mit der interdisziplinären Breite der DH-Forschung einhergehen und denen wir unter anderem durch das Konzept der CRETA-Werkstätten begegnen. Abschließend bietet diese Einleitung in Abschnitt 4 Antworten auf die Fragen, an wen sich der vorliegende Band richtet und wie er gelesen werden kann.

3 Die Teams des Projekts ‚e-Identity' (2012–2015, gefördert durch das BMBF), unter Federführung der Politikwissenschaftlerin Cathleen Kantner (Stuttgart) mit Ko-Projektleitern Ulrich Heid (Hildesheim), Jonas Kuhn (Stuttgart) und Manfred Stede (Potsdam), und des Projekts ‚ePoetics' (2013-2016, ebenfalls vom BMBF gefördert), unter Federführung der Literaturwissenschaftlerin Sandra Richter (Stuttgart) mit Ko-Projektleiterin Andrea Rapp (Darmstadt, Computerphilologie) und Ko-Projektleitern Thomas Ertl und Jonas Kuhn (beide Stuttgart).

4 Zur Entwicklung des Feldes vgl. McCarty 2005, S. 2 ff.

1 Hintergrund: Modelle und Modellierung in der algorithmischen Textanalyse

Bevor wir im Kern dieser Einleitung zur Motivation von CRETA als einem Zentrum für reflektierte algorithmische Textanalyse kommen können, sind einige Bemerkungen zum Modellbegriff angebracht. Die Rede von Modellen und das Verständnis, das mit dem Prozess der ‚Modellierung' verbunden ist, dürften für zahlreiche Missverständnisse in interdisziplinären Arbeitsteams verantwortlich sein. Im Jargon unterschiedlicher wissenschaftlicher Disziplinen ist die Rede von der ‚Modellierung' jeweils mit (zu Teilen impliziten) Annahmen über die etablierten arbeitspraktischen Abläufe und Kontexte verbunden. Zwar mögen sich die meisten interdisziplinären Kooperationsteams dieser terminologischen Problematik grundsätzlich bewusst sein. Doch bleibt es trotz eines solchen Bewusstseins bisweilen schwierig festzustellen, welche der jeweiligen Grundannahmen, die im disziplininternen Alltag unproblematisiert vorausgesetzt werden, im interdisziplinären Zusammenhang wechselseitig vereinbar sind, welche dagegen einer Klärung bedürfen. In CRETA (und auch in diesem Band) hat es sich daher als hilfreich erwiesen, mit dem Modellbegriff vorsichtig umzugehen.

Wenn wir den Begriff ‚Modell' ohne weitere Qualifikation verwenden (oder synonym den Begriff ‚Computermodell'), folgen wir in der Regel dem Methodenverständnis der datenorientierten Computerlinguistik/Informatik. Unter einem Modell wird hier ein (zumeist komplexes) algorithmisches System verstanden, das auf dem Computer eine Eingabe/Ausgabe-Funktion implementiert, die einen empirischen Prozess approximiert – wie etwa die Weiterentwicklung von meteorologischen Messdaten (Temperatur, Luftdruck) bei einem gegebenen Verlauf oder die Zuordnung einer handschriftlich geschriebenen Ziffer zu einer von zehn möglichen Interpretationskategorien (‚0'–‚9'). Das Modell ist also eine vereinfachende Abbildung eines Ausschnitts der Realität (siehe u. a. Thalheim und Nissen 2015, S. 13 ff.) und der Prozess der Modellierung ist die Suche nach einer Modellinstanz, bei der die Abbildung der Realität nach bestimmten Gütekriterien optimal ausfällt.[5] In der angewandten Forschung und der Technologieentwicklung hat sich

5 Großes Gewicht in den DH und verwandten Forschungsgebieten kommt neben der Approximierung von empirischen Prozessen auch der systematischen Charakterisierung und Standardisierung der Relation von Datenelementen untereinander und ihrem Bezug zu repräsentierten Entitäten zu. Auch dieses Aufgabenfeld – die *Datenmodellierung* – wird nicht selten schlicht als ‚Modellierung' bezeichnet. Im Einzelfall dürfte der jeweilige Verwendungskontext die Verwechslungsgefahr minimieren, dennoch ist gerade in den DH terminologisch Vorsicht geboten, da mit beiden Modellbegriffen wichtige Entwicklungen in der Forschungscommunity verbunden sind.

ein Vorgehen etabliert, bei dem in einem frühen Stadium ein geeigneter Realitätsausschnitt bestimmt wird, der als hinreichend repräsentativ angesetzt wird (beispielsweise historische Wetterverlaufsdaten über einen Zeitraum von zehn Jahren oder eine große Sammlung von handschriftlichen Ziffern und deren Interpretation durch menschliche Leserinnen). Die Bestimmung einer geeigneten vereinfachenden Abbildung bezeichnet man hier als ‚Domänenmodellierung'. Um für das jeweilige Anwendungsszenario die besten Ergebnisse zu erzielen, kommt hierbei (wohlreflektierten) pragmatischen Überlegungen häufig erhebliche Bedeutung zu: Soll etwa ein Wettervorhersagemodell für Mitteleuropa entwickelt werden, ist es sinnvoll, den Realitätsausschnitt entsprechend zu wählen; wenn Postleitzahlen in deutschen Adressen erkannt werden sollen, ist die Annahme, dass nur Ziffern vorkommen, sinnvoll – bei britischen Post-Codes müssten hingegen auch Buchstaben erwartet werden. (Zu den pragmatisch gegebenen Faktoren bei der Modellierung kann es auch gehören, dass für den Zielausschnitt aus der Realität selbst keine oder nur unzureichende Entwicklungsdaten vorliegen, so dass behelfsweise andere Daten hinzugezogen werden, die in geeigneter Weise als parallel betrachtet werden können.)

Nach der Bestimmung eines repräsentativen Datenausschnitts und eines geeigneten Abstraktionsgrads für die Eingabe/Ausgabe-Repräsentationen des gesuchten Modells reduziert sich die Suche nach einer optimalen Modellinstanz auf ein mathematisches bzw. algorithmisches Optimierungsproblem: die Vorhersage des Modells soll möglichst nahe an den gewählten empirischen Ausschnitt herankommen – sich also mit dem beobachteten Verlauf bzw. der Kategorisierung decken. Die Festlegung eines Referenz-Datenausschnitts erleichtert die Arbeit mit komplexen Computermodellen erheblich: Zwischen unterschiedlichen denkbaren Modellarchitekturen/Modellklassen muss ausgewählt werden und innerhalb jeder Architektur und Klasse ist eine Vielzahl von Modellparametern anzupassen. Da eine exakte Bestimmung der optimalen Modellparameter häufig unmöglich ist, kommen approximative Optimierungsprozesse zum Einsatz. Die Festlegung von Referenzdaten erlaubt eine Entkopplung der Optimierungsaufgabe von empirischen Fragen der Modelladäquatheit.

Für viele Problemstellungen, für die es eine theoretisch fundierte Forschungstradition gibt (beispielsweise grammatische Kategorisierungen wie die Bestimmung der Wortkategorie – Nomen, Verb, Adjektiv usw. – einer Wortform im Verwendungskontext), hat sich erwiesen, dass Computermodelle, die mit *machine-*

So hebt dem Sammelband Flanders und Jannidis 2019 zur Modellierung von Text und textbasierten Ressourcen in den DH auf vielgschichtige Fragen der Repräsentation von Datenobjekten ab, ausgehend von der *Text Encoding Initiative*.

learning-Verfahren auf repräsentativen Daten trainiert wurden, ein sehr robustes Vorhersageverhalten an den Tag legen. Basierend auf solchen Befunden hat sich das Einsatzgebiet für datenorientierte Computermodelle entwickelt, das nicht nur den meisten neueren Sprachtechnologie-Anwendungen zugrunde liegt, sondern auch die Basis für vielfältige Arbeiten mit einem kognitionswissenschaftlichen Erkenntnisinteresse bildet (eine vergleichsweise lange Tradition der datenbasierte Computermodellierung besteht etwa im Bereich Psycholinguistik, siehe z. B. Altmann 1990). Die Grundidee ist folgende: Um zu entscheiden, welche von mehreren denkbaren Abstraktionen im Zuge der (‚Domänen'-)Modellierung besser geeignet ist, einen forschungsrelevanten Ausschnitt der Realität abzubilden, werden mit den erwähnten Optimierungsverfahren jeweils prädiktive Modellinstanzen trainiert. Wenn nun eine Abstraktionsidee zu einer höheren empirischen Vorhersagegenauigkeit führt als eine andere, kann dies als indirekter empirischer Befund für die Güte der Abstraktionsidee gewertet werden. Die datenorientierte Optimierung von prädiktiven Modellen wird so also aus dem rein technologisch motivierten Kontext gelöst; anwendungsspezifische Überlegungen bei der Wahl von Vereinfachungen/Abstraktionen im Zuge der ‚Domänenmodellierung' werden abgelöst durch theoretische Überlegungen, die sich aus übergeordneten Forschungsproblemen ableiten (wobei weiterhin pragmatischen Faktoren bei der Auswahl von Daten für die Modellentwicklung eine große Rolle spielen).

Eine Variante des skizzierten Vorgehens liegt auch aktuellen datenorientierten Modellierungsansätzen in den DH zugrunde (im Anschluss an das Modellierungsverständnisses etwa von McCarty (2005), wie unten in Abschnitt 2.1 ausgeführt).[6] Es ist hervorzuheben, dass die fortlaufende datenorientierte Modellierung innerhalb eines solchen arbeitspraktischen Selbstverständnisses Mittel zum Zweck auf dem Weg zu einem systematischen Verständnis des Forschungsgegenstandes ist. Unterschiedliche Modelle fokussieren jeweils auf bestimmte relevante Realitätsausschnitte. Ein umfassendes Modell erarbeiten zu wollen, das jenseits einer Einzelaufgabe alle relevanten Überlegungen geeignet erfasst, entspricht nicht dem Modellverständnis der DH.

6 Den Status der Modellierung als einem kreativen Prozess bei der Wissensgenerierung zu einem Gegenstand diskutieren Ciula und Eide 2016 auf Grundlage einer semiotischen Konzeptualisierung.

2 Das Selbstverständnis von CRETA

2.1 ‚Die Digitalisierung' und die Wissenschaftspraxis in den 2010er-Jahren: große Erwartungen

Reflektierte algorithmische Textanalyse erfordert nach dem Selbstverständnis von CRETA ein sehr differenziertes Vorgehen beim Zusammenfügen der Arbeitspraktiken aus Informatik und Textwissenschaften. Zur Motivation der Grundsätze ist es hilfreich, zunächst ein simpleres Zusammenspiel zwischen geistes- und sozialwissenschaftlichen Fragestellungen und den technologischen Möglichkeiten der Informationstechnologie zu skizzieren, das durchaus denkbar ist. Wie sich im folgenden zeigt, ist jedoch auch der erzielbare Nutzen eingeschränkt.

Die weltweite Vernetzung von Informationsquellen, der unkomplizierte Zugriff auf dieses Netz für jeden Interessierten und jede Interessierte und nicht zuletzt die Möglichkeit, dieser gewaltigen Sammlung mit wenig Aufwand eigene Beiträge hinzuzufügen, hat in kaum drei Jahrzehnten nach der Einführung des Internets den Umgang mit Information – etwa bei der Generierung und dem Austausch von Wissen – grundlegend verändert. Neben etablierte Kulturtechniken wie das Lesen, das Schreiben oder die Verwendung von konventionellen Bibliothekskatalogen treten neue werkzeugbasierte Techniken: so ist mit minimalem Aufwand eine satellitenbasierte Geo-Lokalisierung von Bildern oder Textnachrichten möglich; und es entwickeln sich Varianten konventioneller Techniken wie das kollaborative Verfassen etwa von Wikipedia-Artikeln oder die Recherche von Informationsquellen mit Hilfe von modernen Suchmaschinen und den Algorithmen, die sich hinter der Indizierung von Dokumenten und einer nutzerorientierten Optimierung der Ergebnispräsentation verstecken. Wie jeder gesellschaftliche Bereich sind die Wissenschaften von dieser Transformation umfassend betroffen. Unter dem Schlagwort ‚Digitalisierung' wird mit zunehmender Intensität in den unterschiedlichen Feldern und Disziplinen diskutiert, wie mit den Chancen und Herausforderungen der technologischen Möglichkeiten umzugehen sei.

Gerade für die Geistes- und Kulturwissenschaften birgt ‚die Digitalisierung' erhebliches Potenzial – darin sind sich viele Fachwissenschaftlerinnen und -wissenschaftler einig mit Laien, die die Situation beispielsweise aus der Informationstechnologie oder aus der Politik heraus betrachten. Zeitraubende Prozesse des Auffindens, Sichtens und Erschließens von (möglicherweise) relevantem Untersuchungsmaterial und von Befunden aus Forschungsarbeiten, die in einem ganz anderen Fragezusammenhang entstanden sind, sollten sich in einer digital vernetzten Welt erheblich schneller und zuverlässiger umsetzen lassen. So sollte es möglich werden, Fragen zu einem breiteren Spektrum von Untersuchungsge-

genständen empirisch zu untersuchen und eine vielschichtigere Betrachtung von Kontextfaktoren vorzunehmen. Auf den ersten Blick scheint zudem keine spezielle technologische Unterstützung für den Bedarf der Geistes- und Kulturwissenschaften notwendig zu sein. Die Anforderungen an eine Rechercheumgebung erscheinen ähnlich wie bei jeder großen und heterogenen Informationssammlung – etwa dem Intranet eines Unternehmens oder dem Internet: Sind umfassende Sammlungen von Untersuchungsgegenständen und Quellen für die geisteswissenschaftliche Forschung erst einmal in digitalisierter Form erschlossen und vernetzt – so die Überlegung –, sollte sich im Zuge einzelner Forschungsstudien das gerade angedeutete Potenzial mit Standardverfahren z. B. aus dem *information retrieval* abrufen lassen. Oder falls die verfügbaren Standardverfahren für die Textquellen nicht unmittelbar anwendbar sind, sollten sie sich mit überschaubarem Aufwand an deren spezielle Eigenschaften anpassen lassen. Die beiden Nadelöhre lägen danach also in der Überführung der Primärdaten in geeignete digitale Formate und in einer Befähigung der Forschenden zu einem möglichst effektiven Einsatz der technologischen Werkzeugangebote, also der Schulung hin zu einer breiten IT-Kompetenz bzw. *information technology literacy*.

Diese Charakterisierung des Status Quo dürfte *grosso modo* zutreffend sein, wenn sich die Ausnutzung von algorithmischen Verfahren ausschließlich auf Vorarbeiten zu einer im Kern konventionellen geisteswissenschaftlichen Studie beschränkt, deren Textbasis zudem nur geringfügig von den Standard-Entwicklungskorpora für sprachtechnologische Werkzeuge abweicht.

Mindestens vier Überlegungen sprechen jedoch dafür, eine stärkere Verschränkung von Aspekten des geisteswissenschaftlichen Forschungsprozesses mit algorithmischen Verfahren zu erkunden: Erstens bedeutet die Beschränkung des Einsatzes von algorithmischen Verfahren auf Vorarbeiten (etwa die Sichtung eines großen Bestands von potenziell relevanten Texten), dass das Skalierungspotenzial – also die Möglichkeit, eine größere Menge von Untersuchungsgegenständen zu betrachten als dies bei einem manuellen Vorgehen möglich ist – auch nur sehr eingeschränkt ausgenutzt wird: nach Abschluss der Vorarbeiten müssen die Untersuchungsgegenstände mit konventionellen Mitteln untersucht werden. Zwar mag der Eindruck entstehen, dass informationstechnologische Standardverfahren wie etwa eine textuelle Suche nach Schlüsselbegriffen eine erhebliche Ausweitung des erfassbaren Korpus ermöglichen. Die wenigsten zu untersuchenden Konzepte sind jedoch an der Textoberfläche zweifelsfrei zu erfassen, so dass entweder eine manuelle Untersuchung nachgeschaltet werden muss (die der Skalierbarkeit entgegenwirkt) oder die textuelle Suche um einen spezielleren algorithmischen Analyseschritt ergänzt wird – womit der Schritt hin zu einer Verschränkung der algorithmischen Analysemethodik mit Aspekten der fachspezifischen Fragestellung vollzogen wäre.

Zum zweiten erweist sich eine Abkoppelung der algorithmischen Schritte, die ausschließlich im Zuge von Vorarbeiten eingesetzt werden, von der Forschungsfrage, die im eigentlichen Kern verfolgt wird, in der Praxis sehr häufig als nicht praktikabel – gerade unter einem im weitesten Sinne hermeneutischen Vorgehen, bei dem die Auseinandersetzung mit dem Gegenstand während der Sichtungsphase mit einer Verfeinerung der Forschungsfrage einhergeht:[7] Im Zuge eines geisteswissenschaftlich motivierten Vorgehens kommt es immer wieder zu einer aus dem Zusammenspiel von Forschungsfrage und Textempirie resultierenden Kontexterweiterung, die wiederum neues Licht auf deutungsrelevante Textphänomene wirft. Wird eine solche Erweiterung vollzogen, müssten die algorithmischen Schritte aus den vermeintlich abgekoppelten Vorarbeiten auf den Prüfstand gestellt werden: Sind sie mit dem revidierten Vorverständnis immer noch kompatibel oder müssen bestimmte Analyseschritte wiederholt werden? Effektiv würde so also erneut eine Verschränkung des methodischen Vorgehens erzwungen.

Drittens eignen sich – unabhängig vom gerade angesprochenen dynamischen Aspekt einer Kontexterweiterung im Zuge der Textanalyse – viele der etablierten technologischen Standardverfahren für skalierbare Text- und Datenanalyse nur bedingt dazu, Analysebefunde zu einem Untersuchungsgegenstand in einer Weise auf die Forschungsfrage zu beziehen, die der geisteswissenschaftlichen Herangehensweise vollständig gerecht wird. Dies dürfte mit grundsätzlichen Unterschieden im Umgang mit kontextuellen Einflussfaktoren zusammenhängen: Bei der ingenieursmäßigen Konzeption etwa von Sprach- oder Textanalysetechnologien wird versucht, soweit wie möglich von Eigenschaften des Kontexts zu abstrahieren und damit möglichst generische Werkzeuge zu erzeugen. (Wie in Abschnitt 1 angedeutet, ist dieses Arbeitsprinzip fast die Voraussetzung für die Möglichkeit einer systematischen Methodenentwicklung: Indem sich eine Teilcommunity auf klar kontrollierbare Kontextfaktoren einigt, kann sie sich zielgerichtet auf möglichst effektive Werkzeuge für die Problemstellung konzentrieren.) Für die geisteswissenschaftliche Auseinandersetzung mit einem Text oder einem gezielt ausgewählten Untersuchungkorpus sind jedoch nicht selten gerade die idiosynkratischen Kontextfaktoren von Belang – also möglicherweise gerade diejenigen Faktoren, von denen ein generisch angelegtes Werkzeug abstrahiert. Ein

7 Unter einem hermeneutischen Vorgehen verstehen wir in CRETA die geisteswissenschaftlichen Praktiken des Verstehens von Texten, geleitet durch implizite oder explizite Regeln für ein Erschließen von Bedeutungsebenen jenseits des oberflächlichen Informationsgehalts der Texte. Wir verwenden den Begriff ‚Hermeneutik' also weder in der seit der Mitte des 18. Jahrhunderts etablierten Terminologie zur Trennung zwischen Auslegungspraxis (= Exegese) und der dahinterstehenden Theorie (= Hermeneutik), noch beziehen wir uns auf eine spezifische philologische oder philosophische Hermeneutik-Schule.

Vorgehen, das die Entwicklung von algorithmischen Analysekomponenten mit der (Weiter-)Entwicklung der geisteswissenschaftlichen Fragestellung zum Gegenstand verschränkt, kann auf das bestehende Spannungsfeld im Umgang mit Kontextfaktoren reagieren.[8]

Viertens schließlich zeigt die Erfahrung aus der sprachtechnologischen Werkzeugentwicklung, dass bereits moderate Abweichungen der Texte im Anwendungsszenario vom Entwicklungsstandard zu einer erheblichen Verschlechterung der Werkzeugqualität führen;[9] Korpora, die im Fokus geisteswissenschaftlicher Forschung stehen, weichen meist in noch wesentlich stärkerem Maße ab. Somit können selbst vermeintlich ‚unproblematische' Vorverarbeitungsschritte zu einer Verzerrung der Untersuchungsbasis führen. Ein simples Beispiel wäre die Lemmatisierung der Wortformen in Textsammlungen, die eine systematische Suche oder Erzeugung von Subkorpora ermöglicht, indem morphologische Varianten von Schlüsselbegriffen miterfasst werden. Auf Sammlungen, die nicht einer konsistenten orthographischen Konvention folgen – etwa historische Korpora –,

8 Bei der disziplinübergreifenden Arbeit in CRETA versuchen wir, dem Zielkonflikt bei der Konstruktion von Kontextabhängigkeiten durch ein Vorgehen zu begegnen, bei dem einzelne fragerelevante Kontextfaktoren im Rahmen der Spezifikation einer Analyseaufgabe explizit operationalisiert werden. So werden diese Faktoren über die technologischen Standardverfahren hinaus technologisch fassbar (und die Mechanismen der ingenieursmäßigen Werkzeugoptimierung können auf eine entsprechend differenzierte Betrachtung des Untersuchungsgegenstands zugeschnitten werden); anstelle einer generischen Werkzeuglösung tritt ein fragespezifisch angepasstes Analyseinventar.
Die Anpassung der etablierten ingenieursmäßigen Arbeitspraxis an den geisteswissenschaftlichen Differenzierungsbedarf wird besonders plastisch bei der ‚Übersetzung' der Idee von *shared tasks* in ein DH-Umfeld , die ab Seite 391 im Beitrag von Willand et al. (2020) diskutiert wird.
Eine technologische Erfassung der *Interaktion* zwischen mehreren relevanten Kontextfaktoren berührt über das hier Gesagte hinaus die Frage der Modularisierung, die zunächst im Widerspruch zu einem klassisch geisteswissenschaftlichen Zugang zu stehen scheint, der sich der Komplexität der kontextuellen Abhängigkeiten bei der Interpretation des Gegenstands holistisch nähert. Vgl. hierzu Fußnote 11.

9 Sekine 1997 und Escudero et al. 2000 sind Beispiele für systematische Untersuchungen zum Grad der Abhängigkeit sprachtechnologischer Werkzeuge (syntaktisches Parsing und Wortbedeutungsdisambiguierung) von domänenspezifischen Eigenschaften der Trainingsdaten.
In den letzten Jahren entwickelte Verfahren, die mit künstlichen neuronalen Netzen arbeiten, ermöglichen ein Training von kontextualisierten Wortrepräsentationen auf sehr großen unannotierten Korpora (Peters et al. 2018), die dann als ‚Pre-Training' eine anschließende Domänenanpassung von spezialisierten Analysewerkzeugen erheblich erleichtert (für Parsing demonstrieren dies beispielsweise Joshi et al. 2018). Für viele DH-Anwendungsfelder sind jedoch auch unannotierte Text-Ressourcen in der Zieldomäne nicht in größerem Umfang verfügbar, so dass auch zukünftig erhebliche Abstriche bei der erzielbaren automatischen Analysequalität der Normalfall bleiben dürften.

kann der Einsatz eines Standard-Lemmatisierungs-Werkzeugs leicht zu Verfälschungen führen, die sich zudem der Betrachtung auf üblichen Kanälen entziehen, also möglicherweise nie auffallen.

So attraktiv es also aus arbeitspraktischer Sicht erscheint, die Entwicklung von Werkzeuginventaren für gegenstandsunabhängige ‚digitale' Prozesse von der Bearbeitung einzelner geisteswissenschaftlicher Forschungsfragen zu entkoppeln – in der Praxis ist dies nicht ohne Abstriche möglich. Nun gibt es unterschiedliche Strategien, mit dieser Tatsache umzugehen. Etwa kann versucht werden, die untersuchten geisteswissenschaftlichen Forschungsfragen so zu gestalten, dass eine Skalierung auf größere Untersuchungskorpora ausschließlich auf nachweislich gegenstandsunabhängige Analyseschritte aufbaut (z. B. Clustering-Ansätze, die keine Vorannahmen zu Eigenschaften des Untersuchungsgegenstands machen). Die anhaltende Entkoppelung würde also zum Preis einer Beschneidung der effektiv formulierbaren Hypothesen erkauft.

Stärker im Sinne des Verständnisses der DH als eines methodischen Experimentierfelds[10] ist ein Praxismodell, nach dem ein gegenstandssensitives Computerwerkzeug (das möglicherweise einem kanonischen Sprachtechnologie-Inventar entstammt) versuchsweise bei der Untersuchung des Zielkorpus zur Anwendung gebracht wird. Erweist sich der Einsatz als ergiebig für die Verfeinerung der Forschungsfrage, kann das Werkzeug mit weiteren Komponenten kombiniert werden. Treten dagegen problematische Facetten der Werkzeuganwendung zu Tage, werden aus dem inhaltlichen Vorverständnis zum Untersuchungsgegenstand heraus Anpassungen beim Werkzeugeinsatz vorgenommen, die in der Macht der Anwenderinnen und Anwender stehen. (Beispielsweise kann ein zusätzlicher Schritt vorgeschaltet werden: etwa das textuelle Aufspalten von Komposita, die im Deutschen ansonsten dazu führen können, dass Vorkommen von Schlüsselbegriffen für spätere Schritte unauffindbar bleiben; die systematische Ersetzung von Eigennamen in einem Text durch eine künstliche Zeichenfolge etc.) Hier wird die Entkoppelung der disziplinären Felder aufgeweicht: Über die Entwicklung eines immer weiter verfeinerten technischen Instrumentariums für die Bearbeitung einer Forschungsfrage entwickeln die DH-Forschenden ein interdisziplinäres Spezialwissen. In der DH-Konzeption von Willard McCarty ist der Gedanke programmatisch, zyklisch Verbesserungen an – für sich genommen jeweils unzulänglichen – Computermodellen vorzunehmen, um ein verfeinertes Verständnis

10 Ramsay und Rockwell 2012 dokumentieren beispielsweise eine 2011 von Stephen Ramsay ausgelöste Kontroverse, in der über das Verhältnis zwischen ‚building' (als Erstellen von Computerprogrammen) und ‚studying' in den DH debattiert wurde: das Entwickeln von Computerprogrammen wird als möglicherweise konstitutives Element des Erkenntnisprozesses in den DH diskutiert.

des modellierten Gegenstands zu erhalten (McCarty 2005, S. 20–72). Das Vorgehen schließt damit an die Tradition des *Computing in the Humanities* an; es hat seine Wurzeln also deutlich vor der disziplinübergreifenden Anwendung von *textmining*-Verfahren im Zuge des ‚Digitalisierungs'-Diskurses.

In der eben skizzierten pragmatischen Ausprägung kann der Ansatz dann problematisch sein, wenn die Situation entsteht, dass das werkzeuggestützte Analyseergebnis von inhaltlich gerechtfertigten Erwartungen abweicht, ohne dass dies deutlich vor Augen tritt. Ein Beispiel wäre das oben angeführte Lemmatisierungs-Werkzeug, das auf einem historischen Korpus auf vermeintliche Trends zu einer semantischen Verschiebung hindeutet, in Wahrheit jedoch auf Verschiebungen in der orthographischen Konvention reagiert. In bestimmten Szenarien mag das Risiko derartiger Effekte gering sein; mit zunehmendem Abstraktionsniveau der Zielanalysen wächst jedoch die Gefahr von Scheineffekten.

Die Problematik lässt sich methodisch auffangen, indem (a) die Gesamtfragestellung zum Gegenstand in Teilkomponenten mit einer klar spezifizierbaren Analyseaufgabe *modularisiert* wird[11] und (b) für Teilaufgaben, die mit Computerwerkzeugen angegangen werden, eine Validierung der automatischen Analyseer-

11 Wichtig ist hervorzuheben, dass eine Modularisierung entlang geeigneter Schnittstellenrepräsentationen zunächst rein arbeitspraktisch motiviert ist, also weder mit normativen Vorgaben noch eventuellen Hypothesen zur Unabhängigkeit von kognitiven Modulen einhergeht, etwa im Zusammenhang der Debatte über die Modularität des Geistes, Fodor 1983. Ein Beispiel für eine sinnvolle Abstraktionsebene für die praktische Arbeit mit Texten aus unterschiedlichen Sprachstadien und -registern wäre die Lemmatisierung, also die Abbildung jeder Form auf eine orthographisch normalisierte Zitationsform: ‚es hat ihn viel Gelt gekoſt/ und hat daruber noch eine groſſe Wett verlohren' würde so repräsentiert als ‚es haben er viel Geld kosten / und haben darüber noch eine groß Wette verlieren' (Becher 1682, S. 148). Ohne den Abstraktionsschritt wären viele korpusüberspannende Betrachtungen äußerst mühselig – etwa eine thematische Suche nach Textpassagen, in denen über Geld gesprochen wird. Im allgemeinen Fall kann die Lemmatisierungsaufgabe allerdings nicht ohne Textverständnis gelöst werden. (So wird im genannten Beispiel ‚gekoſt' im DTA fälschlich auf ‚kosen' abgebildt, was für die gegenwartssprachliche Form ‚gekost' in der Tat die einzige in Frage kommende Lesart ist.) Jede Modularisierung kann Fehlerquellen einführen, die gegen die erreichbaren Generalisierungen abgewogen werden und gegebenenfalls anhand von Referenzdaten bewertet werden sollten.
Über diese Überlegungen hinausgehend ist es mit neueren datenorientierten Verfahren grundsätzlich möglich, auch solche Wechselbeziehungen robust zu erfassen, welche die gewählten Abstraktionsebenen ‚überspringen'. Für moderne Komponenten in der sprachtechnologischen Analysekette wird etwa mit *joint-inference*-Verfahren gearbeitet, um eine Propagierung von Fehlern auf oberflächennahen Ebenen zu minimieren. Effekte wie die angeführte semantisch begründete Disambiguierung der Form ‚gekost' können modelliert werden, wenn geeignete Daten zur Verfügung stehen. (Systematische Lernexperimente zu einem theoretischen Umgang mit Wechselbeziehungen zwischen Morphologie und Syntax werden z. B. in Seeker und Kuhn 2013 diskutiert.)

gebnisse gegen die Erwartungen der Anwenderinnen vorgenommen wird. Werkzeugen, deren Vorhersagen robust den Erwartungen folgen, kann bei der anschließenden Interpretation der Analyseergebnisse größeres Gewicht gegeben werden; kommt es dagegen bei einem Werkzeug zu großen Schwankungen, kann der Versuch unternommen werden, seine Analyse aus einem anderen Blickwinkel heraus zu ergänzen, um Befunde besser absichern zu können.

Um eine Validierung zu ermöglichen, müssen unabhängig von der Werkzeugentwicklung Testdaten herangezogen werden, die aus theoretischen Überlegungen heraus (oder möglicherweise auch aus einer Intuition im Zuge der hermeneutischen Fortentwicklung eines Vorverständnisses zum Gegenstand) begründet sind. Im Fall einer systematischen Textanalyse wird es sich um eine manuelle Annotation der kontextangemessenen Zielkategorisierung handeln, die an einem geeigneten Testkorpus vorgenommen wird. Derartige Referenzdaten dienen dann im Verlauf der Weiterentwicklung und Kombination von Werkzeugen als Messschnur für eine gezielte Optimierung.[12] Betont sei, dass zwar die Validierung der technischen

---

Für die Behandlung geisteswissenschaftlicher Fragestellungen heißt dies, dass eine Modularisierung auf Basis bestimmter Schnittstellenrepräsentationen (meist aufbauend auf theoretischen Konstrukten) nicht grundsätzlich im Widerspruch zu Situationen steht, in denen aus theoretischen Überlegungen heraus komplexe Interaktionen von Wissensquellen von Belang sind. Eine holistische Betrachtung des Einzelgegenstands wird zwar einem konventionellen ‚mikroanalytischen' Zugang vorbehalten bleiben; systematische ebenenübergreifende Effekte, die zum holistischen Eindruck beitragen, lassen sich jedoch durchaus mit einem modularisierten Inventar von Analysekomponenten beleuchten.

12 Die zentrale Rolle von Referenzdaten für das Vorgehen der reflektierten algorithmischen Textanalyse, das wir in CRETA verfolgen, spiegelt sich in den Beiträgen in diesem Band wider: Pichler und Reiter 2020 diskutieren ab Seite 43 den CRETA-Workflow systematisch, in dessen Zentrum die Annotation eines Referenzkorpus steht. Auch der programmatische Beitrag zur Kombination von Arbeitspraktiken in den DH, Kuhn 2020 ab Seite 63 weist annotierten Referenzdaten eine Schlüsselfunktion zu (Kuhn 2020, S. 72 f.). Pagel et al. 2020 demonstrieren ab Seite 125, dass sich ein einheitlicher Annotations-Workflow mit unterschiedlichen Zielsetzungen nutzen lässt. Reiter 2020 bietet eine konkrete Anleitung für die Erstellung von Annotationsrichtlinien. Ketschik, Blessing et al. 2020 diskutiert ab Seite 204 disziplinübergreifend den gesamten Prozess von der Erstellung von Richtlinien über die manuelle und semiautomatische Annotation bis hin zur Einbindung der Analyseergebnisse in die Forschung zu geistes- oder sozialwissenschaftlichen Fragestellungen; das Annotationsziel ist über die unterschiedlichen Fachdisziplinen hinweg eine Erfassung der für die jeweilige Textgattung und Forschungsfrage relevanten Entitätenreferenzen in Texten (‚Berlin', ‚Angela Merkel', ‚die Kanzlerin', ‚der helt', ‚das Kunstwerk' usw.). Für Klinger et al. 2020, in diesem Band ab Seite 238, stellt die Annotation von Phrasen, die Emotionen hervorrufen, in einem Korpus von fiktionalen Texten einen zentralen Schritt dar. Die Beiträge in Teil IV (Barth 2020; Ketschik, Murr et al. 2020; Willand et al. 2020) dokumentieren gemeinsam ein methodologisches Experiment zur arbeitsgruppenübergreifenden Entwicklung von Annotationsrichtlinien für textanalytisch anspruchsvolle Konzepte – hier das narratologische Konzept

Realisierung eines Analyseversuchs gegen die konzeptionelle Erwartung *per se* nichts an der effektiven Analysequalität ändert. Eine ausdifferenzierte Wahrnehmung der erzielbaren Genauigkeit kann jedoch eine wesentlich aussagekräftigere Interpretation der Analyseergebnisse eines oder mehrerer Werkzeuge auf den Zieldaten ermöglichen – selbst wenn die Fehlerquote der Werkzeuge absolut betrachtet relativ hoch ist, wie dies bei der Computermodellierung vieler anspruchsvoller Aufgaben der Fall ist.

Die geschilderte testdatenbasierte Werkzeugvalidierung ist nichts anderes als eine Anwendung der streng etablierten Standardmethodologie aus der maschinellen Sprachverarbeitung bzw. der werkzeuggestützten Korpuslinguistik. In jeder DH-Kooperation mit einer Beteiligung aus diesen Arbeitsfeldern oder anderen Informatikbereichen der datenorientierten Modellentwicklung werden ihre Vertreterinnen großes Augenmerk auf entsprechende *workflows* legen.

Bedeutet die Anerkennung des erwartungsbezogenen Validierungsbedarfs also, dass DH-Projekte, deren Analysekategorien sich nicht unmittelbar aus oberflächennahen Texteigenschaften ergeben, generell wie ein sprachtechnologisches Entwicklungsprojekt zu behandeln sind? (Den Beteiligten aus den Geisteswissenschaften käme damit die Rolle der Domänenexperten und Domänenexpertinnen zu, die mit ihrem gegenstandsbezogenen Fachwissen zur Spezifikation der gewünschten Zielmodellierung beitragen. Spezialistinnen für Computermodellierung – möglicherweise ohne Hintergrund zum Gegenstand – wären dann für eine möglichst effektive technische Umsetzung zuständig.) Ein solches Vorgehensmodell ist grundsätzlich denkbar und kommt in Szenarien breit zum Einsatz, in denen ein etablierter Konsens zu Zielkategorien der Textanalyse besteht, die studienübergreifend nutzbar sind. Beim Domänenexperten-/Entwickler-Modell ist das arbeitspraktische Vorgehen klar entlang der jeweiligen disziplinären Expertise entkoppelt – allerdings geht dies mangels einer verzahnten dynamischen Entwicklung auf Kosten des ‚hermeneutischen' Charakters der Modellierung, wie McCarty (2005) sie charakterisiert und wie sie für die Erschließung von offenen geisteswissenschaftlichen Fragekomplexen typisch ist. Will man dem fachwissenschaftlichen Erkenntnisinteresse gerecht werden, erscheint die Charakterisie-

---

der Erzählebene. Im Rahmen eines *shared task* wurden von unterschiedlichen Teams Richtlinien entwickelt und anschließend vergleichend evaluiert. Der Beitrag von Willand et al. führt die Konzeption des *shared task* „Systematic Analysis of Narrative Texts through Annotation" (SANTA) aus, die Beiträge Barth 2020 und Ketschik, Murr et al. 2020 stellen jeweils einen Wettbewerbsbeitrag dar.

rung des Entwicklungsprozesses entlang eines Domänenexperten-/Entwickler-Modells mit strikter Kompetenzaufteilung also eher unglücklich.[13]

Im Ergebnis bietet es sich daher an – das ist jedenfalls die Schlussfolgerung der Kooperationspartnerinnen und -partner in CRETA –, bei textanalytisch anspruchsvollen Fragekomplexen aus dem Spektrum der Geistes- und Sozialwissenschaften differenziert vorzugehen: Dort, wo zu Projektbeginn die Spezifikation und Operationalisierung von analyserelevanten Konzepten (und eine daraus ableitbare Modularisierung) zu einem Fragenkomplex noch sehr offen ist, erscheint es sinnvoll, auf eine Entkoppelung der methodischen Sphären zu verzichten. Auf einen geeigneten algorithmisch gestützten Analyseansatz lässt sich in einem solchen Szenario effektiv zuarbeiten, wenn die Anpassung von Zielkategorien und Analyseerwartungen eng verzahnt mit einer versuchsweisen technischen Umsetzung, möglicherweise sogar parallel mit konkurrierenden technischen Modellierungsansätzen, erfolgt. Ein solches Vorgehen erscheint insbesondere dann angezeigt, wenn mit den anstehenden Textanalyseaufgaben auch aus informatischer Sicht offene Forschungsfragen verbunden sind – wenn die Werkzeugentwicklung also nicht auf gut etablierte Lösungspfade zurückgreifen kann. Und in der Tat ist eine Vielzahl der Analyseschritte, die in DH-Szenarien relevant sind, mit echten technischen Herausforderungen verbunden. So sind die verfügbaren Datensätze oft klein und das Textmaterial weicht in vielen Dimensionen von kanonischen Werkzeuganwendungen ab.

Auf der anderen Seiten gibt es Untersuchungsbereiche, in denen sich aus der etablierten geisteswissenschaftlichen Theoriebildung eine Modularisierung der Analysefragestellung ableitet – etwa in Kernbereichen der Erzähltheorie oder Narratologie. In einem solchen Szenario ist es durchaus möglich, bei der Entwicklung des Analyseinventars die methodischen Sphären anhand von präzise spezifizierten Schnittstellen zu trennen. Ein Beispiel für ein solches Vorgehen wird in den Beiträgen Barth (2020), Ketschik, Murr et al. (2020) und Willand et al. (2020) in Teil IV des Bandes ausgeführt. Für das narratologische Konzept der Erzählebene soll eine Referenzdatenannotation die Schnittstelle zwischen literaturwissenschaftlich ausgerichteten Arbeiten (theoriegeleitete Operationalisierung und spätere Anwendung der Analyseergebnisse) und einem informatisch-

13 Im Prinzip ist auch beim Domänenexperten-/Entwickler-Modell ein zyklisches Vorgehen mit regelmäßiger Abstimmung zwischen Domänenexperten/Domänenexpertinnen und Werkzeugentwicklerinnen möglich und auf diese Weise wäre auch eine fortlaufende Weiterentwicklung der Zielspezifikation umsetzbar, einschließlich einer eng damit verbundenen Anpassung der Ergebniserwartung. Zwischen einer strikten arbeitspraktischen Trennung und einer engen Verzahnung besteht ganz offensichtlich ein Kontinuum.

computerlinguistischen Vorgehen (Optimierung von Analysemodellen anhand der Referenzdaten) bilden.

Angestrebt wird also die folgende **Praxis der *reflektierten algorithmischen Textanalyse*:**[14] Die reflektierte Textanalyse verzahnt Arbeitspraktiken, die in den jeweiligen ‚Mutterdisziplinen' methodisch verankert sind – wobei der *Grad* der Verzahnung dem jeweiligen Untersuchungskontext angepasst wird. Aus den textwissenschaftlichen Disziplinen gehen das dem Gegenstand angemessene theoretische Vorwissen sowie die Maßstäbe zur Beurteilung von Befunden in Bezug auf den zu bearbeitenden Fragekomplex ein; aus der datenorientierten Informatik/Computerlinguistik und der Korpuslinguistik wird (a) die Modularisierung der Analysearchitektur und (b) die Methodik einer empirisch fundierten, erwartungsbezogenen Validierung von Modellierungsschritten übernommen. Ausdrücklich eingeschlossen in eine Gesamt-‚Architektur' der Textanalyse können nicht-technische ‚Module' sein, also manuelle Arbeitsschritte, bei denen Wissenschaftlerinnen und Wissenschaftler prinzipiengeleitet interpretatorische Entscheidungen treffen. Und neben ‚Primär'-Werkzeugen, die eine Kategorisierungs- oder Segmentierungs- bzw. Strukturierungsaufgabe lösen, können ‚Sekundär'-Werkzeuge zum Einsatz kommen, die etwa mit *visual-analytics*-Verfahren[15] über Ergebnisse eine Primäranalyse aggregieren.

Kondensiert lässt sich unser Verständnis von reflektierter algorithmischer Textanalyse in folgender Definition aus Pichler und Reiter 2020, S. 58 fassen:

> Reflektierte algorithmische Textanalyse bezeichnet Praktiken der computergestützten Textanalyse, die sich durch ihre interdisziplinär verzahnte Modularisierung kennzeichnen. Bei diesen Modulen handelt es sich um miteinander verknüpfte, manuelle und automatische Arbeitsschritte, die sich auf Begriffe oder Textphänomene beziehen. Die Aufteilung der Module sowie die Interpretation von deren Ergebnissen erfolgt unter Berücksichtigung des gegenstandsbezogenen Vorwissens, der Operationalisierbarkeit der Module sowie deren empirischer Validierung.

14 Der Ansatz ist keine *Erfindung* des interdisziplinären CRETA-Teams, sondern hat sich in vergleichbarer Weise in anderen gleichgewichtig austarierten Kooperationen entwickelt. Einige Gedanken zur wissenschaftstheoretischen Einordnung diskutieren wir in Kuhn et al. 2020 (hier bauen wir insbesondere auf den konzeptionellen Rahmen von Danneberg und Albrecht (2017) auf). Pichler und Reiter 2020 vertiefen das arbeitspraktische Vorgehen anhand eines konkreten Workflows; die Frage der Verzahnung wird u. a. von Kuhn (2020) ausführlich diskutiert. Gerstorfer 2020 arbeitet ein Reflexionsmodell heraus, mit dem er den Status der Schritte im CRETA-Workflow wissenschaftstheoretisch beleuchtet.

15 Baumann et al. 2020 bieten ab Seite 270 einen Einblick in Visualisierungsverfahren, die im Rahmen der CRETA-Kooperationen entwickelt wurden.

2.2 Reflektierte algorithmische Textanalyse und die Zielsetzung in CRETA

Wie bereits deutlich wurde, waren zwei Aspekte des Herangehens an die interdisziplinäre Herausforderung der DH prägend für die methodisch-arbeitspraktische Konzeption von CRETA: Zum einen sollte die Entwicklung und Evaluierung von komputationell-technischen Komponenten nicht hinter die etablierte Arbeitspraxis einer empirisch fundierten Modellierung von Analyseaufgaben zurückfallen – gerade die Computerlinguistik blickt hier auf eine jahrzehntelange Entwicklung zurück, in der theoretisch-konzeptionelle Überlegungen aus der generativen Linguistik ebenso Niederschlag gefunden haben wie systematische Fragen der Korpusstatistik und methodische Prinzipien, welche ein adäquater Einsatz von maschinellen Lernverfahren erforderlich macht. Zum anderen sollten die eingesetzten Computermodelle und -werkzeuge in Analyseschritte eingebunden werden, die Befunde zu geistes- oder sozialwissenschaftlichen Kernfragen liefern – sich also nicht auf den Prozess der Vorverarbeitung des Datenmaterials, auf Suchverfahren zur schnelleren Sichtung des Materials oder ähnliches beschränken. Zur Einbindung in die Bearbeitung der Kernfragen gehört also insbesondere die Möglichkeit, die Modelle zyklisch zu verbessern und im Rahmen eines hermeneutischen Prozesses auf ein verfeinertes Verständnis der relevanten Zusammenhänge anzupassen (s. a. Pichler und Reiter 2020, Abschnitt 6).

Die Umsetzung beider Aspekte ist nur möglich, wenn die methodisch-arbeitspraktischen Standards aus der Informatik wie aus den Geisteswissenschaften bei Design-Entscheidungen zum konkreten Forschungs-*workflow* gleichberechtigt berücksichtigt werden. Häufig ist dies derzeit nur in interdisziplinär besetzten Kooperationsteams möglich[16] und es setzt die beidseitige Bereitschaft voraus, gängige methodische Konventionen der eigenen Fachdisziplin zu hinterfragen, da diese möglicherweise aus Arbeitshypothesen heraus motiviert sind, die im erweiterten interdisziplinären Rahmen nicht ohne Weiteres aufrecht erhalten werden können.

Man mag sich fragen, weshalb solches Gewicht auf eine enge Verschränkung der disziplineigenen Standards gelegt werden soll. Wäre es arbeitsökonomisch

16 Es ist jedoch durchaus möglich, dass Teil-Communities im methodischen Schnittmengenbereich das methodisch-technische Knowhow spezifizieren, das einzelne Vertreterinnen und Vertreter mitbringen müssen, um Projektideen im interdisziplinären Spektrum eigenständig angehen zu können – häufig sicherlich unter Ausnutzung von Modellierungsarchitekturen, die von vielen Beteiligten sukzessive vorangebracht werden. Der Fokus des eigenen innovativen Forschungsbeitrags wird sich dann je nach Spezialisierung unterscheiden, wie das in den meisten interdisziplinären Forschungsfeldern üblich ist.

nicht günstiger, in den DH vorwiegend solche komputationelle Verfahren zur Anwendung zu bringen, die bereits im breiten Stil erprobt sind und deren technische Anwendbarkeit unabhängig von spezifischen Eigenschaften des Forschungsgegenstands (etwa des untersuchten Textkorpus) ist? Es sollte dann genügen, die entsprechenden Schritte sorgfältig zu dokumentieren und sie könnten auch in Forschungsprojekten ohne hochgradig interdisziplinär besetzte Teams zur Anwendung kommen. Darüber hinaus könnten zentrale Service-Einrichtungen Beratung und technische Unterstützung bei der Einrichtung von DH-*workflows* bieten, die sich an vorwiegend geisteswissenschaftlich besetzte Teams wendet.

Verfolgt man das Ziel einer Praxis der reflektierten algorithmischen Textanalyse, wie die Akteure in CRETA dies tun, fällt die Antwort auf diese Überlegung wie folgt aus: Es mag im Prinzip Anwendungskontexte von Computerwerkzeugen geben, für die es nicht nötig wäre, sich zu vergewissern, dass deren Anwendung auf den Untersuchungsgegenstand adäquat ist. (Gerade beim Einsatz von maschinellen Lernverfahren liegt das erklärte Entwicklungsziel schließlich in der Regel in einem verallgemeinerten Systemverhalten, das von Eigenheiten einzelner Datensätze abstrahiert.) Unterscheiden sich jedoch die Anwendungsdaten in relevanten Eigenschaften von den Daten, für die das Werkzeug entwickelt wurde, ist eine verlässliche Abschätzung der zu erwartenden Qualität kaum möglich. Und in typischen DH-Anwendungsszenarien weichen die untersuchten Daten, wie bereits angedeutet, häufig in mehreren Dimensionen von den kanonischen Entwicklungsdaten ab: So werden etwa historische Texte einer bestimmten Gattung und regionalen Herkunft betrachtet; die Verlässlichkeit eingesetzter Werkzeuge wurde jedoch bislang auf modernen Zeitungstexten etabliert. Hinzu kommt, dass die Zielfragestellung einer geisteswissenschaftlichen Untersuchung zumeist eine Kombination von Analyseschritten voraussetzt, zu deren Wechselwirkungen im spezifischen Untersuchungskontext es keine Vorarbeiten gibt.

CRETA ging mit dem Ziel an den Start, in fächerübergreifender Zusammenarbeit zwischen unterschiedlichen geistes- und sozialwissenschaftlichen Disziplinen und der Informatik Arbeitspraktiken und Workflows zu erarbeiten, die es bei der Analyse von Texten und Textkorpora erlauben, formal-komputationelle Analysemodelle methodisch fundiert auf die dem jeweiligen Forschungskontext entsprechende Fragestellung anzupassen und effektiv zur Anwendung zu bringen. Konkret wird ein modulares Inventar von einerseits technischen und andererseits methodisch-arbeitspraktischen Komponenten entwickelt, die jeweils für einen Arbeitsschritt bei der Text- bzw. Korpusanalyse stehen. Zusammengenommen decken die Komponenten den gesamten Weg von der anfänglichen Auseinandersetzung mit einem geistes- oder sozialwissenschaftlichen Fragekomplex in Bezug auf eine Textsammlung bis hin zu einer fundiert interpretierbaren qualitativen oder quantitativen Analyse und deren kritischer Reflexion ab.

Wie bereits eingangs dieser Einleitung hervorgehoben, verfolgt CRETA den Grundsatz der interdisziplinären Zusammenarbeit in zwei Dimensionen zugleich: ‚Horizontal' werden Methoden und Arbeitspraktiken (i) aus einem Cluster von informatiknahen Fächern und (ii) aus textwissenschaftlichen Disziplinen in den Geistes- und Sozialwissenschaften kombiniert. Zudem führt CRETA innerhalb der beiden methodisch verwandten Fächer-Cluster jeweils ‚vertikal' unterschiedliche Blickwinkel zusammen. Mit diesem multidisziplinären Vorgehen werden methodisch-konzeptionelle Komponenten, Workflows und technische Werkzeugmodule entwickelt, mit denen

- umfangreiche Textkorpora durch adaptierbare Filter im Sinne der studienspezifischen Aufgabenstellung befragt und die Ergebnisse visuell kompakt dargestellt werden können,
- die Identifizierung von systematisch gleichartigen Teilaufgaben über Studien hinweg unterstützt wird,
- für gängige Typen von Teilaufgaben ein Instrumentarium zur Verfügung steht, das die Evaluation der Analysequalität, das Auffinden möglicher Fehler sowie die Aggregation und Meta-Analyse von Ergebnissen erlaubt,
- der interaktive Prozess einer Korrektur und Anpassung von Komponenten sowie die Kombination von Ergebnissen mit Hilfe visuell orientierter Interfaces unterstützt wird.

Auf technischer Ebene konnte CRETA breit auf infrastrukturelle Vorarbeiten und laufende Initiativen in der Community[17] sowie Standardverfahren und - bibliotheken aufbauen.

Die Einrichtung von CRETA hat strukturell zu einer nachhaltigen Verankerung der *Digital Humanities* an der Universität Stuttgart geführt – nicht zuletzt

17 Die beiden paneuropäischen Infrastrukturinitiativen CLARIN (*Common Language Resources and Technology Infrastructure*, siehe www.clarin.eu und www.clarin-d.net) und DARIAH (*Digital Research Infrastructure for the Arts and Humanities*, siehe www.dariah.eu und de.dariah.eu), organisiert im *European Research Infrastructure Consortium* (ERIC), gehörten zu den ersten Initiativen, die Lösungen für die Forschungsdateninfrastruktur gezielt vom Blickwinkel der Forschenden her entwickelten. Das Institut für Maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart ist unter Leitung von Jonas Kuhn seit 2011 Mitglied des CLARIN-D-Zentrenverbundes. Seit 2019 besteht unter der Federführung des Deutschen Literaturarchivs Marbach das Datenzentrum SDC4Lit (www.sdc4lit.org), das sich einem nachhaltigen Datenlebenszyklus für digitale Literatur widmet. Neben Marbach sind die Abteilung *Digital Humanities* des Instituts für Literaturwissenschaft der Universität Stuttgart, das IMS sowie das Höchstleistungsrechenzentrum Stuttgart beteiligt.

durch die neu geschaffene W3-Professur *Digital Humanities* und die zugehörige Abteilung im Institut für Literaturwissenschaft sowie den gleichnamigen MA-Studiengang. Die vorhandenen Kooperationsbeziehungen wurden deutlich gestärkt. Im Sinne des Strategieziels des „Stuttgarter Wegs: Vernetzte Disziplinen" der Universität Stuttgart bilden die *Digital Humanities* einen der universitätsweiten interdisziplinären Forschungsschwerpunkte.

3 Strukturelle Herausforderungen für die DH-Forschung

3.1 Die Werkstatt-Idee

Dass Methoden und Arbeitspraktiken für die algorithmische Textanalyse am besten in einem disziplinübergreifenden Kooperationsteam zu entwickeln sind, ist ohne Weiteres einsichtig. Wie kann der Prozess jedoch in der alltäglichen Praxis des akademischen Betriebs umgesetzt werden? Wie in Abschnitt 2.1 nachgezeichnet, will der Zugang zur Textanalyse, den CRETA sucht, methodische Prinzipien aus unterschiedlichen Disziplinen gleichberechtigt nebeneinander stellen. Die Ausdifferenzierung einer geisteswissenschaftlichen Forschungsfrage soll nicht vordringlich durch die technischen Analysemöglichkeiten bestimmt werden, sondern inhaltlichen und theoretischen Überlegungen folgen. Umgekehrt soll jeder komputationell umgesetzte Analyseschritt gegen eine überprüfbare Operationalisierung der zugrundeliegenden Konzepte validiert werden; anhand von Referenzdaten soll eine kontrollierte Optimierung von Werkzeugen ermöglicht werden.

Die Beteiligten an einer solchen interdisziplinären Unternehmung sind sich bewusst, dass ein wechselseitiges Verständnis zu zentralen Konzepten und etablierten Arbeitspraktiken entwickelt werden muss, dass geklärt werden muss, aus welchen übergeordneten wissenschaftlichen Problemen heraus die Disziplinen ihre Agenden entwickeln. Jedes klug geplante interdisziplinäre Projekt wird also eine Phase der Klärung von Begriffen und der Verankerung des Vorgehens in den Einzeldisziplinen vorsehen.

Das vielleicht größte Risiko kann allerdings gerade in Diskrepanzen zu vermeintlich geteilten Begriffen und Arbeitsschritten liegen. Wenn sich etwa zwei kooperierende Disziplinen darin einig sind, dass eine Segmentierung von Texten nach geschilderten Ereignissen vorzunehmen ist, kann möglicherweise übersehen werden, dass tatsächlich sehr unterschiedliche Ereigniskonzepte zugrundeliegen, die nicht mit den gleichen Analysekomponenten bearbeitet werden kön-

nen. Im ungünstigen Fall könnten also in einer systematisch-konzeptionellen Planungsphase mit den besten Absichten Entscheidungen getroffen werden, die eine erfolgreiche disziplinübergreifende Weiterentwicklung schwer machen.

Nun können Irrtümer bei der Planung von Projektverläufen nie ausgeschlossen werden, Revisionen müssen möglich sein. Da CRETA wie oben dargestellt jedoch in mehreren Dimensionen gleichzeitig interdisziplinär arbeitet, erschien es bei der Konzeption des Zentrums besonders wichtig, Mechanismen zu einer weitgehenden Absicherung gegen vermeidbare Missverständnisse vorzusehen. Zur Umsetzung des Ziels wird ein sehr simples Verfahren eingesetzt, das sich als sehr effektiv erwiesen hat: vor Beginn einer systematischen Planung des Vorgehens zu einem Untersuchungsgegenstand setzen sich disziplinär gemischte Kleingruppen konkret mit einer kleinen Stichprobe aus dem Textmaterial auseinander. Sie betreiben ein *brainstorming* zu möglichen Analysezielen und Hilfskategorien, wenden diese probehalber auf Textpassagen an und diskutieren aus den verschiedenen Hintergründen heraus, wo eventuelle Fallstricke und wo Chancen liegen könnten.

In den halbjährlich abgehaltenen zwei- bis dreitägigen CRETA-Werkstätten stellen diese stark textgebundenen *brainstorming*-Sitzungen zu Analysemöglichkeiten ein Kernelement dar. Ähnlich wie bei der Diskussionsform des *world café* können durch Untergliederung des Gesamtteams in Kleingruppen die verschiedenen Hintergründe und Erfahrungshorizonte berücksichtigt werden; Nachwuchswissenschaftler und Projektleiterinnen sind gleichermaßen einbezogen. Nach paralleler Bearbeitung desselben Textmaterials in mehreren, selbst selektierten Kleingruppen wird im anschließenden Plenumsgespräch herausgearbeitet, welches konkrete Ziel und welche Zwischenschritte in der anschließenden Projektphase angegangen werden sollen.

Im Vergleich zu klassischen, *top-down*-orientierten Konzeptionsgesprächen hat das CRETA-Werkstatt-Konzept den Vorteil, dass die angesprochenen versteckten Missverständnisse sehr viel wahrscheinlicher schon früh zutage treten, so dass die Planung gleich darauf Rücksicht nehmen kann. Der erforderliche Zeitaufwand für die Beteiligten ist möglicherweise geringfügig größer, das konzentrierte Zusammenführen der Expertise aus unterschiedlichen Fachhintergründen erzeugt jedoch gerade durch die gemeinsame Arbeit an konkreten Textstellen ein hohes Maß an positiver Energie. (Eine ausführlichere Darstellung des Konzepts der CRETA-Werkstatt und weiterer Instrumente, die zur Verbreitung und Weiterentwicklung der interdisziplinären Praktiken dienen, findet sich ab Seite 467 in diesem Band im Beitrag von Reiter, Kremer et al. (2020).)

Regelmäßig abgehalten kommen den CRETA-Werkstätten selbstverständlich weitere Funktionen im Projektverlauf zu. Über laufende Arbeiten wird berichtet, häufig unter Einbeziehung von externen Gästen, die Anregungen und Kommen-

tare liefern. Die eingesetzten Werkzeuge und *workflows* können über den Verlauf von mehreren Projektphasen hinweg zyklisch verfeinert werden. Im Rahmen von *brainstorming*-Sitzungen kann ein disziplinübergreifendes Bild zu möglichen Zusatzanforderungen und eventuellen Unzulänglichkeiten entwickelt werden.

Der vorliegende Band versucht mit seiner Zusammenstellung von Beiträgen, die die interdisziplinäre Herangehensweise betonen, einen plastischen Eindruck zur Arbeit in einer ‚Werkstatt' für algorithmische Textanalyse für geistes- und sozialwi

Aus dem ‚CRETA-Umfeld' sind einige Nebenprojekte entstanden, ssenschaftliche Fragen zu vermitteln. Neben diesen Arbeiten, in denen methodologische Überlegungen eine zentrale Rolle spielen, haben die an CRETA beteiligten Wissenschaftlerinnen und Wissenschaftler zahlreiche spezifischere Studien durchgeführt, die wir im vorliegenden Band nicht berücksichtigen konnten. Die daraus entstandenen Publikationen sind auf der Webseite des Zentrums[18] aufgelistet.

3.2 Erfahrungen aus CRETA und die breitere strukturelle Perspektive

Die gezielte Einrichtung eines Zentrums wie CRETA, dessen Kernaufgabe die disziplinübergreifende Methodenentwicklung ist, erlaubt es nicht nur, naheliegende Verbindungen zwischen den beteiligten Fächern zu realisieren, für die im akademischen Alltag bis dahin lediglich die Zeit fehlte. Sie bietet darüber hinaus auch die Möglichkeit, weniger offensichtliche Pfade der Zusammenarbeit zu explorieren – Pfade, deren Einstieg für einzelne Beteiligte oder auch für das ganze Team jenseits des abgesicherten Terrains disziplineigener Grundannahmen und Methoden liegt. Solche Experimente sind naturgemäß riskanter, sie können aus unterschiedlichen Gründen scheitern – etwa wenn die Erwartungen seitens Disziplin A, die an die Einbindung einer bestimmten Wissenskomponente aus Disziplin B geknüpft waren, enttäuscht werden, weil notwendige Annahmen zum Untersuchungskontext nicht kompatibel sind. Gelingt eines der Experimente jedoch, kann dies eine stärkere Horizonterweiterung bedeuten als sie mit mehreren gut abgesicherten Projektplänen möglich wäre. Aber selbst aus einem partiellen Scheitern können wertvolle Lehren gezogen werden, die in das nächste Experiment einfließen.

Das Format eines interdisziplinären Zentrums mit einem relativ großen Grad an ‚Beinfreiheit' beim Einsatz der Ressourcen erlaubt es insbesondere, ein ver-

18 https://www.creta.uni-stuttgart.de/publikationen/

breitetes Dilemma für projektfinanzierte Forschung in stark interdisziplinären Bereichen zu vermeiden: Stellt ein Team einerseits eine weitreichende Vision für eine innovative Zusammenarbeit ins Zentrum eines Projektantrags und legt offen, dass der Weg dorthin teilweise erst im Projektverlauf erschlossen werden soll, dann fällt der Antrag möglicherweise im Wettbewerb mit konkurrierenden monodisziplinären Anträgen auf, die klar an jüngste Entwicklungen in der jeweiligen Community anknüpfen; mit der Betonung der Innovation wird ein weniger abgesichertes Arbeitsprogramm und eine eingeschränkte Vergleichbarkeit der erwarteten Ergebnisse einhergehen. Bei begrenzten Gesamtmitteln kann es dann für das Entscheidungsgremium aus Überlegungen der Qualitätssicherung heraus schwer zu rechtfertigen sein, nicht den besser abschätzbaren ‚konventionellen' Ideen den Vorzug zu geben. Versucht ein Team bei der Projektkonzeption andererseits, die Anschlussfähigkeit an abgesichertes Methodenwissen und inhaltliche Vorarbeiten in den Vordergrund zu stellen, kann für Betrachter aus den beteiligten Disziplinen, die selbst nicht mit den speziellen Herausforderungen der interdisziplinären Zusammenarbeit vertraut sind, der Eindruck einer wenig inspirierten Zielsetzung entstehen.[19] Die Problematik wird mitunter noch dadurch verschärft, dass bei einer fächerübergreifenden Beurteilung von Projektanträgen unterschiedliche Kulturen der Begutachtung aufeinandertreffen. Im eigenen disziplinären Kontext fällt es bei der Formulierung von Anträgen vergleichsweise leicht, mögliche Kritikpunkte zu identifizieren (etwa weil auf eine bekanntermaßen kontroverse Annahme aufgesetzt wird) und die Argumentation darauf abzustimmen; dagegen ist es erheblich schwerer zu antizipieren, welche Aspekte einer avisierten interdisziplinären Verbindung etablierter Wissensbereiche den größten Klärungsbedarf erzeugen. Bei einem thematisch und methodisch offenen Förderrahmen mag daher die Abwägung von Aufwand für die Projektkonzeption gegenüber Erfolgschancen manche Forscherinnen und Forscher vor riskanteren interdisziplinären Projektideen zurückschrecken lassen.[20]

19 Vgl. auch die Diskussion des *scheduling dilemma* in Kuhn 2020 in diesem Band.

20 Eine ähnliche Problematik stellt sich Nachwuchswissenschaftlerinnen und -wissenschaftlern, die die Entscheidung treffen müssen, ob sie ihren Forschungsschwerpunkt auf ein stark interdisziplinäres Feld legen sollten. Der große zeitliche Aufwand, der in eine Einarbeitung in sehr verschiedene inhaltliche bzw. methodische Arbeitsbereiche investiert werden muss, könnte zu ‚Lücken' im kanonischen Portfolio der Kerndisziplinen führen, die bei der Besetzung von Stellen mit einer Verankerung in der etablierten Institutsstruktur von Nachteil ist. Im Rahmen von CRETA verfolgen wir nicht zuletzt deshalb die Devise, dass der wissenschaftliche Nachwuchs auch Augenmerk auf die Entwicklungen in der eigenen ‚Mutterdisziplin' legen sollte – wenngleich die angeführte ‚Lücke' unvermeidlich ist.

Der gezielten Unterstützung stark interdisziplinärer Forschung in speziellen Förderlinien kommt daher erhebliches Gewicht zu bei der Erschließung neuer Felder der Grundlagenforschung. Das Stuttgarter Methodenzentrum konnte hier von einer sehr günstigen Konstellation von Umständen profitieren: Wie eingangs der Einleitung skizziert, hatte eine Gruppe von Forscherinnen und Forschern aus informatischen, geistes- und sozialwissenschaftlichen Disziplinen im Rahmen der BMBF-Förderung für eHumanities-Projekte (nach Ausschreibung 2011) Gelegenheit, Kooperationsstrukturen zu etablieren und die Grundlagen für eine erfolgreiche Beantragung eines Methodenzentrums (in Folge der BMBF-Ausschreibung von 2013/14) zu legen. Hinzu kam die aktive Unterstützung durch die Universitätsleitung, die nicht zuletzt durch die strukturelle Maßnahme einer neu geschaffenen Professur und Institutsabteilung *Digital Humanities* die Voraussetzung für die Vergabe der Zentrenförderung schuf. Das *Stuttgart Research Centre for Text Studies*[21] unter der Direktion von Sandra Richter und mit dem stellvertretenden Direktor Claus Zittel bot darüber hinaus den Rahmen für eine fakultätsübergreifende Verortung des Zentrums.

Die fünfjährige BMBF-Förderung für CRETA hat über das Auslösen der strategischen Strukturmaßnahmen an der Universität hinaus zu einer erheblichen Belebung der interdisziplinären Lehre und Forschung beigetragen – nicht zuletzt belegt durch eine Reihe von eingeworbenen Drittmittelprojekten, die zumeist bilaterale DH-Kooperationen vorantreiben (siehe S. 487 im Fazit des Bandes, Reiter und Pichler 2020). CRETA wird nach dem Auslaufen der BMBF-Projektfinanzierung im Rahmen der beteiligten Institutionen weitergeführt – damit steht freilich kein Stellenvolumen in vergleichbarem Umfang zur Verfügung. Der geschilderte Idealzustand der ‚Beinfreiheit' bei der Exploration innovativer Kooperationsideen innerhalb eines sehr weitreichenden interdisziplinären Raums wird mit dem Übergang in spezifische Einzelprojekte ein klein wenig geschmälert. Leider dürfte eine Dauerfinanzierung für ein größer angelegtes ‚Versuchslabor' für disziplinübergreifende Methodenentwicklung wie CRETA während der BMBF-Förderung kaum möglich sein.

Für die Fortführung der angestoßenen Entwicklungen spielt auch die Vernetzung innerhalb der DH-Community eine große Rolle. Die CRETA-Beteiligten haben über die Jahre der Förderung vielfältige bilaterale Kooperationsbeziehungen geknüpft, die aufzuführen hier zu weit führen würde. Besonders hervorzuheben sind enge Kooperationsbeziehungen mit dem Würzburger DH-Zentrum KALLIMACHOS[22], der DH-Gruppe an der Universität Hamburg (u. a. im Zusammenhang mit

21 http://www.srcts.uni-stuttgart.de

22 http://kallimachos.de

dem Projekt ForText[23], und den DH-Standorten an der Universität Trier und der Technischen Universität Darmstadt sowie mit dem SOCIUM (Forschungszentrum Ungleichheit und Sozialpolitik) an der Universität Bremen.

Ein wichtiger Schritt auf dem Weg zur Entwicklung von eigenständigen Teildisziplinen innerhalb der DH gelang mit der Einrichtung eines Schwerpunktprogramms *Computational Literary Studies* der Deutschen Forschungsgemeinschaft unter Federführung von Fotis Jannidis, Würzburg, an dessen Konzeption Forscherinnen und Forscher aus CRETA direkt beteiligt waren.

Der Kernstrang der Methodenentwicklung in CRETA ist bewusst so ausgerichtet, dass er gleichermaßen auf informatischer Seite wie auf Seiten der Geistes- und Sozialwissenschaften Herausforderungen von aktuellem Belang für die jeweilige disziplinäre Forschung angeht. So kann gewährleistet werden, dass die Modellierungsaktivitäten von den neuesten Entwicklungen in den Teildisziplinen profitieren können. Außerdem können die Nachwuchswissenschaftlerinnen und -wissenschaftler parallel zur Profilierung im DH-Spektrum in den Foren der Kerndisziplinen publizieren und etwa in den entsprechenden Fächern promoviert werden.

Einem Weiterverfolgen dieses Kernstrangs kommt in CRETA und anderen DH-Forschungszentren und -projekten auf absehbare Zeit große Bedeutung zu – nicht zuletzt aufgrund fortwährender technischer Neuentwicklungen (u. a. bei künstlichen neuronalen Netzwerkarchitekturen, die in der Lage sind, subtile ebenenübergreifende Muster aus Textsammlungen abzuleiten). Daneben ist es für die DH-Community jedoch ähnlich wichtig, dass Pfade für die Verbreiterung der Werkzeugnutzung vorgezeichnet werden, über die Forschungsprototypen aus einer erfolgreichen ‚experimentellen' Methodenforschung in eine Phase der ‚produktiven' Nutzung im Forschungsalltag übergehen können. Strukturell ist die Abdeckung dieses wichtigen Entwicklungspfades allerdings zum Teil ungeklärt: Innerhalb von projektgebundenen Forschungsaktivitäten, in denen der akademische Nachwuchs die Forschung vorantreibt und parallel auf Qualifikationsarbeiten zuarbeitet, kann eine umfassende softwaretechnische Einbettung von Forschungstypen in robust einsetzbare Werkzeuglösungen nicht geleistet werden. Weniger komplexe Forschungswerkzeuge lassen sich unter Umständen direkt in das technische Angebot eines Forschungsdatenmanagement einbinden, für welches zentrale Einheiten der Universitäten wie Rechenzentren und Bibliotheken in den letzten Jahren erhebliche Ressourcen bereitstellen und das darüber hinaus im Zentrum von Infrastrukturinitiativen steht. Für das vielschichtige Analyseinventar, das in der reflektierten algorithmischen Textanalyse zum Einsatz

23 https://fortext.net

kommt, wären jedoch zusätzliche, personalintensive Schritte notwendig, um von Forschungsprototypen zu breiter einsetzbaren Lösungen zu kommen. Stünden nicht finanzielle Beschränkungen im Wege, ließe sich eine erheblich größere Wirkung durch die DH-Methodenentwicklung erzielen, indem (i) professionelle Softwareentwicklerinnen mit einer Reimplementierung von Prototypen aus der Methodenforschung betraut würden (insbesondere auch in Hinblick auf die Anpassbarkeit an spezifische Korpuseigenschaften), und (ii) kompetent besetzte Service-Teams in Forschungseinrichtungen oder an anderen zentralen Stellen bereitstünden, die eine Anpassung entsprechender Werkzeuge an Gegenstand und Fragestellung in spezifischen Anwendungskontexten unterstützen.

4 An welches Publikum richtet sich dieser Band?

Dieser Band führt Arbeiten aus den verschiedenen Regionen des interdisziplinären Spektrums zusammen, das in CRETA abgedeckt wird. Textanalytische Fragestellungen aus verschiedenen geistes- und sozialwissenschaftlichen Disziplinen werden beleuchtet, unterschiedliche Computermodelle und computergestützte Verfahren aus Teilbereichen der Informatik und Computerlinguistik kommen zur Anwendung. Einige Artikel beziehen auch einen Blickpunkt der Metareflexion und betrachten das Vorgehen in CRETA wissenschaftstheoretisch oder dokumentieren übergreifende Maßnahmen des Zentrums. Gemeinsam ist allen, dass sie ihre zentralen Gedanken jeweils einem breiten Publikum zu vermitteln versuchen, um so in der Zusammenschau des Bandes ein facettenreiches Bild der interdisziplinären Arbeit zu zeichnen.

Für die Lektüre erscheinen uns innerhalb des disziplinären Spektrums der digitalen Geisteswissenschaften drei Zugangsweisen naheliegend. Auf einen zusätzlichen denkbaren Lektürezugang, der über den engeren DH-Fachzusammenhang hinaus führen könnte, gehen wir zum Schluss noch kurz ein.

Innerhalb der DH fügen sich die einzelnen Beiträge erstens mit ihren jeweils spezifischen Konstellationen aus fachwissenschaftlichem Untersuchungsgegenstand, Fragekomplex und methodischem Ansatz zur algorithmischen Textanalyse in unterschiedliche spezielle Diskurse innerhalb der DH und der beteiligten Spezialdisziplinen ein. Dieser Aspekt muss hier nicht weiter diskutiert werden, die primäre Leserschaft eines Beitrags wird sich jeweils selbst angesprochen fühlen (so wenden sich zum Beispiel Krautter (2020), Richter (2020) und Viehhauser (2020) an die DH-Teilcommunity mit einem literaturwissenschaftlichen Fokus, Pichler, Blessing et al. (2020) an eine philosophisch interessierte Leserschaft und gleichzeitig an eine methodologisch definierte Teilcommunity mit Interesse an Verfah-

ren zur Netzwerkanalyse; Kantner und Overbeck (2020) sprechen ein sozialwissenschaftliches Publikum an, Klinger et al. (2020) ein Publikum an der Schnittstelle zwischen komputationeller Emotionsanalyse und Literaturwissenschaften usw.). Das systematische Nebeneinanderstellen von Beiträgen, die in vielfältiger Weise untereinander verwandt sind, lädt die Leserin oder den Leser hoffentlich zu Seitenblicken ein und mag im günstigen Fall hier und da eine inspirierende Wirkung entfalten – ähnlich wie das disziplinübergreifende Umfeld von CRETA für seine Akteure selbst.

Eine zweite mögliche Zugangsweise richtet sich auf Details und praktische Fragen der Umsetzung methodisch-analytischer Problemstellungen in den DH: Welches Vorgehen eignet sich für die Operationalisierung relevanter Konzepte für die Textanalyse? Auf welcher Grundlage geschieht eine iterative Verbesserung von Annotationsrichtlinien? Was sind geeignete Klassen von Computermodellen für einen bestimmten Aufgabentypus und wie können sie etwa für eine semiautomatische Vorhersage eingesetzt werden? Die ausführliche Darstellung einer größeren Zahl von konkreten DH-Aktivitäten in diesem Band – unter weitgehend vergleichbaren methodologischen Grundannahmen – ist auch eine Dokumentation der vielen konkreten Einzelentscheidungen zur DH-Arbeitspraxis, zum Beispiel in den eben genannten fachspezifischen Kapiteln. Anregungen und Erfahrungswerte mögen hier in vielen Fällen auch über die spezifischen Untersuchungskontexte hinweg von Interesse sein. Zu den inhaltlich fokussierten Kapiteln kommen zudem solche, die direkt eine kontextübergreifende Betrachtungsweise zur Methodik verfolgen (wie der Beitrag von Pagel et al. (2020) zum Annotations-Workflow über unterschiedliche Zielsetzungen hinweg oder der von Ketschik, Blessing et al. (2020) zu fachübergreifenden Fragen der Annotation von Entitäten) oder Vorschläge für eine praktische Verfahrensweise beinhalten (wie der Beitrag von Reiter (2020) zur Erstellung von Annotationsrichtlinien). Weder ein einführender Text zu einer DH-Methode, noch ein eher knapper Konferenzbeitrag zu bestimmten Studienergebnissen kann auf Einzelheiten eingehen, die hier im jeweiligen Zusammenhang dargestellt sind. Gerade zu arbeitspraktischen Aspekten, die keinen unmittelbaren Einfluss auf die Hauptergebnisse einer Untersuchung haben, fehlen in knapperen Fachbeiträgen aussagekräftige Hinweise. Sie sind jedoch für Projekte, die sich in der einen oder anderen Weise an Vorarbeiten orientieren wollen, von großer Relevanz. Das Begriffsregister zu diesem Band soll einen entsprechenden Gebrauch dieser Sammlung unterstützen.

Mit der Dokumentation der arbeitspraktischen bzw. methodischen Entscheidungen in diesem Band verbindet sich auch die dritte Zugangsweise, der wir den Weg bereiten wollen. Außer dem skizzierten praktisch orientierten Interesse an der Sammlung konkreter Erfahrungsberichte kann eine solche die Basis für eine Metareflexion des Vorgehens darstellen – etwa für eine vergleichende

Gegenüberstellung verschiedener methodischer Grundansätze und nicht zuletzt für weitergehende wissenschaftstheoretische Betrachtungen. Ausführliche *zurückblickende* Dokumentationen zum konkreten Vorgehen in größeren interdisziplären Kooperationsprojekten finden sich vergleichsweise selten.[24] In der metamethodischen Diskussion scheint es ein – durchaus erklärliches – Übergewicht von *vorweggenommenen* Beiträgen zu geben: programmatische Konzeptpapiere, die sich mit einer Kooperationsidee für interdisziplinäre Konstellationen auseinandersetzen, oder Panel-Beiträge, die weitgehend abstrakt oder hypothetisch über Modelle einer disziplinübergreifenden Zusammenarbeit nachdenken.[25]

Das beobachtete Übergewicht von programmatischen Betrachtungen im Vergleich zu detaillierteren rückblickenden Dokumentationen, wie der vorliegende Band sie anbietet, erklärt sich zu einem guten Teil sicherlich aus dem typischen Publikationsverhalten im Projektverlauf: Vor Projektbeginn und in frühen Projektphasen – bevor konkrete Resultate vorliegen – liegt es nahe, programmatische Überlegungen zu publizieren. Später im Projektverlauf wiederum sind es die handfesten Studienergebnisse, die in den Publikationsorganen den höheren Stellenwert einnehmen und daher oft mit größerer Priorität behandelt werden.

Nicht zuletzt der tatsächliche oder gefühlte Druck, in Publikationen mit konkreten Resultaten an die Fragestellungen anzuknüpfen, die in der jeweiligen Disziplin als aktuell offen problematisiert sind, kann jedoch auch zu einer Unterrepräsentation von arbeitspraktisch-methodischen Bewertungen von interdisziplinären Kooperationen führen. Publikationsorgane für die disziplintypischen Kernbeiträge mögen weniger geeignet erscheinen, um solche Bewertungen zu platzieren – erfordern diese doch abwägende Betrachtungen, die sich aus ihrer Natur heraus nicht vollständig innerhalb des kanonischen methodischen Rasters verhandeln lassen. Somit sind nicht nur Anhaltspunkte zu negativen Studienverläufen, sondern auch durchaus positive arbeitspraktische Einsichten jenseits der ‚harten' interdisziplinären Untersuchungsergebnisse unterrepräsentiert. In Debatten zur Machbarkeit von riskanteren interdisziplinären Unternehmungen kann dies dazu beitragen, dass abstrakten Überlegungen mehr Gewicht zukommt als belegten

24 Ein Beispiel ist Allison et al. 2011.

25 Gerade um die Mitte der 2010er-Jahre wurde der Diskurs zur disziplinären Ausrichtung im Spannungsfeld der DH in vielen Foren geführt – beispielsweise beim Dagstuhl-Seminar 14 301 „Computational Humanities – Bridging the gap between Computer Science and Digital Humanities" im Juli 2014 (Biemann et al. 2014), beim DHd Workshop „Informatik und die Digital Humanities" am 3. November 2014 in Leipzig und bei einer Reihe von Panels auf den DHd-Konferenzen 2014 ff. Ein weiteres disziplinäres Spektrum lag dem DFG-Strukturierungsprojekt „Digitaler Wandel in den Wissenschaften" zugrunde, das 2019 stattfand; viele Fragen der Interdisziplinariät finden jedoch außerhalb der DH eine Entsprechung.

Erfahrungen. Der vorliegende Band schließt den Bericht über Überlegungen und konkrete Erfahrung zu einer effektiven interdisziplinären Arbeitspraxis ein, insbesondere im Beitrag von Reiter, Kremer et al. (2020) (ab S. 467) und dem Fazit des Buches (ab Seite 485).

Damit sind wir bei einer möglichen vierten Zugangsweise zu den Beiträgen in diesem Band angelangt, die jenseits der Kerncommunity der DH liegt: Obgleich die dokumentierten Arbeitsschritte sehr konkret an die methodischen und inhaltlichen Problemstellungen der geistes- und sozialwissenschaftlichen Textanalyse anknüpfen, sind sie durchweg auch Beispiele für den Umgang mit einer stark interdisziplinär ausgerichteten Gesamtkonstellation im Spannungsfeld der ‚Digitalisierung' von Arbeitsfeldern, in denen komplexe und voraussetzungsreiche Problemstellungen auf Basis von heterogenen Datensammlungen bearbeitet werden. Hierzu gehören durchaus auch Fragestellungen in technischen Anwendungsfeldern, die sich einer oberflächennahen Operationalisierung entziehen und/oder für die nur begrenzte Datenmengen verfügbar sind – etwa wenn mögliche Auswirkungen von relevanten Ereignissen auf unterschiedliche Bevölkerungsschichten anhand von dokumentierten Reaktionen untersucht werden oder wenn Trainingsdaten, die zur Entwicklung einer komplexen *machine-learning*-Architektur eingesetzt werden, möglicherweise gesellschaftliche Ungleichheiten (*biases*) widerspiegeln. Und selbst bei der Operationalisierung von technisch-naturwissenschaftlichen Konzepten für eine effektive Wissensextraktion aus großen Textsammlungen ist ein reflektiertes Vorgehen zur iterativen Verbesserung des Systemverhaltens von Vorteil.

Technologische Werkzeuge, die zu Teilen auf maschinellen Lernverfahren basieren, sind mittlerweile in nahezu allen Bereichen von Wissenschaft und Gesellschaft verbreitet – und fast immer sind komplexe Prozesse bei der Datenauswahl und -erweiterung, Spezifikation von relevanten Analysekategorien, bei der Systemoptimierung und bei der Kombination von Teilkomponenten mit möglichen Wechselwirkungen im Spiel – eventuell versteckt. Insofern kommt einer fundierte Reflexionspraxis für komplexe *workflows* und Systemarchitekturen mit datenorientierten algorithmische Komponenten potenziell eine Relevanz zu, die weit über den Bereich der Digital Humanities hinaus reicht. Der Rahmen eines akademischen Zugangs zu Fragekomplexen aus den Geistes- und Sozialwissenschaften stellt dabei ein Experimentierfeld mit Pilotcharakter für die typischen Herausforderungen der vielschichtigen Abhängigkeiten dar: Als mögliches Fundament für die Reflexionspraxis können theoretische Überlegungen aus den jeweiligen Kerndisziplinen angesetzt und geeignet weiterentwickelt werden. So erschließen sich unterschiedliche Explorationspfade für die Suche nach arbeitspraktischen Modellen, die eine fundierte Methodenreflexion in *workflows* einer gemischt technischen/intellektuellen Bearbeitung des Gegenstands unterstützen. Die gewonne-

nen methodologischen Einsichten dürften sich schließlich auch auf Anwendungsfelder übertragen lassen, zu denen keine vielschichtige Tradition der Methodenreflexion vorliegt.

Primärliteratur

Becher, Johann Joachim (1682). *Närrische Weißheit Und Weise Narrheit*. Frankfurt. URL: http://www.deutschestextarchiv.de/becher_narrheit_1682 (besucht am 1. Juni 2020).

Sekundärliteratur

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti und Michael Witmore (2011). „Quantitative Formalism: An Experiment". Pamphlet 1. Stanford Literary Lab. URL: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf (besucht am 1. Juni 2020).

Altmann, Gerry T. M., Hrsg. (1990). *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. ACL-MIT Press series in natural language processing. Cambridge und London: MIT Press.

Barth, Florian (2020). „Annotation narrativer Ebenen und narrativer Akte". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 423–438.

Baumann, Martin, Steffen Koch, Markus John und Thomas Ertl (2020). „Interactive Visualization for Reflected Text Analytics". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 270–296.

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum und Alexander Mehler (2014). „Computational Humanities - bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301)". In: *Dagstuhl Reports* 4.7, S. 80–111. DOI: 10.4230/DagRep.4.7.80.

Ciula, Arianna und Øyvind Eide (2016). „Modelling in digital humanities: Signs in context". In: *Digital Scholarship in the Humanities* 32.suppl_1, S. i33–i46. DOI: 10.1093/llc/fqw045.

Danneberg, Lutz und Andrea Albrecht (2017). „Beobachtungen zu den Voraussetzungen des hypothetisch-deduktiven und des hypothetisch-induktiven Argumentierens im Rahmen einer hermeneutischen Konzeption der Textinterpretation". In: *Journal of Literary Theory* 10.1, S. 1–37.

Escudero, Gerard, Lluis Marquez und German Rigau (2000). „A comparison between supervised learning algorithms for word sense disambiguation". In: *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*. Association for Computational Linguistics, S. 31–36.

Flanders, Julia und Fotis Jannidis, Hrsg. (2019). *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*. London: Routledge.

Fodor, Jerry A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Massachusetts: MIT Press.

Gerstorfer, Dominik (2020). „Entdecken und Rechtfertigen in den Digital Humanities“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 107–123.

Joshi, Vidur, Matthew Peters und Mark Hopkins (2018). „Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples“. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, S. 1190–1199.

Kantner, Cathleen und Maximilian Overbeck (2020). „Exploring Soft Concepts with Hard Corpus-Analytic Methods“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 170–189.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Kirschenbaum, Matthew G. (2012). „What Is Digital Humanities and What’s It Doing in English Departments?“ In: *ADE Bulletin* 150. Hrsg. von Matthew K. Gold, S. 55–61. DOI: 10.1632/ade.150.55.

Klinger, Roman, Evgeny Kim und Sebastian Padó (2020). „Emotion Analysis for Literary Studies“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 238–268.

Krautter, Benjamin (2020). „‚Figurenstil‘ im deutschsprachigen Drama (1740–1930)“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 299–326.

Kuhn, Jonas (2020). „Computational Text Analysis within the Humanities“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 63–106.

Kuhn, Jonas, Axel Pichler, Nils Reiter und Gabriel Viehhauser (2020). „Textanalyse mit kombinierten Methoden – ein konzeptioneller Rahmen für reflektierte Arbeitspraktiken“. In: *DHd 2020 Digital Humanities: Spielräume. Conference abstracts*. Paderborn, S. 223–227.

McCarty, Willard (2005). *Humanities Computing*. London: Palgrave.

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee und Luke Zettlemoyer (2018). „Deep Contextualized Word Representations“. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Volume 1: Long Papers)*. Association for Computational Linguistics, S. 2227–2237.

Pichler, Axel, André Blessing, Nils Reiter und Mirco Schönfeld (2020). „Algorithmische Mikrolektüren philosophischer Texte“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 328–372.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Ramsay, Stephen und Geoffrey Rockwell (2012). „Developing Things: Notes toward an Epistemology of Building in the Digital Humanities“. In: *Debates in the digital humanities*. Hrsg. von Matthew K. Gold. Minneapolis/London: University of Minnesota Press, S. 75–84. DOI: 10.5749/minnesota/9780816677948.003.0010.

Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.

Reiter, Nils, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler (2020). „*Reaching out*: Interdisziplinäre Kommunikation und Dissemination“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 467–484.

Reiter, Nils und Axel Pichler (2020). „CRETA, ein erstes Fazit“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 485–491.

Richter, Sandra (2020). „Reading with the workflow: Arbeitsprozesse in den Computational Literary Studies – Beiträge zur Empirisierung literaturwissenschaftlicher Verfahren“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 143–168.

Seeker, Wolfgang und Jonas Kuhn (2013). „Morphological and Syntactic Case in Statistical Dependency Parsing“. In: *Computational Linguistics* 39.1, S. 23–55.

Sekine, Satoshi (1997). „The domain dependence of parsing“. In: *Proceedings of the 5th conference on applied natural language processing*. Association for Computational Linguistics, S. 96–102. DOI: 10.3115/974557.974572.

Thalheim, Bernhard und Ivor Nissen, Hrsg. (2015). *Wissenschaft und Kunst der Modellierung – Kieler Zugang zur Definition, Nutzung und Zukunft*. Berlin: De Gruyter. DOI: 10 . 1515 / 9781501501234.

Viehhauser, Gabriel (2020). „Zur Erkennung von Raum in narrativen Texten“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 373–388.

Willand, Marcus, Evelyn Gius und Nils Reiter (2020). „SANTA: Idee und Durchführung“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 391–422.

Axel Pichler und Nils Reiter

Reflektierte Textanalyse

Zusammenfassung: In diesem Beitrag diskutieren wir einen abstrakten *workflow* zur reflektierten Textanalyse: Ausgehend von einer disziplinären Forschungsfrage werden geeignete Arbeitsschritte und Teilfragen identifiziert, auf deren Basis dann die für sie zentralen Begriffe über Annotationen und Automatisierungen – wie z. B. *machine-learning* – operationalisiert werden. Die Anwendung dieser Auszeichnungsregeln auf die Korpusdaten führt mehrheitlich zu quantitativen Ergebnissen, die in der Gesamtschau interpretiert werden müssen. In diese Gesamtschau fließen neben den Details der Operationalisierung auch weitere Vorannahmen wie z. B. disziplinäres Domänenwissen ein, deren Konsequenzen für die Interpretation kritisch reflektiert und berücksichtigt werden.

Abstract: This chapter discusses the central CRETA workflow. Starting with a research question from the humanities and/or social sciences, we define work packages and partial questions. On the basis of these questions the central terms are operationalized via annotations and automations – such as machine learning. The application of these labeling rules to the corpus data leads mostly to quantitative results, which are to be interpreted in a holistic fashion. In addition to the details of the operationalization, further assumptions such as domain knowledge are included in this overall view, the consequences of which are critically reflected and considered for interpretation.

1 Einleitung

Das *Center for Reflected Text Analytics* (CRETA) hat sich zum Ziel gesetzt, technische Methoden und Arbeitsablaufpraktiken zur Textanalyse im Forschungsbereich der Digital Humanities zu entwickeln. Die Methoden und Praktiken sollen fachübergreifend für textanalytische Fragestellungen aus der Literatur-, Sprach-, Geschichts- und Sozialwissenschaft sowie der Philosophie anwendbar sein. Bei der folgenden Darstellung handelt es sich um eine Abstraktion der im Rahmen von CRETA tatsächlich realisierten Praktiken. Der in Abbildung 1 dargestellte prototypische Arbeitsablauf kennzeichnet sich durch die enge Verzahnung von traditionellen geisteswissenschaftlichen Textumgangsformen mit komputationellen

Axel Pichler, Stuttgart Research Center for Text Studies, Universität Stuttgart
Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Axel Pichler und Nils Reiter, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-003

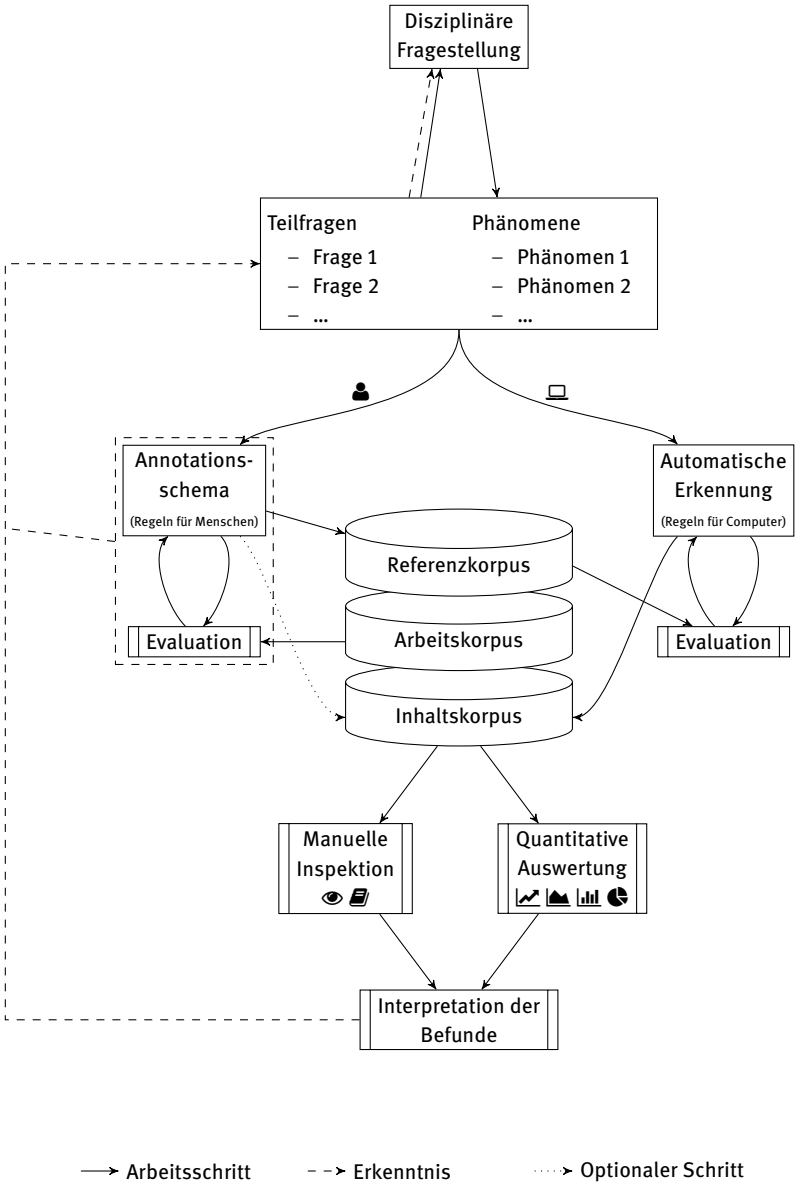


Abb. 1: Prototypischer Arbeitsablauf. Nicht bildlich dargestellt ist Domänenwissen, das als Hintergrund für viele Arbeitsschritte dient.

Verfahren sowie deren permanenter Reflexion. Die Summe dieser Praktiken und ihrer regelgeleiteten Verknüpfung bezeichnen wir als *reflektierte Textanalyse*.[1]

An deren Anfang steht stets eine disziplinäre Fragestellung, die eingangs in einzelne Teilfragen und Teilaufgaben segmentiert wird (siehe dazu den Abschnitt 2). Anwendungs- und damit Untersuchungsgegenstand sind Texte bzw. ausgewählte Korpora, deren Rollen in Abschnitt 3 diskutiert werden. Zur Beantwortung der leitenden Fragestellung sind diejenigen (Textoberflächen-)Phänomene auszuzeichnen, die für die Beantwortung der Frage relevant sind. Dies erfolgt auf Basis von Annotationsregeln, die sowohl von Menschen (Abschnitt 4) als auch automatisch (Abschnitt 5) umgesetzt werden können. Diese Regeln stellen das Resultat eines experimentellen und iterativen Arbeitens dar, das wir in Abschnitt 6 nochmal gesondert besprechen. Die solcherart gewonnen Befunde werden in einem letzten Schritt ausgewertet und so – falls möglich – die eingangs gestellte Frage beantwortet, wie wir in Abschnitt 7 erläutern.

2 Disziplinäre Fragestellung und Modularisierung

DH-Forschungsprojekte gehen häufig von bereits existierenden fachspezifischen Fragestellungen aus.[2] Derartig knüpfen sie nicht nur unmittelbar an den Forschungsstand der jeweiligen Disziplin an, sondern bauen auf deren *Domänenwissen* auf. Das *Domänenwissen* eines Faches umfasst sowohl implizite Vorannahmen bezüglich seiner Gegenstände als auch explizites fachspezifisches Hintergrundwissen bezüglich derselben, wie zum Beispiel das Wissen über interpretationsrelevante Kontexte.[3] Zu ihm zählt auch ein entweder als ausformulierte Methode vorliegendes oder bloß inkorporiertes Wissen, wie mit den Untersu-

1 ‚Reflexion' bezeichnet dabei die Theoretisierung der impliziten Vorannahmen und Handlungsregeln des Praxisvollzuges.

2 In den textorientierten Geisteswissenschaften existiert hingegen eine Vielzahl an unterschiedlichen Textumgangsformen. So hat z. B. Carlos Spoerhase in Anknüpfung an Andrew Abbot das fälschlicherweise als Inbegriff des geisteswissenschaftlichen Lesens verstandene kontinuierlich-narrative Lesen – d. h. die vollkommen vereinnahmte und nicht-unterbrochene Lektüre – den Praktiken des meditativen, argumentativen und scannenden Lesens sowie der Stellenlektüre gegenübergestellt (Spoerhase 2015). Diese Praktiken erfüllen unterschiedliche Funktionen in der geisteswissenschaftlichen Forschung. Prinzipiell können sie jedoch alle sowohl einem bestimmten Interpretationsziel folgen als auch zur Erzeugung einer Fragestellung zum Einsatz kommen.

3 Im Hinblick auf den Kontextbegriff folgen wir dessen Ausdifferenzierung durch Danneberg (2000). Danneberg unterscheidet zwischen intra-, infra-, inter- und extratextuellen Kontexten, wobei unter intratextuellem Kontext die „Beziehung eines Teiles eines Textes zu anderen Ausschnitten desselben Textes in (a) thematischer oder (nur) (b) sequentieller Hinsicht", unter in-

chungsgegenständen des Faches sowie den genannten Kontexten umzugehen ist. Die hier genannten Formen des *Domänenwissens* können ausdifferenziert werden in Wissen, das in Form von Aussagesätzen vorliegt (= ‚wissen, dass' bzw. *knowing that*) und prozedurales Wissen (= ‚wissen, wie' bzw. *knowing how*), das insgeheim die Forschungspraxis leitet und potenziell explizit gemacht werden kann.[4]

Für die leitenden Fragestellungen von textorientierten DH-Projekten folgt daraus, dass sie in hohem Grade fachspezifisch beeinflusst sind: Wonach und auf welche Art und Weise gefragt wird, hängt von einer Vielzahl disziplinärer Faktoren ab, die im Regelfall nicht vollständig freigelegt werden.[5] Nach unserem Verständnis unterscheiden sich textorientierte DH-Projekte von traditionelleren geistes- und sozialwissenschaftlichen Arbeiten darin, dass sie ihre Forschungsaktivitäten in klar bestimmte Teilschritte gliedern müssen, deren Abfolge und Zusammenhang ebenso präzise bestimmt werden müssen. Bewusst oder unbewusst orientieren sie sich dabei an einer schwachen Variante des seit den zwanziger Jahren des vorigen Jahrhunderts intensiv diskutierten Operationalismus,[6] dem zufolge die Operationalisierung eines Begriffes sowohl zu dessen Schärfung beiträgt, als auch seine empirische Überprüfbarkeit vereinfacht.[7] Operationalisierung bezeichnet im

---

fratextuellem Kontext die „Beziehung eines Textes oder Textabschnittes zum Textganzen", unter intertextuellem Kontext die „Beziehung eines Text(ausschnitt)s zu (a) bestimmten Textklassen oder (b) zu anderen Texten bzw. Textausschnitten" und unter extratextuellem Kontext die „Beziehung eines Textes zu nichttextuellen Gegebenheiten" verstanden wird.

4 Zum Stand der Diskussion dieses Verhältnisses in Philosophie und Wissenschaftstheorie siehe einführend: Fantl 2017.

5 Wie eine konkrete Frage entsteht, die von der *scientific community* als relevant erachtet wird, ist selbst Forschungsgegenstand, da es dafür mehrheitlich keine expliziten methodischen Anleitungen gibt. Die sich damit beschäftigenden Forschungsrichtungen sind Wissenschaftstheorie und historische Epistemologie, welche die Geschichte wissenschaftlicher Ideen rekonstruiert, indem sie deren Entstehung, Rechtfertigung, Verbreitung, u. a. m. in unterschiedlichen Kontexten untersucht (cf. Rheinberger 2007). Vergleiche hierzu auch den Beitrag von Gerstorfer (2020) zur Unterscheidung zwischen Entdeckung und Rechtfertigung, in diesem Band ab Seite 107.

6 Die Grundannahme der starken Version des Operationalismus lautet in ihrer ersten Ausformulierung durch den späteren Physiknobelpreisträger Percy Bridgman (1927, S. 5): „[W]e mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations". Bridgman hat das darin nahegelegte bedeutungstheoretische Verständnis des Operationalismus später selbst zurückgewiesen und im Zuge der Auseinandersetzung mit seinen Kritikern letztendlich eine schwache Version des Operationalismus entwickelt, die Operation und Aktivität gleichsetzt und Wissenschaft als eine in unterschiedliche Aktivitäten untergliederbare Praxis versteht (cf. Bridgman 1938). Das wissenschaftstheoretische Potential dieser Auffassung sowie die für dessen Entfaltung zu erfüllenden Desiderata sind jüngst von Hasok Chang ausformuliert worden (Chang 2019).

7 Hans Reichenbach hat 1928 im Zuge seiner Auseinandersetzung mit der ‚Zuordnungsdefinition' – d. i. die Zuordnung von Begriffen zu wirklichen Dingen – gezeigt, dass Zuordnungen will-

Folgenden dementsprechend sowohl die (Teil-)Schritte bzw. Aktivitäten, die notwendig sind, um einen Begriff empirisch überprüfbar bzw. messbar zu machen, als auch das Resultat dieser Aktivitäten.[8] Für die konkrete DH-Projektarbeit hat das zur Konsequenz, dass diese damit einsetzt, komplexe fachspezifische Fragen in weniger komplexe Teilfragen zu gliedern, um im Anschluss daran die dafür relevanten Begriffe zu operationalisieren, indem diese – eventuell über Zwischenschritte – mit Textoberflächenphänomenen verknüpft werden. Im Zuge dessen kann sich herausstellen, dass die für eine bestimmte fachspezifische Fragestellung relevanten Begriffe nicht operationalisierbar sind. Die Gründe dafür sind vielfältig. Zu ihnen zählen neben der Vagheit und Ambiguität von bestimmten Fragen bzw. der für diese relevanten Begriffe insbesondere die zu ihnen führenden Vorannahmen. Ob sich die für eine bestimmte Frage relevanten Begriffe operationalisieren lassen, zeigt sich letztendlich nur im Zuge des Versuches, dies zu tun.

Der im vorigen Absatz beschriebenen Operationalisierungspraxis entspricht ein Verständnis von Algorithmus, nach welchem es sich bei einem solchen um eine eindeutige Handlungsvorschrift zur Beantwortung einer Frage bzw. zur Lösung eines Problems handelt. In diesem weiten Verständnis von Algorithmus können Handlungsschritte formalisiert werden, müssen es aber nicht. Damit hängt die Frage der Granularität dieser Handlungsschritte zusammen. Letztlich müssen bestimmte Operationen als ‚atomar' gegeben angenommen werden, die keine weitere Operationalisierung erfordern. Bei Algorithmen, die in Programmcode überführt werden, sind dies die Konstrukte und Operationen, die eine Programmiersprache und ihre Standardbibliothek bereitstellen (etwa: Multiplikation zweier Zahlen). Bei Handlungsvorschriften für Menschen könnten dies Handlungen sein, die Menschen gemeinhin direkt ausführen können. Derartige Handlungsvorschriften bestehen aus endlich vielen Teilschritten, die in einzelnen Modulen abgehandelt werden. Die folgende Darstellung orientiert sich an jenen Teilschritten, die sich im Rahmen von CRETA als zielführend erwiesen haben. Die in diesem Band beschriebenen Projekte sind von disziplinären Fragestellungen ausgegangen, bei denen der quantitative Blick einen heuristisch wertvollen Beitrag zu leis-

kürlich erfolgen, dementsprechend selbst nicht empirisch überprüft werden können, jedoch die empirische Überprüfbarkeit der von ihnen abhängigen Begriffe ermöglichen (Reichenbach 1928, S. 23 ff.). Die Operationalisierung von Begriffen hängt also davon ab, ob für das Begriffssystem, aus dem diese Begriffe stammen, bereits derartige Zuordnungsdefinitionen bestehen.

8 Als ‚Operationalisierung' kann sowohl der Prozess zur Entwicklung der Messung als auch deren Umsetzung bezeichnet werden. Im Beitrag von Pichler et al. (2020) etwa geht es um philosophische Begriffe: Deren Messbarkeit wird erarbeitet, daher beschreibt der Artikel insgesamt die Entwicklung einer Mess-Praxis für philosophische Begriffe.

ten versprach.[9] Gegenstand dieser Projekte sind Texte bzw. Textkorpora, die im Zuge einer reflektierten Textanalyse unterschiedliche Funktionen übernehmen, wie wir im folgenden Abschnitt zeigen werden.

3 Korpora

In Abbildung 1 sind drei verschiedene Korpora dargestellt: Das Arbeits-, Referenz- und Inhaltskorpus. Diese Bezeichnungen beziehen sich auf die *Funktion* der Korpora für die jeweiligen Arbeitsschritte. In der Praxis werden oft mehrere Funktionen von einem konkreten Korpus ausgefüllt. Generell verstehen wir unter einem Korpus eine maschinenlesbare Sammlung von Text- oder Sprachdaten. In verschiedenen Disziplinen existieren Korpora, in denen bereits bestimmte Phänomene annotiert wurden.[10]

Als **Arbeitskorpus** wird in Abbildung 1 das Korpus bezeichnet, das zur Entwicklung der Annotationsrichtlinien, also zur Arbeit am Phänomen, verwendet wird. Hier ist sehr auf die stilistische, inhaltliche und formale Passung zu achten: Änderungen an den Richtlinien (z. B. die Ergänzung von Kategorien) machen die bestehenden Annotationen unbrauchbar, erfordern also manuelle Nacharbeiten oder Neuannotationen. Das **Referenzkorpus** wird verwendet, um die automatische Erkennung des fraglichen Phänomens zu evaluieren. Es muss dazu Annotationen enthalten, mit denen die Ergebnisse einer automatischen Annotation verglichen werden können. Damit verbindet das Referenzkorpus die manuelle mit der automatischen Annotation. Hier ist auch eine zeitliche Ordnung impliziert: Ohne Referenzkorpus ist die Entwicklung einer validen automatischen Erkennung nicht möglich. Gerade um den sich hier abzeichnenden ‚*scheduling conflict*' (Kuhn 2020a, S. 81) zu vermeiden, kann auch zunächst ein bereits exis-

9 So fußen zum Beispiel die algorithmischen Analysen in Braun und Ketschik 2019 auf der literaturwissenschaftlichen These, dass Figurenbeziehungen je nach Textsorte variieren und für die Kennzeichnung dieser Textsorten von Relevanz sind. Da sich mit Methoden der Netzwerkanalyse Figurenbeziehungen quantitativ fassen lassen, verspricht eine solche Analyse hier einen Mehrwert.

10 Die Auszeichnung eines Textes bzw. einer Textsammlung mit maschinenlesbaren, strukturellen Informationen kann auch als *Datenmodellierung* bezeichnet werden. Datenmodelle sind formale Modelle, die textliche und textexterne Informationen maschinenlesbar machen, wofür sie sich einem konkreten Datenformat wie z. B. XML bedienen. Die hier vorgestellte Art der Textanalyse hat jedoch die Analyse von Daten zum Ziel, bei denen allenfalls Metadaten sowie Textstrukturinformationen (wie z. B. Kapitel, Bühnenanweisungen) modelliert wurden. Zur Einführung in die Datenmodellierung, deren Funktion und deren Grundbegriffe siehe Jannidis 2017.

tierendes Referenzkorpus für einen anderen Text oder eine andere Textsorte herangezogen werden. Änderungen an der automatischen Erkennung lassen sich im Regelfall durch den nochmaligen Aufruf eines Programms umsetzen. Spätere Änderungen an den Annotationsrichtlinien hingegen sind aufwändiger als Änderungen an der automatischen Erkennung. Das **Inhaltskorpus** bezeichnet das Korpus, das am Ende Gegenstand der inhaltlichen Analysen wird, also dasjenige Korpus, auf das sich die Fragestellung bezieht.

In der Praxis sind die Korpora oft nicht so klar getrennt wie hier dargestellt. Es können sehr gut Teile eines einzelnen Werkes als Referenz- und andere als Arbeitskorpus dienen.

4 Annotationsrichtlinien: Regeln für Menschen

Ausgangspunkt einer jeden Annotationsrichtlinie sind theoretische Vorüberlegungen oder eine ausgearbeitete Theorie über das relevante Phänomen. Auf dieser Basis wird eine erste Version von Annotationsrichtlinien erstellt, und einige Festlegungen bezüglich des Arbeitsablaufes getroffen: Welches Annotationswerkzeug wird verwendet? Welche Texte werden annotiert? Welche Annotationskategorien werden verwendet und wie ist ihr Verhältnis zu den theoretischen Konzepten? Wie werden Textoberflächenphänomene den Kategorien zugeordnet?[11]

Die weitere Arbeit erfolgt dann iterativ, im Wechselspiel aus testweiser, manueller Anwendung der Richtlinien durch Menschen und der Analyse der erzielten Annotation.[12] In frühen Stadien kann es sinnvoll sein, auf Papier zu annotieren, weil dabei flexibel auf Änderungen an den Annotationskategorien o. ä. reagiert werden kann. Sobald die Kategorien einigermaßen stabil sind, kann ein digitales Annotationstool verwendet werden. Wenn die Annotationen parallel und unabhängig erfolgen, können direkt die Annotationsunterschiede ausgewertet, analysiert oder untersucht werden. Dabei können digitale Werkzeuge eine große Hilfe sein, indem sie die Unterschiede direkt in ihrem Kontext anzeigen. Zur Analyse kann weiterhin auf quantitative Metriken des *inter-annotator agreement* zurück-

11 Dabei handelt es sich um eine Folge der Tatsache, dass textorientierte DH-Projekte an der Textoberfläche der von ihnen untersuchten Korpora ansetzen. Diese Zugangsweise ist nicht trivial, da einerseits die Art und Weise wie mit Textoberflächenphänomenen umgegangen wird, bedeutungstheoretische Konsequenzen zeitigt, andererseits bedeutungstheoretische Vorannahmen Anforderungen für den Umgang mit Textoberflächenphänomenen nach sich ziehen. Dieser Sachverhalt ist bei der konkreten Textanalyse stets mitzureflektieren.

12 Im Beitrag von Reiter (2020), ab Seite 193 des vorliegenden Bandes wird eine detailliertere Anleitung bereitgestellt.

gegriffen werden, die der Übereinstimmung eine Zahl zuweisen (z. B. im Interval [-∞;1]). Die Berechnung der Übereinstimmung sollte aber durch eine diskursive, qualitative Auswertung komplementiert werden, unter anderem indem die Annotierenden ihre Annotationsentscheidungen erklären (müssen).[13]

Ergebnis dieses Schrittes sind primär die Annotationsrichtlinien, die allerdings nur innerhalb eines gewissen Kontextes (intern) valide sind. Annotationsrichtlinien sind typischerweise zunächst auf die Textsorte bezogen, für die sie erstellt wurden, und können mehr oder weniger generisch sein (Annotationsrichtlinien, die z. B. für Dramen erstellt wurden, können sich andere Merkmale zunutze machen als Richtlinien für Prosatexte). Daneben können auch Tools entwickelt oder angepasst werden, die spezifisch für diese Richtlinien optimiert sind, und z. B. die Visualisierung von Annotationsunterschieden unterstützen.

5 Automatische Erkennung: Regeln für Computer

Für die automatische Erkennung von Textphänomenen können grundsätzlich zwei unterschiedliche Verfahren eingesetzt werden, die verschiedene Vor- und Nachteile aufweisen. Die Erkennung auf Basis von manuell erstellten, aber formalisierten **Regeln** eignet sich besonders für Phänomene, bei denen die Kontextabhängigkeiten überschaubar und/oder gut bekannt sind. Eine Regel könnte etwa spezifizieren, dass großgeschriebene Wörter in einem mittelhochdeutschen Text einer bestimmten Edition auf außersprachliche Entitäten verweisen, wenn sie nicht am Anfang eines Absatzes stehen. Diese einfach zu implementierende Regel greift dabei auf zwei Textkriterien zurück, die für ihre Anwendung bekannt sein müssen: (i) Großschreibung und (ii) Absätze. Während Groß- und Kleinschreibung einfach, d. h. ‚atomar', zu unterscheiden sind, müssen für die korrekte Erkennung von Absätzen die Textdaten so eingelesen worden sein, dass Absätze identifizierbar sind (etwa durch Leerzeilen, da einfache Absatzzeichen oft uneinheitlich verwendet werden).

Regeln können natürlich auch komplexere Kriterien verwenden als im Beispiel, sie müssen aber in jedem Fall maschinenlesbar sein. Linguistische Kriterien wie z. B. Satzgrenzen oder inhaltliche Kriterien wie Themenwechsel sind dies nur bei entsprechender Vorverarbeitung oder -annotation, z. B. mit computerlinguistischen Werkzeugen wie Satzgrenzen- oder Wortartenerkennern, die dann ihrer-

13 Siehe hierzu auch den Beitrag von Pagel et al. (2020), ab Seite 125 in diesem Band.

seits als atomar betrachtet werden können. Die Ergebnisse dieser Vorverarbeitung können jedoch fehlerbehaftet sein.[14]

Grundsätzlich spricht nichts dagegen, dass die Computer-Regeln den Regeln für manuelle Annotationen folgen. Sind die Annotationsregeln so beschaffen, dass sie sich direkt in einen deterministischen Algorithmus überführen lassen, ist die (manuelle) Annotation zur Datenerzeugung im Grunde überflüssig. Interessante Phänomene sind allerdings meistens nicht dieser Natur, sondern selbst in der manuellen Annotation hochkomplex.

Mit steigender Komplexität von Regelsystemen steigt auch ihre Intransparenz, so dass eine systematische Evaluation geboten ist. Dies liegt vor allem an möglichen Interaktionen von Regeln, die für Menschen schwer zu überblicken sind. Zur Evaluation sollten daher Referenzdaten vorgehalten bzw. erzeugt werden, mit denen die Qualität der Erkennung durch Regeln überprüft werden kann (s. u. zur Evaluation). Sind Referenzdaten vorhanden, können Regelsysteme in einem iterativen Verfahren entwickelt werden. Dabei wird zunächst eine Version der Regeln erstellt und auf die Referenzdaten angewendet. Die Ergebnisse dieser Anwendung können dann mit den Referenzdaten abgeglichen werden, so dass Regelschwächen identifiziert werden können. Diese können in einer neuen Version der Regeln verbessert werden.

Mit **maschinellen Lernverfahren** können ebenfalls Textphänomene erkannt werden. Die Verfahren eignen sich insbesondere für Phänomene, deren Kontextabhängigkeiten unbekannt oder unübersichtlich sind. Zum Beispiel können Menschen Anaphern[15] in den meisten Fällen auf Basis des Kontextes auflösen, ohne dass es klar definierte und bekannte Regeln gäbe, denen dabei gefolgt wird. Um maschinelle Lernverfahren einzusetzen, müssen Merkmale definiert werden, die bereits maschinenlesbar sind oder automatisch aus den Daten extrahiert werden können. Außerdem muss das Referenzkorpus groß genug sein, so dass ein Teil der Daten zum Training und der andere zum Testen verwendet werden kann. Der Lernalgorithmus untersucht nun systematisch die Merkmalsausprägungen der Instanzen in den Trainingsdaten. Ermittelt wird, welche Kombination von Ausprägungen mit welcher Zielkategorie gemeinsam auftritt. Der Algorithmus könnte z. B. lernen, dass kleingeschriebene Wörter in einem mittelhochdeutschen Text

14 Eine fehlerbehaftete Vorverarbeitung macht darauf aufbauende Verarbeitungsschritte nicht unmöglich. Solange die Fehler regelmäßig sind und sich in einem gewissen Rahmen bewegen, können sie durch folgende Arbeitsschritte auch wieder ausgeglichen werden. Außerdem wirken sich Fehler je nach Verarbeitungsschritt unterschiedlich aus (cf. Reiter 2014, S. 54 f.).

15 Anaphern sind rückbezügliche Satzteile, die nicht ohne Rückgriff auf vorher stehende Satzteile interpretiert werden können. Neben Pronomen fallen auch einige Nominalphrasen und Adverbien (z. B. „dort“) unter die Anaphern.

mit einer Wahrscheinlichkeit von 55 % Referenzen auf außersprachliche Entitäten sind,[16] wenn sie dem Possessivpronomen ‚mîn' folgen. Folgen sie auf ein anderes Wort, ist womöglich die Wahrscheinlichkeit eine andere oder ein weiteres Wort aus dem Kontext muss zur Entscheidung herangezogen werden. Die so gelernten Abhängigkeiten werden in einem Modell gespeichert.[17] Dieses Modell kann dann auf ‚neue' Daten angewendet werden, d. h. auf Daten die noch nicht annotiert wurden (etwa die aus dem Inhaltskorpus, dem unsere eigentlichen Fragen gelten).

In den letzten Jahren ist mit *deep learning* eine Variante des maschinellen Lernens populär geworden, bei der die Entscheidungen nicht auf Basis von für Menschen interpretierbaren Merkmalen und ihren Ausprägungen erfolgt. Stattdessen werden die Instanzen in hochdimensionalen Vektorräumen repräsentiert. Im Lernalgorithmus werden dann verschiedene *layer* mit mathematischen Operationen verschaltet und auf der Basis von Trainingsdaten Gewichtungen ermittelt. Mit diesen können dann neue Instanzen klassifiziert werden. Abgesehen von der nicht vorhandenen Transparenz wird beim *deep learning* das gleiche iterative Vorgehen verfolgt wie beim „klassischen" *machine learning* und auch bei der Entwicklung von Regelsystemen: Ein erstes Modell wird trainiert und auf den Testdaten evaluiert. Fehler, die das Modell bei der Vorhersage gemacht hat, lassen sich dann analysieren, und z. B. kategorisieren. Manche Fehlerkategorien sind mit maschinellen Lernverfahren womöglich nicht lösbar (etwa Tippfehler in den Daten), für andere können neue Merkmale oder Dimensionen definiert werden. Daran schließt sich ein zweites Experiment an, in dem mit der neuen Merkmalsmenge trainiert wird, woraufhin wieder evaluiert wird, etc.

Für beide, regelbasierte und *machine-learning*-basierte Wege, ist die **Evaluation** eine Schlüsselkomponente. Die Evaluation kann und sollte zunächst quantitativ erfolgen: Dazu werden die vom Modell gemachten Vorhersagen mit den ‚korrekten' Vorhersagen in den Testdaten verglichen. Im einfachsten Fall berechnet man daraus die sog. *accuracy*, also den Anteil der korrekt erkannten Instanzen. Je nach Fragestellung sind andere Metriken aufschlussreicher, etwa der Dreiklang aus *precision*, *recall*, und F_1 (cf. Manning und Schütze 1999, S. 267 ff.). An die quantitative Evaluation sollte eine Fehleranalyse anschließen. Im Rahmen der Fehleranalyse wird analysiert, welche Fehler das Modell gemacht hat, und war-

16 Siehe hierzu auch den Beitrag von Ketschik, Blessing et al. (2020), ab Seite 204 in diesem Band.

17 ‚Modell' wird hier in der Bedeutung eines *machine-learning*-Modells verwendet, in dem Zusammenhänge kodiert sind, die aus Trainingsdaten extrahiert wurden. Das Modell kann im Sinne einer Ein-/Ausgabe-Funktion verwendet werden, um Vorhersagen für neue Instanzen zu liefern. Siehe zu den unterschiedlichen Modellbegriffen in den DH auch die kurze Überblicksdarstellung in Kuhn 2020b, S. 12.

um. Je nach Datenmenge kann dies nur bei einer Stichprobe gemacht werden. Die Fehler in Kategorien einzuteilen, ist ein guter Weg um festzustellen, welche Fehlerursachen es sich lohnt zu beheben. Seltene Fehler können vielleicht in Kauf genommen oder regelbasiert behoben werden (womit sich bereits eine Möglichkeit zeigt, regelbasierte Systeme und maschinelle Lernverfahren zu kombinieren).

Bei der Evaluation ist generell zu bedenken, dass es zu sogenannten *overfitting*-Effekten kommen kann. Damit ist gemeint, dass ein Modell zwar für einen bestimmten Datensatz sehr gute Vorhersagen liefert, aber beim Training keine (oder nur wenig) Generalisierung stattgefunden hat. Das Modell ist sozusagen ‚überangepasst' auf einen bestimmten Datensatz. Eine strenge Aufteilung von Trainings-, Test-, und Developmentdaten ist ein Weg dem *overfitting* entgegenzuwirken, allerdings lässt sich ohne großflächige Tests oft nicht von vornherein sagen, ob *overfitting* stattgefunden hat.

6 Experimentelles und iteratives Arbeiten

Sowohl die manuelle Annotation als auch die Entwicklung automatischer Erkenner bedienen sich Kreisläufen, die in Abbildung 2 isoliert dargestellt werden. Auch wenn in den beiden Kreisläufen unterschiedliche Ziele verfolgt werden, folgen diese einem ähnlichen Muster: Eine erste Version von Regeln bzw. Parametern (für das maschinelle Lernverfahren) wird erstellt und auf Testdaten angewendet. Die Ergebnisse der Anwendung werden evaluiert und fließen in die nächste Version ein. Dieses iterative Vorgehen ist ein zentrales Element reflektierter Textanalyse. Damit die dabei erzielten Einzelergebnisse vergleichbar sind, muss auf ein explizit definiertes Testszenario geachtet werden. Die wiederholten Tests, die im Rahmen der Weiterentwicklung von Richtlinien, Werkzeugen oder Modellen durchgeführt werden, müssen möglichst unter gleichen Bedingungen durchgeführt werden, indem mögliche Einflussfaktoren (Sprachstufe, Länge, Edition, Gattung, ...) kontrolliert werden.

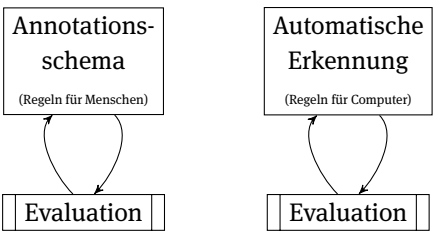


Abb. 2: Iterative Arbeitsabläufe aus Abbildung 1.

Für die Annotation von Entitätenreferenzen z. B. könnte eine Annahme sein, dass Nomen, die Städtenamen sind, immer auf Entitäten der Kategorie ‚Location' referieren. In den Annotationsrichtlinien würde dann eine entsprechende Regel formuliert. Wie das folgende Beispiel zeigt, entspricht diese Regel jedoch nicht dem tatsächlichen Sprachgebrauch:

(1) a. Dann reiste er nach Berlin.
b. Im Finanzstreit blieb Berlin hart.

Die Regel könnte auf Sätze wie 1a erfolgreich angewendet werden, ‚Berlin' referiert hier auf einen Ort. In Sätzen wie 1b dagegen referiert ‚Berlin' nicht auf einen Ort, sondern metonymisch auf die Regierung Deutschlands. In diesem Fall schlägt die Regel fehl und muss, nach Identifikation des Problems, geändert werden. Die in den Richtlinien formulierte Annahme wird also abgelehnt und eine neue aufgestellt, die dann als nächstes getestet wird.[18]

Dieses Vorgehen kann als experimentelles Arbeiten bezeichnet werden, da seine Struktur demjenigen einer wissenschaftlichen Erklärung im Sinne von Hempel und Oppenheim (1948) entspricht. Nach Hempels und Oppenheims hypothetisch-deduktivem Modell besteht ein derartiges Experiment aus einer Hypothese (oben als Annahme bezeichnet), bei der es sich gemeinhin um ein Konditional handelt, und notwendigen Randbedingen. Aus diesen beiden Prämissen können dann deduktiv empirisch überprüfbare Vorhersagen abgeleitet werden.

7 Interpretation der Befunde

Wir verstehen unter der Interpretation der Befunde die Auswertung der zuvor gewonnenen Informationen innerhalb des am Anfang eines Projektes abgesteckten Rahmens in Hinblick auf die damit einhergehenden Teilfragen.[19] ‚Auswertungen von Informationen' steht dabei für diejenigen Prozesse, in denen besagte

18 Bei der Aufstellung derartiger Regeln wird häufig auf den Schlussmodus der Abduktion zurückgegriffen. Siehe dazu den Beitrag von Gerstorfer (2020), ab Seite 107 in diesem Band.

19 In den textorientierten geisteswissenschaftlichen Disziplinen herrscht bezüglich des Begriffes ‚Interpretation' weder Übereinstimmung darüber, worin dessen spezifische Merkmale bestehen, noch, ob derartige Merkmale überhaupt existieren. Siehe dazu exemplarisch den Überblick der aktuellen literaturwissenschaftlichen Diskussion in Descher et al. 2015. In Anbetracht der anhaltenden geisteswissenschaftlichen Debatten über den Interpretationsbegriff, verzichten wir hier auf eine Auflistung seiner für CRETA spezifischen Merkmale und beschränken uns auf

Informationen als Ausdruck bzw. Indikator von *etwas* bestimmt werden. Dies sei anhand eines hypothetischen Beispiels erläutert, der Anwendung der narratologischen Erzählebenenbestimmung, deren Annotationsrichtlinien im Beitrag von Ketschik, Murr et al. (2020),ab S. 440 in diesem Band, entwickelt werden. Dessen Anwendung auf drei Korpora ausgewählter deutscher Prosatexte aus klar definierten Zeiträumen könnte zu dem Resultat führen, dass es in einem dieser Korpora zu einem statistisch relevanten Anstieg an Prosatexten kommt, die mit Rahmen- und Binnenerzählungen arbeiten. Dieser Sachverhalt wäre dann mit dem gängigen literaturwissenschaftlichen Wissen zur Prosa aus den besagten Zeiträumen abzugleichen und könnte auf diese Weise die diesbezüglichen Auffassungen bestätigen oder unterlaufen. Eine Bestätigung literaturwissenschaftlichen Wissens kann zunächst als (weitere) Validierung der Methode angesehen werden, sollte aber dennoch gründlich reflektiert werden. Womöglich wurden auf Basis des (gleichen) literaturwissenschaftlichen Wissens implizite Vorannahmen getroffen, die die Ergebnisse letztlich zirkulär machten. Kann literatur- bzw. geisteswissenschaftliches Wissen nicht direkt bestätigt werden, ist ebenfalls eine gründliche Reflexion des textanalytischen Prozesses sowie auch des Fundamentes des geisteswissenschaftlichen Wissens nötig. Je nach Quellenlage kann vielleicht identifiziert werden, woher die Annahme kam und warum sie bisher noch nicht widerlegt wurde. In jedem Fall ist das Ergebnis nicht das Ende der Forschungstätigkeit, sondern oft der Auftakt zu neuen Fragestellungen.

Wie dieses Beispiel zeigt, kommen bei der Auswertung der Befunde unterschiedliche Arten von Informationen zusammen, die in der Gesamtschau interpretiert werden müssen: Einerseits kann hier nun auf quantitative Ergebnisse zugegriffen werden, andererseits liefert die (manuelle) Inspektion von Einzelstellen weitere Informationen, etwa zu möglichen intratextuellen Kontexten. Beides muss vor dem Hintergrund von disziplinärem Domänenwissen ‚gelesen' werden, wobei hier auch das Domänenwissen aus der Informatik relevant ist, bei dem es sich vorwiegend um Methodenwissen handelt.

Die Auswertung der quantitativen Ergebnisse ist abhängig von messtheoretischen Vorannahmen und den damit einhergehenden Skalierungen, sowie einer Reihe konkreter Entscheidungen die im Zuge der Auswertung oder einer visuellen Darstellung getroffen werden. Für die Auswertung der quantitativen Ergebnisse sind Visualisierungen in Form von Balkendiagrammen, Zeitreihen oder Boxplots typisch. Dazu sind Metadaten entscheidend, mit denen die quantitativen Ergebnisse korreliert bzw. assoziiert werden sollen (etwa Erscheinungsjahre, Autoren,

---

die oben gegebene abstrakte Beschreibung dessen, was geschieht, wenn Befunde ‚interpretiert' werden.

oder andere, extratextuelle Kontexte)[20]. Einfache Visualisierungen sind ohne tiefergehende Kenntnisse in statistischen Verfahren interpretierbar. Weitergehende Auswertungen (etwa zur statistischen Signifikanz der Funde) setzen dagegen statistische Kenntnisse voraus.[21] Für die praktische Bereitstellung dieser Analysen gibt es verschiedene Möglichkeiten, die je nach Personal- und Projektsituation unterschiedlich geeignet sind. So kann z. B. auf korpuslinguistische Tools wie CQPweb[22] oder AntConc[23] zurückgegriffen werden, die solche Analysen bereitstellen können. Eine andere Möglichkeit besteht darin, die quantitativen Ergebnisse als CSV-Dateien zu exportieren und die weitere Analyse mit einer Tabellenkalkulation vorzunehmen (z. B. OpenOffice oder Microsoft Excel).[24] Nicht zuletzt können auch spezifische graphische Benutzeroberflächen entwickelt werden.

In den meisten Fällen sind gerade die Ausreißer interessant, also Instanzen die außerhalb des ‚Erwartbaren' liegen. Sie sind in jedem Fall manuell zu untersuchen, allerdings kann auch eine stichprobenartige Inspektion beliebiger Datenpunkte lohnenswert sein. Das Aufspüren ‚interessanter' Einzelstellen ist in vielen Fällen keine ganz einfache Aufgabe und kann als neues Teilphänomen betrachtet werden, dessen Untersuchung einen neuen Strang reflektierter Textanalyse eröffnen würde.

Die kombinierte Interpretation von quantitativen Ergebnissen und einer Analyse von Einzelstellen ist nicht trivial. Eine der größten Gefahren hierbei ist, dass Ergebnisse überinterpretiert werden, was hauptsächlich in *confirmation* und *selection bias* münden kann. Ein *selection bias* kann etwa auftreten, wenn die quantitativen Ergebnisse, die notwendigerweise auf einem endlichen Datensatz beruhen, als allgemein gültig angesehen werden. Dabei wird ausgeblendet, dass die Auswahl von Textkorpora einer Vielzahl von Selektionsmechanismen unterliegt, von denen nur wenige explizit markiert sind. Bei literaturwissenschaftlichen Fragestellungen, die über eine klar definierte Menge an Texten hinausgehen (etwa die Texte eines Autors), stellt sich so häufig die Frage der Kanonisierung:

20 Mit dem Begriff ‚Korrelation' wird ein Verhältnis zwischen zwei numerischen Variablen beschrieben, also hier z. B. Jahreszahl und Anzahl der Entitätenreferenzen. ‚Assoziation' beschreibt ein Verhältnis, bei dem mindestens eine der Variablen kategoriellen Charakter hat, also etwa den Zusammenhang zwischen Genre und Anzahl der Entitätenreferenzen (cf. Dormann 2017, Kapitel 5).

21 Als Einstieg in die Lektüre seien hier Bortz und Schuster (1977), Dormann (2017) und Gries (2013) genannt.

22 http://cwb.sourceforge.net/cqpweb.php

23 https://www.laurenceanthony.net/software/antconc/

24 Im Projekt QuaDramA wird z. B. die Software RStudio verwendet, und alle Mitarbeitenden in der Programmiersprache R geschult, womit eine gewisse Art der Interaktivität erreicht werden kann (cf. Reiter et al. 2017).

Aus einem literaturwissenschaftlichen Kanon kann eben nur unter bestimmten zusätzlichen Annahmen auf allgemeine literarische Phänomene geschlossen werden. Ein *confirmation bias*, also die Neigung, dass wir vor allem wahrnehmen, was bestehende Haltungen bestätigt, könnte dagegen vorliegen, wenn aus einer größeren Menge an quantitativen Ergebnissen diejenigen herausgepickt werden, die vorher bekanntes Wissen stützten, während andere, bewusst oder unbewusst, ignoriert werden. Quantitative Ergebnisse müssen stets in ihrer Gesamtheit rezipiert und interpretiert werden.

Die Vermeidung dieser Gefahren führt in vielen Fällen dazu, dass die Ergebnisse nicht eindeutig sind. Regelmäßig ergeben quantitative Untersuchungen ein gemischtes Bild, das eher neue Fragen aufwirft als sie abschließend klärt. Dies liegt auch daran, das Mechanismen aus der Statistik, etwa Signifikanztests, für nicht-repräsentative Daten nur beschränkt anwendbar sind. ‚Repräsentativität' lässt sich für historische Daten generell nur schwer festlegen oder erreichen, und auch kulturelle Daten stellen hier Herausforderungen dar.

Für die Interpretation der Befunde kann es daher kein Patentrezept geben. Sie muss, darin liegt ein wesentlicher Teil reflektierter Textanalyse, im Bewusstsein der vorherigen Schritte und der sich daraus ergebenden Einschränkungen erfolgen. Gleichzeitig muss existierendes Wissen über den Gegenstand, das in den jeweiligen Disziplinen besteht, Berücksichtigung finden. Die Resultate der Interpretation können schließlich – je nach der globalen Zielsetzung des Forschungsprojektes – für sich stehen oder als Antwort auf eine Teilfrage einer umfangreicheren innerdisziplinären Fragestellung mit dieser zusammengeführt werden.

8 Fazit

Wie der dargestellte Arbeitsablauf zeigt,[25] ist das, was wir unter *reflektierter Textanalyse* verstehen keine Theorie, sondern eine Praxis, die sich durch drei Aspekte kennzeichnet: (i) Die Aufteilung einer Fragestellung in Teilfragen berücksichtigt stets die Operationalisierung bzw. Operationalisierbarkeit der für sie relevanten

25 Die Vorstellung eines solchen *workflows* wirft die Frage nach technischer Unterstützung in Form von Tools und Anwendungen auf. In der Tat wäre es wünschenswert, wenn ein Werkzeug den vollständigen Arbeitsablauf unterstützen würde, so dass sich Forschende auf die ‚interessanten' Fragen konzentrieren können. Eine Herausforderung dabei ist allerdings, dass der beschriebene Arbeitsablauf generisch angelegt ist und dementsprechend flexibel umgesetzt werden kann. Welche Arten von z. B. technischer Modellierung im Einzelnen Verwendung finden muss flexibel bleiben, um bestmögliche Ergebnisse zu erzielen. Daher ist es in der Praxis meist fruchtbarer sich bei solchen Arbeitsabläufen einer Vielzahl von Einzelwerkzeugen zu bedienen.

Begriffe. Dadurch ergibt sich ein gegenseitiges Bedingungsverhältnis zwischen der Aufteilung in Teilfragen und der Operationalisierung besagter Begriffe. (ii) Bei der empirisch fundierten Validierung der Operationalisierung(en) und der Interpretation der Befunde werden explizite und implizite sowie theoretische und praktische Vorannahmen berücksichtigt. (iii) Alle Arbeitsschritte werden vor dem Hintergrund von disziplinärem Domänenwissen vorgenommen, das ebenfalls kritisch – in Hinblick auf die Analysepraxis und -resultate – reflektiert werden muss.

Reflektierte algorithmische Textanalyse bezeichnet Praktiken der computergestützten Textanalyse, die sich durch ihre interdisziplinär verzahnte Modularisierung kennzeichnen. Bei diesen Modulen handelt es sich um miteinander verknüpfte, manuelle und automatische Arbeitsschritte, die sich auf Begriffe oder Textphänomene beziehen. Die Aufteilung der Module sowie die Interpretation von deren Ergebnissen erfolgt unter Berücksichtigung des gegenstandsbezogenen Vorwissens, der Operationalisierbarkeit der Module sowie deren empirischer Validierung.

Literatur

Bortz, Jürgen und Christof Schuster (1977). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.

Braun, Manuel und Nora Ketschik (2019). „Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen". In: *Das Mittelalter* 21.1, S. 54–70. DOI: 10.1515/mial-2019-0005.

Bridgman, Percy Williams (1927). *The Logic of Modern Physics*. New York: The Macmillan Company.

Bridgman, Percy Williams (1938). „Operational Analysis". In: *Philosophy of Science* 5.2, S. 114–131. DOI: 10.1086/286496.

Chang, Hasok (2019). „Operationalism". In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2019/entries/operationalism/ (besucht am 1. Juni 2020).

Danneberg, Lutz (2000). „Artikel Kontext". In: *Reallexikon der deutschen Literaturwissenschaft*. Hrsg. von Harald Fricke. Bd. 2. Berlin, New York: De Gruyter, S. 333–337.

Descher, Stefan, Jan Borkowski, Felicitis Ferder und Philipp David Heine (2015). „Probleme der Interpretation von Literatur. Ein Überblick". In: *Literatur interpretieren. Interdisziplinäre Beiträge zur Theorie und Praxis*. Hrsg. von Stefan Descher, Jan Borkowski, Felicitis Ferder und Philipp David Heine. Münster, Germany: mentis, S. 11–70.

Dormann, Carsten (2017). *Parametrische Statistik*. 2. Aufl. Statistik und ihre Anwendungen. Berlin/Heidelberg: Springer.

Fantl, Jeremy (2017). „Knowledge How". In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2017/entries/knowledge-how/ (besucht am 1. Juni 2020).

Gerstorfer, Dominik (2020). „Entdecken und Rechtfertigen in den Digital Humanities“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 107–123.

Gries, Stefan (2013). *Statistics for Linguistics with R*. 2. Aufl. Berlin und Boston: De Gruyter.

Hempel, Carl Gustav und Paul Oppenheim (1948). „Studies in the Logic of Explanation“. In: *Philosophy of Science* 15.2, S. 135–175. DOI: 10.1086/286983.

Jannidis, Fotis (2017). „Grundlagen der Datenmodellierung“. In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 99–108.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Kuhn, Jonas (2020a). „Computational Text Analysis within the Humanities“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 63–106.

Kuhn, Jonas (2020b). „Einleitung“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 9–40.

Manning, Christopher D. und Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts und London, England: MIT Press.

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.

Pichler, Axel, André Blessing, Nils Reiter und Mirco Schönfeld (2020). „Algorithmische Mikrolektüren philosophischer Texte“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 328–372.

Reichenbach, Hans (1928). *Philosophie der Raum-Zeit-Lehre*. Berlin, Leipzig: Walter de Gruyter.

Reiter, Nils (2014). „Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms“. Diss. Heidelberg University. DOI: 10.11588/heidok.00017042.

Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.

Reiter, Nils, Jonas Kuhn und Marcus Willand (2017). „To GUI or not to GUI?“ In: *INFORMATIK 2017*. Bd. 275. Lecture Notes in Informatics (LNI). Chemnitz, Germany: Gesellschaft für Informatik e.V., S. 1179–1184.

Rheinberger, Hans J. (2007). *Historische Epistemologie zur Einführung*. Hamburg: junius.

Spoerhase, Carlos (2015). „Gegen Denken? Über die Praxis der Philologie“. In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89, S. 637–646.

Jonas Kuhn

Computational Text Analysis within the Humanities

How to combine working practices from the contributing fields?

Abstract: This position paper is based on a keynote presentation at the COLING 2016 Workshop on Language Technology for Digital Humanities (LT4DH) in Osaka, Japan. It departs from observations about working practices in Humanities disciplines following a hermeneutic tradition of text interpretation vs. the method-oriented research strategies in Computational Linguistics (CL). The respective praxeological traditions are quite different. Yet more and more researchers are willing to open up towards truly transdisciplinary collaborations, trying to exploit advanced methods from CL within research that ultimately addresses questions from the traditional humanities disciplines and the social sciences.

The article identifies two central workflow-related issues for this type of collaborative project in the Digital Humanities (DH) and Computational Social Science: (i) a *scheduling dilemma*, which affects the point in the course of the project when specifications of the core analysis task are fixed (as early as possible from the computational perspective, but as late as possible from the Humanities perspective) and (ii) the *subjectivity problem*, which concerns the degree of intersubjective stability of the target categories of analysis. CL methodology demands high inter-annotator agreement and theory-independent categories, while the categories in hermeneutic reasoning are often tied to a particular interpretive approach (viz. a *theory of literary interpretation*) and may bear a non-trivial relation to a reader's pre-understanding.

Building a comprehensive methodological framework that helps overcome these issues requires considerable time and patience. The established computational methodology has to be gradually opened up to more hermeneutically oriented research questions; resources and tools for the relevant categories of analysis have to be constructed. This article does not call into question that well-targeted efforts along this path are worthwhile. Yet, it makes the following additional programmatic point regarding directions for future research: It might be fruitful to explore – in parallel – the potential lying in DH-specific variants of the concept of rapid

Note: This article is a slightly revised version of: Jonas Kuhn (2019). "Computational text analysis within the Humanities: How to combine working practices from the contributing fields?" In: *Language Resources & Evaluation* 53, pp. 565–602. DOI: 10.1007/s10579-019-09459-3.

Jonas Kuhn, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
https://doi.org/10.1515/9783110693973-004

prototyping from Software Engineering. To get an idea of how computational analysis of some aspect of text might contribute to a hermeneutic research question, a prototypical analysis model is constructed, e. g. from related data collections and analysis categories, using transfer techniques. While the initial quality of analysis may be limited, the idea of *rapid probing* allows scholars to explore how the analysis fits in an actual workflow on the target text data and thus provide early feedback for the process of refining the modeling. If the rapid probing method can indeed be incorporated in a hermeneutic framework to the satisfaction of well-disposed Humanities scholars, a swifter exploration of alternative paths of analysis would become possible. This may generate considerable additional momentum for transdisciplinary integration. It is as yet too early to point to truly Humanities-oriented examples of the proposed rapid probing technique. To nevertheless make the programmatic idea more concrete, the article uses two experimental scenarios to argue how rapid probing might help addressing the scheduling dilemma and the subjectivity problem respectively. The first scenario illustrates the transfer of complex analysis pipelines across corpora; the second one addresses rapid annotation experiments targeting character mentions in literary text.

Zusammenfassung: Dieses Positionspapier betrachtet zwei Problemstellungen, die sich für den Workflow in interdisziplinären Kooperationsprojekten der Digital Humanities (DH) und komputationellen Sozialwissenschaften ergeben: (i) ein Terminierungsdilemma (*scheduling dilemma*), das den Zeitpunkt im Projektverlauf betrifft, zu dem die präzise Spezifikation der zentralen Analyseaufgabe erfolgen sollte (aus komputationeller Perspektive so früh, aus geisteswissenschaftlicher Perspektive hingegen so spät wie möglich) und (ii) das Subjektivitätsproblem, das den Grad der erzielbaren intersubjektiven Übereinstimmung zu Zielkategorien der Analyse betrifft (die computerlinguistische Methodenkonvention fordert beispielsweise hohes *inter-annotator agreement* und möglichst theorieunabhängige Kategorien, während die Analysekategorien eines geisteswissenschaftlichen Zugangs zum Textverständnis meist an eine spezielle Interpretationstheorie gebunden sind, spezifische Kontextfaktoren berücksichtigen und das Vorverständnis der Leserin oder des Lesers einbeziehen).
Mittel- bis langfristig muss die Methodik der rechnergestützten Disziplinen umfassend für geisteswissenschaftliche Fragenkomplexe geöffnet werden, und dazu müssen Ressourcen und Werkzeuge für die relevanten Kategorien sorgfältig konzipiert und aufgebaut werden. Dieser Aufsatz regt zu zusätzlichen Überlegungen an: Inwieweit ließe sich bereits kurzfristiger das Konzept des *Rapid Prototyping* aus dem Software Engineering auf eine DH-spezifische Weise ausgenutzen? Um eine Vorstellung davon zu bekommen, ob und wie ein komputationeller Textana-

lyseansatz zu einer geisteswissenschaftlichen Forschungsfrage beitragen *könnte*, wird rasch ein prototypisches Analysemodell erstellt – z. B. auf verwandten Textsammlungen und mit Hilfe von Transfertechniken. So kann frühzeitiges Feedback in den Optimierungsprozess der Computermodelle eingespeist werden, und zu konkurrierenden Analysestrategien kann eine (zumindest partiell) informierte Entscheidung getroffen werden. Die Idee wird anhand von zwei experimentellen Szenarien illustriert: (1) der korpusüberspannenden Übertragung komplexer Analyse-Pipelines für eine maßgeschneiderte Informationsextraktion und (2) prototypische Textannotation zum Ausloten von Hypothesen zur Erzählperspektive in Texten eines spezifischen Autors.

1 Introduction

1.1 Preliminaries

Many years of research and tool development in the fields of Natural Languages Processing (NLP) and Computational Linguistics (CL) have led to (1) the availability of numerous mature tools for text analysis in the major languages, such as lemmatizers, part-of-speech taggers, parsers, etc., but also tools for specific tasks beyond linguistic annotation such as sentiment analysis, translation, purpose-specific information extraction, etc. Alongside the technical machinery (2) an advanced methodology has been developed, defining appropriate workflows for training, adapting, evaluating, and employing new analysis components where off-the-shelf tools are not readily available–because the text corpus under consideration diverges from the development standard (earlier language stage, special text genre or content domain, under-resourced language, etc.), and/or because the analytical task involves steps not covered so far (e. g., the identification of passages of scenic narration in novels and other narrative texts).

Both (1) the use of existing tools and (2) the adaptation/augmentation of analysis systems is supported by resource infrastructures such as CLARIN[1] (providing access to interoperable tools and to corpora, e. g., for training data), and by publicly available code libraries. In principle, it now takes manageable effort to build or adapt text analysis systems for arbitrary combinations of text corpora and research questions (as is demonstrated by the rapid developments in Natural Language Processing over the past 5–10 years, particularly accelerated by recent

1 Common Language Resources and Technology Infrastructure, see http://www.clarin.eu, and http://www.clarin-d.net for the German partner within the European initiative.

successes in the application of artificial neural net models–"Deep Learning"). A number of recent contributions show that CL techniques can be expanded to construct analysis systems for literary texts. They for instance support the extraction of Social networks among literary characters (Elson et al. 2010), an analysis of the text-internal dynamics of inter-character relationships (Chaturvedi et al. 2016; Iyyer et al. 2016) or aspects of plot structure (Goyal et al. 2010), they induce types of characters from large text collections (e. g., Bamman et al. 2014) or help understand the stylistic characterization of certain character types (Brooke et al. 2017). Over the past few years, small communities of researchers pushing targeted computational modeling techniques have evolved in several field-specific branches of DH.[2] Yet, computational modeling components of this kind[3] are still rarely used within the core areas of the classical Humanities disciplines like Literary Studies or History, which generally take a hermeneutic approach[4] to text interpretation and, moreover, textual criticism, which is aimed at the *significance* of a text–following Hirsch's 1967 separation of text meaning and significance, where the latter comprises the "relationship between [the text] meaning and a person, or a conception, or a situation, or indeed anything imaginable" (Hirsch 1967, p. 8).[5] Under a hermeneutic approach, a literary scholar may for instance try to under-

2 For Computational History, see for instance the contributions in Bozic et al. 2016; for Computational Literary Studies, many researchers came together for the 2017 International DFG Symposion in Literary Studies at Villa Vigoni, Italy 9–13 October 2017, which was dedicated to Digital Literary Studies. A volume of the contributions is in preparation (Jannidis in preparation).
The development of the subfield of Computational Literary Studies in Germany is quite dynamic. In 2018, the Deutsche Forschungsgemeinschaft (DFG), established a priority program *Computational Literary Studies* (SPP 2207) with a runtime of 6 years.

3 This is not to say that digital editions and computational support for corpus search and exploration have not been widely adopted. Indeed, fields like Literary Studies have changed with the omnipresence of computational tools and resources, hence the observation of a "computational turn" (Berry 2011). Most influential have arguably been the contributions by Moretti (2007) and Jockers (2013). The emphasis in this article is however on the integration of formal/algorithmic models in the core argumentation of scholarly research in the Humanities disciplines, which up until today are still in its early stages.

4 Our focus is not on Philosophical Hermeneutics (associated with Friedrich Schleiermacher, Wilhelm Dilthey, Hans-Georg Gadamer a.o.), but on hermeneutics as a broadly adopted practical method in Literary Studies and other disciplines targeting the interpretation of works of art. As Newton (1989) points out, the term, which has long been central in the German tradition, gained prominence in the Anglo-American tradition with the work by Hirsch (1967). "The central concern of hermeneutics is the problem created by the fact that texts written in the past continue to exist while their authors and the historical context which produced them pass away in time. Reading such texts therefore becomes inseparable from the question of interpretation." (Newton 1989, p. 116)

5 See also the discussion of the aims of text interpretation in Mantzavinos 2016.

stand the significance of a group of novels from a particular epoch and cultural background against its historical context, possibly taking into account the sociological situation at the time etc. There may thus be multiple valid interpretations (at the level of its significance) for the same text. Often, such "polyvalence" is seen as a constitutive property of literary texts, and it indicates that the standard CL methodology of corpus annotation and computational modeling, which aims at determining a single, intersubjectively stable target, cannot be applied at this level (we will come back to this in Sect. 5).[6]

Despite the special character of text interpretation as the final objective of hermeneutic research, much of the evidence that the scholar can build on is available in the form of preserved texts and other sources, so–from the point of view of Computational Linguists–the use of advanced corpus-based methods as a means to ensure systematicity in the process would seem an evident choice nonetheless (of course, the scholar's pre-understanding would have to be reflected in the formulation of the corpus analysis task). On the other hand, the relevant analytical

6 An anonymous reviewer questions that there is any systematic difference between the computational modeling situation (1) for a linguistic analysis task (such as part-of-speech tagging) and (2) in the context of a hermeneutic approach to some object of study in the Humanities. He or she points out that the need to place a modeling approach within a space of competing interpretive frameworks, whose categories are mutually incompatible, is familiar from linguistic tasks: competing linguistic theories go along with different conceptual frameworks that cannot be easily reconciled in corpus-oriented operationalization. While I agree that the modeling challenges may not be categorically different between a linguistic context and for an embedding in a hermeneutic context, the two different cultural/methodological umbrellas *do* go along with substantially different weight that needs to be placed on establishing a framework of analytical concepts that (a) support an intersubjectively agreed-upon operationalization (e. g., for categories of text analysis) and at the same time (b) allow researchers to put hypotheses to a test that have been considered relevant in the recent literature. In the Humanities, the distinction between the *meaning* of a text versus its *significance* relative to various possible conceptions and situations (Hirsch 1967) discussed above goes along with a much more diverse landscape of distinct frameworks of reference, and even when text analysis is limited to the simpler notion of literal text meaning, pre-existing theoretical frameworks with a similar degree of formal specification as is common in Linguistics are rarely found in the Humanities. At the risk of oversimplifying things, one may ascribe this to modern Linguistics following a methodological agenda modeled after the natural sciences and generally working in relatively settled paradigms of "normal science" (in terms of T. Kuhn's 1962 terminology), whereas much of the research in the Humanities does not aspire to forming a paradigm with a disciplinary consensus; the work practices there rather follow the patterns that T. Kuhn uses to characterize the "pre-paradigm" period.
Of course, it is not impossible to approach literary texts and historical source collections with an approach following the methodology established in Linguistics; but this is a different thing from trying to build on top of insights from (many distinct existing) hermeneutic branches in the established Humanities disciplines. And here, considerable cultural differences cannot be denied.

questions are likely to be different from study to study. Moreover, computational tools for the questions will rarely be readily available (few questions being directly correlated with the linguistic form of the text). Hence, it is not surprising when a specialist scholar keeps relying on their erudition and manual analysis rather than investing time into the development or refinement of analytical tools that may have just a one-time application.[7]

Given this understandably conservative tendency in the core Humanities disciplines, emerging DH fields such as Digital Literary Studies, tend to focus on questions that have not been at the center of traditional research (e. g., stylometric research[8] or corpus-oriented research on the historical development of key genres[9]) rather than trying to augment the methodological spectrum for addressing classical key questions of text interpretation. To sum up, from the CL perspective it seems that the potential of computational models for Humanities research is currently underexploited. This article, which expands on a keynote presentation at the COLING 2016 Workshop on Language Technology for Digital Humanities (LT4DH) in Osaka, Japan,[10] contributes some more detailed considerations of how this status quo can be explained and whether it could (and should) be changed. The basis for characterizing the status quo as underexploiting a potential are mostly the author's personal exchanges on many occasions–with scholars from various disciplines in the Humanities and Social Sciences. Numerous studies could in principle benefit from computational corpus analysis targeting special, non-trivial analytical categories, but given the considerable development effort and unclear chances of success (in terms of supporting innovative conclusions), it often seems wiser to follow simpler approaches.

For thoughts about potential paths along which the situation might be changed, the LT4DH keynote and this article rely mainly on experiences from collaborative DH projects involving the author himself, as well as from collaborative research involving CL and Linguistics.[11] This is because the emphasis here is not on research results as typically reported in publications, but more on obser-

7 Below we will also come to the problem that even the best conceivable computational tools will display a level of accuracy that will make their use unattractive to scholars. And with tentatively small datasets underlying Humanities study, quality issues are amplified.

8 Viz. the success story of the metric of Burrows's Delta (Burrows 2002), e. g., applied to German literary history by Jannidis and Lauer (2014), and scrutinized methodologically by Evert et al. (2015).

9 See e. g., Underwood 2015, but also earlier work such as Biber and Finegan 1989.

10 The title of the keynote presentation was *Flexible and Reliable Text Analytics in the Digital Humanities–Some Methodological Considerations*.

11 The author's recent DH projects are the following: CLARIN-D (common language resource infrastructure for the Digital Humanities, 2011–2020, in which Stuttgart has been a center within

vations about practices, gathered along the way of collaborative project work and in dedicated methodological explorations. This view makes this article a fairly subjective contribution which cannot claim to describe the status quo in a systematic way. Nor is there a claim of exclusivity. But hopefully, this contribution will stimulate further methodological discussions and developments in an exciting interdisciplinary and transdisciplinary area.

1.2 What this article aims to achieve

This article asks what are the reasons for the observed underexploitation of advanced models from DH and CL/NLP in the Humanities and Social Sciences. A brief explanation could be that there is simply no interest in computational methods within the core disciplines–but this is clearly not the case; the fields have always been eager to adopt new approaches and in Literary Studies, for instance, the "computational turn" (Berry 2011) is considered to be in full swing. Contributions in Literary Studies make use of visualization techniques, network analysis and other methods (however not taking advantage of the full spectrum of modeling options as the computational linguist would see it).

A different explanation might be that the respective methodological prerequisites are too far apart to be reconciled. There is probably a lot to this explanation, and throughout the years there have been numerous blogs in DH forums, discus-

---

the German network, led by the University of Tübingen, Erhard Hinrichs (the Stuttgart center has been co-funded by the German Federal Ministry of Education and Research BMBF and the state ministry MWK in Baden-Württemberg); the eHumanities project "e-Identity" (2012–2015, funded by the Federal Ministry BMBF, with leading PI Cathleen Kantner from Political Science and additional Co-PI's Ulrich Heid and Manfred Stede); the eHumanities project "ePoetics" (2013–2016, funded by the Federal Ministry BMBF, with leading PI Sandra Richter from Literary Studies and additional Co-PI's Thomas Ertl and Andrea Rapp); the Digital Humanities "Center for Reflected Text Analytics" CRETA (2016–2020, with a group of 10 PI's from various disciplines in the Humanities and Social Sciences, directed by the author and funded by Federal Ministry BMBF); the "DebateExplorer" project (2016–2017, a data-driven journalism project funded by Volkswagen Foundation, with Co-PI Eva Wolfangel, a journalist); the "RePlay-DH" (2016–2019, on research data management for the Digital Humanities, funded by the state ministry MWK in Baden-Württemberg, with Co-PIs Helge Steenweg and Stefan Wesner).

Collaborative project experience regarding links between CL and Linguistics include for instance involvement 2007–2010 in SFB 632 *Information structure: The linguistic means for structuring utterances, sentences and texts* (University of Potsdam and Humboldt University Berlin, funded by Deutsche Forschungsgemeinschaft, DFG) and 2010–2018 in the DFG-funded SFB 732 *Incremental Specification in Context* (University of Stuttgart, in which the author led several subprojects and was deputy director 2012–2015 and director 2015–2018).

sion panels and position papers observing the two cultures problem as a major obstacle.[12] But granted the cultural divide, it is surprising that is so hard to overcome this obstacle that after many years, there is still no best-practice recipe for teams of interdisciplinary collaborators to follow. Could it be that what Computational Linguistics has to offer in terms of deeper analytical means is generally insufficient to be integrated into hermeneutically oriented research? An anonymous reviewer expresses the suspicion that Humanities scholars are unlikely to ever accept error rates with automatic analysis tools that are significantly above human inter-annotator discrepancies. Indeed, it seems plausible that scholars would not want to move themselves into a worse starting position than when relying on "close-reading type" manual analysis. And CL tools for analysis tasks that are more complex than part-of-speech tagging do go along with considerably higher error rates, even with contemporary newswire datasets. So how would Humanities scholars ever use automatic tools for complex tasks in the text domains of their interest? There is typically much fewer training data and hence error rates are bound to be much higher. And the closer we get to interpretive questions in a hermeneutic approach, the more extreme it appears to get. With this reasoning it seems useless to seek for collaborative workflows that help modeling deeper and deeper analysis tasks: If no tool can be expected to reach acceptable error rates, one would essentially waste time. Energy seems to be better spent on improving machine learning from small datasets.

I think one should not follow this reasoning, but rather acknowledge both as important goals: machine learning from fewer data and interdisciplinary integration of work practices. For one thing, entirely postponing the latter would imply that after successes in the former, there would still be a very long way before the Humanities can take advantage of them. But, more importantly, I believe that even the application of analysis models with comparatively high error rates could find a reasonable home in some next-generation hermeneutic approach. Imagine for instance a scholar working on a key text from some German nineteenth century author. She suspects that this text reflects influence from the author's reception of a contemporary French text (and she wants to use this to argue for some production aesthetic thesis, pointing out that the author uses certain text features to signal the intertextual references). There are some indications in diary entries that make the assumption plausible, but no certain evidence. Now, one might envisage training a computational model for intertextual links on known cases of

12 The discussion of the status of different methodologies contributing to DH is often linked with the question of definition (see e. g., the contributions in Gold and Klein 2016). Regarding the issues stemming from the combination of different cultural traditions, the discussion in Hammond et al. 2013 seems to me to be very much to the point.

text pairs. On indirect links, this model will have a fairly high error rate, but if it corroborates the scholar's suspicion by predicting several passages to be likely intertextual links, it does provide a valuable additional indication for argumentation (essentially following an abductive reasoning pattern). As a matter of fact, in any historic context scholars are very much used to dealing with combinations of sources with variable reliability.[13] An effective strategy for minimizing the risk of incorrect inferences drawn from imperfect analysis components could use several strands of analysis in parallel– chosen so the error sources are likely to be independent of each other. As a result, cases of mutual agreement are very unlikely to be analytical artifacts.

It may seem a little disturbing that it is only on hypothetical grounds that we can decide whether or not one should pursue a more integrative methodology. But if it is really true that for now there are major roadblocks that prevent an effective application of deeper computational models in hermeneutic research, it would come as no surprise that there are no examples yet that show an everyday use of the idea. The present article makes the observation that there are at least two workflow-related issues that hermeneutically oriented DH projects face even when they bear a real potential for exploiting the data-driven methodology from CL: (1) a *scheduling dilemma*, which affects the point in the course of the project when specifications of the core analysis task are fixed (as early as possible from the computational perspective, but as late as possible from the Humanities perspective); (2) the *subjectivity problem*, which concerns the degree of intersubjective stability of the target categories of analysis. CL methodology demands high inter-annotator agreement and theory-independent categories, while the categories in hermeneutic reasoning are often tied to a particular interpretive approach (viz. a *theory of literary interpretation*) and may bear a non-trivial relation to a reader's pre-understanding. Building a comprehensive methodological framework that helps overcome these issues requires considerable time and patience.

The established computational methodology has to be gradually opened up to more hermeneutically oriented research questions; resources and tools for the relevant categories of analysis have to be constructed. In many cases, this includes coming up with an inventory of descriptive categories appropriate for sharing across specific research frameworks. This article does not call into question that well-targeted efforts along this path are worthwhile. Yet, it makes the following additional programmatic point: It might be fruitful to explore–in parallel–the potential lying in DH-specific variants of the rapid prototyping idea from Software

13 Similar methodological considerations for Digital History are made in ter Braake et al. 2016.

Engineering. If a method of *rapid probing of analysis models* can be incorporated in a hermeneutic framework to the satisfaction of well-disposed Humanities scholars, a swifter exploration of alternative paths of analysis would become possible. This may generate considerable additional momentum for transdisciplinary integration.

It is as yet too early to point to truly Humanities-oriented examples of the proposed rapid probing technique. To nevertheless make the programmatic idea more concrete, the article uses two experimental scenarios to argue how rapid probing might help addressing the scheduling dilemma and the subjectivity problem respectively. The first scenario illustrates the transfer of complex analysis pipelines across corpora; the second one addresses rapid annotation experiments targeting character mentions in literary text.

Section 2 briefly reviews the standard methodology of data-oriented model development; Sect. 3 makes some observations about the different working practices in approaches from the Humanities versus the Computational Sciences. Against this background, Sects. 4 and 5 address the scheduling dilemma and the subjectivity problem respectively, discussing ways in which they might be tackled with the idea of rapid probing. Section 6 presents a short conclusion.

2 Background

No automatic language-technological tool achieves one hundred percent correct results[14]–not even when it is applied to texts whose properties correspond exactly to the corpus used in tool development. And as soon as the application context deviates from the development scenario (be it due to differences in historical language stage, register, text genre, or content domain), the error rate will increase–possibly to a considerable degree, depending on circumstances (cp. Sekine 1997). By chaining up several analysis steps in a pipeline, in which each component re-

14 It has to be noted of course that a perfect match is not even achieved (in non-trivial analysis tasks) when the performance of two human annotators is compared. Over the past years, the fields of corpus linguistics and CL have placed considerable emphasis on the development of high methodological standards for corpus annotation work. This has resulted in various strategies for dealing with difficult annotation decisions; across the board, a clear operationalization of the decision criteria and an empirical evaluation of *inter-annotator agreement*, as an upper bound for the prediction quality that one might expect, is required. For some analysis tasks, such as part-of-speech tagging and parsing, highly developed automatic approaches have reached accuracy levels in testing that come close to the human consensus (see Giesbrecht and Evert 2009; Manning 2011).

ceives as input the output of another automatic analysis, the risk of error is potentiated. Computational approaches in the Digital Humanities that address "deeper" analytical questions (e. g., questions closer to text interpretation/textual criticism in Literary Studies) are likely to employ relatively long chains of analysis[15] and are thus particularly exposed to error propagation. As an additional issue when moving away from traditional research on the text material towards automated analysis of larger collections of source texts, the human view on each single document is eliminated from the process. This also eliminates a free "sanity check": traditionally, sampling errors or other issues in the source selection procedure would have been noticed as a side effect of actually looking every text and applying some manual analysis.

Hence, it is not surprising when quick attempts to apply existing analysis tools to a corpus of text material taken from some Humanities scholar's key research area lead to a certain degree of disappointment–an effect that is not uncommon in DH pilot studies: it is likely that the tools will get some obvious cases wrong (besides the unnoticed ones they get right) and whatever catches the eye as an aggregate outcome of the automated analysis will typically appear to replicate findings that are (apparently) "obvious" from scholarly research using conventional approaches. Such disappointments can have multiple reasons; an important one lies in the fact that the direct application of unmodified existing tools greatly underexploits the potential lying in the computational methodology.[16] For the remainder of this article, our focus shall be on possible adjustments and extensions of the tool-based working practices in response to higher-level research questions from the Humanities.

15 This is not necessarily the case. It may be possible to employ "end-to-end" modeling techniques, which have proven very successful in recent NLP work (see e. g., Zhou and Xu 2015)– e. g., the guiding research question rests on the emotional perception of text passages from a corpus, and explicit records of such reception may be available for a set of representative texts (say, through marginal annotation in surviving text witnesses). In this case, computational models could (in principle) be trained to assign target emotions to candidate texts, without intermediate steps of linguistic analysis, such as the identification of relevant lemmas, etc. A practical problem for application of this technique in the Digital Humanities is that for robust and reliable predictions, a relatively large set of training data is needed, which is typically unavailable for the analytical questions at hand. With small datasets, the classical pipeline approach, using intermediate steps of analysis, has the advantage that general linguistic knowledge (e. g., from lexical resources) can be taken into account to form generalization beyond the few available data points.

16 In many cases, straightforward tool application will nevertheless advance the analytical basis in corpus-oriented work and be highly beneficial (e. g., taking advantage of apparently small modifications of the view on the text, such as lemmatization or part-of-speech tagging, or through the use of unsupervised computational methods such as latent topic modeling).

2.1 The reference data-based methodology in Computational Linguistics

In data-oriented natural language processing (NLP), a standard methodology has been established that uses independently annotated "gold standard" data to avoid an overly impressionistic view of the usefulness of some analysis system for one's own purposes. Indeed, before making use of any automatic model predictions it is of key importance to question the quality of a system relative to one's corpus and the defined target analysis task: if the error rate is below a certain threshold, it may be safe to draw certain inferences despite the system being imperfect; on the other hand, if the system is unreliable for core categories of analysis, alternative approaches should be considered or problematic components should be fixed, etc.

Such an informed model application can be achieved with a conceptually simple procedure, which does however take some extra effort: whenever one plans to apply some analysis system to a new type of text data, a prior step of reference data-based quality assessment has to be performed. This means that a sample from the target-specific corpus data has to be annotated manually with the intended target analysis (where the sample of reference data is representative for the application case[17]).

Of course, this study-specific annotation step is associated with non-negligible effort: the annotation guidelines for the original task need to be adjusted, annotators have to be trained, and a sufficiently large amount of data needs to be annotated, ideally with multiple annotations per data-point, so inter-annotator agreement measures can be taken into account (see Hovy and Lavid 2010).

When transferring existing tools and tool chains to a new task and/or target corpus, it is tempting to skip the step of reference data annotation and rather do a post hoc assessment of the system output. However, it is definitely methodologically superior to adhere to prior manual annotation of test data for evaluating quality. With a post hoc assessment of system output, there is a known bias for the analysis that is presented (Fort and Sagot 2010), so not all system errors affecting *precision* may be reliably detected–i.e., data instances that have been incorrectly assigned to the target category. System errors affecting *recall*, i.e., missing instances in the system prediction are even harder to detect without prior data annotation. The relevant instances are, by their very nature, missing in the system prediction that undergoes the post hoc assessment. Relevant cases may only

17 The representativeness of test data is critical for methodological validity. Dealing with it in a DH context is far from trivial; however, the present article cannot focus on this [some considerations are made in Kuhn in preparation].

be detected by chance, in case they appear in close proximity to another data instance.

To conclude, independent corpus-based evaluation following the standard methodology is a reliable way for assessing the usefulness of one or more available systems for a task at hands and for indicating where possible adjustments are needed.

2.2 System adaptation driven by reference data: Perspectives for the Digital Humanities?

The outlined quality assessment approach is not only applied in an academic context. In language-technological practice dealing with large amounts of web data, it is also quite commonly applied–at least in a rudimentary form: when a provider of language-technological web analytics is approached by a new customer who is interested in user acceptance for their products or services as reflected in web forums (maybe restaurant reviews), the provider will assemble a corpus of customer-specific development data for the task at hand. In our example this would be a sentiment analysis task, i.e., the detection and categorization of positive or negative subjective text passages in the restaurant reviews. The provider can then optimize their system for the customer, with a clear optimization objective: precision *and* recall of the automatically retrieved web documents can be maximized–possibly with a bias for one or the other depending on the analytical goals. If the new data (the target corpus) is sufficiently similar to the development data used for the established standard systems, only minor adjustments may be needed (maybe there is a dataset of hotel reviews, which turns our to be relatively similar); otherwise, it has to be decided what components need adjustment. In extreme cases, it may be necessary to rebuild all components. (Under a supervised machine learning approach, this may mean that not just a relatively small sample of test data needs to be annotated manually, but a relatively large set of training data.)

Now, approaching the technical challenges for automatic text analysis in DH and computational Social Science, it would seem natural to apply the same procedure: Given a large collection of digitized source documents (our target corpus), a representative sample is drawn as a test corpus. For this sample, the text-analytical decisions that are supposed to feed higher-level research questions are hand-annotated, following the annotation methodology from NLP (see Hovy and Lavid 2010). The test data can then drive the further system development process,

similarly as sketched above.[18] The approach works very well in cases where (a) the target corpus is electronically available at the beginning of the project, and (b) the text analysis steps that are needed to contribute to the main research goals are known and can be related to existing NLP tasks (e. g., named entity recognition or sentiment analysis). This is for instance a realistic scenario for extensions of the established method of Content Analysis in Social Science (cp. Krippendorff 1980), which is based on manual text annotation (or "coding") and has always placed emphasis on a research design that can be broken down into operationalized analysis questions.[19] It also works for corpus-oriented text studies that build directly on surface text properties, i.e., empirical research in Theoretical Linguistics and structuralist approaches in Literary Studies.

However, for many research scenarios from the spectrum of Humanities disciplines for which one can expect benefits from the use of computational modeling approaches, the reference data-based standard methodology cannot be straightforwardly applied: the relevant input/output relations for analysis models that may be used are not known at the beginning of the project. It is in fact one of the major project tasks to determine appropriate analytical devices informing the higher-level research question. Many scholars would point out that the hermeneutic approach they follow is in opposition to a methodologically driven preconception of the overall research agenda as a structured set of sub-questions and analysis tasks.[20]

In order to better understand the implications of this circumstance for computational working practices, it is worthwhile clarifying the working assumptions and preferred practices of "classical" research in the Humanities versus the typical approach from CL. I will present a schematic sketch to this end in the following section.

18 In Kuhn and Reiter 2015, we argue for a working practice in text-analytically challenging digital Humanities projects which puts the annotation of reference data at the center. The Stuttgart DH center CRETA (Center for Reflected Text Analytics, https://www.creta.uni-stuttgart.de) implements this strategy in a transdisciplinary framework.

19 Viz. methodological developments in the e-Identity project and in CRETA (Blessing, Glaser, et al. 2014; Blessing, Kliche, et al. 2015; Overbeck 2018).

20 For instance Reichert (2014) in his preface to the anthology *Big Data* observes the common concern in data-driven research that priority is given to what is methodologically feasible on the available data; Reichert speaks of an "evidence-based concentration on what is feasible with the data" (Reichert 2014, p. 20, translation: J.K.).

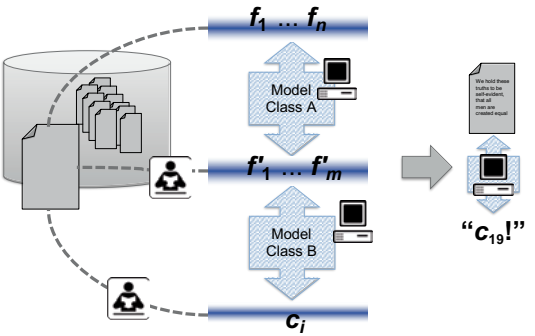


Fig. 1: Standard approach to data-driven text analysis in Computational Linguistics: (1) Left side: Texts or text elements from training data are manually annotated at levels of description that are considered relevant for analysis (possibly including intermediate levels that can serve as features for downstream pipeline components, e. g., part-of-speech tags serving as input for a syntactic parser). (2) Middle: Using appropriate model classes from machine learning and feature sets $f_1 \ldots f_n$ (extracted from the text data), model parameters are estimated based on the training corpus, for instance learning to assign category labels such as c_i to the input. Alternatively, rule-based components may implement an input-output function or part of it, which is then also evaluated against the manually annotated reference data. (3) Right side: The resulting (pipeline of) models can be applied to unseen text elements, predicting an analysis according to the learned function, i.e. assigning a label from a set of possible target categories

3 Working practices in the Humanities versus Computational Linguistics

Before going into a juxtaposition of the typical research strategies, practices and workflows in the two broad scholarly fields contributing to DH, it should be noted that generalizing across "the Humanities" is certainly problematic. There is no single methodological framework across sub-disciplines in the Humanities, and even for each specific discipline, such as Literary Studies, there is a pluralism of research approaches. Yet there are commonalities in working practice clearly contrasting with the standard methodology in Computer Science and Computational Linguistics, which I here take as a basis for reflections on how insights from the two sides can best be combined.

As another proviso, note that what follows should neither be seen as a normative characterization of best practice in the disciplines, nor as an exhaustive attempt to describe working practices. It simply serves to bring out differences across fields in the *typical* approach to breaking down one's research ideas into an agenda–this may not do full justice to the approaches, leaving commonalities

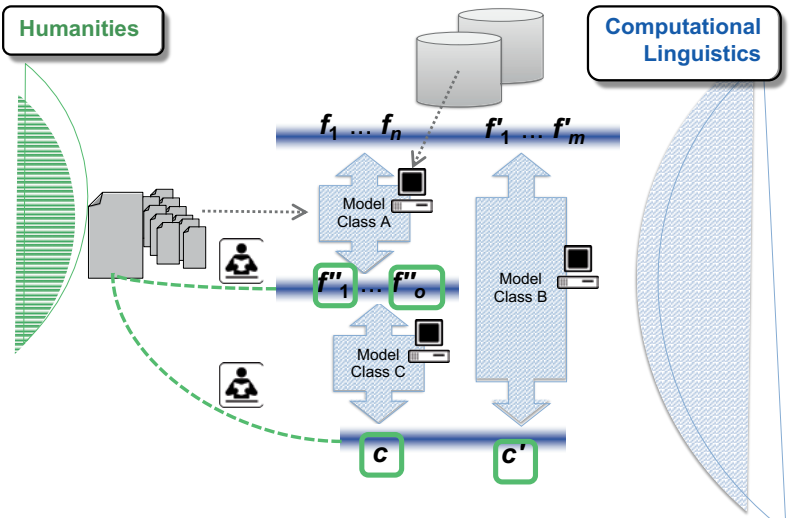


Fig. 2: Natural set-up for collaborative research in DH as seen from the CL angle: Humanities scholars suggest relevant corpora, help in the identification of relevant levels and categories of analysis and perform manual annotation of a subsample of the corpus which acts as reference data; computational linguists do machine learning experiments with candidate model classes, including additional tool or data resources where appropriate (e. g., additional training data that are sufficiently similar and can be included using model transfer techniques); the reference data annotated by Humanities scholars are used for the target of optimization

across working practices aside. But for the purpose of identifying common roadblocks in transdisciplinary approaches, this should be acceptable.

Figure 1 provides a schematic characterization of the core process of developing analysis models in modern data-oriented Computational Linguistics, here showing the decomposition of some target analysis function into two submodules. As noted in Sect. 2, annotated reference data play a key role in driving the project agenda.

The computational standard approach provides clear interfaces for the integration of expert knowledge about the data under consideration: gold standard annotations of the input/output relation in key modular components can be devised in close collaboration with the "domain experts"; components for which a given discipline has strong theoretical accounts can even be modeled as a rule-based system (or as a hybrid rule-based/statistical model). Consequently, the picture that computer scientists view as a natural and fruitful collaboration scheme is sketched in Fig. 2: In exchanges with the domain experts (in the DH scenario, Humanities scholars) at the beginning of a project, the requirements for text analysis components are established, and subsequently the experts develop annotation

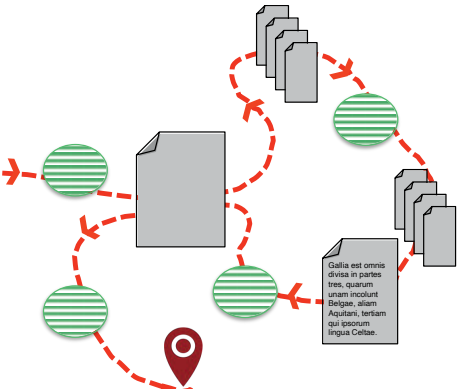


Fig. 3: Schematic depiction of characteristics in the hermeneutic research process in the Humanities (small hatched ovals symbolize theses for which the scholar has gathered argumentative support): starting out with some pre-understanding, informed a.o. by a particular literary theory of interpretation that the scholar adopts, she/he approaches the central object of study (one text or a relatively small group of texts), identifying the need for additional research into other relevant texts (possibly an established canon). This process draws attention to a further group of texts, which is next taken into consideration; this again prompts interpretive work on one particular other text, etc. Insights gathered along the way lead to a revision of the pre-understanding and ultimately the proposal of a (novel) literary interpretation of the object of study

guidelines and supervise an annotation process that leads to a reliable gold standard, capturing the targeted input/output relation for computational analysis in a precise, empirically grounded way.

On this basis, the computational linguists can experiment with different algorithmic modeling approaches and optimize model parameters, so the computational system they "deliver" at the end of this process achieves the best possible quality (measured through gold standard evaluation, including the application of tests to estimate statistical significance). Of course, the development can be implemented as a cyclic process of (a) specification, (b) preliminary development and (c) expert testing to obtain more informed specifications over time, but early architectural design decisions will always carry a major importance.

Let us now move to common work practices in Humanities disciplines. Figure 3 tries to provide a schematic picture of a typical research process in disciplines following a broadly hermeneutic approach. It essentially "rolls out" the familiar concept of Friedrich Schleiermacher's hermeneutic circle across a map of the terrain suggesting the textual material that is being considered in the evolutionary research process. Since contrary to the situation in CL, the formal shape

of the project outcome (such as an implemented input/output function) is not known at the outset, the agenda is less pre-structured.

The meandering dashed red line suggests a (consciously) open course of development of the research process.[21] The evolution is driven by a combination of pre-understandings and novel insights obtained from approaching the text under consideration–typically on the basis of some particular literary theory of interpretation and taking into account the relevant context. Thus, the process may lead to a cyclic revision of scholarly understanding of the text at hand (incorporating an application of the hermeneutic circle).

Based on this general conception, the natural way to integrate results from computational analysis components in the research process is seen in Fig. 4 (showing three distinct research contexts at the same time, where each targets a particular object of study). Where appropriate tools and models are available, computational analysis steps contributing insights about certain aspects of texts or text corpora can be integrated quite easily in the hermeneutic process (as suggested by the employment of blue-hatched input-output devices available from a research infrastructure or trained on available data).

Both Figs. 2 and 4 present straightforward extensions of the respective disciplinary self understanding, and at first glance, each extension seems to capture the requirements of a DH project exhaustively while providing a natural role for the respective partner discipline. When we compare the two resulting pictures, it becomes clear however that the two types of envisaged collaborative DH project look very different. So neither Fig. 2 nor Fig. 4 can fully meet the expectations of the respective partner discipline.

The problem of the view favored from the computational angle in Fig. 2 is the following: when it is applied in a typical project cycle comprising two or three years of funding, there is a danger that the tools and models developed will not meet any real analytical need from a Humanities context: in order to allow for a thorough model development, the corpus design, specification of analytical categories and reference data annotation has to happen at an early stage. With many design decisions, it will be hard to revise them later on when the hermeneutic process approaching specific questions has revealed different analytical interests (in terms of corpus choice or analytical task). The fact that decisions on analytical targets have to happen early makes it natural to focus on relatively generic tasks and stay with readily available, well-studied corpora. This could again re-

21 Some DH scholars point out the important role that the uncovering of seemingly random connections (through "Serendipity") can play in the hermeneutic process (see e. g., Quan-Haase and Martin 2012).

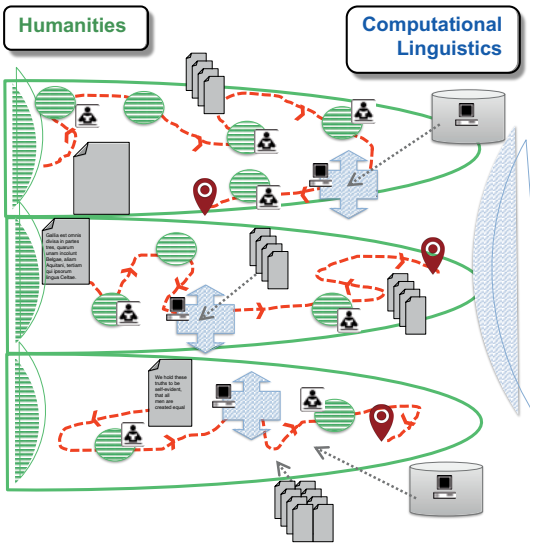


Fig. 4: Natural set-up for collaborative research in DH as seen from the Humanities angle (showing three distinct projects at the same time, each enclosed in a large partial oval): as Humanities scholars progress in their hermeneutic research process, they formulate hypotheses about a text or text corpus, which can be addressed through recourse to computational tools or models (e. g., using corpus collocation statistics to establish whether or not a key term in the text under consideration patterns with a collection of candidate texts or a background corpus). The analysis results are then incorporated in the overall argumentation and may trigger classical "close reading" steps or further steps involving computer models. Depending on the nature of the analytical step, available tools from CL or customized model solutions may be employed (drawing on additional resources); typically, even standard analysis tasks such as lemmatization and part- of-speech tagging will require tool adaptation in the DH context since the texts are not from canonical NLP domains, genres and language stages

inforce the impression among skeptics addressed at the beginning of Sect. 2 that computational models can at best replicate well-known results. Since the amount of available data for less studied targets of analysis will most likely be very small, they are less attractive for systematic model development.[22] Finally, using target categories for analysis that are dependent on specific interpretive assumptions (which can be more helpful than generic descriptive categories in the course of hermeneutic work based on the respective pre-understanding) is not something that the strictly systematic overall approach will encourage.

22 Machine learning on small datasets goes along with a high risk of overfitting, i.e., the system will memorize the observed patterns rather than pick up systematic generalizations.

When we look at the picture that seems favorable under a Humanities perspective in Fig. 4, we can make complementary observations: computational analyses are only prompted as the need arises in a hermeneutic process, hence it is desirable to allow for each course of reasoning to draw on completely different types of analysis. Also, dependence on quite specific interpretive assumptions should in principle be possible. Practically speaking however, unless the project can take indefinite time (in which case a subproject following the scheme in Fig. 2 could be triggered each time a new analysis model is required), the methodological principle of allowing appeal to some analysis procedure at any arbitrary point of the hermeneutic research process places serious limits on the depth of analysis that can be realistically performed. Entirely distinct contexts for computational analysis as suggested by the three abstract project scenarios shown in Fig. 4 will only be possible with highly generic surface-oriented tools–which means in practice that tools relying on language-specific knowledge (such as lemmatization) may already stand against methodological transfer from one scenario to the other; more corpus or task specific dimensions are even more unlikely to be sharable across the scenarios. This is not only unfortunate because it underexploits the computational potential, it also implies that critical reflection of methodological implications of computational analysis cannot build on any systematic observations across contexts of application. The latter is the basis for developing principles of 'tool criticism', as ter Braake et al. (2016) put it.

So in short, neither of the two scenarios is a satisfactory basis when trying to take full advantage of the strengths of both sides. Certain issues affect both scenarios in the same way, especially those relating to the small size of available data for the most relevant analysis target, which will lead to *overfitting* in training and *issues of limited accuracy of machine-learned tools*.[23] As suggested in Sect. 1.1 there may be ways of embedding models with limited accuracy in a multi-strand methodology relying on abductive reasoning–if the component models match the analytical requirements. So let us leave the **small data problem** aside despite its importance from the CL angle and ask whether there could be a better synthesis of the respective working practices that helps avoid the other issues listed, among which as far as I see two types of problem are very central: I will call these the ***scheduling dilemma*** and the ***subjectivity problem***. Section 4 is dedicated to the former, Sect. 5 to the latter.

23 Thanks to an anonymous reviewer for pointing out that this point should not be skipped; the great advances in (neural) machine learning from the past few years mostly rely on the availability of very large amounts of training data. So, the usefulness of analytical models in many Humanities contexts will in part depend on progress made towards improved machine learning on very small datasets.

4 The scheduling dilemma

The ***scheduling dilemma*** arises from the opposing principles of maximal flexibility in the content-driven choice of where to apply computational modeling components (responding to the needs arising in the hermeneutic evolution of an understanding of the textual material) versus the systematicity in the specification and decomposition of text-analytical tasks (representing a necessary basis for methodologically valid analysis models with a predictable quality). Reliable predictions in complex analysis tasks can only be achieved with a significant development effort, which requires careful planning of the analytical decision points and categories. The epistemic interest within the Humanities on the other hand can only be reasonably pursued if the procedure can react flexibly to observations which only come to attention in the course of the study through the engagement in evolving analyses of the source material.

The scheduling dilemma could in principle be solved by very generous project runtimes: whenever an open partial question arises which has some corpus-related dimension, a "proper" computational development process could be triggered (including manual corpus annotation and computational model optimization). However, in practice this is not a realistic scenario: since the usefulness of a type of computational model in a hermeneutic process is not clear until concrete analyses are available, one would often have to trigger multiple model development processes for parallel exploration, ready to discard most of them– which could create the impression of wasting valuable research time among the collaborators (who are each under publication pressure within their disciplinary home) and might lead to principled doubts about the benefit of computational models.

4.1 Approaching the dilemma with systematic bottom-up resource building

For a realistic integration of approaches, each of the two pictures from Figs. 2 and 4 have to be adjusted to the justified requirements of the partner discipline: corpus choice and annotation by Humanities scholars has to be embedded in a hermeneutic context, and vice versa there has to be room (and expert input) for systematic model development in the contexts that are deemed relevant. In other words both sides have to move (possibly moving out of their comfort zone as Hammond et al. (2013) put it).

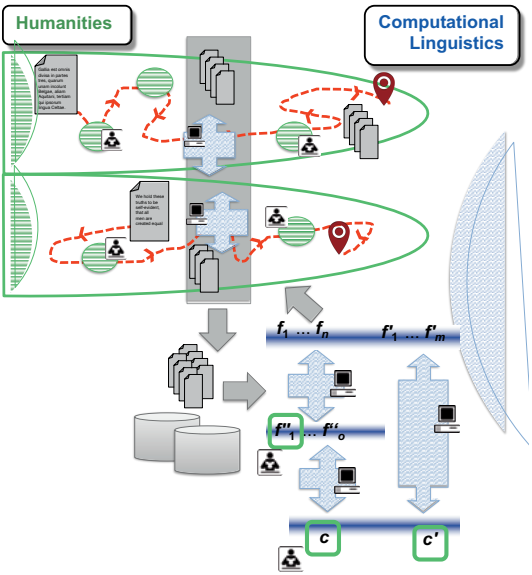


Fig. 5: Schematic depiction of a scenario that would allow for successful synthesis of working practices: when the analysis contexts for computational models in distinct Humanities projects are sufficiently similar, computational optimization efforts can be to the benefit of more than one application case. Besides leading to better tools (most likely), this will provide richer contexts for reflecting the analytical task itself, both from a Humanities perspective and from the computational perspective

Figure 5 depicts the idea of a combined workflow, again schematically (this time limiting attention to just two distinct Humanities project contexts in the upper part).

As the gray box underlying the hermeneutic "trajectories" suggests, the complete independence of the analytical focus from considerations about computational methodological is given up: Humanities scholars commit to experimenting with computational analysis that matches a particular pattern for which machine learning model classes are partially understood and which find correspondences in other Humanities project contexts–thus generating the grounds for systematic exploration both from the technical side and from the point of view of hermeneutic integration. The computational specialists on the other hand commit to adjusting the scope of their machine learning experiments to the needs dictated by the actual context(s) of application, including choice of corpus, focus of analysis task and possibly the emphasis on theory-dependent target categories with rather limited intersubjective stability.

How can this schema be implemented in practice? Within the spectrum of possibilities there is one that requires considerable time and patience, but avoids risks with regard to potentially missing out an important component within complex analytical scenarios: building up a full pipeline or network of related subtasks in a careful bottom-up manner (necessarily making some selective choice regarding target discipline, language, genre etc.). Such a **systematic bottom-up resource building approach** is effectively what a lot of computationally oriented projects in the Digital Humanities have been taking in the past 5–10 years (see e. g., Biemann et al. 2014; Kuhn, Alexiadou, et al. 2016; ter Braake et al. 2016; Gurevych et al. 2018), with varying dynamic flexibility in the interaction between the computational side and Humanities scholars. Over time, the inventory of readily available tools is growing, such that access to reliable computational analysis models in the course of a hermeneutic process will become less and less constrained.

A disadvantage of this path (if one wants to call it a disadvantage) is that the systematic build-up of methodological insights involves extensive phases of groundwork that do not make substantial contributions to the classical core areas in the Humanities. As a consequence, there tends to be limited recognition of this groundwork. Many DH researchers however view the process as a longer-term enterprise that requires some patience. Implications for the core areas should become noticeable once the analytical machinery has been carefully formalized and modeling approaches have been adjusted to the specific needs of the field.

A risk of the systematic bottom-up approach is the following: given the loose connection with dominant research questions from the core fields, targeting text interpretation, the DH agenda may develop a momentum of its own that could push the point of convergence between computationally oriented work and the

traditional fields further and further into the future. Also, it has to be noted that in many cases, the resource building for relevant subtasks cannot rely on an established inventory of descriptive categories appropriate for sharing across specific research frameworks, so the process has to be interleaved with theoretical groundwork.

4.2 An alternative strategy: rapid probing of analysis models

Without questioning the merits of the longer-term agenda of building up a more or less exhaustive pipeline of analysis models, I would here like to discuss a different strategy that would be worth while to explore in parallel. Simply put, it can be regarded as an attempt to translate the long-established concept of rapid prototyping from Software Engineering to the transdisciplinary field of computationally advanced DH, permitting for what we might call **rapid probing** of computational analysis models within a hermeneutic context. To avoid the pitfalls of the plain picture from Fig. 4, constraining principles about the choice of target models have to be assumed (essentially the gray box from Fig. 5). However, a full bottom-up regime is not necessary before assessing the usefulness of an analytical step. When a sufficiently similar model is available for rapid adaptation, relevant aspects of the behavior of the real tool (if it was built) can be anticipated. This could help make a choice among alternative candidate options, possibly saving considerable development effort in fruitless directions.

Integration of rapid probing within a truly hermeneutic approach is as yet still a programmatic idea. In such a context the assessment of prototype models, which share only certain properties of the target analysis scenario, may be harder than in typical cases of language technology development where such a strategy is more commonly applied. But if it could be made to work, the rapid probing idea has an enormous potential. As a parallel strategy besides systematic resource building, it could generate the dynamics that the patient bottom-up path tends to lack– indicating the potential that lies in deep computational analysis.

What does it take to make rapid probing work practically? The idea depends crucially on positive answers to two questions: (a) Is it possible to migrate existing complex analysis pipelines across text collections and (partial) analytical questions? To be useful, the required technical effort should be rather limited while at the same time allowing researchers to analyze a significant part of the target corpus in a robust way (though not necessarily with the highest possible quality). (b) Can the evolutionary unfolding of content-related questions in the Humanities be augmented to incorporate experiments with preliminary corpus analysis steps? These experimental analyses, along with independent analytical considerations,

should help estimate the viability of expanding the preliminary model (and hence make an informed decision when alternative modeling options are available).

Neither question can be fully answered independently of the other one–a quick technical solution to (a) that does not provide appropriate starting points for critical reflection of the preliminary analysis may make addressing issue (b) essentially impossible, even for the most highly motivated team of Digital Humanities collaborators. Nevertheless focusing mainly on question (a) in this article, I would like to make the point that there can be a positive answer: for a number of non-trivial analysis tasks, analysis chains *can* indeed be ported to a different analysis task and different corpus of source material with a reasonable effort. This is particularly so for language-technological analysis tasks that build on a background pipeline of Computational Linguistics tools and uses their output representations as features for machine learning methods, which can be flexibly adjusted to study-specific content analyses. Such methods can often be "retrained" for modified target objectives (provided that the linguistic material in the target texts does not radically violate assumptions underlying the standard tools). The gray box in Fig. 5 can be seen as the axis along which rapid adaptation across projects can be performed.

4.3 Illustrating rapid probing with the "Textual Emigration Analysis" system

It is best to illustrate the abstract strategy of rapid adaptation of analysis chains with a concrete example. The web application "Textual Emigration Analysis" (TEA, Blessing and Kuhn 2014),[24] was designed as an example platform showcasing the exploration of biographical information using tools from Computational Linguistics. Fokkens et al. (2014) and ter Braake et al. (2016) discuss a similar system and the methodological framework it takes to integrate it in Computational History. TEA is a good example for illustrating the present methodological point since it facilitates tool chain transfer across contexts. So, the potential ways of realizing a truly Humanities-centered rapid probing scenario can be explained rather clearly with this system–even though in our examples it is still computational linguists that have experimented with the adaptations. The point of the

24 The platform is available at http://clarin01.ims.uni-stuttgart.de/tea/; the different underlying text collections discussed in this article can be selected in a pull-down menu above the world map. Blessing and Kuhn 2014 describes technical and methodological details of the approach; a tutorial for web application can be found in http://clarin01.ims.uni-stuttgart.de/tutorial/tea.html.

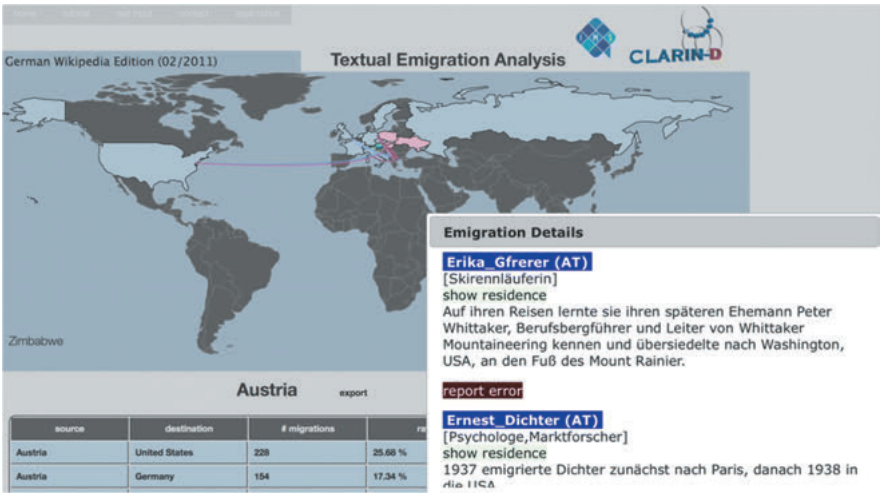


Fig. 6: Web application "Textual Emigration Analysis": screen view after having selected of the Wikipedia-based extraction results for Austria and furthermore having activated of the detailed text instances for migration from Austria to the United States

present article remains programmatic; technical experiments are provided to make abstract methodological ideas more concrete.

TEA is based on automatic extraction of specific biographical events from large text collections, providing an interactive visualization for aggregated extraction results. Textually extracted facts provide an enormous potential for further exploration and aggregation of distributed detail information. We chose the description of a person emigrating or relocating to a different country as an appropriate test case for this platform. This type of biographical event is (a) of interest for a variety of broader analytical studies; (b) it occurs relatively frequently; and (c) it can be visualized in aggregated form geographically. There are quite a few linguistic formulations for emigration events that can be found:

- sie übersiedelte nach Warschau
  "she relocated to Warsaw"
- Der Weg in die Emigration [...] führte über die Schweiz und England letztlich in die USA.[25]

25 From the German Wikipedia entry for Alfred Hauptmann (1881–1948).

"The path to emigration [...] led through Switzerland and England, finally to the USA"

– Später ging sie nach Norwegen, wo sie zu den prominentesten deutschen Emigranten gehörte.[26]
"Later she went to Norway, where she was among the most prominent German emigrants"

From a given text collection, textual descriptions of emigration from country A to country B can be extracted automatically, and the overall relation can be visualized in an interactive world map by countries of origin and destination. Figure 6 shows a screen shot of our web application with the mouse pointer over Austria. Countries that are the origin of a relocation to Austria are light red; destination countries of a relocation from Austria are light blue. A table (at the bottom) shows the absolute numbers and the relative distribution among the source and destination countries and provides hyperlinks pointing to a list of the underlying text snippets that formed the basis of the extraction.

The snippets are displayed in a pop-up window (labeled "Emigration Details"), and are again linked to the full text source. The hyperlinking makes it straightforward for users, for example, to reassure themselves that there are no errors in the automatic extraction.

The extraction of relevant event descriptions is based on a complex analysis pipeline, starting with preprocessing of the text base, followed by a sequence of standard natural language processing (NLP) steps–tokenization, part-of-speech tagging, lemmatization, syntactic parsing–and, lastly, task-specific steps, which combine textual information with metadata or data available in (semi-) structured format, such as the country of birth of a person. The actual determination of instances of the emigration relation (or an emigration event)–here a three-place relation between the person, the linguistic description of the place or country of origin and the description of the destination place or country–from the various distinct linguistic realization variants, is based on supervised machine learning, i.e., the mapping is induced from a collection of manually marked training examples, taking advantage of the generalizations captured in the linguistic analyses output by the NLP tools.

As is also discussed in Blessing, Glaser, et al. 2014; Blessing, Kliche, et al. 2015; Kuhn and Blessing 2018, a rapid adaptation of TEA's original analysis chain to other corpora and research questions is feasible (although the chain is relatively

26 German Wikipedia entry for Hanna Sandtner, née Ritter (1900–1958).

complex), and has turned out useful in practical experiments. We can distinguish a number of adaptation scenarios:

(I) Adjusting the target relation

The target relation for which textual instances are extracted can be adjusted in an interactive training process. So, instead of emigration events the corpus can be searched for another type of event. Extraction is realized as a machine learning classifier. The features on which the classifier is based include the output of language-technological preprocessing tools (incl. tokenization, lemmatization, part-of-speech tagging and syntactic dependency parsing) which are run independently of the target task. The learning process can thus exploit new generalizations for the adjusted target relation exploiting the interactive training. An example relation for which we retrained the system is membership of a person in political parties and associations.[27]

When the text material is very divergent from the typical newspaper data for which the preprocessing tools were developed, the quality of analysis degrades. However under certain circumstances the trained relations extraction component can display acceptable behavior despite faulty underlying analyses, since the machine learning may be able to compensate for systematic errors in the preprocessing output (which has no deterministic influence on the classifier decision). This means that migration of machine-learned tools can be included in exploratory prototype experiments even when the corpus material is deviant– providing some indications for the decision for or against a full adjustment. In this scenario the research team will consider the preliminary analysis results primarily from a method-oriented point of view, abstracting away from details of the content analysis: Could higher-quality results of the same kind be useful in the process of realization? Suppressing certain aspects of sample data for some conceptual considerations is routine in Computer Science and Computational Linguistics as method-oriented disciplines. However, under a Humanities perspective such a move is highly unusual. Perhaps this point is one of the biggest hurdles for Implementierungimplementation of the proposed synthesis of procedural practices. And it can be seen as a central task for the recently implemented

27 Of course, the aggregation steps and visualization for the emigration analysis cannot be retained for arbitrary target relations. But with a sufficiently modular design it is conceivable to include aggregation mechanisms in the prototype-oriented exploration. In general the adjustments are however non-trivial and require close dialogue between the technical and content point of view (for discussion of the interdisciplinary approach see Kuhn 1999; Kuhn, Alexiadou, et al. 2016).

Bachelor and Master programmes in Digital Humanities to train students in this aspect of the method-oriented abstraction from certain characteristics of the research object.

In many DH projects, the text corpus to be studied is not available in fully digitized form at the project start. Even in such cases, the exploration-through-prototype-migration approach can be carried out: one or more existing corpora that are similar to the later target corpus in relevant dimensions can be used to approximate the ultimate corpus for the purposes of assessing analytical options. (Of course, the abstraction skills are strained even more in this case.)

(II) Migrating the analysis pipeline to other text sources

The second type of pipeline adjustment which Blessing, Glaser, et al. (2014) and Blessing, Kliche, et al. (2015) carried out for the TEA system pertains to the adjustment of a system from the original development corpus (in our case, the collection of all biographical articles in the German version of Wikipedia) to a different underlying text collection. In our experiments it was effectively possible to perform a rapid technical migration of the full pipeline to the text collection underlying the Austrian Biographical Lexicon (Österreichisches Biographisches Lexikon, ÖBL) in a very short time (about four hours),[28] and similarly for the German Biography (Deutsche Biographie). This means that the aggregation and visualization functions can be readily used on different source bases. In this context, the modular design of the system architecture and the use of web services from the CLARIN-D infrastructure pays off (Mahlow et al. 2014) .

The prototype is fully appropriate for an estimate of the potential argumentative benefits to be expected from a more thorough migration, for which we argue in this article. Moreover, the idea of Linked Data can here be exploited, i.e., the textual analysis of biographies for specific people can be juxtaposed or merged in cases where more than one collection contains an entry. This invites the exploration of discrepancies in the text sources or peculiarities in the source selection process.

Above all however, the merger of the available resources provides a valuable post-analytical framework for evaluating the tools and methods. As discussed in Sect. 2.2, the practical development of analytical models often suffers from the notorious difficulty of detecting recall problems. Inspecting the system output does draw attention to precision errors, but to detect a recall error, which is an omission by the system, one would have to know the set of results. And for rare phenomena,

28 We are grateful to Eveline Wandl-Vogt and her colleagues at the Austrian Academy of Sciences for providing the text data and sharing their expertise.

one can only approach it by putting considerable effort into random sampling of data.

If two systems are available for which one would expect the same result (at least for some of the data, e. g., a person who appears in two different biographies), a systematic comparison of the system results can be used to detect (certain types of) recall problems: For example, if system A predicts the emigration of a person X to country L, but system B (which is based on another biographical collection) does not, the origin of the discrepancy can be easily checked in the pipeline.[29] The following examples illustrate the comparison between the extraction of emigration events from Wikipedia and ÖBL.[30] In (1) and (2), the extraction results from ÖBL indicate that there were recall errors in the Wikipedia-based extraction: the construction *zog mit ihm nach England* ("moved with him to England") in (1b), embedded in a coordination structure was not recognized; in (2b) there is a complex coordination structure too; in addition, the Wikipedia article was missing punctuation (a period after *zwangspensioniert* ("forced to retire").

(1) Stokes, Marianne; née Preindlsberger (*1855 in Graz; †1927 in London), painter

   a. [ÖBL-Artikel:] *Während ihrer Stud.reisen in die Bretagne lernte sie Adrian S. kennen und übersiedelte mit ihm nach London, wo S. seitdem regelmäßig ausstellte (u. a. Fine Art Society, Grosvenor Gallery, New Gallery, Royal Acad.).* [**detected**]
   "During her educational journey to Bretagne, she met Adrian S. and moved with him to London, where S. since had regular expositions (a.o. F.A.S. ...)"
   b. [Wikipedia-Artikel:] *1883 lernte sie bei einem Aufenthalt in Pont-Aven in der Bretagne den englischen Maler Adrian Scott Stokes (1854–1935) kennen. Ihn heiratete sie 1884 und zog mit ihm nach England.* [**not detected**]
   "1883 she met the English painter A.S.S. in Pont-Aven in Bretagne. She married him in 1884 and moved with him to England."

(2) Marburg, Otto (*1874 in Römerstadt; †1948 in New York City), neurologist
   a. [ÖBL-Artikel:] *1919 wurde er als Nachfolger Obersteiners Vorstand des Neurolog. Inst. 1938 emigrierte er in die USA und arbeitete als Prof. für Neurol. am College of Physicians and Surgeons der Columbia Univ., wo er*

29 One has to be aware however that those gaps in recall which exist in both systems (possibly for systematic reasons), stay undetected. A comprehensive evaluation of the system quality can only be done on independently sampled and hand-annotated test data.

30 Typographical emphasis by the author of this article.

ein eigenes Laboratorium hatte. [**detected**]
"1919 he become director of the Neurological Institute, as Obersteiner's successor. 1938 he emigrated to the USA and worked as professor for neurology at the College of Physicians and Surgeons der Columbia Univ., where he had his own laboratory."

b. [Wikipedia-Artikel:] *Nach dem Anschluss Österreichs 1938 wurde Marburg wie zahlreiche andere Dozenten der Wiener Universität aufgrund seiner jüdischen Herkunft zwangspensioniert Marburg und seine Frau verließen das Land und emigrierten mit Unterstützung Bernhard Sachs' über England in die Vereinigten Staaten.* [missing punctuation: sic] [**not detected**]
"After the annexion of Austria in 1938, Marburg like many other lecturers from the university of Vienna was forced to retire due to his Jewish origin. Marburg and his wife left the country and Bernhard Sachs helped them to emigrate the the United States, via England."

In (3)–(5) a discrepancy in extraction results indicates a precision error: in (3) a mentioned *failed* emigration attempt led to erroneous extraction in (3b) (but not in (3a)). (4) is similar in that (4a) posits a real emigration, while the biography entry refers to incorrect reports. (5b) talks about "inner emigration", which triggered an erroneous extraction; this can be explained by a special heuristic in the analysis chain: whenever a sentence with a trigger for the emigration relation (like the noun *Emigration*) lacks information on the countries of origin and destination (or place of birth), the system falls back on the place (or country) of birth or death from the structured data. In case the countries of origin and destination are distinct, an emigration movement is postulated. In this particular case, Trient was part of Austria-Hungary in Nikodem's youth, but our prototype system uses present-day boundaries to map place names to countries–hence the move within the country is erroneously detected as a transnational relocation. (The mentioned heuristic seems somewhat risky, but it helps overcome substantially more precision errors than there are recall errors that it introduces. Nevertheless, the erroneous extraction of "inner emigration" cases could be overcome by a retraining of the classifier.)

(3) Klang, Heinrich Adalbert (*1875 in Wien (Vienna); †1954 in Wien), Law scholar

a. [ÖBL-Artikel:] *Bemühungen um die Ausreise sowie ein Fluchtversuch nach Ungarn scheiterten.* [**correctly not extracted**]
"Requests for emigration and an attempt to escape to Hungary failed."

b. [Wikipedia-Artikel:] *Mehrere Versuche legal zu emigrieren, so in die USA, Kuba und nach China, scheiterten.* [**erroneously extracted**]

"Several attempts to emigrated legally, e. g., to the USA, Cuba and China, failed."

(4) Kossak, Leon (*1815 in Nowy Wiśnicz; †1877 in Krakau), Polish officer and painter
   a. [ÖBL-Artikel:] *1848 nahm K. am ung. Aufstand teil, kämpfte im Rgt. der poln. Ulanen und nahm an der Schlacht bei Világos teil. Wo er sich dann aufhielt, ist nicht bekannt, in der Literatur wird irrtümlich angegeben, er sei nach Australien ausgewandert.* [**erroneously extracted**]
   "In 1848, K. participated in the Hungarian revolt, fought with the regiment of the Polish Ulans and participated in the battle near Világos. It is unknown where he lived afterwards, the literature mentions erroneously that he emigrated to Australia"
   b. [Wikipedia-Artikel:] [**no mention of putative emigration in the text**][31]

(5) Nikodem, Arthur (*1870 in Trient; †1940 in Innsbruck), painter
   a. [ÖBL-Artikel:] [**no mention of the move in the text, only travels**]
   b. [Wikipedia-Artikel:] *Nikodem begab sich daraufhin in eine Art "innere Emigration"; nur ihm sehr Nahestehende hatten die Möglichkeit, seine Arbeiten zu sehen.* [**erroneously extracted**]
   "Nikodem then went into a kind of 'inner emigration'; only close friends and relatives had a chance to see his work."

A third dimension of adjustment discussed in Blessing, Glaser, et al. 2014; Blessing, Kliche, et al. 2015 is system transfer across languages.

To conclude this section, we discussed ideas for overcoming a dilemma resulting from divergent scheduling priorities: A Humanities-centered view would prefer to avoid an early commitment to specific types of analytical models, while under a CL-centered view, best results are obtained with an early detailed specification of the input/output relation in specific analysis steps. A standard strategy for resolving this conflict in practice is the recourse to maximally generic analysis tools, which can be applied without adaptation. This however clearly underexploits the potential lying in the development of more targeted analysis tools with knowhow from CL.

The rapid probing idea advocated here resolves the scheduling dilemma differently, aiming to encourage explorations of more complex modeling approaches to feed the hermeneutic process. A (rapid and preliminary) migration of complex

31 The Polish Wikipedia however mentions a 1-year stay in Australia, with his brother Władysław Kossak, who did emigrated there himself (according to Polish Wikipedia).

analysis pipelines across structurally related subject areas can help Humanities scholars to integrate the choice of computational analysis steps in their hermeneutic process of unfolding research questions–without effectively implying a commitment to the prediction results etc. and hence not presenting a limitation, but an enrichment of the procedural practice. In the further development and fine-tuning of those computational components that seem promising for the ultimate argumentation, the rapid prototyping approach brings the advantage that details of the models can be discussed and expanded in close interaction between the Humanities scholars and the computer scientists.

5 The subjectivity problem

Orthogonal to the scheduling dilemma, an application of the data-oriented standard methodology from Computer Science/Computational Linguistics in hermeneutically oriented research contexts may run up against what one may call the ***subjectivity problem***. As laid out in Sect. 2, within the computational disciplines the "proper use" of computational modules in an analysis chain has to adhere to the established annotation-based methodology for specifying the modules' input/output relations: annotation guidelines have to operationalize the categories of annotation, such that an intersubjectively stable observation about language use in context is captured. By measuring inter-annotator agreement in multiple annotation experiments, the effectiveness of guidelines can even be tested empirically. Target categories leading to low levels of agreement in human annotation are generally considered problematic for data-driven modeling.

Now, when aspects of literary or historical text interpretation are targeted in a text study, the postulate of intersubjectively stable "results" becomes highly controversial. In the hermeneutic context, the process of text interpretation/textual criticism (targeting the relational notion of *significance*) is not aimed at a single, "correct" target for a given text–even if the full text production context is taken into account in all facets. Rather, throughout the reception history of important texts, new interpretations have been and will be obtained, taking different points of view such as a psychological dimension, societal considerations, production aesthetics, emphasis on intertextual links with other works, etc. In most cases, a new interpretation does not invalidate earlier interpretations. In Literary Studies, a broadly shared hypothesis is that literary texts are inherently ambiguous or "poly-

valent".[32] As a consequence, for text properties connected up with interpretive differences, intersubjectively stable interpretation results cannot be assumed. What does this imply for the applicability of the standard annotation-based methodology in the study of literary or historical texts? A plausible reaction would seem to be to completely exclude the sphere of interpretation (in the literary or historical sense) from the scope of formal/computational modeling–leaving it to traditional hermeneutics–and rather concentrate the operationalized annotation guidelines and computational modeling efforts on descriptive categories for surface-related text properties, for which intersubjective agreement can generally be reached.[33] The annotation approach in the heureCLÉA project (Gius and Jacke 2016), focusing on narrative literary texts, implements such an approach, including reconciliation steps for resolving disagreements.

At the same time, the exclusion of those text properties from formalized annotation that are contingent on interpretive decisions seems awkward too: one of the purposes of the traditional practice of (individual, subjective) text annotation has been for the reader/annotator to record one's subjective reading impression: These may provide the basis for observing systematic patterns among text properties in a second pass. The objective of systematicity in annotation and the concession that certain annotations are influenced by subjective judgements do not necessarily exclude each other. It would seem that a computationally enhanced hermeneutic approach could benefit from computational models based on subjective annotations–even though these do not follow the rules of "proper" data-driven modeling.

5.1 Base for illustration: point of view in narrative text

The desideratum to address the subjectivity problem in more lenient ways becomes particularly clear when considering the interplay across levels of "depth" in text analysis. As I argue in Kuhn in preparation, most categories of text analysis

32 Jannidis (2003) provides a critical discussion of the polyvalence thesis; he argues against a strong version of the thesis, which claims that because of inherent polyvalence of literature it is impossible to argue–given *any* two interpretations of a text–that one is more appropriate than the other. Hammond et al. 2013 discuss the difference between the two cultures of Computational Linguistics vs. Literary Studies from the experience of an annotation project targeted at literary texts–they also identify the status of ambiguity as a key issue.

33 The separation between the two distinct levels of text meaning and literary significance advocated by Hirsch (1967) could be used as a basis–although according to Jannidis, Lauer, et al. (2003) the assumption of an intersubjectively stable referential meaning of a text is itself controversial in the hermeneutic tradition.

that would under most circumstances be considered plain descriptive–i.e., candidates for inclusion in the strict annotation methodology–can appear in ambiguous patterns, which effectively open up a disambiguation choice that depends on preference among alternative interpretations.

Consider for instance the classification of narrative point of view in narrative texts by the Austrian author Arthur Schnitzler (1862–1931). Many of his shorter narrative texts (e. g., *Berta Garlan/Frau Bertha Garlan*, 1900) are written in third-person narrative voice, limited to the subjective viewpoint of the focal character.[34]

(6) *It was striking nine from the tower of the Church of St. Michael when George stood in front of the café. He saw Rapp the critic sitting by a window not completely covered by the curtain, with a pile of papers in front of him on the table. He had just taken his glasses off his nose and was polishing them, and the dull eyes brought a look of absolute deadness into a face that was usually so alive with clever malice. Opposite him with gestures that swept over vacancy sat Gleissner the poet in all the brilliancy of his false elegance, with a colossal black cravat in which a red stone scintillated. When George, without hearing their voices, saw the lips of these two men move, while their glances wandered to and fro, he could scarcely understand how they could stand sitting opposite each other for a quarter of an hour in that cloud of hate.*
[Arthur Schnitzler: *The Road to the Open*, translated by Horace Samuel[35] (Chapter 2)]

(7) [George has just asked Heinrich a question]
Heinrich nodded. [...] He sank into meditation for a while, thrust his cycle forward with slight impatient spurts and was soon a few paces in front again. He then began to talk again about his September tour. He thought of it again with what was almost emotion. Solitude, change of scene, movement: had he not enjoyed a threefold happiness? "I can scarcely describe to you," he said, "the feeling of inner freedom which thrilled through me [...]."
George always felt a certain embarrassment whenever Heinrich became tragic. "Perhaps we might go on a bit," he said, and they jumped on to their machines.

34 The notion of *focalization* as a differentiated narratological concept goes back to Genette (1972) (see Köppe and Kindt 2014, pp. 208 sqq.); numerous variants for a definition have been proposed in narratological theory. For the purposes of making the relevant interpretive distinctions in the annotation experiments I discuss below, a relatively simple characterization of psychological point of view similar to the one Wiebe (1994) is sufficient (see also Sect. 5.2 below).

35 The text passages are taken from the English translation of Arthur Schnitzler's *The Road to the Open* available at http://www.gutenberg.org/ebooks/45895.

[Arthur Schnitzler: *The Road to the Open*, translated by Horace Samuel (Chapter 3)]

In his novel *The Road to the Open* (*Der Weg ins Freie*, 1908), the viewpoint of the third-person narration alternates to a certain degree between several characters' subjective viewpoints and an objective viewpoint (predominant is narration from the narrow subjective scope of the Catholic composer George von Wergenthin-Recco, but occasional passages also take the Jewish writer Heinrich Bermann's and other characters' subjective viewpoint).

At first glance, the following two passages from chapter 2 and from chapter 3 appear to be typical depictions of George's and Heinrich's viewpoint respectively. Passage (6) directly and indirectly conveys sensory perceptions by George (e. g., him seeing Rapp polishing his glasses). Seemingly similar, passage (7) depicts the mental state of Heinrich, in part through direct attribution ("*He thought of it again with what was almost emotion.*"), in part through free indirect discourse ("*Solitude, change of scene, movement: had he not enjoyed a threefold happiness?*").

Annotating the subjective viewpoint accordingly would hence seem to be relatively uncontroversial (George for (6), Heinrich for (7)). However, when looking at the novel as a whole (and at Schnitzler's shorter, limited-viewpoint narrations), it turns out that there are many passages with an extended build-up establishing one character's inner view, which can then be kept up for quite some time, including the perception of other character's actions. Since formally, we find different variants of third-person narration, the transposition of whose viewpoint we are being presented is quite subtle. Assuming that Schnitzler likes to play with this uncertainty (which is an interpretive postulate!), passage (7) can be convincingly analyzed as depicting George's perception of Heinrich's actions: Heinrich's pushing of the bike is deictically related to George's position ("*a few paces in front*"), and we do not learn about the content of Heinrich's meditations until he begins to speak (so we can hear him through George's ears). The most misleading sentence is "*He thought of it again with what was almost emotion.*" What appears like a switch of the narrator's voice towards Heinrich's inner view can of course also be free indirect discourse– conveying George's perception of Heinrich saying "*I think of it again with what is almost emotion*".[36] There is nothing in the passage about

36 For the two sentences about the September tour, the English translation makes the "transposed viewpoint" interpretation somewhat less salient than the German original: *Dann begann er wieder von seiner Septemberreise zu sprechen. Beinahe mit Ergriffenheit dachte er an sie zurück.* So one might speculate that the translator did not have the transposed interpretation in mind. systematic modeling efforts (or formal annotations) to conditionally depend on the acceptance of some subjective pre-understanding.

Heinrich's mental state that is not conveyed through an indirect or direct quote of what Heinrich uttered in the situation. The closing sentence "George always felt a certain embarrassment whenever Heinrich became tragic" resolves the tension, revealing whose viewpoint we were confronted with earlier on in passage (7). (Note that this is an interpretation of the aesthetics of the passage, which presumably cannot be defended on intersubjectively uncontroversial grounds, although it could–hopefully–be made plausible by appealing to fine-grained distinctions in the linguistic form and comparisons with other passages in the novel and other texts by the author, i.e., elements of a hermeneutic process.)

So, what we can observe when analyzing text passages using largely descriptive narratological categories is the following: the inherent ambiguity of many linguistic characterizations can easily lead to situations where "deeper" interpretive decisions percolate down to more superficial ones. (In our sample scenario, an interpretive hypothesis percolates down the recursive embedding of narrative levels: are we seeing one character's inner state or is it another character's perception on the first one talking about his inner state?)

If one takes the subjectivity problem to exclude a formal annotation approach (because no sufficient inter-annotator agreement can be reached), then, the possibility of such percolation happening implies that there might be no level of descriptive text analysis that is perfectly "safe" from interpretive biases. Vice versa, one might take it as a plausibility argument for an approach taking certain

5.2 Modeling subjective categorizations: Another place for rapid probing?

For the subjectivity problem, the rapid probing idea of methodological integration presented in Sect. 4.2 can also be realized based on a standard NLP analysis chain, augmented with a task-specific machine learning classifier that is trained with the rapid prototyping idea, similar as in the previous section. The corpus data and research questions are narrative literary texts, on which narratological categorizations are performed that may be correlated with interpretive decisions.

In Kuhn in preparation, pilot experiments on a corpus of Schnitzler texts are discussed, targeting annotation of character-specific viewpoint in the narration. The idea is to explore the implications of (different subjective) interpretive preassumptions by integrating them in a machine learning classifier.

The experiments adopt a straightforward mention-based operationalization of point of view that is compatible with the formally precise descriptive framework worked out by Wiebe (1994) for predicting psychological point of view in narrative texts. Her model takes the form of an algorithm that predicts at each mention of

a character in the linear text sequence, whether or not the previously established point of view stays the same, or whether it is shifted to the character now mentioned. The algorithm is formulated deterministically, taking into account a differentiated set of linguistic features; so whenever there are competing interpretation options, Wiebe's algorithm would enforce a decision. However, the decision relies on the auxiliary notion of *subjective elements*, which would be the natural place for including non-determinism in the algorithm.

With modern machine learning techniques, a simple mention classification framework is a sufficient basis for rapidly probing experiments testing the effects of a model that follows a particular approach to reading point of view in Schnitzler's text. Linguistic indicators (explicit attribution of speech or thought, deictic elements, adverbial modifications, reference to sensory perception, etc.) and contextual build-up, including certain patterns of character references, are included in the feature set, and so are style indicators (as Brooke et al. (2017) showed in a detailed analysis of free indirect discourse in the writings of Virginia Woolf and James Joyce).

Due to the subtle interactions, we cannot expect a machine learning approach to reliably and robustly predict the "correct" subjective viewpoint. However, following the rapid probing idea, the behavior of alternative predictive models trained on manually annotated viewpoint annotations can be systematically compared, potentially allowing for conclusions about the role different factors play; similarly, models trained on distinct texts can provide indications for a contrastive analysis.

The relevant datapoints in the machine learning classification are defined to be all mentions of characters, in their respective context. For (6), an excerpt of the above text passage, there would for instance be seven datapoints. The annotation decision is a binary decision: whether or not character referred to by the mention is the focus of perception at the current point of narration–where what counts is the informed reading impression, i.e., readers who have the individual impression that frequent switches of viewpoint occur will make different annotations than readers who perceive long build-ups of embedded narration levels (as discussed above). For (8), the annotation would be uncontroversial: the first two mentions refer to the focus of perception, the remaining ones do not.

(8) *[George] stood in front of the café. [He] saw [Rapp the critic] sitting by a window not completely covered by the curtain, with a pile of papers in front of [him] on the table. [He] had just taken [his] glasses off [his] nose*

Subjective annotations of this kind can be performed quite fast; for a pilot study, about a thousand data points could be annotated within a few hours. Note that

Tab. 1: Experimental results from pilot study on narrative viewpoint classification

		Logistic regression classifier			
		Precision	**Recall**	**F_1**	**Accuracy**
(A) Training: *Road*, Test: *Garlan*	foc.	0.77	0.71	0.74	0.78
	non-foc.	0.79	0.84	0.82	
(B) Training: *Road*, Test: *Road*	foc.	0.73	0.77	0.75	0.75
	non-foc.	0.76	0.73	0.74	
(C) Training: *Road + Garlan*, Test: *Road*	foc.	0.76	0.73	0.75	0.79
	non-foc.	0.81	0.83	0.82	

no distinction between explicit thought attribution, free indirect discourse, etc., are made, since–by design–emphasis in this pilot study is placed on the pattern of switches in viewpoint.

On the dataset, supervised classifiers can be trained using a standard machine learning library (e. g., the Python library scikit-learn http://scikit-learn.org/). As features, the output from multiple NLP analysis tools can be used,[37] including syntactic structure (which is important for detecting attributions of speech and thought), co-reference, but also verb class membership (which may lead to better generalizations).

Besides training the classifier on manually annotated data, one can also experiment with systematic automatized annotations. As mentioned above, some shorter Schnitzler narrations are entirely told from a single character's viewpoint, e. g., *Berta Garlan*. Using automatic co-reference resolution, with a few minutes of manual post-correction, a dataset marking all mentions referring to the main character as "focal"[38] can be generated, and one can experiment with an "intertextual" model transfer: the automatized *Berta Garlan* data are used to train a supervised classifier, and this is applied to data from *Road to the Open*, in which psychological point of view varies more among the protagonists.

37 The experiments, which were run on the English translation of the Schnitzler texts, were based on the output oft he Stanford CoreNLP tools (https://stanfordnlp.github.io/CoreNLP/). In addition, the verb classification from VerbNet (https://verbs.colorado.edu/verb-index/), and the stylistic profile vectors from Brooke et al. 2017 were used, which are included in the Gutentag tools suite (www.cs.toronto.edu/~jbrooke/gutentag/; the style profile lexicon is available as built_in_lexicons/sixstyleplus.txt; thanks to Julian Brooke).

38 What we call "focal" here is operationalized as evoking the reader's interpretation of the character being the present exponent of subjective point of view.

Table 1 displays evaluation results of a number of different training experiments on held-out test data–the idea being to give some indication of what kind of considerations can be taken. The rows in the table ((A) through (C)) vary the training and test data in the experiments, and in the columns, evaluation results for classifiers trained with Logistic Regression are shown.[39]

In scenario (A), a classifier is trained on *Road to the Open* and tested on manually annotated test data from *Berta Garlan*. The fact that relatively decent accuracy scores (0.78) can be reached in transfer across texts seems to indicate that the model picks up a certain level of abstraction (across texts, it cannot be highly text-specific clues that help in testing).

In scenario (B), with training and test data from the same text (but with a smaller training set than in (A)), the prediction accuracy is slightly lower than in (A) (0.75).[40] Scenario (C) includes "mixed" training data, testing on the same data as in (B). The classifier benefits from the increased amount of training data–which could be an indication for relatively homogeneous patterns of narrative viewpoint across texts.

As a last thing, it can be interesting for a text study aiming at interpretive aspects to check the machine-learned classifier on other texts or parts of the development text that were not taken into account in the annotation. About *Road to the Open* it has for example been observed that George's mistress, Anna Rosner, is very rarely focalized. One can now look for text passages in which the classifier (trained on viewpoint contexts for other characters) nevertheless predicts a subjective viewpoint for clusters of references to Anna. Passage (9) (from the end of chapter 2) is an example of such a passage. This can be compared with passages where references to her are assigned low scores by the subjective viewpoint classifier, e. g., (10) (from the beginning of chapter 3).

(9) *She had for the first time in her life the infallible feeling that there was a man in the world who could do anything he liked with her.*

(10) *Anna had given herself to him without indicating by a word, a look or gesture that so far as she was concerned, what was practically a new chapter in her life was now beginning.*

39 F-Score is the harmonic average of precision and recall.

40 In experiments not shown in the table, we trained a different classification algorithm, Gaussian Naive Bayes, which generally led to lower accuracy. However it benefitted more from scenario (B); here, scenario (A) resulted in an accuracy of 0.67, compared to 0.74 for (B). Comparisons of this kind, plus error analyses etc., may help to develop a better understanding of the analysis problem.

(9) is indeed one of the few passages in which the narrator merges with Anna's subjective perception, whereas in (10), the subjective viewpoint is intuitively George's. So, we do find indications that a machine-learned classifiers, which a scholar can adjust to his or her individual pre-understanding within a few hours, could indeed be of use for advanced text-analytical explorations.

Of course, if a pilot study converges on certain correlations, structural patterns, etc., tentative insights from rapid probing have to be followed up by efforts for building relevant components more systematically and subjecting the models to a strict empirical evaluation on independently obtained target annotations.

6 Conclusion

This position paper took as its point of departure the observation that in research targeted at the literary or historic interpretation or *significance* of texts in a broadly hermeneutic sense, the use of complex computational modeling components is still an exception–despite substantial advances in the development of highly adaptable computational frameworks and infrastructures supporting re-use of tools, corpus resources and annotations. I argued that this can be explained (at least in part) through diverging working practices in humanities disciplines (predominantly following a hermeneutic tradition of text interpretation) versus the method-oriented research strategies in Computational Linguistics (with a relatively strict conception of the "proper use" of text analysis models). The diverging methodological principles pose a challenge for a joint methodology/working practice. Specifically, we can identify a scheduling dilemma that makes it hard to deploy sophisticated computational analysis chains in specialized hermeneutic studies, and the subjectivity problem. The latter originates from the constraint that in standard data-driven modeling, gold-standard annotations have to be grounded in operationalized categories leading to high inter-annotator agreement. But hermeneutic text analysis may explore implications of pre-understandings of a subjective kind, grounded for instance in aesthetic judgements.

There are ongoing efforts of systematically extending the basis for computational models targeted at research questions from the Humanities that avoid methodological clashes, which were not in the focus of this contribution. I conjecture that moving along this "main" path of systematic bottom-up resource building will expand the established computational methodology to gradually open up to more hermeneutically oriented research questions. Yet, as a swifter way of exploring the potential for methodological innovations, it may be fruitful to occa-

sionally run rapid probing experiments, in which the established constraints on the "proper use" of computational models are consciously weakened in order to facilitate experimental investigations–also in scenarios for which no carefully developed datasets and tool chains are in place, or where researchers would like to explore implications of their own or another scholar's subjective pre-understanding. Ideally, the increased flexibility will inspire experimentation in more risky interdisciplinary terrain, and the successes (and failures) help assess the value of novel ideas that are otherwise too far out of the established methodological reach.

Advocating a rapid probing approach of course bears a certain risk. Swiftly achieving preliminary analytical results on a text or corpus of interest may tempt researchers to jump to conclusions. A rapidly transferred prototype model may have flaws that make the text analysis (incorrectly) appear to corroborate a scholar's pre-understanding regarding an important hypothesis (where the pre-understanding may or may not be erroneous). It is hence crucial to apply the rapid probing approach under a regime that follows highest methodological standards, which means that no substantial conclusions may be drawn from a tool's prediction unless it has been evaluated against independently annotated test data in the target domain.

What we have here is a small-scale version of the re-occurring methodological dispute: should it only be the firmly established methods that drive the research agenda–methods from the core of the research paradigm for which the community has a clear consensus (potentially at the cost of missing out on questions that cannot be fully stated within the framework)? Or is it legitimate to pursue more unorthodox research ideas exploring the borderlines of the established consensus (without sacrificing crucial standards of validity)? Rapid probing with careful subsequent empirical evaluation seems a viable instrument for proponents of the latter self-understanding. And I believe that with the current emphasis on heavily data-driven approaches to language and text, it can be healthy and fruitful to encourage such alternative practices. Indeed, a (broadly) hermeneutic approach differs from the mainstream paradigm in that the research process typically starts out from the object of study–a text or text collection–and generates questions by taking contextualizing views of the object of study (not applying filters of what are feasible methodological ways for addressing the questions until later in the process). It will still be only possible to address certain questions on a computational basis; but an approach that subordinates choice of methods to the collection of questions prompted by contextualizing the object of study will generate a different awareness of biases and systematic gaps.

Ultimately, the established (continuously growing) catalogue of NLP tasks that can be "solved" with measurable quality bears its own risk: it is easy to be overly optimistic about the degree of scholarly and scientific understanding that

trainable predictive models give us about language, culture and cognition. Here too, alternative paths in research practice may help overcome biases and unveil gaps.

Acknowledgment: The author would like to thank the organizers of the COLING 2016 Workshop on Language Technology for Digital Humanities (LT4DH) in Osaka, Japan, for the opportunity to engage in timely discussions about possible contributions that Language Technology and Computational Linguistics can make to the Digital Humanities. Many thanks go to five anonymous reviewers for very valuable comments and suggestions on manuscript versions of this article and to the colleagues in Stuttgart for contributions, discussions and comments, in particular to André Blessing, Nils Reiter and Gabriella Lapesa. This work was partially supported through a grant from the German Federal Ministry of Education and Research (BMBF) for the Center for Reflected Text Analytics (CRETA). Parts of Sects. 3 and 4 are based on methodological discussions in Kuhn and Blessing 2018 (contributed by the first author); an English translation is used with friendly permission by the editors of the volume "Europa baut auf Biographien. Aspekte, Bausteine, Normen und Standards für eine europäische Biographik" (Bernád et al. 2018).

References

Bamman, David, Ted Underwood, and Noah A. Smith (2014). "A Bayesian mixed effects model of literary character". In: *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*. Baltimore: Association for Computational Linguistics, pp. 370–379.

Bernád, Ágoston Z., Christine Gruber, and Maximilian Kaiser, eds. (2018). *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*. Vienna: New Academic Press.

Berry, David M. (2011). "The computational turn: Thinking about the Digital Humanities". In: *Culture Machine* 12, pp. 1–22.

Biber, Douglas and Edward Finegan (1989). "Drift and the evolution of English style. A history of three genres". In: *Language* 65, pp. 487–517.

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler (2014). "Computational Humanities - bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301)". In: *Dagstuhl Reports* 4.7, pp. 80–111. DOI: 10 . 4230 / DagRep.4.7.80.

Blessing, André, Andrea Glaser, and Jonas Kuhn (2014). "Biographical data exploration as a testbed for a multi-view, multimethod approach in the Digital Humanities". In: *Proceedings of the 1st conference on biographical data in a digital world*. Amsterdam, pp. 53–60.

Blessing, André, Fritz Kliche, Ulrich Heid, Cathleen Kantner, and Jonas Kuhn (2015). "Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien". In: *Grenzen und Möglichkeiten der Digital Humanities*. Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1.

Blessing, André and Jonas Kuhn (2014). "Textual emigration analysis (TEA)". In: *Proceedings of the 9th international conference on language resources and evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA), pp. 2089–2093.

Bozic, Bojan, Gavin Mendel-Gleason, Christophe Debruyne, and Declan O'Sullivan, eds. (2016). *Computational history and data-driven Humanities: Second IFIP WG 12.7 international workshop*. Springer.

Brooke, Julian, Aadam Hammond, and Graeme Hirst (2017). "Using models of lexical style to quantify free indirect discourse in modernist fiction". In: *Digital Scholarship in the Humanities* 32, pp. 234–250.

Burrows, John (2002). "'Delta': A measure of stylistic difference and a guide to likely authorship". In: *Literary and linguistic Computing* 17, pp. 267–287.

Chaturvedi, Snigdha, Shashank Srivastava, Hal Daumé III, and Chris Dyer (2016). "Modeling evolving relationships between characters in literary novels". In: *Proceedings of the 30th AAAI conference on artificial intelligence*, pp. 2704–2710.

Elson, David K., 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*. Stroudsburg, PA: Association for Computational Linguistics, pp. 138–147.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Steffen Pielstrom, Christof Schoch, and Thorsten Vitt (2015). "Towards a better understanding of Burrows's Delta in literary authorship attribution". In: *Proceedings of the 4th workshop on computational linguistics for literature*. Denver, pp. 79–88.

Fokkens, Antske, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber (2014). "BiographyNet: Methodological issues when NLP supports historical research". In: *Proceedings of the 9th edition of the language resources and evaluation conference (LREC)*. Reykjavik, pp. 3728–3725.

Fort, Karën and Benoît Sagot (2010). "Influence of pre-annotation on POS-tagged corpus development". In: *Proceedings of the 4th linguistic annotation workshop, ACL 2010*. Uppsala, pp. 56–63.

Genette, Gérard (1972). *Figures III*. Paris: Seuil.

Giesbrecht, Eugenie and Stefan Evert (2009). "Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus". In: *Proceedings of the 5th web as corpus workshop*, pp. 27–35.

Gius, Evely and Janina Jacke (2016). *Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets. Version 2*. Hamburg.

Gold, Matthew K. and Lauren F. Klein, eds. (2016). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Goyal, Amit, Ellen Riloff, and Hal Daumé III (2010). "Automatically producing plot unit representations for narrative text". In: *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 77–86.

Gurevych, Iryna, Christian M. Meyer, Carsten Binnig, Johannes Fürnkranz, Kristian Kersting, Stefan Roth, and Edwin Simpson (2018). "Interactive data analytics for the Humanities". In:

Proceedings of the 18th international conference. Lecture Notes in Computer Science (Vol. 10761). Berlin: Springer.

Hammond, Adam, Julian Brooke, and Graeme Hirst (2013). "A tale of two cultures: Bringing literary analysis and computational linguistics together". In: *Proceedings of the NAACL 13 workshop on computational linguistics for literature*. Atlanta, GA, pp. 1–8.

Hirsch Eric Donald, Jr. (1967). *Validity in interpretation*. New Haven: Yale.

Hovy, Eduard and Julia Lavid (2010). "Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics". In: *International Journal of Translation* 22(1), pp. 13–36.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III (2016). "Feuding families and former friends: Unsupervised learning for dynamic fictional relationships". In: *Proceedings of NAACL-HLT 2016*, pp. 1534–1544.

Jannidis, Fotis (2003). "Polyvalenz – Konvention – Autonomie". In: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin: de Gruyter, pp. 3–30.

Jannidis, Fotis (in preparation). *Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets. Version 2*. Stuttgart: Metzler.

Jannidis, Fotis and Gerhard Lauer (2014). "Polyvalenz–Konvention–Autonomie". In: *Distant readings. Topologies of German culture in the long nineteenth century*. Rochester, pp. 29–54.

Jannidis, Fotis, Gerhard Lauer, Matías Martínez, and Simone Winko (2003). "Der Bedeutungsbegriff in der Literaturwissenschaft. Eine historische und systematische Skizze". In: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin: de Gruyter, pp. 3–30.

Jockers, Matthew L. (2013). *Macroanalysis. Digital media and literary history*. Illinois: University of Illinois Press.

Köppe, Tilmann and Tom Kindt (2014). *Erzähltheorie*. Stuttgart: Reclam.

Krippendorff, Klaus (1980). *Content analysis; An introduction to its methodology*. Beverly Hills, CA: Sage.

Kuhn, Jonas (1999). *The challenge for the computational sciences in Digital Humanities: Establishing a common meta-methodological framework*. Website. Leipzig. URL: http://dhd-wp.hab.de/files/book_of_abstracts.pdf (visited on May 9, 2019).

Kuhn, Jonas (2019). "Computational text analysis within the Humanities: How to combine working practices from the contributing fields?" In: *Language Resources & Evaluation* 53, pp. 565–602. DOI: 10.1007/s10579-019-09459-3.

Kuhn, Jonas (in preparation). "Empirie–Beschreibung–Interpretation: über den Platz von Computermodellen in den hermeneutisch-historisch orientierten Literaturwissenschaften". In: *Article based on contribution to the international symposium of the Deutsche Forschungsgemeinschaft (DFG) on digital literary studies, 9–13 October 2017, Villa Vigoni, Italy*. Berlin, pp. 3–30.

Kuhn, Jonas, Artemis Alexiadou, Manuel Braun, Thomas Ertl, Sabine Holtz, Cathleen Kantner, Catrin Misselhorn, Sebastian Padó, Sandra Richter, Achim Stein, and Claus Zittel (2016). "CRETA (Centrum für reflektierte Textanalyse) – Fachübergreifende Methodenentwicklung in den Digital Humanities". In: *Konferenzabstracts zur Konferenz Digital Humanities im deutschsprachigen Raum*. Leipzig, pp. 340–343. URL: http://dhd2016.de/boa.pdf (visited on May 9, 2019).

Kuhn, Jonas and André Blessing (2018). "Die Exploration biographischer Textsammlungen mit computerlinguistischen Werkzeugen–methodische Überlegungen zur Übertragung komplexer Analyseketten in den Digital Humanities". In: *Europa baut auf Biographien: Aspekte,*

Bausteine, Normen und Standards für eine europäische Biographik. Vienna: New Academic Press, FEHLEN.

Kuhn, Jonas and Nils Reiter (2015). "A plea for a method-driven agenda in the Digital Humanities". In: *Global Digital Humanities*. Sydney, FEHLEN.

Kuhn, Thomas S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Mahlow, Cerstin, Kerstin Eckart, Jens Stegmann, André Blessing, Gregor Thiele, Markus Gärtner, and Jonas Kuhn (2014). "Resources, tools, and applications at the CLARIN center Stuttgart". In: *Proceedings of the 12th edition of the KONVENS conference*. Universitätsverlag Hildesheim, pp. 11–21.

Manning, Christopher D. (2011). "Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?" In: *Computational Linguistics and intelligent text processing, 12th international conference, CICLing 2011, proceedings, part I. Lecture notes in computer science (Vol. 6608)*. Berlin: Springer, pp. 171–189.

Mantzavinos, Chrysostomos (2016). *Hermeneutics*. Ed. by Edward N. Zalta. Website. URL: https://plato.stanford.edu/archives/win2016/entries/hermeneutics/ (visited on May 9, 2019).

Moretti, Franco (2007). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.

Newton, Ken M. (1989). "Hermeneutics and modern literary criticism". In: *The British Journal of Aesthetics* 29(2), pp. 116–127.

Overbeck, Maximilian (2018). "Vom Beobachter zum Teilnehmer". In: *Das Narrativ von der Wiederkehr der Religion*. Wiesbaden: Springer, pp. 231–260.

Quan-Haase, Anabel and Kim Martin (2012). "The continuing role of serendipity in historical research". In: *Proceedings of the annual conference of CAIS*.

Reichert, Ramón, ed. (2014). *Big data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*. Bielefeld: Transcript.

Sekine, Satoshi (1997). "The domain dependence of parsing". In: *Proceedings of the 5th conference on applied natural language processing*. Association for Computational Linguistics, pp. 96–102. DOI: 10.3115/974557.974572.

Ter Braake, Serge, Antske Fokkens, Niels Ockeloen, and Chantal M. van Son (2016). "Digital History: Towards New Methodologies". In: *Computational history and data-driven Humanities: Second IFIP WG 12.7 international workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, revised selected papers. IFIP Advances in Information and Communication Technology (Vol. 482)*. Springer, pp. 23–32.

Underwood, Ted (2015). *Understanding genre in a collection of a million volumes*. NEH White Papers.

Wiebe, Janyce M. (1994). "Tracking point of view in narrative". In: *Computational Linguistics* 20(2), pp. 233–287.

Zhou, Jie and Wei Xu (2015). "End-to-end learning of semantic role labeling using recurrent neural networks". In: *Proceedings of the 53rd annual meeting of the Association for Computational linguistics and the 7th international joint conference on natural language processing*, pp. 1127–1137.

Dominik Gerstorfer

Entdecken und Rechtfertigen in den Digital Humanities

Zusammenfassung: In diesem Beitrag werde ich die Unterscheidung zwischen Entdeckungs- und Rechtfertigungszusammenhang diskutieren und für die Methodenreflexion in den digitalen Geisteswissenschaften fruchtbar machen. Ziel ist es, eine adäquate Begrifflichkeit zu entwickeln, mit der die komplexen interdisziplinären und kollaborativen Forschungsprozesse analysiert werden können.

Abstract: In this contribution I will discuss the distinction between context of discovery and context of justification in order to make them fruitful for methodological reflection in the digital humanities. The aim is to develop adequate terminology, with which the complex interdisciplinary and collaborative research processes can be analysed.

1 Einleitung

Dieser Beitrag beschäftigt sich aus wissenschaftstheoretischer Perspektive mit den Elementen des Forschungsprozesses in den Digital Humanities (DH) am Beispiel der im *Center for Reflected Text Analytics* (CRETA) entwickelten reflektierten algorithmischen Textanalyse. Das Ziel ist es, einen begrifflichen Rahmen zu erarbeiten, um die fächerübergreifende Abstimmung der Arbeitsschritte zu erleichtern. Dies soll im Rückgriff auf den wissenschaftstheoretischen Diskurs über Entdecken und Rechtfertigen in den Wissenschaften geschehen. Dabei werden die Begriffe so weit ausdifferenziert, dass die von Pichler und Reiter (2020), in diesem Band ab Seite 43, ausgearbeiteten Arbeitsschritte funktional charakterisiert werden können. Die schematische Einordnung der infrage stehenden Arbeitsschritte kann in der Folge die interdisziplinäre Verständigung erleichtern, da es möglich wird, die jeweiligen Module hinsichtlich ihrer Leistungen und Funktionen zu thematisieren, ohne das Vokabular, die Methodologie und das Domänenwissen der entsprechenden Fachdisziplin replizieren zu müssen. Die funktionale Charakterisierung der Arbeitsschritte kann als Schnittstelle des interdisziplinären Austauschs dienen, wobei die Kompetenzen zur Bewertung der Schritte in den jeweiligen Fachdisziplinen verbleiben können. Somit wird eine genuin arbeitsteilige Vorgehensweise erleichtert, bei der sich nicht alle Beteilig-

Dominik Gerstorfer, Institut für Philosophie, Universität Stuttgart

Open Access. © 2020 Dominik Gerstorfer, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-005

ten das gesamte Expertenwissen der anderen Fachdisziplinen aneignen müssen, sondern sich kompetent über die Schritte zur Erreichung des gemeinsamen Forschungsziels austauschen können.

Eine solche Anwendung wissenschaftstheoretischer Konzepte auf eine konkrete Forschungspraxis ist mit zwei Schwierigkeiten konfrontiert: Zum einen beschäftigt sich die Wissenschaftstheorie primär mit der abgeschlossenen Forschung, d. h. sie analysiert wissenschaftlichen Prozesse und Ergebnisse erst im Nachhinein und trägt nur indirekt dazu bei, Prozesse zu optimieren und zukünftige Forschung zu verbessern. Zum anderen ist das Abstraktionsniveau wissenschaftstheoretischer Untersuchungen, die zumeist größtmögliche Allgemeinheit anstreben, sehr hoch und kann nicht direkt auf spezifische Praktiken angewendet werden. Darum wird es nötig sein, die allgemeinen Konzepte so zu konkretisieren, dass sie im Kontext der reflektierten Textanalyse eingesetzt werden können. Der hier vorgeschlagene Begriffsrahmen nimmt eine vermittelnde Stellung zwischen den arbeitspraktischen Erwägungen des CRETA-*workflows* und den theoretischen Überlegungen der Wissenschaftstheorie ein. So kann eine weitere Reflexionsebene gewonnen werden, um erkenntnistheoretische Bewertungen schon bei der Planung und Durchführung von konkreten Projekten zu berücksichtigen.

In Abschnitt 2 werde ich Ziele und Elemente wissenschaftlicher Forschung im Allgemeinen skizzieren und die speziellen Eigenarten der interdisziplinären Forschung in den Digital Humanities (DH) herausarbeiten, um die Unterscheidung zwischen Entdeckung und Rechtfertigung motivieren, die ich dann in Abschnitt 3 weiter ausführe und diskutiere. In Abschnitt 4 und Abschnitt 5 werde ich die Leitunterscheidung Entdecken/Rechtfertigen weiter ausdifferenzieren, um sie dann in Abschnitt 6 auf einzelne Komponenten des Arbeitsablaufes anzuwenden.

2 Der Forschungsprozess

Allgemein kann der Forschungsprozess als Praxis aufgefasst werden, die das Ziel hat, gehaltvolle und wahre Aussagen über ihren Forschungsgegenstand zu machen.

Diese knappe Charakterisierung bedarf einer Erläuterung: Gehaltvoll heißen Aussagen dann, wenn aus ihnen weitere Aussagen über den Untersuchungsgegenstand abgeleitet werden können. Damit soll sichergestellt werden, dass nicht beliebige Aussagen als wissenschaftliche Erkenntnisse über die Untersuchungsgegenstände akzeptiert werden, insbesondere sollen triviale Tautologien vermieden werden, welche wahr, aber nicht gehaltvoll sind. Wahr heißen Aussagen dann, wenn ihr Gehalt auch tatsächlich zutrifft. Damit soll sichergestellt werden, dass

keine gehaltvollen, aber falschen Aussagen als wissenschaftliche Erkenntnisse akzeptiert werden.[1] Das Kriterium der Wahrheit einer Aussage kann in diesem Zusammenhang sehr weit gefasst werden. Es ist nicht nötig, sich auf dieser Ebene auf eine bestimmte Wahrheitskonzeption festzulegen, abhängig vom Gegenstandsbereich oder der sprachlichen Konventionen der Fachdisziplin können auch Validität, Richtigkeit, Plausibilität oder Korrektheit als Kriterien verwendet werden. Unabhängig von der verwendeten Terminologie ist jedoch der Umstand, dass Wahrheitsansprüche einer Begründung oder Rechtfertigung bedürfen.

Wissenschaftliche Erkenntnisse umfassen den so abgesteckten Bereich der gehaltvollen Wahrheit. Hypothesen sind gehaltvolle Aussagen, die im Forschungsprozess als Kandidaten für wissenschaftliche Erkenntnis aufgestellt werden. Der wissenschaftliche Prozess hat nun die Aufgabe, Hypothesen aufzustellen und zu überprüfen.

Diese formale Bestimmung des wissenschaftlichen Prozesses macht keinerlei Aussagen darüber, wie in der Forschungspraxis Hypothesen aufgestellt und überprüft werden könne oder sollen. Es soll auch nicht suggeriert werden, dass es eine einzige wissenschaftliche Methode gebe.[2] Vielmehr soll die Diskussion wissenschaftstheoretischer Positionen dazu führen, angesichts der faktischen Methodenpluralität in den Wissenschaften, die Elemente und Zusammenhänge der wissenschaftlichen Arbeitsabläufe besser zu verstehen und gegebenenfalls durch weitere Methodenreflexion im Einzelfall zu korrigieren.

Doch wie lässt sich der konkrete Forschungsprozess adäquat fassen? William Whewell macht den Vorschlag, den Prozess und das Produkt der Forschung als wissenschaftliche Entdeckung, *scientific discovery*, folgendermaßen in drei Phasen aufzuteilen (Whewell 1840):

1. Am Anfang steht ein glücklicher Einfall.
2. Dieser Einfall wird artikuliert und zu einer Hypothese ausformuliert.
3. Die Hypothese wird im Fall ihrer erfolgreichen Überprüfung als Erkenntnis anerkannt.

---

1 Zur selben Problemklasse gehören auch überangepasste Modelle, die sehr gute Ergebnisse auf den Trainingsdaten, aber nur schlechte Ergebnisse auf den Testdaten erzielen. Das Modell macht gehaltvolle Aussagen über die Testdaten, die jedoch nicht zutreffen.

2 Die Behauptung, dass die Wissenschaft durch eine einzige, universelle Methode bestimmt wird, hat schon Karl Popper zu deutlichen Worten inspiriert. So soll er am ersten Vorlesungstag in den Hörsaal gekommen sein und gerufen haben: "I am a professor of scientific method, and I have a problem. There is no such thing as scientific method!" (Nickles 2006, S. 177)

Der glückliche Einfall (1) oder *happy thought*, wie ihn Whewell nennt, zeichnet sich durch das Fehlen von Regeln aus:

> Scientific Discovery must ever depend upon some happy thought, of which we cannot trace the origin; some fortunate cast of intellect, rising above all rules. No maxims can be given which inevitably lead to discovery. (Whewell 1840, S. 186)

Es kann ihm zufolge nicht eine einzige Maxime oder Methode geben, die unweigerlich zu einer Entdeckung führt, es gibt keine Anleitung, um *Heureka*-Momente zu erzwingen. Die *happy thoughts* sind nicht determiniert, allerdings sind sie auch nicht beliebig, wie Schickore (2018, Abs. 3) bemerkt:

> In this sense, happy thoughts are accidental. But in an important sense, scientific discoveries are not accidental. The happy thought is not a wild guess. Only the person whose mind is prepared to see things will actually notice them. The "previous condition of the intellect, and not the single fact, is really the main and peculiar cause of the success. The fact is merely the occasion by which the engine of discovery is brought into play sooner or later. It is, as I have elsewhere said, only the spark which discharges a gun already loaded and pointed; and there is little propriety in speaking of such an accident as the cause why the bullet hits its mark." (Whewell 1840, S. 189)

Die *engine of discovery* sollte nicht als Mechanismus verstanden werden, der automatisch Erkenntnisse generiert, sondern als Hintergrundtheorien, Domänenwissen und bekannte Methoden, die – auch in den digitalen Geisteswissenschaften – den Kontext konstituieren, der Einfälle überhaupt möglich macht. Hier kann die CRETA-Werkstatt (vgl. Einleitung dieses Bandes) als Beispiel dienen, die Einfälle und *Heureka*-Momente haben sich oft bei Thementischen und *brainstorming*-Sitzungen eingestellt, wenn Domänenexpert:innen aus den verschiedenen Fachdisziplinen über gemeinsame Themen nachdenken und dabei Hintergrundwissen, Theorien und Methoden miteinander verbinden. Durch solches Vorgehen kann kein Ergebnis (und erst recht kein Ergebnisinhalt) vorausbestimmt werden, jedoch kann der Arbeitsprozess so strukturiert werden, dass die Wahrscheinlichkeit eines Einfalls steigt. Der *happy thought* (1) ist keine mysteriöse Eingebung, die dem wissenschaftlichen Genie zufällt, sondern Teil der Forschungspraxis.

Der Übergang vom *happy thought* zur Erkenntnis (2) ist der für den Arbeitsablauf wichtigste und gleichzeitig am wenigsten bestimmte Schritt. Hier stellt sich die Frage, ob er rational analysierbar ist oder ausschließlich durch erkenntnisexterne Faktoren bestimmt wird. Der Arbeitsablauf der reflektierten Textanalyse ist, wie Kuhn (2020) und Pichler und Reiter (2020) in diesem Band ausgeführt haben, hinreichend systematisch, um eine Aufteilung in Module und Handlungsanweisungen zu ermöglichen.

Die Rechtfertigung in (3), wo die Hypothesen systematisch evaluiert werden, findet in der reflektierten Textanalyse jedoch an verschiedenen Stellen statt, zum einen zeitlich getrennt am Ende des prototypischen Arbeitsablaufes, wenn die Befunde interpretiert werden zum anderen logisch in den jeweiligen Teilmodulen, die entweder ihre eigenen Validierungsverfahren mitbringen oder selbst durch Rückführung auf die Standards der Fachdisziplinen gerechtfertigt sind.

Die Abbildung der reflektierten Textanalyse auf das 3-Phasen-Modell Whewells zeigt auch dessen Schwächen: Zum einen folgt der tatsächliche Arbeitsablauf nicht streng den drei Phasen, zum anderen ist nicht klar, wie genau zwischen Prozess oder Ergebnis differenziert werden kann. Gehört die Rechtfertigung zur Entdeckung oder ist sie unabhängig?

3 Entdeckungs- und Rechtfertigungszusammenhang

Eine Antwort auf die Frage, wie Rechtfertigung und Entdeckung zusammenhängen bietet Hans Reichenbach an, der in *Experience and Prediction* (1938) die begriffliche Unterscheidung zwischen *context of discovery* (Entdeckungs- oder Entstehungszusammenhang) und *context of justification* (Rechtfertigungs- oder Begründungszusammenhang) einführt, um die Geltung wissenschaftlicher Erkenntnisse von ihren sozialen, ökonomischen und psychologischen (und anderen erkenntnisexternen) Entstehungsumständen zu trennen.

Der Entdeckungszusammenhang umfasst die tatsächlichen Denkvorgänge und Prozesse, die zu einer Entdeckung geführt haben, und ist Gegenstand der Psychologie, Soziologie und Geschichtswissenschaft, nur der Rechtfertigungszusammenhang ist nach Reichenbach genuiner Gegenstand der Wissenschaftstheorie, er umfasst die gesamte Bewertung und Überprüfung der wissenschaftlichen Erkenntnisse. Dieser Unterscheidung liegt die Vorannahme zugrunde, dass die Wissenschaftstheorie ein normatives Projekt sei, das formallogisch verfährt und dessen Gegenstände die rationalen Rekonstruktionen wissenschaftlicher Theorien sind. Rationale Rekonstruktion bedeutet, dass die Bewertung der Gültigkeit und Zuverlässigkeit einer Theorie anhand einer fiktionalen Konstruktion durchgeführt wird, bei der die tatsächlichen Denkvorgänge und sprachlichen Äußerungen durch logische Konstruktionen ersetzt werden, da selbst die vollständig ausformulierten wissenschaftlichen Ergebnisse in Veröffentlichungen noch zu unpräzise sind oder noch subjektive Motive enthalten (Reichenbach 1938, S. 6). Hierzu bemerkt Reichenbach:

> If a more convenient determination of this concept of rational reconstruction is wanted, we might say that it corresponds to the form in which thinking processes are communicated to other persons instead of the form in which they are subjectively performed. [The] well-known difference between the thinker's way of finding [a] theorem and his way of presenting it before a public may illustrate the difference in question. I shall introduce the terms *context of discovery* and *context of justification* to mark this distinction. Then we have to say that epistemology is only occupied in constructing the context of justification. But even the way of presenting scientific theories is only an approximation to what we mean by the context of justification. Even in the written form scientific expositions do not always correspond to the exigencies of logic or suppress the traces of subjective motivation from which they started. (Reichenbach 1938, S. 6–7)

Das Problem, das durch die Unterscheidung von Entdeckungs- und Rechtfertigungszusammenhang primär gelöst werden soll, zeigt sich am Beispiel August Kekulés deutlich, der die Struktur des Benzolrings im Traum erkannt haben will:

> Ich drehte den Stuhl nach dem Kamin und versank in Halbschlaf. Wieder gaukelten die Atome vor meinen Augen. Kleinere Gruppen hielten sich diesmal bescheiden im Hintergrund. Mein geistiges Auge, durch wiederholte Gesichte ähnlicher Art geschärft, unterschied jetzt grössere Gebilde von mannigfacher Gestaltung. Lange Reihen, vielfach dichter zusammengefügt; Alles in Bewegung, schlangenartig sich windend und drehend. Und siehe, was war das? Eine der Schlangen erfasste den eigenen Schwanz und höhnisch wirbelte das Gebilde vor meinen Augen. Wie durch einen Blitzstrahl erwachte ich; auch diesmal verbrachte ich den Rest der Nacht um die Consequenzen der Hypothese auszuarbeiten. (August Kekulé nach Schultz 1890, S. 1306)

Auch wenn die Entdeckungen in den digitalen Geisteswissenschaften selten im Traum gemacht werden, kommt die Unterscheidung zwischen Entdeckung und Rechtfertigung auch hier zum Tragen: Die Ergebnisse komputationeller Werkzeuge können nicht für sich allein stehen, sondern müssen in eine Argumentation eingebunden sein und interpretiert werden. Dieser Aspekt der Rechtfertigung findet sich im Schema der reflektierten Textanalyse als „Interpretation der Befunde" (Pichler und Reiter 2020, S. 54). Da die Interpretation der Befunde Teil des Arbeitsablaufes ist und nicht nach der Formulierung der Endergebnisse einsetzt, ist es nötig, genauer auf das Verhältnis zwischen Entdeckungs- und Rechtfertigungszusammenhang einzugehen.

In der „orthodoxen" (Feigl 1970) Lesart,[3] welche die wissenschaftstheoretischen Diskussionen lange geprägt hat, werden Entdeckung und Rechtfertigung als zeitlich und logisch streng getrennt aufgefasst: Etwas muss zuerst entdeckt werden, bevor es gerechtfertigt werden kann. Dabei spielen Logik und normati-

3 Wie Reichenbach selbst die Kontextunterscheidung verstanden hat, ist Gegenstand einer eigenen Debatte, vgl. Nickles 1980 und Schiemann (2006).

ve Analysen ausschließlich im Rechtfertigungszusammenhang eine Rolle. Diese Lesart ist mit vier Schwierigkeiten konfrontiert:

1. Wissenschaftliche Prozesse laufen in der Regel nicht in genau zwei Schritten ab: Häufig alterieren Teilentdeckungen und Teilrechtfertigungen (Hoyningen-Huene 2006, S. 120–121).
2. Die beiden Phasen lassen sich nicht immer trennscharf unterscheiden: Neue Messmethoden, die die Verbesserung eines empirischen Gesetztes zur Folge haben, können sowohl als Entdeckung betrachtet werden, als auch als Beitrag zur Rechtfertigung des Gesetzes (Hoyningen-Huene 2006, S. 120–121; Arabatzis 2006, S. 216).
3. Entdeckungen werden zumeist nicht zufällig gemacht, sondern folgen logischen und methodologischen Überlegungen und Verfahren (Nickles 1980, S. 14).
4. Der Begriff der Entdeckung selbst ist schon normativ aufgeladen, denn nicht jede Behauptung ist eine Entdeckung, sie wird es erst, wenn sie durch die Forschungsgemeinschaft als solche anerkannt wird (Arabatzis 2006, S. 217).

Die statische Auffassung, dass die Entdeckung zeitlich vor der Rechtfertigung einsetzt, muss so erweitert werden, dass komplexere Arbeitsweisen berücksichtigt werden können. Für die reflektierte Textanalyse bedeutet das, dass Entdeckung und Rechtfertigung auf Modulebene relativ zu den beteiligten Fachdisziplinen angesetzt werden. Jedes Modul, d. h. jeder Arbeitsschritt, der ein Ergebnis produziert, muss selbst gerechtfertigt sein, um zu gewährleisten, dass das Endergebnis wahre, gehaltvolle Aussagen über den Untersuchungsgegenstand enthält.

In den nächsten beiden Abschnitten werden nun die systematischen Aspekte von Entdeckung (Abschnitt 4) und Rechtfertigung (Abschnitt 5) weiter ausdifferenziert, um die Arbeitsabläufe der reflektierten Textanalyse adäquat abbilden zu können.

4 Entdeckungszusammenhang

Die Unterscheidung zwischen Entdeckungs- und Rechtfertigungszusammenhang findet eine Entsprechung in der Formulierung der hypothetisch-deduktiven (H-D)

Methode, wobei die Formulierung von Hypothesen[4] (H) dem Entdeckungszusammenhang und die deduktive Überprüfung (D) dem Rechtfertigungszusammenhang zugeordnet werden kann. Bei der deduktive Überprüfung der Hypothesen werden logische Konsequenzen deduktiv abgeleitet und mit Beobachtungen verglichen. Dieses Verfahren findet sich zum Beispiel bei der Evaluation formaler Modelle, etwa bei der Bestimmung von *precision*- und *recall*-Metriken.

In der wissenschaftstheoretischen Untersuchung wird zumeist die Hypothesenbildung ausgeklammert, da es für die deduktive Überprüfung irrelevant ist, wie die zu prüfende Hypothese gefunden wurde. Bei der Analyse wissenschaftlicher Arbeitsabläufe ist die Art und Weise, wie Hypothesen aufgestellt werden, jedoch hochgradig relevant.

Einen Schritt hin zu einer Logik der Entdeckung geht Norwood Hansons in *The Logic of Discovery*:

> H-D [hypothetico-deductive] accounts all agree that physical laws explain data, but they obscure the initial connexion between data and laws; indeed, they suggest that the fundamental inference is from higher-order hypotheses to observation statements. This may be a way of setting out one's reasons for accepting an hypothesis after it is got, or for making a prediction, but it is not a way of setting out reasons for proposing or for trying an hypothesis in the first place. Yet the initial suggestion of an hypothesis is very often a reasonable affair. It is not so often affected by intuition, insight, hunches, or other imponderables as biographers or scientists suggest. Disciples of the H-D account often dismiss the dawning of an hypothesis as being of psychological interest only, or else claim it to be the province solely of genius and not of logic. They are wrong. If establishing an hypothesis through its predictions has a logic, so has the conceiving of an hypothesis. (Hanson 1958, S. 71)

Das vorrangige Interesse an einer Entdeckungslogik besteht darin, den Entdeckungsprozess durch reliable Arbeitsabläufe so zu gestalten, dass möglichst richtige Hypothesen generiert werden. Doch auch wenn die Möglichkeit einer Entdeckungslogik eingeräumt wird, könnte bestritten werden, dass sie relevant für den Erkenntnisprozess ist, da die aufgestellten Hypothesen zusätzlich noch gerechtfertigt werden müssen, gleichgültig durch welche Prozedur sie generiert wurden.

Dieser Einwand kann jedoch entkräftet werden (Kelly 1987): Eine Prämisse des Einwands besagt, dass die Prozeduren, die Hypothesen aufstellen, und Prozeduren, die Hypothesen testen, unabhängig sind. Darum ist es notwendig, jede Hypothese zu testen, egal wie sie zustande gekommen ist. Gegeneinwand: Man kann zeigen, dass Überprüfung und Erzeugung der Hypothesen symmetrisch

4 Hypothese soll sehr weit als Kandidat für eine Erkenntnis verstanden werden, d. h. als Aussage, deren Richtigkeit unterstellt, aber noch nicht gezeigt ist.

sind, denn jede Testprozedur kann in einen Generator transformiert werden. Kelly (1987, S. 440) zeigt das an einem trivialen Beispiel: Gegeben ist eine Relation $\varphi(x,y)$, die $y=x^2$ berechnet. Der Hypothesengenerator könnte einfach für jedes x die Reihe der natürlichen Zahlen durchlaufen, während die Testprozedur die falschen Werte aussondern und die richtigen Werte behalten würde. Mit anderen Worten, die Testprozedur berechnet intern $y=x^2$, d. h. sie benutzt einen effizienten Generator als Subroutine. Dieses Beispiel zeigt, dass die strikte Unterscheidung nicht prinzipiell aufrechterhalten werden kann.

Ein noch allgemeineres Argument für die Möglichkeit einer Entdeckungslogik entwickelt Jantzen (2015): Angenommen, eine Entdeckungslogik wäre unmöglich, dann wäre der Prozess der Hypothesenbildung äquivalent mit einem vollständig zufälligen Prozess. Wird eine Hypothese zufällig aus der unendlichen Menge aller möglichen Hypothesen ausgewählt, dann tendiert die relative Wahrscheinlichkeit, die richtige Hypothese zu ziehen, gegen null. Die Wissenschaftsgeschichte ist jedoch voll von erfolgreichen Beispielen, in denen nach nur sehr wenigen Versuchen die richtige Hypothese gefunden wurde. Nur ein Wunder könnte die Tatsache erklären, dass die Wissenschaften empirisch so erfolgreich sind. Wunder sind als Erklärungsgründe ausgeschlossen. Daraus folgt, dass es eine Entdeckungslogik geben muss.

Diese beiden Argumente betonen die Relevanz der logischen Untersuchung des Entdeckungszusammenhangs hinsichtlich effizienter Methoden, Hypothesen zu generieren. Dennoch sollte die Möglichkeit einer Entdeckungslogik nicht so missverstanden werden, dass sie die Möglichkeit einer deterministischen Entdeckungsmaschine (Curd 1980, S. 207) implizieren würde. Also einer Maschine, die unweigerlich wahre und gehaltvolle wissenschaftliche Hypothesen generieren könnte. Stattdessen soll lediglich gezeigt werden, dass es Prozeduren, Algorithmen und Heuristiken der Theoriekonstruktion und Hypothesenbildung gibt, die systematisch untersucht werden können und so zu einer Methodenreflexion beitragen.

Doch was soll unter Entdeckungslogik verstanden werden? Die klassische deduktive Logik kommt offensichtlich nicht infrage, da deduktive Kalküle ausschließlich wahrheitserhaltend von den Prämissen unter Zuhilfenahme von Schlussregeln zu den Konklusionen übergehen, d. h. in den Konklusionen findet sich nichts, was nicht schon in den Prämissen gewesen wäre. Es wird nichts Neues hinzugefügt. Entdeckungen sind jedoch per Definition neu und zuvor unbekannt. Um Entdeckungen zu modellieren, werden also logische Schlussmuster benötigt, die den Gehalt der Prämissen inferenziell erweitern können.

Solche Schlussmuster finden sich bei Charles Sanders Peirce, der (1) Deduktion, (2) Induktion und (3) Abduktion unterscheidet. Peirce erläutert die drei Schlussmuster anhand des folgenden Beispiels (Peirce 1931–1958, CP 2.623): In

einem Raum befinden sich auf einem Tisch eine Handvoll weißer Bohnen und in einer Ecke ein Beutel mit ausschließlich weißen Bohnen. Mithilfe der Bohnen (x) und der Prädikate „stammt aus dem Beutel“ (B) und „ist weiß“ (W) können nun alle drei Schlussmuster gebildet werden.

1. Im Fall der Deduktion ist bekannt, dass alle Bohnen im Beutel weiß sind und dass die Bohnen auf dem Tisch aus dem Beutel stammen. Es kann mit Sicherheit geschlossen werden: Diese Bohnen sind weiß. (Der Schluss ist gewiss, wahrheitserhaltend und nicht erkenntniserweiternd.)

$$\begin{array}{lll} 1. & \forall x(Bx \rightarrow Wx) & \text{Regel} \\ 2. & Bx & \text{Fall} \\ \hline \therefore & Wx & \text{Resultat} \end{array} \tag{1}$$

2. Im Fall der Induktion ist bekannt, dass die Bohnen aus dem Beutel stammen und dass sie weiß sind. Es kann die Regel erschlossen werden, dass alle Bohnen im Beutel weiß sind. (Der Schluss ist erkenntniserweiternd, ungewiss und nicht wahrheitserhaltend.)

$$\begin{array}{lll} 1. & Bx & \text{Fall} \\ 2. & Wx & \text{Resultat} \\ \hline \therefore & \forall x(Bx \rightarrow Wx) & \text{Regel} \end{array} \tag{2}$$

3. Im Fall der Abduktion ist bekannt, dass alle Bohnen im Beutel weiß sind und dass die Bohnen auf dem Tisch weiss sind. Es kann auf die Hypothese geschlossen werden: Die Bohnen stammen aus dem Beutel. (Der Schluss ist erkenntniserweiternd, ungewiss und nicht wahrheitserhaltend.)

$$\begin{array}{lll} 1. & \forall x(Bx \rightarrow Wx) & \text{Regel} \\ 2. & Wx & \text{Resultat} \\ \hline \therefore & Bx & \text{Fall} \end{array} \tag{3}$$

Oft werden die beiden erkenntniserweiternden, unsicheren und nicht wahrheitserhaltenden Schlussmuster, Induktion und Abduktion, zusammengefasst, wobei die Abduktion als Sonderfall der Induktion gesehen wird. Es gibt jedoch einen wichtigen Unterschied zwischen beiden Schlussmustern hinsichtlich des Ziels. Die Induktion zielt auf zukünftig beobachtbare Ereignisse ab, während die Abduktion unbeobachtete Ursachen oder explanatorische Gründe zum Ziel hat (Schurz 2008, S. 202).

Mit Magnani (2001, S. 20) können zwei Typen der Abduktion unterschieden werden: Selektive und kreative Abduktion. Die selektive Abduktion wählt aus

einer gegebenen Menge von Hypothesen die beste Erklärung aus, während die kreative Abduktion neue Begriffe, Hypothesen oder Modelle erzeugt. Das Beispiel der Krankheitsdiagnose verdeutlicht den Unterschied: Die selektive Abduktion geht von den beobachteten Symptomen zur Diagnose, indem aus einer Liste mit Krankheitsbildern dasjenige ausgewählt wird, das den Symptomen am besten entspricht. Kreative Abduktion erzeugt hingegen eine neue Krankheitsdefinition aus dem beobachteten Krankheitsbild. Selektive Abduktion liefert immer minimal plausible Hypothesen, die dann durch weitere deduktiven Tests überprüft werden können.

In diesem Sinne kann selektive Abduktion als Schluss auf die beste Erklärung (SBE) verstanden werden, die sich jedoch von der ursprünglichen Konzeption Harmans (1965) hinsichtlich ihrer entdeckenden Funktion unterscheidet, denn Harmans SBE ist nur dann erfolgreich, wenn alle möglichen Alternativen zur Verfügung stehen. Da das jedoch nicht der Fall ist, kann nur auf die beste *verfügbare* Erklärung geschlossen werden, dieser Schluss auf die beste verfügbare Erklärung (SBVE) unterliegt folgender Einschränkung: Es ist nicht gewährleistet, dass die beste verfügbare Erklärung auch eine gute, d. h. zutreffende, Erklärung ist. Gerade bei neuen oder schlecht verstandenen Phänomenen ist die beste Erklärung oft reine Spekulation.[5]

Eine Lösung dieses Problems besteht darin, die entdeckende Rolle der Abduktion genauer zu betrachten. Jedes Schlussmuster hat eine rechtfertigende und eine entdeckende Funktion. Die rechtfertigende Funktion der Deduktion ist maximal, während ihre rechtfertigende Funktion minimal ist, d. h., sie liefert keine neuen Erkenntnisse, sondern nur maximale Gewissheit über die Wahrheit der Konklusion, da diese nichts enthält, was nicht schon in den Prämissen gewesen wäre. Bei der Abduktion verhält es sich genau umgekehrt, die Gewissheit der von ihr generierten Hypothesen ist minimal[6], während ihre entdeckende Funktion maximal ist. In diesem Zusammenhang wäre es jedoch falsch, die minimale Rechtfertigungsfunktion der Abduktion so zu verstehen, dass sie gar keine begründende Kraft hätte, denn in der Praxis werden meistens nach erstaunlich wenigen Versuchen befriedigende Lösungen für ein Problem gefunden, was nahelegt, dass die abduktiven Suchprozesse sogar (relativ) effizient sind. Versuchte man, alle möglichen Erklärungen komputationell zu finden, würde der Suchraum exponentiell anwachsen, darum ist es notwendig, die Suche so einzuschränken, dass Lösun-

5 Schurz (2008, S. 203) führt hier intentionales Handeln übernatürlicher Wesen zur Erklärung von Naturphänomenen als Beispiel an.

6 Peirce (1903, CP 5.171) hält die abduktiv gewonnenen Hypothesen nicht einmal für wahrscheinlich, sondern lediglich für möglich.

gen in kurzer Zeit gefunden werden können.[7] Die effiziente Einschränkung des Suchraums kann nun wiederum als schwache Rechtfertigung für die gewonnenen Ergebnisse gelten (vgl. Schurz 2008, S. 204).

Die logischen Schlussmuster Deduktion, Induktion sowie selektive und kreative Abduktion erlauben eine differenzierte Beurteilung der Module und Teilschritte der reflektierten Textanalyse und können sowohl im Design also auch in der Umsetzung von konkreten Forschungsprojekten eingesetzt werden, um die inferenzielle Stringenz des Arbeitsablaufes zu bewerten.[8]

5 Rechtfertigungszusammenhang

Der Rechtfertigungszusammenhang enthält bei Reichenbach nur die epistemische Validierung der im Entdeckungszusammenhang gebildeten Hypothesen. Doch bei konkreten Arbeitsabläufen zeigt sich, dass nicht alle normativen Bewertungen epistemisch sind.

Im Forschungsprozess sind ständig Entscheidungen zu treffen und Evaluationen vorzunehmen. Doch nicht alle diese Evaluationen und Entscheidungen haben das Ziel, die Wahrheit einer Hypothese zu stützen. Eine nützliche Differenzierung des Rechtfertigungszusammenhangs nimmt Nickles (2006) vor, der darauf hinweist, dass nicht alle Evaluationen *epistemische* Rechtfertigungen sind. Er schlägt vor, zwei Bewertungstypen zu unterscheiden: (1) epistemische und (2) heuristische Bewertungen.

Epistemische Bewertungen betreffen alle Fragen nach der Wahrheit von Behauptungen. Heuristische Bewertungen hingegen sind Evaluationen und Entscheidungen, die aus pragmatischen oder forschungsökonomischen Gründen getroffen werden. Die meisten Forschungsentscheidungen werden nicht allein aus epistemischen Gründen getroffen, da es immer auch nötig ist, die Fruchtbarkeit und Durchführbarkeit der Forschungsansätze zu bedenken.

Bei der Begutachtung von Forschungsprojekten oder Antragsstellung steht die Wahrheit einer Hypothese selten im Vordergrund der Überlegungen (auch oder gerade, weil die Forschungs*ergebnisse* noch nicht vorliegen). Vielmehr wird die Durchführbarkeit des Forschungsvorhabens, die Qualität des Forschungsdesigns oder die Wahrscheinlichkeit, dass der *principle investigator* das Projekt zu

7 Schurz (2008, S. 211) schlägt hier technische Lösungen wie Bestensuchen und Tableaukalküle als aussichtsreiche Kandidaten vor.

8 Zur Rolle der Abduktion als Heuristik in DH-Projekten siehe Gius und Jacke 2015 und Gius 2019.

Ende bringen kann, beachtet. Es wird gefragt: Wie interessant oder wichtig ist die Fragestellung? Ist sie anschlussfähig? Sind Folgeprojekte möglich?

Heuristische Bewertungen beinhalten Überlegungen, die von epistemischen Beurteilungen entweder unabhängig sind oder vom Standpunkt zukünftiger Fruchtbarkeit her urteilen. Einstein verwendete zum Beispiel in seinem Aufsatz *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt* (1905) die Formel für die ‚schwarze Strahlung' von Wilhelm Wien, obwohl sie schon damals als nicht gültig verworfen war. Allerdings war sie für Einstein höchst fruchtbar, denn er konnte sie benutzen, um seine eigene Argumentation voranzubringen (Nickles 2006, S. 162). Und umgekehrt verlassen Wissenschaftler ihre Forschungsgebiete, obwohl sich die dort erlangten Erkenntnisse nicht als falsch erwiesen haben und obwohl es noch offene Probleme gibt, weil sie größere zukünftige Erfolge in anderen Gebieten erwarten. Max Delbrück kehrte beispielsweise der Physik den Rücken zu und betätigte sich in der Biologie, nicht etwa, weil die Physik fehlgeschlagen wäre oder es keine offenen Probleme mehr gab, sondern weil er die Möglichkeit, selbst fruchtbar zu arbeiten, dort ungünstiger einschätzte als in der Biologie (Nickles 2006, S. 162).

In der Planung und Organisation spielt die Wahl des nächsten Forschungsvorhabens eine kritische Rolle. Welches der vielen offenen Probleme, welche der vielen interessanten Fragen soll als nächstes bearbeitet werden? Hier hilft die rein epistemische Maxime ‚Finde die Wahrheit!' nicht weiter. Hier helfen nur heuristische Überlegungen (Nickles 2006, S. 163).

Gerade in multidisziplinären Kooperationen, die die digitalen Geisteswissenschaften prägen, ist der Bereich der Heuristik von großer Bedeutung, unter anderem bei der Wahl der Korpora und der komputationellen Werkzeuge.[9]

6 Anwendung

Der im CRETA-Projekt entwickelte prototypische Arbeitsablauf versteht sich vorrangig als methodisch-arbeitspraktische Herangehensweise, um die Probleme mehrdimensionaler, interdisziplinärer Forschungsvorhaben zu bewältigen. Sie beansprucht dabei nicht, vollendete Theorie oder letztgültige Methodologie zu sein, sondern versucht vielmehr, anhand eines differenzierten, prototypischen Arbeitsablaufes eine reflektierte wissenschaftliche Praxis zu entwickeln. Dementsprechend sollen die oben angestellten wissenschaftstheoretischen Überlegungen nicht spezifische Ergebnisse oder Entdeckungen thematisieren, sondern

9 Zur Rolle von Heuristiken in den Digital Humanities vgl. Gius und Jacke 2015.

einen begriffliche Rahmen zur Verfügung stellen, um interdisziplinäre Abstimmungsprozesse zu erleichtern.

Eine Besonderheit des CRETA-*workflows* ist die Modularisierung des Arbeitsablaufs, welche als Zerlegung in eindeutige Handlungsvorschriften konzipiert ist (Pichler und Reiter 2020, S. 43). Der Forschungsprozess wird also normativ strukturiert, um die Ziele der jeweiligen konkreten DH-Projektarbeit zu erreichen. Es geht aus wissenschaftstheoretischer Perspektive darum, die einzelnen Module so zu verzahnen, dass eine lückenlose Rechtfertigungskette hergestellt wird, die von den neuen Entdeckungen zurück zu bereits akzeptierten Erkenntnissen in den beteiligten Fachdisziplinen reicht.

So zeigt zum Beispiel Trevor Owens (2012) anhand einer Diskussion über *topic modeling*, dass die Bedingungen, unter denen ein technisches Instrument Daten produziert hat, notwendige Voraussetzungen sind, um beurteilen zu können, inwieweit jene Daten als Mittel zur Rechtfertigung einer These oder der Verteidigung einer Interpretation dienen können. Er bemerkt in diesem Zusammenhang:

> When we separate out the context of discovery and exploration from the context of justification we end up clarifying the terms of our conversation. There is a huge difference between "here is an interesting way of thinking about this" and "This evidence supports this claim." Both scientists and humanists make both of these kinds of assertions. (Owens 2012)

Die Unterscheidung zwischen Entdeckung und Rechtfertigung kann dabei helfen, zu klären, was ein jeder Einzelschritt zum gesamten Forschungsprozess beiträgt. Liefert er etwas Neues? Handelt es sich um eine Entdeckung? Oder hilft er dabei, eine gemachte Entdeckung abzusichern? Liefert er Gründe für die Gültigkeit einer bestimmten Aussage?

Die oben gewonnenen Begriffe können nun für die Methodenreflexion im CRETA-Projekt angewendet werden. Der prototypische Arbeitsablauf der reflektierten Textanalyse kann auf das obige Modell wie folgt abgebildet werden:

1. Die Aufteilung der *disziplinären Fragestellung* in Teilfragen und Module wird von heuristischen Bewertungen geleitet: In erster Linie geht es darum, herauszufinden, welche Probleme mit den verfügbaren Ressourcen bearbeitet werden können und welche Fragestellungen fruchtbar erscheinen. Ausschlaggebend ist unter anderem, welche Mittel zur Verfügung stehen (sowohl Techniken als auch Mitarbeiter), welche Themen gerade in der wissenschaftlichen Gemeinschaft diskutiert werden und welche Interessen die Kooperationspartner haben. Ein großes Projekt mit vielen Hilfskräften, die Annotationen machen können, wird andere Fragen auswählen als ein kleines Projekt mit einem neuen Visualisierungstool.

2. Die *Operationalisierung* der Begriffe erfolgt zunächst abduktiv: Ein Begriff ist denn operationalisiert, wenn eine Menge an Indikatoren (z. B. Textoberflächeneigenschaften) und Regeln gefunden wurde, die mit hinreichender Wahrscheinlichkeit auf die korrekte Verwendung des Begriffs hinweisen. Immer dann, wenn die Indikatoren und ihre regelhaften Verknüpfungen am Text aufgewiesen werden können, ist auch die Verwendung des Begriffs adäquat.
3. Ein Kernstück des prototypischen Arbeitsablaufes ist die Konstruktion der Korpora und die Erstellung der *Annotationsrichtlinien*: Die Erstellung der Annotationsrichtlinen kann als Schluss auf die beste Erklärung verstanden werden. Die Wahl der Texte und Annotationsrichtlinien folgt heuristischen Bewertungen, die auf den zukünftigen Erfolg der eingesetzten Mittel zielen. Die iterative Ausarbeitung und Stabilisierung der Annotationen erfolgt im Wechsel von abduktiver oder induktiver Regelaufstellung und deduktiver Prüfung der Regeln. Annotationen können als Inferenzen über den Text verstanden werden, die durch abduktive Schlüsse auf Basis der Annotationsrichtlinien erzeugt werden. Die Revision der Richtlinien selbst ist wiederum eine Instanz der selektiven Abduktion: Es wird aus den vorliegenden Annotationen diejenige ausgewählt, welche die beste Erklärung für die gegebenen Textphänomene liefern kann. Entscheidungsgründe der Annotator:innen und quantitative Metriken liefern anschließend die deduktive Rechtfertigung der jeweiligen Hypothese.

Jeder der Teilschritte ist entweder unabhängig gerechtfertigt oder als Schluss auf die beste Erklärung abgesichert. Die Aufteilung des Arbeitsablaufes in einzelne Komponenten kann somit garantieren, dass jeder Schritt, der einen weiteren rechtfertigt, selbst gerechtfertigt ist.

7 Fazit

Die digitalen Geisteswissenschaften haben sich vom romantischen Bild des solitären Genius weit entfernt. Die Arbeitsabläufe ähneln mehr denen der Naturwissenschaften, die in Laborsituationen arbeitsteilig verfahren, nur unter den verschärften Bedingungen der Interdisziplinarität. Die Verschärfung besteht in der Hauptsache darin, dass sich Erkenntnisziele, Methoden und Vokabulare oft signifikant in den Teildisziplinen unterscheiden. Um die methodische Verständigung zu erleichtern, wurde eine Reihe von Begriffsunterscheidungen entwickelt, die die

Reflexion der eigenen Praxis unter dem Vorzeichen der Wissenschaftstheorie ermöglichen.

Mithilfe der Unterscheidung zwischen Entdeckungs- und Rechtfertigungszusammenhang, kreativer und selektiver Abduktion, Schluss auf die beste Erklärung sowie heuristischer und epistemischer Bewertung ist es möglich, praktische Fragen und Probleme auf einer theoretischen Ebene zu thematisieren.

Literatur

Arabatzis, Theodore (2006). „On the Inextricability of the Context of Discovery and the Context of Justification". In: *Revisiting discovery and Justification. Historical and philosophical perspectives on the context distinction*. Hrsg. von Jutta Schickore und Friedrich Steinle. Dordrecht: Springer, S. 215–230.

Curd, Martin (1980). „The Logic of Discovery: An Analysis of Three Approaches". In: *Scientific Discovery, Logic, and Rationality*. Hrsg. von Thomas Nickles. Dordrecht: Reidel, S. 201–219.

Einstein, Albert (1905). „Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt". In: *Ann. Phys.* 322.6, S. 132–148. DOI: 10.1002/andp.19053220607.

Feigl, Herbert (1970). „The ‚orthodox' view of theories: Remarks in defense as well as critique". In: *Analyses of Theories and Methods of Physics and Psychology*. Hrsg. von Michael Radneri und Stephen Winokur. Bd. 4. Minnesota Studies in the Philosophy of Science. Minneapolis, MN: University of Minnesota Press, S. 3–16.

Gius, Evelyn (2019). „Computationelle Textanalysen als fünfdimensionales Problem: Ein Modell zur Beschreibung von Komplexität". Pamphlet 8. Digital Humanities Cooperation. URL: https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/12/pamphlet_gius_2.0.pdf (besucht am 1. Juni 2020).

Gius, Evelyn und Janina Jacke (2015). „Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse". In: *Zeitschrift für digitale Geisteswissenschafteni* 1: *Sonderband der Zeitschrift für digitale Geisteswissenschaften*. Hrsg. von Constanze Baum und Thomas Stäcker. DOI: 10.17175/sb001_006.

Hanson, Norwood Russell (1958). „The Logic of Discovery". In: *Journal of Philosophy* 55.25, S. 1073–1089.

Harman, Gilbert (1965). „The Inference to the Best Explanation". In: *The Philosophical Review* 74.1, S. 88–95.

Hoyningen-Huene, Paul (2006). „Context of Discovery versus Context of Justification and Thomas Kuhn". In: *Revisiting discovery and Justification. Historical and philosophical perspectives on the context distinction*. Hrsg. von Jutta Schickore und Friedrich Steinle. Dordrecht: Springer, S. 119–132.

Jantzen, Benjamin C. (2015). „Discovery without a ‚logic' would be a miracle". In: *Synthese* 193.10, S. 3209–3238.

Kelly, Kevin (1987). „The Logic of Discovery". In: *Philosophy of Science* 54.3, S. 435–452.

Kuhn, Jonas (2020). „Einleitung". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 9–40.

Magnani, Lorenzo (2001). *Abduction, Reason and Science. Processes of Discovery and Explanation*. New York: Kluwer.

Nickles, Thomas (1980). „Introductory Essay: Scientific Discovery and the Future of Philosophy of Science". In: *Scientific Discovery, Logic, and Rationality*. Hrsg. von Thomas Nickles. Dordrecht: Reidel, S. 1–59.

Nickles, Thomas (2006). *Heuristic Appraisal: Context of Discovery or Justification? Historical and philosophical perspectives on the context distinction*. Hrsg. von Jutta Schickore und Friedrich Steinle. Dordrecht.

Owens, Trevor (2012). *Discovery and Justification are Different: Notes on Science-ing the Humanities*. URL: http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/ (besucht am 1. Juni 2020).

Peirce, Charles Sanders (1903). „Lectures on Pragmatism". In: *Collected Papers*. Hrsg. von Charles Hartshorne, Paul Weiss und Arthur Burks. 8 Bde. Cambridge, MA: Harvard University Press, S. 5.14–5.212.

Peirce, Charles Sanders (1931–1958). *Collected Papers*. Hrsg. von Charles Hartshorne, Paul Weiss und Arthur Burks. 8 Bde. Cambridge, MA: Harvard University Press.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Reichenbach, Hans (1938). *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*. Chicago: Chicago University Press.

Schickore, Jutta (2018). „Scientific Discovery". In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2018/entries/scientific-discovery/ (besucht am 1. Juni 2020).

Schiemann, Gregor (2006). „Inductive Justification and Discovery. On Hans Reichenbach's Foundation of the Autonomy of the Philosophy of Science". In: *Revisiting discovery and Justification. Historical and philosophical perspectives on the context distinction*. Hrsg. von Jutta Schickore und Friedrich Steinle. Dordrecht: Springer, S. 23–40.

Schultz, Gustav (1890). „Feier der Deutschen Chemischen Gesellschaft zu Ehren August Kekulé's". In: *Berichte der deutschen chemischen Gesellschaft* 23.1, S. 1265–1312.

Schurz, Gerhard (2008). „Patterns of abduction". In: *Synthese* 164.2, S. 201–234.

Whewell, William (1840). *The Philosophy of the Inductive Sciences founded upon their history*. Bd. 2. London: John Parker.

Janis Pagel, Nils Reiter, Ina Rösiger und Sarah Schulz

Annotation als flexibel einsetzbare Methode

Zusammenfassung: In der Computerlinguistik (CL) wird Annotation mit dem Ziel eingesetzt, Daten als Grundlage für maschinelle Lernansätze und Automatisierung zu sammeln. Gleichzeitig nutzen Geisteswissenschaftler Annotation in Form von Notizen beim Lesen von Texten. Wir behaupten, dass mit der Entwicklung der Digital Humanities (DH) die Annotation zu einer Methode geworden ist, die als Mittel zur Unterstützung der Interpretation und Entwicklung von Theorien eingesetzt werden kann. In diesem Beitrag zeigen wir, wie diese verschiedenen Annotationsziele in einem einheitlichen Workflow abgebildet werden können. Wir reflektieren die Komponenten dieses Workflows und geben Beispiele, wie Annotation im Rahmen von DH-Projekten einen Mehrwert schaffen kann.

Abstract: In computational linguistics (CL), annotation is used with the goal of compiling data as the basis for machine learning approaches and automation. At the same time, in the Humanities scholars use annotation in the form of notetaking while reading texts. We claim that with the development of Digital Humanities (DH), annotation has become a method that can be utilized as a means to support interpretation and develop theories. In this paper, we show how these different annotation goals can be modeled in a unified workflow. We reflect on the components of this workflow and give examples for how annotation can contribute additional value in the context of DH projects.

1 Einführung

Unter Annotation verstehen wir eine Methode, um Textdaten mit zusätzlichen Daten anzureichern. Diese zusätzlichen Daten sind an mehr oder weniger klar abgegrenzte Textstellen gebunden. Im Sprachgebrauch changiert der Begriff ‚Annotation' zwischen einer Bezeichnung für einen Prozess und dessen Ergebnis. In diesem Artikel liegt unser Fokus auf der Methode, und nicht auf den erstellten Annotationen als Datenobjekt oder Analysegegenstand. Wir konzentrieren uns zu-

Anmerkung: Dieser Beitrag ist eine übersetzte und überarbeitete Fassung von: Janis Pagel, Nils Reiter, Ina Rösiger und Sarah Schulz (2018). „A Unified Text Annotation Workflow for Diverse Goals". In: *Proceedings of the Workshop for Annotation in Digital Humanities (annDH)*. Sofia, Bulgarien, S. 31–36. URL: http://ceur-ws.org/Vol-2155/pagel.pdf (besucht am 1. Juni 2020)

Janis Pagel, Nils Reiter, Ina Rösiger, Sarah Schulz, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Janis Pagel, Nils Reiter, Ina Rösiger und Sarah Schulz; publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-006

dem auf Annotationsaufgaben, die interpretative oder assoziative Aspekte haben (d. h., sich auf den expliziten oder impliziten Inhalt eines Textes beziehen).[1]

Annotationsprojekte in der Computerlinguistik (CL) haben eine große Anzahl an Korpora hervorgebracht, die mit linguistischen Konzepten annotiert sind (z. B. Wortarten oder semantische Rollen). Computerlinguistische Annotationsprojekte legen darüber hinaus Wert auf konsistente und konsensuale Entscheidungen zwischen den Annotator*innen, da die getroffenen Entscheidungen oft als Trainings- oder Testdaten für (überwachte) maschinelle Lernmethoden verwendet werden.

Annotationen in den Geisteswissenschaften folgen nicht dem gleichen Paradigma wie in der Computerlinguistik, in der die Annotationen auch als Methode verwendet werden, um intersubjektive Entscheidungen zu treffen. Vielmehr spielt in den Geisteswissenschaften die Annotation als Hinzufügen von Notizen am Rand oft eine Rolle, um individuelle Interpretationen z. B. eines literarischen Textes für den eigenen Gebrauch zu visualisieren, wenn auch manchmal implizit. Die explizite Darstellung dieses Prozesses hat Vorteile, da explizite Annotationen die Interpretation unterstützen können, indem sie sie klarer, eindeutiger und auch nachvollziehbarer machen. Darüber hinaus könnte eine Zukunftsperspektive für die Geisteswissenschaften ein kollaborativer Prozess der Theorieentwicklung sein. Ein Ansatz zur Erreichung dieses Ziels ist die Integration der Annotationsmethode in die geisteswissenschaftliche Forschung, bei der theoretische Begriffe durch den Annotationsprozess formalisiert und iterativ geschärft werden.

Das vorliegende Kapitel vergleicht die in der CL vorherrschenden Annotationsprozesse mit den in den (digitalen) Geisteswissenschaften verwendeten Prozessen. Wir argumentieren, dass die Annotationsprozesse zwar unterschiedlichen Zielen dienen und unterschiedliche Prioritäten setzen, aber viele Gemeinsamkeiten aufweisen und in ein gemeinsames konzeptionelles Modell integriert werden können. Darüber hinaus argumentieren wir, dass Annotation ein produktives Werkzeug sein kann, um theoretische Begriffe und Definitionen in den Geisteswissenschaften zu präzisieren und verbessern, was eine neue Art der Verwendung von Annotation darstellt.

1 Im Gegensatz dazu stehen strukturelle Annotationen, die z. B. Formatierungsinformationen oder Metadaten markieren. Auch wenn hier technisch etwas Ähnliches passiert, ist es im Regelfall kein interpretativer Akt und steht (oft) in keinem Zusammenhang zum Textinhalt. Die Grenzen zwischen interpretativen und strukturellen Annotationen sind aber in Einzelfällen schwer zu ziehen.

2 Verschiedene Annotationsziele

Die verschiedenen Annotationsziele, die in der Literatur diskutiert werden, lassen sich grob in vier Bereiche sortieren, die sich teilweise überlappen:

Die **explorative Annotation** bietet zunächst die Möglichkeit, einen Text (oder ein anderes Datenobjekt) strukturiert kennenzulernen, ohne sich vorab auf Kategorien oder Schemata festzulegen. Diese Art der Auszeichnung ist der langjährigen Tradition der Kommentierung in den traditionellen Geisteswissenschaften am nächsten (Bradley 2008), bei der Forscher*innen ihnen wichtig erscheinende Aspekte sowie Ideen, die beim Lesen entstanden sind, am Rand einer Seite notieren. Bradley führt aus:

> [T]his kind of annotation, indeed note-taking more generally, provides one of the bases for much scholarly research in the humanities. In this view note-taking fits into the activity of developing a personal interpretation of the materials the reader is interested in. (Bradley 2012, Abs. 11)

Ziel dieser Art von Annotation ist es also, am Ende ein vorläufiges Textwissen samt möglicher Assoziationen zu erhalten, das es den Forscher*innen ermöglicht, eine konkretere Forschungsfrage oder Hypothese zu formulieren. Diese Frage oder Hypothese kann später vor dem Hintergrund einer theoretischen Grundlage bearbeitet werden, während die erste Lektüre ohne spezifische Annahmen oder Fragen erfolgt.

Zweitens zielt die **konzeptualisierende Annotation**[2] darauf ab, die Definition von theoretischen Begriffen oder prä-theoretischen Beobachtungen, die erklärungsbedürftig sind, zu verbessern. Beide werden oft in Sekundärliteratur beschrieben, aber selten so definiert, dass sie auf neue Texte ‚anwendbar' sind. Der Versuch, die Begriffe oder Beobachtungen systematisch in weiteren Texten zu finden, fördert zunächst Definitionslücken oder -schwierigkeiten zutage. Durch das Beheben dieser Lücken oder Schwierigkeiten werden die Definitionen besser und breiter anwendbar. Primäres Mittel hierbei ist es, Fälle von Meinungsverschiedenheiten zwischen verschiedenen Annotator*innen zu identifizieren und die Definitionen zu verfeinern, bis eine ausreichende Übereinstimmung erreicht wird.

Drittens zielt die **erklärende Annotation** darauf ab, eine kondensierte und ggf. formalisierte Darstellung der textlichen Grundlage für eine Interpretationshypothese zu liefern. Während Interpretationshypothesen (z. B. in der Literaturwissenschaft) typischerweise auf Textnachweisen basieren, sind die entsprechen-

2 Konzeptualisierende Annotation ist Ziel des *shared tasks* SANTA, der ab Seite 390, in Teil IV des Bandes vorgestellt wird.

den Textstellen nicht explizit markiert und der Argumentationspfad von Textstelle zur Interpretation bleibt implizit. Durch die Erläuterung in Form von Annotationen werden diese Schritte explizit und formalisiert. Diese Annotationen sind nicht auf ein einzelnes Phänomen beschränkt, sondern decken alle Phänomene ab, die für eine Interpretation als relevant erachtet werden. Bei dieser Konstellation besteht das Hauptziel nicht darin, eine einzige ‚wahre' Annotation zu erstellen, sondern verschiedene plausible Annotationen, die unterschiedliche Lesarten des Textes darstellen. Erläuternde Annotationen sind grundsätzlich in verschiedenen Formalisierungsgraden denkbar. Im unformalisierten Fall bestehen die Annotationen aus Fußnoten, die Text enthalten, der sich an Menschen richtet. Am anderen Ende des Spektrums sind Annotationen denkbar, in denen maschinenlesbare Informationen geliefert werden.

Viertens zielt die **automatisierungsorientierte Annotation**[3] (cf. Hovy und Lavid 2010; Pustejovsky und Stubbs 2012) auf die Zusammenstellung konsistent annotierter Daten, die als Trainings- und Testmaterial für automatische Annotationswerkzeuge verwendet werden. Automatisierte Annotationswerkzeuge benutzen dabei typischerweise Verfahren des überwachten maschinellen Lernens, bei dem Modelle trainiert werden, die den Zusammenhang zwischen Oberflächenmerkmalen und Zielkategorien aus großen Datenmengen erschließen. Die Konsistenz der Annotation ist für die Automatisierung von größter Bedeutung, da Inkonsistenzen die Klassifikationsleistung negativ beeinflussen. Annotationsprojekte, die Trainings- oder Testdaten generieren, legen Wert auf ein hohes *inter-annotator agreement* (IAA), also auf eine hohe Übereinstimmung zwischen den Annotator*innen.

Die vier genannten Anwendungsfälle der Annotationsmethode schließen sich nicht gegenseitig aus. Tatsächlich ist es schwierig, nicht zumindest einige Aspekte der anderen Ziele zu berühren, auch wenn nur ein einziges Ziel momentan im Fokus liegt. Annotation zur Generierung von Trainings- oder Testdaten z. B. legt oft Probleme in den Definitionen und Annotationsrichtlinien offen, die verfeinert werden müssen. Dabei werden Entscheidungen getroffen, die Rückwirkungen in die konzeptuelle Ebene haben, selbst wenn diese nicht intendiert, sondern die Entscheidungen vor allem pragmatisch motiviert sind. Die Entscheidung z. B., zwei Kategorien zu verschmelzen, mag pragmatisch motiviert sein. Sie ist aber Teil der Konzeptarbeit, die innerhalb des Annotationsprojektes stattfindet.

3 Annotation zum Ziel der automatisierten Erkennung des Zielphänomens ist außerdem Diskussionsgegenstand des Beitrages von Pichler und Reiter (2020), S. 43ff. in diesem Band.

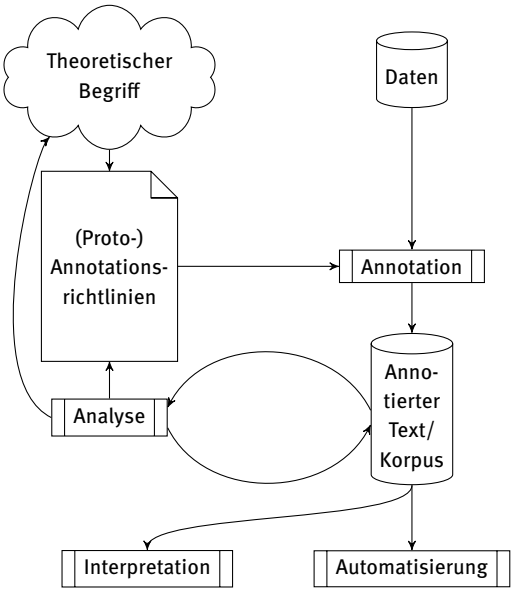


Abb. 1: Annotations-Workflow-Schema. Pfeile zeigen eine (grobe) zeitliche Abfolge an.

3 Ein einheitlicher Annotationsworkflow

Abbildung 1 veranschaulicht ein Modell für einen Annotationsworkflow, der auf die im vorigen Abschnitt genannten Ziele ausgerichtet ist. Es beschreibt sowohl das in der CL vorherrschende Annotationsmodell als auch Annotationsmodelle aus den Geisteswissenschaften und Anwendungsfälle, die neu und spezifisch für die DH sind. Die Bezeichnung als Workflow bedeutet nicht, dass jedes Annotationsprojekt jeden Teilschritt verwenden oder dass der gesamte Workflow innerhalb eines einzelnen Projekts abgeschlossen werden muss. Die in einem Projekt festgelegten Annotationsvorgaben können problemlos im nächsten Projekt fortgesetzt oder ausgearbeitet werden. Je nach Ziel des Projektes werden verschiedene Bereiche betont oder ignoriert. Generell können die verschiedenen Annotationsprozesse auch als Phasen betrachtet werden, die ein Phänomen durchläuft, bis ein intersubjektives Verständnis erreicht worden ist.

Der Ausgangspunkt des Workflows ist ein **theoretischer Begriff**. Wir verwenden hier die Bezeichnung ‚Begriff', um eine Vielzahl von Fällen einzubeziehen: Der Begriff kann auf der Grundlage einer vollwertigen Theorie (z. B. Wortarten oder Erzählebenen) beschrieben/vorhergesagt werden, aber er kann auch auf einer Beobachtung in Textdaten beruhen, die erklärt werden muss oder in existie-

render Fachliteratur diskutiert wurde (z. B. Ähnlichkeiten in der Figurendarstellung bei Adaptionen eines literarischen Stückes). Der theoretische Begriff wird in der Abbildung durch eine Wolke dargestellt, um anzuzeigen, dass er oft ‚unscharfe Kanten' aufweist, und die Anwendung auf textuelle Daten Interpretationsschritte beinhaltet. Der Begriff kann natürlich, etwa durch Sichtung von Forschungsliteratur, weiter und feiner spezifiziert werden, bevor er in den Annotationszyklus eintritt.

Theoretische Begriffe interagieren auf vielfältige Weise mit **Daten**: Viele Begriffe beruhen auf an Daten gemachten Beobachtungen, selbst wenn sie indirekt oder nur durch den bisherigen wissenschaftlichen Diskurs übermittelt werden. Daher spielt das Zustandekommen der Beobachtung ebenfalls eine Rolle: Eine konkrete Datensammlung kann nie wirklich zufällig ausgewählt werden und setzt daher zumindest die (breite) Beschränkung auf ein Interessengebiet voraus. Die Auswahl der Daten führt zu einer Verzerrung und schränkt den Raum für mögliche Beobachtungen ein. Dazu führen Kanonisierungs- und Standardisierungsprozesse zu einer Verengung des Blickwinkels und machen bestimmte Phänomene grundsätzlich nicht beobachtbar. Dies ist unabhängig vom genauen Status des theoretischen Begriffs. Daher muss der Datenauswahl große Aufmerksamkeit geschenkt werden und Kriterien für die Auswahl müssen für die Nutzer*innen der Sammlung explizit festgelegt werden, um die Forschung transparent und nachvollziehbar zu machen.

Die eigentliche Annotation basiert immer auf **Annotationsrichtlinien**, selbst wenn diese nicht expliziert sind (und es vielleicht auch nie werden). Wenn ein theoretisches Konzept zum ersten Mal annotiert wird, können die Richtlinien zunächst nur eine Fixierung auf ein bestimmtes theoretisches Werk (für die Narratologie z. B. Genette 1980) oder einen Teil davon (z. B. narrative Ebenen) sein. Iterationen im Annotations-Workflow können zu immer aufwendigeren Annotationsvorgaben führen, die vom theoretischen Konzept abweichen können. Für das Alltagsgeschäft der Annotation dienen Richtlinien als Vermittler zwischen theoretischen Begriffen und der eigentlichen Annotationspraxis. Im Idealfall können so auch Nicht-Expert*innen (z. B. studentische Hilfskräfte oder Crowdsourcing-Arbeiter*innen) die Annotationen durchführen. Annotationsrichtlinien sollten zwar so generisch wie möglich gehalten werden, durch die Art ihres Zustandekommens sind sie jedoch oft ‚optimiert' für bestimmte Texte oder Korpora. Wenn theoretische Konzepte für Nicht-Expert*innen aufgeschlüsselt werden, werden sie in Bezug auf das zu kommentierende Korpus beschrieben; schwierige, aber irrelevante Aspekte können völlig ignoriert werden.

Der eigentliche **Annotationsprozess** besteht dann darin, Texte zu lesen, Textteile zu markieren/auszuwählen und mit den in den Richtlinien definierten Kategorien zu verknüpfen. Manchmal werden zusätzliche Merkmale einer

Instanz des theoretischen Begriffs annotiert. Abhängig von den Zielen der Annotation können Annotationen parallel durchgeführt werden, d. h. mehrere Annotator*innen bearbeiten den gleichen Text parallel. Dies ermöglicht den direkten Vergleich der Annotationen, um mögliche Mängel der Annotationsrichtlinien aufzudecken. Ein weiterer Parameter im Annotationsprozess ist die Annotationseinheit: Einige Annotationen funktionieren wortweise (d. h., jedes Wort ist annotiert), andere basieren auf Sätzen oder Phrasen. Um (nicht-triviale) linguistische Einheiten zuverlässig als Basiseinheit zu verwenden, ist es sinnvoll, diese manuell oder automatisch zu erzeugen und als Vorannotation bereitzustellen oder klare Kriterien für diese ebenfalls in den Richtlinien zu verankern. Auch wenn die Annotationen (je nach Ziel) grundsätzlich auf Papier erfolgen können, können computergestützte Annotationswerkzeuge den Annotationsprozess unterstützen, indem sie Annotationskandidaten vorschlagen oder die Nachnutzung der annotierten Daten ermöglichen.

Das unmittelbare Ergebnis des Annotationsprozesses ist ein **annotiertes Korpus**, das auf verschiedene Arten genutzt werden kann. Eine naheliegende Art der **Analyse** ist es, bestimmte Hypothesen oder Annahmen mit den neu erstellten Daten zu vergleichen. Diese Art der Analyse fördert ein besseres Verständnis der Theorie, z. B. in Form von präziseren theoretischen Begriffen. Die Analyse von konkreten Daten kann auch dazu führen, dass Belege für oder gegen bestimmte theoretisch motivierte Hypothesen gefunden werden. Diese Ergebnisse können dann genutzt werden, um die zugrundeliegende Theorie zu verfeinern. Eine andere Art von Analysen basiert auf den Meinungsverschiedenheiten zwischen mehreren Annotator*innen. Das Hauptziel dieser Art von Analyse ist es, sicherzustellen, dass i) die Annotationsrichtlinien ausreichend genau und klar definiert sind und ii) sie von den Annotator*innen gelesen, verstanden und befolgt wurden. Ein zielführendes Vorgehen dabei besteht darin, diejenigen Annotationen manuell zu überprüfen, bei denen die Annotator*innen verschiedener Meinung waren, d. h., unterschiedliche Annotationsentscheidungen getroffen haben. Dies kann durch die Annotator*innen selbst oder durch die ‚Vorgesetzten' erfolgen. Quantitativ kann die Menge an Meinungsverschiedenheiten als *inter-annotator agreement* ausgedrückt werden, das typischerweise in der Dokumentation zu einer Korpusveröffentlichung enthalten ist. Während die Messung von IAA eine recht lange Tradition hat (frühe Veröffentlichungen dazu stammen aus den 1960ern, wie etwa Cohen 1960), ist die Diskussion darüber, wie sich IAA genau quantifizieren lässt, noch nicht abgeschlossen (Mathet et al. 2015). Unterschiedliche Aufgaben, die auf unterschiedlichen Einheiten basieren erfordern zudem unterschiedliche Metriken für das *inter-annotator agreement*. Die quantitative Messung von IAA ist beim Vergleich verschiedener Annotationsrichtlinien oder annotierter Korpora zentral und das gemessene IAA kann auch als Obergrenze für die Maschinenleis-

tung dienen. Wenn das Ziel der Annotation darin besteht, theoretische Konzepte zu entwickeln, ist die Überprüfung der tatsächlichen Meinungsverschiedenheiten der Annotator*innen aufschlussreicher als die Zählung derselben. Gius und Jacke (2017) schlagen vor, Meinungsverschiedenheiten in vier Kategorien einzuteilen, basierend auf ihren Ursachen: i) Annotationsfehler, ii) Schwächen in den Annotationsrichtlinien, iii) abweichende Vorannahmen und iv) ‚echte' Mehrdeutigkeiten. Annotationsfehler basieren auf Nachlässigkeiten und können sofort behoben werden, die Meinungsverschiedenenheiten der Kategorien ii) und iii) erfordern eine Anpassung der Annotationsrichtlinien. Wenn Meinungsverschiedenheiten der Kategorie iv) nicht durch Berücksichtigung eines zusätzlichen Kontextes gelöst werden können, bleiben sie als zwei Lesarten und Annotationsmöglichkeiten im Korpus erhalten.

Sobald ein annotiertes Korpus verfügbar ist, sind zwei verschiedene Folgeschritte möglich: Interpretation und Automatisierung. Die **Interpretation** eines Textes auf der Grundlage von Annotationen führt zu zusätzlichen Lesarten, die auf konkreten Datenpunkten beruhen, was letztlich auch zu einer intersubjektiveren Interpretation des Textes beitragen kann. Wir werden hier nicht im Detail auf den **Automatisierungsprozess** eingehen, aber er erfordert typischerweise annotierte Daten (siehe in diesem Band den Beitrag von Pichler und Reiter (2020) für eine abstrakte Beschreibung und den von Klinger et al. (2020) für ein konkretes Fallbeispiel).

Eine Annahme in der CL ist, dass die Annotationen eindeutig sind, dass also alle Meinungsverschiedenheiten gelöst wurden. Wie mit Unstimmigkeiten, die auch durch korrekte Anwendung der Annotationsregeln nicht eindeutig entschieden werden können in Bezug auf die Automatisierung umgegangen werden kann, ist noch nicht abschließend geklärt. Gius und Jacke (2017) schlagen unterschiedlich parametrisierte Modelle für die automatische Vorhersage vor, zumindest für die Uneinigkeitskategorie (iii). So kann beispielsweise die Annotation einer bestimmten Kategorie eine Entscheidung über eine grundlegendere verwandte Kategorie erfordern. In einem Tool zur automatischen Erkennung eines bestimmten Begriffs kann dieser Parameter manuell eingestellt werden, um einen bestimmten Messwert zu erzwingen. Gius und Jacke lassen jedoch die Frage offen, wie dies bei Meinungsverschiedenheiten, die sich aus einer gültigen Mehrdeutigkeit des Textes ergeben, realisiert werden kann.

Der robuste Umgang statischer oder quantitativer Methoden mit echt mehrdeutigen Daten bleibt also ein Desideratum, gerade auch in Hinblick auf Daten aus den Geistes- und Sozialwissenschaften.

4 Beispielhafte Annotationsprojekte

Wir besprechen nun mehrere DH-Projekte, um die verschiedenen Ziele von Annotation zu veranschaulichen und verschiedene Wege aufzuzeigen, die Projekte mit unserem Annotationsworkflow einschlagen können.

4.1 Explorative Annotation

Ein Beispiel für explorative Annotation ist das Anfertigen von Notizen. McCarty (2020) beschreibt seinen *note-taking*-Workflow, der aus dem Sammeln von Notizen in digitaler Form, deren Ausdrucken auf kleinen Kärtchen und anschließend der manuellen Sortierung besteht.

Ein frühes Projekt, das ein solches Vorgehen in einem Tool implementierte, ist das Pliny-Projekt (Bradley 2008). Pliny ist eine 2009 veröffentlichte Software, um neue Möglichkeiten der Annotation in der digitalen Welt zu erforschen. Es soll den traditionellen geisteswissenschaftlichen Workflow (Bradley 2008) unterstützen, indem es den Prozess des *note-taking* und die Aufzeichnung erster Reaktionen auf einen Text mit dem Ziel einer nachfolgenden Phase, in der eine Forschungsfrage entwickelt wird, ermöglicht. Die Entwickler*innen geben das Beispiel einer Webseite[4], auf der die Benutzer*innen Beobachtungen notieren, die sie während des Besuchs der Seite machen. In unserem Workflow entspricht diese Phase der Annotation einer prä-theoretischen Stufe, in der Daten die Annotation auslösen. Dies kann in einem nächsten Schritt möglicherweise zur Analyse des annotierten Textes führen, was wiederum zu Annotierungsrichtlinien führen kann. Obwohl die Verantwortlichen behaupten, dass Pliny den ‚traditionellen Weg' des Notierens in die digitale Welt verlagere, scheint Pliny in der DH-Wissenschaftswelt keine Akzeptanz zu finden: Es gibt – wenn überhaupt – nur wenige Projekte, die das Tool nutzen (und es dokumentiert haben). Dies könnte aber auch ein Hinweis auf eine wenig entwickelte Tradition der geisteswissenschaftlichen Methodendiskussion sein, die zu einem Mangel an Publikationen über den Annotationsprozess innerhalb bestimmter Projekte führt.

Ein neueres Projekt zur Unterstützung von explorativen Annotationen ist das 3DH-Projekt[5], bei dem zunächst Prototypen entwickelt werden (Kleymann et al. 2018). Eine Möglichkeit besteht darin, Textstellen tentativ gruppieren zu können, ohne sie direkt benennen zu müssen. Dies wiederum soll in einer intuitiv

4 Aus den „Proceedings of the Old Bailey": http://www.oldbaileyonline.org
5 http://threedh.net/3dh/

benutzbaren und durchdachten Oberfläche auch technischen Laien möglich sein, wodurch ein Arbeitsschritt *vor* einer kategoriengeleiteten Annotation unterstützt würde.

Daneben wird eine explorative Annotation auch in Version 6 des Annotationswerkzeuges CATMA[6] unterstützt: Durch simples Hervorheben von Textstellen, sowie durch eine freie Kommentarfunktion kann explorative Annotation leicht umgesetzt werden (Horstmann und Jacke 2020). Durch das iterative Ineinandergreifen mehrerer Annotationsmodi (Hervorhebung, Kommentierung, Kategorisierung) kann zudem ein Teil des in Abbildung 1 gezeigten Workflows direkt innerhalb eines Tools umgesetzt werden.

Für explorative Annotation ist das *inter-annotator agreement* nicht wichtig, da sie in erster Linie dem Ziel dient, ein Verständnis für wichtige Konzepte und mögliche Ansatzpunkte für ein Forschungsprojekt zu entwickeln.

4.2 Konzeptualisierende Annotation

Als Beispiel für die Konzeptualisierung von Annotation wollen wir zunächst Moretti (2013) nennen. Er beschreibt die Abweichung von der Definition des ‚character-space' von Woloch (2003). Die Operationalisierung dieser Literaturtheorie durch Annäherung an den Textraum, den eine Figur einnimmt, durch Kenntlichmachung dessen, wie viele Wörter diese in einem dramatischen Text spricht, stärkt die zugrunde liegende Theorie und führt „back from theories, through data, to the empirical world." (Moretti 2013, S. 4). Moretti hält dies für entscheidend für literarische Theorien, weil es einige Konzepte „‚actual' in the strong sense of the word" mache (Moretti 2013, S. 4). In unserem Arbeitsablauf konzentriert sich dieses Projekt stark auf die Formalisierung eines theoretischen Begriffs, also die Übersetzung vom Konzept des Zeichenraums in den Raum des eigentlichen Textabschnitts. Die Annotation selbst ist trivial, aber der annotierte Text wird dann als Grundlage für die Interpretation verwendet.

Ein gründlicherer Versuch, die Annotation zur Entwicklung theoretischer Konzepte zu nutzen, wurde von Bögel et al. (2015) unternommen. Ziel des Projekts heureCLÉA ist es, zeitbezogene narrative Phänomene in literarischen Texten zu annotieren. Die veröffentlichten Richtlinien[7] sind bereits spezifischer als die zugrundeliegende Theorie, da sie den Umgang mit z. B. hypothetischen Prolepsen definieren. Dieser Prozess der Verfeinerung des theoretischen Begriffs durch

6 https://catma.de

7 http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf

Annotation kann auch als *shared task* durchgeführt werden (siehe auch Willand et al. 2020 in diesem Band, für einen *shared task* mit Fokus auf Erzählebenen).

Potenziell wird die systematische Konfrontation der Theorie mit Textbeispielen zu Implikationen für diese Theorie führen. Für diese Art von Annotation bildet das *inter-annotator agreement* eine Grundlage, um vordefinierte theoretische Konzepte intersubjektiv zu diskutieren. So ist IAA eine Metrik, die Informationen darüber liefern kann, wie spezifiziert eine Theorie ist und in welchem Maße sie die Definition von Indikatoren zur intersubjektiven Überprüfung ermöglicht.

Ein weiteres Beispiel für die konzeptualisierende Annotation ist die Koreferenzannotation. Die Annotation von Koreferenzen ist in der Computerlinguistik gut etabliert und wird durch bereits bestehende theoretische Begriffe und Richtlinien unterstützt (Pradhan et al. 2007; Dipper und Zinsmeister 2009; Riester und Baumann 2017). Die Anwendung dieser Richtlinien auf ‚neue' Textsorten zeigt jedoch, dass sie weiter verbessert werden müssen. Im Rahmen des QuaDramA-Projektes[8] wurden dramatische Texte mit Koreferenzketten annotiert. Dabei ergaben sich neue Fragen, die in Weiterentwicklungen der Annotationsrichtlinien mündeten (Rösiger et al. 2018). Das Projekt entspricht dem Workflow wie folgt: Bestehende Annotationsrichtlinien wurden übernommen und der Annotationsprozess eingeleitet. Nach der Annotation der ersten Texte wurde eine Analyse durchgeführt und die Richtlinien wurden an die Daten und spezifischen Probleme angepasst. Weitere Texte wurden dann mit der neuen Version der Richtlinien annotiert. Bei der Anpassung der Richtlinien handelt es sich um einen Konzeptualisierungsschritt, da die neuen Richtlinien neue Erkenntnisse widerspiegeln, die aus der Betrachtung konkreter Koreferenzphänomene gewonnen wurden.

Ein Beispiel, bei dem die theoretische Vorarbeit eine größere Rolle spielt, ist die Annotation von Gender-Stereotypen im m*w-Projekt (Schumacher und Flüh 2020)[9]. Ausgehend von theoretischen Unterscheidungen von Butler (zwischen Genus, Gender Identity und Gender Performanz) wurde weitere Literatur herangezogen, um zu annotierbaren Kategorien zu gelangen, bevor die eigentliche Annotation startete (die dann ihrerseits wieder zu kategoriellen Verschiebungen führte).

4.3 Erklärende Annotation

Ein Beispiel für ein erklärendes Annotationsprojekt in einem frühen Stadium ist die in Nantke und Schlupkothen 2018 vorgestellte Arbeit. Die Autor*innen kon-

8 https://quadrama.github.io

9 https://msternchenw.de

zentrieren sich auf die Kommentierung von intertextuellen Referenzen, um mögliche Interpretationen eines Textes zu formalisieren. Nur eine Teilmenge der vorgeschlagenen Formalisierungen sind tatsächlich textuelle Annotationen im engeren Sinne – andere sind Beziehungen zwischen textuellen Annotationen oder zwischen textuellen Annotationen und (digitalen Darstellungen von) historischen Kontexten. Auf technischer Ebene werden sowohl die Annotationen als auch die Beziehungen durch *semantic-web*-Technologien dargestellt. Wichtig ist, dass diese Annotationen kein einzelnes Phänomen, sondern eine große Anzahl an ‚grundlegenden Annotationen' zu verschiedenen Phänomenen abdecken. Angesichts der Komplexität dieser Annotationen erscheint ein groß angelegtes Annotationsprojekt schwierig zu realisieren – solche Annotationen werden hauptsächlich für einen einzigen Text erstellt. Da zudem keine festen Kategorien existieren (müssen), ist ein *inter-annotator agreement* schwer zu berechnen.

Als zweites Beispiel sei hier das Projekt TEASys[10] genannt. Ziel des Projektes ist es, erklärende Annotationen systematisch zur Unterstützung des Leseverständnisses zu erproben. Erklärungsbedürftigen Stellen englischer Literatur werden dabei (u. a. in Seminaren, also im Unterrichtskontext) Erklärungen beigefügt, die auf verschiedenen Ebenen strukturiert sind. So können u. a. historische Verwendungen von Wörtern oder Formulierungen, intertextuelle Verweise oder Interpretationsansätze ergänzt werden (Bauer und Zirker 2017).

In Bezug auf den in Abbildung 1 dargestellten Arbeitsablauf verwenden erklärende Annotationen theoretische Begriffe zur grundlegenden Bestandsaufnahme von Textnachweisen (möglichst unter Verwendung von Annotationsrichtlinien), ohne darauf abzuzielen, diese zu verbessern. Stattdessen folgen diese Projekte dem rechten Pfad im Workflow, was zu einem annotierten Text führt, gefolgt von einer Interpretation oder einer Begründung der Interpretation mithilfe der Annotationen.

4.4 Automatisierungsorientierte Annotation

Die letzte Art der Annotation, die wir diskutieren wollen, ist die automatisierungsorientierte Annotation, die in der Computerlinguistik weit verbreitet ist. Dabei besteht das Ziel der Annotation darin, Daten bereitzustellen, um statistische Zusammenhänge zwischen Oberflächenmerkmalen und den jeweiligen Zielkategorien zu ermitteln. Die annotierten Daten fungieren dann zum einen als Trainings- und

10 http://www.annotation.es.uni-tuebingen.de

zum anderen als Testdaten, also für die Evaluation der erstellten automatischen Erkenner.

Ein prominentes Beispiel für Annotationen, die als Input für einen vollautomatischen Ansatz verwendet werden, ist die Annotation von Wortarten. Part-of-Speech-Tagging (PoS-Tagging) ist eine konzeptionell klare Aufgabe, was sich im hohen *inter-annotator agreement* zeigt, das für diese Aufgabe erreicht wird. Das kürzlich veröffentlichte GRAIN-Korpus (Schweitzer et al. 2018) beispielsweise enthält Annotationen von drei Annotator*innen für deutsche Radiointerviews, die recht komplexe und spontane Rede umfassen. In ihrer Arbeit geben die Autoren ein *pairwise* Cohen's κ (Cohen 1960) von 0,97 an, das allgemein als fast perfekte Übereinstimmung gilt. Die Tatsache, dass die Annotation vom Menschen konsequent durchgeführt werden kann, ist eine notwendige Voraussetzung für die Entwicklung von automatischen Werkzeugen. In der Folge war das PoS-Tagging eine der ersten CL-Aufgaben, bei denen die Leistung von automatischen Werkzeugen mit einer Genauigkeit von über 97 % ein zufriedenstellendes Niveau erreicht hat (cf. Manning 2011) und heute zumindest für Standardtexte als eine fast gelöste Aufgabe gilt.

Das PoS-Tagging wurde auch auf Texte aus der DH-Domäne angewendet, z. B. historische Texte, bei denen die Leistung von Standardwerkzeugen natürlich nicht zufriedenstellend ist. Schulz und Kuhn (2016) konnten jedoch zeigen, dass für Mittelhochdeutsch bereits eine geringe Menge an annotierten Daten (z. B. rund 200 Sätze) zu tragfähigen Ergebnissen von automatischen Systemen führen kann.

Ein weiteres Beispiel für Annotation, die letztlich der Automatisierung dient, ist in dem Beitrag von Klinger et al. (2020), in diesem Band ab S. 238, dargestellt. Dabei ist das Ziel, Emotionen zu kategorisieren, die von literarischen Figuren empfunden werden. Dazu wurde ein Korpus annotiert und anschließend als Trainingsmaterial für ein maschinelles Lernverfahren verwendet. Auftretende Hürden bei der Annotation wurden gelöst, indem die Annotationsrichtlinien erweitert wurden.

5 Diskussion und Schlussfolgerungen

In diesem Beitrag zeigen wir, dass Annotation nicht nur als Mittel zur Erstellung von Trainingsmaterial für maschinelle Lernansätze dienen kann. Eine als Forschungsmethode verstandene Annotation kann zur Entwicklung eines fokussierten Verständnisses relevanter Konzepte, die in Texten zu finden sind, sowie als Instrument zur Spezifikation und Verifikation theoretischer oder prätheore-

tischer Konzepte dienen. Dies ist besonders fruchtbar für Disziplinen wie die Literaturwissenschaft, in denen Konzepte im wissenschaftlichen Diskurs oft implizite Vorannahmen mitbringen und daher unterspezifiziert wirken, was eine intersubjektive Anwendung erschwert.

Im Allgemeinen kann die Kommentierung von nicht-standardisierten (aus Sicht der Computerlinguistik) Texten dazu beitragen, neue Phänomene aufzudecken, die eine Anpassung oder Erweiterung von Annahmen erfordern. So muss beispielsweise die computerlinguistische Annahme, dass es eine *ground truth* – eine einzige korrekte Annotation – gibt, für literarische Textkonzepte möglicherweise ausgeweitet werden, da das Lesen und Interpretieren eines Textes unterschiedliche und dennoch korrekte Lesarten zulässt. Es bleibt eine Herausforderung für die Methoden des maschinellen Lernens, wie mit diesen ‚echten' Unklarheiten in Bezug auf das Training und die Bewertung automatischer Systeme umgegangen werden kann.

Eine weitere Überlegung, die diese verschiedenen Arten von Annotationen auslösen, ist die Wahl des Annotationswerkzeugs: Annotationswerkzeuge, die in der Computerlinguistik entwickelt wurden (z. B. WebAnno, Yimam et al. 2013, oder MMAX2, Müller und Strube 2006), treffen selbstverständlich die dort verwendeten Annahmen. So beinhalten sie typischerweise eine Methode zum Vergleichen von Annotationen und die Annotationskategorien und -richtlinien werden im Voraus definiert.

Explorative und auch erklärende Annotation haben andere Anforderungen: Das Annotationswerkzeug CATMA 6 unterstützt die Annotation freier Kommentare oder simpler Hervorhebungen, das im Rahmen des 3DH-Projekts (Kleymann et al. 2018) entwickelte Tool ermöglicht die Markierung beliebiger Textbereiche und Interaktion mit ihnen (z. B. Gruppierung und/oder Visualisierung). Das Annotationssystem TEASys verwendet Kategorien, aber Kategorien für Freitext-Kommentare.

In jedem Fall ist das Verhältnis zwischen der vom Tool angebotenen Funktionalität und dem Ziel des Annotationsprozesses noch ein wenig erforschter Bereich. Leider scheint es kaum Publikationen zu geben, in denen die Verwendung der Annotationsmethodik als Mittel zur Erforschung neuer Texte oder zur Schärfung von Forschungsfragen dokumentiert wird, während Beschreibungen und Diskussionen von automatisierungsorientierten Annotationen häufig zu finden sind.

Zusammenfassend haben wir einen Workflow für Annotationen beschrieben, die in den DH durchgeführt werden. Der Workflow soll so offen und flexibel wie möglich sein, um den verschiedenen möglichen Perspektiven und Bereichen, die in den DH zusammenkommen, Rechnung zu tragen, während er sich gleichzeitig auf Schritte konzentriert und solche erfordert, die von allen Annotationsunterneh-

mungen geteilt werden sollten. Wir definieren vier Hauptziele, die die verschiedenen Bereiche von DH verfolgen könnten: Explorative, konzeptionelle, erklärende und automatisierungsorientierte Ziele. Wir besprechen den Zweck und die Unterschiede der einzelnen Ziele auf allgemeiner Ebene, gefolgt von einer Übersicht konkreter Projekte in den DH mit einem dieser Ziele. Diese Übersicht zeigt auch den Einsatz des Workflows in verschiedenen Situationen und unterstreicht seine Flexibilität. Wir glauben, dass unser Workflow für alle Arten von DH-Zielen allgemein anwendbar ist und hoffen, dass in Zukunft weitere Projekte Annotation nutzen werden, um alte Fragen der Geisteswissenschaften aus einer neuen Perspektive zu betrachten.

Literatur

Bauer, Matthias und Angelika Zirker (2017). „Explanatory Annotation of Literary Texts and the Reader: Seven Types of Problems". In: *International Journal of Humanities and Arts Computing* 11.2, S. 212–232. DOI: 10.3366/ijhac.2017.0193.

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris und Jannik Strötgen (2015). „Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative". In: *DHCommons* 1. DOI: 10.5281/zenodo.3240591.

Bradley, John (2008). „Thinking about interpretation: Pliny and scholarship in the humanities". In: *Literary and Linguistic Computing* 23.3, S. 263–279. DOI: 10.1093/llc/fqn021.

Bradley, John (2012). „Towards a Richer Sense of Digital Annotation: Moving Beyond a ‚Media' Orientation of the Annotation of Digital Objects". In: *Digital Humanities Quarterly* 6.2. URL: http://www.digitalhumanities.org/dhq/vol/6/2/000121/000121.html (besucht am 1. Juni 2020).

Cohen, Jacob (1960). „A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, S. 37–46.

Dipper, Stefanie und Heike Zinsmeister (2009). „Annotating Discourse Anaphora". In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. ACL-IJCNLP. Singapore, S. 166–169.

Genette, Gérard (1980). *Narrative Discourse – An Essay in Method*. Übers. von Jane E. Lewin. Ithaca, New York: Cornell University Press.

Gius, Evelyn und Janina Jacke (2017). „The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis". In: *International Journal of Humanities and Arts Computing* 11.2, S. 233–254. DOI: 10.3366/ijhac.2017.0194.

Horstmann, Jan und Janina Jacke (2020). „Interpretationsspielräume. Undogmatisches Annotieren literarischer Texte in CATMA 6". In: *DHd 2020 Digital Humanities: Spielräume. Conference abstracts*. Paderborn, S. 154–158.

Hovy, Eduard und Julia Lavid (2010). „Towards a ‚Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In: *International Journal of Translation Studies* 22.1, S. 13–36.

Kleymann, Rabea, Jan Christoph Meister und Jan-Erik Stange (2018). „Perspektiven kritischer Interfaces für die Digital Humanities im 3DH-Projekt“. In: *Abstracts der DHd: Kritik der digitalen Vernunft*. Köln: Digital Humanities im deutschsprachigen Raum e.V., S. 279–284.

Klinger, Roman, Evgeny Kim und Sebastian Padó (2020). „Emotion Analysis for Literary Studies“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 238–268.

Manning, Christopher D. (2011). „Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?“ In: *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science. Berlin/Heidelberg: Springer, S. 171–189. DOI: 10.1007/978-3-642-19400-9_14.

Mathet, Yann, Antoine Widlöcher und Jean-Philippe Métivier (2015). „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment“. In: *Computational Linguistics* 41.3, S. 437–479. DOI: 10.1162/COLI_a_00227.

McCarty, Willard (2020). „Making and studying notes: Towards a cognitive ecology of annotation“. In: *Annotations in Scholarly Editions and Research*. Hrsg. von Julia Nantke und Frederik Schlupkothen. Berlin: De Gruyter, S. 269–295.

Moretti, Franco (2013). „‚Operationalizing‘: or, the function of measurement in modern literary theory“. Pamphlet 6. Stanford Literary Lab. URL: http://litlab.stanford.edu/LiteraryLabPamphlet6.pdf (besucht am 1. Juni 2020).

Müller, Christoph und Michael Strube (2006). „Multi-level annotation of linguistic data with MMAX2“. In: *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Hrsg. von Sabine Braun, Kurt Kohn und Joybrato Mukherjee. New York: P. Lang, S. 197–214.

Nantke, Julia und Frederik Schlupkothen (2018). „Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches‘ Semantic Web“. In: *Abstracts der DHd: Kritik der digitalen Vernunft*. Köln: Digital Humanities im deutschsprachigen Raum e.V., S. 345–349.

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2018). „A Unified Text Annotation Workflow for Diverse Goals“. In: *Proceedings of the Workshop for Annotation in Digital Humanities (annDH)*. Sofia, Bulgarien, S. 31–36. URL: http://ceur-ws.org/Vol-2155/pagel.pdf (besucht am 1. Juni 2020).

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Pradhan, Sameer S., Lance Ramshaw, Ralph Weischedel, Jessica Macbride und Linnea Micciulla (2007). „Unrestricted coreference: Identifying entities and events“. In: *ICSC ’07: Proceedings of the International Conference on Semantic Computing*, S. 446–453. DOI: 10.1109/ICSC.2007.106.

Pustejovsky, James und Amber Stubbs (2012). *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Sebastopol, Boston, Farnham: O’Reilly Media.

Riester, Arndt und Stefan Baumann (2017). „The RefLex Scheme – Annotation Guidelines“. SinSpeC. Working papers of the SFB 732 14. University of Stuttgart.

Rösiger, Ina, Sarah Schulz und Nils Reiter (2018). „Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena“. In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, USA, S. 129–138.

Schulz, Sarah und Jonas Kuhn (2016). „Learning from Within? Comparing PoS Tagging Approaches for Historical Text." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slowenien: European Language Resources Association (ELRA), S. 4316–4322.

Schumacher, Mareike und Marie Flüh (2020). „m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts". In: *DHd 2020 Digital Humanities: Spielräume. Conference abstracts*. Paderborn, S. 162–166.

Schweitzer, Katrin, Kerstin Eckart, Markus Gärtner, Agnieszka Faleńska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien und Jonas Kuhn (2018). „German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), S. 2887–2895.

Willand, Marcus, Evelyn Gius und Nils Reiter (2020). „SANTA: Idee und Durchführung". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 391–422.

Woloch, Alex (2003). *The One Vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton, New Jersey: Princeton University Press.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho und Chris Biemann (2013). „WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, S. 1–6. URL: http://www.aclweb.org/anthology/P13-4001 (besucht am 1. Juni 2020).

Sandra Richter

Reading with the Workflow

Arbeitsprozesse in den Computational Literary Studies – Beiträge zur Empirisierung literaturwissenschaftlicher Verfahren

Zusammenfassung: Ausgehend von der Debatte über den ‚new rigor' in den Geisteswissenschaften fragt der Beitrag nach Arbeitsprozessen der Computational Literary Studies (CLS). Als heuristische Suchvorgabe dient der Begriff der Empirisierung. Der Beitrag fragt, inwiefern und in welchen Hinsichten die CLS Gegenstände, Verfahren, Begriffe und Ziele der Literaturwissenschaft schärfen und für andere Disziplinen anschlussfähig machen. Als Beispiel dienen Ansätze aus einem Projekt über Goethes „Werther" und sogenannte Wertheriaden, das im Rahmen des Center for Reflected Text Analytics lief. Das Projekt zeigt, dass Arbeitsprozesse in den CLS nicht nur durch big data, sondern auch durch feingranular aufbereitete small data Empirisierungsleistungen für die Literaturwissenschaft erbringen können.

Abstract: Based on the debate about the 'new rigor' in the humanities, this article examines the working processes of Computational Literary Studies (CLS). The term 'empiricism'' serves as a heuristic search criterion. The article asks to what extent and in what respects the CLS sharpen the objects, procedures, concepts and goals of literary studies and make them compatible with other disciplines. Approaches from a project on Goethe's "Werther" and so-called Wertheriads, which ran within the framework of the Center for Reflected Text Analytics, serve as a basis. The project shows that work processes in the CLS can provide empirical evidence for literary studies not only through big data, but also through finely granulated small data.

Obwohl die Digital Humanities (DH) bereits einige Jahrzehnte alt sind, sind ihre Arbeitsprozesse bislang nur wenig theoretisiert (Krämer und Huber 2018). Dieser Befund erstaunt unter methodologischem, forschungspraktischem und wissenschaftspolitischem Aspekt. Zugleich entspricht er dem status quo eines Arbeitsfeldes, das sich zwischen den Disziplinen entwickelt, sich durch digital aufbereitete Korpora, computationelle Verfahren, Visualisierungsformen und dergleichen

Sandra Richter, Institut für Literaturwissenschaft, Universität Stuttgart, und Deutsches Literaturarchiv Marbach

Open Access. © 2020 Sandra Richter, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-007

behauptet, um seinen eigenen Ort im Wissenschaftssystem zu suchen. In diese Dynamik hinein möchte dieser Beitrag Beobachtungen für künftige Theoretisierungen von Arbeitsprozessen in den computationell erweiterten Geisteswissenschaften beisteuern. Er blickt zu diesem Zweck aus der Literaturwissenschaft auf einen bedeutenden Zweig der Digital Humanities: die Computational Literary Studies (CLS).

Seit ihrer Entstehung im 19. Jahrhundert bemüht sich die Literaturwissenschaft um ihre Selbstbeschreibung als Wissenschaft sowie um Praktiken, die sie als solche legitimieren. Dabei kommt es wiederkehrend und orientiert an Entwicklungen anderer Wissenschaften zu Bewegungen der Rationalisierung und Empirisierung des Faches (Richter 2010). Beides ist nicht dasselbe, doch überlappen sich die Bemühungen um Rationalisierung und Empirisierung. Beide gehen in der Regel – mitunter in Opposition zu anderen Ansätzen, die vornehmlich auf den literarischen Text und seine Besonderheit gerichtet sind – von Defizitanalysen aus: Das Fach gilt als zu eng, zu fokussiert auf die mehr oder minder subjektive Lektüre ‚schöner' Texte, verknappt unter den Stichworten der Hermeneutik, des ‚*close reading*' oder dergleichen attackiert. Dieses so oder ähnlich beschriebene Defizit versuchen Vertreter der Rationalisierung mit Hilfe eines Analysevokabulars beispielsweise aus der Philosophie oder Linguistik und durch exaktere Beschreibungen der fachlichen Arbeitsprozesse zu durchdringen. Vertreter der Empirisierung setzen – oft mit Hilfe historiografischer, soziologischer oder psychologischer Ansätze – auf die Anreicherung der Lektüren durch ergänzende Informationen (Kontexte, Daten u. dgl.), um sie zu verdichten, zu erhärten oder überprüfbar zu machen.

Auch die Gründungsideen der CLS verdanken sich solchen Impulsen zur Rationalisierung und Empirisierung der Literaturwissenschaft (Lauer 2019). Dabei erscheinen die Erwartungen, die CLS möge die Literaturwissenschaft rationalisieren und empirisieren, unterkomplex, bedenkt man, dass die informationswissenschaftlichen Fächer, die längere Erfahrung mit maschinellen Verfahren haben, nicht nur auf Rationalität und Empirie, sondern auch auf das mehr oder minder methodisch kontrollierte Herstellen von computationellen Modellen und Werkzeugen zielen. Den Rationalisierungs- und Empirisierungseffekten stehen zunächst scheinbar Komplexitätssteigerungen und eine neue Unübersichtlichkeit entgegen. Gerade dies macht aber möglicherweise den Reiz und den Impuls für die vielen Selbstreflexionen der CLS aus. Oft nehmen diese auf spezifische Werkzeuge oder Analyseverfahren Bezug; eine tatsächliche Bilanz des Erreichten fällt schwer.

Auch deshalb und parallel zu ähnlichen Phänomenen in anderen Fächern werden möglicherweise Rufe nach rigorosen, wissenschaftlichen Evaluationen lauter: Die empirische und experimentelle Psychologie etwa konstatiert für ihr

Feld eine Reproduktions- oder Replikationskrise (Yong 2018; Piper 2019a; Clayton 2020). Dort hat sich offenbar gezeigt, dass einige Studien, die andere und für das Fach zentrale Studien nachzubauen suchen, nicht gelingen. In der Folge stehen mehrere Fragen im Raum: die Frage nach der wissenschaftlichen Validität der jeweiligen psychologischen Thesen ebenso wie diejenige nach der Kontextabhängigkeit der auf sie bezogenen psychologischen Studien.

Den DH selbst entstammt die Idee, einen ‚*new rigor*' walten zu lassen, um neue Evaluationsideen für DH-Projekte, DH-Publikationen und DH-Personal zu entwickeln (Parham 2018; Klein 2019). Die Anglistin, Film- und Medienwissenschaftlerin Marisa Parham trug diese Idee im Jahr 2018 vor und zielte damit auf einen großangelegten und produktiven Neuentwurf der Geisteswissenschaften: auf einen Entwurf, der die Möglichkeiten dieser Fächergruppe hinsichtlich neuer Verfahren ebenso austestet wie ihr Versprechen, innovativ, experimentell, kritisch und grenzüberschreitend zu agieren. Parhams Fragen lauteten (Parham 2018, S. 683): „What constitutes the terrain of today's academy? Who, actually, do we want to be able to be, and how might assessment practices support that growth?"

Diese sinnvollen und hilfreichen Fragen, die mit ‚*new rigor*' nur verknappt gekennzeichnet sind, wurden durch eine Diskussion über die CLS im Jahr 2019 vorschnell und in desillusionierender Weise beantwortet. Auf der einen Seite stand Nan Z. Da, eine Spezialistin für die Kritische Theorie, amerikanische und chinesische Literatur, die in der bekannten Zeitschrift *Critical Inquiry* einen CLS-kritischen Artikel veröffentlichte (Da 2019); der renommierte Literatur- und Rechtswissenschaftler Stanley Fish sekundierte ihr (Fish 2019). Auf der anderen Seite fanden sich die Vertreter*innen der CLS, die im Online Forum von *Critical Inquiry*[1] und in anderen Periodica widersprachen. Der Streitwert war hoch und betraf gleich mehrere zentrale Aspekte der CLS. Da attestierte den CLS einen undifferenzierten Literaturbegriff, hielt selbst jedoch an einer romantischen Variante desselben fest (Jannidis 2019). Sie zweifelte die empirische Validität computationeller Verfahren an, weil sie sich mehr oder minder auf *data mining* beschränkten, nicht auf statistische Signifikanz zielten und sprach explorativen Verfahren der Datengewinnung generell wissenschaftliche Relevanz ab – ein Urteil, das verstört, bedenkt man die empirische Relevanz explorativer Verfahren (Algee-Hewitt 2019; Jannidis 2019; Piper 2019b). Dabei bezog sich Da außerdem auf eine nur geringe empirische Basis von insgesamt acht Studien, die sie selektiv auswählte und darstellte (Bode 2019a,b; Underwood 2019a). Da klagte die CLS an, vor allem rhetorische Floskeln zu produzieren, viele Fördermittel und zu viel

1 https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/

Publikationsplatz zu beanspruchen. Sie wollte Zuwendungsgebern und Herausgebern deshalb Kriterien für die Bewertung von Anträgen oder Studien aus den CLS an die Hand geben – ein vermessenes Unterfangen, auch vor dem Hintergrund der bedenklichen Beweisführung gegen die CLS (Jannidis 2019; Piper 2019b).

Dieser massive Angriff kehrte sich aus den genannten Gründen gegen die Angreifende und war letztlich wenig fruchtbar (Herrmann et al. 2019). Doch zugleich legten die Einwände gegen Da ein weiteres Mal und konzentriert zahlreiche Herausforderungen für die CLS offen, die – verknappt formuliert – mit Fragen der Empirisierung und mit dem ‚*new rigor*' zu tun haben. Ich will einige davon aufgreifen, sofern sie Arbeitsprozesse der CLS betreffen. Die Leitfrage ist dabei, inwiefern Arbeitsprozesse der CLS zur Empirisierung der Literaturwissenschaft beitragen – und inwiefern sie auch gegenläufige Tendenzen anstoßen. Solche gegenläufigen Tendenzen können u. a. als Verstärkung von Unübersichtlichkeit und unangemessenem Reduktionismus beschrieben werden. Den Aspekt der Rationalisierung der Literaturwissenschaft durch CLS will ich hier zunächst zurückstellen, denn dieser weist in etwas andere Richtung.

Um Arbeitsprozesse der CLS in Hinblick auf ihre Empirisierungsleistungen zu untersuchen, will ich der Debatte über die verschiedenen Begriffe des *reading* einen weiteren beifügen: den Begriff *reading with the workflow* (Richter 2017a), der einen genuinen Beitrag computationeller Ansätze zur Literaturwissenschaft auszudrücken erlaubt. Es geht mir dabei nicht darum, den vielfach verhandelten Objektumgangsnormen für das digitale Traktieren von Literatur eine weitere beizufügen. Weder sollen also *close*, *micro*, *deep*, *distant*, *wide*, *macro* oder *scalable* als optimale Formen des *reading* empfohlen werden (Schruhl 2014), sondern vielmehr geht es um die empirische Frage, wie – pauschal gesagt – Lesen, Analyse, Interpretation und maschineller *workflow* so miteinander einhergehen (können), dass sie empirische Antworten auf literaturwissenschaftliche und vielleicht sogar neue Forschungsfragen ermöglichen können. Allerdings ist nur eine Teilmenge literaturwissenschaftlicher Fragen auf diesem Wege beantwortbar, sodass CLS-Projekte eingangs zunächst abwägen müssen, was genau für den computationellen Ansatz operationalisierbar ist. ‚Empirie' kann dabei in unterschiedlichen Stadien der CLS vermutet werden und zwar sowohl *in praxi* – also beim Interpretieren und Programmieren – als auch beim Beobachten beider Vorgänge, beim makrologischen Blick auf ein großes Textkorpus ebenso wie bei der mikrologischen Sichtung eines kleinen Korpus.

1 Was ist und wozu dient Empirisierung?

Der Empiriebegriff seinerseits ist unscharf und zwar sowohl hinsichtlich seiner Intension und Extension. Wer empirisch arbeitet, will Annahmen über einen Gegenstand oder Sachverhalt prüfen, die in einem mehr oder minder expliziten Sinne theoretisch fundiert und in hypothetischer Form formuliert sind. Forschung unter den Auspizien der Empirie erweist sich dabei vor allem als Versprechen, etwas zu leisten, was für die Literaturwissenschaft in gewissen Masse erstaunlich oder anders ist: die insistierende Frage nach der Genese der jeweiligen Daten, nach ihren Zusammenhängen und nach ihrer Überprüfbarkeit durch eine *scientific community*, zu der Einzelne kooperativ ihre jeweilige Expertise beitragen. ‚Theorie' allerdings fungiert dabei nicht als Gegenbegriff zu Empirie, sondern Empirie und Theorie gehen vielmehr ein komplexes Wechselverhältnis ein. Lässt man die zahlreichen Kontroversen von Wissenschaftstheorie und -philosophie beiseite, kann man dieses Wechselverhältnis so beschreiben: Empirisch lässt sich testen, inwiefern eine Theorie oder Hypothese zutrifft oder falsifiziert werden muss, inwiefern sie oder bestimmte Annahmen zu korrigieren sind oder eine andere Theorie oder Hypothese zu entwickeln ist. Empirische Ergebnisse können dabei Impulse liefern, die ihrerseits aber zu interpretieren sind.

Suchte man Gegenbegriffe zur ‚Empirie', so wären dies Begriffe aus dem Spektrum subjektiver Eindrücke. Die Lektüre von Literatur setzt viele solcher Eindrücke frei: etwa, wenn Leser über ironische Textstellen in Christoph Martin Wielands Romanen lachen oder sich beim Lesen von Mary Shelleys *Frankenstein* angesichts des Mordes an Victor Frankenstein durch seine Kreatur gruseln. Literaturwissenschaft aber muss solche Eindrücke beobachten und untersuchen, wenn sie sich vom Genuss- oder Gebrauchslesen unterscheiden und eben Wissenschaft sein will. Für die Literaturwissenschaft der 1950er- und beginnenden 60er-Jahre lag die Reaktion auf diese Anforderung auf der Hand: Das meisterhafte und zunehmend metasprachliche Interpretieren von Texten war als die wissenschaftliche Leistung im Fach zu bewerten. Verglichen mit soziologischen und psychologischen Ansätzen des ausgehenden 19. und frühen 20. Jahrhunderts war dieses Verständnis allerdings eng (Richter 2010), was einigen Kollegen auffiel, die deshalb einen auch polemischen Begriff von empirischer Literaturwissenschaft dagegenstellten. Sie schlugen bislang wenig beachtete Einheiten für das literaturwissenschaftliche Studium vor, nämlich die Leser*innen. Norbert Groeben, damals Professor für Allgemeine Psychologie und Psycholinguistik an der Universität Heidelberg, entwarf die *Rezeptionsforschung als empirische Literaturwissenschaft* (Groeben 1977).

In seiner Monografie *Grundriss der Empirischen Literaturwissenschaft* beklagte der Germanist und Kommunikationswissenschaftler Siegfried J. Schmidt (1991), Groebens Ansatz (1977, 1982) fortentwickelnd, das aus seiner Sicht eklatante faktische Empirie- (und Theorie-)defizit der Literaturwissenschaft. Im Ausgang der Theoriediskussionen der 1970er-Jahre vertrat er die Auffassung, die Literaturwissenschaft sollte sich nicht mehr als hermeneutische Geisteswissenschaft, sondern als empirische Sozial- und Medienwissenschaft verstehen. Einen ausgefeilten Empiriebegriff erarbeitete Schmidt jedoch nicht, sondern vielmehr entnahm er diesen der Sozialwissenschaft. Eingehend befasste er sich hingegen mit der Frage, wie sich Theoriestrukturen in der Literaturwissenschaft explizit machen lassen sollten, wie theoretische Aussagen empirisch prüfbar sein und gesellschaftliche Relevanz erhalten könnten. Fragen nach dem Sozialsystem Literatur, seinen Aktanten und Entwicklungen standen im Vordergrund.

Damit aber litt die empirische Literaturwissenschaft wiederum unter einem Defizit, nämlich dem eines zwar emphatisch vorausgesetzten, aber wissenschaftstheoretisch und auch praxeologisch nicht weiter ausbuchstabierten Konzepts von Empirie. Empirie war in gewisser Weise eine vielversprechende Leerformel, die in den 1990er-Jahren durch konkrete Ansätze wie diejenigen der empirischen Psychologie oder der systemtheoretisch orientierten Soziologie substituiert wurde. Auch war die Gegenüberstellung von empirischen und nicht-hermeneutischen, also strukturanalytischen, semiotischen oder in anderer Weise linguistisch orientierten Verfahren der Textanalyse (Ort 2019, S. 104–122, hier 107) wenig hilfreich, weil die empirische Literaturwissenschaft durchaus potenzielle methodische Allianzen hätte suchen können.

Im Ausgang von diesen kontroversen Diskussionen näherten sich die Positionen einander an, differenzierten sich aber zugleich auch aus. Für die textorientierte Literaturwissenschaft arbeitete Harald Fricke 1986 unterschiedliche Formen von Erfahrung heraus, um die Breite literaturwissenschaftlichen Erfahrungswissens zu verdeutlichen und wissenschaftlich fruchtbar zu machen. Er unterschied zu diesem Zweck ‚philologische', ‚historische' und ‚experimentelle' Erfahrung (Fricke 1986). Aus Anlass des siebzigsten Geburtstages des auch empirisch arbeitenden Literaturwissenschaftlers Karl Eibl nahmen seine Schüler*innen u. a. auf die Unterscheidung Frickes Bezug, um drei Formen der Empirie in der Literaturwissenschaft zu identifizieren: Die „Empirie des Textes, eines ‚Kontextes' oder [...] anderweitiger Erfahrung" (Ajouri et al. 2013, S. 12). Unter Textempirie fallen danach Fragen der Textgenese und Textkonstitution ebenso wie Aspekte der Textstruktur, die sich sowohl mit Hilfe des *close reading* als auch durch standardisierte, darunter u. a. maschinelle Verfahren bearbeiten lassen. Die Empirie von Kontexten bezieht sich auf die zumeist historische Umwelt, die für die Analyse und Interpretation eines Textes relevant ist. Dazu gehören auch Fragen der

Quellenkritik. ‚Anderweitige Erfahrung' lässt sich aus interdisziplinären Studien ermitteln, die u. a. Aspekte der Rezeption mit Hilfe experimenteller Studien behandeln können (Ajouri et al. 2013, S. 12–17).

Parallel dazu arbeitete die rezeptionsorientierte, sich als empirisch bezeichnende Literaturwissenschaft ihrerseits auf eine Öffnung der eigenen Ansätze hin. Die von Schmidt im Jahr 1987 gegründete *International Society for the Empirical Study of Literature* (IGEL) hat die Gegenüberstellungen von traditionellem *close reading* und empirischer Literaturwissenschaft weitgehend abgebaut. Zum einen definiert sie Literatur breit: „as all cultural artifacts that embody literary devices, such as narrative genre, stylistic variations, and figurative language. The domain includes novels, short stories, and poetry, but also theater, film, television, and digital media" (IGEL Society 2018). Zum anderen setzt sich die IGEL heute umfassend zum Ziel, wissenschaftliche Methoden zur Untersuchung von Struktur und Funktion der Literatur, „especially its aesthetic function", anzuwenden (IGEL Society 2018).

Historisch betrachtet hat sich der Empiriebegriff der (empirischen) Literaturwissenschaft also erweitert. Darüber hinaus hat er sich vielfach ausdifferenziert: Die Medienwissenschaft ist zu einer eigenen Disziplin geworden. Aus der Kultursoziologie heraus hat sich ein kleines Feld der soziologisch inspirierten empirischen Literaturwissenschaft entwickelt, die sich für Handlungsfelder der Literatur wie für Institutionen und Organisationen des Literaturbetriebs interessiert – von der Buchmarkt- bis zur Preisforschung. Darüber hinaus fragt eine anthropologisch orientierte Literaturwissenschaft nach der Varianz literarischer Wahrnehmung in unterschiedlichen Kulturen. Außerdem ist, von der Wissenschaftsforschung ausgehend, ein eigener Zweig der Wissenschaftsforschung entstanden, der Praxeologie des Faches und analytische Literaturwissenschaft im Sinne eines Studiums literaturwissenschaftlicher Argumentationen und Begrifflichkeiten betreibt.

Die psychologisch interessierte Literaturwissenschaft hat sich ihrerseits ausdifferenziert. Mindestens drei Richtungen lassen sich identifizieren: erstens die vor allem psychologische Leseforschung, die aufgrund der Anforderungen des medialen Wandels hin zu digitalen Leseformen auch in der Öffentlichkeit gefragt ist, zweitens die Cognitive Poetics, die sich an die kognitive Linguistik anlehnt und literarische Texte mit Hilfe kognitionswissenschaftlicher Ansätze analysieren und interpretieren will. Als dritte Ausrichtung hat sich die empirische Ästhetik (mit einem literaturwissenschaftlichen Zweig) entwickelt, die den Bogen psychologischer Ansätze bis hin zur Neurowissenschaft spannt, u. a. verhaltenspsychologisch und experimentell vorgeht (Leder, Belke et al. 2004; Leder und Nadal 2014). Hinsichtlich der Literatur zielt sie weniger auf die Texte selbst als vielmehr auf die ‚literarische Erfahrung', die Menschen beim Lesen oder, weiter formuliert: Wahr-

nehmen von Literatur machen (Referat Forschung Deutsches Literaturarchiv Marbach 2019).

Nimmt man empirische Praktiken in der Literaturwissenschaft und die sich dezidiert als empirisch verstehende Literaturwissenschaft für eine möglichst breite Arbeitsbeschreibung von ‚Empirie' in der Literaturwissenschaft zusammen, dann lässt sich ‚Empirisierung' in einem weiten Sinne sowohl auf die (1.) Objekte und (2.) Ziele als auch auf die (3.) Verfahren bzw. ‚Methoden', denen sich ein Beitrag zuordnet, und (4.) Ergebnisse der Literaturwissenschaft beziehen. Hinzu kommen die (5.) faktischen, spezifische Verfahren auch unterlaufende Praktiken des Forschens, seine Voraussetzungen, Präsentationsformen wie der (6.) Habitus der CLSler*innen selbst – Aspekte also, die der Empirisierung Vorschub leisten, aber auch Gegenteiliges bewirken können. Die Reflexion solcher Empirisierungsstrategien kann dabei helfen, sie produktiv einzusetzen, konzeptionell und auch methodisch immer wieder auf den Prüfstand zu stellen.

Die CLS nehmen nun ihrerseits zumeist vom Objekttypus der Textempirie (mitunter aber auch am Objekttypus der Rezeptionsempirie) ihren Ausgang. Die Ziele der CLS sind mit denjenigen der (empirischen) Literaturwissenschaft in Teilbereichen deckungsgleich, bzw. sie könnten es sein, wobei sie darüber hinaus und um die eigene Arbeit zu ermöglichen, nicht nur bestimmte Verfahren, sondern eben auch je spezifisch zu diskutierende Tools für die Textanalyse einsetzen und dabei auf Standards aus anderen Fächern zurückgreifen, die nicht unbedingt empirisch sind. Die Entwicklung solcher Tools und das Prüfen der mit ihnen ermittelten Arbeitsergebnisse ist ein komplexes Thema. Hinzu kommen die sich in den CLS einspielenden Praktiken und der für sie charakteristische Habitus.

2 Empirische Anteile der CLS

Von Teilen der Informatik hat die CLS den lässigen Habitus des „*toolsmith*" geerbt (Brooks 1977, S. 1), der an der *usefulness* seiner Computer, Algorithmen oder Softwaresysteme bastelt. Der Habitus des ‚*toolsmith*' schlägt sich im Kleidungsstil nieder. Es paart sich dieser Habitus mit dem des Denkers, der einerseits programmierend ‚mit der Hand am Arm', zugleich aber mit dem Kopf tätig sein will, um die Welt in ihrer Komplexität zu erfassen oder virtuell mögliche Welten zu schaffen. Die ‚Inklusionsaffinität' der CLS widerspricht in manchem der exklusiven ‚Aura' der analogen Literaturwissenschaft (Willand 2017, S. 78). Schon in der persona der CLSer*innen ist ein Mix von Eigenschaften und Erwartungen angelegt, die nicht alle gleichzeitig und in gleichem Umfang realisiert werden können, zugleich aber Dynamik versprechen. Aus literaturwissenschaftlicher Sicht kann

solche Selbstbeschreibung Effekte der Empirisierung freisetzen: Die Literaturwissenschaft und ihre Akteure werden gewissermaßen neu konfiguriert, und zwar aus dem Versuch heraus, ihre Objekte hinsichtlich der quantitativen und qualitativen Komponenten zu vermessen, die maschinellen Arten der Analyse und ihrer Modellierung zugänglich sind.

Diese Neukonfigurierung verspricht der Rhetorik und dem Bekenntnis nach eine Maximierung von Möglichkeiten, die Uprichard (2014) polemisch, aber vergleichsweise umfassend wie folgt summiert: „predict; steer, shape; harvest, harness, mine; sort, store, synthesize; track and trace; innovate and transform; optimize, maximize, visualize; and so on.“. Konstruktiv reformuliert weisen Tätigkeitsbeschreibungen wie diese jeweils in unterschiedliche Richtungen, die mit Versprechen der Empiriserung zu tun haben. Auf die CLS treffen sie in ebenso unterschiedlichem Maße zu: Das ‚*predict*‘ zielt auf empirisch begründete Vorhersagen, etwa aus der Analyse von Datenkomplexen. Für die CLS kommen diese nur bedingt in Betracht. Zentraler sind andere datenorientierte Dimensionen der CLS, die sich vornehmlich auf Aspekte der Textempirie sowie auf die Methodik beziehen:

1. *Identifikation und Explikation von (Text-)Daten*. Die CLS haben dazu beigetragen, dass die Literaturwissenschaft zunehmend ihre Kladden öffnet. Textanalysen, Exzerpte u. dgl., die zuvor als Beiwerk der Literaturwissenschaft im Vorfeld der Interpretation galten, bekommen einen eigenen Stellenwert. Denn bevor ein maschinengestützter *workflow* ansetzen kann, gilt es zu beschreiben, was überhaupt maschinell analysierbare literaturwissenschaftliche Daten sind und wie sie sich ermitteln lassen. Speziell der kontextabhängige Bereich der Einstellungen ist dabei kompliziert. Etwa ist es äußert schwierig festzustellen, wo im Text ein Phänomen wie Ironie vorkommt. Es fragt sich, ob Ironie überhaupt von Textmerkmalen abhängt und maschinell analysiert werden kann oder ob das Feststellen von Ironie auf Kontextmerkmale reagiert, die sich nur schwer identifizieren lassen. Andere Textmerkmale lassen sich leichter maschinell analysieren: Anführungsstriche etwa lassen sich leicht zählen, um Passagen mit wörtlicher Rede herauszufiltern. Aber wie erkennt man wörtliche Rede, die nicht durch Anführungszeichen markiert ist? Mit gutem Grund bildeten sich gewisse Schwerpunkte ausgehend von narratologischen Analysebegriffen heraus; auch das in der Computerlinguistik bewährte *topic modelling* fand oft Verwendung, um thematische und ggf. historisch relevante Themencluster in literarischen Texten zu identifizieren. Doch all diese Studien sind nur möglich durch intensive Diskussionen über die Art und Weise der Suchparameter und der Ergebnisse. Diese wiederum sind nach nachvollziehbaren – empirischen – Verfahren zu gewinnen und ent-

sprechend zu beschreiben. Literaturwissenschaftler*innen, die mit Grauzonen der Beschreibung umzugehen gewohnt sind, scheint dieses kleinteilige Verfahren der *checks and balances* oft langatmig. Zugleich aber schärft es auch das literaturwissenschaftliche Auge für Details, sodass die empirischen Aspekte der Datenidentifikation und -explikation umgekehrt auf literaturwissenschaftliche Interpretationen ausstrahlen und also zusätzlich empirische Effekte haben können, was die Genauigkeit der Analysen und Interpretationen betrifft.

2. *Auszeichnung von (Text-)Daten.* Das Auszeichnen von Daten ist eine eigene Tätigkeit innerhalb und jenseits der CLS. Dabei geht es weniger um die oft beschworenen Datenmengen (*big data*) als vielmehr um die Qualität von Daten (Rat für Informationsinfrastrukturen 2019). Gütekriterien (etwa: Validität, Reproduzierbarkeit, Dokumentation der Datenprovenienz usf.) sind dabei nicht klar formuliert, sondern vielmehr Teil eines wissenschaftlichen Prozesses, der Prozesse der Korpuserstellung, Digitalisierung, OCRisierung usf. einschließt. Dabei arbeiten verschiedene wissenschaftliche und infrastrukturelle Akteure (und solche, die beiden Gruppen angehören) parallel, leider jedoch nicht immer koordiniert. Etwa stützen sich einige Arbeiten in den CLS auch auf bibliothekarische Normdaten (erhoben nach der Gemeinsamen Normdatei im deutschen Raum oder vergleichbaren Vorgaben andernorts). Im Forschungsprozess der CLS tauchen diese Daten oft unter dem Begriff ‚Metadaten' auf, womit sachlich Autornamen, Namen des Publikationsortes oder Publikationsdatum gemeint sein können. Der Zweck der Datenauszeichnung weicht jedoch ab: Während eine Bibliothek nach dem Prinzip des Bestands agiert und vor allem das Objekt und seinen Fundort identifizieren will, interessiert sich die Forschung für die Daten selbst. Gerade aus einer engen Zusammenarbeit zwischen CLS und Infrastrukturen wäre für die Prozesse der Datengenese, der Gewinnung von Datenmengen und der Erhöhung der Datenqualität noch viel herauszuholen, was der Literaturwissenschaft in jeder Form zugute käme. Etwa wäre auch an weitere Kategorien für sogenannte Normdaten zu denken, die für literaturwissenschaftliche Fragen relevant sind und die diesen eine neue empirische Grundlage verschaffen könnten. Wie wäre es etwa, wenn man die Provenienzen antiquarischer Bücher verzeichnete und auf diese Weise der Provenienzforschung neue Wege ebnete (Bögel et al. 2015; Jessen 2019)?
3. *Zugänglichkeit und Nachhaltigkeit von (Text-)Daten.* Wie andere Geisteswissenschaften bedürfte auch die Literaturwissenschaft dafür avancierter bibliothekarischer Kataloge oder, besser: Datendienste (Schneider 2020). Solche Datendienste müssten sich zunehmend vom bibliothekarischen Bestandsprinzip verabschieden und auf das Teilen von Daten setzen, um für die

Wissenschaft geeignete Plattformen für die Suche nach ihren Daten oder Materialien zu schaffen. Speziell in Deutschland mit seinem ausgefeilten Bibliothekssystem und den Einrichtungen, die in unterschiedlicher Trägerschaft von Bund und Land sind, scheint diese Aufgabe besonders komplex. Einen umfassenden, seit den Anfängen der Literatur in deutscher Sprache gepflegten, Nationalkatalog oder dergleichen gibt es nicht – anders als etwa in Frankreich oder England, wo die großen Nationalbibliotheken dergleichen in gewissem Maße bieten. Auch kommen Vorhaben der Digitalisierung der Bestände nur langsam voran, weil die Gelder dafür fehlen. Ein zusätzliches Thema ist dasjenige der Nachhaltigkeit von Daten. Mit den in der Auswahl- und Einrichtungsphase befindlichen Konsortien der Nationalen Forschungsdateninfrastruktur hat auch die CLS die Chance, ein wissenschaftsgetriebenes Netzwerk von Forschungsdatendiensten zu entwickeln, das stärker als die bisherigen Dienste der Bibliotheken und Archive auf ihre Bedürfnisse zugeschnitten ist. Wie aber die nachhaltige Nutzung und wissenschaftsweite Teilhabe gesichert werden kann, ist noch weitgehend ungeklärt. Besonders reizvoll ist dabei die Idee, Datenlebenszyklen zu ermöglichen, die nicht nur das Erstellen, Verarbeiten und Analysieren, sondern auch das nachhaltige Nutzen von Daten umfassen. Im Rahmen solcher Datenlebenszyklen werden sich Daten noch einmal ausdifferenzieren. Sie können unterschiedliche Formen von Empirie umfassen und ermöglichen: Text- und Verstehensempirie, kuratierte Datensammlungen ebenso wie Ko-Publikationen von Daten und auf diesen Daten basierender Forschung.

4. *Verfahren der Datenanalyse und Datenentwicklung aus Texten und über Texte*. Diese Daten werden mit Hilfe von Verfahren wie *text mining*, *topic modelling*, Stilometrie, Netzwerkanalysen u. dgl. ermittelt – voraussetzungsreichen Verfahren also, die weit über die literaturwissenschaftliche Grundkompetenz hinausgehen. Verstehensfragen, wie sie auch die Literaturwissenschaft beschäftigen, sind zu komplex, um sie dem Fach allein zu überlassen, notiert Underwood (2019b) dazu so nüchtern wie korrekt. Da es angesichts der Datenmenge oft schwierig ist, statistisch valide Ergebnisse zu erreichen, gehen die CLS explorativ vor und nutzen Kontrollverfahren: Computationelle Analyseverfahren werden etwa mit Hilfe von manuellen Annotationen kontrolliert (Kuhn 2020, S. 72). Denn die maschinellen Werkzeuge sind fehleranfällig und produzieren mitunter Ergebnisse, die im Auge von Leser*innen an der Sache vorbeiführen oder korrigiert werden müssen. Dabei handelt es sich bei diesen Annotationen nicht um umfassende (‚hermeneutische') Textdeutungen oder auch: Interpretation, die etwa im traditionellen Sinne auf Vollständigkeit der Interpretation achtet oder auch nur ein kohärentes *close reading* präsentieren will, aber Annotationen können – je nach Annotationstyp – Leseein-

drücke von Individuen aufnehmen (Gius und Jacke 2017). Einige Bereiche der CLS vollziehen damit also eine doppelte Empirisierung der Textanalyse: Sie versuchen, Texten maschinell und zugleich manuell auf den Grund zu gehen, Muster und Ambivalenzen zu erkennen. Komplexe und mehrstufige Verfahren wie diese werden womöglich umso bedeutender, je schneller sich Verfahren des *deep learning*, also der künstlichen neuronalen Netze, entwickeln (Manning 2015, Kuhn 2019, S. 567). Diesen gilt Literatur als besonders interessante Domäne, weil ihr Verstehen noch mehr als das Verstehen von Normalsprache auf Kontexte und Vorwissen angewiesen ist. Zugleich hätten *deep-learning*-Verfahren erhebliche Konsequenzen für die Literaturwissenschaft, da sie die Grenze zwischen quantitativen und qualitativen Analyse- und Deutungsverfahren vermutlich weiter verschieben und unterschiedliche Typen von Empirie miteinander verbinden.

5. *Verfahren der Aufbereitung von (Text-)Daten*: Eine andere große Herausforderung, gerade auch hinsichtlich der Empirie sind Verfahren der Visualisierung. Mit ihnen ist die Literaturwissenschaft (abgesehen von Tabellen, Baumdiagrammen o. dgl.) bislang wenig vertraut. Tatsächlich ist der Einsatz von Visualisierung im Fall von Literatur voraussetzungsreich: Ihre Textdaten sind multidimensional und unstrukturiert (Gius und Petris 2015); eine Visualisierung setzt bereits eine Strukturierung der Daten, also auch gewissermaßen eine Interpretationsentscheidung voraus. Visualisierungen können „die Präsentation, die konfirmative Analyse sowie die explorative Analyse“ umfassen oder für das *storytelling* anhand von Daten eingesetzt werden (Jannidis et al. 2017, S. 331). Sie changieren dabei zwischen Datenpräsentation, Interpretation und *enabling technology*, insofern sie (im Sinne der *visual analytics*) durch ihre Strukturierung bestimmte Analyse- und Deutungsergebnisse erst nahelegen. Dabei lassen sich Grade der Unsicherheit, auf denen visuelle Modelle beruhen, produktiv machen, und zwar sowohl für die Visualisierung als auch für das Erkunden der Textempirie (John et al. 2017). Mitunter erkennt man aus einer Visualisierungsform Zusammenhänge, die sich beim linearen Lesen nicht ergaben. Das Theodor Fontane Archiv etwa hat die Benutzungsspuren und Glossen von Theodor Fontane aus dem Bestand seiner Bibliothek digital zugänglich gemacht und so visualisiert, dass man die möglichen Quellen von Fontanes Romanen beinahe per Mausklick ausmachen kann (Dörk et al. 2019).

3 Zur Empirie von Interpretationen und dem produktiven Irritationspotenzial der CLS

Die beschriebenen empirischen Dimensionen der CLS fließen jedoch nicht ohne Weiteres in die Literaturwissenschaft ein. Zwar ist deutlich, dass bessere Kataloge, Datendienste und gut angelegte Plattformen der Textbeschaffung sowie auch der Dokumentation aufhelfen, aber mehr folgt daraus zunächst nicht. Blickt man auf das Kerngeschäft der Interpretation, ist der Hiat zwischen analoger Literaturwissenschaft und CLS erheblich, zugespitzt formuliert: Maschinelle Textanalysen können zwar interessante Ergebnisse liefern, die für sich genommen aber nur Indizien dafür sind, dass sich mit Hilfe einer entsprechend fokussierten Interpretation etwas aus einem Text herausholen ließe. Die Ergebnisse solcher Analysen haben selbst keine Erklärungskraft. Mit ihnen verbindet sich zunächst einmal keine These oder Theorie, die sich prüfen und ggf. modifizieren ließe.

Doch ist das nicht nur charakteristisch für die CLS, sondern auch in der analogen Literaturwissenschaft gibt es einen Hiat zwischen Theorie und Textinterpretation. Wenn die Theorie (oder, vorsichtiger formuliert: der Ansatz) etwa besagt, dass Literatur einen besonderen Beitrag zur Debatte über die Relevanz von Emotionen liefert, weil sie es erlaubt, diese in besonderer und intimer Weise zu schildern, und die These lautet, dass Goethes *Werther* ein bedeutender Kandidat für solche Schilderungen ist, dann klaffen zwischen Theorie, These und Text eine oder mehrere Lücken. Sie werden mit Hilfe einschlägiger Analyse- und Interpretationskonzeptionen gelöst, die metasprachlich und aus der Kenntnis von Praktiken des Interpretierens darüber informieren, wie man von Textbeobachtungen zum Belegen oder Testen von Theorien und Thesen gelangt.

Dabei allerdings lässt sich in der Literaturwissenschaft von einem weichen Verständnis von solchem Belegen oder Testen sprechen, das zumeist auf Anschlusskommunikationen aus ist. Ausgehend von der These, dass literarische Texte, wenn sie ästhetisch taugen, polyvalent sind (Jannidis 2003), setzt die Literaturwissenschaft auf das Prinzip unabschließbarer Interpretation großer Werke. Diese Werke hält sie zwar für singulär, zugleich behandelt sie aber auch Gattungen oder andere Korpora. Die Interpretation von einzelnen Werken wie von Korpora basiert ihrerseits auf Textanalysen, die zumeist qualitativer Art sind, also etwa den Umgang mit Sprache, Raum, Zeit, Figuren, Handlung, Perspektiven in einem Text betreffen. Dabei können jedoch auch quantitative Aspekte wie mehrfache Raumwechsel, das wiederholte Auftreten einer Figur im Drama oder redundante Handlungen eine wichtige, die Interpretation stützende Rolle spielen.

Das Verhältnis zwischen qualitativen Aussagen oder, allgemeiner formuliert qualitativen Daten und quantitativen Daten ist also schon in analogen Interpretationen vermischt. Auch qualitative Textdeutungen kommen nicht ohne quantitative Beobachtungen von Textmerkmalen aus, wie die CLS ihrerseits hervorgehoben hat (Ramsay 2011, S. 16, Willand 2017). Manche Interpretationen zielen geradezu auf vermischte Methoden: auf einen Methodenmix etwa aus anthropologischen Verfahren der ‚tiefen Beschreibung' und mit literaturhistorischen Kontexten, aus semiotischen Analysen mit solchen der Literaturinterpretation usf. Was den Umgang mit quantitativen Daten betrifft, nehmen analoge Interpretationen jedoch selten numerische Befunde zum Anlass einer Interpretation. Vielmehr werden quantitative Daten u. a. als Belege für bestimmte Interpretationen herangezogen. Eine Ausnahme bilden die im Fach bislang eher marginalen Traditionen quantitativer Literaturwissenschaft (Bernhart et al. 2018). Im Zusammenhang mit der Konjunktur des Strukturalismus fanden solche Ansätze aber gleichwohl Anerkennung durch den literaturwissenschaftlichen Mainstream. Mehr noch: Sie haben sich neben hermeneutischen (kontextorientierten) Ansätzen als zweite Methodik im Fach etabliert und schlagen sich in den begrifflich wie methodisch differenzierten Ansätzen der Narratologie wie der Dramen- und Lyrikanalyse nieder.

Blickt man aus diesen Momentaufnahmen der Entwicklung von Interpretationspraktiken in den Literaturwissenschaften auf ihre computationelle Variante, dann fällt auf, dass der oft postulierte kategoriale Unterschied zwischen qualitativer und quantitativer Literaturwissenschaft vielmehr bloß graduell ist und im Fach selbst auf bereits eingespielte, wenn auch erweiterbare und methodisch weiter zu reflektierende Praktiken trifft. Doch gelingt es nicht ohne Weiteres, Ansätze der CLS damit zu verbinden. Oft bedarf es einer gestaffelten Methodik. Durch die Staffelung kann mit jedem Schritt methodisch etwas anders angesetzt werden; jeder einzelne Schritt lässt sich durch Hypothesen verbinden. Der analogen Literaturwissenschaft mag die dafür notwendige Dokumentation auch kleiner Schritte und die Reflexion methodischer Entscheidungen unelegant und umständlich erscheinen, aber möglicherweise überzeugen erste Ergebnisse oder regen an, auf dieselbe oder auf andere, gegebenenfalls auch analoge Weise oder im Sinne von Methodenkombinationen weiterzuarbeiten. In jedem Fall trägt das Dokumentieren auch inkrementaler Schritte dazu bei, Ergebnisse replizierbar und d. h.: empirisch überprüfbar zu machem.

Möglich wird diese Staffelung auch, weil die epistemologischen Ziele beider, der analogen Literaturwissenschaft und der CLS, zu einem erheblichen Teil deckungsgleich sind, auch wenn ihre Verfahren und Begrifflichkeiten teilweise voneinander abweichen. Auch der CLS geht es um die angemessen komplexe Beschreibung von Texten und Korpora. Anders als die Informatik nimmt sie auf das Kriterium der *usefulness* vornehmlich Bezug, wenn sie ihre Werkzeuge beur-

teilt: Diese sind nützlich zum literaturwissenschaftlichen Zweck der Datensammlung oder -analyse oder nicht. Viele Studien mit dem Ziel der tiefen Textanalyse schließen dabei an etablierte literaturwissenschaftliche Fragestellungen an. Sie widmen sich etwa der Untersuchung von Figuren und Figurenkonstellationen im Drama (lina.digital[2]; QuaDramA/Q:TRACK[3]), der Figureninteraktion, bestimmten Plot-Formationen wie dem ‚happy end' (Zehe et al. 2016), der Beschreibung von Themen in der Novelle des 19. Jahrhunderts (Weitin 2017) usf.

Andere Arbeiten der CLS weisen mehr Irritationspotenzial für die Literaturwissenschaft auf und dieses Irritationspotenzial speist sich u. a. aus den fünf zuvor skizzierten Empirisierungsleistungen der CLS: Die identifizierten und ausgewerteten Daten entsprechen in solchen Fällen nicht den in der Literaturwissenschaft erwarteten Interpretationsleistungen. Vielmehr arbeiten sie mit Daten, die bislang eher mit Nachbardisziplinen wie der Kultursoziologie, der Buch- oder Übersetzungswissenschaft verbunden wurden. Laura McGrath beispielsweise konnte in einer Studie zu Verlagsempfehlungen zeigen, dass 478 der 500 für Vertreter und Buchhändler meistempfohlenen Bücher New Yorker Verlage zwischen 2013 und 2019 von weißen Autor*innen geschrieben wurden (McGrath 2019). Berit Glanz und Nicole Seifert zeigten in einer Studie über Verlagsprogramme aus dem Jahr 2018, dass Frauen dort weniger häufig vertreten sind (Glanz und Seifert 2019). Sie lösten damit eine kontroverse Debatte über Sinn und Zweck solcher Zählungen, über literarische Werte, große und kleine Werke und die Bedeutung von Genderrepräsentanz in Verlagsprogrammen aus.

Diese Dimensionen umfassen jeweils Aspekte empirischer Forschung in der Literaturwissenschaft: Die CLS stärken die Reflexion der Ziele, Voraussetzungen, der Verfahren und Ergebnisse von Literaturwissenschaft im Sinne von wissenschaftlicher Kontrolle, jedenfalls im Idealfall. Doch ist diese Stärkung aus mehreren Gründen nicht notwendig unilinear. Ein erster Grund dafür ist pragmatischer Natur. Die Ansätze der CLS scheitern mitunter daran, dass es speziell aus dem Feld der deutschsprachigen Literatur ebenso wie der normalsprachlichen Textquellen noch zu wenig digital verfügbare Texte gibt. Auf diese Weise haben die CLS hier nur geringe Chancen, ihre Ansätze auszuprobieren. Zwar ist auch die tiefe maschinelle Analyse einzelner ‚großer' Texte reizvoll, aber speziell komplexe statistische Verfahren bewähren sich erst an größeren Textmengen (Underwood 2019c). Darüber hinaus bedarf es, je nachdem, was genau man wissen möchte, der Vergleichskorpora. Wer etwa untersuchen möchte, inwiefern Literatur von der Alltagssprache oder privat und öffentlich diskutierten Themenfeldern abweicht, be-

2 https://dlina.github.io

3 https://quadrama.github.io

nötigt alltagssprachliche Korpora, etwa digitalisierte Zeitungen oder andere Periodika. Zwar bilden diese nicht den Durchschnitt des Ausdrucksvermögens der Sprecher des Deutschen ab, aber sie helfen doch einen Schritt weiter, will man Literatur und Alltagsprosa korrelieren und auf diese Weise auch auf die jeweiligen Besonderheiten von Literatur schließen.

Zweitens – und dieser Aspekt ist systematisch essenziell – schaffen die CLS Daten, für deren Reflexion die Literaturwissenschaft noch keine Standards oder *best-practice*-Beispiele entwickelt hat. Dadurch entsteht methodisch ein erheblicher Erfassungs-, Beschreibungs- und Vergleichsdruck, auf den manch analoge Literaturwissenschaftler*in reserviert reagiert. Vorwürfe lauten hier gewöhnlich, dass die CLS letztlich unbrauchbare Datenmengen erzeugen, die dem methodischen Kern des Faches, der Textinterpretation, nicht weiterhelfen. Der Vorwurf überlappt dabei mit dem alten Positivismus-Vorwurf (‚Datensammeln sei geistlos'), der hier bloß aktualisiert wird. Umgekehrt fällt durch die CLS und beim Arbeiten in diesem interdisziplinären Gebiet auf, dass literaturwissenschaftliche Fragen oft, um nicht zu sagen: notorisch unscharf und schwer operationalisierbar sind. Gerade aus dem komplexen Empirismus der CLS aber erwachsen reizvolle Anforderungen, nicht nur für die Literaturwissenschaft, sondern auch für die DH oder neuere *deep-learning*-Ansätze.

4 *Reading with the workflow*

Ausgehend von dieser Beschreibung produktiver Irritationspotenziale will ich zur der Frage beitragen, wie CLS und analoge Literaturwissenschaft mehr als bisher zusammenkommen können, um Ziele, Verfahren und Ergebnisse von Textanalysen zu reflektieren, die Analyse und Interpretation von Literatur also zu empirisieren (siehe dazu auch den Beitrag von Pichler und Reiter (2020), ab Seite 43 in diesem Band). Es geht mir darum, ein Miteinander unterschiedlicher Fachrichtungen in dynamischen Situationen des Analysierens und Interpretierens von Texten zu beschreiben. Dieses Miteinander nenne ich ‚*Reading with the workflow*' – ein Begriff, den ich an anderer Stelle bereits gebraucht habe, dort aber nur kursorisch ausführen konnte (Richter 2017a, S. 25).

Den Ausgangspunkt dafür bildet der Zweifel an der Produktivität bestimmter Objektumgangsnormen (Schruhl 2014). Ob man sich nun zu dem von Franco Moretti ausgerufenen ‚*distant reading*' oder zum bloß scheinbar verzopften *close reading* bekennt, scheint mir sekundär. Zum einen fallen unter interpretationstheoretische Schlagworte wie diese je unterschiedliche Analyse- und Interpretati-

onsverfahren. Zum anderen scheint es mir relevanter, das für die jeweilige Fragestellung beste Verfahren zu wählen oder Verfahren zu kombinieren.

Der Einsatz maschineller Verfahren setzt dabei ein initiales Abwägen von Aufwand und Ertrag voraus. Will man eine komplexe und zumeist kooperative Versuchsanordnung riskieren und, wenn ja, wie genau soll diese verlaufen? Sollen analoge und computationelle Literaturwissenschaft parallel zueinander agieren, besteht das Risiko, dass ihre Ergebnisse schlussendlich unverbunden nebeneinander stehen. Das mag im Einzelfall lehrreich sein, weil so die Heterogenität computationeller und analoger Ansätze augenfällig wird, ist für dauerhafte Praktiken der CLS jedoch wenig erfolgversprechend. Ein integraler Ansatz ist in vielen Fällen hilfreich, wenn auch aufwendig. Er bedarf der Abstimmung von Vorstellungen und Methoden, die nicht deckungsgleich sind. *Reading* und *workflow* sind nicht entweder literaturwissenschaftlich oder computationell, sondern beides auf einmal, wenn auch in unterschiedlichem Maße und auf unterschiedliche Weise. Beim *reading* sind Chancen und Grenzen des jeweiligen *workflow* zu bedenken, beim *workflow* die Forschungsfragen und Wahrnehmungen des *reading* einzubauen. Zwar greifen dabei beide Prozesse ineinander, aber sie driften zugleich in unterschiedliche Richtung, geleitet durch die Dynamik des Datenmaterials einerseits, der Verfahren und *tools* andererseits. Das Ergebnis ließe sich als eine Art komplexe Balance aus unterschiedlichen Arbeitsschritten beschreiben, die an bestimmten Punkten zu heuristisch interessanten Schnittmengen kommen.

Um diese Überlegungen zu veranschaulichen, will ich knapp zwei unterschiedliche Ansätze skizzieren, um Goethes *Werther* zu untersuchen. *Werther* zählt zu den Texten, an dem einige Digital Humanists bereits ihre Instrumente erprobt haben. Andrew Piper und Mark Algee-Hewitt beispielsweise fragten – ausgehend von typischen Wörtern aus Goethes *Werther* – nach einem ‚*Werther Effect*' für Goethes Werk selbst wie für den nachfolgenden literarischen Diskurs (Piper und Algee-Hewitt 2014). Gemeinsam mit dem studentischen Team „German Literature in the World" konnte ich selbst anhand von Metadaten zu Übersetzungsorten und -jahren die globale Verbreitung von *Werther*-Übersetzungen kartografisch darstellen (Richter 2017b, Richter 2017a, S. 495).

Hier will ich in der gebotenen Kürze zunächst eine hypothetische Studie skizzieren, die von einem numerischen Befund ausgeht und diesen ex post mit literaturwissenschaftlichen Forschungsfragen verbindet und eine mögliche mehrstufige Studie entwickelt. Der zweite Ansatz entstammt einem mehrstufigen Projekt zu Goethes *Werther* und einigen sogenannten Wertheriaden im Rahmen des Stuttgarter *Center for Reflected Text Analytics*.

4.1 Studie 1

Die hypothetische Studie orientiert sich an einer bemerkenswerten Erschließungs- und Aufbereitungsleistung, nämlich dem Deutschen Textarchiv, das TEI-konforme Texte mit dem Analysetool Voyant verbindet. Das DTA hat die zweibändige Erstausgabe des Goethe-Romans *Die Leiden des jungen Werthers, Leipzig* (1774) gewählt. Zwar führt schon dies zu einer selektiven Wahrnehmung des Textes: Die durch den Autor selbst veränderte und entschärfte Version der *Leiden des jungen Werther* (ohne Genitiv-s) aus dem Jahr 1787 gerät dabei ins Hintertreffen. Lässt man diesen Umstand aber einmal beiseite, so liefert die Analyse des *Werther* mit Voyant (also quasi: per Knopfdruck) nach der transliterierten Version ein quantitativ wie qualitativ reizvolles Ergebnis: Der meistfrequente Begriff ist ‚jch', nimmt man die Funktionswörter aus. In Band eins taucht er 128 mal auf, in Band zwei 121 mal. Danach folgen Begriffe wie ‚seyn' (Band 1: 50 mal, Band 2: 41 mal) und ‚all' (Band 1: 41 mal, Band 2: 50 mal).

Um diese Ergebnisse zu deuten, wären sie in existierende Interpretationen einzubauen, vielleicht so: Der meistfrequente Begriff des *Werthers* in seiner ersten Fassung ist ‚jch'. Dieser Befund passt zur Interpretation des Textes als Roman des Sturm und Drang – also als eines Textes, der dem ‚jch' in besonderer Weise zum Ausdruck verhilft. Doch was genau dies für den Roman bedeutet, ließe sich (im Fall von Voyant) erst durch genaue Textlektüre klären: Wo genau taucht das ‚jch' auf? Bildet sein gehäuftes Auftauchen eine Art Gelenkstelle im Text? Der Graph zum Vorkommen von ‚jch' im *Werther* nach Voyant weist auf bestimmte Textstellen hin, die durch häufiges Auftauchen des Begriffs gekennzeichnet sind. Darunter fallen vor allem Segment sechs aus dem ersten Band und Segment drei aus dem zweiten Band. Segment sechs aus dem ersten Band beschreibt den Besuch von Lotte bei einer sterbenden Dame, die ihre Gesellschaft sucht. Werthers ‚jch' wiederum sinniert anlässlich des Besuchs über den Sinn und Zweck von Geselligkeit, also nach dem Ort des ‚jch' in der Gemeinschaft. Segment drei aus dem zweiten Band enthält u. a. Werthers Dimissionsgesuch, seine Absage an die vorgesehene Karriere und seine Andeutung, dass er künftig Gesellschafter des Fürsten werde.

Beide Textstellen sind also in Hinblick auf die Thematik des ‚jch', des sich über sich, sein Umfeld und ‚die Gesellschaft' schlechthin verständigenden Subjektes reizvoll. Doch sind sie zunächst einmal nicht mehr und nicht weniger als ein Indiz dafür, dass das ‚jch' für den Text vielleicht besonders relevant sein könnte. Es bedarf verbindender Ansätze und Hypothesen, um sie mit bisherigen Interpretationen des *Werther* (also: mit dem literaturwissenschaftlichen Forschungsstand) ins Gespräch zu bringen. Die unter den vielen denkbaren Ansätzen oben angedeutete Hypothese, dass das häufige Vorkommen des ‚jch' für einen Text des Sturm

und Drang charakteristisch sein könnte, wäre dabei nicht nur aus dem *close reading* einzelner Textstellen weiter zu prüfen, sondern auch aus dem Textvergleich.

Ein solcher Textvergleich müsste andere Texte des Sturm und Drang und auch solche der Vorzeit heranziehen, um die Bedeutung der Vokabel ‚jch' im Sinne eines punktuellen Tests literarischer Semantik historisch zu prüfen. Da ein großes Vergleichskorpus, aus dem ausgewählt werden könnte, nicht zur Verfügung steht, käme etwa Christoph Martin Wielands *Geschichte des Agathon* als Vergleichstext in Betracht. Der Roman wurde 1767 publiziert und gilt als Ausdrucksform aufklärerischer Selbst- und Fremdbildung, als philosophischer Roman. Im Deutschen Textarchiv ist auch dieser Text vorhanden und zwar in der Form der Erstauflage in zwei Bänden.

Eine Voyant-Analyse des ersten Teils des *Agathon* weist das ‚jch' als fünfthäufigsten Ausdruck aus. Es kommt 188 mal vor; die Analyse des zweiten Teils zählt das ‚jch' nicht zu den fünf häufigsten Wörtern. Vielmehr fällt auf, dass der Name der Hauptfigur in beiden Bänden als zweithäufigster Ausdruck gilt (Band 1: 252 mal, Band 2: 258 mal), was im Wesentlichen an der Perspektive liegt, aus der erzählt wird. Wielands *Agathon* legt mehrere Schichten übereinander, um Distanz vom erfundenen Geschehen zu erzeugen: Der Autor präsentiert sich als Herausgeber einer biografischen Erzählung griechischen Ursprungs, deren Wahrheitsgehalt ihm selbst Rätsel aufgibt und eine längere Einlassung über den moralischen Gehalt des historischen Exempels hervorruft. Über dieses wiederum erzählt ein Erzähler, der eine eigene ironische und vermittelnde Position sucht. Goethe wiederum nutzt ebenfalls das Distanzierungsmittel der Herausgeberfiktion, um in seinem Briefroman das werthersche Ich selbst sprechen zu lassen und zu kommentieren. Nimmt man die Länge der Texte hinzu, dann macht das ‚jch' im ersten Band des *Agathon* (Gesamtlänge: 86 178 Wörter) ca. 0,22 % des Texts aus, im ersten Band des *Werther* (Gesamtlänge: 17 383 Wörter) sind es ca. 0,74 %.

In beiden Fällen ist das Vorkommen des Wörtchens ‚jch' trotz der Häufigkeitsanzeige also gering – und für die Literaturwissenschaft trotzdem interessant oder jedenfalls: nicht zu ignorieren. Jenseits der exakten statistischen Bedeutung dieses Vorkommens ließe sich aus dem explorativen Befund mehr entwickeln: eine Versuchsanleitung dazu, wie – sofern mehr Texte aus dem relevanten Bereich (Romane des Sturm und Drang und der Vorzeit) digital verfügbar wären – mit Hilfe etwa von Voyant sowie im Rahmen ergänzender Studien Aspekte der Bedeutung eines solchen Begriffs für den einen Text und benachbarte Korpora ermittelt werden könnten. Viele explorative Erkundungen wie diese bleiben noch im Konjunktiv, was im konkreten Fall an mangelnden Vergleichskorpora liegt. Liegen mehr Texte digital vor, so ließe sich ein nächster Schritt unternehmen. Einstweilen ist es aber dennoch möglich, das maschinelle Instrumentarium zunächst an der Einzeltextanalyse oder an der Analyse weniger Texte zu schärfen.

4.2 Studie 2

Im Rahmen von CRETA hat unser Projekt zu *Werther* und den Wertheriaden dies exemplarisch versucht.[4] Ausgehend von der Forschungsliteratur zu *Werther* und den Wertheriaden einerseits, der Forschung zu Formen des seriellen Schreibens andererseits, zielte das Projekt darauf, Merkmale unterschiedlicher Texte miteinander zu vergleichen, die – der Forschung zufolge – eng miteinander zusammenhängen, weil sie sich auf denselben *master text* beziehen.

Das Projekt verlief also in mehreren Stufen: In einem ersten Schritt analysierte es Goethes *Werther* in Hinblick auf seine zentralen Merkmale, darunter die Dreiecksbeziehungen von Werther, Lotte und Albert, den melancholische Werther-Charakter, die Herausgeberfiktion, die Rolle von Naturmotiven, die ‚Krankheit zum Tode' und der Suizid (Martens 1985; Horré 1997). Diese Merkmale dienten uns als Suchvorgabe, um *Wertherness* auszuzeichnen: Merkmale also, deren Vorkommen darauf hinweisen könnte, dass es sich bei einem literarischen Text bis zu einem gewissen Grad um eine affirmative oder kritische Imitation von Goethes *Werther* handeln könnte.

In einem zweiten Schritt sondierten wir ein Korpus von 150 deutsch- und 30 englischsprachigen sogenannten Wertheriaden. Ausgewählte Texte daraus untersuchten wir in Hinblick auf ihre *Wertherness*, dies beginnend mit einzelnen Merkmalen wie etwa der Dreiecksbeziehung (Barth und Murr 2017). Ausgehend von ersten digitalen Ansätzen in dieser Richtung (Moretti 2011; Trilcke 2013; Hettinger et al. 2015) untersuchten wir die Nähe von lexikalischen Einheiten wie ‚Lotte' (‚Lottgen', ‚Lottchen') zu anderen Einheiten dieser Art. Dabei konnten wir zeigen, wie die jeweiligen Dreiecksbeziehungen einander ähneln und durch die Arten und Weisen ihrer Netzwerke voneinander abweichen. Um das jeweilige Netzwerk und vor allem individuelle Figurenbeziehungen näher charakterisieren zu können, ergänzten wir einen auf *word clouds* basierenden Zugang. Er erlaubte es, auch die semantischen Komponenten der Figurenkommunikation zu erfassen: Geht es etwa (wie in den Gesprächen von Lotte und Werther) um Liebe oder ist die Kommunikation (wie in der Auseinandersetzung zwischen Werther und Albert) durch düstere Vokabeln charakterisiert (Barth, Kim et al. 2018)?

Weitere Schritte, die das Merkmalsbündel *Wertherness* zu testen erlauben, stehen noch aus. Dazu zählen u. a. manuelle Annotationen, die einzelne Aspekte von *Wertherness* durch individuelle Leseeindrücke prüfen und ergänzen sollen. Annotationen wie diese sind auch bei maschinellen Ansätzen unverzichtbar, weil sie Vagheiten zu entdecken erlauben und auf Ungenauigkeiten der maschinellen

4 Das Kernteam des Projekts besteht aus Sandra Murr, Florian Barth und mir.

Analyse hinweisen. Dabei ist bei literarischen Texten möglicherweise noch mehr als bei pragmatischen Kommunikationen in Alltagssprache davon auszugehen, dass das in der Computerlinguistik angestrebte *inter-annotator agreement* schwer zu erreichen ist. Diese Auffassung allerdings wäre durch Vergleichsstudien zu erhärten. Wie es speziell im Fall des *Werther* aussieht, steht ebenfalls zu fragen. Möglicherweise werden in der Annotation ähnliche Konflikte auftauchen, wie bei der Interpretation des Textes: Was etwa deutet tatsächlich auf die ‚Krankheit zum Tode' hin und inwiefern ist sie für den Suizid der Hauptfigur entscheidend?

Unser Ansatz bedient sich also einer Kombination der Analyseformen, und dies in mehreren Stufen. Die Interpretationskonzeption dieses *reading with the workflow* knüpft an etablierte Ergebnisse und heuristische Ziele der Literaturwissenschaft an, will sie aber strukturanalytisch verfeinern und zwar sowohl auf maschinellem wie auf manuellem Weg. Die methodische Anlage dieses Versuchs ist damit komplexer und aufwendiger als das, was die Literaturwissenschaft bislang dazu bot. Umgekehrt aber verspricht das Vorgehen, ein feinkörniges Analysemodell für Texte zu bieten, deren Merkmale durch einen gemeinsamen Bezugstext überlappen (können) und die dadurch eine Serie oder, anders gesagt, ein merkmalspezifisches Korpus bilden. Im Fall der Wertheriaden lässt sich ausgehend von diesem kombinierten Ansatz zeigen, wie groß die Abweichungen bei den jeweiligen Figurenkonstellationen und den sie auszeichnenden Semantiken sind und wie also die für den Werther und seine Nachfolgetexte spezifische Dreieckskonstellation aussieht. Dieses Zeigen ist allerdings nicht identisch mit einer Interpretation. Vielmehr deutet es – ähnlich und doch anders als die Funde über das Vorkommen des Ausdrucks ‚jch' – auf einen Aspekt hin, der für eine Interpretation des *Werthers* und der Wertheriaden bedeutsam ist und der zusammen mit anderen Befunden zu einer Interpretation dieser Textverhältnisse aggregiert werden könnte.

Hinsichtlich einer Empirisierung der Literaturwissenschaft nimmt das *reading with the workflow* Objekte und Forschungsfragen der Literaturwissenschaft auf, in diesem Fall einen kanonisierten literarischen Text und mitunter wenig bekannte Texte, die sich auf ihn beziehen und ein über Textmerkmale verbundenes Korpus bilden. Diese Objekte werden ihrerseits in ihre Bestandteile zerlegt, um sie selbst zu analysieren und die Hypothese von ihrer Verbundenheit als serielles Korpus zu testen. Der empirische Vorzug und damit auch das Ziel dieses Vorgehens besteht dabei vor allem in der Feingranularität des Vorgehens und der Ergebnisse. Diese ließen sich zwar potenziell auch auf ein großes Korpus anwenden, doch wäre dies zum gegenwärtigen Zeitpunkt jedenfalls hinsichtlich der Korpusbildung und Analyse sehr aufwendig.

Im *reading with the workflow* werden Untersuchungsgegenstand und Ziele der Analyse so justiert, dass sie ggf. in mehreren Stufen auf einen Befund (das Vor-

kommnis des ‚jch' im *Werther*) oder eine Fragestellung (Wie hängen *Werther* und die Wertheriaden zusammen?) antworten können. Ein *reading with the workflow* kann sich dabei im heuristischen Rahmen der Literaturwissenschaft bewegen, diesen aber methodisch und terminologisch erweitern und damit zugleich zur Entwicklung der empirischen Literaturwissenschaft beitragen. Der Beitrag liegt dabei weniger in der Beobachtung von Dingen, die man bislang gar nicht gesehen hat, die also von den Textbeobachtungen der analogen Literaturwissenschaft abweichen, sondern vor allem in der feingranularen Textempirie. Diese feine Granularität auch auf größere Korpora zu übertragen, ist oft noch Zukunftsmusik. Doch steht und fällt die Leistung des *reading with the workflow* damit nicht. Einem alten Diktum der Literaturwissenschaft folgend, zählt eben gerade die Analyse auch des singulären Textes, die in unserem Fall am Beispiel der Wertheriaden zugunsten eines kleinen Korpus überschritten wurde.

Eine andere und komplizierte Frage ist dabei, welche Rigidität hilfreich ist, um den feingranularen Ansatz weiterzutreiben und welche Granularität die Literaturwissenschaft zu schätzen bereit ist. Anders gesagt fragt sich, wieviel Textempirie das Fach verträgt und wieviel Empirie jenseits des Textes es aushält. Blickt man auf die wechselvolle Geschichte seiner Rationalisierungs- und Empirisierungsbestrebungen zurück, so reibt sich die Literaturwissenschaft immer wieder an diesen Fragen. Sie gehören, praxeologisch betrachtet, zu seinem methodischen und systematischen Reflexionsbestand, der immer wieder aktualisiert sein will, damit sich das Fach als Wissenschaft bewähren kann.

Primärliteratur

Von Goethe, Johann Wolfgang (1774). *Die Leiden des jungen Werthers*. 2 Bde. Leipzig: Weygand.

Sekundärliteratur

Ajouri, Philip, Katja Mellmann und Christoph Rauen (2013). „Einleitung". In: *Empirie in der Literaturwissenschaft*. Hrsg. von Philip Ajouri, Katja Mellmann und Christoph Rauen. Münster: Mentis, S. 9–17.

Algee-Hewitt, Mark (2019). „Day 1 Response: Criticism, Augmented". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses/ (besucht am 1. Juni 2020).

Barth, Florian, Evgeny Kim, Sandra Murr und Roman Klinger (2018). „A Reporting Tool for Relational Visualization and Analysis of Character Mentions in Literature". In: *Abstracts der*

DHd: Kritik der digitalen Vernunft. Köln: Digital Humanities im deutschsprachigen Raum e.V., S. 123–127.

Barth, Florian und Sandra Murr (2017). „Digital Analysis of the Literary Reception of J.W. von Goethe's Die Leiden des jungen Werthers". In: *Digital Humanities 2017: Conference Abstracts*. Montreal, S. 540–542.

Bernhart, Toni, Marcus Willand, Sandra Richter und Andrea Albrecht (2018). *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin, Boston: de Gruyter.

Bode, Katherine (2019a). „Day 1 Response". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-2/ (besucht am 1. Juni 2020).

Bode, Katherine (2019b). „Day 2 Response". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-3/ (besucht am 1. Juni 2020).

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris und Jannik Strötgen (2015). „Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative". In: *DHCommons* 1. DOI: 10.5281/zenodo.3240591.

Brooks Frederick P., Jr. (1977). „The Computer ‚Scientist' as Toolsmith. Studies in Interactive Computer Graphics". In: *Information Processing 77*. hier zit. n. Wiederabdruck TR 88-041, University of North Carolina, Chapel Hill 1988., S. 625–634.

Clayton, Aubrey (2020). „Die Replikationskrise". In: *Merkur. Deutsche Zeitschrift für europäisches Denken* 74.849, S. 79–87.

Da, Nan Z. (2019). „The Computational Case against the Computational Literary Studies". In: *Critical Inquiry* 45.3, S. 601–639.

Dörk, Marian, Peer Trilcke und et al. (2019). *Fontanes Handbibliothek*. URL: https://uclab.fh-potsdam.de/ff/ (besucht am 1. Apr. 2020).

Fish, Stanley (2019). „Afterword". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-5/ (besucht am 1. Juni 2020).

Fricke, Harald (1986). „Zur Rolle von Theorie und Erfahrung in der Literaturwissenschaft". In: *Colloquium Helveticum* 4, S. 5–21.

Gius, Evelyn und Janina Jacke (2017). „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis". In: *International Journal of Humanities and Arts Computing* 11.2, S. 233–254.

Gius, Evelyn und Marco Petris (2015). „Die explorative Visualisierung von Texten. Von den Herausforderungen der Darstellung geisteswissenschaftlicher Primär- und Annotationsdaten". In: *Abstracts der DHd: Von Daten zu Erkenntnissen*. Graz, S. 85–92.

Glanz, Berit und Nicole Seifert (2019). *Wenn es unterhaltsam wird, sind die Frauen dran*. Spiegel Online. URL: https://www.spiegel.de/kultur/literatur/vorschauenzaehlen-anteil-von-autorinnen-in-den-fruehjahrsprogrammen-a-1301975.html (besucht am 1. Juni 2020).

Groeben, Norbert (1977). *Rezeptionsforschung als empirische Literaturwissenschaft. Paradigmadurch Methodendiskussion*. Kronberg: Athenäum.

Herrmann, Berenike, Anne-Sophie Bories, Rebora Frontini Francesca, Simone und Jan Rybicki (2019). „Response by the Special Interest Group on Digital Literary Stylistics to Nan Z. Da's Study". In: *Journal of Cultural Analytics* 3.5.2019. DOI: 10.22148/001c.11827.

Hettinger, Lena, Martin Becker, Isabella Reger, Fotis Jannidis und Andreas Hotho (2015). „Genre Classification on German Novels". In: *Proceedings of the 12th International Workshop on*

Text-based Information Retrieval. Valencia, Spanien, S. 249–253. DOI: 10.1109/DEXA.2015.62.

Horré, Thomas (1997). *Werther-Roman und Werther-Figur in der deutschen Prosa des Wilhelminischen Zeitalters*. St. Ingbert: Röhrig.

IGEL Society (2018). *The Society's Mandate*. URL: https://sites.google.com/igelassoc.org/igel2018/home/the-igel-mandate (besucht am 1. Juni 2020).

Jannidis, Fotis (2003). „Polyvalenz – Konvention – Autonomie". In: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin/New York: De Gruyter, S. 305–328.

Jannidis, Fotis (2019). „On the perceived complexity of literature. A response to Nan Z. Da". In: *Journal of Cultural Analytics*. DOI: doi.10.22148/001c.11829 .

Jannidis, Fotis, Hubertus Kohle und Malte Rehbein, Hrsg. (2017). *Digital Humanities. Eine Einführung*. Stuttgart: Metzler.

Jessen, Caroline (2019). *Bücherspuren. Karl Wolfskehls deutsch-jüdische Bibliothek*. Bd. 2/13. Münchner Beiträge zur jüdischen Geschichte und Kultur. München: LMU München. Abteilung für jüdische Geschichte und Kultur.

John, Markus, Steffen Koch und Thomas Ertl (2017). „Uncertainty in Visual Text Analysis in the Context of the Digital Humanities". In: *Designing for Uncertainty in HCI: When does Uncertainty help? Workshop on CHI 2017*. Denver, Colorado: Association for Computing Machinery.

Klein, Lauren F. (2019). „Day 1 Response: What the New Computational Rigor Should Be". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-5/ (besucht am 1. Juni 2020).

Krämer, Sybille und Martin Huber (2018). „Dimensionen digitaler Geisteswissenschaften". In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/sb003_013.

Kuhn, Jonas (2019). „Computational text analysis within the Humanities: How to combine working practices from the contributing fields?" In: *Language Resources and Evaluation* 53.4, S. 565–602. DOI: 10.1007/s10579-019-09459-3.

Kuhn, Jonas (2020). „Computational Text Analysis within the Humanities". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 63–106.

Lauer, Gerhard (2019). „Über den Wert der exakten Geisteswissenschaften". In: *Geisteswissenschaften – was bleibt? Zwischen Theorie, Tradition und Transformation*. Hrsg. von Hans Joas und Jörg Noeller. Freiburg/München: Karl Alber, S. 152–173.

Leder, Helmut, Benno Belke, Andries Oeberst und Dorothee Augustin (2004). „A model of aesthetic appreciation and aesthetic judgements". In: *British Journal of Psychology* 95, S. 489–508.

Leder, Helmut und Marcos Nadal (2014). „Ten years of a model of aesthetic appreciation and aesthetic judgements. The aesthetic episode –- Developments and challenges in empirical aesthetics". In: *British journal of Psychology* 105, S. 443–464.

Manning, Christopher D. (2015). „Last Words: Computational Linguistics and Deep Learning". In: *Computational Linguistics* 41.4, S. 701–707. DOI: 10.1162/COLI_a_00239.

Martens, Lorna (1985). *The Diary Novel*. Cambridge: Cambridge University Press.

McGrath, Laura (2019). *Comping White*. Website. URL: https://lareviewofbooks.org/article/comping-white/ (besucht am 1. Juni 2020).

Moretti, Franco (2011). „Network Theory, Plot Analysis". In: *New Left Review* 68, S. 80–102.

Ort, Claus-Michael (2019). „Texttheorie – Textempirie – Textanalyse. Zum Verhältnis von Hermeneutik, empirischer Literaturwissenschaft und Literaturgeschichte". In: *Empirische Litera-*

turwissenschaft in der Diskussion. Hrsg. von Achim Bartsch, Gebhard Rusch und Reinhold Viehoff. Frankfurt M.: Suhrkamp, S. 104–122.

Parham, Marisa (2018). „Ninety-Nine Problems: Assessment, Inclusion, and Other Old-New Problems". In: *American Quarterly* 70.3.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Piper, Andrew (2019a). „Day 1 Response: The Select". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-7/ (besucht am 1. Juni 2020).

Piper, Andrew (2019b). „Do We Know What We Are Doing?" In: *Journal of Cultural Analytics* 1.4.2019. DOI: doi.10.22148/001c.11826 .

Piper, Andrew und Mark Algee-Hewitt (2014). „The Werther-Effect I: Goethe Topologically". In: *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Hrsg. von Matt Erlin und Lynn Tatlock. Rochester: Camden House, S. 155–184.

Ramsay, Stephen (2011). *Reading machines. Toward an algorithmic criticsm*. Urbana, Chicago und Springfield: University of Illinois Press.

Rat für Informationsinfrastrukturen (2019). *Herausforderung Datenqualität –- Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel*. 2. Aufl. URN:nbn:de:101:1-201911201154165773273 7. Göttingen: Rat für Informationsinfrastrukturen.

Referat Forschung Deutsches Literaturarchiv Marbach (2019). URL: https://www.dla-marbach.de/forschung/kooperationen/netzwerk-literarische-erfahrung/ (besucht am 3. Apr. 2020).

Richter, Sandra (2010). *A History of Poetics. German Scholarly Aesthetics and Poetics in International Context, 1770–1960. With Bibliographies by Anja Zenk, Jasmin Azazmah, Eva Jost, Sandra Richter*. Berlin, New York: De Gruyter.

Richter, Sandra (2017a). *Eine Weltgeschichte der deutschsprachigen Literatur*. München: C. Bertelsmann.

Richter, Sandra (2017b). *German Literature Global*. Team ‚German Literature in the World': Jasmin Azazmah, Florian Barth, Steffen Burk, Dilan Cakir, Falk Erdmann, Philipp Heiter, Martin Kuhn, Sandra Murr, Merisa Taranis. URL: http://www.germanliteratureglobal.com (besucht am 3. Apr. 2020).

Schmidt, Siegfried J. (1991). *Grundriss der Empirischen Literaturwissenschaft*. Frankfurt am Main: Suhrkamp.

Schneider, Ulrich Johannes (2020). „Deutsche Nationalkataloge – Herausforderungen an das deutsche Bibliothekssystem. Was aus der Perspektive der Digital Humanities zu tun wäre". In: *ABI Technik* 40.1, S. 40–51. DOI: 10.1515/abitech-2020-1005.

Schruhl, Friederike (2014). „Objektumgangsnormen in der Literaturwissenschaft". In: *Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden*. Hrsg. von Martin Huber und Sybille Krämer. Rochester: Camden House, S. 155–184. DOI: 10:17175/sb003_012.

Trilcke, Peer (2013). „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In: *Empirie in der Literaturwissenschaft*. Hrsg. von Philip Ajourie, Katja Mellmann und Christoph Rauen. Münster: Mentis, S. 201–247.

Underwood, Ted (2019a). „Day 1 Response". Online Forum. Critical Inquiry. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-8/ (besucht am 1. Juni 2020).

Underwood, Ted (2019b). „Dear Humanists: Fear Not the Digital Revolution. Advances in computing will benefit traditional scholarship – not compete with". In: *The Chronicle of Higher Education* 65.29. URL: https://www.chronicle.com/article/Dear-Humanists-Fear-Not-the/245987 (besucht am 1. Juni 2020).

Underwood, Ted (2019c). *Distant Horizons. Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Uprichard, Emma (2014). „Big-Data Doubts". In: *Chronicle of Higher Education* 13. URL: https://www.chronicle.com/article/Big-Doubts-About-Big-Data-/149267 (besucht am 1. Juni 2020).

Weitin, Herget (2017). „Falkentopics. Über einige Probleme beim Topic Modeling literarischer Texte". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, S. 29–48. DOI: 10.1007/s41244-017-0049-3.

Willand, Marcus (2017). „Hermeneutische Interpretation und digitale Analyse. Eine Verhältnisbestimmung". In: *Lektüren. Positionen zeitgenössischer Philologie*. Hrsg. von Luisa Banki und Michael Scheffel. Trier: Wissenschaftlicher Verlag Trier, S. 77–100.

Yong, Ed (2018). „Psychology's Replication Crisis Is Running Out of Excuses". In: *The Atlantic* 19.11.2018.

Zehe, Albin, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger und Fotis Jannidis (2016). „Prediction of Happy Endings in German Novels based on Sentiment Information". In: *Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda*. Riva del Garda, Italien, S. 9–16.

Cathleen Kantner and Maximilian Overbeck

Exploring Soft Concepts with Hard Corpus-Analytic Methods

Abstract: Corpus-analytic studies are currently experiencing a veritable boom in the social sciences. However, three crucial barriers impede the methodological quality and long-term reputation of these promising new technologies: Firstly, creating and pre-processing very large text corpora is still a laborious and costly enterprise. Secondly, the semantically valid operationalization of complex theoretical concepts remains a problem. Thirdly, scholars need flexible data output and visualization options to connect the data generated by computational methods with the discipline's existing research. We will conclude that it is possible to solve these problems. However, hermeneutically sensitive uses of computational methods will take much more time, work, and creativity than often assumed. The reflected appropriation of big data methods in the social sciences has only just begun.

Zusammenfassung: In diesem Beitrag argumentieren wir, dass drei methodische Barrieren die Verbreitung korpusanalytischer Methoden in den Sozialwissenschaften erschweren. Erstens bereitet es immer noch einen sehr hohen Aufwand, große Textkorpora zu erstellen und aufzubereiten. Zweitens ist das Problem der semantisch validen Operationalisierung komplexer geistes-, sozial- und kulturwissenschaftlicher Begriffe noch völlig unzureichend gelöst. WissenschaftlerInnen sind daran interessiert, über die Analyse manifester Textinhalte komplexe gesellschaftliche Sinnzusammenhänge zu rekonstruieren. Drittens erlauben viele der für linguistische Fragestellungen designten Tools kaum eine sozialwissenschaftlich anschlussfähige Ergebnisdarstellung. Die Effizienz der neuen Methoden bleibt hinter den hohen Erwartungen zurück, wenn es nicht gelingt, diese mit jedem neuen Forschungsprojekt verbundenen zeitraubenden Arbeitsschritte zu standardisieren. Allerdings kann es hierfür keine one size fits all Lösungen geben,

Note: An earlier version of this contribution in German language can be found in: Cathleen Kantner and Maximilian Overbeck (2018). "Die Analyse 'weicher' Konzepte mit 'harten' korpuslinguistischen Methoden". In: *Computational Social Science: Die Analyse von Big Data*. Ed. by Andreas Blätte, Joachim Behnke, Kai-Uwe Schnapp, and Claudius Wagemann. Baden-Baden: Nomos Verlag, pp. 163–189.

Cathleen Kantner, Department of International Relations and European Integration, University of Stuttgart

Maximilian Overbeck, Department of Communication and Journalism, The Hebrew University of Jerusalem

Open Access. © 2020 Cathleen Kantner und Maximilian Overbeck, published by De Gruyter BY-NC-ND
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license.
https://doi.org/10.1515/9783110693973-008

weil aus der Perspektive unterschiedlicher wissenschaftlicher Forschungsfragen unterschiedliche methodische Entscheidungen zu treffen bleiben.

1 Introduction

The increasing accessibility of textual data, as well as numerous innovations in computational text analysis tools, are currently creating new opportunities for big data research in the social sciences.[1] Yet the great potential of corpus-analytic[2] methods and tools remains largely unused until now, although a growing number of researchers are interested in large or even big data[3] content analyses.

In this paper, we discuss the opportunities and challenges of conducting computerized corpus-analytic research guided by complex theoretical concepts. We will call them "soft concepts" because they are structurally contested in the social scientific research community. We will put special emphasis on how to use the new, seemingly easy to use corpus-analytic methods without violating the exigent quality standards of empirical research in the social sciences. In 'classical', 'qualitative', or manual textual analysis, there are well-established best practices that comply with the methodological standards of reliability, validity, and representativeness.[4] These standards have been established in long foundational debates and the methodological development of the discipline. For the analysis of

1 In this paper, we discuss the methodological challenges from a social sciences perspective. However, we assume that our argument is equally valid for the humanities and other research disciplines that analyze abstract theoretical concepts within larger text corpora.

2 We will use the terms "corpus-linguistic" and "corpus-analytic" instead of "computational-linguistic" since computers support the data processing, yet these methods relate to linguistic properties of human language, expressed in a *corpus* of text. We define a corpus as "a collection of spoken and written expressions. The corpus material is digitized, which means it is saved on computers and can be read by machines. The corpus elements consist of the material itself as well as metadata, describing the text material, and of linguistic annotations which were applied to the textual material" (Lemnitzer and Zinsmeister 2006, p. 8, *author translation*).

3 There does not exist a strict numerical threshold in text analysis for when to speak of small or large corpora. We will refer to a corpus as small if the documents can be manually annotated with reasonable efforts and if the researcher has an overview over the full corpus by close reading. In contrast, we refer to a corpus as large if it cannot longer be manually explored and annotated, and it therefore requires some kind of computer assisted support in order to gain an understanding of the corpus. Large corpora deal with masses of text or even big data.

4 *Reliability* deals with the question of whether data generation remains stable over time and can be precisely reproduced by different researchers. (Semantic) *validity* enquires whether the collected data actually measures the theoretically derived concepts (Krippendorff 2018, pp. 277, 361). *Representativeness*, or "sampling validity" (Krippendorff 2018, pp. 368–373) verifies whether

large amounts of textual data, however, similar 'hard methods' have yet to be developed.

In our research project, *eIdentity*, we created and analyzed a multilingual corpus of 460 917 newspaper articles dealing with wars and humanitarian military interventions since the end of the Cold War. We specifically focused on multiple collective identities as well as discussions on religious identities within these public debates. The analyzed corpus includes a conservative and a liberal newspaper each from six countries (Germany, Austria, France, United Kingdom, Ireland, and USA) extending over the entire period from January 1990 to December 2012. The corpus includes controversial debates about well-known conflicts like those in the former Yugoslavia, Iraq, or Libya, but also less intensively discussed conflicts, such as armed conflicts in Kashmir, Sierra Leone, or Kurdistan. For our analyses, we developed generic tools and best practices for future empirical research: the *Exploration Workbench*, designed to create and preprocess large text corpora, and the *Complex Concept Builder*, which supports researchers in their analysis and operationalization of theoretical concepts.[5]

We will argue that three crucial methodological obstacles impede the dissemination of corpus-analytic methods in the social sciences. Firstly, it still takes a lot of effort to create, manage, and preprocess large text corpora before the actual analysis can even begin. The performance of the new methods will lag behind high-flying expectations until we succeed in standardizing those time-consuming and recurring processes. Having said that, there can be no *'one size fits all'* solution for corpora creation and preprocessing because scholars must adjust these processes to their respective research questions.

Secondly, adequate attention has not yet been paid to the semantically valid operationalization of theoretical concepts. Social scientists are not *per se* interested in the lexical or grammatical patterns of a text or a corpus of texts. Instead, they want to learn about society from manifest text content (Krippendorff 2018, pp. 29–40). Moreover, they are conducting their research usually from a very specific theoretical point of view within a scientific community characterized by theoretical pluralism. Therefore, social scientists search for expressions by 'real world'

a selected set of textual data represents the whole spectrum of variability inside a given (textual) population (Biber 1993, p. 243).

5 For further information, see: https://www.sowi.uni-stuttgart.de/abteilungen/ib/forschung/eldentity/ (last accessed on April 16th 2020). Follow-up projects such as PolMine and CRETA already reuse parts of our tools. Moreover, in the medium term, some of the tools will be accessible in the CLARIN-D infrastructure.

speakers in everyday language which can with good reasons count as indicators for the occurrence of *abstract theoretical* concepts (e. g. 'national identity').[6]

Thirdly, many of the tools designed for linguistic research are barely suitable for specific social scientific needs. Flexible data output and visualization options are required to relate corpus-linguistic data to existing research in the discipline. This is a central prerequisite for making meaningful contributions to ongoing debates and gaining acceptance for the new methods.

In this paper, we will proceed as follows: We will start by presenting three generations of text-analytic research. Every new generation improves upon the preceding one through more sophisticated research techniques. Then we focus on the youngest generation of textual analysis, as it most profoundly deals with the challenging task of combining innovative computational tools with interpretative techniques in a reflected way. This task is crucial because we could seriously harm the methodological quality and the long-term reputation of these methods in our disciplines if we do not manage to find viable solutions for *the three fundamental problems identified above*.

To date, there are a considerable number of scholars who are motivated to employ computational, corpus-analytic methods for their own research questions but who often surrender because of three seemingly insurmountable barriers. Social scientists need tools that address the three barriers and find solutions to them. We conclude that it is possible to find solutions for these challenges but, even then, using corpus-analytic methods in the social sciences will take much more time, work, and creativity than is widely expected.

2 Three Generations of Computer-Assisted Text Analysis

In comparison to other fields, social scientists are generally very open-minded about the use of new computational methods for their own research questions. This is certainly because text analysis already has a long-standing tradition in the social sciences. The increasing availability of digitized text archives and innovative computational tools has important repercussions for social science research. Simultaneously, computational methods have spread and developed in recent

6 One would for instance rarely express 'national identity' directly like "because of my Spanish identity, I..." but rather in expressions such as "Spain should..." or "our country faces the challenge of..."

Tab. 1: Three generations of computer-assisted text analysis

Generation	Corpus size	Methodology	Strengths	Problems
1st generation *A Good Piece of Craftsmanship*	Small	Manual annotation of small corpora with qualitative methods (e.g. content analysis, discourse analysis)	High validity & reliability	Low representativeness of the analyzed data (small corpora, opaque sampling) Low (inter-coder-) reliability, if interpretations are not made inter-subjectively transparent
2nd generation *The Lawn Mower Approach*	Large / big data	'Counting words' within large corpora instructed by quantitative mostly corpus linguistic methods (e. g. dictionaries, key terms, collocates, etc.)	High reliability	Low semantic validity of research instruments
3rd generation *The Swiss Army Knife Approach*	Large / big data	Combination of qualitative-hermeneutic and quantitative text analyses on large corpora via: – Corpus linguistics / text mining & discourse analysis (inductive) – Semantic field analysis (deductive) – Semi-automatic classification and text mining via machine learning algorithms	Full samples or large representative corpora Analyzing complex concepts on large or big data corpora validly, reliably and transparently	Reliability sometimes uncertain, in case of untransparent usage of complex algorithms (*black box*)

years. Overall, this has led to a steady improvement in the research instruments. In the following section, we distinguish between three generations of computer-assisted text analysis and we present their methodological strengths and weaknesses. We will assess how each generation deals with the trade-offs between the quality criteria of validity, reliability, and representativeness (see Table 1 for an overview).

2.1 The First Generation: A Good Piece of Craftsmanship

Max Weber, already in 1910, suggested "measuring with scissors and compass, how the content of newspapers has shifted [...] from what is now published as news compared to what was previously published as news" (Weber 1910, *author translation*; cf. Rössler 2010, p. 13). It was impossible, though, to carry out systematic analyses on a larger scale. Things changed in the early 1990s when the first software packages such as *ATLAS.ti* or *MAXQDA* appeared on the market. User-friendly interfaces for qualitative text analysis and manual annotation enabled scholars to use their limited resources (i. e. money, time, and annotators) more efficiently.

The first generation had direct contact with text. Manual work and attention to detail and quality characterized this approach – almost as in traditional craftsmanship. Scholars manually annotated small text corpora using qualitative content or discourse analysis methods. If annotation guidelines (codebooks) were well elaborated and pretested, coders well trained, and intercoder reliability checks performed continuously, corpus-analytic data could attain high quality in terms of validity and reliability. Many political science scholars used to apply first-generation methodological approaches within their research. Their usage can be illustrated within the particularly vivid debate on transnational public spheres (see e. g. Pfetsch and Heft (2015) for an overview).[7]

The first generation, however, faced the severe issue of sampling bias in order to get a somehow manageable amount of text. Ideally, researchers should draw a text-*sample* from the relevant corpus which represents a "down-scaled but

7 Further examples of first-generational research practice can be found in the publications of the Europub project (http://europub.wzb.eu) and the SFB 597 project "The Transnationalization of Public Spheres in the EU: Citizens' (re)actions" (http://www.sfb597.uni-bremen.de/pages/forProjektBeschreibung.php?SPRACHE=en&ID=9). Moreover, the study by Rucht et al. (1992) used first-generational approaches within their exceptionally large-scale "Prodat" project (https://www.wzb.eu/en/research/completed-research-programs/civil-society-and-political-mobilization/projects/prodat-dokumentation-und-analyse-von-protestereignissen-in-der-bundesrepublik, last accessed on April 16th 2020).

structurally equivalent image of the basic population" (Rössler 2010, p. 58, *author translation*).[8] Studies, however, often selected samples which were not representative of the whole spectrum of variability inside a given population. Many researchers did not even specify their selection criteria for what they called the 'relevant' texts. Therefore, if the sample is not representative of the investigated social phenomenon, the conclusions of the study cannot make valid claims about society.

As soon as more texts were analyzed within a study and more than one coder was annotating the texts, another problem occurred: intercoder reliability. In most projects, however, intercoder reliability was checked – if at all – superficially and only during the preparation of the coding phase. Continuous intercoder reliability tests were and are still an exemption. This limits, of course, the validity of any results if one cannot tell whether differences found are indeed in the data or rather due to different coders' readings.

2.2 The Second Generation: The Lawn Mower Approach

Having learned from the first generation that sampling corpora for qualitative analysis carries a risk of sampling bias, second generation scholars took up the challenge of analyzing larger corpora via text statistical methods (e. g. simple word counts, key terms analyses or collocates). For their quantitative analyses, scholars often used dictionaries.[9] Dictionaries are lists of words or combinations of words that represent an abstract concept in a text corpus and can be subsequently counted by the computer (Neuendorf and Skalski 2009, p. 207). Such term lists could be either self-generated or based on externally available sources.

The second generation, while having employed perfectly replicable methods to categorize linguistic features along social-scientific concepts, still encountered limits. Keyword lists and dictionaries have rarely been cleaned to account for word ambiguities and misleading semantic connotations.[10] It was tempting to use second-generation computational tools and run them – like a lawn mower

8 Analysts used different sampling techniques (e. g. relevance based, full random, or snowball sampling). Sample sizes in the first generation usually vary from several hundred to a few thousands of newspaper articles.

9 Moreover, dimensional scaling methods such as 'Wordscores' (Laver et al. 2003; Lowe 2008) or 'Wordfish' (Slapin and Proksch 2008) have become popular methods within quantitative text analysis.

10 An exception is the study from Klüver (2009) who compared the validity of the quantitative 'Wordfish' and 'Wordscores' techniques with manual annotation on a small control sample.

– over vast amounts of digitized text data at the expense of quality. The quickly generated quantitative results were often of low semantic validity.[11]

2.3 The Third Generation: The Swiss Army Knife

Grounded in a long-standing tradition of research experience, third generation studies are taking up the challenge of combining the assets of qualitative and quantitative text analysis. Past research has shown that they cannot build valid indicators for 'soft concepts' without interpretive procedures and manual annotation. Therefore, they opt for the systematic combination of quantitative and qualitative steps of analysis.

Third-generational text analysts combine, for example, quantitative corpus linguistic approaches with qualitative discourse analysis (e. g. Baker and McEnery 2005; Gabrielatos and Baker 2006; Kutter 2007; Bayley and Williams 2012). Compared to content-analytical studies, these analyses generally tend to have an inductive-exploratory character and do not attempt to operationalize complex theoretical concepts. Their aim is to identify typical linguistic patterns and to deconstruct them via in-depth discourse analysis. In comparison to such exploratory approaches, *semantic field analysis* (Kutter and Kantner 2012) is a theory driven approach. Quantitative analyses are systematically combined with qualitative steps of validation and disambiguation in order to identify relevant terms and word combinations representing complex theoretical concepts within large text corpora (for a more detailed description, see also 3.2).[12] We have now the chance to create platforms or workflows of tools that provide – like a Swiss-army-knife – various options for qualitative analysis (manual coding of codebook guided hermeneutic interpretation) as well as for quantitative, semi-automatic, or automatic categorization of text.

11 For a critical stance in this regard, see Krippendorff (2018, pp. 215–276, 361–382) and Roberts (1989).

12 Kutter and Kantner (2012) developed semantic field analysis for the examination of international discussions on wars and humanitarian military interventions within a carefully cleaned newspaper corpus of about 500 000 newspaper texts. They created semantic fields for a semi-automatic identification of discussions on international organization, or humanitarian military interventions. For further applications of *semantic field analysis* in media analysis, see Kantner, Kutter, and Renfordt (2008), Kantner (2011a,b, 2014), Overbeck (2014), and Kantner (2016).

The recent rapid computational developments have the potential to advance hermeneutic text analyses significantly. Supervised or unsupervised[13] machine learning techniques and other algorithmic methods can be used and combined in many different ways (for a helpful overview, see Wiedemann 2016, pp. 49–54). They are currently used for example in text-clustering (Hall et al. 2008; Janasik et al. 2009; Grimmer 2010; Evans 2014), or in (semi-) automatic classification methods operating on the level of entire documents or specific text segments (see e. g. Stalpouskaya and Baden 2015).

The growing amount and increasing availability of digitized text on electronic text archives offers new opportunities for social scientists. Commercial software for text analysis is continuously improving and becoming more user-friendly and affordable (Alexa and Zuell 2000; Krippendorff 2018). Moreover, there are countless open-source applications, often created in the context of research projects or doctoral theses, or freely combinable computer-linguistic modules for self-programming in R or Python. National ministries for research and other research funding institutions are currently financing the establishment of comprehensive national and international platforms (e. g. DSPIN, CLARIN, ERIC, and DARIAH) as future research infrastructures, making cutting-edge corpus-analytic solutions permanently available for all scientific users.

Yet, why do scholars still make relatively little use of corpus analytic methods of the third generation? Why do first-generation research designs with small corpora or second generation 'lawn mower' approaches still dominate our discipline? Is it due to a deficiency in user friendliness, or to poor promotion?

3 Before Take-Off: Three Barriers and Prospects for Solutions

Social scientists today cannot yet exploit the full potential of third generation methodology due to three fundamental barriers:

1. Researchers tend to underestimate the considerable efforts that are necessary to create text corpora and preprocess large amounts of text. The more heterogeneous the texts and their sources (i. e., because multiple languages and

13 Whereas *supervised* classifications rely on training data, e. g. manually classified text segments, *unsupervised* text classifications process textual corpora automatically without training data (e. g. Pollak et al. 2011, p. 659).

sources from different countries are compared), the more problems will occur. Moreover, there are no common software applications to support these tasks.
2. Social scientists want to learn about society, not language *per se*. Each researcher is embedded in a scientific debate in which different theoretical paradigms compete. Each researcher will certainly try to define her concepts with 'hard' arguments; however, those concepts will always be contested in the scientific community and insofar they will remain 'soft'. Moreover, these abstract concepts are rarely expressed in direct terms in popular language. We cannot capture these 'soft' concepts by mere quantitative methods such as simple word counts or dictionary based tools.
3. Many corpus-analytic applications and tools are limited in terms of their data output and visualization options. In order to connect corpus-analytic data to the discipline's existing research (e. g. opinion polls or official statistics), tools must consider the specific needs for subsequent statistical analyses.

3.1 The First Barrier: Corpus Creation and Processing

The easy access to large text corpora often turns out to be a trap. The preprocessing of large corpora poses various problems before the actual analysis can even begin (Krippendorff 2018, p. 115). Yet even the best maintained electronic text archives offer only limited search and retrieval options. Keyword strategies that are based on selected search terms and simple Boolean operators inevitably contain semantic ambiguities. The common result is an abundance of *sampling errors* within the collected corpora.

Sampling errors are documents within the 'raw corpus'[14] which do not deal with the topic or the primary text genre of interest (Kantner 2016, pp. 65–67). To give an example from our project, news articles about completely off-topic issues such as sports events were erroneously included in our corpus because of their metaphorical usage of 'war'-terms (e. g. "civil war like conditions reigned in the soccer stadium when the Bosnian team scored in the 56th minute").[15]

A second source of sampling errors or "white noise" (Gabrielatos 2007, p. 6) in the raw sample is the occurrence of *duplicates*. Duplicates are fully or partly similar or identical documents that appear twice or several times in the raw corpus due to a) badly maintained archives, b) different versions of documents

14 A 'raw corpus' is the entirely unprocessed amount of text delivered by a certain keyword query in an electronic text archive.

15 These and other preprocessing challenges are described more in detail in Kantner, Kutter, Hildebrandt, et al. 2011 and Kantner 2016.

in the database that are not relevant to the research question, or c) complex Boolean keyword searches which have to be split into parts for searches in professional data archives like Nexis or Factiva (Kantner 2016, pp. 63–65).[16] To keep these documents within a corpus would impair the empirical results' validity. No publicly available software is currently available to remove these documents. Conventional tools just assume that scholars possess a perfectly prepared and preprocessed corpus from the outset. Most researchers therefore just work with the 'raw corpus' or they write their own scripts for corpora storage, management, and preprocessing within individual databases. This limits the potential range of big data analysts to researchers who have either programming skills or are collaborating with computer scientists. The vast majority of social scientists would however welcome easily usable tools and shared best practices for corpus creation and preprocessing.

In our research, we developed innovative solutions for these recurring problems. For the removal of sampling errors, we developed an innovative procedure based on topic modeling and supervised text classification.[17] There are various approaches to automatic corpora classification based on lexical features. In recent years, the so-called Latent Dirichlet Allocation method (LDA, Blei et al. 2003) has become one of the most popular usages for automatic document classification. The approach is based on the idea that each document contains a certain number of latent variables ('topics') that are themselves represented by co-occurring structures of words and phrases. Apart from determining some formal parameters, the detection of these latent variables is an entirely automatic procedure within LDA. Clustering an entire corpus into small sub collections has decisive advantages. It enables the researcher to become aware of systematic features and peculiarities within a large amount of textual data to an extent reaching beyond purely keyword-based procedures. However, the main difficulty within this approach is that the created document clusters do not represent 'topics' in the public debate's sense of the term.[18] In our case, collections of names such as "*Merkel* or 'Blair'", rubrics such as "*reviews*", as well as terms relating to places or events were among the topics.

16 Some online archives provide automatic duplicate-removal options; however, these options lack transparency since they do not define what counts as a duplicate (with respect to a specific research question) and how they automatically identify and remove them.

17 For more detailed information on our sampling cleaning process, see Blessing, Sonntag, et al. (2013) and Blessing, Kliche, et al. (2015b).

18 According to Wessler (1998, p. 666) topics or 'issues' are controversial objects of political communication that must be distinguishable from other issues, and on which the public forms opinions in the course of a debate.

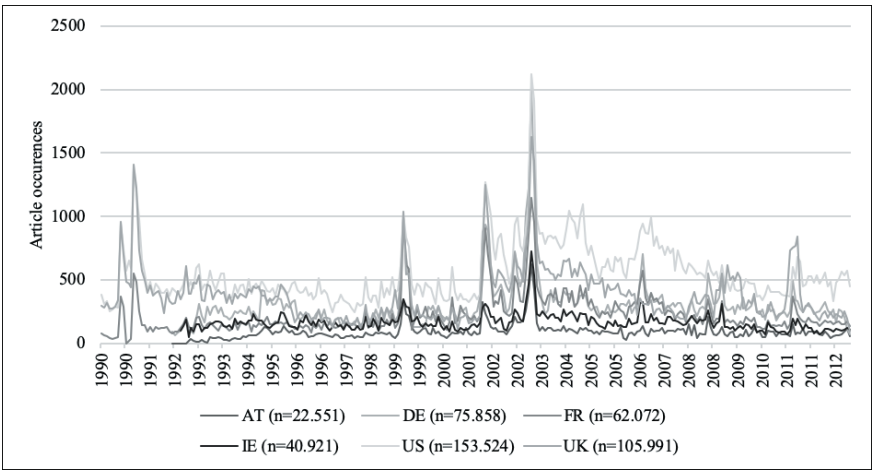


Fig. 1: The e-Identity corpus cleaned issue cycle (article occurrences per month in absolute numbers, 1990–2012, N=460 917)

We first removed duplicates from the raw corpus within our *Exploration Workbench*.[19] We then used LDA topic models to cluster each language-specific corpus (English, French, and German) into 200 sub collections. In a second step, trained coders manually classified[20] a large number of articles within each of the 200 topics as 'relevant' or 'irrelevant' according to our issue of interest ('wars' and 'humanitarian military interventions').[21] We used these manual annotations as train-

19 The e-Identity *Exploration Workbench* has been developed for creating, managing and preprocessing large text corpora (see Kliche et al. 2014; Blessing, Kliche, et al. 2015a). It allows importation of texts of different document types, from a wide variety of sources, with diverse text and character formats, and in different languages. By selecting the relevant options from a menu, the *tool* also offers flexibility according to one's research question. The workbench features selectable options for the duplicate removal according to individual research questions. For our project, for example, we considered identical texts as non-duplicates if they were published on different days or in different media. In order to measure the similarity of two newspaper articles, we used the methods of 'fingerprinting' and 'shingling' by Manning et al. (2008) (for more information, see Kliche et al. (2014, pp. 694–696).

20 The quality, reliability, and transparency of the annotation process were ensured through a codebook with detailed annotation instructions and examples. We moreover held weekly coder trainings, and continuously verified intercoder reliability and respectively inter-annotator agreement.

21 For each topic, we selected articles from 20 different topic ranks (e. g. 1st, 2nd, 3rd order topic), and the whole range of topic related percentages (e. g. the whole article dealt to 70 % with a 1st topic, to 30 % with a 2nd topic etc.). With this stratified selection strategy, we expected a higher

ing data within a subsequent machine learning approach that operated on the compositions of topics within 'relevant' and 'irrelevant' articles (for more details, see Blessing, Kliche, et al. (2015a), Blessing, Sonntag, et al. (2013), and Dick et al. (2015). The quality of the semi-automatically labelled articles was then further manually validated. The final cleaned and fully preprocessed corpus of 460 917 articles (see Figure 1) was half as large as the original raw corpus (901 856 articles).

3.2 The Second Barrier: Operationalization of Complex Theoretical Concepts

Social scientists are mostly not interested in "hard facts" that could be measured by an objective uninvolved observer.[22] Especially the most interesting and most intensively debated subjects of our research are, in the words of one of the French founding fathers of sociology, "*social facts*" (Durkheim 1984 [1895]). Concepts such as 'identity', 'social values', 'the state', 'love', 'social justice', 'beauty', 'art' and so forth do not exist outside of the practice of communication and interaction in which people think about these abstract objects, argue about them, so to speak 'socially construct' them, and (re-)interpret them collectively from generation to generation in discourse. In the medium of such discourses, cultures, societies, and communities reproduce themselves (Tietz 2002).

By analyzing text and text corpora, social scientists seek answers to questions about society that are relevant against the background of their scientific theories. They do not simply want to describe *how* something is communicated linguistically, but rather they observe 'real-world' communication processes from specific theoretical viewpoints with complex theoretical concepts in mind. Those complex concepts are usually *not* expressed directly in everyday language by non-scientists.

Since Thomas S. Kuhn (Kuhn 1962, 1970) and the post-analytical theory of science, we know that sciences do not approach truth asymptotically, but that we always engage in critique and counter-critiques of concepts and conceptual contents. This leads to a notorious pluralism of the concepts within the humanities,

---

chance to cover all 'mixture' combinations of topics as compared to a full random sampling. Rules for all coder decisions were defined in a codebook in advance.

22 Such 'hard facts' exist of course. They are generated via processes like demographic or socio-economic data collection and statistical analysis. However, in many cases social scientists are interested in associating 'hard' data (e. g., distribution of education, income, unemployment, etc.) or political preferences (e. g., survey results, election results) with their communicative context in order to understand why people act as they do.

social sciences and cultural studies, and in empirical terms, it leads to a pluralism of ways to operationalize these concepts: Terms that according to one theory might count as semantically valid indicators for, as an example, 'national identity' might be irrelevant or they might measure something different according to another theory.

Social scientists are usually conducting their research from a very specific theoretical point of view, or theoretical paradigms, within a scientific community characterized by theoretical pluralism. Taking a firm and innovative theoretical standpoint within these debates is one of the criteria for scientific originality in our field. This notorious pluralism needs to be translated into options for various operationalizations: a researcher studying collective 'identity' in a Habermasian perspective looks for different linguistic expressions than one who follows a Foucauldian perspective.

While linguists, computational linguists, or commercial operators tend towards identifying universal linguistic features within text, scholars in the humanities, social sciences, or cultural studies are instead interested in reconstructing complex social meaning through the analysis of manifest textual content (Krippendorff 2018).[23] They do not look for simple terms or term lists. The geographic concept of 'Europe' (as needed, e. g., for the purposes of navigation software for cars) might be operationalized by a list of words such as Paris, Rome, Helsinki, Rhine, Via Repubblica, Karpaty, Vltava, etc. Yet, for complex expressions of abstract concepts like 'political Europe' that are rarely expressed *directly* in everyday language we need to choose more hermeneutically sensitive approaches. Typical representations for the phenomena of interest for social scientists often are paraphrases (e. g., "our historical responsibility as Europeans", "Brussels has to", "Europe should", "therefore we are obliged to").

Moreover, one term might relate to the specific concept of interest only in a very specific context, whilst in other contexts it might not reach beyond its literal meaning. 'National identity' thus might be differently expressed in security policy, environmental policy, or a cultural discourse. From a social science perspective, it is therefore not practical to use standardized word lists or other prefabricated indicators.

In *eIdentity*, we addressed this challenge and developed three workflows for identifying and exploring expressions of complex concepts in large corpora (see Figure 2). In the first workflow, we explore corpora by simple term or collocation

23 Given these specific needs for analysis, the attempt by commercial software to standardize and integrate social science concepts within fixed lexica makes them intransparent and unusable (e. g. in the software *SPSS Clementine* that is very useful for many other purposes).

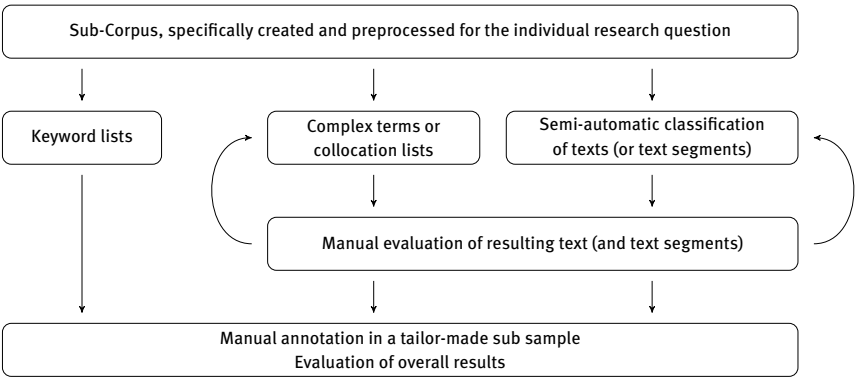


Fig. 2: Three different analysis workflows for sub-corpora creation

lists. This approach, although sharing some of the weaknesses of the second generation of text analysis, proves useful for an initial overview of the corpus. Simple and more complex search queries based, for example, on quantity or distance operators, can be linked to the selected search terms. Each search query can be manually checked and validated within its context in the full sentence. The results can be exported and sorted by metadata (e. g., by newspaper or data), and visually examined through processes like diachronic visualizations.

In the second workflow, we apply the methodological approach of *Semantic Field Analysis* (Kutter and Kantner 2012; Kantner 2016, pp. 67–70, 71–75) to create comprehensive lists of terms and collocations referring to a complex concept. This methodology follows a four-step approach: In the first step of concept *specification* we identify key terms and associated expressions via external sources such as encyclopedias and thesauri. In the second step of concept *identification* we trace, within the corpus of interest, lexical proxies of key terms and lexical variants. We moreover translate relevant terms from one language to another to identify additional potential terms until we reach a satisfying level of exhaustion. In the third step of concept *assignment* we apply the list to our corpus. In the fourth step of *quantitative analysis* we evaluate and statistically quantify our results.[24] The iden-

24 Our *Complex Concept Builder* provides a tool for the creation and management of even sophisticated term lists by combining various query operators with simple or multiple terms and collocates (e. g. by connecting quantity- or distance operators with selected terms) (for more details, see Blessing, Kliche, et al. (2015a), Blessing, Sonntag, et al. (2013). The tool moreover offers the possibility to examine every single query concerning its semantic validity in a sentence context (via KWIC-lists or whole sentences). In the data export, the results can be sorted by metadata (e. g. name of newspaper or date) and visualized in time series analyses.

tification and disambiguation of semantically valid indicators usually turns out to be an extremely difficult and time-consuming enterprise. Sometimes even the most sophisticated term-lists will not suffice to catch the exact meaning of a concept of interest – some terms are inherently ambiguous.

To deal with such semantically ambiguous terms, we therefore used, as a third workflow, the previously presented semi-automatic approach for removing sampling errors (see Section 3.1). This approach has the advantage of considering a large amount of semantic context. It proved suitable to create tailored subsamples of articles dealing explicitly with, for example, the role of religious identities in armed conflicts (Overbeck 2015).

The automatic clustering procedure of the overall corpus through LDA topics supports the difficult and ambitious task to identify the relevant semantic fields as indicators for the complex theoretical concepts.[25] The manual validation and classification of 'relevant' and 'irrelevant' articles subsequently helps to sort out articles that might contain religious keywords but do not explicitly address the role of religious identities in armed conflicts. The supervised machine learning procedure eventually helps to identify similar 'relevant' articles in the overall corpus of yet un-annotated articles.

In applying these three workflows, highly specific sub-corpora containing *only those texts* of the cleaned corpus *speaking to our theoretical research question* could be created and further be examined via qualitative content analysis, as practiced by the first generation.[26]

3.3 The Third Barrier: Enabling Suitable Result Visualization for Social Scientists

Quantitative text analyses – in one way or another – result in data. Once the textual analysis is completed, social scientists require flexible data arrangement and visualization options organized along the units of analysis that are relevant for their research questions and their formulated hypotheses. Sometimes we make statements about the whole corpus, sometimes about different countries or actors or (sub-)issues; sometimes we compare countries or aggregated results along time

25 We integrated filter options which helped to identity the most 'relevant' topics for one's research question via key term searches.

26 We coded, e. g., references to 'multiple collective identities' in a randomly drawn subsample of the cleaned corpus (n = 6539). Based on the experience of previous projects, we generated a detailed codebook and performed continuous intercoder reliability tests on every eleventh newspaper article (cf. Renfordt 2011, pp. 94–101).

periods. The data output of many, even widely used commercial software packages, however, is often based on linguistically relevant units of analysis (e. g., the whole corpus, specific words, collocates, phrases, sentences).

Such outputs however are rarely interpretable for social scientists, and they cannot be directly connected to other data in social sciences (e. g., event-statistics, socio-demographics, public opinions, election results, etc.). As shown in the previous parts, social scientists look for textual references of their relevant complex concepts and subsequently count them in relation to different units of analysis. Social scientists *describe* the social world (e. g., as mirrored in texts), but they also want to *explain* it. We want to make claims about relations between variables. The indicators for these variables are to be found in different sources such as metadata (e. g. dates, geographic locations, names of newspapers), lists of terms or collocates as generated by second-generation quantitative approaches, or counts of occurrences (per texts, text passages, or other text units) of sophisticated measures for complex theoretical concepts.

During statistical analysis, these units of analysis must often be aggregated to larger dimensions (e. g. time periods, geographic locations, newspaper types). Yet the data output of many computational tools often does not allow compiling the data according to these needs. It is also often very difficult to switch from one data compilation (e. g., how often were religious identities discussed in different countries per year?) to another one (e. g., how big is the share of articles addressing religious identities in the European newspapers as compared to the American papers?).

By providing flexible data output options, tools should enable researchers to connect various variables with each other. Of course, the statistical analysis can be done with common statistics software; however, it should become easier to import text analytical data into these packages. Since every empirical sub-question together with its theoretically informed definition of dependent and independent variables requires a different kind of data arrangement, tools should provide very flexible options for data output and visualization.[27]

27 The *Complex Concept Builder* developed in *eIdentity* provides flexible data output (e. g., as CSV or Excel files which can be easily imported into SPSS or R for statistical analysis) and visualization options. Annotated data can be aggregated on different levels (e. g., temporal units such as days, months, years) and units of analysis (e. g., text, segment, sentence-level), and filtered by specific sub-samples (e. g., newspaper, language, country).

4 Conclusion

Because it seems so simple to 'run' innovative computational tools on millions of electronically available text data, it is tempting to use these procedures with great enthusiasm but little care. Especially in this take-off phase of big (textual) data research in our discipline, social scientists need to promote and defend the hard-won methodological standards of their scientific discipline. The analysis of large corpora will not, *per se*, lead to more knowledge or methodological progress. If the current boom of computer-linguistic methodology in the social sciences shall not turn into great disappointment soon, we must develop best practices of theory-guided empirical research that can significantly contribute to important scientific debates. For the application of computational tools and methods in the social sciences this necessarily involves time-intensive steps of manual annotation and qualitative validation, and a reflective selection of suitable corpora and appropriate (computational) methods.

In this contribution, we outlined the establishment of quality standards and best practices in reliable and valid text analysis in the social sciences since the 1990s. By focusing on the third, still ongoing phase of textual analysis, we argued that three hurdles impede the proliferation of quality-oriented corpus-analytic research in the social sciences. Flexible generic tools should standardize recurring tasks in order to make them much easier and more efficient, particularly during corpus creation, cleaning, and preprocessing. Our overall suggestion is to develop methods and workflows that systematically combine corpus-linguistic and qualitative steps of analysis to encourage the very individual theory driven operationalization of complex social scientific concepts in semantically valid ways. The better the new tools respond to the needs of flexible data-export and -output options for subsequent statistical analyses, the more these new methods will be able to prove themselves in discipline specific discourse and position themselves at eye level with established methods.

Funding: This article results from the collaborative interdisciplinary research project "Multiple Collective Identities in International Debates on War and Peace since the End of the Cold War" (e-Identity) that has been conducted by Prof. Dr. Cathleen Kantner, Prof. Dr. Jonas Kuhn, Prof. Dr. Manfred Stede and Prof. Dr. Ulrich Heid (funding period 2012–2015, funding code: 01UG1234A-C), as well as the "Center for Reflected Text Analytics" (CRETA, 2016–2020, funding code: 01UG1601). We would like to thank the Federal Ministry of Education and Research (BMBF) for the generous support of both projects. Last but not least, we want to thank Caitlin Reynolds for the meticulous and insightful language editing.

References

Alexa, Melina and Cornelia Zuell (2000). "Text analysis software: Commonalities, differences and limitations: The results of a review". In: *Quality and Quantity* 34 (3), pp. 299–321.

Baker, Paul and Tony McEnery (2005). "A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts". In: *Journal of Language and Politics* 4 (2), pp. 197–226.

Bayley, Paul and Geoffrey Williams (2012). *European identity: what the media say*. Oxford; New York: Oxford University Press.

Biber, Douglas (1993). "Representativeness in corpus design". In: *Literary and linguistic computing* 8.4, pp. 243–257.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022.

Blessing, André, Fritz Kliche, Ulrich Heid, Cathleen Kantner, and Jonas Kuhn (2015a). "Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorie". In: *Grenzen und Möglichkeiten der Digital Humanities. Sonderband der Zeitschrift für digitale Geisteswissenschaften*. Ed. by Constanze Baum and Thomas Stäcker. Vol. 1. DOI: 10.17175/sb001_013.

Blessing, André, Fritz Kliche, Ulrich Heid, Cathleen Kantner, and Jonas Kuhn (2015b). "Die Exploration großer Textsammlungen in den Sozialwissenschaften". In: *CLARIN Newsletter* 2015 (8), pp. 17–20.

Blessing, André, Jonathan Sonntag, Fritz Kliche, Ulrich Heid, Jonas Kuhn, and Manfred Stede (2013). "Towards a tool for interactive concept building for large scale analysis in the humanities". In: *Proceedings of the 7th Workshop on Language Technology for Cultural heritage, Social Sciences, and Humanities*, pp. 55–64.

Dick, Melanie, André Blessing, and Ulrich Heid (2015). "Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen". In: *Abstracts der DHd: Von Daten zu Erkenntnissen*. Graz.

Durkheim, Emile (1984 [1895]). *Die Regeln der soziologischen Methode*. Frankfurt: Suhrkamp.

Evans, Michael (2014). "A computational approach to qualitative analysis in large textual datasets". In: *PloS one* 9 (2), pp. 1–10.

Gabrielatos, Costas (2007). "Selecting query terms to build a specialised corpus from a restricted-access database". In: *ICAME Journal* 31, pp. 5–43.

Gabrielatos, Costas and Paul Baker (2006). "Representation of refugees and asylum seekers in UK newspapers: Towards a corpus-based analysis". In: *Joint Annual Meeting of the British Association for Applied Linguistics and the Irish Association for Applied Linguistics (BAAL/IRAAL 2006): From Applied Linguistics to Linguistics Applied: Issues, Practices, Trends*.

Grimmer, Justin (2010). "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases". In: *Political Analysis* 18 (1), pp. 1–35.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). "Studying the History of Ideas Using Topic Models". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 363–371. URL: https://www.aclweb.org/anthology/D08-1038 (visited on June 1, 2020).

Janasik, Nina, Timo Honkela, and Henrik Bruun (2009). "Text mining in qualitative research application of an unsupervised learning meth-od". In: *Organizational Research Methods* 12 (3), pp. 436–460.

Kantner, Cathleen (2011a). "Debating humanitarian military interventions in the European public sphere". In: *RECON Online Working Paper* 2011 (30), pp. 1–19.

Kantner, Cathleen (2011b). "European Identity as Commercium and Communio in Transnational Debates on Wars and Humanitarian Military Interventions". In: *RECON Online Working Paper* 2011 (37), pp. 1–22.

Kantner, Cathleen (2014). "The European public sphere and the debate about humanitarian military interven-tions". In: *European Security* 23 (4), pp. 409–429.

Kantner, Cathleen (2016). *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. London: Routledge.

Kantner, Cathleen, Amelie Kutter, Andreas Hildebrandt, and Mark Püttcher (2011). "How to get rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts". In: *International Relations Online Working Paper* 2011 (2), pp. 1–23.

Kantner, Cathleen, Amelie Kutter, and Swantje Renfordt (2008). "The Perception of the EU as an Emerging Security Actor in Media Debates on Humanitarian and Military Interventions (1990-2006)". In: *RECON Online Working Paper* 2008 (19), pp. 1–24.

Kantner, Cathleen and Maximilian Overbeck (2018). "Die Analyse 'weicher' Konzepte mit 'harten' korpuslinguistischen Methoden". In: *Computational Social Science: Die Analyse von Big Data*. Ed. by Andreas Blätte, Joachim Behnke, Kai-Uwe Schnapp, and Claudius Wagemann. Baden-Baden: Nomos Verlag, pp. 163–189.

Kliche, Fritz, André Blessing, Jonathan Sonntag, and Ulrich Heid (2014). "The e-identity exploration workbench". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14*, pp. 691–697.

Klüver, Heike (2009). "Measuring interest group influence using quantitative text analysis". In: *European Union Politics* 10 (4), pp. 535–549.

Krippendorff, Klaus (2018). *Content Analysis: An Introduction to Its Methodology (4th Edition)*. Thousand Oaks, CA: Sage.

Kuhn, Thomas S. (1962). *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main: Suhrkamp.

Kuhn, Thomas S. (1970). "Reflections on my critics". In: *Criticism and the Growth of Knowledge*. Ed. by Imre Lakatos and Alan Musgrave. Cambridge: Cambridge University Press, pp. 231–278.

Kutter, Amelie (2007). "Petitioner or partner? Constructions of European integration in Polish print me-dia debates on the EU Constitutional Treaty". In: *Discourse and Contemporary Social Change*. Ed. by Norman Fairclough, Giuseppina Cortese, and Patrizia Ardizzone. Bern: Peter Lang, pp. 433–457.

Kutter, Amelie and Cathleen Kantner (2012). "Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention". In: *International Relations Online Working Paper* 2012 (01), pp. 691–697.

Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data". In: *American Political Science Review* 97 (2), pp. 311–331.

Lemnitzer, Lothar and Heike Zinsmeister (2006). *Korpuslinguistik: Eine Einführung*. Tübingen: Gunter Narr.

Lowe, Will (2008). "Understanding wordscores". In: *Political Analysis* 16 (4), pp. 356–371.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, Cambridge: Cambridge University Press.

Neuendorf, Kimberly A. and Paul D. Skalski (2009). "Quantitative Content Analysis and the Measurement of Collective Identity". In: *Measuring Identity: A Guide for Social Scientists*. Ed. by Rawi Abdelal, Yoshiko M. Herrea, Alastair I. Johnston, and Rose McDermott. Cambridge: Cambridge University Press, pp. 203–236.

Overbeck, Maximilian (2014). "European debates during the Libya crisis of 2011: shared identity, divergent ac-tion". In: *European Security* 23 (4), pp. 583–600.

Overbeck, Maximilian (2015). "Observers turning into participants: Shifting perspectives on religion and armed conflict in Western news coverage". In: *The Tocqueville Review/La revue Tocqueville* 36 (2), pp. 95–124.

Pfetsch, Barbara and Annett Heft (2015). "Theorizing communication flows within a European public sphere". In: *European Public Spheres. Politics is Back*. Ed. by Thomas Risse. Cambridge: Cambridge University Press, pp. 29–52.

Pollak, Senja, Roel Coesemans, Walter Daelemans, and Na-da Lavrač (2011). "Detecting contrast patterns in newspaper articles by combining discourse analy-sis and text mining". In: *Pragmatics*, pp. 674–683.

Renfordt, Swantje (2011). *Framing the Use of Force: An International Rule of Law in Media Reporting. A Comparative Analysis of Western Debates about Military Interventions, 1990-2005*. Baden-Baden: Nomos.

Roberts, Carl W. (1989). "Other than counting words: A linguistic approach to content analysis". In: *Social Forces* 68 (1), pp. 147–177.

Rössler, Patrick (2010). *Inhaltsanalyse*. Stuttgart: UTB.

Rucht, Dieter, Peter Hocke, and Thomas Ohlemacher (1992). *Dokumentation und Analyse von Protestereignissen in der Bundesrepublik Deutschland (Prodat): Codebuch*. Berlin: WZB, Abt. Öffentlichkeit und Soziale Bewegung.

Slapin, Jonathan B. and Sven-Oliver Proksch (2008). "A scaling model for estimating time-series party positions from texts". In: *American Journal of Political Science* 52 (3), pp. 705–722.

Stalpouskaya, Katsiaryna and Christian Baden (2015). "To Do or Not to Do: the Role of Agendas for Action in Analyzing News Coverage of Violent Conflict". In: *Proceedings of the First Workshop on Computing News Storylines*. Beijing, China: Association for Computational Linguistics, pp. 21–29. DOI: 10.18653/v1/W15-4504.

Tietz, Udo (2002). *Die Grenzen des "Wir". Eine Theorie der Gemeinschaft*. Frankfurt am Main: Suhrkamp.

Weber, Max (1910). "Geschäftsbericht auf dem ersten Deutschen Soziologentage in Frankfurt 1910". In: *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*. Ed. by Marianne Weber. Tübingen: JCB Mohr, pp. 431–449.

Wessler, Hartmut (1998). "Issue". In: *Politische Kommunikation in der demokratischen Gesellschaft. Ein Hand-buch mit Lexikonteil*. Ed. by Otfried Jarren, Ulrich Sarcinelli, and Ulrich Saxer. Vol. 1. Opladen: Westdeutscher Verlag, p. 666.

Wiedemann, Gregor (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences*. Wiesbaden: VS Verlag.

Nils Reiter

Anleitung zur Erstellung von Annotationsrichtlinien

Zusammenfassung: In diesem Kapitel wird eine kurze Anleitung in die Entwicklung von Annotationsrichtlinien gegeben, unter der Annahme dass die Richtlinien ein Phänomen abdecken sollen, das bereits theoretisch beschrieben wurde. Ziel des Prozesses ist dann, Annotationsrichtlinien sowohl so generisch wie möglich als auch so präzise wie möglich zu gestalten. Menschliche Annotatorinnen und Annotatoren sollen auf Basis der Richtlinien sicher und zuverlässig annotieren können.

Abstract: This chapter gives a brief practical introduction into the development of annotation guidelines, for the scenario that new guidelines are created for a phenomenon or concept that has been described theoretically. In a single sentence, the goal of annotation guidelines can be formulated as: Given a theoretically described phenomenon or concept, describe it as generic as possible but as precise as necessary so that human annotators can annotate the concept or phenomenon in any text without running into problems or ambiguity issues.

1 Einleitung

Annotationsrichtlinien sollen ein Phänomen oder theoretisch gegebenes Konzept so generisch wie möglich, aber gleichzeitig so genau wie nötig beschreiben, damit menschliche Annotatorinnen und Annotatoren zuverlässig und intersubjektiv annotieren können. Mehrdeutigkeiten, also Stellen an denen mehrere Annotationskategorien möglich sind, sollen zumindest bemerkt werden und nicht stillschweigend untergehen. Die Erstellung der Richtlinien erfolgt dabei in einem iterativen Prozess: Sobald eine erste (Proto-)Version erstellt wurde, kann sie getestet werden. Dies erfolgt durch Anwendung, also die Annotation von Texten. Dadurch werden Unzulänglichkeiten sichtbar, die in einer nächsten Version der Richtlinien behoben werden. Diese wiederum werden getestet, was weitere Unzulänglichkeiten sichtbar macht. Diese können dann in einer weiteren Version behoben werden.

Anmerkung: Diese Anleitung wurde im Rahmen des *shared tasks* SANTA, Teil IV dieses Bandes, erstellt und zuerst auf https://sharedtasksinthedh.github.io/2017/10/01/howto-annotation/ publiziert. Für den Abdruck hier wurde sie übersetzt, überarbeitet und erweitert.

Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Nils Reiter, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-009

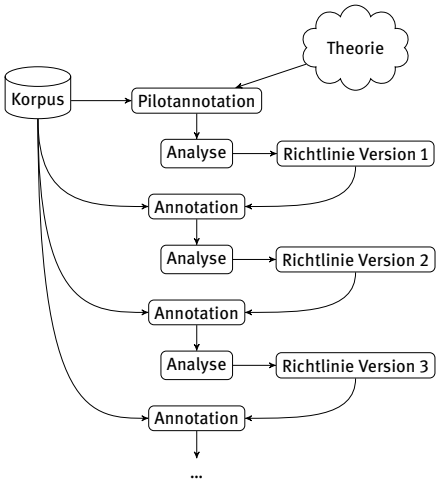


Abb. 1: Allgemeiner Arbeitsablauf. Der Prozess endet, wenn die Analyse ergibt, dass die Richtlinien hinreichend zuverlässig angewendet werden können.

Dieser Prozess ist schematisch dargestellt in Abbildung 1. Im folgenden wird genauer beschrieben, wie Richtlinien von einer Version zur nächsten weiterentwickelt werden können. Die wichtigste Idee dabei ist, dass *der gleiche Text von mehreren Annotatorinnen und Annotatoren unabhängig voneinander bearbeitet wird*.

Prinzipiell kann das gesamte Verfahren mit oder ohne Einsatz von Computern erfolgen. Digitale Annotationswerkzeuge erleichtern einige Arbeiten allerdings massiv, etwa den Vergleich von Annotationen verschiedener Annotatorinnen und Annotatoren. Außerdem zwingen sie die Annotatorinnen und Annotatoren dazu, sich darauf festzulegen, welche Wörter *genau* gemeint sind. Dies ist schwer auf Papier umzusetzen. Papier-basierte Annotationen sind dagegen zugänglicher und leichter aufzusetzen, können aber auch dazu führen, dass die Annotationen weniger exakt lokalisiert werden, oder ihrerseits nicht eindeutig sind (etwa weil Kategorien einfach am Seitenrand stehen). Für viele Arbeitsschritte, die sich an die Annotation anschließen, sind ohnehin exakte Grenzen nötig. Daher sollte man auch bei Papier-basierten Annotationen auf diese achten.

2 Pilotannotationen

Die erste Annotationsrunde wird am Besten von Annotatorinnen und Annotatoren durchgeführt, die mit der Theorie, gemäß der annotiert werden soll, vertraut

sind. Die Annotation sollte gleichzeitig und unabhängig voneinander durchgeführt und insbesondere sollten ad-hoc-Absprachen zwischen den Annotatorinnen und Annotatoren vermieden werden. Eine erste Annotationsrunde kann mit minimaler Vorbereitung starten. Im Regelfall sollte es genügen, eine oder mehrere wissenschaftliche Grundlagenwerke zur Orientierung und den zu annotierenden Text oder Textabschnitt festzulegen.

Ist ein Text oder Textabschnitt annotiert, sollten die Annotationen verglichen und besprochen werden. Dass die Annotationen schriftlich fixiert sind, zwingt die Annotatorinnen und Annotatoren dazu, ihre Entscheidungen ggf. begründen zu müssen. *Die Annotationsunterschiede zu diskutieren, ist ein zentraler Mechanismus, Unschärfen in den Richtlinien offenzulegen.* Gerade am Anfang sind viele Parameter noch nicht festgelegt und daher Ursache von Annotationsunterschieden. In der Diskussion sollte der Fokus zunächst auf den folgenden, eher formalen Punkten liegen. Ist der Annotationsprozess weiter fortgeschritten, verschiebt sich der Fokus zu inhaltlicheren Fragen (s. u.):

- Was genau wird annotiert? Jeder Absatz, jeder Satz, jedes Wort? Nur Einheiten, die eine bestimmte Bedingung erfüllen? Hier sollte eine möglichst klare und interpretationsunabhängige Einheit gewählt werden, so dass die eigentliche Entscheidung in der Kategorisierung der Einheiten liegt.
- Was genau sind die Annotationskategorien? Wie sind sie untereinander verwandt? Schließen sie sich zum Beispiel gegenseitig aus? In manchen Fällen hilft es vielleicht, die Kategorien zu hierarchisieren (z. B. *finites Verb* ist eine Unterkategorie von *Verb*).
- Mit welchem Tool wird annotiert? Müssen tool-spezifische Festlegungen getroffen werden? Sind alle Annotatorinnen und Annotatoren mit der Benutzung vertraut? Wie können Lesezeichen oder Kommentare gesetzt werden?

Diese ersten Entscheidungen und ihre Gründe sollten **dokumentiert** werden. Aus dieser Dokumentation ergeben sich dann nach und nach die Annotationsrichtlinien. Annotationsrichtlinien zeichnen sich in der Praxis durch eine Vielzahl von Beispielen aus. Es ist nie zu früh, mit dem Sammeln von interessanten, schwierigen und aufschlussreichen Beispielen anzufangen. Beispiele aus ‚echten' Texten sind zwar hilfreicher als ausgedachte, allerdings kann und sollte der ‚Knackpunkt' herausgearbeitet werden (z. B. indem Teile des Beispiels weggelassen werden, die für die Entscheidung nicht relevant sind).

3 Richtlinien verbessern

Um Richtlinien in diesem Szenario zu verbessern, müssen zuerst die Annotationen, die mit der früheren Version der Richtlinien erstellt wurden, analysiert werden. Dabei sollte der Fokus sinnvollerweise auf den Annotationsunterschieden liegen, also den Stellen, an denen die Annotatorinnen und Annotatoren unterschiedliche Entscheidungen getroffen haben. Diese können natürlich quantifiziert werden (s. u.), aber gerade zu Beginn ist es ergiebiger mit den Annotatorinnen und Annotatoren über ihre Entscheidungen zu sprechen. Dabei können die folgenden Fragen eine Orientierung geben:

- Welche Kategorien waren besonders einfach zu annotieren, welche besonders schwer?
- In welchen Fällen widersprachen Entscheidungen, die aufgrund der Richtlinien getroffen wurden, der eigenen Intuition?
- In welchen Fällen konnte keine Entscheidung getroffen werden, weil mehrere Kategorien ‚passten'?
- In welchen Fällen konnte keine Entscheidung getroffen werden, weil keine Kategorie ‚passte'?

Sind die Annotatorinnen und Annotatoren eingearbeitet und trainiert, genügt womöglich ein quantitativer Blick auf die Annotationsunterschiede zur Weiterarbeit. Dafür wurden eine Reihe von Metriken unter den Stichworten *inter-annotator agreement* oder *inter-rater reliability* vorgeschlagen (cf. Artstein und Poesio 2008, für eine Übersicht). Die meisten Metriken sind eine Kombination aus tatsächlich beobachteter (*observed*) und erwarteter (*expected*) Übereinstimmung. Die beobachtete Übereinstimmung wird aus den Annotationen ermittelt, etwa als Anteil der paarweisen Übereinstimmungen für die Metrik Fleiss' κ (Fleiss 1971). Erwartete Übereinstimmung zu messen, ist eher erklärungsbedürftig.

Im Falle von Fleiss' κ wird mit der erwarteten Übereinstimmung gemessen, welche Übereinstimmung erzielt worden wäre, wenn alle Annotatorinnen und Annotatoren zufällige Entscheidungen getroffen hätten. Die Überlegung dahinter ist, dass auch bei rein zufälligen Entscheidungen noch ein gewisses Maß an Übereinstimmung erzielt wird – wie hoch diese ist, hängt stark von der Anzahl der Kategorien ab. Um eine realistische Einschätzung der echten Übereinstimmung zwischen den Annotatorinnen und Annotatoren zu erhalten, wird nun die zufällige Übereinstimmung mit der beobachteten verrechnet. Das Ergebnis davon ist eine Zahl, etwa im Intervall zwischen $-\infty$ und 1, wobei ein Wert von 0 ausdrückt, dass

die Übereinstimmung der Annotatorinnen und Annotatoren *nicht besser* ist als die zufällige Übereinstimmung.[1]

Nur wenige Metriken differenzieren die Arten der Nicht-Übereinstimmung. In der Praxis ist es aber relevant, ob eine Nicht-Übereinstimmung aus Unachtsamkeit oder Missinterpretation der Definitionen erfolgt ist. Entscheidungen aus Unachtsamkeit sind leicht reparierbar und erfordern auch keine Neudefinition der Richtlinien, während Entscheidungen aus Missinterpretation der Definitionen sehr wohl Anlass geben, die Definitionen zu überdenken: Ein zusätzliches Beispiel kann diese klarer machen, oder auch eine Neuformulierung der Definition selbst. Gegebenenfalls kann auch die Klärung von Begriffen notwendig sein. Gius und Jacke (2017) unterscheiden in diesem Zusammenhang noch eine weitere Kategorie von Nicht-Übereinstimmungen, die je nach Komplexitätsgrad der Annotationsaufgabe benötigt wird: Divergierende Vorannahmen beziehen sich darauf, dass die Entscheidung für oder gegen eine Annotationskategorie von Interpretationen des Textes abhängen kann, die nicht Teil der Annotationsaufgabe sind. Es ist darüber hinaus zu erwarten, dass nicht alle Annotationsentscheidungen eindeutig sein werden. Einige Annotationsunterschiede basieren auf echten Mehrdeutigkeiten von Textstellen oder Texten und lassen sich auch nicht ausräumen.

Viele Annotationsunterschiede werden in der Praxis von Fällen verursacht, die in den Richtlinien bisher nicht abgedeckt sind. In diesem Fall muss entweder eine existierende Definition/Kategorie übertragen (und entsprechend erweitert) oder eine neue Kategorie definiert werden. Bei allen Änderungen an den Kategorien müssen deren Auswirkungen auf vergangene Annotationen mitgedacht werden. Hätte es die Annotationen verändert, wenn die Kategorie schon bestanden hätte? Teilweise muss dann ein Teil des Korpus neu annotiert (oder nicht mehr weiter verwendet) werden.

Die Diskussionen über konkrete Texte sind erfahrungsgemäß intensiv und verlaufen ungesteuert teilweise assoziativ. Eine gute Gesprächsführung und Moderation ist daher wichtig, um die verschiedenen Annotationsprobleme sauber zu trennen und auch die Entscheidungen zu dokumentieren. Nicht alle Probleme lassen sich auf einmal lösen, manchmal ist das Zurückstellen bestimmter Fragen die einzig richtige Antwort – wenn mehr Beispiele zum gleichen Phänomen diskutiert wurden und vorliegen, kann ggf. leichter eine Entscheidung getroffen werden.

Während der Prozess auf diese Weise mehrere Phasen durchläuft, passieren zwei Dinge gleichzeitig und miteinander verwoben: Sowohl die Annotationsrichtlinien, als auch die Annotatorinnen und Annotatoren werden besser, wobei ‚bes-

1 Verschiedene Metriken verwenden unterschiedliche Skalen.

ser' im letzten Fall auch heißen kann, dass sie besser antizipieren, was gewünscht wird, ohne dass es expliziert wurde. Eine höhere Übereinstimmung muss also nicht zwangsläufig bedeuten, dass die Richtlinien besser werden. Um das wirklich zu testen, bietet es sich an, gelegentlich neue Annotatorinnen und Annotatoren einzubeziehen, um deren Feedback zu bekommen. Dadurch werden auch ungeschriebene Regeln, die sich innerhalb eines Projektes entwickeln, transparent und sichtbar.

4 Annotationsrichtlinien finalisieren

Der in Abbildung 1 gezeigte Workflow kann theoretisch ohne Ende fortgeführt werden. Ist es das Ziel der Arbeit, theoretische Erkenntnisse über die annotierten Konzepte zu erzielen, kann die Annotation über sehr lange Zeiträume auch ein angemessenes Vorgehen sein (cf. Pagel et al. 2020, in diesem Band). Werden die Richtlinien zu dem Zweck entwickelt, möglichst viele Annotationen in einem Folgeschritt zu verwenden, muss eine Richtlinie zu einem Zeitpunkt als fertig deklariert werden. Eine Möglichkeit dazu ist, ein Mindestmaß an *inter-annotator agreement* zu definieren, das auf jeden Fall erreicht werden soll.

In Abbildung 2 findet sich eine Empfehlung für eine Gliederung einer Richtlinie. Zu beachten ist dabei, dass es sich bei Annotationsrichtlinien nicht primär um wissenschaftliche Texte handelt – die Richtlinien sollen zuvörderst von den Annotierenden verwendet werden. Dabei kann man drei unterschiedliche Verwendungsweisen unterscheiden: (i) Bei neu angefangenen Annotationsprojekten dienen die Richtlinien dazu, die zu annotierenden Einheiten und Kategorien erstmalig zu definieren. Sind die Annotierenden mit den Grundlagen hinreichend vertraut, entwickeln sich die Richtlinien weiter zu (ii) einem Nachschlagewerk, in dem vergangene Annotationsentscheidungen dokumentiert sind. Stößt eine Annotationskraft auf ein neues Beispiel, für das eine bereits getroffene Entscheidung ebenfalls angewendet werden kann, muss diese Entscheidung schnell auffindbar sein. Dies kann mit einem Index, einem großzügigen Layout oder einer Volltextsuche unterstützt werden. (iii) Ein dritter Anwendungsfall ist das Einarbeiten neuer Annotierender. Bei länger laufenden Annotationsprojekten lässt sich nicht vermeiden, dass die Annotierenden wechseln, und neue Annotierende eingelernt werden müssen. Diese werden die Richtlinien zunächst wieder von vorne lesen, aber ohne den Entwicklungsprozess miterlebt zu haben. Es ist also darauf zu achten, dass auch bei vielen kleineren Änderungen an den Richtlinien, die einleitenden Teile noch zum Kern der Richtlinien passen. Diesem Anwendungsfall ähnlich

1. Einleitung
   - Worum geht es? Was soll annotiert werden?
   - Worauf basieren die Richtlinien?
   - Wer hat sie wann in welchem Projekt erstellt?
   - An wen richten sie sich? Welche Kenntnisse werden auf Seiten der Annotierenden vorausgesetzt?
2. Annotationseinheiten
   - Was sind die zu annotierenden Einheiten (z. B. Wörter, Sätze, Absätze, …)?
   - Sind alle Einheiten zu annotieren oder nur bestimmte? Woran erkennt man sie?
3. Annotationskategorien
   - Welche Kategorien werden den Einheiten zugewiesen?
   - Woran erkennt man eine Kategorie?
   - Welche Kategorien sind ggf. leicht zu verwechseln?
   - Gibt es Abhängigkeiten zwischen den Kategorien?
4. Problematische Fälle
   - Was sind schwierige Fälle?
   - Wie wurde bei denen entschieden, und warum?
5. Praktische Umsetzung
   - Wie genau ist das Annotationstool zu verwenden?
   - Falls das Annotationstool eine eigene Anleitung bereitstellt, kann auch auf diese verwiesen werden. In dem Fall sollte aber ggf. eine ‚Übersetzung' der Begriffe mitgeliefert werden.
6. Änderungsprotokoll
   - Welche Stellen wurden in welcher Iteration geändert?

Abb. 2: Gliederungsempfehlung für Annotationsrichtlinien

ist auch die Weitergabe an andere Forschende, die sich für das annotierte Korpus, die annotierten Konzepte oder die Annotationsrichtlinien als solche interessieren.

Stilistisch empfehlen wir, eher Stichpunkte und Aufzählungen als Fließtext zu verwenden. Die Nummerierung von Beispielen macht es leichter mit ihnen zu arbeiten. Wissenschaftliche Referenzen können im Sinne der Nachvollziehbarkeit angegeben werden, es sollte aber eher vermieden werden, deren Rezeption zur Voraussetzung zu machen. Für das Sammeln und Erstellen der Richtlinien hat sich in CRETA ein Wiki als geeignet erwiesen, es sollte allerdings darauf geachtet werden, dass eine Druckversion extrahierbar ist (in längeren Richtlinien ist die Orientierung und Zugänglichkeit eine Herausforderung, und im Druck fällt das vielen Menschen leichter).

5 Beispiele für Annotationsrichtlinien

Beispiele für Annotationsrichtlinien zum literarischen Phänomen der Erzählebene finden sich in der *Cultural-Analytics*-Sonderausgabe zum *SANTA shared task* (Gius, Reiter et al. 2019). Mit Barth (2020) und Ketschik, Murr et al. (2020) befinden sich zwei von ihnen auch in diesem Band ab Seite 423. Daneben dokumentiert der vorliegende Band auch die interdisziplinäre Entwicklung einer Annotationsrichtlinie für Entitätenreferenzen (Ketschik, Blessing et al. 2020, S. 204 ff.), sowie die Konzeptualisierung von Annotationen für Emotionen in Erzähltexten (Klinger et al. 2020, S. 238 ff.). Eine Sammlung computerlinguistischer Annotationsrichtlinien wurde von Ide und Pustejovsky (2017) publiziert.

Danksagung: Am Inhalt dieser Anleitung haben Evelyn Gius und Marcus Willand im Rahmen des SANTA-*shared tasks* (Willand et al. 2020, in diesem Band) mitgewirkt.

Literatur

Artstein, Ron und Massimo Poesio (2008). „Inter-Coder Agreement for Computational Linguistics“. In: *Computational Linguistics* 34.4, S. 555–596.

Barth, Florian (2020). „Annotation narrativer Ebenen und narrativer Akte“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 423–438.

Fleiss, Joseph L. (1971). „Measuring nominal scale agreement among many raters“. In: *Psychological Bulletin* 76.5, S. 420–428.

Gius, Evelyn und Janina Jacke (2017). „The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis“. In: *International Journal of Humanities and Arts Computing* 11.2, S. 233–254.

Gius, Evelyn, Nils Reiter und Marcus Willand, Hrsg. (2019). *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*.

Ide, Nancy und James Pustejovsky, Hrsg. (2017). *Handbook of Linguistic Annotation*. Berlin/Heidelberg: Springer.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Klinger, Roman, Evgeny Kim und Sebastian Padó (2020). „Emotion Analysis for Literary Studies“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 238–268.

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.

Willand, Marcus, Evelyn Gius und Nils Reiter (2020). „SANTA: Idee und Durchführung“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 391–422.

Nora Ketschik, André Blessing, Sandra Murr,
Maximilian Overbeck und Axel Pichler

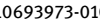
Interdisziplinäre Annotation von Entitätenreferenzen

Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung

Zusammenfassung: Der Beitrag präsentiert die Ergebnisse der interdisziplinären Auseinandersetzung mit Entitäten. Wir erarbeiten disziplinübergreifende Richtlinien für eine reliable und semantisch valide Annotation von Entitätenreferenzen sowie einen gemeinsamen methodischen Workflow für eine semi-automatische Klassifikation großer Textmengen. Aus vier sozial- und geisteswissenschaftlichen Perspektiven heraus diskutieren wir anschließend Herausforderungen und Lösungsmöglichkeiten, die bei der Anwendung unseres Ansatzes auf die jeweiligen fachspezifischen Textkorpora auftreten. Wir zeigen, dass die interdisziplinäre Kooperation zur methodischen Stringenz verpflichtet und zu einer reflexiven Entwicklung der Analysekonzepte und Operationalisierungsprozesse beiträgt.

Abstract: This contribution presents the results of our interdisciplinary engagement with entities. We developed interdisciplinary guidelines for a reliable and semantically valid annotation of entity references, and a methodological workflow for their semi-automatic identification within large amounts of textual data. From the perspective of four different disciplines within the Humanities and Social Sciences, we discuss challenges related to the application of a generic workflow to heterogeneous text corpora, and present possible solutions. We conclude that the interdisciplinary collaboration enhances the overall methodological stringency and fosters conceptual reflections within each participating discipline.

Nora Ketschik, Institut für Literaturwissenschaft, Universität Stuttgart
Maximilian Overbeck, Department of Communication and Journalism, Hebrew University of Jerusalem
Sandra Murr, Axel Pichler, Stuttgart Research Center for Text Studies, Universität Stuttgart
André Blessing, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Nora Ketschik, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler; publiziert von De Gruyter  Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-010

1 Einleitung

In diesem Beitrag präsentieren wir die Resultate der interdisziplinären methodischen Auseinandersetzung mit Entitäten und deren Referenzen, die im Rahmen des DH-Projekts CRETA erzielt wurden.[1] Ursprünglich stammt der Begriff ‚Entität' aus der Philosophie und zwar aus der allgemeinen Metaphysik bzw. Ontologie, die die elementaren Grundstrukturen der Wirklichkeit untersucht. Hierin bezeichnet er ein einzelnes unteilbares Seiendes. Entitäten spielen insbesondere in den Kategorienlehren eine zentrale Rolle, im Rahmen derer die Philosophie seit der Antike den Gesamtbereich des Seienden zu klassifizieren versucht.[2] Derartige Klassifikationspraktiken sind auch in anderen Disziplinen relevant. So untersucht die Politikwissenschaft beispielsweise, wie in den verschiedenen Arenen des politischen und gesellschaftlichen Lebens auf Entitäten wie politische Akteure, Parteien oder Organisationen referiert wird. Für die germanistische Mediävistik und die Neuere Deutsche Literatur steht die Erfassung von Referenzen auf Figuren und deren Relationen zueinander in den fachspezifischen Texten im Fokus. Im Rahmen von CRETA wurde ein Umgang mit Entitäten und deren Referenzen angestrebt, der für Fragestellungen aus den unterschiedlichsten Fächern von Interesse und Nutzen sein sollte. Im Unterschied zur *named entity recognition* (NER), die das Ziel verfolgt, Eigennamen im Text zu erkennen (Jurafsky und Martin 2008, S. 793–746; Carstensen et al. 2010, S. 596–599; Clark et al. 2013, S. 518–522)[3] sind für uns neben Eigennamen auch Gattungsnamen von Interesse. Zudem erweitern wir das Repertoire der gängigen Entitätenklassen PER (Personen, Figuren), LOC (Orte) und ORG (Organisationen) um die Klassen WRK (Werke) und CNC (Abstrakte Konzepte). Im Kontext der hier vorgestellten Projekte bezeichnet ‚Entität' folglich isolierte reale, fiktive oder mögliche Objekte, auf die von der Textoberfläche mittels eines einzelnen Wortes oder einer abgegrenzten Wortfolge referiert werden kann.

Während sich einzelne Disziplinen aus unterschiedlichen inhaltlichen Gesichtspunkten für Entitäten interessieren, lassen sich auf der Oberfläche eines Textes nur Entitäten*referenzen* erfassen, also diejenigen sprachlichen Ausdrücke, die auf eine bestimmte Entität in der realen, fiktiven oder einer möglichen Welt

1 Siehe dazu auch Kuhns Einleitung des Bandes ab Seite 9.

2 Seit dem 19. Jahrhundert herrscht in der Philosophie Skepsis darüber, ob philosophische Kategorienlehren ihre Minimalanforderungen – d. s. die Vollständigkeit und Exaktheit der Kategorien – erfüllen bzw. welche Kriterien zur Überprüfung dieses Sachverhaltes anzuwenden sind. Dies bedeutet jedoch nicht, dass die Philosophie gegenwärtig auf die Arbeit an Kategoriensystemen verzichtet.

3 Die NER wurde v. a. durch die Message Understanding Conferences (MUC) 6 und 7 (1994–1998) populär.

referieren. Die Annotation von Entitätenreferenzen erfolgt im Rahmen von CRETA vor dem Hintergrund verschiedener Forschungsfragen und in Textkorpora unterschiedlicher Disziplinen: in einem sozialwissenschaftlichen Korpus bestehend aus Plenardebatten des deutschen Bundestags, in einer Sammlung von *Werther*-Adaptionen und ihrer Vorlage, Goethes *Die Leiden des jungen Werthers*, in einem altgermanistischen Korpus mittelhochdeutscher Artusromane, sowie in Schriften zur philosophischen Ästhetik des 20. Jahrhunderts.

Die Auseinandersetzung mit Entitäten und deren Referenzen stellte ein übergreifendes Forschungsinteresse aller an CRETA beteiligten Fächer dar und wurde bewusst an den Anfang des Förderzeitraums gesetzt, um eine gemeinsame Grundlage für darauf aufbauende Analysen zu schaffen. Zudem sollte sie dazu beitragen, disziplinäre Fragestellungen weiterzuentwickeln und zu präzisieren.[4]

Die Struktur unseres Beitrags orientiert sich an der Chronologie unseres interdisziplinären Workflows, so wie er bei der Analyse der Entitäten und ihrer textuellen Referenzen entstanden ist. Im ersten Abschnitt gehen wir ausführlich auf die gemeinsame Entwicklung der Annotationsrichtlinien ein. Mit den Richtlinien wird das Ziel verfolgt, die im Rahmen von CRETA untersuchten Entitätenklassen auf Basis valider Textindikatoren zu erfassen. Wir veranschaulichen den disziplinübergreifenden Workflow, der qualitative und quantitative Analyseschritte systematisch miteinander kombiniert und die Identifikation von Entitätenreferenzen zunächst auf kleinen und anschließend großen Textkorpora anleitet. Im zweiten Abschnitt gehen wir genauer auf die vier geistes- und sozialwissenschaftlichen Zugänge zu Entitäten ein und widmen uns kritisch den Problemen und Herausforderungen, die bei der Erfassung und Analyse von Entitätenreferenzen in unterschiedlichen Textkorpora entstanden sind. Im dritten Abschnitt fassen wir in einer Schlussbetrachtung unsere Forschungsergebnisse zusammen.

2 Interdisziplinärer Austausch und gemeinsame methodische Umsetzung

2.1 Annotationsrichtlinien

Am Beginn des interdisziplinären Annotationsprozesses stand die gemeinsame Entwicklung fachübergreifender Annotationsrichtlinien. Sie wurden im dialogi-

4 Siehe zu interdisziplinärer Kommunikation auch den Beitrag „*Reaching out*: Interdisziplinäre Kommunikation und Dissemination“ ab Seite 467 (Reiter, Kremer et al. 2020).

schen Austausch und auf der Basis fachspezifischen Wissens über die ausgewählten Textkorpora und deren jeweilige Spezifika entwickelt.[5] Grundlegend für uns war dabei zunächst, einen gemeinsamen Zugriff auf Entitäten und ihre Referenzen zu erarbeiten. Wir gingen dabei von der NER aus und orientierten uns an den gängigen Annotationspraktiken aus der Computer- und Korpuslinguistik. Im Zuge der interdisziplinären Erarbeitung von Annotationsrichtlinien erwies sich für die in CRETA versammelten Disziplinen eine Entitäten-Konzeption als sinnvoll, die sich zwischen der Eigennamenerkennung (NER, Jurafsky und Martin 2008, S. 739–746; Carstensen et al. 2010, S. 596–599; Clark et al. 2013, S. 518–522) und der Koreferenzresolution (Jurafsky und Martin 2008, S. 694, 707–724; Clark et al. 2013, S. 522 f.) bewegt: Die CRETA-Annotationsrichtlinien gehen über die reine Eigennamenerkennung hinaus, da auch Entitätenreferenzen via Gattungsnamen berücksichtigt werden, sie bleiben aber hinter einer vollständigen Koreferenzresolution zurück, da weder pronominale Referenzen erfasst noch alle Entitätenreferenzen zu Koreferenzketten gruppiert werden.

Neben solch grundlegenden Entscheidungen hatten die Richtlinien vor allem den Zweck, präzise Handlungsanweisungen für das Annotieren von Entitäten zu formulieren: So geben die CRETA-Annotationsrichtlinien z. B. vor, immer syntaktische Einheiten, also maximale Nominalphrasen zu annotieren, so dass auch Artikel, Adjektive, Appositionen und Relativsätze Teil des Referenzausdrucks sein können (z. B. „meine liebste Lotte, mein Engel“). Dies ist zum einen mit der gängigen Praxis in der Computerlinguistik konform (Reznicek 2013a,b) und hat zum anderen den Grund, dass semantische, die Entität näher bestimmende Informationen für weiterführende Analysen miterfasst werden. Durch die Annotation der maximalen Nominalphrase kann es zu ‚Verschachtelungen' von Entitätenreferenzen kommen, wobei immer alle Referenzausdrücke – alle eingebetteten wie der maximale – zu annotieren sind (z. B. „**[**Angela Merkel, die Kanzlerin **[**der Bundesrepublik Deutschland**]]**“). Darüber hinaus wird eine Entitätenreferenz immer in ihrem jeweiligen semantischen Kontext interpretiert und annotiert, um eventuelle Ambiguitäten aufzulösen – z. B. kann sich „Washington“ je nach Kontext auf einen Ort (LOC), eine Organisation (ORG, hier die US-Regierung), oder auf die Person (PER, ‚George Washington') beziehen.

Generell wird in CRETA eine Vollannotation angestrebt, was bedeutet, dass alle Entitätenreferenzen in einem Text unabhängig von ihrem Status oder ihrer Relevanz annotiert werden. So kann für den späteren Einsatz maschineller Lernverfahren eine vollständige Datengrundlage geschaffen werden. Die Vollannotation

5 Unsere Annotationsrichtlinien orientieren sich grob an Reznicek (2013b) und Reznicek (2013a). Die vollständigen Annotationsrichtlinien sind im Anhang beigefügt.

macht es notwendig, gegebenenfalls in einem zweiten Schritt über die Relevanz einer Entitätenreferenz für die jeweilige Forschungsfrage zu urteilen oder weitere Merkmale zu ihrer Differenzierung (z. B. die Unterscheidung von historischer Person und fiktiver Figur) hinzuzufügen.

Die Annotationsrichtlinien werden durch Mitarbeitende in einer ersten Version aufgesetzt und an Textbeispielen der verschiedenen Disziplinen geprüft. In regelmäßigen Projekttreffen werden die Anwendungsfälle besprochen und die Richtlinien erweitert, präzisiert und korrigiert. Durch die direkte Anwendung werden sehr praxisnahe Richtlinien erstellt, die einerseits durch Beispiele konkretisiert, andererseits durch die heterogenen Beispieltexte allgemein gehalten werden und so für verschiedene Disziplinen nutzbar sind.

Ein wesentlicher Beitrag des interdisziplinären Vorgehens liegt in der Entwicklung eines Workflows für die transparente und intersubjektiv nachvollziehbare manuelle Annotation der Referenzausdrücke. Die zentrale Herausforderung besteht darin, die *Replizierbarkeit* der Annotationen zu ermöglichen, so dass unterschiedliche Annotierende mit derselben Annotationsrichtlinie zu einheitlichen Annotationsergebnissen gelangen. Darüber hinaus sollen die annotierten Referenzausdrücke die Entitätentypen semantisch valide abbilden.

2.2 Technische Umsetzung und Workflow

2.2.1 Entitätenannotation in CRETAnno

Die Annotation der Entitätenreferenzen wird in einem eigens für das Projekt entwickelten Annotationstool – CRETAnno[6] – vorgenommen, das den Annotierenden über eine intuitive graphische Benutzeroberfläche die Möglichkeit gibt, mittels Markierung einen Textausschnitt (ein oder mehrere Wörter) als Entitätenreferenz auszuzeichnen (Abbildung 1). Bei der Auszeichnung wird über ein ‚Pop-up'-Fenster zugleich die Entitätenklasse festgelegt, der die Entität angehört. Bei Unklarheiten kann die Kategorie ‚Wiedervorlage' angewählt werden, die es erlaubt, die problematischen Fälle zu kommentieren und sie in einem späteren Schritt nochmals separat anzeigen zu lassen.

Für die Inspektion paralleler Annotationen ein und derselben Textstelle durch mehrere Annotierende steht darüber hinaus eine spezielle Ansicht (*Gold-Overview*) zur Verfügung. Sie hebt nicht-übereinstimmende Annotationen hervor

6 http://hdl.handle.net/11022/1007-0000-0007-E1BE-5

Abb. 1: Benutzeroberfläche zur Entitätenannotation in CRETAnno

und gibt den Annotierenden so die Möglichkeit, diese Fälle zu diskutieren und infolgedessen eine der Annotationen als korrekt festzulegen (Abbildung 2).

2.2.2 Manueller Annotationsprozess

Der Arbeitsablauf zur manuellen Annotation ist in Abbildung 3 zusammengefasst. Die Annotationsrichtlinien werden zunächst einem iterativen Prozess aus Anwendung und Überarbeitung unterzogen, indem a) die Richtlinien händisch durch mehrere Annotierende auf ausgewählte Textstellen angewendet (Parallelannotation), b) die Annotationen anschließend verglichen und c) die Richtlinien entsprechend überarbeitet werden.[7] Bei diesem wiederholenden Vorgehen entstehen verschiedene Versionen der Richtlinien, die aufeinander aufbauen und schließlich in einer finalen Version münden. Für die Überarbeitung der Richtlinien erweist sich der systematische Abgleich nicht-übereinstimmender Annotationen als fruchtbar, da dieser die ‚Schwachstellen' der Richtlinien – etwa unzureichende Präzision oder nicht bedachte Fälle – offenlegt. Die vergleichende Parallelannotation wird so lange fortgeführt, bis die Entitätenreferenzen adäquat erfasst und eine möglichst hohe Übereinstimmung zwischen den Annotationen erzielt wird. Der zykli-

7 Vgl. hierzu auch die Anleitung zur Erstellung von Richtlinien von Reiter (2020), ab Seite 193 in diesem Band

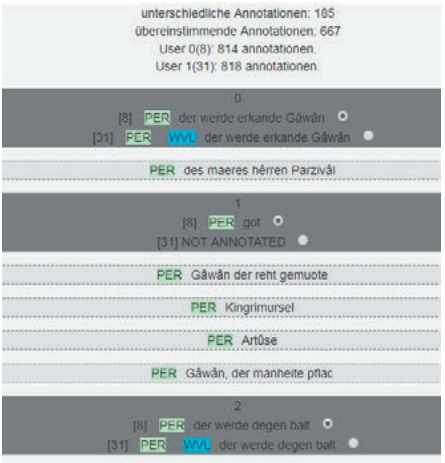


Abb. 2: *Gold-Overview* in CRETAnno mit Auflistung übereinstimmender und nicht-übereinstimmender Annotationen

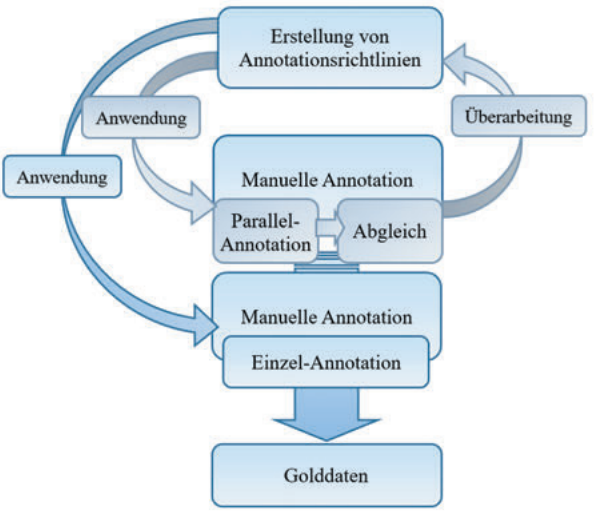


Abb. 3: Workflow zur manuellen Annotation

sche Prozess verfolgt damit zum einen den Zweck, Unklarheiten in den Richtlinien zu beseitigen; zum anderen schult die wiederholte Anwendung die Annotierenden und befähigt sie, die Richtlinie anschließend autonom auf neue Texte anzuwenden, so dass im Anschluss an die Parallelannotation die Annotation in Einzelarbeit fortgeführt werden kann.

Die intensive manuelle Annotation führt nicht nur zu einer Schärfung der verwendeten Kategorien und Konzepte, sondern sie deckt auch text- bzw. korpusspezifische Besonderheiten auf. Das hat zur Folge, dass die unterschiedlichen Teilgruppen die generischen Annotationsrichtlinien stellenweise genauer an ihre jeweiligen Textsorten anpassen müssen. Für die Annotation des sozialwissenschaftlichen Plenardebatten-Korpus wurden z. B. Zusatzrichtlinien und Beispiele eingefügt, die die Disambiguierung zwischen Orten und Organisationen erleichtern; für die fiktionalen Texte in den literaturwissenschaftlichen Korpora wird die Annotation von Orten weiter geschärft und für die mittelhochdeutschen Artusromane muss zusätzlich über die Nennung von Gott, Heiligen oder menschenähnlichen Handlungsträgern (Riesen, Zwerge, Feen) entschieden werden. Darüber hinaus machen sprachliche Besonderheiten des Mittelhochdeutschen (z. B. Wortverschmelzungen) es notwendig, die Grenzen des zu annotierenden Ausdrucks genauer zu bestimmen. Insgesamt wird der disziplinübergreifende Zugriff auf Entitäten also um disziplinspezifische Zusätze ergänzt, die den Besonderheiten der jeweiligen Korpora Rechnung tragen.

2.2.3 Semi-automatischer Annotationsprozess

Durch die händische Annotation der Entitätenreferenzen in den verschiedenen Korpora werden hochqualitative Daten (sog. ‚Golddaten') erzeugt, die daraufhin zum Training eines Entitätenreferenzerkenners (ERT) verwendet werden. Während die Annotationsrichtlinien für die Entitätenannotation disziplinübergreifend entwickelt werden können, erweist es sich für den maschinellen Lernprozess als sinnvoll, disziplinspezifische Modelle zu trainieren, so dass das Vorhersagemodell die Besonderheiten der jeweiligen Texte erlernen kann (vgl. Schulz 2018, S. 35–39).

Abbildung 4 beschreibt die Vorgehensweise der semi-automatischen Annotation, die im Anschluss an die händische Annotation angestoßen wird: Basierend auf den manuell annotierten Daten einer bestimmten Disziplin wird ein Entitätenreferenztagger trainiert, der Vorhersagen über Entitätenreferenzen in neuen Texten derselben Disziplin generiert. Daraufhin werden die automatisch erzeugten Annotationen einem Korrekturprozess unterworfen, bei dem die Annotationen als korrekt oder falsch bestimmt und darüber hinaus mögliche fehlende Entitäten-

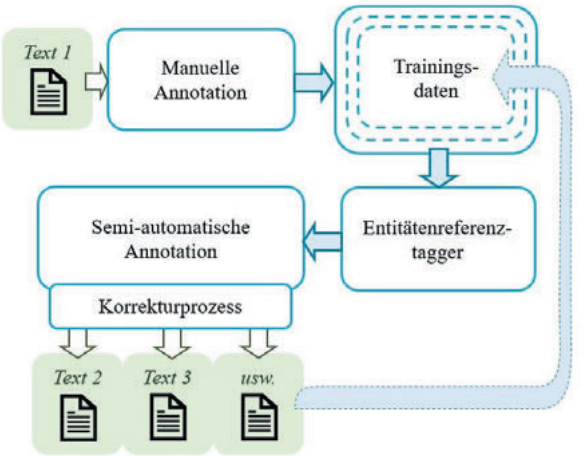


Abb. 4: Workflow zur semi-automatischen Annotation

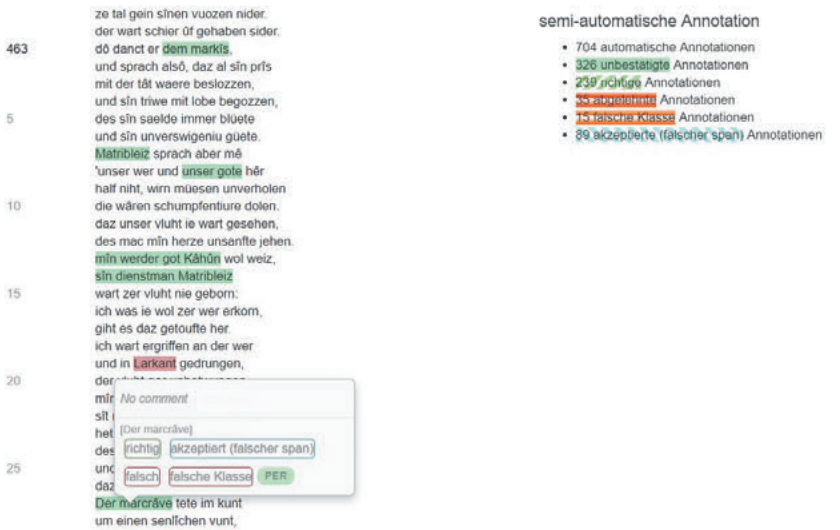


Abb. 5: Benutzeroberfläche zur semi-automatischen Annotation in CRETAnno

Tab. 1: Anzahl der Tokens und Entitätenreferenzen in den manuell annotierten CRETA-Subkorpora.

Subkorpus	Tokens	Entitäten
Literaturwissenschaft (Werther)	41 505	331
Philosophie (Adorno)	13 233	929
Mediävistik (Parzival)	30 491	2001
Sozialwissenschaft (Bundestagsdebatten)	6371	488

Korpus	Cohen's κ
Literaturwissenschaft (*Werther*)	0,66
Philosophie (Adorno)	0,89
Mediävistik (*Parzival*)	0,8
Sozialwissenschaft (Bundestagsdebatten)	0,78

Tab. 2: IAA-Werte (Cohen's κ) in den parallel annotierten CRETA-Subkorpora

referenzen händisch nachannotiert werden (Abbildung 5). Auf diese Weise kann die Qualität der Daten gewährleistet werden, so dass diese anschließend zum (Re)Trainieren verwendet werden können. Der sukzessive Aufbau der Trainingsdaten bietet den Vorteil, den Entitätenreferenzerkenner nach und nach auf neue Texte anwenden zu können und hierbei einem Absinken der Performance entgegenzuwirken, die typischerweise aus der Divergenz zwischen Quell- und Zieldaten hervorgeht.

Durch die semi-automatische Vorgehensweise wird einerseits eine erhebliche Zeitersparnis gegenüber der händischen Annotation erzielt, andererseits aber die hohe Qualität der Daten sichergestellt, so dass die anschließenden Untersuchungen auf einer adäquaten Datengrundlage aufbauen können.

2.3 Korpusstatistik und Evaluation

Die Modelle des Entitätenreferenzerkenners basieren auf heterogenen Daten, die in Hinblick auf den Umfang der Tokens und Entitäten stark variieren können. Tabelle 1 gibt einen Überblick über die Größe der manuell annotierten Korpora.

Aus den parallel annotierten Textausschnitten lässt sich ein *inter-annotator agreement* (IAA) errechnen, das die übereinstimmenden Annotationen in ein Verhältnis zu den nicht-übereinstimmenden setzt und damit einen Richtwert für die Klarheit und Intersubjektivität der Guidelines sowie für die Reproduzierbarkeit der Annotationen darstellt (vgl. Carstensen et al. 2010, S. 152). In unserem Fall wird das IAA-Maß Cohen's κ (Cohen 1960) verwendet. Generell wird bei einem Wert von $\kappa \leq 0{,}8$ eine hohe Reliabilität angenommen. Aus den Subkorpora des

Tab. 3: Evaluation des Entitätenreferenztaggers (ERT) im Vergleich zur NER-*baseline*

Klasse	Korpus	NER (baseline)			ERT		
		precision	*recall*	F_1	*precision*	*recall*	F_1
PER	Bundestagsdebatten	31,58	0,31	0,61	40,62	0,67	1,32
	Adorno	35,34	2,42	4,53	45,45	0,52	1,03
	Parzival	42,86	0,46	0,91	63,24	21,56	32,16
	Werther	59,09	5,36	9,83	65	12,07	20,26
LOC	Bundestagsdebatten	33,82	3,48	6,31	72,22	1,97	3,89
	Parzival	28,57	0,3	0,59	63,16	10,91	18,61
	Werther	28,21	1,67	3,15	69,8	11,52	19,78
ORG	Bundestagsdebatten	3,7	0,32	0,59	42,28	37,1	39,52
WRK	Bundestagsdebatten	0	0	0	42,86	4,05	7,25
	Werther	0	1,35	0	0	0	0
CNC	Adorno	0	0	0	60,66	45,73	52,18

CRETA-Projekts ergeben sich IAA-Werte zwischen 0,66 und 0,89 (vgl. Tabelle 2), wobei die Berechnung die zeitliche Dimension des Annotationsprozesses unberücksichtigt lässt und diesbezüglich davon auszugehen ist, dass die anfängliche Übereinstimmung geringer ausfiel und der Wert nach Schärfung der Richtlinien und Training der Annotierenden anstiegen ist.

Die Sichtung der nicht-übereinstimmenden Annotationen zeigt, dass inkongruente Annotationen dreierlei Ursachen haben können: Erstens können Unstimmigkeiten darüber herrschen, ob es sich bei einem Ausdruck um eine Entitätenreferenz handelt oder nicht; zweitens können die Annotierenden zwar den gleichen Ausdruck annotiert, dabei aber verschiedene Entitätenklassen ausgewählt haben (was insbesondere metonymische Verwendungsweisen betrifft); drittens können Inkongruenzen darauf zurückgehen, dass die Grenzen eines Referenzausdrucks nicht übereinstimmend annotiert wurden. Die iterative Überarbeitung der Richtlinien kann generell nur diejenigen Inkongruenzen beseitigen, die auf unzureichend geschärften Definitionen oder auf unterschiedlichen Vorverständnissen seitens der Annotierenden zurückzuführen sind (vgl. Gius und Jacke 2017).

Wie in Abschnitt 2.2 beschrieben, werden auf Grundlage der manuell annotierten Daten korpusspezifische Entitätenreferenztagger trainiert. Diese werden

im Vergleich zu einer *baseline* evaluiert.[8] Die *baseline* repräsentiert den Status quo für eine Aufgabe und bildet damit einen Ausgangspunkt für die Evaluation neuer Modelle. Für die Aufgabe der Entitätenreferenzerkennung wurde als *baseline* ein Modell des *Stanford Named Entity Recognizer* (Finkel et al. 2005) für das Deutsche verwendet (Faruqui und Padó 2010). Tabelle 3 fasst die Evaluation der Modelle zusammen. Der *precision*-Wert steht für die Genauigkeit eines Systems und gibt an, wie viele der vorhergesagten Annotationen korrekt sind, der *recall* misst die Vollständigkeit, indem er angibt, wie viele der zu identifizierenden Objekte von einem System gefunden worden sind (vgl. Manning und Schütze 1999, S. 267–269). Im F_1-*score* werden die Werte *precision* und *recall* zusammengefasst (harmonisches Mittel, vgl. Manning und Schütze 1999, S. 269). Als korrekt (*true positive*) wurden in dieser Evaluation nur die vorhergesagten Referenzen gezählt, die mit den Referenzdaten *exakt* übereinstimmen und z. B. auch anhängende Relativsätze korrekt erkennen.

Der in CRETA verwendete Entitätenreferenzerkenner basiert auf dem Verfahren *conditional random fields* (Lafferty et al. 2001), bei dem Kontextabhängigkeiten zwischen den Entscheidungen für einzelne Wörter mit berücksichtigt werden. Im Gegensatz zu einer einfachen Klassifikation, bei der individuelle Wörter unabhängig von den Entscheidungen im Kontext klassifiziert würden, kann das Modell erlernen, dass ein Wort eher Teil einer Entitätenreferenz ist, wenn davor z. B. ein definiter Artikel steht. Die verwendeten Merkmale sind die Wortform, die (automatisch erkannte) Wortart, eine Angabe darüber, ob die Wortform groß- und kleingeschrieben im Korpus vorkommt, eine Namensliste (die korpusspezifische häufige Namen enthält), sowie eine Charakterisierung der Buchstabenmuster, die im Wort vorkommen.[9] Für die technische Umsetzung wurde die Bibliothek clearTk[10] (Bethard et al. 2014) verwendet, die ihrerseits das Tool mallet[11] (McCallum 2002) einbindet. Trainiert und getestet wurde mit fünffacher Kreuzvalidierung.

Die Evaluation zeigt, dass die korpusspezifisch trainierten Entitätenreferenztagger insgesamt deutlich bessere Ergebnisse erzielen können als das *state of the*

8 Die Evaluation des Entitätenreferenztaggers erfolgte im Rahmen des Workshops „CUTE: CRETA Unshared Task zu Entitätenreferenzen" bei der DHd 2017 (Reiter, Blessing et al. 2017). Das annotierte Korpus steht über folgenden Link zum Download zur Verfügung: https://www.creta.uni-stuttgart.de/cute/datenmaterial/.

9 Dabei wird jedes Zeichen des Wortes ersetzt durch seine Unicode-Kategorie (z. B. das Zeichen „a" durch ‚Letter lowercase', Ll). Danach werden nachfolgend gleiche Kategorien zusammengefasst. Aus der Zeichenfolge „Parzival" wird also ‚LuLl'.

10 http://cleartk.github.io/cleartk/

11 http://mallet.cs.umass.edu

art NER-Modell.[12] Die stark variierenden Ergebnisse für die verschiedenen Textkorpora sind u. a. dem unterschiedlich großen Umfang der Trainingsdaten (Tabelle 1) und der Verteilung der Entitätenklassen geschuldet. Die besten Ergebnisse werden bei einem häufigen Vorkommen einer Entitätenklasse im Trainingskorpus erzielt, beispielsweise bei Figuren (PER) in den mittelhochdeutschen Artusromanen oder bei Organisationen (ORG) in den Bundestagsdebatten. Insgesamt fällt auf, dass die *precision* des Entitätenreferenztaggers deutlich höher ist als der *recall*. Das bedeutet, dass der Tagger zwar gefundene Objekte korrekt klassifiziert, er aber die Mehrheit der relevanten Objekte nicht findet. Vor dem Hintergrund, einen zeitsparenden semi-automatischen Annotationsprozess anzustoßen, ist es sinnvoll, den *recall*-Wert zu verbessern (ggf. auch zulasten der *precision*), also möglichst viele Objekte in einem Text zu finden.

Für das mittelhochdeutsche Romankorpus wird unter Hinzunahme weiterer Features die Entwicklung des Entitätenreferenztaggers fortgeführt (Blessing et al. 2017). Durch die Implementierung eines mittelhochdeutschen Wortartenerkenners (Echelmeyer et al. 2017)[13], von Namenslisten und eines Features zur Groß-/Kleinschreibung wird eine deutliche Verbesserung der automatischen Vorhersagen erzielt. Des Weiteren werden verschiedene Evaluationssettings getestet und miteinander vergichen: Während im oben verwendeten engen Evaluationssetting eine Annotation nur dann als korrekt bewertet wird, wenn sowohl die Entitätenklasse (PER, LOC, ORG, WRK, CNC) als auch die Grenzen des Referenzausdrucks exakt übereinstimmen, gilt im weiten Evaluationssetting eine Annotation als korrekt, sobald die Entitätenklasse stimmt und sie mindestens ein Token Überlappung mit der Goldannotation aufweist.

Die Ergebnisse in Tabelle 4 zeigen, dass zum einen die Implementierung zusätzlicher Features zu erheblichen Verbesserungen der Entitätenreferenzerkennung führt und dass zum anderen viele Fehler auf inkorrekte Grenzen des Referenzausdrucks zurückgehen, was über das weite Evaluationssetting aufgefangen wird. Für das Untersuchungsvorhaben, die Vorkommen von Entitäten aus ei-

12 Anzumerken ist, dass das verwendete Stanford NER-Modell nur die Klassen PER, LOG und ORG umfasst und daher bei den Klassen WRK und CNC keine Entitäten findet. Des Weiteren ist er für eine Eigennamenerkennung entwickelt, so dass Entitätenreferenzen über Gattungsnamen nicht erfasst werden.

13 Unser *part-of-speech*-Modell für das Mittelhochdeutsche kann von der Webseite des Instituts für Maschinelle Sprachverarbeitung heruntergeladen werden: https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/pos_tag_mhg/; es ist auch im CLARIN Virtual Language Observatory (https://vlo.clarin.eu), einer Suchmaschine für linguistische Ressourcen und Werkzeuge, nachhaltig auffindbar. Eine Online-Version des Taggers steht zur Verfügung unter: http://clarin05.ims.uni-stuttgart.de/mhdtt/index.html.

Evaluation	*precision*	*recall*	F_1
eng	73,94	54,79	62,94
weit	94,85	73,82	83,02

Tab. 4: Evaluation des weiterentwickelten Entitätenreferenztaggers (ERT2). Getestet wurden PER-Referenzen auf Parzival.

nem Text zu extrahieren, scheint generell die weite Evaluation geeigneter zu sein, da – sofern eine Entitätenreferenz erfasst wurde – die exakten Grenzen des Ausdrucks vernachlässigbar sind. Für semantisch motivierte Untersuchungen kann hingegen die enge Evaluation, die die Erfassung der gesamten Nominalphrase evaluiert, sinnvoller sein. Prinzipiell sollte die Evaluation maschineller Verfahren immer vor dem Hintergrund des spezifischen Forschungsinteresses erfolgen und entsprechend reflektiert werden.

3 Multidisziplinäre Zugänge zu Entitäten in den Geistes- und Sozialwissenschaften

3.1 Sozialwissenschaftlicher Zugang

Politische Parteien, internationale Organisationen oder Institutionen sind seit jeher zentrale Analyseobjekte der empirischen sozialwissenschaftlichen Forschung. Sie werden spätestens seit dem *linguistic turn* (Rorty 1992) in den Sozialwissenschaften auch mehr und mehr mittels textanalytischer Methoden in zunächst kleinen und zunehmend größeren Mengen von Textdokumenten untersucht, beispielsweise in Parteiprogrammen, offiziellen Regierungsdokumenten oder Zeitungstexten. Innerhalb transnationaler europäischer Mediendebatten wird darüber hinaus die Sichtbarkeit und die semantische Bewertung verschiedener politischer Entitäten (z. B. der Europäischen Union) untersucht (z. B. Kantner 2015).

Sprachliche Bezüge auf unterschiedliche Entitätentypen sind also Bestandteil zahlreicher sozialwissenschaftlicher Forschungsinteressen und Fragestellungen. In unserem Projekt haben wir die Nennung verschiedener Entitätentypen als Bestandteil multipler kollektiver Identitäten im politischen Diskurs untersucht. Kollektive Identität verstehen wir als gemeinsames normatives Selbstverständnis, das von den Mitgliedern einer Wir-Gemeinschaft geteilt wird (Tietz 2002; Kantner 2004). In modernen Gesellschaften sind Individuen in ein komplexes Geflecht multipler Zugehörigkeiten eingebunden, die je nach sozialem oder politischem Kontext unterschiedliche Bedeutung gewinnen können (Kantner und Tietz 2013). Über den sprachlichen Bezug auf bestimmte Entitäten können Sprecher*innen

hierbei ihre Identifikation mit einem bestimmten Kollektiv zum Ausdruck bringen. Zwar sind nicht alle sprachlichen Bezüge auf Entitäten automatisch auch Äußerungen von kollektiver Zugehörigkeit, doch setzt die Äußerung kollektiver Identifikation oftmals einen sprachlichen Bezug auf Entitäten voraus. Im Sinne eines mehrstufigen Analyseprozesses, der sich der semantischen Komplexität des Konzepts kollektiver Identität schrittweise annähert, bildet die (semi-automatische) Annotation von Entitätenreferenzen daher eine sinnvolle Möglichkeit, große Textkorpora auf die für die Fragestellung relevanten Texte und Textpassagen einzugrenzen.

Wir haben für unsere Analyse das PolMine[14]-Korpus verwendet(Blätte und Blessing 2018). Hierbei handelt es sich um ein Textkorpus, das die Plenarprotokolle des Deutschen Bundestags im Zeitraum von 1996 bis 2016 enthält. Die Auswahl dieses Textmaterials ermöglichte es uns, Referenzen auf politische Entitäten innerhalb alltagssprachlicher Debatten zu analysieren. Um Trainingsdaten für die semi-automatische Annotation zu generieren, annotierten wir sämtliche Nennungen politischer Entitäten auf einer Teilmenge der transkribierten Plenarprotokolle. Annotiert wurden Referenzen auf Personen (PER), Organisationen (ORG), Orte (LOC) und Werke (WRK). Bei der Anpassung der universellen Annotationsrichtlinien an das ausgewählte Textmaterial haben wir weitere, ergänzende Zusatzrichtlinien erstellt. Dies erleichterte die Identifikation relevanter Textstellen und erhöhte insgesamt die Einheitlichkeit der Annotationsergebnisse.

Für die verschiedenen Klassen der fokussierten Entitäten enthielten die Zusatz-Guidelines allgemeine Definitionen und Beispiele, wie in konkreten Sätzen oder Passagen auf die verschiedenen Entitäten referiert werden kann. Organisationen wurden beispielsweise anhand von Referenzen wie „die Europäische Union“, „die deutsche Bundesregierung“ benannt. Auf Personen wurden über Ausdrücke wie „Frau Merkel“ oder „die deutsche Bundeskanzlerin“ referiert. Referenzen auf Orte fanden sich, oftmals gemeinsam mit lokalen Präpositionen, in Nennungen wie „Paris“, „Berlin“ oder „Brüssel“. Auf Werke wurde in Ausdrücken wie z. B. „die Kopenhagener Kriterien“ referiert.

(1) a. Die Deutschen zwingen den Griechen einen rigiden Sparkurs auf.
b. Berlin *muss* sich seiner internationalen Verantwortung stellen.

Probleme, die bei der Annotation auftraten, betrafen oftmals die Disambiguierung mehrdeutiger Textstellen. Organisationen wurden z. B. häufig über Referenzen auf konkrete Orte wie „Berlin“, „Brüssel“ oder „New York“ adressiert. Auch

14 Für weitere Informationen zum PolMine Projekt, siehe https://polmine.github.io/

Nennungen von Personen oder Gattungen von Personen referierten inhaltlich in manchen Fällen auf Organisationen, z. B. wenn „die Deutschen“ „den Griechen“ „einen rigiden Sparkurs“ aufzwangen (1a), war eigentlich eine Organisation gemeint (in diesem Fall die Bundesregierung). In vielen Fällen ließen sich solche Mehrdeutigkeiten über den jeweiligen Satzkontext disambiguieren. Organisationen wurden oftmals per Städte- oder Ländernamen adressiert, wenn diese mit Modalverben wie „muss“, „sollte“ oder „kann“ kookkurrierten (1b). Auf Orte wurde hingegen öfter über lokale Präpositionen wie „in“, „nach“ oder „aus“ referiert. Um die Einheitlichkeit der Annotationen zu erhöhen, wurden Entscheidungshilfen für rekurrierende ambivalente Fälle in die Zusatz-Guidelines aufgenommen.

In anderen Fällen kam es zu voneinander abweichenden Annotationen ein und derselben Textstelle, die auch im Rahmen der Adjudikation (Zusammenführung und Auflösung der Annotationsunterschiede) nur schwer aufzulösen waren. Im vorherigen Beispiel der ‚Deutschen‘ und ‚Griechen‘ war es tatsächlich schwierig, einheitlich zu bestimmen, ob nun dem griechischen Staat (ORG) oder vielmehr den Griechen als Gruppe von Personen (PER) ein rigider Sparkurs aufgezwungen wird. Neben der inhärenten Multireferentialität einzelner Terme, war die ‚Bestimmtheit‘ von Gattungsreferenzen eine weitere Quelle für abweichende Annotationen. Ob mit ‚den Deutschen‘, die eine ‚historische Verantwortung haben‘ eine bestimmte oder eine unbestimmte Gruppe von Personen gemeint ist, ist eine Interpretationsfrage. Um die Einheitlichkeit der Annotation zu erhöhen, wurden bei solch mehrdeutigen Textstellen möglichst generische Entscheidungen getroffen, die anschließend als ergänzende Guidelines in die Annotationsrichtlinien aufgenommen wurden.

3.2 Mediävistischer Zugang

Die mediävistische Fachgruppe untersucht Figuren und ihre Relationen in einem Korpus mittelhochdeutscher Artusromane (*Erec* und *Iwein* Hartmanns von Aue, *Parzival* Wolframs von Eschenbach)[15] (Ketschik in Vorbereitung). Figuren sind, allgemein gesprochen, „mit ihrer sinnkonstitutiven und handlungsprogressiven Funktion ein[] elementare[r] Baustein der fiktiven Welt“ (Platz-Waury 1997a, S. 587) und daher für viele Forschungsfragen grundlegend. In mediävistischen Kontexten wird vor allem die charakteristische Figurenzeichnung als „flache“ Fi-

15 Es werden folgende Editionen verwendet: Hartmann von Aue 1963, Hartmann von Aue 1968 und Wolfram von Eschenbach 1891.

guren[16] (Forster 1949, S. 77) oder „Handlungsträger“ (Propp 1972) diskutiert, derzufolge den Figuren nur eine „Minimalausstattung“ (Haferland 2013, S. 105–114) an Individualität und Innerlichkeit zugesprochen wird. Stattdessen werden Figuren stärker über ihr Handeln definiert, was mit der strukturalistischen Figurenkonzeption einhergeht, die ihre Funktionsgebundenheit betont (vgl. Schulz 2012, S. 12).

In den ausgewählten mittelhochdeutschen Artusromanen sollen die Figuren vor allem hinsichtlich ihrer Zentralität, ihrer Relationen und ihrer Bedeutung für die Handlung untersucht werden. Hintergrund ist ein Vergleich zwischen den Artusromanen Hartmanns von Aue (*Erec* und *Iwein*) und dem deutlich komplexeren Artusroman Wolframs von Eschenbach (*Parzival*). Mittels datenbasierter Verfahren und unter Rückgriff auf die Methode der sozialen Netzwerkanalyse[17] wird das Figureninventar der Erzählungen analysiert und ausgewertet (vgl. Braun und Ketschik 2019). Dabei soll eine dynamische Visualisierung der Netzwerke ermöglichen, die Figurenvorkommen und -relationen über den Verlauf der Handlung hinweg nachzuvollziehen. Auch der sukzessive Aufbau des Figureninventars und mögliche Veränderung in den Positionen einzelner Knoten über die Erzählzeit hinweg werden auf diese Weise sichtbar. Ergänzend werden netzwerkanalytische Metriken herangezogen, die u. a. Auskunft über die Größe und Verbundenheit der Netzwerke und die strukturelle Relevanz der Akteure geben, so dass die ausgewählten Texte quantitativ miteinander verglichen werden können.

Weiterführend soll auch das Zusammenspiel von Figuren und Räumen in den Blick genommen werden (vgl. hierzu auch Viehhauser 2020 ab Seite 373 in diesem Band). Der Handlungsraum ist in der mittelhochdeutschen Epik meist dichotom strukturiert und semantisiert (Lotman 1972) – so steht im höfischen Roman der kultivierte höfische Raum dem gefährlichen, unkultivierten Abenteuerraum konträr gegenüber. In den ausgewählten Artusromanen sollen die Räume zum einen in Bezug auf ihre handlungsstrukturierende Funktion und zum anderen bezüglich möglicher Wechselwirkungen mit der Kategorie der Figur untersucht werden. Hierbei interessiert, welche Figuren bestimmten semantischen Räumen fest zugeordnet sind, und welche Figuren in der Lage sind, die Raumgrenzen zu überschreiten.[18]

16 ‚Flache‘ Figuren verkörpern – im Gegensatz zu ‚runden‘ Figuren – nur ‚eine einzige Idee oder Eigenschaft‘. Neutralere Begriffe als ‚flach‘ und ‚rund‘ führen Lahn/Meister ein, die stattdessen von „einfachen“ und „komplexen“ Figuren sprechen (Lahn und Meister 2016, S. 239).

17 Zur sozialen Netzwerkanalyse vgl. allgemein Jansen (2003) sowie stärker literaturwissenschaftlich Trilcke (2013).

18 Ein wesentliches Strukturmerkmal von Räumen in Texten ist Lotman zufolge ihre unüberschreitbare Grenze, die sich nur für den Helden als überwindbar erweist (Lotman 1972, S. 341).

Zur Durchführung der Analysen ist es notwendig, die Vorkommen der Figuren sowie die Handlungsräume im Text adäquat zu erfassen, was über die Annotation der Entitätenreferenzen der Klassen PER und LOC erfolgt.

Für das Vorhaben, Figurenvorkommen und -relationen im Korpus ausgewählter mittelhochdeutscher Artusromane zu untersuchen, ist die Erfassung von Figurenreferenzausdrücken grundlegend. Nach Umsetzung des Untersuchungsvorhabens auf die oben beschriebene Weise können folgende Ergebnisse, aber auch Herausforderungen und Grenzen herausgestellt werden. Zunächst ist festzuhalten, dass die Erfassung von Figurenreferenzausdrücken (via Eigen- und Gattungsname) für die angestrebten Analysen grundlegend ist und aus den folgenden Gründen geeignet erscheint: Die Figurenvorkommen lediglich über Eigennamen zu erfassen, führt nach eingehender Analyse der Daten zu verzerrenden Ergebnissen, da die Figuren zum einen deutlich häufiger über Gattungsnamen erwähnt werden als über Eigennamen, und da zum anderen die Verteilung der Eigen- und Gattungsnamen je nach Figur unterschiedlich ausfällt. So werden die Protagonisten in der Regel häufig mit Gattungsnamen bezeichnet („der ritter“, „der helt“), während auf andere Figuren häufiger mit ihrem Eigennamen referiert wird. Die Berücksichtigung der Gattungsnamen, die mit den disziplinübergreifenden Richtlinien korreliert, ist für das mittelhochdeutsche Textkorpus somit essenziell. Dieser Umstand macht allerdings zur weiteren Nutzung der Daten einen Folgeschritt notwendig, der darin besteht, die Ausdrücke auf die Figur, auf die sie referieren, aufzulösen (*entity grounding*). Nur durch diesen zusätzlichen Schritt kann die notwendige Information, welche Figureninstanz genannt wird, mit der textuellen Oberflächenform verknüpft werden. Da ein Großteil der Appellativbezeichnungen mehrdeutig ist („die vrouwe“ kann je nach Textstelle auf verschiedene Figuren referieren), muss die Auflösung der Begriffe unter Berücksichtigung des jeweiligen Kontextes vorgenommen werden, was trotz technischer Unterstützung einen hohen Zeitaufwand erfordert.

Ferner ist anzumerken, dass – anders als die generischen Richtlinien es vorsehen – die Erfassung der maximalen Nominalphrase für die weiteren Analysen nicht notwendig ist. Um zu extrahieren, wann welche Figur im Text genannt wird, ist es lediglich wichtig, einen Teil des Referenzausdrucks (wenn möglich den syntaktischen Kopf) zu erfassen – vor- oder nachgestellte Adjektive, Determinanten, Appositionen oder andere Ergänzungen sind hingegen vernachlässigbar. Die Richtlinien wurden im Laufe des Annotationsprozesses also dahingehend angepasst, dass auch eine unvollständige Entitätenreferenz als ausreichend betrachtet wird. Lediglich bei verschachtelten Referenzausdrücken blieb es wichtig, die Grenzen korrekt zu erfassen, um später die eingebetteten Ausdrücke identifizieren zu können.

Da im Rahmen des Untersuchungsvorhabens große Textmengen zu bewältigen waren, wurde nach der anfänglichen manuellen Annotation ein maschinelles Lernverfahren angestoßen, um einen Entitätenreferenzerkenner für mittelhochdeutsche Texte zu trainieren (vgl. Abschnitt 2.2). Mithilfe des Entitätenreferenztaggers konnte der Annotationsprozess erheblich zeiteffizienter gestaltet werden, da sich die manuelle Arbeit im Anschluss darauf beschränkt, die Vorhersagen zu kontrollieren und eventuelle fehlende Referenzausdrücke nachzuannotieren.

Eine Herausforderung der (semi-)automatischen Annotation bestand in der Adaption des Modells an neue Texte, in denen andere Eigennamen und Appellative vorkommen oder ein anderer (autor-/gattungsspezifischer) Schreibstil vorliegt als in den Daten, mit denen ein Modell trainiert wurde. Um die Differenz zwischen Quell- und Zieldomäne zu überbrücken, wurde – sobald der Entitätenreferenzerkenner auf einen neuen Roman angewandt wurde – je ein Teil des Zieltextes von Grund auf annotiert und in die Trainingsdaten inkludiert. Dadurch musste zwar zunächst mehr manuelle Arbeit investiert werden, die manuelle Mehrarbeit konnte aber dem Absinken der Erkennungsrate entgegenwirken und somit den Annotationsprozess des gesamten Textes effizienter gestalten.

Auch der Schritt des *entity grounding* wurde semi-automatisch unterstützt, indem Vorhersagen darüber generiert wurden, auf welche Figur ein bestimmter Ausdruck an einer bestimmten Stelle referiert. Über eine spezielle Benutzeroberfläche im Annotationstool CRETAnno wurden sowohl der annotierte Referenzausdruck mit einem Kontextfenster als auch die sechs wahrscheinlichsten Figureninstanzen angezeigt, so dass aus dieser Liste nur noch die korrekte Instanz ausgewählt oder – sofern nicht enthalten – über eine zuschaltbare Figurenliste herausgesucht werden musste. Trotz der technischen Unterstützung handelte es sich um einen aufwändigen, für das Forschungsinteresse aber unverzichtbaren Arbeitsschritt.

Generell ist festzuhalten, dass die Umsetzung des angestrebten Untersuchungsvorhabens deutlich komplexer ist, als zuvor angedacht. Die Erfassung der Referenzausdrücke ist für das Vorhaben zwar essenziell, es ist aber eine Vielzahl weiterer Schritte notwendig. So muss neben dem bereits skizzierten *entity grounding* auch der Status der Entitäten berücksichtigt werden, um Figuren der fiktiven Welt von anderen Figuren oder Personen (z. B. von historischen Personen) zu unterscheiden. Ferner ist zu differenzieren, ob eine Entitätenreferenz in Figurenrede steht oder nicht, wofür alle Passagen direkter Rede identifiziert werden müssen. Und zuletzt müssen eingebettete Referenzausdrücke erkannt und herausgefiltert werden, da sie in der Regel nicht auf Figuren referieren, die an der entsprechenden Stelle des Textes agieren. Somit ist für eine adäquate Umsetzung des Untersuchungsvorhabens die reine Erfassung der Entitätenreferenzen nicht ausreichend. Die Integration der weiteren Schritte führte dazu, das ein komple-

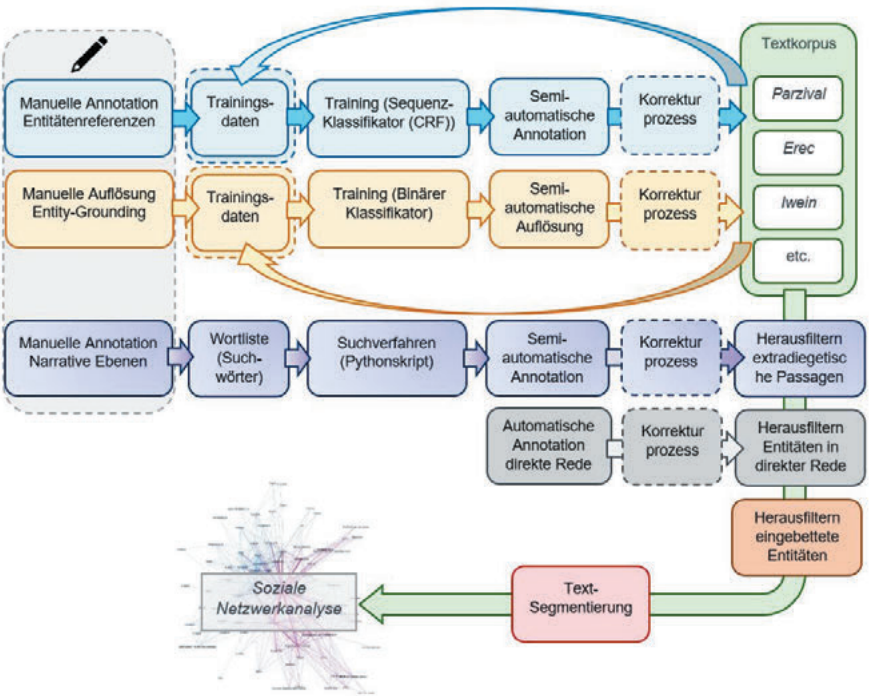


Abb. 6: Gesamtworkflow zur Sozialen Netzwerkanalyse mittelhochdeutscher Artusromane

xer Workflow entstand, der über den gemeinsamen methodischen Ansatz weit hinausgeht (vgl. Abbildung 6 und Ketschik in Vorbereitung).

Eine weitere Herausforderung bestand in der Annotation der Orte (LOC), da das Konzept des Ortes oder Handlungsraums in fiktionalen Texten unscharf definiert ist und weitere Unterscheidungskriterien, etwa ‚Ereignisregion' und ‚erwähnte räumliche Gegebenheit', benötigt werden (vgl. Barth und Viehhauser 2017 und Viehhauser 2020, ab Seite 373 in diesem Band). Zusätzliche Schwierigkeiten bestanden darin festzulegen, wie zum einen mit der Granularität von Räumen oder Raumobjekten umgegangen werden soll (der Ort Nantes ist eindeutig ein Handlungsraum, aber wie sieht es mit einem Zimmer in der Burg von Nantes oder mit einem Bett im Burgzimmer aus?) und wie zum anderen Objekte zu kategorisieren sind, die zur räumlichen Verortung herangezogen werden (ist z. B. ein Stuhl als Ortsreferenz zu zählen, wenn eine Figur darauf sitzt?). Und zuletzt kann in fiktionalen Texten potenziell jedes Objekt zum Raum werden (man denke an einen Flaschengeist, der die Flasche zum Lebensraum hat), so dass jeweils kontextspezifisch über den Status eines Objekts entschieden werden musste. Diese

Herausforderungen machten es notwendig, text- und korpusspezifische Richtlinien für den Umgang mit Raumobjekten einzuführen.

Für die folgenden Analysen von Figuren- und Raumkonstellationen wurde ferner festgestellt, dass die annotierten Raumausdrücke nicht zwangsweise Aufschluss über den Schauplatz der Handlung geben (wurde beispielsweise eine ‚Kemenate' annotiert, fehlte die Information, wo sich dieses Schlafgemach befindet). Für die Erfassung der Handlungsräume erwies es sich daher als geeigneter, eine Segmentierung des Textes in Raumsegmente vorzunehmen, bei der ganze Handlungsabschnitte einem bestimmten Raum zugeordnet wurden. Auf diese Weise konnte für die anschließenden Analysen die Information extrahiert werden, welche Figuren in welchen Handlungsräumen agieren.

3.3 Zugang der Neueren Deutschen Literatur

Für die Neuere Deutsche Literatur steht die Untersuchung und Charakterisierung von Figuren und deren Konstellation im Fokus, allen voran der Dreiecksbeziehung, die sich in Johann Wolfgang von Goethes maßgebendem Briefroman *Die Leiden des jungen Werthers* manifestiert und zu einem Bezugspunkt in der literarischen Rezeption des Werks wird.

Die Adaptationen, die in Referenz auf Goethes Briefroman entstanden sind, orientieren sich in unterschiedlicher Hinsicht am Originalwerk. So wurde der *Werther* seit der Veröffentlichung seiner ersten Fassung 1774 in verschiedenen Gattungen literarisch und kritisch verarbeitet, satirisch verfremdet, parodiert oder schlicht nachgeahmt. Insgesamt konnten bisher rund 140 deutsche Werke identifiziert werden,[19] (vgl. hierzu Appell 1882; Atkins 1949; Scherpe 1970; Richter 2017) die als ‚Wertheriaden' gelten.[20] Die Bezüge auf das Originalwerk fallen in den Adaptationen unterschiedlich stark aus und und können sowohl formaler, struktureller als auch inhaltlicher Art sein. (Martens 1985; Horré 1997). Die im Rahmen von CRETA vorgenommenen Analysen konzentrieren sich auf die

19 Goethes *Werther* erfuhr auch eine starke internationale Rezeption. 1776 wurde der Briefroman bereits ins Französische und 1779 ins Englische übersetzt. Zu Lebzeiten Goethes erschienen zudem Übersetzungen ins Russische, Italienische, Niederländische, Schwedische, Spanische, Dänische und Portugiesische (vgl. Goedeke und Goetze 1979). Es lassen sich auch fremdsprachige Wertheriaden nachweisen, wie unter anderem die 1785 anonym veröffentlichte englischsprachige Adaptation *Eleonora. From the Sorrows of Werther. A Tale*, in dem Werther und Lotte ihre Rollen tauschen, vgl. hierzu u.a. Richter 2017.

20 Die Anzahl der deutschen Wertheriaden setzt sich aus 33 Dramentexten, 29 Gedichten und 78 Prosatexten zusammen. Von den 78 Prosatexten werden in unserer Analyse 30 Texte, darunter Brief- und Tagebuchromane, Novellen und Erzählungen berücksichtigt.

Gattung der Prosatexte. Der Zugang zu diesem heterogenen Textkorpus erfolgt über die Entitätenklassen Figuren/Personen (PER), Orte (LOC) und (literarische) Werke (WRK). Die Annotation der Entitätenreferenzen beschränkt sich zunächst auf das Originalwerk, mit dem Vorhaben, die Annotation auf die überarbeitete Fassung des *Werthers* von 1787 sowie weitere Wertheriaden auszuweiten, um auch hierin Figuren, Orte und intertextuelle Verweise zu erfassen und so eine Vergleichbarkeit der Werke zu ermöglichen. Vorrangiges Ziel ist die Erfassung von Figurenkonstellationen sowie die Untersuchung und die Beantwortung der Frage, in welchen Wertheriaden sich die charakteristische Dreiecksbeziehung ebenfalls identifizieren lässt. Zur Analyse werden Netzwerke des Originaltexts und der Werther-Adaptationen generiert.

Die Analyse der Figurenkonstellationen in literarisch-fiktionalen Texten ermöglicht es, die dynamische Struktur der Interaktion zwischen den Figuren zu erfassen. Die jeweiligen Konstellationen übernehmen dabei nicht nur eine „handlungsgestaltende Funktion", die Figuren werden durch die „Kontrast- und Korrespondenzrelationen" (Platz-Waury 1997b, S. 591) zudem auch strukturiert und charakterisiert (Pfister 2001, S. 232–235). Dadurch bilden bestimmte Figurenkonstellationen die Grundlage für Konfliktstrukturen, etwa die Gegenüberstellung eines Protagonisten mit einem Antagonisten. In Goethes *Werther* ist diese ‚klassische' Konstellation um eine weitere Figur, Lotte, erweitert. Diese bildet neben dem emotionalen Protagonisten Werther, der sich in sie verliebt, und ihrem Verlobten Albert, der den Widersacher Werthers darstellt, eine Grenzgängerfigur. Viele Wertheriaden greifen diese tragische Liebesgeschichte auf und stellen sie in den Fokus, um eine deutliche Referenz auf den Originaltext zu erzeugen.

Der komplexe Aufbau von Goethes Briefroman birgt Herausforderungen für die Annotation der Entitätenklassen, insbesondere der Figuren (PER), und hat dadurch in unterschiedlicher Hinsicht zur Schärfung dieser Klasse beigetragen.

Der Roman beginnt mit den mahnenden Worten des fiktiven Herausgebers:

> Was ich von der Geschichte des armen Werther nur habe auffinden können, habe ich mit Fleiß gesammelt [...] Und du gute Seele, die du eben den Drang fühlst wie er, schöpfe Trost aus seinem Leiden, und laß das Büchlein deinen Freund sein, wenn du aus Geschick oder eigener Schuld keinen nähern finden kannst. (Goethe 1899, S. 4)

Diese Worte dienen nicht nur der Rahmung der folgenden Briefe Werthers, sondern werfen auch erste Fragen an die Richtlinien auf: Werden Ansprachen an fiktive Adressaten („gute Seele") als PER annotiert?

In Hinblick auf den Status der Entitätenreferenz wird in den CRETA-Richtlinien zunächst nicht unterschieden, ob solche Ansprachen innerhalb der Figurenrede, einer Protokollanmerkung, der Exegesis, Diegesis oder Metadiegesis erwähnt wer-

den. Eine entsprechende Bestimmung wird erst nach Abschluss der Vollannotation und im Kontext der darauf aufbauenden Analyse getroffen. Für die Annotation ist jedoch entscheidend, ob sich aus dem Zusammenhang entscheiden lässt, ob sich eine Entitätenreferenz auf eine der Figuren im Text bezieht oder nicht. Mit der „gute[n] Seele“[21] (vgl. Lahn und Meister 2016, S. 17) wird keine der Figuren im Text adressiert, weshalb diese Ansprache folglich unannotiert bleibt.

Die Diegese im Roman bilden vorrangig 81 Briefe, die Werther an seinen Freund Wilhelm richtet und in denen er von seinem abgeschiedenen Leben in Walheim und seiner aufkeimenden Liebe zur Amtstochter Lotte berichtet. Antwortbriefe seines Briefpartners sind nicht enthalten, wodurch eine autodiegetische Erzählweise und Monoperspektive auf die Geschehnisse entsteht.

Der Herausgeber führt die Geschichte Werthers zu Ende, nachdem sich dieser durch einen Schuss aus der Waffe seines Gegenspielers Albert das Leben genommen hat. Dieser Aufbau macht sich bei der Annotation der Figuren dahingehend bemerkbar, dass ‚Werther‘ nur selten erwähnt wird, da er sich als Erzähler in seinen Briefen nicht selbst beim Namen nennt. Eine Nennung häuft sich erst in den abschließenden Passagen aus dem Teil „Der Herausgeber an den Leser“ oder findet sich in den indirekt wiedergegebenen oder zitierten Aussagen anderer Figuren. Für die Generierung eines Netzwerkes und die Identifikation der charakteristischen Dreiecksbeziehung hat dies zur Folge, dass als heuristischer Ansatz zunächst alle Vorkommen des Pronomens ‚Ich‘ mit Werther synonym gesetzt werden müssen (Barth und Murr 2017). Eine exakte Zuordnung des Pronomens ermöglicht erst eine Auflösung der Koreferenzen.

Bei der Generierung des Netzwerkes gilt es zudem zu berücksichtigen, dass manchen Figuren im *Werther* ein anderer Status zukommt, da sie einer unterschiedlichen fiktiven Welt angehören. Denn es sind nicht nur die Naturbeschreibungen, die das Seelenleben des Protagonisten widerspiegeln, sondern auch die intertextuellen Verweise (WRK). So liest Werther, wenn es ihm gut geht, in seinem *Homer*, erkennt in dem gemeinsamen Ausruf „Klopstock“ am Ball die scheinbare Seelenverwandtschaft zu Lotte und seine Ausweglosigkeit nach seinem Selbstmord wird durch das aufgeschlagene Drama Lessings *Emilia Galotti* verdeutlicht. Bevor Werther diesen drastischen Schritt wagt, liest er am Vorweihnachtsabend Lotte aus den Gesängen Ossians vor. Diese insgesamt drei Passagen (Colma, Ryno und Alpin) sind gespickt mit Figurennennungen, deren Häufung auch in der Visualisierung und dem Vergleich der (unterschiedlichen) Annotationen deutlich

21 Bei der „gute[n] Seele“, die vom Herausgeber angesprochen wird, handelt es sich um den „‚Hörer‘ [bzw. hier Leser] der herausgegebenen Briefe und der nachgeordneten Erzählung des fiktiven Herausgebers, an den dieser sich direkt wendet.“

hervortritt (vgl. hierzu auch den Beitrag von Baumann et al. (2020) ab Seite 270 in diesem Band). Bei der Analyse des Netzwerkes des gesamten Textes gilt es somit, eine Unterscheidung von aktiv präsenten Figuren der Diegese und Figuren aus anderen fiktiven Welten zu treffen. Bezüglich der Annotation der aktiv präsenten Figuren kommt hinzu, dass der Erzähler Werther in seinen Briefen nicht alle Figuren ausschließlich bei ihrem Namen nennt, sondern in diesen Fällen u. a. zu Kosenamen übergeht. Umso stärker sich beispielsweise das Verhältnis von Werther und Lotte intensiviert, desto öfter referiert Werther mittels Gattungsnamen auf sie und nennt sie einen „Engel des Himmels" und sieht in ihr den Lebenssinn „O der Engel! Um deinetwillen muss ich leben!" usf. Diese spezifischen Referenzausdrücke auf einzelne Figuren können ausschließlich aus dem Kontext erschlossen werden.

Im Hinblick auf viele Schauplätze und Figuren finden sich teilweise nur vage oder zensierte Angaben („mit dem Gesandten nach *** gehen soll") im Roman, was unter anderem dazu beitrug, dass er in der zeitgenössischen Rezeption als Bezugnahme auf persönliche Erlebnisse des Autors (den Selbstmord Karl Wilhelm Jerusalems und seine Beziehung zur bereits verlobten Charlotte Buff) (Vorderstemann 2007, S. 53, 71) hin gelesen wurde. Für die Annotation hat dieser Kunstgriff Goethes zur Folge, dass sich solche Namensnennungen auf zwei unterschiedliche Entitäten beziehen können: „Graf von M.", „die übergnädige Dame von S.". Entsprechend erfolgt hier eine Doppelannotation, um der Verschachtelung von PER und LOC gerecht zu werden. Bei der Annotation der Orte (LOC) kamen ähnliche Fragen wie im mediävistischen Arbeitsbereich auf. Ausgehend von der Intention, durch die Annotation der Orte eine Abfolge dieser sowie eine Gegenüberstellung von Natur- und Stadtraum sichtbar machen zu können, war ebenfalls eine korpusspezifische Eingrenzung notwendig. Sie sollte zeigen, ob ausschließlich Handlungsräume („Lieblingsplätzgen", „fruchtbare Tal", „Wirtshaus") und Städtenennungen („Walheim") erfasst werden oder ob es notwendig ist, feingranularer zu annotieren, um damit auch Räume und Objekte wie den Tisch oder das Bett Werthers zu erfassen. Als sinnvoll hat sich erwiesen, Objekte nur dann zu annotieren, wenn sich Werther an genau diesem einem ‚Brunnen' wiederholt aufhält und er und Lotte in dieser ‚Kutsche' zum Ball fahren. Weitere, unbestimmte Nennungen solcher Objekte blieben bei der Annotation unberücksichtigt.

In der Annotationspraxis hat sich darüber hinaus gezeigt, dass die Erfassung von maximalen Nominalphrasen beim *Werther* eine deutliche Herausforderung darstellt. Der Briefroman ist durch einen expressiven Schreibstil gekennzeichnet, der sich sowohl in einem emotionalen Vokabular manifestiert als auch in der gewollt expressiven Art zu schreiben, die durch unvollständige und verschachtelte Sätze, Parenthesen, Ellipsen und Ausrufe die Emotionalität und Unmittelbarkeit des Geschriebenen unterstreicht (Martens 1985, S. 96). Aus diesem Grund sind

auch Relativsätze wie „Sie ist schon vergeben, antwortete jene, an [einen sehr braven Mann, der weggereist ist, seine Sachen in Ordnung zu bringen nach seines Vaters Tod, und sich um eine ansehnliche Versorgung zu bewerben]“ keine Besonderheit, bergen jedoch ein hohes Fehlerpotenzial, wenn es darum geht, die exakten Grenzen der maximalen Nominalphrase zu bestimmen (vgl. Abschnitt 2.3).

3.4 Philosophischer Zugang

Die Auseinandersetzung mit dem philosophischen Textkorpus im Rahmen von CRETA verfolgte zwei Ziele:

1. die textkritische Edition ausgewählter Überlieferungsträger von Adornos *Ästhetischer Theorie* (vgl. Endres et al. 2013 und Adorno o.D.);
2. die computergestützte Textanalyse ausgewählter philosophischer Texte von Adorno.

Zu Förderbeginn von CRETA stand noch nicht fest, welche Fragen an das philosophische Textkorpus gestellt werden sollten. Im Zuge der interdisziplinären Arbeit an einem gemeinsamen Annotationsworkflow bestätigte sich jedoch sehr schnell die Vermutung, dass sich die Philosophie von den anderen beteiligten Disziplinen sowohl in ihrem Umgang mit Entitäten als auch in ihren Textumgangsformen, d. h. den disziplimmanent üblichen Formen der Lektüre und Interpretation von philosophischen Texten, unterscheidet. Der philosophische Umgang mit Entitäten ist am Anfang dieses Aufsatzes bereits kurz skizziert worden. Die Textumgangsformen in der Philosophie unterscheiden sich je nach philosophischem Ansatz. Fast alle dieser Ansätze kennzeichnen sich jedoch gegenwärtig dadurch, dass der Text als für sich stehende und inhaltlich zusammenhängende Folge von Aussagen kaum eine Rolle spielt.[22] Paradigmatisch zeigt sich dies an demjenigen philosophischen Ansatz, der die Philosophie die letzten Jahrzehnte dominierte: der *analytischen Philosophie*. Deren Grundannhamen werden exemplarisch in Jay F. Rosenbergs Einführung in die Philosophie artikuliert.[23] Rosenberg kennzeichnet die Philosophie als eine Disziplin „zweiter Ordnung“, welche Tätigkei-

22 Eine kritische Auseinandersetzung mit derartigen Zugängen sowie etwaige Alternativen bieten Endres et al. 2017.

23 Die deutsche Übersetzung von Rosenbergs 1984 erstmals auf Englisch publizierter Einleitung liegt mittlerweile in der sechsten Auflage vor (Rosenberg 2009). Die Verifikation ihres mutmaßlich paradigmatischen Charakters stellt selbst ein potenzielles DH-Forschungsprojekt dar: Eine Möglichkeit, zu bestimmen, welches Verständnis von Philosophie gegenwärtig dominiert, be-

ten „erster Ordnung“, wie zum Beispiel Kunst, Literatur, Physik oder Psychologie untersucht. Dabei orientiere sie sich an zwei Gruppen von Fragen: Bedeutungsfragen und Rechtfertigungsfragen. Erstere widmen sich insbesondere der Klärung von Verständnisschwierigkeiten, die sich in Hinblick auf die Behauptungen von Disziplinen ‚erster Ordnung‘ ergeben können. Letztere untersuchen deren Begründungsmodi. Aus den beiden Fragen und ihren Gegenstandsbereichen lassen sich in Anknüpfung an Rosenberg zwei Praktiken ableiten, welche die Philosophie dominieren: Argumentations- und (Begriffs-)Analyse .

Ein derartiges Philosophieverständnis zeitigt weitreichende Konsequenzen für den Umgang mit philosophischen Texten. Dieser wird sich auf jene Textmerkmale konzentrieren, in welchen sich die argumentative Struktur und der terminologische Gehalt dieser Texte manifestieren. Für den Umgang mit *Entitätenreferenzen* in philosophischen Texten folgt daraus, dass eine Textanalyse im Sinne von Rosenbergs Philosophieverständnis primär nur an einer der fünf im Rahmen von CRETA entwickelten Entitätenkategorien interessiert sein wird: den abstrakten Konzepten (CNCs). Was ein abstraktes Konzept bzw. ein Begriff ist, ist in der Philosophie jedoch genauso schulabhängig wie das Philosophieverständnis selbst. Daraus folgt, dass, bevor eine textanalytische Fragestellung in Hinblick auf das philosophische Textkorpus entwickelt bzw. präzisiert werden kann, zu klären ist, welcher Begriff des Begriffes dieser zugrundegelegt wird.

Was bedeutete dieser Sachverhalt für den Umgang mit Entitäten und deren Referenzen im Zuge der Auseinandersetzung mit dem philosophischen Textkorpus im Rahmen von CRETA? Obwohl für diese primär die abstrakten Konzepte von Interesse waren, wurde aus den in diesem Aufsatz genannten arbeitsdynamischen und explorativen Gründen der Arbeitsablauf im Umgang mit Entitäten und deren Referenzen aufrechterhalten, d. h., es wurden die Referenzen auf sämtliche der im Zuge von CRETA entwickelten Entitätenkategorien annotiert.[24] Da jedoch in Adornos Werk die Frage nach dem philosophischen Umgang mit Begriffen selbst eine zentrale Rolle spielt, wurden letztendlich in Anknüpfung an die aktuelle Adornoforschung (vgl. zum Beispiel Hogh 2015) – Adornos Selbstverständnis und dem Fokus der Hauptströmungen der Gegenwartsphilosophie entsprechend – die Be-

steht darin, diejenigen (Lehr-)Bücher zu sammeln, die gegenwärtig in Einführungsveranstaltungen der Philosophie verwendet werden, um zu untersuchen, welches Verständnis von Philosophie in diesen propagiert wird.

24 Dabei bestätigte sich die naheliegende Vermutung, dass auch zwei weitere Entitätenkategorien – Werke und Eigennamen – für historisch-hermeneutische Fragestellungen von Bedeutung sind: So ließe sich über deren Verhältnis zu den abstrakten Konzepten zum Beispiel eruieren, welche Werke welcher Autoren welche Relevanz in den ästhetischen Theorien Adornos und anderer Philosophen besitzen.

griffe ins Zentrum der konkreten Textanalyse gestellt (vgl. hierzu den Beitrag von Pichler, Blessing et al. 2020 ab Seite 328 in diesem Band).

Mit diesem Fokus wurde jene Frage akut, die bereits am Anfang der manuellen Annotation aufgekommen war: Wie verhält sich die in CRETA etablierte Entitätenkategorie ‚abstrakte Konzepte' (CNC) zu den Begriffskonzeptionen der Philosophie? Die Beantwortung dieser Frage setzte definitorische Festlegungen voraus, da einerseits in CRETA die Kategorie ‚abstrakte Konzepte' bewusst vage bestimmt wurde – in den Richtlinien ist von „Referenzen auf Konzepte, die für die Analyse wichtig sind", die Rede – andererseits in der Philosophie selbst – wie bereits angemerkt – keine Einigkeit darüber herrscht, was einen Begriff kennzeichnet. Festzulegen war also erstens, welche Begriffe als für die Analyse relevant erachtet werden, was wiederum, zweitens, eine Klärung dessen voraussetzte, was im Folgenden als Begriff erachtet werden sollte, und wie – drittens – auf diese in den Texten referiert wird.

Die Bestimmung dessen, was bei der Analyse des philosophischen Textkorpus als Begriff verstanden werden sollte, orientierte sich an der im Zuge des Projektes präzisierten disziplinspezifischen Fragestellung. Forschungsziel war es, zu überprüfen, ob Adorno Begriffe seinen Selbstbeschreibungen entsprechend verknüpft und ob sich diese Verknüpfungsform von derjenigen anderer Philosophieschulen unterscheidet. Aus dieser Fragestellung resultierten zwei Anforderung an das zu etablierende Begriffsverständnis: Erstens sollten die Begriffe eine Entitätenkategorie bilden, deren Referenzen sich ohne eine allzu große Anzahl an Zusatzbestimmungen relativ einfach, wenn möglich sogar automatisch, an der Textoberfläche festmachen und annotieren ließen. Zweitens galt es, eine Konzeptualisierung von ‚Begriff' zu entwickeln, die keine der Begriffskonzeptionen, welche die in das Textkorpus aufgenommenen Philosophien vertraten, bevorzugte. Nur so konnte vermieden werden, dass die Beantwortung der leitenden Fragestellung in eine zirkuläre Argumentation mündete und damit als typisches Resultat eines *confirmation bias* erachtet werden würde.

Bei der Suche nach einer diesen Voraussetzungen entsprechenden Begriffskonzeption zeigte sich, dass eine Übernahme des in der sprachanalytisch orientierten Gegenwartsphilosophie dominierenden Verständnisses von Begriffen nicht nur eine umfangreiche Ergänzung der in CRETA entwickelten Annotationsrichtlinien von Nöten gemacht hätte, sondern auch von Vorannahmen geprägt ist, gegen die sich Adorno mit seinem Denken wandte.[25] Die sprachanalytisch orientierte Philosophie unterscheidet – grob gesagt – zwischen singulären Begrif-

25 Adorno wandte sich insbesondere gegen den vermeintlichen ‚Zwangscharakter' eines an der formalen Logik orientierten Denkens, vgl. zum Beispiel Adorno 2003.

fen, die ein konkretes Einzelding bezeichnen (z. B. Eigennamen), und generellen Begriffen bzw. Termini, die durch ihre Anwendung auf singuläre Begriffe letztere bestimmen, indem sie atomare Sätze bilden (z. B.: „Sokrates ist ein Philosoph"). Generelle Termini werden daher auch als Prädikate bezeichnet. Bezüglich derselben können Eigenschaftsbegriffe (z. B.: „x ist grün") von Artbegriffen (z. B.: „x ist ein Philosoph") unterschieden werden. Generelle Termini können sowohl durch Nomen und Adjektive als auch durch Verben sprachlich realisiert werden (z. B.: „x liebt y"). Dieser Sachverhalt sowie die Tatsache, dass es sich bei generellen Termini um Funktionsausdrücke handelt, für die das Relationsverhältnis von zentraler Bedeutung ist, erlauben zwar deren formale Transkription, erschweren jedoch zugleich ihre Auszeichnung auf der Textoberfläche. Letzteres hätte, wie bereits angemerkt, eine Erweiterung der CRETA-Annotationsrichtlinen notwendig gemacht. Abgesehen von diesen pragmatischen Gründen erschien es nicht sinnvoll, für ein Begriffsverständnis und die damit einhergehende Form der Begriffsverknüpfung, für die bereits eine etablierte Formalisierungspraxis existiert – die Prädikatenlogik –, eine alternative Form zu entwickeln.

In Anbetracht dieser Gründe wurde letztendlich auf ein Begriffsverständnis zurückgegriffen, das sich nahe am alltäglichen Verständnis von ‚Begriff' bewegt und das zudem problemlos in die CRETA-Entitätenkonzeption integriert werden konnte: Als Referenzen auf Begriffe wurden sämtliche Nomina erachtet, bei denen es sich nicht um eine der vier anderen in CRETA entwickelten Entitätenkategorien handelte.

Abschließend ist festzuhalten, dass sich im philosophischen Arbeitsbereich disziplinspezifische Vorannahmen und damit einhergehende Textumgangsformen weitaus stärker auf die Arbeit mit den in CRETA entwickelten Entitätenklassen auswirkten als bei den anderen hier vorgestellten Arbeiten. Diese Vorannahmen sowie die im Zuge des Projektes entwickelte Fragestellung führten letztendlich dazu, dass nur eine der fünf CRETA-Entitätenklassen als heuristisch relevant erschien – und auch das erst, nachdem diese Klasse, die abstrakten Konzepte (CNC), der Disziplin entsprechend präzisiert wurde. Diese Präzisierung und die mit ihr einhergehende Begriffsarbeit führte zu einer Selbstreflexion zentraler Vorannahmen der Auseinandersetzung mit philosophischen Texten, wie sie bei der Interpretation philosophischer Texte nicht immer üblich ist, und generierte so einen zusätzlichen Erkenntnisgewinn für das Projekt. Dieser Mehrwert ist eine Folge des interdisziplinären Dialogs sowie der interdisziplinären Arbeit an einem gemeinsamen Workflow, ohne die die zu ihm führenden Reflexionen wahrscheinlich gar nicht erst angestoßen worden wären.

4 Schlussbetrachtung

Der in CRETA verfolgte und in diesem Artikel beschriebene Ansatz ging von der Annahme aus, dass mit der Annotation von Entitäten ein Konzept in den Blick genommen wird, das für alle beteiligten Disziplinen von Relevanz und für disziplinspezifische Fragestellungen anschlussfähig ist. Dabei standen zwei Ziele im Fokus: Erstens sollte durch den interdisziplinären Austausch eine methodische Umsetzung entwickelt werden, die im Gesamten oder zumindest in Teilen disziplinenunabhängig einsetzbar ist und so dazu beiträgt, eine disziplinenunabhängige Begrifflichkeit für CRETA zu entwickeln. Zweitens sollte die Auseinandersetzung mit Entitäten und ihren Referenzen zur Entwicklung und Schärfung der disziplinären Fragestellungen beitragen. Beide Zielsetzungen werden hier abschließend evaluiert und reflektiert.

In Hinblick auf die einheitliche methodische Umsetzung lässt sich zunächst festhalten, dass die enge Zusammenarbeit es ermöglicht hat, sowohl unterschiedliche Entitätenklassen disziplinübergreifend in Annotationsrichtlinien zu definieren als auch einen gemeinsamen Arbeitsablauf für die manuelle und semiautomatische Annotation zu entwickeln. Dabei ist allerdings anzumerken, dass die Annotationsrichtlinien zunächst wiederholt auf die verschiedenen Textkorpora angewendet werden mussten, um text- bzw. disziplinspezifische Konkretisierungen oder Anpassungen vorzunehmen. Die notwendigen Ergänzungen reichten von der Hinzunahme weiterer Referenzausdrücke (etwa Pronomina im *Werther*-Korpus) über die Berücksichtigung zusätzlicher Entscheidungskriterien (z. B. die Unterscheidung von fiktiver Figur und historischer Person) bis hin zur Implementierung weiterer umfassender Arbeitsschritte (*entity grounding*). Die zunächst tendenziell generischen Annotationsrichtlinien waren jedoch grundlegend für die disziplinären Spezifikationen und erleichterten diese erheblich. Die bei der Entitätenannotation entwickelten Arbeitsabläufe erwiesen sich somit für alle Disziplinen als effektiv und bestätigten somit zentrale methodische Grundannahmen von CRETA (vgl. auch den Beitrag von Pichler und Reiter 2020 ab Seite 43 in diesem Band).

Die wiederholte Auseinandersetzung mit dem Konzept der Entität und ihrer Referenz trug darüber hinaus zu einem tieferen Verständnis innerhalb der einzelnen Disziplinen bei. Dabei erwies sich insbesondere der Prozess der Parallelannotation als fruchtbar, den disziplininternen Dialog zu fördern, Vorannahmen offenzulegen und sie dadurch zur Diskussion zur stellen. Zudem ermöglichte die unmittelbare Anwendung der Annotationsrichtlinien auf die jeweiligen Textkorpora, die fachspezifischen Fragestellungen so zu präzisieren, dass eine Operationalisierung und Beantwortung dieser im Zuge einer reflektierten algorithmischen

Textanalyse möglich war. Der semi-automatische Annotationsprozess (vgl. Abbildung 4) ist zwar (bisher) nur für das mediävistische Projekt zum Einsatz gekommen, er wurde in diesem Kontext jedoch bereits auf verschiedene Texte und Textgattungen angewendet, was nahelegt, dass er auch ohne Weiteres auf Texte anderer Disziplinen übertragen werden kann.

In summa profitierten sämtliche Teilnehmenden von den Synergieeffekten des interdisziplinären Dialogs (vgl. hierzu auch den Beitrag von Reiter, Kremer et al. (2020), ab Seite 467 in diesem Band). Zwar wurde im Laufe des Projektes deutlich, dass alle Fachbereiche von disziplinspezifischen Vorannahmen geprägt waren, die einen bestimmten Umgang mit den Untersuchungsgegenständen nahelegen und zum Verständnis aller wiederholt diskutiert wurden. Dieser vermeintliche Nachteil schlug jedoch in einen Vorteil um, da die interdisziplinäre Kooperation die Teilnehmenden von Anfang an dazu verpflichtete, besagte Vorannahmen offenzulegen, wodurch sich ein gegenseitiges Verständnis bilden konnte. Denn nur unter Berücksichtigung derselben konnte die Entwicklung eines interdisziplinären Entitätenkonzeptes und diesem entsprechender Entitätenklassen sowie eines damit einhergehenden Annotationsworkflows realisiert werden. Diese Form des Austauschs trug letztendlich sowohl zur Etablierung neuer Arbeitspraktiken und Methoden als auch zu einem vertieften Selbstverständnis der teilnehmenden Disziplinen bei. Insofern hat sich das Vorgehen für alle Disziplinen als fruchtbar erwiesen und eignet sich darüber hinaus als methodische Grundlage für zukünftige Kooperationen an der Schnittstelle von Sozial- und Geisteswissenschaften einerseits und Computerlinguistik andererseits.

Danksagung: Wir bedanken uns bei Florian Barth, Dilan Cakir, Anaïck Geissel, Annika Holzer, Alina Palesch, Fabian Schan, Thomas Frank und Jan Velimsky für ihre wertvolle Hilfe bei der manuellen Annotation.

Primärliteratur

Goethe, Johann Wolfgang (1899). *Die Leiden des jungen Werther*. I. A. d. Großherzogin Sophie von Sachsen. Weimar: Hermann Böhlaus Nachfolger.

Hartmann von Aue (1963). *Erec*. Hrsg. von Albert Leitzmann. 3. Aufl. Tübingen: Niemeyer.

Hartmann von Aue (1968). *Iwein*. Hrsg. von Georg F. Benecke, Karl Lachmann und Ludwig Wolff. 7. Aufl. Berlin: de Gruyter.

Wolfram von Eschenbach (1891). *Parzival*. Hrsg. von Karl Lachmann. 5. Aufl. Berlin.

Sekundärliteratur

Adorno, Theodor W. (2003). „Einleitung zum Positivismusstreit in der deutschen Soziologie". In: Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 8. Gesammelte Schriften. Frankfurt am Main: Suhrkamp, S. 280–353.

Adorno, Theodor W. (o.D.). *Schein - Form - Subjekt - Prozeßcharakter - Kunstwerk. Textkritische Edition der letzten bekannten Überabreitung des III. Kapitels der 'Kapitel-Ästhetik'*. Hrsg. von Martin Endres, Axel Pichler und Claus Zittel.

Appell, Johann Wilhelm (1882). *Werther und seine Zeit*. Oldenburg: Schulzesche Hofbuchhandlung.

Atkins, Stuart P. (1949). *The testament of Werther in poetry and drama*. Cambridge, Massachusetts: Harvard University Press.

Barth, Florian und Sandra Murr (2017). „Digital Analysis of the Literary Reception of J.W. von Goethe's Die Leiden des jungen Werthers". In: *Digital Humanities 2017: Conference Abstracts*. Montreal, S. 540–542.

Barth, Florian und Gabriel Viehhauser (2017). „Digitale Modellierung literarischen Raum". In: *Abstracts der DHd: Digitale Nachhaltigkeit*. Bern: Digital Humanities im deutschsprachigen Raum e.V., S. 128–132.

Baumann, Martin, Steffen Koch, Markus John und Thomas Ertl (2020). „Interactive Visualization for Reflected Text Analytics". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 270–296.

Bethard, Steven, Philip Ogren und Lee Becker (2014). „ClearTK 2.0: Design Patterns for Machine Learning in UIMA". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), S. 3289–3293.

Blätte, Andreas und André Blessing (2018). „The GermaParl Corpus of Parliamentary Protocols". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, S. 810–816.

Blessing, André, Nora Echelmeyer, Markus John und Nils Reiter (2017). „An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver, Canada: Association for Computational Linguistics, S. 57–67. DOI: 10.18653/v1/W17-2208.

Braun, Manuel und Nora Ketschik (2019). „Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen". In: *Das Mittelalter* 24, S. 54–70. DOI: 10.1515/mial-2019-0005.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde und Hagen Langer, Hrsg. (2010). *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Heidelberg: Spektrum. DOI: 10.1007/978-3-8274-2224-8.

Clark, Alexander, Chris Fox und Shalom Lappin, Hrsg. (2013). *The handbook of computational linguistics and natural language processing*. Chichester: Wiley-Blackwell.

Cohen, Jacob (1960). „A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, S. 37–46.

Echelmeyer, Nora, Nils Reiter und Sarah Schulz (2017). „Ein PoS-Tagger für ‚das' Mittelhochdeutsche". In: *Abstracts der DHd: Digitale Nachhaltigkeit*. Bern: Digital Humanities im deutschsprachigen Raum e.V., S. 141–147.

Endres, Martin, Axel Pichler und Claus Zittel (2013). „‚Noch offen'. Prolegomena zu einer Textkritischen Edition der *Ästhetischen Theorie* Adornos". In: *editio* 27, S. 173–204.

Endres, Martin, Axel Pichler und Claus Zittel, Hrsg. (2017). *Textologie. Theorie und Praxis interdisziplinärer Textforschung*. Berlin: De Guyter.

Faruqui, Manaal und Sebastian Padó (2010). „Training and Evaluating a German Named Entity Recognizer with Semantic Generalization". In: *Proceedings of KONVENS*, S. 129–133.

Finkel, Jenny Rose, Trond Grenager und Christopher Manning (2005). „Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, USA: Association for Computational Linguistics, S. 363–370. DOI: 10.3115/1219840.1219885.

Forster, Edward Morgan (1949). *Ansichten des Romans*. Übers. von Walter Schürenberg. Frankfurt am Main: Suhrkamp.

Gius, Evelyn und Janina Jacke (2017). „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis". In: *International Journal of Humanities and Arts Computing* 11.2, S. 233–254.

Goedeke, Karl und Edmund Goetze (1979). *Grundriss zur Geschichte der deutschen Dichtung aus den Quellen*. Nendeln, Lichtenstein: Kraus Reprint.

Haferland, Harald (2013). „Psychologie und Psychologisierung: Thesen zur Konstitution und Rezeption von Figuren. Mit einem Blick auf ihre historische Differenz". In: *Erzähllogiken in der Literatur des Mittelalters und der Frühen Neuzeit. Akten der Heidelberger Tagung vom 17. bis 19. Februar 2011*. Hrsg. von Florian Kragl und Christian Schneider. Heidelberg: Winter, S. 91–117.

Hogh, Philip (2015). *Kommunikation und Ausdruck: Sprachphilosophie nach Adorno*. Weilerswist: Velbrück.

Horré, Thomas (1997). *Werther-Roman und Werther-Figur in der deutschen Prosa des Wilhelminischen Zeitalters*. St. Ingbert: Röhrig.

Jansen, Dorothea (2003). *Einführung in die Netzwerkanalyse. Grundlagen, Methoden, Forschungsbeispiele*. Opladen: Leske & Budrich. DOI: 10.1007/978-3-663-09875-1.

Jurafsky, Daniel und James H. Martin (2008). *Speech and language processing. An introduction to natural language processing*. 2. Aufl. Upper Saddle River, NJ: Prentice Hall.

Kantner, Cathleen (2004). *Kein modernes Babel: Kommunikative Voraussetzungen europäischer Öffentlichkeit*. Wiesbaden: Springer VS.

Kantner, Cathleen (2015). *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. London/New York: Routledge.

Kantner, Cathleen und Udo Tietz (2013). „Identitäten und multiple Identitäten. Über die wertrationale Integration der Gemeinschaften unter den Bedingungen der Moderne". In: *Dialektik–Arbeit–Gesellschaft. Festschrift für Peter Ruben*. Hrsg. von Erhard Crome und Udo Tietz. Potsdam: Welt-Trends, S. 47–63.

Ketschik, Nora (in Vorbereitung). „Mittelhochdeutsche Großepik im Lichte computergestützter Methoden (Arbeitstitel)". Diss. Universität Stuttgart.

Kuhn, Jonas (2020). „Einleitung". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 9–40.

Lafferty, John, Andrew McCallum und Fernando C. N. Pereira (2001). „Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. Burlington/San Francisco: Morgan Kaufmann Publishers, S. 282–289.

Lahn, Silke und Jan Christoph Meister (2016). *Einführung in die Erzähltextanalyse*. 3. Aufl. Stuttgart: Metzler.

Lotman, Juri M. (1972). *Die Struktur literarischer Texte*. Übers. von Rolf-Dietrich Keil. München/Stuttgart: UTB.

Manning, Christopher und Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. 2. Aufl. Cambridge, Massachusetts: MIT Press.

Martens, Lorna (1985). *The Diary Novel*. Cambridge: Cambridge University Press.

McCallum, Andrew Kachites (2002). „MALLET: A Machine Learning for Language Toolkit". http://mallet.cs.umass.edu. (Besucht am 1. Juni 2020).

Pfister, Manfred (2001). *Das Drama: Theorie und Analyse*. 11. Aufl. Bd. 580. Uni-Taschenbücher Literaturwissenschaft. München: Fink.

Pichler, Axel, André Blessing, Nils Reiter und Mirco Schönfeld (2020). „Algorithmische Mikrolektüren philosophischer Texte". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 328–372.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Platz-Waury, Elke (1997a). „Figur". In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Hrsg. von Klaus Weimar. Berlin: De Gruyter, S. 587–589.

Platz-Waury, Elke (1997b). „Figurenkonstellation". In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Hrsg. von Klaus Weimar. Berlin: De Gruyter, S. 591–593.

Propp, Vladimir Jakovlevič (1972). *Morphologie des Märchens*. Russisch übers. von Karl Eimermacher. München: Hanser.

Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.

Reiter, Nils, André Blessing, Nora Echelmeyer, Steffen Koch, Gerhard Kremer, Sandra Murr und Max Overbeck (2017). „CRETA Unshared Task zu Entitätenreferenzen (CUTE)". In: *Abstracts der DHd: Digitale Nachhaltigkeit*. Bern: Digital Humanities im deutschsprachigen Raum e.V., S. 19–22.

Reiter, Nils, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler (2020). „*Reaching out*: Interdisziplinäre Kommunikation und Dissemination". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 467–484.

Reznicek, Marc (2013a). „Linguistische Annotation von Nichtstandardvarietäten - Guidelines und 'Best Practices'. Guidelines Koreferenz. Version 1.1". Annotationsrichtlinien. Humboldt-Universität zu Berlin. URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-cor-1.1 (besucht am 1. Juni 2020).

Reznicek, Marc (2013b). „Linguistische Annotation von Nichtstandardvarietäten - Guidelines und 'Best Practices'. Guidelines NER. Version 1.5". Annotationsrichtlinien. Humboldt-Universität zu Berlin. URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5 (besucht am 1. Juni 2020).

Richter, Sandra (2017). *Eine Weltgeschichte der deutschsprachigen Literatur*. München: C. Bertelsmann Verlag.

Rorty, Richard M. (1992). *The Linguistic Turn. Essays in Philosophical Method*. University of Chicago Press: Chicago.

Rosenberg, Jay (2009). *Philosophieren. Ein Handbuch für Anfänger*. Übers. von Brigitte Flickinger. 6. Aufl. Frankfurt am Main: Vittorio Klostermann.

Scherpe, Klaus R. (1970). *Werther und Wertherwirkung. Zum Syndrom bürgerlicher Gesellschaftsordnung im 18. Jahrhundert*. Bad Homburg: Athenaion.

Schulz, Armin (2012). *Erzähltheorie in mediävistischer Perspektive*. Berlin: De Gruyter.

Schulz, Sarah (2018). „The Taming of the Shrew. Non-standard text processing in the Digital Humanities". Diss. Universität Stuttgart. DOI: 10.18419/opus-9685.

Tietz, Udo (2002). *Die Grenzen des "Wir". Eine Theorie der Gemeinschaft*. Frankfurt am Main: Suhrkamp.

Trilcke, Peer (2013). „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In: *Empirie in der Literaturwissenschaft*. Hrsg. von Philip Ajouri, Christoph Rauen und Katja Mellmann. Bd. 8. Poetogenesis - Studien zur empirischen Anthropologie der Literatur. Münster: Mentis, S. 201–247. DOI: 10.30965/9783957439710_012.

Viehhauser, Gabriel (2020). „Zur Erkennung von Raum in narrativen Texten". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 373–388.

Vorderstemann, Karin (2007). *Ausgelitten hast du - ausgerungen ...": Lyrische Wertheriaden im 18. und 19. Jahrhundert*. Heidelberg: Winter.

Roman Klinger, Evgeny Kim, and Sebastian Padó

Emotion Analysis for Literary Studies

Corpus Creation and Computational Modelling

Abstract: Most approaches to emotion analysis in fictional texts focus on detecting the emotion class expressed over the course of a text, either with machine learning-based classification or with dictionaries. These approaches do not consider who experiences the emotion and what triggers it and therefore, as a necessary simplicifaction, aggregate across different characters and events. This constitutes a research gap, as emotions play a crucial role in the interaction between characters and the events they are involved in. We fill this gap with the development of two corpora and associated computational models which represent individual events together with their experiencers and stimuli. The first resource, REMAN (Relational EMotion ANnotation), aims at a fine-grained annotation of all these aspects on the text level. The second corpus, FANFIC, contains complete stories, annotated on the experiencer-stimulus level, i. e., focuses on emotional relations among characters. FANFIC is therefore a character relation corpus while REMAN considers event descriptions in addition. Our experiments show that computational stimuli detection is particularly challenging. Furthermore, predicting roles in joint models has the potential to perform better than separate predictions. These resources provide a starting point for future research on the recognition of emotions and associated entities in text. They support qualitative literary studies and digital humanities research. The corpora are freely available at http://www.ims.uni-stuttgart.de/data/emotion.

Zusammenfassung: Die meisten Ansätze zur Emotionsanalyse in fiktionalen Texten konzentrieren sich auf das Erkennen der in Text ausgedrückten Emotion, entweder mit maschinellem Lernen oder mit Wörterbüchern. Diese Ansätze berücksichtigen in der Regel nicht, wer die Emotion erlebt und warum. Dies stellt eine Vereinfachung dar, die dazu führt, dass über verschiedene Charaktere und Ereignisse aggregiert wird. Emotionen spielen aber eine entscheidende Rolle in der Interaktion zwischen den Charakteren und den Ereignissen, in die sie verwickelt sind. Wir füllen diese Lücke mit der Entwicklung von zwei Korpora und den zugehörigen Berechnungsmodellen, die einzelne Emotionen (Emotionsereignisse) und die dazugehörigen Charaktere bzw. Stimuli in Beziehung setzen. Die Ressource REMAN (Relationale EMotionsANnotation) zielt auf eine feinkörnige Annotati-

Roman Klinger, Evgeny Kim, Sebastian Padó, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Roman Klinger, Evgeny Kim und Sebastian Padó; published by De Gruyter BY-NC-ND
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license.
https://doi.org/10.1515/9783110693973-011

on all dieser Aspekte auf der Textebene. Unser FANFIC-Korpus enthält komplette Geschichten, annotiert auf der Erlebnis-Stimulus-Ebene, wobei aber Stimuli jeweils durch andere Figuren realisiert sind. Diese Ressource konzentriert sich deshalb auf die Relationen zwischen Figuren, während REMAN zusätzlich Ereignisbeschreibungen betrachtet. Unsere Modelle zeigen, dass insbesondere die automatische Erkennung von Stimuli eine Herausforderung ist. Weiterhin hat die gemeinsame Modellierung das Potential, besser zu funktionieren als getrennte Vorhersagen. Unsere Ressourcen bilden einen Ausgangspunkt für zukünftige Forschung zur Erkennung von Emotionen und assoziierten Entitäten im Text. Sie unterstützen qualitative Literaturwissenschaft und digitale geisteswissenschaftliche Forschung. Die Korpora sind frei verfügbar unter http://www.ims.uni-stuttgart.de/data/emotion.

1 Introduction

The analysis of affect in text in general became popular in computational linguistics as well as in application areas like social media mining or computational literary studies with the work by Wiebe (2000), who aimed at distinguishing subjective language from objective, factual statements. Based on this groundbreaking work, several subtasks have emerged, including sentiment analysis (classifying positive vs. negative statements). Another related subfield from this domain is to analyze emotions that are associated with text. This field recently attracted increasing attention, with the creation of corpora and automatic models. One focus is to analyze social media, as it is easy to access and process, and commonly full of relevant instances (Mohammad 2012a; Mohammad, Zhu, et al. 2014; Klinger, De Clercq, et al. 2018).

Emotions are also a crucial component of compelling narratives (Oatley 2002; Ingermanson and Economy 2009; Hogan 2015). Not only do emotions help readers to understand texts (Barton 1996; Robinson 2005) but they also improve readers' abilities of empathy and understanding of others' lives (Mar et al. 2009; Kidd and Castano 2013). This makes literature an interesting field for the study of emotions, as evidenced by the growing interest in emotion-oriented text analysis among digital humanities scholars.

Most research in this regard is based on annotated data, in different variations, either directly for the analysis of the text, the development of automatic systems, or the evaluation of such systems. Emotion annotation can be defined at different textual levels. For example, the corpus which originates from the ISEAR project (Scherer and Wallbott 1994) is annotated on the level of (short) documents, each of which contains the description of an emotionally charged situation. Ex-

amples of resources with sentence-level annotation include the work by Alm et al. (2005), a corpus of children stories, and Strapparava and Mihalcea (2007), who label news headlines. While these studies do not annotate any explicitly textual markers (also called cues) of emotion (Johnson-Laird and Oatley 1989), Aman and Szpakowicz (2007), who annotate blogposts, do include such textual markers. Wiebe et al. (2005) annotate a corpus of news articles with emotions at a word and phrase level. Mohammad, Zhu, et al. (2014) annotate emotion cues in a corpus of 4058 electoral tweets from US via crowdsourcing. Similar in annotation procedure, Liew et al. (2016) curate a corpus of 15 553 tweets and annotate it with 28 emotion categories, valence, arousal, and cues.

A number of studies, including the ones named in the previous paragraph, have explored automatic emotion analysis. In the context of literary studies, relatively simple setups have been used (see Kim and Klinger 2019), notably classification, where a single emotion label is assigned to a segment of text. This corresponds directly to the emotion annotation schemes sketched above. For instance, Kim, Padó, et al. (2017) show that emotions, recognized with dictionaries or bag-of-words models, can serve as features for genre classification in fiction. The predictive power of these models, however, remains generally limited.

We believe that the very simplicity of classification is one of the reasons for the limited performance: Such approaches ignore the semantic role-like structure of emotion, which are not textual categories but rather events. Obviously, the semantic roles in fiction should not be disconnected from their narratological embeddings: when there is an emotion, there is typically somebody who feels the emotion, a target for the motion, and a cause for it (Russell and Barrett 1999; Scarantino 2016). Consider the sentence "*Jack is afraid of John because John has a knife*". Following structural approaches to defining emotional episodes, the sentence can be rephrased as "emotion of fear is experienced by Jack (experiencer) because John (target) has a knife (cause)". Here, dictionary-based or bag-of-words approaches would probably capture that this sentence describes fear, but would fail in assigning the correct semantic roles to John and Jack. This could lead us to conclude, incorrectly, that their emotional experiences are the same.

Compared to classification approaches, there is a rather limited amount of both annotation and modelling work which considers emotions from a structured point of view. There are a few studies on English (Mohammad, Zhu, et al. 2014; Gao et al. 2015; Ghazi et al. 2015; Kim and Klinger 2018) and a considerable number on Mandarin Chinese (Gui, Yuan, et al. 2014; Li and Xu 2014; Gao et al. 2015; Gui, Wu, et al. 2016; Cheng et al. 2017; Gui, Hu, et al. 2017; Xu, Hu, et al. 2017; Chen et al. 2018; Ding et al. 2019; Xia and Ding 2019; Xia, Zhang, et al. 2019; Xu, Lin, et al. 2019). Notably, the corpus by Mohammad, Zhu, et al. (2014) considers experiencers, the stimuli, and targets. However, in the case of tweets, the experi-

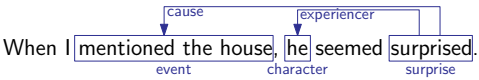


Fig. 1: Example annotation in REMAN for a sentence from Hugo (1885), with one character, an emotion word, and event and cause and experiencer annotations.

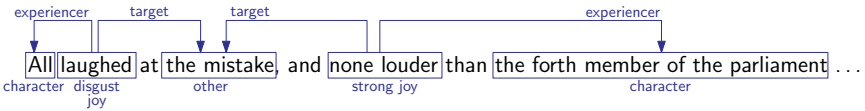


Fig. 2: Example annotation in REMAN for a sentence from Stimson (1943), with two characters who are experiencers of different emotions. Disgust and joy are annotated as a mixture of emotions. Both emotions have the same target.

encer is mostly the author of the tweet. Another recent resource of news headlines annotated via crowdsourcing is GoodNewsEveryone (Bostan et al. 2020). This corpus includes annotations of the perceived emotion of the reader in addition to the direct text realization of the emotion.

With this article, we present two corpora and modelling experiments which contribute to this situation, particularly for literature. In the REMAN corpus presented here (Relational Emotion Annotation), we aim for a more comprehensive analysis of emotion events in terms of the semantic roles that these events possess. Our work loosely follows the concept of directed emotions, as defined in FrameNet (Fillmore et al. 2003), and extends the work of Ghazi et al. (2015), who focus on detecting emotion stimuli in the FrameNet exemplary sentences annotated for emotions and causes. In REMAN, we annotate and extract who feels (*experiencer*) which emotion (*cue*, *class*), towards whom the emotion is expressed (*target*), and what is the event that caused the emotion (*stimulus*). Our study is the first one to apply this idea to literary texts. Figures 1 and 2 show examples of the more complex annotation in REMAN.

A second aspect which we believe to be understudied is the role of emotions in characterizing interpersonal relations. This direction links up emotion analysis with social network analysis, an important strand of research in computational literary studies (Agarwal et al. 2013; Nalisnick and Baird 2013; Piper et al. 2017, i. a.). The REMAN resource covers this direction to some extent, since some emotion stimuli happen to be characters, but does not do so in a focussed manner. Starting from the idea that structured emotion representations can serve as a basis for inferring relations between experiencers and stimulus characters, we create a second resource, the FANFIC corpus. In FANFIC, all emotion experiencers are annotated with the emotions they perceive and, if available, with the character which

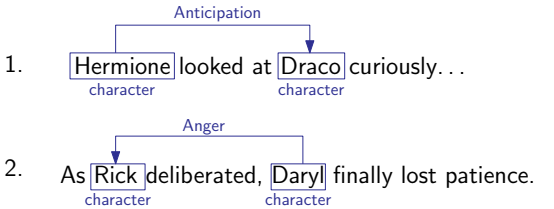


Fig. 3: Examples for emotional character interaction in the FANFIC corpus. Example (1) taken from Apryl_Zephyr (2016), example (2) from EmmyR (2014). The arrow starts at the experiencer and points at the causing character.

plays a role in causing the emotion. Figure 3 depicts two examples for emotional character interactions at the text level.

In the remainder of this chapter, we use a discrete set of emotions, based on fundamental emotions as proposed by Plutchik (2001), a common choice of emotion inventory. This model has previously been used in computational analysis of literature (Mohammad 2012b, i. a.). We refer the reader to social psychology literature for more details about alternative emotion theories (such as Ekman 1992) and on the emotional relationships among persons (Gaelick et al. 1985; Burkitt 1997).

Our work has potential to support literary scholars, for example, in analyzing differences and commonalities across texts. As an example, one may consider Goethe's *The Sorrows of Young Werther* (Goethe 1774), a book that gave rise to a plethora of imitations by other writers, who attempted to depict a similar love triangle between main characters found in the original book. The results of our study provides a computational methodology on the basis of which derivative works can be compared systematically with the original.

Our main contributions are therefore: (1) We discuss and make available two resources of fictional texts annotated for emotions, experiencers, causes, and targets as well as for emotional character relations; (2) We analyze the corpora and show which emotions are realized more often with stimuli than others. (3) We provide results of computational methods which automatize the annotation process of emotion words, roles and relations, and further (4) show that the prediction performance of all subtasks benefits from joint modelling, similar to the process of human reading, which is also not entirely linear but considers relations in the text to develop an understanding.

2 Annotation Task

We describe the creation and modelling of two resources, the REMAN corpus of emotion semantic role labeling and the FANFIC corpus of character relations.

2.1 REMAN: Semantic Role Labeling for Emotion Recognition in Literature

The REMAN corpus is a dataset of excerpts from fictional texts annotated for the phrases that evoke emotions, the experiencer of each emotion (a character in the text, if mentioned), the target and its cause, if mentioned (e. g., an entity, or event). An example of such an annotation is shown in Figures 1 and 2. Each annotation includes textual span labels such as emotions, characters, and events, as well as relational annotations that establish relations among text spans (viz., cause, experiencer, target). We now describe the conceptual background for each annotation layer in detail. The complete annotation guidelines are available online together with the corpus at http://www.ims.uni-stuttgart.de/data/emotion.

2.1.1 Conceptualization

We conceptualize ***emotions*** as an individual's experiences that fall in the categories in Plutchik's classification of emotions, namely *anger*, *fear*, *trust*, *disgust*, *joy*, *sadness*, *surprise*, and *anticipation*. In addition, we permit annotations with the class *other emotion* to capture cases when the emotion expressed in the text cannot be reliably categorized into one of the predefined eight classes. This emotion is not meant to extend the existing set of labels, but aims to cover ambiguous and vague emotion expressions. A list of the emotions along with example realizations can be found in Table 1.

Annotators are instructed to preferentially annotate individual key words (e. g., "afraid"), except in cases when emotions are expressed by complete phrases (e. g., "tense and frightened", "wholly absorbed") or by contextual realizations of emotion expression (e. g., "the corners of her mouth went down"). Additionally, emotion spans can be marked as intensified (i. e., amplified, "very happy"), diminished (i. e., downtoned, "a bit sad") and negated ("not afraid") without marking the modifier span or including the modifier word. Spans can be associated with one or more emotion labels (exemplified in Figure 2).

Tab. 1: Concepts used for the phrase annotation layer in REMAN together with examples.

	Concept Value	Examples
Emotion	Anger	*angry, defend themselves by force, break your little finger, loss of my temper*
	Anticipation	*want, wish, wholly absorbed, looked listlessly round, wholly absorbed*
	Disgust	*repellent, cheap excitement, turn away from, beg never to hear again*
	Fear	*horrified, tense and frightened, shaking fingers*
	Joy	*cheerful, grateful, boisterous and hilarious, violins moved and touched him*
	Sadness	*failed, despair, the cloudy thoughts, staring at the floor*
	Surprise	*perplexing, suddenly, petrified with astonishment, loss for words, with his mouth open*
	Trust	*honor, true blue, immeasurable patience*
	Other	*careful, brave, had but a tongue, break in her voice, bit deeply into his thumb*
Modifier	strong	*I loved her the more*
	weak	*with a little pity*
	negated	*could not be content*
Entity	character	*the chairman of the board*
	event	*marry a man I did not love, because of his gold*
	other	*Lily's beauty*

As a preparation for relation annotation, we annotate ***entities***, which are of a clear identity, for instance of a person, object, concept, state, or event (see Table 2). We only annotate them in the context of relations. The subtypes we are particularly interested in are:

Character An entity that acts as a character in the text. Character annotation should not omit important information (e. g., the annotation of "the man with two rings of the Royal Naval Reserve on his sleeve" is preferred over only annotating 'the man').

Event An event is an occasion or happening that plays a role in the text. Events can be expressed in many ways (see Table 2 for examples from the annotated dataset) and annotators are instructed to label the entire phrases including complementizers or determiners.

Other This is an umbrella concept for everything else that is neither a character nor an event, but participates in a relation.

Next, we annotate ***relations***, links between an emotion and other text spans and can be of type *experiencer*, *cause*, and *target*. They can be thought of as the roles that entities play with regard to specific emotions. These relations can only originate from the emotion annotations. In addition, we partially annotate *coreferences* to link personal pronouns to proper nouns.

Tab. 2: Typical linguistic realization of entities.

Entity type	Linguistic realiz.	Examples
Character	noun phrase	*his son*
	adjectival phrase	*old man*
Event	verb phrase	*Mrs. Walton had got another baby.*
	adverbial phrase	*Jesus spoke unkindly to his mother when he said that to her.*
	prepositional phrase	[...] *giving her up.*
	clause	[...] *what she said to him* [...]
	noun phrase	*the journey*
Other	adjectival phrase	[...] *old age* [...]
	noun phrase	[...] *the heavens and the earth.*
	tense phrase	*She was the only treasure on the face of the Earth that my heart coveted.*

Roles which are part of relations are:

Experiencer The experiencer relation links an emotion span and entity of type *character* who experiences the emotion. If the text contains multiple emotions with multiple experiencers, they all are subject to relation annotation.

Target The target relation links an emotion span and entity of any type towards which the emotion experienced by the experiencer is directed. If there are multiple targets of the emotion, then all of them should also be included in the relation annotation. See Figure 2 for the example of a target annotation.

Cause The cause relation links an emotion span and entity of any type, which serves as a stimulus, something that evokes the emotional response in the experiencer. If there are multiple causes for the emotion, then all of them are included in separate relation annotations.

Coreference The annotators are instructed to annotate as an experiencer the character that is the closest to the emotion phrase in terms of token distance. If the closest mention of the character is a pronoun and the text provides a referent that has a higher level of specificity than the pronoun (i. e., a proper noun or a noun denoting a group or class of objects), the annotators are asked to resolve the coreference. The coreference annotation can be used later to evaluate downstream task applications which associate emotions with a unique character instead of a pronoun.

2.1.2 Corpus Construction and Annotation

Selection

The corpus of 200 books is sampled from Project Gutenberg[1]. All books belong to the genre of fiction and were written by authors born after the year 1800[2]. We sample consecutive triples of sentences from this subsample of books. A triple is accepted for inclusion for annotation if the middle sentence includes a word that is known to be associated with an emotion, even in isolation. This increases the probability that actual emotion-role-relevant content is present in the instance. To realize this, we use the so-called NRC Emotion Dictionary, which consists of 14 183 linguistic units with an associated emotion (Mohammad and Turney 2013). We consider this middle sentence the target sentence and the annotators are instructed to label emotions in this second sentence only. Experiencers, causes and targets are annotated in the whole sentence triple if they refer to an emotion in the target sentence.

When selecting texts for annotation, there is a trade-of between short passages which are easy to parse but might not contain all relevant roles around an emotion expression, and longer passages which are more likely to contain all relevant relations but are more time-consuming to annotate. Ghazi et al. (2015), for instance, annotate only one sentence and speculate whether adding one sentence before and after will lead to better results. To check their hypothesis, we conduct a small pre-study experiment by extracting 100 random sentences from Project Gutenberg with the NRC dictionary and analyze how often the roles of experiencer, cause, and target are found in the target sentence and in the window of up to five sentences before and after. The analysis shows that 98 % of the snippets include the experiencer in the target sentence, while cause and target are found in the target sentence in 67 % of the texts. Another 29 % of the texts include cause and target in the window of one sentence before and after the target sentence. The remaining texts include cause and target in the window of two (2 %), three (1 %), and four (1 %) sentences around the target sentence. Evidently, three-sentence spans provide enough information regarding 'who feels what and why' without creating excessive annotation overhead (cumulatively, 96 % of cause and target are found in such sentence triples in the pre-study). We therefore opt for the anno-

1 http://www.gutenberg.org/. Note that Project Gutenberg is currently not available in Germany due to an ongoing legal dispute. None of the texts under discussion regarding copyright are part of our corpus.

2 We wanted to work with the texts from the nineteenth and twentieth centuries. However, metadata available to us does not include the book publication date, but specifies the birth year of the author.

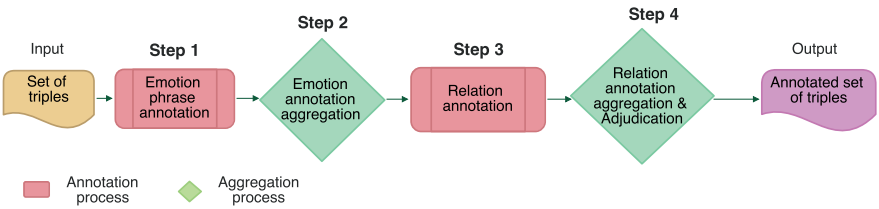


Fig. 4: A visualization of the multi-step annotation process.

tation of sentence triples with one sentence before and one after an emotion cue (preselected with the NRC dictionary).

Annotation Procedure

The annotations are generated in a multistep process visualized in Figure 4. The people involved in the annotation were either *annotators* or *experts*, whose roles did not overlap. The annotations (of spans and relations) were performed by three graduate students of computational linguistics (two native English speakers, one non-native speaker) within a three-month period. Arising questions were discussed in weekly meetings with the experts (the two first authors of the paper) and the results documented in the annotation guidelines. Further, the experts perform manual adjudication in cases where automatic annotation aggregation is not possible (see below). We use WebAnno[3] (Yimam et al. 2013) as annotation environment. In the following, we discuss the four steps of generating the corpus.

Step 1: Emotion phrase annotation The *annotators* are asked to first decide whether the text to be annotated expresses an emotion and which emotion it is. If any exists, they label the phrase that led to their decision. The annotators are instructed to search for emotions that are expressed either as single words or phrases.

Step 2: Emotion phrase aggregation In the previous step, each annotator generates a set of annotations. In this step, the *expert* heuristically aggregates all spans that overlap between annotators in a semi-automatic process: Concrete emotions are preferred over the 'other-emotion' category, annotations with modifier are preferred over annotations without, and shorter spans are preferred over longer spans. Overlapping annotations with different emotion labels are all accepted.

Step 3: Relation annotation *Annotators* receive the texts that they annotated for emotions in Step 1, including the aggregated annotation from Step 2 based on

3 https://webanno.github.io/webanno/

all individual annotation. Thus, all annotators see the same texts and annotations in this step. For each emotion, the task is now to annotate entities that are experiencers, targets, or causes of the emotion and relate them to their respective emotions. The annotators are instructed to tag only those entities that have a role of an experiencer, cause, and target. The decision on the entity and relation annotation is made simultaneously: For each emotion, the annotators need to identify who experiences the emotion (which *character*) and why (because of an event, object, or other character).

Step 4: Relation aggregation and adjudication This final step is a manual *expert* step: Aggregate the relation annotations provided by the annotators. Heuristically, we prefer shorter spans for entities, but guide ourselves with common sense. For instance, consider the phrase "[...] *wishing rather to amuse and flatter himself by merely inspiring her with passion*". "*Wishing*" is labelled as emotion. One annotator tagged "*to amuse and flatter himself by merely inspiring her with passion*" as event, another tagged only "*by merely inspiring her with passion*", which is incomplete, as the target of the emotion is the act of amusing and flattering oneself.

Note that we do not discard the rejected annotations but publish all annotations of all annotators.

2.2 FANFIC: Character Relations in Fanfiction

The goal of the FANFIC corpus is different from REMAN but shares the goal of moving beyond identifying emotion labels for stretches of text. In REMAN, emotions are assigned to cues that are related to experiencers and stimuli. This is a detailed approach of modeling the structure of emotions in literature, but it might be too complex for computational modeling approaches to perform well. Further, we do not link the relational structure of emotions to the field of social network analysis.

This is what we aim at with the FANFIC corpus. On the one side, we opt for a simpler formulation of emotion structures, namely emotional relations between characters. On the other side, we aim at an evaluation on the document level, in the spirit of social network analysis.

2.2.1 Conceptualization

FANFIC is centered completely around interpersonal emotions: emotions are understood as relations between characters in a text. Formally, each emotion rela-

tion is a triple $(C_{\text{exp}}, e, C_{\text{cause}})$ in which the character C_{exp} feels the emotion e (mentioned in text explicitly or implicitly). The character C_{cause} is part of an event which triggers the emotion e. We consider the eight fundamental emotions defined by Plutchik (2001) (anger, fear, joy, anticipation, trust, surprise, disgust, sadness). Each character corresponds to a token sequence for the relation extraction task. In a social network analysis setting, these characters correspond to normalized entities. Note that, in contrast to REMAN, we do not annotate the exact span that triggers the emotion (which is difficult in cases of implicit emotion descriptions), nor do we annotate causes that are events or objects.

2.2.2 Data Collection and Annotation

To be able to evaluate on the social network interaction level, the annotation of complete stories is required. We therefore annotate a sample of 19 complete English fan-fiction short stories, retrieved from the Archive of Our Own project[4] (due to availability, the legal possibility to process the texts and a modern language), and a single short story by Joyce (1914), "Counterparts". All fan-fiction stories were marked by the respective author as complete, are shorter than 1500 words, and include at least four different characters. They are tagged with the keywords 'emotion' and 'relationships' as metadata in the repository.

The annotators were instructed to mark every character mention with a canonical name and to decide if there is an emotional relationship between the character and another character (again, using WebAnno, as in the creation of the REMAN corpus). If so, they marked the corresponding emotion phrase with the emotion labels (as well as indicating if the emotion is amplified, downtoned or negated). Based on this phrase annotation, they marked two relations: from the emotion phrase to the experiencing character and from the emotion phrase to the causing character (if available, i. e., C_{cause} can be empty). One character may be described as experiencing multiple emotions.

We generate a 'consensus' annotation by keeping all emotion labels by all annotators. This is motivated by the finding by Schuff et al. (2017) that aggregating with the goal of achieving a high recall leads to better performance for emotion prediction. As we focus here on emotion relationships, we retain all emotion labels from all annotators, providing for the richest possible emotion representation.

4 https://archiveofourown.org

3 Corpus Analyses

3.1 REMAN

In the following, we first discuss the annotation and then provide results of models trained on our resources.

3.1.1 Inter-Annotator Agreement and Consistency of the Annotations

The first step of our analysis is the evaluation of the quality of the annotations. For that, we make use of the Cohen's Kappa coefficient (κ, Artstein and Poesio 2008), a measure to measure the agreement of independent annotators. In this measure, the main ingredient is the the probability that two annotators agree, which is calculated by counting how often they made the same annotation choices. This value is the same as 'accuracy'. However, in contrast to this simple fraction of correct annotations, Cohen's kappa normalizes by the *expected agreement*.

We calculate this measure at the token level. In addition, we calculate F_1 score on the phrase level both with exact match (where all tokens need to be the same between annotators) and with fuzzy match (where one token overlap is sufficient such that the annotation counts as being the same). In this manner, we calculate agreement both for the phrase annotation and the relation annotation.

Table 3 reports the IAA agreement scores for emotion, entity, and relation annotations for each pair of annotators. Among all emotions, *joy* has the highest number of instances (336) and the highest agreement scores (average κ=0.35), followed by *fear* (κ=0.30) and *sadness* (κ=0.24). *Other emotion* has the lowest agreement with average κ=0.07 – not surprisingly, given the nature of this label as category for difficult cases. For entity annotation, especially for *character* annotation, the agreement is higher, with the highest agreement between two annotators being κ=0.63. The agreement on the *event* and *other* entities is low (κ=0.23 and 0.14 and F_1=25 and 14, respectively). This is presumably the case because event annotations are often comparably long which makes it hard to achieve exact match. This also holds, to a lesser extent, for *character* annotations. If we allow partial overlaps to count as a match, the average F_1 increases to 57 for *character* (an increase of 4 percentage points (pp)), 44 for *event* (increase by 19 pp), and 23 for *other* category (increase by 9 pp).

For relation annotations, fuzzy evaluation also leads to higher agreement scores (F_1 increase for *experiencer*, *cause* and *target* by 10 pp, 7 pp, and 12 pp respectively). These results are in line with previous studies on emotion cause an-

Tab. 3: Pairwise inter-annotator agreement for phrase annotation and relation annotation in REMAN (annotators a, b, c). F_1 is in %. Regarding the relation scores, in strict F_1, a TP holds if the relation label agrees and the entity it points to has the same label and span. In fuzzy F_1 ($\approx F_1$), a TP holds if the relation and the entity it points to have the same lavel, but the span boundary of the entity may differ.

		a vs. b			b vs. c			a vs. c		
	Type	**κ**	**strict F_1**	**$\approx F_1$**	**κ**	**strict F_1**	**$\approx F_1$**	**κ**	**strict F_1**	**$\approx F_1$**
Emotion	anger	.25	25	39	.15	15	38	.18	18	33
	anticipation	.09	9	23	.07	7	20	.18	18	39
	sadness	.32	32	41	.22	23	41	.19	20	29
	joy	.38	39	50	.40	40	55	.28	28	44
	surprise	.26	26	43	.22	23	33	.27	27	37
	trust	.17	17	26	.14	14	21	.12	13	32
	disgust	.23	23	41	.10	10	26	.19	19	31
	other	.07	7	7	.06	6	11	.08	8	22
	fear	.35	35	48	.28	28	35	.28	28	41
Entity	character	.63	63	68	.48	49	51	.48	48	54
	event	.29	31	60	.09	10	30	.32	34	44
	other	.11	12	28	.11	11	18	.20	21	23
Relation	experiencer		65	73		48	57		46	55
	cause		20	28		34	39		26	32
	target		27	36		18	29		14	28

notation (Russo et al. 2011), and show that disagreements mainly come from the different choices about the precise annotation spans, while the spans typically overlap.

3.1.2 Assessing Low Agreement

As we show in Section 3.1.1, the agreement across all annotation layers is relatively low, even for a semantic annotation task. There are several reasons. Indeed, emotion categorization is highly subjective and emotions often co-occur (Schuff et al. 2017). In addition, the cause and target of the emotion are not always clearly recognizable in the text and are also subjective categories (two annotators may find two different causes for the same emotion), problems that emotion role annotation inherits from general semantic role annotation (Ellsworth et al. 2004) – hence the low agreement scores across all categories. The only exception are *experiencer* annotations, which are the most reliable among all annotations and match the substantial agreement scores of character annotation (the only type of entities that can be involved in an experiencer relation).

Tab. 4: REMAN corpus statistics for emotions annotation. Columns indicate the frequency of each emotion.

				Modifier			Annotation Length			
	Type	**Total**	**Adjudic.**	**strong**	**weak**	**neg.**	**1 token**		**≥ 2 token**	
Emotions	anger	192	156	5	12	7	106	68%	50	32%
	anticipation	248	201	5	3	11	161	80%	40	20%
	disgust	242	190	2	7	14	144	76%	46	24%
	fear	254	183	11	16	17	145	79%	38	21%
	joy	434	336	31	20	28	289	86%	47	14%
	sadness	307	224	10	2	13	168	75%	56	25%
	surprise	243	196	12	4	7	156	80%	40	20%
	trust	264	232	3	3	33	191	82%	41	18%
	other emotion	432	207	4	4	4	133	64%	41	36%
Entities	character	2072	1715				1288	75%	427	25%
	event	858	615				38	6%	577	94%
	other	771	485				114	24%	371	76%

We illustrate the difficulties the annotators face when annotating emotions with roles with the following example:

> They had never seen ... what was really hateful in his face; ... they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below ...'.

In our study, both annotators agree on the character ("they") and the emotion ('hateful' expressing disgust). They also agree that the disgust is related to properties of the face which is described. However, one annotator marks "his face" as target, the other marks the more specific but longer "the arched brows and the long emphatic chin gave it always a look of being lit from below" as cause.

If we abstract away from the text spans, both annotators agree that the emotion of disgust has something to do with "his face", however they disagree on the target annotation and the cause annotation. We take such cases to indicate that while the annotation task is indeed difficult, the surface-oriented inter-annotator agreement measures that we compute arguably underestimate the amount of conceptual agreement among annotators. It is on this basis that we consider our annotation to be meaningful despite the low agreement.

Tab. 5: REMAN corpus statistics for relation annotation. Rows indicate the total frequency of each relation and each relation-entity combination.

	Entities involved		
Relation	**char.**	**event**	**other**
experiencer	1704		
cause	87	398	343
target	444	315	257
overall relations	2238	717	601

3.1.3 Corpus Statistics

Tables 4 and 5 show the total number of annotations for each annotation category, broken down by different criteria. In Table 4, the *Total* column shows the overall number of annotations generated by all annotators, while the *Adjudic.* column shows the number of accepted annotations. The REMAN corpus consists of 1720 sentence triples, 1115 of which include an emotion. This is a comparably low number, given that we picked the triples based on words that are associated with an emotion, according to an emotion dictionary. But this also shows, that the annotators of our corpus do not agree with an emotion assignment only based on a dictionary, challenging the use of such simple approach for emotion detection. Still, our corpus is densely populated with emotions, with 64 % of triples having an emotion annotation.

Joy has the highest number of annotations, while *anger* has the lowest number of annotations. *Joy*, in addition, is modified as *strong* and *weak* more often than other emotions, while *trust* is negated more often compared to other emotions. In most cases, emotion phrases are single tokens (e. g., 'monster', 'irksome'), out of which 47 % on average are found in the NRC dictionary. *Other emotion* has the largest proportion of annotations that span more than one token (36 % out of all annotations in this category), which is in line with our expectation that lower levels of specificity for emotion annotation make it more difficult to find a single token that indicates an emotion.

In Table 5, we see that, based on the definition of the annotation task, the role of experiencers can only be filled by characters. Causes and targets can be filled by characters, events, and other entities. Interestingly, characters are more often the target of an emotion than the cause. Events are more likely to cause the emotion than being the target of it.

Tab. 6: FANFIC corpus: F_1 scores at different levels in % for agreement between annotators (a1, a2, a3).

	a1–a2	a1–a3	a2–a3
Instances labelled	24	19	24
Instances unlabelled	33	27	29
Graph labelled	66	69	66
Graph unlabelled	90	93	92

3.2 FANFIC

3.2.1 Inter-Annotator Agreement

Recall that the goal of FANFIC is to use emotion annotation for the construction of social networks. From this perspective, it makes sense to define inter-annotator agreement not just in terms of the textual surface, but also at the level of the network computed from the annotations.

Therefore, we calculate the agreement along two dimensions, namely unlabelled vs. labeled and instance vs. graph-level. Table 6 reports the pairwise results for three annotators. In the *Instances labelled* setting, we accept an instance being labeled as true positive if both annotators marked the same span of text to label the characters as experiencer and cause of an emotion and classified their interaction with the same emotion. In the *Instances unlabelled* case, the emotion label is allowed to be different. On the graph level (*Graph labelled* and *Graph unlabelled*), the evaluation is performed on an aggregated graph of interacting characters, i. e., a relation is accepted by one annotator if the other annotator marked the same interaction somewhere in the text. We use the F_1 score to be able to measure the agreement between two annotators on the span levels. For that, we treat the annotations from one annotator in the pair as correct and the annotations from the other as predicted.

As Table 6 shows, agreement on the textual level is low with values between 19 and 33 % (depending on the annotator pair), which also motivated our previously mentioned aggregation strategy. The values for graph-labelled agreement, which arguably provide a more relevant picture for our use-case of network generation, are considerably higher (66 % to 93 %). This shows that annotators agree when it comes to detecting relationships, regardless of where exactly in the text they appear.

Emotion	All	Rel.
anger	258	197
anticipation	307	239
disgust	163	122
fear	182	120
joy	413	308
sadness	97	64
surprise	143	129
trust	179	156
total	**1742**	**1335**

Tab. 7: FANFIC corpus: Statistics of emotion and relation annotation. 'All' indicates the total number of emotion annotations. 'Rel.' indicates the number of emotional relationships (including a causing character) instantiated with the given emotion.

3.2.2 Corpus Statistics

Table 7 summarizes the aggregated results of the annotation. The column 'All' lists the number of experiencer annotations (with an emotion), the column 'Rel.' refers to the counts of emotion annotations with both experiencer and cause. In this sense, 'Rel.' column is a subset of 'All' column.

Joy has the highest number of annotated instances and the highest number of relationship instances (413 and 308 respectively). In contrast, *sadness* has the lowest number of annotations with a total count of instances and relations being 97 and 64 respectively. Overall, we obtain 1335 annotated instances, which we use to build and test our models.

4 Computational Modeling

We now come to the computational modeling part of our study, where we investigate how difficult the manually annotated emotion structures described above are to predict automatically with current NLP methods. Given the differences between the annotation schemes of REMAN and FANFIC, the models we develop differ to an extent. In the case of REMAN, we phrase the prediction of the emotional structure, including its arguments (experiencer, cause, emotion cue, target) as a sequence prediction task. We further analyze if the information about one of the roles is helpful to recognize another. In the case of FANFIC, we phrase the computational modeling as classification which relation exists between all pairs of characters from a given novel. This is a standard formulation for relation extraction.

Tab. 8: Experimental results for the REMAN corpus: Results for predicting Emotions and Roles (column *Predict.*) in Exp. 1–3 (column *Exp*).

					Strict			Fuzzy		
Predict.	**Exp**	**# Ann.**	**Model**	**Features**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
	1		Rule-based	dict	19	83	31			
	1		MLP	BOW	55	21	31			
Emotion	2	1925	CRF	all + dictionary	56	6	11	56	6	11
	3		CRF	all + dict + exp	55	9	16	69	12	20
	2		biLSTM-CRF	embeddings	57	35	43	62	39	48
	2		CRF	all + person	0	0	0	0	0	0
Cause	2	1550	biLSTM-CRF	embeddings	0	0	0	0	0	
	2		CRF	all + person	50	2	4	50	2	4
Exp'cer	3	1717	CRF	all + person + emo.	74	15	24	78	15	26
	2		biLSTM-CRF	embeddings	49	21	30	49	21	30
Target	3	1017	CRF	all + emo.	50	3	6	50	3	6
	3		biLSTM-CRF	embeddings	0	0	0	0	0	0

4.1 REMAN: Role Identification as Sequence Labeling

4.1.1 Experiment 1: Coarse-grained emotion classification

We start with our experiments by first studying how well we can identify the emotion in our sentence triples, without looking at the role labeling. This is therefore a standard approach to emotion analysis. The task is to assign one emotion to a sentence triple (target sentence plus two context sentences). We consider this task as the first step towards the full structured prediction tasks as defined above: It confirms that we can at least correctly predict the core of the emotion structure, namely the emotion itself.

We compare a dictionary-based approach and a bag-of-words-based classifier. For the dictionary-based classification, we take the intersection between the words in the triple and the NRC dictionary and assign the triple with the corresponding emotion labels. The F_1 score is calculated by comparing the set of labels predicted by dictionaries against the set of gold labels for each triple. The gold labels come from the annotation of words and phrases within each triple. For the BOW approach, we convert each triple into a sparse matrix using all words in the corpus as features. We then classify the triples with a multi-layer perceptron with three hidden layers, 128 neurons each, with an initial learning rate of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001.

The results of all experiments are summarized in Table 8. Experiment one corresponds to the first two rows (labeled with a '1' in the column 'Exp'). We evaluate our models in the same two ways as for inter-annotator agreement: Either by accepting a TP if it is exactly found (exact match) or if at least one token is overlapping with the annotation (fuzzy match).

Emotion classification with dictionaries and bag of words shows mediocre performance. The recall with the dictionary classification is comparably high (F_1=0.83), which is due to the fact that texts were sampled using these dictionaries. However, as we said earlier, annotators are free to label any words and phrases as emotion-bearing, hence low precision, recall, and consequently F_1 score. The MLP with BOW features does not perform better but shows increased precision at the cost of lower recall.

The experiments therefore show a comparably low performance for the classification setting. However, given that for emotion classification in social media (e. g., Schuff et al. 2017), the results are also typically around 0.60 F_1, this is a reasonable result, as literature can be considered a linguistically more complex genre.

4.1.2 Experiment 2: Fine-grained emotion and role detection

We now turn to the setting of recognizing the words that correspond to the role and which trigger the emotion. We phrase this as a sequence labeling task, i. e., a sequence of input tokens is assigned a corresponding sequence of output labels. A classical example from NLP is part-of-speech tagging, where each word is assigned a part of speech. We can phrase emotion structure prediction on the REMAN data as a sequence prediction task, where the input is again the sentence, and each word is assigned either an emotion (if it is a cue for an emotion event), one of the labels *experiencer*, *target*, or *cause*, if it is part of a phrase that fills the corresponding role, or *none*, if it does not participate in the emotion structure. Note that we lose the explicit relation between emotion event and its roles; however, since few sentences contain multiple emotion events, we simply assume that all roles and emotion events within a sentence belong to one another.

Consider the example depicted in Figure 1: The phrase "I mentioned the house" is labelled as an event and is assigned a role of a *cause* for the emotion of *surprise*, and the word "he" is labelled as a character and is assigned a role of an *experiencer* of the same emotion. We represent these relationships by tagging "I

mentioned the house" as *cause* and "he" as *experiencer*, capturing the text spans that are linked by relations with an emotion.[5]

We use conditional random fields (CRF) (Lafferty et al. 2001) and bidirectional long short-term memory networks with a CRF layer (biLSTM-CRF), which are both known to provide generally good performance in sequence prediction tasks (Benikova et al. 2014; Huang et al. 2015). Conditional random fields can be considered an extension of hidden Markov models in the sense that they also are probabilistic and that they also model transition probabilities. However, they do that in the spirit of maximum entropy classifiers, which can deal with many correlated features (see Klinger and Tomanek 2007 for more details). BiLSTM-CRF are essentially conditional random fields, but extract feature representations with a deep neural network. Remarkably, in the biLSTM unit of this network, the Markov property is relaxed such that long distant relations can be considered.

We evaluate the performance of fine-grained emotion and role (experiencer, target, and cause) prediction in a sequence labelling fashion. We train separate CRF and biLSTM-CRF models for each relation, as some annotations overlap (e. g., experiencers can also be targets/causes). The CRF uses part-of-speech tags (detected with spaCy[6], Honnibal 2013), the head of the dependency, if it is capitalized, and offset conjunction with the features of previous and succeeding words as features. For the *emotion* category, we use the presence in the NRC dictionary in addition and, for *experiencer*, the presence in a list of English pronouns. We train for 500 iterations with L-BFGS (Liu and Nocedal 1989) and L1 regularization.

The biLSTM-CRF model uses a concatenated output of two biLSTM models (one trained on word embeddings with dimension 300, and one trained on character embeddings from the corpus with dimension 100) as an input to a CRF layer. The word embeddings that we use as input are pre-trained on Wikipedia[7] using *fastText*. We use Adam as activation function, a dropout value of 0.5, and train the model for 100 epochs with early stopping if no improvement is observed after ten consecutive epochs.

The results for this experiment are also shown in Table 8, with the corresponding rows marked with '2' in the 'Exp' column. As results of this experiment show, the recall is low for all predicted categories. Presumably, a major reason, discussed in Section 3, is that substantial numbers of emotion annotations are words or phrases that are not found in the NRC dictionary. On average, only 46 % of emotion annotations are single tokens that can be found in the NRC dictionary,

5 In more detail, we use the inside-outside-beginning (IOB) encoding which is standard in sequence prediction (Ramshaw and Marcus 1995).

6 https://spacy.io/

7 As available at https://github.com/facebookresearch/fastText (Bojanowski et al. 2017).

but for some emotions this number is much lower (only 14 % of *anticipation* cues). For the categories cause and target, their realizations tend to be rather long spans of text (e. g., 94 % of target events are multiword expressions). Faced with the large amount of variability in the training data, the model often abstains from making any predictions whatsoever for these categories. This explains the zero F_1 score for cause prediction with CRF and biLSTM-CRF. We see a somewhat better performance for target prediction with CRFs, which is attributable to the fact that most target relations are triggered by characters, 75 % of which are single tokens.

The highest precision and F_1 across all categories is observed for the *emotion* category with biLSTM-CRF (strict F_1=43 and fuzzy F_1=48). The strict F_1 is by 12 pp higher than predicted with dictionaries and with BOW in text classification experiment.

The *experiencer* category is second best, even though the recall for this category is still very low. This can be explained by the fact that experiencers are expressed in the text mostly as personal pronouns. Since the number of personal pronouns in our texts is relatively low (13 % of all tokens in a sentence triple on average), and only a small fraction of them act as experiencers (<1 % of all tokens in a sentence triple on average), the classifier cannot learn when an entity is an experiencer or not.

4.1.3 Experiment 3: Potential for joint modeling of emotion and role prediction

In the final experiment on REMAN, we analyze if there is a potential for *joint modeling* of relations to improve over learning each relation separately. Joint modeling means that the different parts of the emotion structure are not predicted individually, as is the case in simple models, but at the same time. In this manner, joint models can take into account interdependencies between different parts of the structure. They can be thought of as attempting to arrive at a global understanding, similar to human readers of a text.

To that end, we analyze the potential interactions between predictions with gold labels of all other predictions. Specifically, when training our models, we provide the classifier with the information which sequence of tokens is an experiencer (in the case of emotion phrase prediction) and which sequence of tokens is an emotion (in case of experiencer, cause, and target detection). Since this information is taken from the manual annotations, this does not constitute a 'real' joint model, but a so-called oracle: its results constitute an upper bound for the performance when more knowledge is available.

Recall that the goal of this experiment is to estimate if joint modeling of emotion and roles yields a benefit beyond individual prediction. Table 8 shows that for

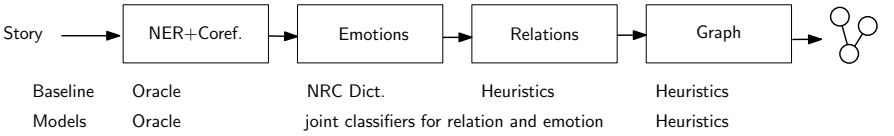


Fig. 5: FANFIC: Models for the emotional relationship prediction. Oracle: a set of character pairs from the gold data.

the *emotion* category, F_1 increases by 5 pp in strict and by 9 pp in fuzzy evaluation if we provide the classifier with the information which sequence of tokens is an *experiencer*. For *experiencer* prediction, F_1 increases by 20 pp in strict and by 22 pp in fuzzy evaluation if we tell the classifier which word or sequence is labelled as emotion.

These results indicate the complementarity of both categories. A qualitative study on a subsample of linguistic properties of emotions and experiencers shows that when the emotion expression and experiencer are parts of the same phrase (verb or adjectival phrase), the emotion word serves as a head to the word that represents an experiencer. Hence, the classifier is able to partially learn that any phrase that is a part of the emotion phrase, whose head is a personal pronoun or a proper name, is a potential *experiencer*. The same applies to *experiencer*: if the head of the governing phrase is an emotion, then the head of the current phrase is a potential *experiencer*. However, due to variability of emotion expressions, this cannot always be the case.

As we have seen, the task of predicting parts of the emotion structure, such as *experiencer*, *cause*, and *target* is a difficult one. In addition to previous observations that informing the classifier about an emotion simplifies the *experiencer* prediction, we have also observed that in many cases characters are experiencing emotions because of other characters. This observation is interesting on its own, as focusing on emotion character relationships potentially adds an interesting angle of analysis to the study of emotions in text. This motivates the focus of the FANFIC model below.

4.2 FANFIC: Emotional Character Relationships

As stated above, the ultimate goal in FANFIC is to predict graphs whose nodes are the characters of a text and whose edges are labeled with emotions. The creation of such graphs requires substantial processing beyond what we have seen for REMAN, as shown in Figure 5: First, references to characters have to be recognized and aggregated (Named Entity Recognition + Coreference). Since our focus

Tab. 9: FANFIC: Example of different indicator conditions. *No-Ind.*: no positional indicators are added. *Role*: uses tags <e> (experiencer) and <c> (cause). *Entity*: uses only tag <et> (entity). The *M-* conditions mask the lexical material (i. e., name) of the entity.

Indicator condition	Example
No-Ind.	Alice is angry with Bob
Role	<e>Alice</e> is angry with <c>Bob</c>
MRole	<e/> is angry with <c/>
Entity	<et>Alice</et> is angry with <et>Bob</et>
MEntity	<et/> is angry with <et/>

is on emotion prediction, we do not automate this step here, but instead rely on gold character annotations. Then, Emotions have to be recognized and mapped onto character pairs (Relations). Finally, all character–emotion relations have to be aggregated into a graph.

We cast the relation detection as a classification task in which each instance consists of two character mentions with up to *n* tokens context to the left and to the right of the character mentions and the classes are the set of emotions. We use a two-layer GRU neural network (Chung et al. 2014) with max and averaged pooling with different variations of encoding the character positions with indicators (inspired by Zhou et al. 2016, who propose the use of positional indicators for relation detection). Our variations are exemplified in Table 9. The goal of an indicator is to mark a character that is either an experiencer or a cause of the expressed emotion in text. We use different encodings of these roles: 'Role' and 'MRole' (masked role) indicators inform the classifier about these roles, while 'Entity' and 'MEntity' (masked entity) indicators do not (they only indicate that marked characters are entities in the relationship). Note that the prediction of directed relations is simpler in the 'Role' and 'MRole' cases, compared to 'Entity' and 'MEntity', as the model has access to gold information about the relation direction.

We obtain word vectors for the deep learning models from GloVe (pre-trained on Common Crawl, d=300, Pennington et al. 2014) and initialize out-of-vocabulary terms with zeros (including the position indicators).

Given that we have comparably limited data on the story-level, we perform cross-story validation, where each story is used as one separate test/validation source. For model selection and meta-parameter optimization, we use 50 % randomly sampled annotations from this respective test/validation instance as a validation set and the remainder as test data.

We evaluate on three different levels of granularity: Given two character mentions, in the instance-level evaluation, we only accept the prediction to be correct if exactly the same mention has the according emotion annotation. We then ag-

Model	Instance	Story	Graph
NoInd	26	25	35
Role	33	33	41
MRole	38 (38)	39 (39)	40 (42)
Entity	23	22	39
MEntity	28	28	39

Tab. 10: FANFIC: Cross-validated results for different models as F_1 scores. 'Instance': aggregated over all instances in the dataset. 'Story': performance averaged over all stories. 'Graph': performance on graph level averaged over all stories. See Table 9 for the examples of the indicator implementation. Results on independent test data shown in brackets.

gregate the different true positive, false positive and false negative values across all stories before averaging to an aggregated score (similar to micro-averaging). On the story-level, we also accept a prediction to be a true positive the same way, but first calculate the result precision/recall/F_1 for the whole story before averaging (similar to macro-averaging). On the graph-level, we accept a prediction for a character pair to be correct without considering the exact position.

Results

Table 10 shows the results on development data and independent test data for the best models. The GRU+MRole model achieves the highest performance on the instance and story levels, and shows a clear improvement over the GRU+NoInd. model. GRU+Role achieves the highest performance on the graph level. As expected, we observe a better performance on a graph level for all models. The absolute numbers, however, are not very high, but the increase from the instance to the graph level shows that constructing the graph is somewhat 'forgiving': Not each individual prediction at the textual level needs to be correct for a correct interaction graph to emerge.

This is shown in practice on Figure 6, which illustrates a fully predicted network from a fan fiction story based on the *Star Wars* universe (Miralana 2015). The error analysis on the predicted network shows that the mistakes made by the model are not immediately obvious. One example is the *trust* relationship between Finn and Rey. Although the textual instance used to classify the interaction contains 'trust' vocabulary ("they could **help** and be **supportive**"), the overall tone suggests that Finn *anticipates* Rey asking for his help rather than directly imposing trust on her. However, as we do not take into account the exact positions, this mistake is still considered a true positive, as a *trust* relationship is present in the gold data. Another example is the *anticipation* relationship between Rey and Leia that is tagged with *sadness* in the gold data. Consider the following text that was used to classify the relationship: "She adored the older woman and enjoyed her company ..., there were certain things that she didn't want to share with her ...." The text implies that though Rey is pious towards Leia, some aspects of their rela-

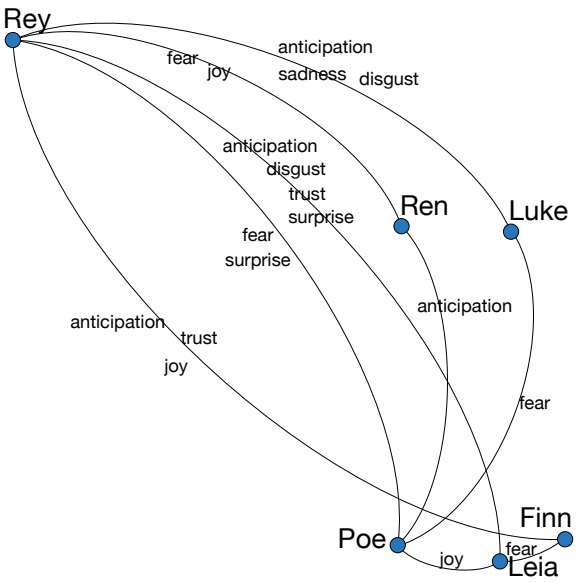


Fig. 6: Example of a predicted network.

tionship do not allow her to be fully open with that woman, hence sadness. The erroneous relationship assignment is then presumably triggered due to specific words such as "adored", "enjoyed" and "share", which often indicate *joy* and *anticipation*. This prediction does not count as true positive, as the gold data does not contain *anticipation* among correct relationship between Rey and Leia.

In general, we find that the sequential and embedding information captured by a GRU as well as additional positional information are all relevant for a substantial performance, at least on the fine-grained emotion prediction task. At the same time, we note that the results presented in this section are based on a setting where names of characters come directly from the annotation. This is an unrealistic scenario, as it is not possible to get character annotations for all books we might be interested in analyzing. We shall address this question in our future work.

5 Discussion, Conclusion, and Future Work

As both the inter-annotator agreement numbers and the results of our computational models show, the tasks of annotating emotions and corresponding roles manually and automatically are both difficult. Contributing factors are the high

lexical variability of emotion expressions (see Table 1) and of the linguistic form of cause and target expressions. At the same time, the resources we present provide useful and valuable insights in the language of emotion expression and, therefore, should be useful to the communities in linguistics, NLP, and literary studies who are interested in the study of textual expressions of emotion.

Developing such resources has its limitations: Due to the subjective nature of emotions, it is extremely challenging to develop annotation guidelines which would lead to annotations with less variation among annotators, in particular if the annotation includes complex structural choices. That is in line with previous research. For instance, both Schuff et al. (2017) and Russo et al. (2011) find that aggregating labels of multiple annotators not by majority vote, but by forming the union, leads to datasets that are, surprisingly, easier to model computationally.

In REMAN, we tackle this problem by employing a multi-step procedure that helps to improve the agreement of the relation annotation. This does not help in the emotion annotation itself, but helps in the role assignment. The introduction of our multi-step annotation procedure lead to an increased inter-annotator agreement for *experiencer* and *cause* annotations by 13 pp and 5 pp in strict evaluation. This indicates that the task seems easier to annotators if they perform role assignment with predefined emotion annotations.

Another difficulty arises from the nature of the texts we work with. Fictional texts are highly metaphoric and full of allusions, which requires thoughtful reading (often reading between the lines) and a global understanding. However, this is something that our annotators cannot develop in the REMAN case: they only have access to one sentence pre- and post-context each. Therefore, it is not always possible to annotate the cause, target, or even the experiencer. This is a trade-off: On the one side, we did not want to annotate full books to cover a range of sources with manageable annotation effort. On the other side, more context might have improved results. Future work will therefore aim at better understanding how to preselect the relevant context that is needed for reliable annotation and secondly use such knowledge for a follow-up annotation project.

Some of the challenges that are posed with the REMAN corpus are addressed in a different way with the FANFIC approach: Here, we formulated the task of emotional character network extraction from fictional texts. We argued that joining social network analysis of fiction with emotion analysis leverages simplifications that each approach makes when considered independently. However, it should be noted that these evaluations are hard to compare, as the tasks and data sets are different.

In ongoing work, we aim at the development of a a real-world application pipeline in which character pairs are not given by an oracle, but rather extracted from text automatically using named entity recognition. To better understand the

relation between instance and graph levels, we explore the best strategy for edge labeling either by a majority vote or accepting the edges with the highest confidence scores. Further, modeling the task in an end-to-end learning setting from text to directly predict the graph, in the spirit of multi-instance learning, is one of the next steps. To that end, we suggest obtaining more gold data with character relations and optimize the pipeline towards the best performance on additional data.

Acknowledgment: This research has been conducted within the CRETA project (http://www.creta.uni-stuttgart.de/) which is funded by the German Ministry for Education and Research (BMBF). This research was partially funded by the German Research Council (DFG), project SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1–1). We thank Laura Ana Maria Bostan and the CRETA consortium for fruitful discussions.

References

Agarwal, Apoorv, Anup Kotalwar, and Owen Rambow (2013). "Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan, pp. 1202–1208.

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat (2005). "Emotions from Text: Machine Learning for Text-based Emotion Prediction". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, BC, pp. 579–586.

Aman, Saima and Stan Szpakowicz (2007). "Identifying Expressions of Emotion in Text". In: *Proceedings of Text, Speech and Dialogue*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 196–205.

Apryl_Zephyr (2016). *Friends*. URL: https://archiveofourown.org/works/8081986 (visited on June 1, 2020).

Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.

Barton, James (1996). "Interpreting character emotions for literature comprehension". In: *Journal of Adolescent & Adult Literacy* 40.1, pp. 22–28.

Benikova, Darina, Chris Biemann, Max Kisselew, and Sebastian Pado (2014). "GermEval 2014 Named Entity Recognition Shared Task: Companion Paper". In: *Workshop Proceedings of the 12th edition of the KONVENS conference*. Hildesheim, Germany, pp. 104–112.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Bostan, Laura Ana Maria, Evgeny Kim, and Roman Klinger (2020). "GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception".

In: *Proceedings of the 12th International Conference on Language Resources and Evaluation*. Marseille, France, pp. 1547–1559. (Visited on June 1, 2020).

Burkitt, Ian (1997). "Social relationships and emotions". In: *Sociology* 31.1, pp. 37–55. URL: https://www.jstor.org/stable/42855768?seq=1 (visited on June 1, 2020).

Chen, Ying, Wenjun Hou, Xiyao Cheng, and Shoushan Li (2018). "Joint Learning for Emotion Classification and Emotion Cause Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 646–651.

Cheng, Xiyao, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou (2017). "An emotion cause corpus for chinese microblogs with multiple-user structures". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 17.1, p. 6.

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *Proceedings of the Deep Learning and Representation Learning Workshop at NIPS 2014*. Montreal, Canada.

Ding, Zixiang, Huihui He, Mengran Zhang, and Rui Xia (2019). "From Independent Prediction to Reordered Prediction: Integrating Relative Position and Global Label Information to Emotion Cause Identification". In: *The Thirty-Third AAAI Conference on Artificial Intelligence*. New Orleans, LA, pp. 6343–6350.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition & Emotion* 6.3-4, pp. 169–200.

Ellsworth, Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó (2004). "PropBank, SALSA and FrameNet: How Design Determines Product". In: *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora at LREC*. Lisbon, Portugal.

EmmyR (2014). *PianoP*. URL: https://archiveofourown.org/works/2481311 (visited on June 1, 2020).

Fillmore, Charles J., Miriam R.L. Petruck, Josef Ruppenhofer, and Abby Wright (2003). "Framenet in Action: The Case of Attaching". In: *International Journal of Lexicography* 16.3, pp. 297–332.

Gaelick, Lisa, Galen V. Bodenhausen, and Robert S. Wyer (1985). "Emotional communication in close relationships." In: *Journal of Personality and Social Psychology* 49.5, p. 1246.

Gao, Kai, Hua Xu, and Jiushuo Wang (2015). "A rule-based approach to emotion cause detection for Chinese micro-blogs". In: *Expert Systems with Applications* 42.9, pp. 4517–4528.

Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz (2015). "Detecting emotion stimuli in emotion-bearing sentences". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. Cairo, Egypt, pp. 152–165.

Goethe, Johann Wolfgang von (1774). *Die Leiden des jungen Werthers*. URL: http://www.deutschestextarchiv.de/book/show/goethe_werther01_1774 (visited on June 1, 2020).

Gui, Lin, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du (2017). "A Question Answering Approach for Emotion Cause Extraction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, pp. 1593–1602.

Gui, Lin, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou (2016). "Event-Driven Emotion Cause Extraction with Corpus Construction". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pp. 1639–1649.

Gui, Lin, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou (2014). "Emotion cause detection with linguistic construction in chinese weibo text". In: *Natural Language Processing and Chinese Computing*. Springer, pp. 457–464.

Hogan, Patrick Colm (2015). "What Literature Teaches Us About Emotion: Synthesizing Affective Science and Literary Study". In: *The Oxford Handbook of Cognitive Literary Studies*. Ed. by Lisa Zunshine. Oxford University Press. Chap. 13, pp. 273–290.

Honnibal, Matthew (2013). *A Good Part-of-Speech Tagger in about 200 Lines of Python*. Online: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991*.

Hugo, Victor (1885). *Les Misérables*. Only accessible outside of Germany. Project Gutenberg: http://www.gutenberg.org/ebooks/135. (Visited on June 1, 2020).

Ingermanson, Randy and Peter Economy (2009). *Writing fiction for dummies*. Indianapolis, IN: John Wiley & Sons.

Johnson-Laird, Philip Nicholas and Keith Oatley (1989). "The language of emotions: An analysis of a semantic field". In: *Cognition & Emotion* 3.2, pp. 81–123.

Joyce, James (1914). *Dubliners*. London: Grant Richards.

Kidd, David Comer and Emanuele Castano (2013). "Reading literary fiction improves theory of mind". In: *Science* 342.6156, pp. 377–380.

Kim, Evgeny and Roman Klinger (2018). "Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, pp. 1345–1359.

Kim, Evgeny and Roman Klinger (2019). "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies". In: *Zeitschrift für Digitale Geisteswissenschaften* 4. DOI: 10.17175/2019_008.

Kim, Evgeny, Sebastian Padó, and Roman Klinger (2017). "Investigating the Relationship between Literary Genres and Emotional Plot Development". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver, BC, pp. 17–26.

Klinger, Roman, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur (2018). "IEST: WASSA-2018 Implicit Emotions Shared Task". In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium, pp. 31–42.

Klinger, Roman and Katrin Tomanek (2007). "Classical Probabilistic Models and Conditional Random Fields". Tech. rep. TR07-2-013. ISSN 1864-4503, Technical Report. Department of Computer Science, Dortmund University of Technology.

Lafferty, John, Andrew McCallum, and Fernando Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the International Conference on Machine Learning*. Williamstown, MA, pp. 282–289.

Li, Weiyuan and Hua Xu (2014). "Text-based emotion classification using emotion cause extraction". In: *Expert Systems with Applications* 41.4, pp. 1742–1749.

Liew, Jasy Suet Yan, Howard R. Turtle, and Elizabeth D. Liddy (2016). "EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 1149–1156.

Liu, Dong C. and Jorge Nocedal (1989). "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3, pp. 503–528.

Mar, Raymond A, Keith Oatley, and Jordan B Peterson (2009). "Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes". In: *Communications* 34.4, pp. 407–428.

Miralana (2015). *What's so strange about a down-home family romance?* URL: https : / / archiveofourown.org/works/5474927 (visited on June 1, 2020).

Mohammad, Saif (2012a). "# Emotional tweets". In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. Montréal, QC, pp. 246–255.

Mohammad, Saif (2012b). "From once upon a time to happily ever after: Tracking emotions in mail and books". In: *Decision Support Systems* 53.4, pp. 730–741.

Mohammad, Saif and Peter Turney (2013). "Crowdsourcing a word–emotion association lexicon". In: *Computational Intelligence* 29.3, pp. 436–465.

Mohammad, Saif, Xiaodan Zhu, and Joel Martin (2014). "Semantic Role Labeling of Emotions in Tweets". In: *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Baltimore, Maryland, pp. 32–41.

Nalisnick, Eric T and Henry S Baird (2013). "Extracting sentiment networks from Shakespeare's plays". In: *Proceedings of the 12th International Conference on Document Analysis and Recognition*. Washington, DC, pp. 758–762.

Oatley, Keith (2002). "Emotions and the story worlds of fiction". In: *Narrative impact: Social and cognitive foundations* 39, p. 69.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, pp. 1532–1543.

Piper, Andrew, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths, and Hardik Vala (2017). "Studying Literary Characters and Character Networks". In: *Digital Humanities 2017: Conference Abstracts*. Montreal, Canada, pp. 119–122.

Plutchik, R. (2001). "The Nature of Emotions". In: *American Scientist* 89.4, pp. 344–350.

Ramshaw, Lance and Mitch Marcus (1995). "Text Chunking using Transformation-Based Learning". In: *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, MA, pp. 82–95.

Robinson, Jenefer (2005). *Deeper than reason: Emotion and its role in literature, music, and art*. Oxford University Press on Demand.

Russell, James A and Lisa F Barrett (1999). "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." In: *Journal of Personality and Social Psychology* 76.5, pp. 805–819.

Russo, Irene, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco (2011). "Emocause: an easy-adaptable approach to emotion cause contexts". In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 153–160.

Scarantino, Andrea (2016). "The philosophy of emotions and its impact on affective science". In: *The handbook of emotions*, pp. 3–65.

Scherer, Klaus R. and Harald G. Wallbott (1994). "Evidence for universality and cultural variation of differential emotion response patterning." In: *Journal of Personality and Social Psychology* 66.2, p. 310.

Schuff, Hendrik, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger (2017). "Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus". In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark, pp. 13–23.

Stimson, Frederic Jesu (1943). *The King's Men: A Tale of Tomorrow*. Project Gutenberg: http : //www.gutenberg.org/ebooks/18960. (Visited on June 1, 2020).

Strapparava, Carlo and Rada Mihalcea (2007). "SemEval-2007 Task 14: Affective Text". In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic, pp. 70–74.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie (2005). "Annotating Expressions of Opinions and Emotions in Language". In: *Language Resources and Evaluation* 39.2, pp. 165–210.

Wiebe, Janyce M. (2000). "Learning Subjective Adjectives from Corpora". In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. Austin, TX, pp. 735–740.

Xia, Rui and Zixiang Ding (2019). "Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 1003–1012.

Xia, Rui, Mengran Zhang, and Zixiang Ding (2019). "RTHN: A RNN-Transformer hierarchical network for emotion cause extraction". In: *arXiv preprint arXiv:1906.01236*.

Xu, Bo, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu (2019). "Extracting Emotion Causes Using Learning to Rank Methods From an Information Retrieval Perspective". In: *IEEE Access* 7, pp. 15573–15583.

Xu, Ruifeng, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui (2017). "An ensemble approach for emotion cause detection with event extraction and multi-kernel SVMs". In: *Tsinghua Science and Technology* 22.6, pp. 646–659.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann (2013). "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria, pp. 1–6.

Zhou, Peng, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu (2016). "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany, pp. 207–212.

Martin Baumann, Steffen Koch, Markus John, and Thomas Ertl

Interactive Visualization for Reflected Text Analytics

Abstract: In this chapter, we discuss the contribution of visualization research to the CRETA project. We give a definition of visualization as we conduct it as partners within the context of the project. Then, we introduce the opportunities and challenges of collaboration in interdisciplinary projects involving visualization research and delineate a number of conflicting requirements that need to be balanced in such a setting. The workflow pipeline that is defined by the concept of reflected text analytics can be separated in two phases, roughly subsumed under the notions 'annotation' and 'interpretation'. For each of these phases, we present one approach that supports the respective analysis tasks by means of interactive visualization: an approach dealing with text annotation data by multiple annotators and an approach dealing with networks of characters in narrative texts. We describe the representation and interaction features of these approaches and exemplify their workings in a series of usage scenarios. Finally, we assess our experiences in developing these approaches with respect to the balancing of requirements mentioned above.

Zusammenfassung: In diesem Kapitel erörtern wir den Beitrag der Visualisierungsforschung zum CRETA-Projekt. Wir definieren den Begriff von Visualisierung, so wie wir sie als Partner im Projektkontext betreiben. Dann besprechen wir die Chancen und Herausforderungen bei der Zusammenarbeit in interdisziplinären Projekten, die Visualisierungsforschung einschließen, und wir umreißen eine Reihe von widersprüchlichen Anforderungen, die in einem solchen Umfeld austariert werden müssen. Der Arbeitsablauf, der durch das Konzept der reflektierten Textanalyse definiert wird, kann in zwei Phasen unterteilt werden, die sich in etwa durch die Begriffe 'Annotation' und 'Interpretation' fassen lassen. Für jede dieser Phasen stellen wir jeweils einen Ansatz vor, der die jeweiligen Analyseaufgaben durch interaktive Visualisierung unterstützt: einen Ansatz, der sich mit Textannotationen von mehreren Annotatoren befaßt und einen Ansatz, der sich mit Figurenkonstellationen in narrativen Texten befasst. Wir beschreiben die Repräsentations- und Interaktionsmerkmale dieser Ansätze und exemplifizieren ihre Mechanismen anhand einer Reihe von Nutzungsszenarien. Schließlich

Martin Baumann, Steffen Koch, Markus John, Thomas Ertl, Institute for Visualization and Interactive Systems, University of Stuttgart

Open Access. © 2020 Martin Baumann, Steffen Koch, Markus John und Thomas Ertl; published by De Gruyter
BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license.
https://doi.org/10.1515/9783110693973-012

bewerten wir unsere Erfahrungen beim Entwickeln dieser Ansätze hinsichtlich des oben erwähnten Austarierens von Anforderungen.

1 Ways of Collaboration

In this section, we expound our view on the opportunities and challenges that arise in collaborative research projects between visualization scientists and scholars from text related disciplines (which, in the following are referred to as 'domains'). Which kinds of research tasks on which kinds of data in these domains could be meaningfully supported by visualization approaches? What characterizes a domain problem as a potentially rewarding subject for visualization research? What do visualization researchers bring to the table, what can be expected from them in terms of expertise and possible solutions – and what not? Figure 1 schematically summarizes the relation between visualization research and domain research in general and within the CRETA project in particular.

1.1 A Definition of Visualization

In order to tackle these questions, we first give a definition of visualization as we conduct it as partners within the context of the CRETA project. In doing so, we build upon the definitions given by Card et al. (1999) and Munzner (2014). We consider visualization as a discipline of computer science, where we conceptualize, implement and evaluate approaches to automatically create interactive, visual representations of abstract data. These representations are designed with the goal to amplify human cognition in order to carry out exploration tasks on large data sets more effectively. In the following, we further unpack and comment some parts of this definition.

By 'exploration tasks' we mean research tasks with the goal to detect structures within a data set and formulate hypotheses about their workings. 'Confirmatory tasks' that seek to validate or falsify such hypotheses also fall into the reach of our field – even though they are attached less importance – whereas the support of 'presentation tasks' that seek to visually communicate facts that are already known about a data set and its structure (e. g. like in the case of newspaper infographics) is not of interest for our research agenda.

The task of free exploration of a mostly unknown, large data set is characteristic for those situations in which visualization can be useful. In the case of a clear-cut research question that can be solved by a purely algorithmic approach, or if the (textual) data set is so small that the respective question can be answered

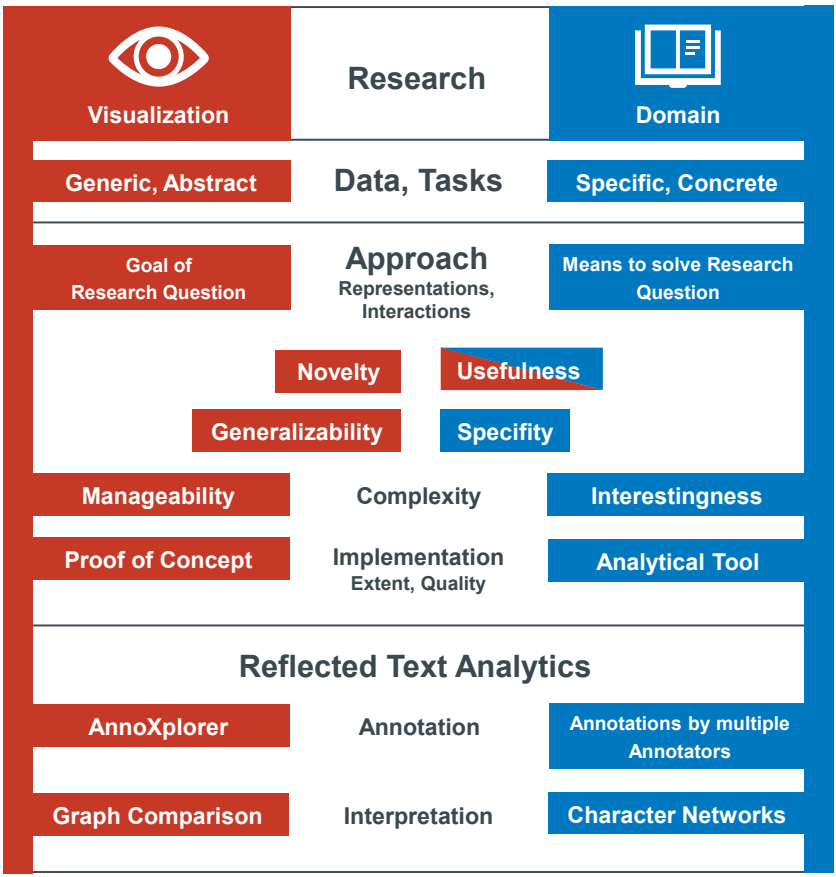


Fig. 1: Schematic comparison of visualization research and domain research with respect to data and tasks, the developed approaches, and the problems of reflected text analytics

by a thorough close reading, then visualizing the data would create a potentially error-prone overhead. However, for an ill-specified problem, whose tackling requires human world knowledge, intuition, experience, and ingenuity, and that is to be pursued on a data set that is too large in order to handle it without computational support, in such a situation, visualization can provide a means to leverage the human visual apparatus and decision-making capabilities.

In connection with these considerations, the notion of 'effectiveness' is also important. In contrast to other instances of crafting visual artifacts, visualization as understood here should produce approaches that aspire not primarily to aesthetic quality, but that put this quality in the service of the higher goods of correctness and accuracy (see Munzner 2009).

The broad notion of visualization as a computer science discipline is commonly further separated into the fields of 'scientific visualization', 'information visualization', and 'visual analytics'. Without going into further details, we want to point out a basic differentiating characteristic of these fields. In scientific visualization, data are dealt with that have inherent spatial relations (e. g. the air flow behavior along the surface of a car). In information visualization and visual analytics, however, data are dealt with that are 'abstract' in the sense that they lack such inherent spatial relations. Hence, in these fields, a spatial metaphor has to be invented (and learned by the viewer) in order to visually represent the data (e. g. in the case when numerical values are depicted as a bar chart). This abstractness is given in particular for textual data. Hence, when speaking of 'visualization' in the following, we refer jointly to the fields of information visualization and visual analytics.

Finally, we would like to emphasize the difference between our visualization research and the research done by domain experts who employ these visualizations to solve a problem in their domain. Thus, while we do research on visualization approaches for these domains, we do not conduct research in the domains themselves. Of course, however, we need to understand the domain tasks in order to conceptualize and implement a visualization solution, and we may need to conduct analysis tasks in order to evaluate our approaches.

1.2 The Interplay of Visualization and Application Domains

Taking up this last point, we now elaborate further on the relation between visualization researchers and domain experts. Munzner (2009) differentiates between, on the one hand, data and tasks that originate in the respective domain and that are described in domain terms and, on the other hand, a set of more generic data and tasks that are described in the vocabulary of the visualization discipline. She

explains that one of the first steps in the creation of a visualization would be to map the raw domain data to their visualization-side abstractions (e. g. by binning or thresholding), and to map the concrete, specific domain tasks to more abstract visualization tasks, which can be organized in a series of general taxonomies. A visualization researcher's goal could then consist in conceiving a visualization approach for dealing with these abstracted data and tasks.

The quest for the generalizability of the developed approach, that is discernible in this process, is characteristic for research efforts in computer science in general and visualization in particular. However, for many domains – especially from the humanities – the data, the analysis tasks, and the approach required to deal with them appear very specific to the respective guiding domain research question at hand. This disparity raises problems for collaborative projects. For the visualization side, it can be unrewarding and expensive to build approaches, that can deal only with one specific set of data and tasks. What visualization researchers typically want is to publish the concept of an approach that is primarily new and secondarily can be proven as useful for the domain problem at hand and potentially for many more related abstract task / data combinations. The domain experts, however, are primarily interested in a useful tool that helps them to solve their problems, and often times, an implementation of a custom-tailored combination of old and time-proven visualization techniques may be the best suited option for their needs.

With domain problems from the humanities, another difficulty can arise from the fact that when problems begin to get interesting for the domain expert, they may be already very hard for the visualization researcher to support, leaving only a small (or sometimes empty) intersection of prospective subjects for collaboration.

Finally, the usage of a visualization prototype for the purpose of design evaluation generally puts lower demands on the implementation quality (in terms of stability, speed, longevity, versatility, accessibility, ease of use etc.) than the usage for the purpose of conducting real, long-term analyses. Furthermore, there may be potential features whose implementation could be very useful for an analyst but that contribute little to nothing to the visualization core idea of the approach at hand. The additional effort required to condition an initial prototype to productive use by a domain expert can be substantial and may provide only little benefit for the visualization researcher in terms of publishable results.

These conflicting requirements of generalizability vs specificity, of different goals with respect to novelty vs usefulness, of what can be done (visualization) and what should be done (domain), and of different expectations with respect to the implementation quality of the visualization approach, can substantially narrow the range of collaboration opportunities. Van Wijk (2006) speaks in this con-

text of an 'interest gap' between visualization researchers and domain experts. He goes on to describe a user-centered design approach as the preferable way to deal with this problem. The problems touched in this section are also discussed more recently and specifically from a digital humanities / visualization perspective by Jänicke (2016).

1.3 Visualization within CRETA

In the following sections, we describe two examples of visualization approaches that we designed in the frame of the CRETA project. In doing so, we sought to come up with new and generalizable approaches that are useful for the tasks in the project context as they are laid out in the chapter by Pichler and Reiter (2020), p. 43 ff. in this volume. Furthermore, we sought to implement them in a way that they can be used productively and be further extended in the process of this usage; to this end, both examples are implemented to be used in a web browser.

We support the workflow depicted in Figure 1 on page 44 in the phase of the phenomenon recognition / annotation as well as in the phase of the analysis and interpretation of these data under the guiding questions of the text-related domain disciplines. For the former phase, we developed an approach to support the creators and analysts of annotations in the visual exploration of large annotation data sets (see Section 2). This approach was published in Baumann et al. 2020[1]. For the latter phase, we developed an approach for the visual analysis of the temporal development of a text's network of characters (see Section 3). This approach was published in John et al. 2019.

An essential common feature of both approaches is the close interlocking of the respective visualization (distant views) with the text itself (close views). As mentioned in the chapter by Pichler and Reiter (2020), this allows for the consideration of annotated phenomena within their textual context and for the combined interpretation of quantitative results together with a detailed analysis of particular text passages.

2 Visualization of Annotations

As it was mentioned in the chapter by Pichler and Reiter (2020), the exploratory analysis of text annotations is an important task at several points of the CRETA

1 Presented at IVAPP 2020, http://www.ivapp.visigrapp.org/?y=2020

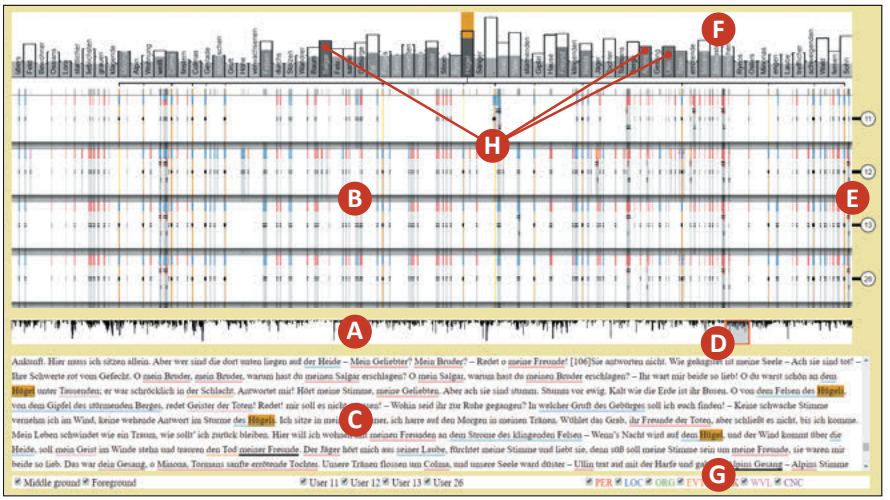


Fig. 2: The approach with Goethe's *Die Leiden des jungen Werthers*. The overview window (A) with the brush at (D). The annotation window in transit mode with four annotator bands at (B), the annotator IDs at (E), the annotated types' distribution at (F), and a series of name types at (H) (see Section 2.2). The text window (C) with underlined annotations. The filter controls at (G). Interaction: Hovering on the type bar 'Hügel' highlights the respective tokens in the annotation and text windows.

workflow. In the following, we understand a text annotation as a piece of meta information that links a list of tags to a segment of text (i. e. a sequence of tokens). In general, such data can pose severe scaling challenges within multiple dimensions of complexity: The source texts can be long, up to hundreds of thousands of tokens, and bear annotations in similar orders of magnitude. Then – as in the case of the CRETA project – there may be multiple annotators, human or algorithmic, who possibly disagree in their assessments. And finally, the annotation themselves can be complex: They may contain multiple tags, or text segments that overlap or whose lengths vary substantially.

Domain tasks that are to be performed on such data can encompass: Analyze the disagreement between the contributing annotators in order to refine and disambiguate the team's annotation guidelines; compare annotations generated by competing algorithms in order to assess systematic differences; or consider high annotator disagreement as an indicator of textual complexity that may warrant further investigations regarding possible patterns in the annotation data and how these patterns may relate to the source text. These tasks can involve a comparison across multiple annotators, the detection of patterns with respect to the annotations' extensions or tags, an analysis of the frequency- and position distribution

of the annotated tokens, or a consideration of annotation segments within their surrounding source text.

Our contribution to deal with such data sets and analysis tasks is an approach to visually and interactively browse text annotations. As can be seen in Figure 2, it consists of three interconnected views: an overview of the whole text (*overview window* at A), a view upon a portion of the annotated text strings (*text window* at C), and an abstract representation of the annotation data in some freely chosen granularity (*annotation window* at B). The latter constitutes the main component, and it shows the data simultaneously in two dimensions: vertically across all contributing annotators and horizontally across the whole token sequence of a freely selected passage of the underlying text. The core concept that we employ in order to deal with the data type's complexity dimensions mentioned above is a combined geometric / semantic zooming mechanism; it governs all other aspects of representation and interaction. We will describe it in Section 2.1 and exemplify its workings in Section 2.2.

We now give a few examples of related approaches that deal with task / data situations similar to ours. For a thorough discussion of the current research embedded into the wider area of text visualization, we refer to Baumann et al. (2020). While these approaches deal with one or more of the complexity dimensions of annotation data that we mentioned in above, none of them encompasses all at the same time. In Table 1, we list them under the aspect as to which of the features they offer that seem important to us in order to support the tasks laid out above.

Correll et al. (2011) present an approach that supports scholars in exploring densely annotated texts. They enhance the scalability by employing a focus+context approach to minimize contexts of less importance with respect to the chosen focus and by allowing to filter dense occurrences of interesting annotations by means of a line chart based abstraction. In the *NEREx* approach, El-Assady et al. (2017) analyze named entities by combining a text with annotation highlights, an abstraction of the text's sentences as lines together with the annotated entities as sequences of glyphs, and a series of node link diagrams that encode relational aspects of the entities. The *VarifocalReader* approach (Koch et al. 2014) is close to ours in that smooth transitional interactions between overview and detail views are supported. However, VarifocalReader does not aim at preserving a consistent representation over different levels of abstraction, which we see as an important benefit of our approach for the depicted tasks. Chandrasegaran et al. (2017) present in their visual analytics approach for the support of open coding of natural language texts a series of interconnected views, that closely resemble some of the views we present: a text view with highlights on tokens, a series of overviews showing meta data in the whole text, and an interactive word cloud. They also use similar means of view coordination, allowing the syn-

Tab. 1: Feature comparison of annotation approaches. ○ = unsupported, ◐ = partially supported, ● = supported. Column labels: Range = continuous range of granularities, Annot. = comparison of multiple annotators, Config. = segment overlap or multiple tags, Text = interconnected text view, Token = distribution of annotated tokens, Edit = create and change annotations.

	Range	Annot.	Config.	Text	Token	Edit
Correll et al. 2011	◐	○	○	●	○	○
Widlöcher and Mathet 2012	◐	◐	◐	●	●	●
Landragin et al. 2012	○	○	◐	●	●	●
Koch et al. 2014	◐	○	○	●	●	◐
Meister et al. 2016	○	○	●	●	●	●
Eckart de Castilho et al. 2016	○	◐	◐	●	○	●
Chandrasegaran et al. 2017	◐	○	○	●	●	◐
El-Assady et al. 2017	◐	○	○	●	●	◐
Kleymann et al. 2018	◐	○	○	●	○	○
Baumann et al. 2020	●	●	●	●	●	○

chronization of positional and pattern information across views. However, they neither offer a continuous range of granularities for the annotation abstractions, nor visual support for the display of multiple annotators. The *CATMA* annotation tool uses stacked, colored underlines to encode annotations within the running text (Meister et al. 2016). Its line spacing can be stretched in order to make room for overlapping annotations or annotations with multiple tags whose segments can even be discontinuous. The *Glozz* annotation tool, like most of its competitors, uses highlights on the text to encode annotations (Widlöcher and Mathet 2012). Overlapping segments are shown as overlapping highlights – an approach that does not scale as well as CATMA's stacked underlines. An uncommon feature is the aligner view of Glozz. Here, different annotators are arranged in a series of rows, and their annotations are strung as colored bars on a line that represents the text. Quite a unique idea for the encoding of annotations is followed by the design study of Kleymann et al. (2018). Besides using an underlined text, they encode each annotation as a glyph consisting of a vertical bar (representing the text), and a colored horizontal bar (representing the tag, the segment length and the position of the annotation). These glyphs can be arranged freely on a canvas, and upon hovering, a tooltip shows the immediate surrounding text. Further examples of annotation tools employing similar means would be *ANALEC* (Landragin et al. 2012), or *WebAnno* (Eckart de Castilho et al. 2016).

2.1 AnnoXplorer: Representations and Interactions

The overview window (see Figure 2A) contains a bar chart representing the whole text as a sequence of token bins. Each bar corresponds to a bin, and its height indicates the number of annotations that any of the bin's tokens takes part in. The user can now select a text passage by brushing over a sequence of token bins (see the red rectangle at Figure 2D), and the corresponding annotation data will then be displayed in the annotation window and the text will be focused in the text window.

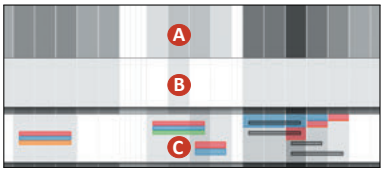


Fig. 3: Cut-out of the annotation window with three annotators. The band of the first annotator is hidden; only the common background layer is shown at (A). Also hidden is the foreground layer in the band of the second annotator at (B). For the third annotator at (C), seven annotations can be seen in the foreground layer.

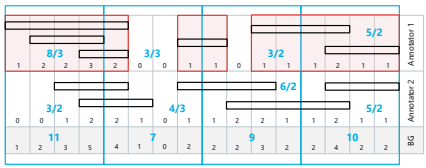
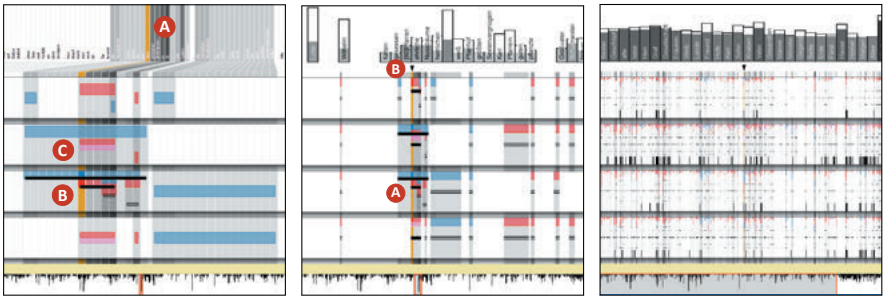


Fig. 4: Schematic example for two annotators (upper and middle row). Chunks for annotator 1 are highlighted in red; bins framed in blue. Annotation counts are given per token (black figures) and per bin (blue figures: annotation count / cut annotations). Lower row: counts for the background layer.

The objects in the annotation window represent the selected tokens and the annotations thereupon. While their exact mode of representation depends on the amount of the selected text, these objects are always organized in the same way: They live in three stacked layers, shown in Figure 3 where the upper layers are partially hidden in order to better show the lower layers. The objects in the lower and middle layer represent tokens or bins of tokens and form the basic grid of the window. They are concerned with counting how many annotation segments a token is contained in and encode this information as values along a gray scale. An example of how annotations are counted is given in Figure 4. The objects of the upper layer represent mainly annotation segments. They are concerned with displaying the segment extensions and tags and employ a color encoding in doing so. Together, the objects in the middle and upper layers represent the annotations of each contributing annotator. Per annotator, they form horizontally stretched bands over the whole window's width (see Figure 2B), and all of these bands (which are freely vertically arrangeable) float over the lower background layer that represents the accumulated information over all annotators. There is a series of

advantages to this way of abstractly representing annotations: First, since there is always a uniform grid of tokens or bins, annotations can be immediately compared across all annotators. Second, since the token strings' lengths are not taken into account, the visual weight of an annotation depends only on the number of tokens that its extension contains. And third, these geometric objects – as opposed to text strings – can be easily stretched or transformed, as we will explain next in connection with our geometric / semantic zooming mechanism.

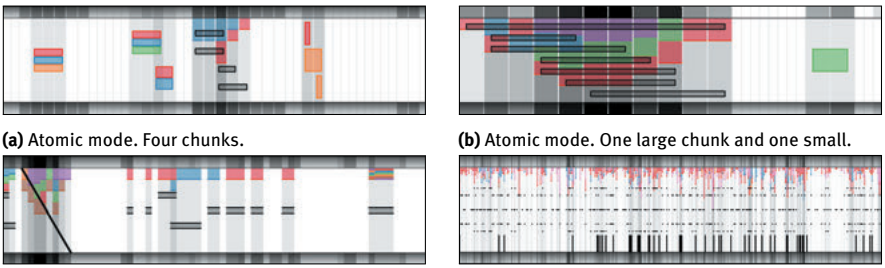


(a) Atomic mode. Hover at (B) **(b)** Transit mode. Hover in text. **(c)** Aggregated mode. Hover in text

Fig. 5: Zooming out from left to right. A text passage containing the token 'dem' (at the yellow highlight / triangle) in the three zooming modes.

There are three major modes of representation with respect to the number of tokens selected by the brush; henceforth, we call them *atomic*, *transit* and *aggregated* mode (see Figure 5 for a juxtaposition of them). In the atomic and transit modes, the vertical grid objects of the lower and middle layers represent tokens. The main difference between these two modes is that the atomic mode applies when the selection brush is small enough that we can display the text string for each token legibly above the annotator bands (see Figure 5a at A). In transit and aggregated mode however, these token strings are replaced by a depiction of the annotated types, that we explain later. Another commonality of the transit and aggregated modes that distinguishes them from the atomic mode is that the grid units are of uniform width and cannot be generally discerned as discrete objects. When the selection brush gets larger than the tool window's width contains pixels, the aggregated mode applies, and here the grid units no longer represent tokens but bins of tokens. Within each mode, brushing (or zooming) is a continuous experience since only the selection and size of the shown elements changes (geometric zoom); when crossing a mode border, however, the representation of the annotation information is adjusted in a way such as to optimize the available screen space for the display of the selected data (semantic zoom). In this way, the user

can get an overview of a text of arbitrary length without the need for scrolling (aggregated mode), continuously drill down to a detailed view upon any short selection of text (atomic mode), and always be presented the annotation information of the current selection with the finest resolution at hand.



(a) Atomic mode. Four chunks.

(b) Atomic mode. One large chunk and one small.

(c) Transit mode. The second chunk is too large for displaying the stack of annotation extensions.

(d) Aggregated mode. The black dots (up to four) indicate the number of annotations cut by a bin.

Fig. 6: Cut-outs of annotator bands in the different zooming modes. In atomic mode, the representation of chunks with less than four annotations is not split.

The problem of displaying constellations where many annotation segments overlap and contain multiple tags is dealt with in different ways in the three zooming modes. In atomic mode and with little overlap, an annotation is encoded as a rectangle spanning its segment's tokens, which is furthermore subdivided into a series of sub-rectangles that are colored according to the annotation's tags (see Figure 5a at C and Figure 6a). For dealing with massive overlap, we introduce the notion of 'chunks', which are chains of annotations by a specific annotator that are linked through their overlap (see Figure 4). When a chunk is too large for the whole stack of its annotation rectangles to be displayed in the available band height in the manner described above, the representation of annotations is split (see Figure 5a at B, Figure 6b): Vertical colored bars on each token encode the number and tags of annotations on this token, whereas thin gray rectangles (or a diagonal line, if a second limit of the stack size is exceeded) indicate the annotations' extension. For the transit mode, we chose to exclusively apply this split representation (see Figure 6c). Finally, in the aggregated mode, the colored bars of the split representation are now used to encode the corresponding information for whole bins, and above them, the number of annotation segments that are cut by a bin is encoded as a number of black dots (see Figure 5c and Figure 6d).

As mentioned above, in the transit and aggregated modes, there is not enough horizontal space to show the text strings of all selected tokens in the upper part of the annotation window. What we do instead is to represent the spatial and

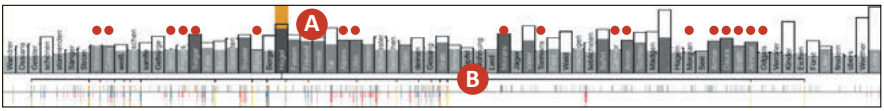


Fig. 7: Types in aggregated mode. Interaction: Hover on 'Hügel' next to (A) highlights the type bar and shows 20 local bin positions in the text selection (tick marks at B). Red dots (added for reference) mark names (see Section 2.2).

frequency distribution of the most annotated types within the selection (see Figure 2F and Figure 7). First, in the set of all tokens that are part of some annotation (globally), we create a partition by subsuming all tokens that share a common lemma under this term. For each of these type terms, we then sum up the annotation counts of each of its tokens (the *global type count*), and for each type that contains tokens in the current selection (the *local types*), we sum up the annotation counts of each of its *local tokens*, i. e. the selected tokens, which gives us the *local type count*. In aggregated mode, we furthermore look for those bins, that contain at least one local token (the *local bins*). Next, we sort the local types according to their local and their global type count, select a number of types to be displayed from the top, and compute the ideal middle position of their bars as the median of the middle positions of all of their local tokens or bins. We now go through this sorted selection of types and place each type bar as closely as possible to its ideal position. The bars themselves are comprised of three elements: The gray portion of the bar encodes in its height and in its lightness the local type count. The black frame (which contains a gap for long type strings) encodes in its height the global type count. And finally, the type string, which is shortened to an ellipsis for long strings, is shown in a manner to provide maximal contrast: black on the white background and black or white on the bar, depending on its gray value. Finally, below the type bars at Figure 7B, a series of tick marks indicate upon mouse hover the number and position of a type's local tokens or bins. This representation of types allows the user to analyze the frequency- and position-distributions of the most annotated tokens in selected passages and compare them to the respective global values.

As mentioned in Section 1.3, an essential part of our approaches is to provide the user with the opportunity to consider the textual data (here: the annotations) within the surrounding text for close reading. In the text window, the text is shown together with annotation marks, and it is connected with the other windows by means of the mechanisms of *scrolling*, *highlighting* and *freezing*. As for scrolling, the window can be scrolled manually, but after altering the brush selection, the most relevant portion of the text is scrolled into view automatically. As for the highlights, they consist of static parts in the text window (colored underlines) and

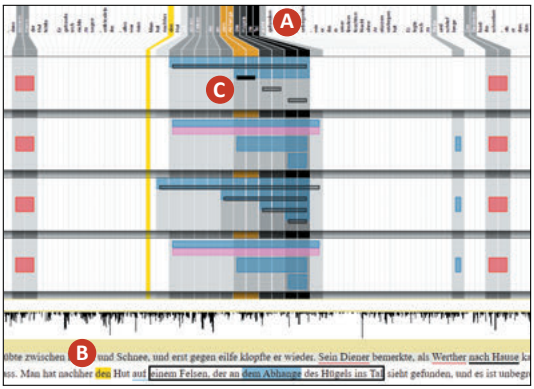


Fig. 8: Highlighting and freezing interaction. Hover and click the annotation at (C), then hover the token 'den' (B).

responsive parts in the text and annotation windows that are interconnected by means of mouse hover actions. The central problem with highlighting in the text is that a token can carry multiple annotations by multiple annotators with multiple tags. Hence, we have to differentiate between hovering over elements that are assigned to a specific annotator and elements that are not. For details on how this is done in the different zooming modes, we refer to Baumann et al. (2020), but in Figure 8 we show a simple example in atomic mode: Here, the static underline encompasses the union of all annotators' chunks: 'auf einem Felsen, der an dem Abhange des Hügels ins Tal'. It is black on all tokens except the first, since the tags are not unique across all annotators. Hovering on the annotation rectangle at (C) darkens it, extends the underline in the text window to a frame, but only on the annotator's chunk, and highlights the annotation extension ('dem Abhange') with the tag color. Hovering on the token 'den' at (B) however, highlights the participating objects of all annotators in the annotation window. Finally, with freezing we mean that the user can lock or unlock the highlights by clicking on an element. Hovering over other elements may then add further highlights, and this state is not lost by changing the selection or by entering into a different zooming mode.

2.2 Usage Scenarios: Annotations of *Werther*

Next, we present two usage scenarios operating on data from the CRETA project. The source text of about 42 000 tokens, the 18th century novel *Die Leiden des jungen Werthers* by Johann Wolfgang von Goethe, was annotated by Sandra Murr (see Ketschik et al. 2020, p. 204 ff. in this volume) and her team of annotators for en-

tities in the categories of 'person', 'location', 'organization', 'event', 'work', and 'concept' (about 6500 annotations).

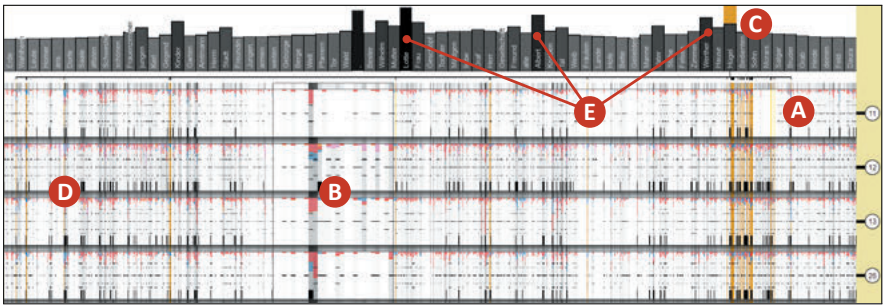


Fig. 9: The *Werther* data set. Fully zoomed out annotation window in aggregated mode. Interaction: The type 'Hügel' at (C) is hovered and the highlights are frozen; at (B), the lens is brought up. Sparse region to the left of (A), dense region at (D).

The first task considered here is to compare the annotations' tags and extensions across the four annotators and to detect and analyze patterns of annotation. From Figure 9 – showing the whole data set in aggregated mode – we can already deduce some general facts: Mainly 'person'-entities were annotated (red), the largest type bars correspond to the novel's main characters 'Lotte', 'Albert', and 'Werther' (see at E), the marks of annotator no. 11 are more sparse (see at A), and there are some quite densely annotated areas (see at D). For the latter, a few quick inspections with the lens (see at B) identify the most interesting candidates for a closer look, and we zoom in at two of them, shown in atomic mode in Figure 5a and Figure 8. Both passages show constellations with a lot of overlap and disagreement. Besides the red 'person'- and the blue 'location'-tags, pink tags mark uncertain annotations for resubmission. The tokens carrying the most annotations are 'St.' in Figure 5a and 'ins Tal' in Figure 8, as can be seen from the dark background color (see at A in both figures respectively). In both passages, annotators no. 12 and 26 (the second and fourth band, counting from top) show a strong agreement, and they were also the ones tagging passages for resubmission. Furthermore, annotator no. 13 (the third band) nested his annotations more strongly in both passages, whereas annotator no. 11 (the first band) annotated more separatingly. Based on these observations, we can construct a hypothesis describing the annotators' preferences. We look for further support by first bringing annotators no. 12 and 26 in immediate adjacency and then slide the brush through the overview window at different zoom levels while also keeping an eye

on the brushed passages in the text window. If we can corroborate our conjecture, we may use it to further refine the project's annotation guidelines.

The second task is to analyze the distribution of the annotated tokens and types and their context. Starting again with the overview of Figure 9, we can hover over the largest type bars and see that while for the top three of them – the main characters' names (E) – the annotations are spread all over the text, the annotations of the fourth type – 'Hügel' (C) – are concentrated in a small region towards the end, as the tick marks and yellow highlights indicate. We freeze these highlights by clicking and zoom into a passage large enough to encompass all 'Hügel' occurrences except for a few outliers; this is shown in transit mode in Figure 2. Some of the most annotated types in this passage are character names that appear almost exclusively in annotations in this passage, like 'Salgar', 'Alpin', or 'Ullin' (H). Reading the respective passage in the text window shows that it is part of a short embedded narrative. We enlarge the brush slightly in the overview window, such that it contains the embedded narrative, and can see in the corresponding type bars a series of further character names, that were almost exclusively annotated in this short passage, while the main characters' names do not show up at all (see red dots in Figure 7). Furthermore, 'Hügel' is still the most annotated type here. Based on these findings, it is now an interesting task for an analyst to consider the role that a hill or hills play in this embedded narrative and how the remaining references to hills in the rest of the text relate to that.

2.3 Visual Annotation Analysis in the CRETA Workflow

As it is described in the chapter by Pichler and Reiter (2020) in this volume, the notion of reflected text analytics consists of a succession of interrelated practices that is sketched there in Figure 1 (page 44). An early, large group of these practices is concerned with tagging a set of phenomena on the text surface that operationalize the respective domain problem. With AnnoXplorer, we provide an approach to support the explorative analysis of these annotation data. As shown in the usage scenarios above, this analysis helps to iteratively enhance the model assumptions reflected in the project's annotation guidelines. And it helps to refine the domain research question in the sense of a hermeneutic procedure as described in the introduction of this volume.

3 Visualization of Character Networks

The later stages of the workflow pipeline of reflected text analytics are about interpreting the phenomena that were operationalized in the earlier stages. When the person entities in a narrative text as well as the verbs and adjectives that characterize their relationships are annotated (manually or automatically), one may next want to analyze the ensemble of the plot's characters under the guiding questions of a literary interpretation. Typical domain tasks in this endeavor may encompass to characterize the protagonists of a plot, their attitudes and development, to collect information about the relationships of the different characters and see how they evolve over the course of the plot, and to compare character constellations at different timestamps with each other.

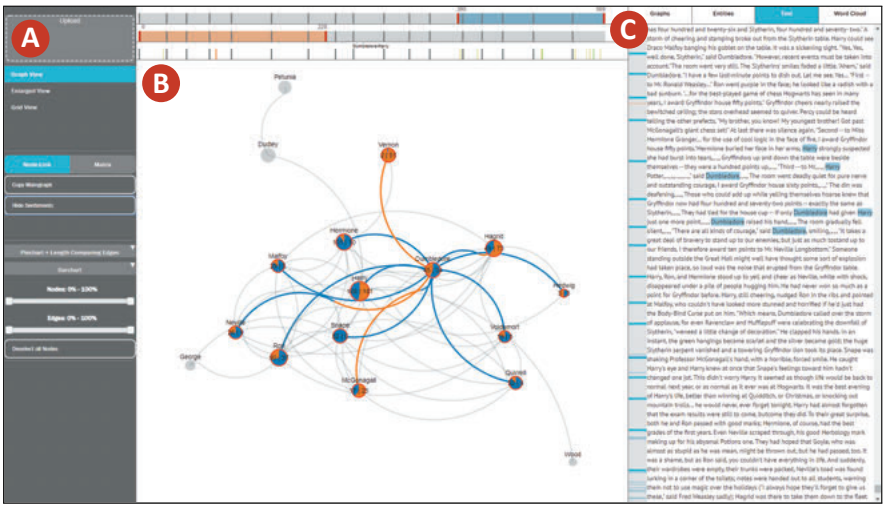


Fig. 10: The main workspace comprises (A) a menu that enables users to switch between the different visualizations or to filter the character network, (B) the main view where users can inspect the character networks with the aid of matrix-based or node-link visualizations, (C) the tab view where users can switch between a graph desktop, an entity list, a text view, and a word cloud view.

Our contribution to support such tasks is an interactive visualization approach that enables users to select different ranges of a text document and to visually analyze and compare the character networks occurring within them. It provides visual abstractions and interactive features like focusing or filtering devices, that facilitate the analysis of information about the change of relationships

between characters as well the analysis of the corresponding text passages. Textual summarizations of the context in which a collection of characters appear characterize their relationships semantically. Here, the concept of 'relationship' between two characters and also the concept of 'characterization' of a character through a verb or adjective are defined as their respective common mention in a certain text range. The extension of such a range depends on the text considered; in the case of *Parzival* – which we will look at in Section 3.2 – it is one paragraph. The main workspace of the approach can be seen in Figure 10. Its main view in the center displays a matrix-based or a node-link visualization of the character network as appearing in the selected text passages. These text selections are represented as color-encoded brush rectangles in the *document bands* above the visualization. Further auxiliary views are collected in a tabbed panel to the right, and controls for filtering and the mode of representation are collected in a menu to the left. We will describe the details of these representations and interactions in Section 3.1 and again exemplify their usage in a scenario in Section 3.2.

Our visual approach deals with the representation and comparison of graphs as well as with the visual analysis of characters in narrative texts. In the following, we discuss some previous works from these three fields and how they influenced our approach or differ from it. For a more detailed and wider discussion of the relevant research context, we refer to John et al. (2019).

First, let us consider the challenges of the visual comparison of graphs. Gleicher et al. (2011) present a survey that defines three main categories for visual comparison. Based on the defined categories, Beck et al. (2017) slightly revise them for graph comparison challenges. The revised categories comprise 'juxtaposition', 'superimposition', and 'integration with explicit visual encoding'. In a juxtaposition approach, the different states of the graph are depicted side by side, whereas a superimposition technique represents the different states on top of each other. In an integrated approach, the different pieces of information about the different states – in our case: mentioned characters in text passages – are an integral part of the visual representation. Some examples for the different techniques are Jianu et al. (2010) for superimposition; Albers et al. (2011) and Blascheck et al. (2017) for juxtaposition; Shi et al. (2011) for integration with explicit encoding; and Federico et al. (2011) for the combination of these techniques.

Second, let us have a look at approaches dealing with the visualization of character networks. These networks differ from social networks – for which there have been derived a series of approaches in recent years – in that they are typically much smaller and show a complex evolution of relationships between their entities. The *NEREx* approach by El-Assady et al. (2017) supports the visual comparative analysis of discussions between several persons by offering a node-link visualization of these persons and their relationships. Kim et al. (2011) present an

approach that combines word clouds and node-link diagrams to visualize named entities and their relationships in text collections. Kurzhals et al. (2016) propose a visual movie analytics approach, which allows users to annotate and analyze characters from movies on different levels of detail. Finally, the *Jigsaw* approach by Stasko et al. (2008) offers multiple coordinated views that enable users to track characters and their relations in large text corpora.

Third, we focus on approaches that integrate similar visualization techniques for visual encodings as our approach. Henry et al. (2007) present *NodeTrix*, which allows users to select different parts of a graph and to seamlessly switch between a matrix-based and a node-link visualization. Cao et al. (2015) present the *g-Miner*, which also supports a matrix-based and node-link visualization to represent network data. Both visualization techniques use similar visual encodings to highlight the differences between variant states: for example a pie chart in the node-link visualization, or bar charts in the matrix-based representation to depict the different numbers of occurrences.

3.1 Character Networks: Representations and Interactions

Let us start with a description of the main view at Figure 10B. Here, users can select one or two text ranges (with the latter, we enter *comparison mode*) by brushing in the document bands and choose to present the respective character networks either as a node-link- or as a matrix-based visualization; or they can chose do display a grid view with selected graphs here (see Figure 12). The gray lines in the document bands represent segments of the document structure like parts or chapters; hovering over them shows a segments' name in a tooltip. Below the document bands, a *sentiment band* can be activated. Here, green bars represent positive and red bars represent negative verbs or adjectives that occur in the context of single characters or of pairs (relationships) of characters. In the main view, we offer four different visual encodings for the node-link diagram and three for the matrix-based visualization in the comparison mode. Furthermore, users can apply filter techniques that support the comparative exploration of the character networks.

The node-link representation is a widely used technique to visualize smaller graphs. It has the advantage to be easily understandable and that paths through the network can be immediately recognized. It consists of vertices (characters) and edges (relationships) to represent the graph structure in a force-directed layout. The four encodings result from a free combination of two encodings for the vertices and two for the edges (see Figure 11). Users can use either a bar or a circle to represent characters. The size of the circles and the length of the bars are scaled

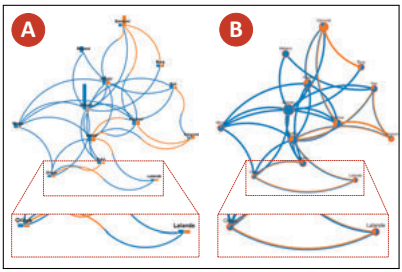


Fig. 11: Four different node-link diagram encodings: (A) juxtaposed bar charts with length comparing edges and (B) pie charts with width comparing edges.

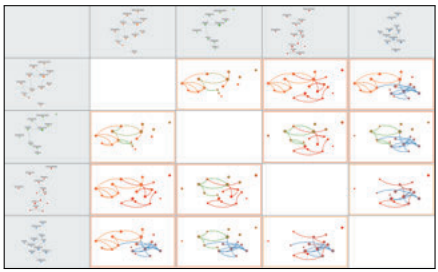


Fig. 12: The grid view provides an overview of all selected character networks as juxtaposed small multiples in an adjacency matrix. The frames' colors represent the similarity between the respective character networks.

proportionally to the number of the characters' occurrences. In the node-link comparison mode (i. e. if two text passages are selected), we map the number of the characters' occurrences in the respective passages to either slices in a pie chart or to bars in a bar chart. For the representation of the relationships, users can chose between length comparing edges and width comparing edges. With the former, the numbers of co-occurrences of the connected characters in the two respective passages are mapped to the lengths of the respective edge segments (Figure 11A); with the latter, they are mapped to the widths of the two parallel lines that compose an edge (Figure 11B). A double-click on a vertex or edge opens the text with the corresponding highlights in the text view, the corresponding related verbs and adjectives in the word cloud view, and the corresponding sentiment results in the sentiment band.

As with all node-link diagrams, the legibility of the character network deteriorates quickly as the number of edges and vertices grows. To alleviate the occlusion problems, we provide filter controls. By hovering over a character, all characters and relations not directly connected to it are grayed out as can be seen in Figure 10. Users can also gray out particular characters and their corresponding relations by clicking on them. Additionally, they can set a percentage to filter characters or relationships that only occur in one of the selected text ranges or that frequently occur in both via the sliders in the menu. However, these filter techniques do only support users up to a certain extent, and therefore, we also provide a matrix-based visualization that eliminates visual clutter.

A matrix-based visualization provides a more scalable approach to represent larger networks. The rows and columns of the matrix represent the characters and the cells their relationships. To represent the number of co-occurrences between

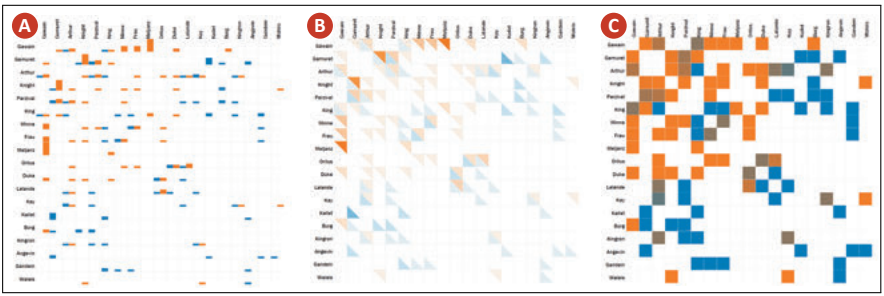


Fig. 13: Three different visual encodings for the comparison mode of the adjacency matrix: (A) juxtaposed color-coded bar charts, (B) split crosswise cells, and (C) color transition.

two characters, we use a color saturation from white to the respective color of the graph as depicted in Figure 14B. If the users activate the comparison mode, they can choose between three visual encodings as depicted in Figure 13. From (A) to (C), the encodings offer less information but can be employed for smaller cells (scalability). With the juxtaposed color-coded bar charts (A), the length of each bar is scaled proportionally to the number of co-occurrences of the two characters in the respective passages. With the split crosswise encoding (B), this information is given by a color gradient in the two triangles that partition each cell. And with the color transition (C), we mark the ratio of the two values by a cell's color. Similarly as with the node-link diagram, users can hover over a cell to show the number of co-occurrences in a tooltip, and by clicking on an entity label or cell, we depict the occurrences or co-occurrences of the characters in the text view as well as the corresponding verbs and adjectives in the word cloud view. We also offer the possibility to filter characters and their relationships that, for example, are only mentioned in one of the selected parts via the sliders in the menu.

In contrast to the node-link visualization, where the users have to examine the extracted verbs and adjectives in a linked view, we enable users to investigate these terms directly in the matrix visualization. To this end, we integrated a focus+context ('fisheye') technique. A fisheye technique uses a spatial distortion to focus on particular areas without loosing their overall context, all within a single view. The focused regions show a higher level of detail, whereas the non-focused areas show fewer details, albeit they are still visible. In our integrated fisheye technique, users can switch between three zoom levels. We use the additional space in the focused areas to display word clouds that represent the verbs and adjectives as depicted in Figure 14A. We sort the terms of the word cloud by their frequency, and if the user is in comparison mode, we use stacked bars below the terms in the colors of the respective networks to show the different occurrences. If the sentiment

Fig. 14: (A) Our focus+context approach in the comparison mode with two different selected parts of the romance *Parzival*. (B) The graph desktop containing four user-selected graphs.

mode is activated, we represent negative terms in red, neutral terms in black, and positive terms in green. Thus, users can freely explore the word clouds of the characters and their relationships within a single view. If there is not enough space to display a word cloud in a cell, we display one of the above-mentioned visual encodings.

The third representation that can be chosen to be displayed in the main view is the grid view. This view provides an overview of all stored character networks as small multiples as depicted in Figure 12. The rows and columns represent the stored character networks and the cells the comparison networks of them. To indicate the similarity of the networks, we integrate a simple measure that compares the number of characters' occurrences and co-occurrences. Afterward, we map the resulting values to values on a color gradient from red to green and show the color mapping on the respective cell frames. Green colors indicate similar and red colors very different networks. Thus, users receive a similarity overview, which can serve as a starting point for a further analysis.

Next, we want to consider the four auxiliary views that are shown in the tabbed panel on the right (see Figure 10C). The *graph desktop* contains small multiples of graphs corresponding to the text ranges selected by the users, either in the form of a matrix or a node-link representation, depending on the selected mode (see Figure 14B). Here, users can also pan, zoom, and rearrange the small

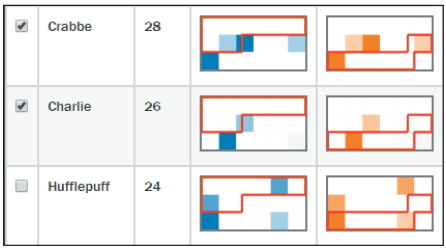


Fig. 15: The entity list enables users to select and deselect characters from the graph, shows the number of characters' occurrences, and fingerprint visualizations that represents the distribution of the entities in the different parts of the text document.

multiples, and with drag and drop, they can focus the small multiple in the main view.

The *entity list* provides an overview of the annotated and extracted characters (see Figure 15). The list is sorted by the number of the entities' occurrences in the text and users can (de)select them from the network dynamically. In addition, we show a pixel-based fingerprint visualization to represent the distribution of the character appearances in the texts. Each cell depicts, for example, a part or chapter of a text document, and the color saturation represents the number of occurrences. The red frame around the cells indicates the currently selected range of the text document.

Another auxiliary view is the *text view*, which is linked with the matrix-based and node-link visualization. This view is equipped with a vertical fingerprint visualization, which highlights the occurrences (or co-occurrences) of selected (pairs of) characters next to the side bar as depicted in Figure 10C. Blue bars represent parts of the text document, and the (co-)occurrences of the characters are depicted in the color of the respective network. Users can navigate to the corresponding text passages by clicking on an occurrence. Additionally, we also highlight the occurrences of the characters in the text. Thus, users can easily find and inspect the occurrences of the characters in the text.

Finally, the *word cloud view* is also linked with the matrix-based and node-link visualization and enables users to investigate the related verbs and adjectives of the selected characters. In the comparison mode, we again use stacked bar charts to represent the occurrences of the terms in the different selected text ranges. If the sentiment mode is activated, we also color the terms according the aforementioned sentiment assignments.

3.2 Usage Scenario: Characters in *Parzival*

In this scenario we consider an inquiry that was conducted by our project partners in CRETA on the Middle High German text *Parzival* by Wolfram von Eschenbach, which they had annotated manually (see the chapter by Ketschik et al. 2020, in this volume). In order to ensure a better understanding of this inquiry, we reproduce it here on the basis of an English translation of the text by J. Weston, where we extracted the respective entities by means of NLP. Therefore, the use case's results differ slightly from the original inquiry's.

As a first step, the project partners load the first nine books of Parzival into our system and start their analysis with exploring the characters in the node-link visualization. During this analysis, they can identify all essential characters of the nine books. Afterward, they want to compare the characters' co-occurrences within the first books with their co-occurrences within the subsequent books. For this purpose, the project partners create two character networks, where they select the first three books and the following four with the respective document bands. For the visual analysis, they switch to the matrix-based representation (see Figure 13) and inspect the character co-occurrences with the different visual encodings.

In a further analysis step, the project partners want to investigate the evolution of the characters' networks over the plot of the first books. With this analysis they want to investigate their hypothesis that the character networks are not similar. Therefore, they create a character network for each of the first four books and store them on the graph desktop as depicted in Figure 14B. Afterward, they switch to the node-link visualization again and activate the grid view to get an overview of the character networks as depicted in Figure 12. With the help of the similarity highlighting, they can easily compare the four networks and confirm that the character networks are not similar.

As a following analysis step, they want to compare the character networks in more detail. Therefore, they focus the different character networks in the main view. They find out that the character networks of the first four books are only connected via a few characters, such as Parzival or Arthur, as depicted in Figure 11.

Subsequently, they want to further investigate the co-occurrences of the characters. Therefore, they switch to the matrix-based representation and activate the fisheye technique as depicted in Figure 14A. While they freely navigate and analyze the word clouds of the character co-occurrences, they come across some verbs and adjectives that they did not expect. Those verbs and adjectives could now serve as the starting point of a deeper analysis in the text view.

3.3 Visual Character Analysis in the CRETA Workflow

As discussed in the chapter by Pichler and Reiter (2020), the later stages of the reflected text analytics workflow pipeline are concerned with the analysis of the information that was collected in the earlier stages, under the guiding questions of the respective domain. In the case that this information concerns the character entities of a narrative text, these questions aim for such aspects as the characters' attitudes and development, their relations among each other, or the evolution of character constellations over the course of the plot. The goal of detecting unexpected phenomena is well supported by our approach, since it allows to interactively explore this graph data set, and the close interleaving with the source text allows to easily inspect those passages that may best support a hypothesis regarding such a phenomenon.

4 Conclusion

As mentioned in Section 1.3, one central goal we pursued while developing the approaches described here, was to balance the conflicting requirements listed in Section 1.2. As for the novelty aspect, since both approaches were published on renowned visualization conferences requiring approval in a peer-reviewing process, their novelty was assessed favorably by the visualization community. As for the usefulness within the project context, we laid out where in the process of reflected text analytics we sought to offer support, and we exemplified a few ways of usage in the scenarios described in Sections 2.2 and 3.2. However, the effort we were able to allocate to the implementation of our approaches beyond their visualization core ideas was limited, which, in turn, limited their usefulness for our project partners. For example, in its present state, AnnoXplorer is not a fully-fledged annotation tool that would allow for the creation of annotations in a complex multi-user setup, but an approach for the visual analysis of such data. Likewise, the capacities of our approach for the comparison of character networks is limited as far as factors are concerned that are not directly related to its central visualization concepts, such as stability, performance or import capabilities. Finally, the approaches' generalizability was a great concern for us, especially as a connecting factor for future developments. For example, AnnoXplorer is agnostic as far as the source (human or algorithmic) or form (the exact meaning of the tags) of the annotations is concerned. Even the source text does not need to be natural language text at all, but just a sequence of units that can be tagged, such as the sequence of base pairs in a gene. For our character network approach, one cen-

tral contribution is that it allows for the comparison of constellations based on two different text ranges; however, the concept of the approach would allow for the simultaneous comparison of more than just two constellations. For a detailed discussion of the current intentions for these and more future developments, we refer to Baumann et al. (2020) and John et al. (2019) respectively. To sum up our experiences, we think that while it can be difficult to match the initial expectations of the visualization and the domain side – say, with respect to the extent and quality of the implemented software prototypes, or with respect to the availability of the domain data – the learning process resulting from an iterative and collaborative development process can be of great benefit for both sides.

Acknowledgment: The authors would like to thank their research assistants Harutyun Minasyan, Jena Satkunarajan, and David Schütz for their work. They also would like to thank their project partners from CRETA for the valuable feedback and the data provided, especially Sandra Murr and Nora Ketschik and their teams of annotators, as well as André Blessing and Nils Reiter.

References

Albers, Danielle, Colin Dewey, and Michael Gleicher (2011). "Sequence Surveyor: Leveraging Overview for Scalable Genomic Alignment Visualization". In: *IEEE Trans. Vis. and Comput. Graph.* 17.12, pp. 2392–2401. DOI: 10.1109/TVCG.2011.232.

El-Assady, Mennatallah, Rita Sevastjanova, Bela Gipp, Daniel A. Keim, and Christopher Collins (2017). "NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations". In: *Comput. Graph. Forum* 36.3, pp. 213–225. DOI: 10.1111/cgf.13181.

Baumann, Martin, Harutyun Minasyan, Steffen Koch, Kuno Kurzhals, and Thomas Ertl (2020). "AnnoXplorer: A Scalable, Integrated Approach for the Visual Analysis of Text Annotations". In: *Proc. 15th Int. Jt. Conf. Comput. Vis., Imaging and Comput. Graph. Theory and App. - IVAPP*. Valletta, Malta, pp. 62–75. DOI: 10.5220/0008965400620075.

Beck, Fabian, Michael Burch, Stephan Diehl, and Daniel Weiskopf (2017). "A Taxonomy and Survey of Dynamic Graph Visualization". In: *Comput. Graph. Forum* 36.1, pp. 133–159. DOI: 10.1111/cgf.12791.

Blascheck, Tanja, Markus Schweizer, Fabian Beck, and Thomas Ertl (2017). "Visual Comparison of Eye Movement Patterns". In: *Comput. Graph. Forum* 36.3, pp. 87–97. DOI: 10.1111/cgf.13170.

Cao, Nan, Yu-Ru Lin, Liangyue Li, and Hanghang Tong (2015). "g-Miner: Interactive Visual Group Mining on Multivariate Graphs". In: *Proc. 33rd Ann. ACM Conf Hum. Fact. Comput. Syst. (CHI '15)*. Seoul, Republic of Korea, pp. 279–288. DOI: 10.1145/2702123.2702446.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman (1999). "Information Visualization". In: *Readings in Information Visualization*. Ed. by Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. Morgan Kaufmann, pp. 1–34.

Chandrasegaran, Senthil, Sriram Karthik Badam, Lorraine Kisselburgh, Karthik Ramani, and Niklas Elmqvist (2017). "Integrating Visual Analytics Support for Grounded Theory Practice in Qualitative Text Analysis". In: *Comput. Graph. Forum* 36.3, pp. 201–212. DOI: 10.1111/cgf.13180.

Correll, Michael A., Michael Witmore, and Michael Gleicher (2011). "Exploring Collections of Tagged Text for Literary Scholarship". In: *Comput. Graph. Forum* 30.3, pp. 731–740. DOI: 10.1111/j.1467-8659.2011.01922.x.

Eckart de Castilho, Richard, Chris Biemann, Anette Frank, Iryna Gurevych, Silvana Hartmann, Eva Mujdricza-Maydt, and Seid Muhie Yimam (2016). "A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures". In: *Proc. Workshop Lang. Technol. Resour. and Tools for Digit. Humanities (LT4DH)*. Osaka, Japan, pp. 76–84.

Federico, Paolo, Wolfgang Aigner, Silvia Miksch, Florian Windhager, and Lukas Zenk (2011). "A Visual Analytics Approach to Dynamic Social Networks". In: *Proc. 11th Int. Conf. Knowl. Manag. and Knowl. Tech. (i-KNOW '11)*. Graz, Austria, pp. 1–8. DOI: 10.1145/2024288.2024344.

Gleicher, Michael, Danielle Albers Szafir, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts (2011). "Visual comparison for information visualization". In: *Inf. Vis.* 10.4, pp. 289–309. DOI: 10.1177/1473871611416549.

Henry, Nathalie, Jean-Daniel Fekete, and Michael J. McGuffin (2007). "NodeTrix: A Hybrid Visualization of Social Networks". In: *IEEE Trans. Vis. and Comput. Graph.* 13.6, pp. 1302–1309. DOI: 10.1109/TVCG.2007.70582.

Jänicke, Stefan (2016). "Valuable Research for Visualization and Digital Humanities: A Balancing Act". In: *VIS4DH*. Balitmore, MD.

Jianu, Radu, Kebing Yu, Lulu Cao, Vinh Nguyen, Arthur R. Salomon, and David H. Laidlaw (2010). "Visual integration of quantitative proteomic data, pathways, and protein interactions". In: *IEEE Trans. Vis. and Comput. Graph.* 16.4, pp. 609–620. DOI: 10.1109/TVCG.2009.106.

John, Markus, Martin Baumann, David Schuetz, Steffen Koch, and Thomas Ertl (2019). "A Visual Approach for the Comparative Analysis of Character Networks in Narrative Texts". In: *Proc. IEEE Pac. Vis. Symp. (PacificVis)*. Bangkok, Thailand, pp. 247–256. DOI: 10.1109/PacificVis.2019.00037.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck, and Axel Pichler (2020). "Interdisziplinäre Annotation von Entitätenreferenzen". In: *Reflektierte Algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Kim, KyungTae, Sungahn Ko, Niklas Elmqvist, and David S. Ebert (2011). "WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora". In: *Proc. 44th Hawaii Int. Conf. Syst. Sci. (HICSS)*. Kauai, HI, pp. 1–8. DOI: 10.1109/HICSS.2011.499.

Kleymann, Rabea, Jan Christoph Meister, and Jan-Erik Stange (2018). "Perspektiven Kritischer Interfaces für die Digital Humanities im 3DH-Projekt". In: *Proc. 5th DHd*. Cologne, DE, pp. 279–283.

Koch, Steffen, Markus John, Michael Wörner, Andreas Müller, and Thomas Ertl (2014). "VariFocalReader: In-Depth Visual Analysis of Large Text Documents". In: *IEEE Trans. Vis. and Comput. Graph.* 20.12, pp. 1723–1732. DOI: 10.1109/TVCG.2014.2346677.

Kurzhals, Kuno, Markus John, Florian Heimerl, Paul Kuznecov, and Daniel Weiskopf (2016). "Visual Movie Analytics". In: *IEEE Trans. Multimedia* 18.11, pp. 2149–2160. DOI: 10.1109/TMM.2016.2614184.

Landragin, Frédéric, Thierry Poibeau, and Bernard Victorri (2012). "ANALEC: A New Tool for the Dynamic Annotation of Textual Data". In: *Proc. 8th Int. Conf. Lang. Resour. and Eval. (LREC)*. Istanbul, Turkey, pp. 357–362.

Meister, Jan Christoph, Marco Petris, Evelyn Gius, and Janina Jacke (2016). *CATMA 5.0: Software for Text Annotation and Analysis*. URL: http://catma.de/ (visited on Mar. 15, 2019).

Munzner, Tamara (2009). "A Nested Model for Visualization Design and Validation". In: *IEEE Trans. Vis. and Comput. Graph.* 15.6, pp. 921–928. DOI: 10.1109/TVCG.2009.111.

Munzner, Tamara (2014). *Visualization Analysis & Design*. CRC Press.

Pichler, Axel and Nils Reiter (2020). "Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Shi, Lei, Chen Wang, and Zhen Wen (2011). "Dynamic network visualization in 1.5D". In: *Proc. IEEE Pac. Vis. Symp. (PacificVis)*. Hong Kong, China, pp. 179–186. DOI: 10.1109/PACIFICVIS.2011.5742388.

Stasko, John T., Carsten Görg, and Zhicheng Liu (2008). "Jigsaw: Supporting Investigative Analysis through Interactive Visualization". In: *Inf. Vis.* 7.2, pp. 118–132. DOI: 10.1057/palgrave.ivs.9500180.

Van Wijk, Jarke J. (2006). "Bridging the Gaps". In: *IEEE Comput. Graph. and Appl.* 26.6, pp. 6–9. DOI: 10.1109/MCG.2006.120.

Widlöcher, Antoine and Yann Mathet (2012). "The Glozz Platform: A Corpus Annotation and Mining Tool". In: *Proc. 2012 ACM Symp. Doc. Eng. (DocEng '12)*. Paris, France, pp. 171–180. DOI: 10.1145/2361354.2361394.

Benjamin Krautter

‚Figurenstil' im deutschsprachigen Drama (1740–1930)

Eine stilometrische Annäherung

Zusammenfassung: Der Beitrag eruiert mittels stilometrischer Analysen, ob Autor*innen ihre dramatischen Figuren durch eine je individuell gestaltete Figurenrede charakterisieren. Diese Hypothese liegt nahe, unterliegt die Figurenrede und ihre sprachstilistische Realisierung doch vielfältiger poetologischer Reflexionen. Untersuchungsgegenstand sind dabei 60 deutschsprachige Dramen, die zwischen 1740 und 1930 veröffentlicht oder uraufgeführt wurden. Die Auswertung der Analysen zeigt, dass die stilometrische Differenzierung der Figurenrede von den jeweiligen Autor*innen, dem Veröffentlichungszeitraum und poetologischen Überlegungen abhängt. Die Ergebnisse verdeutlichen auch, dass die Kopräsenz von Figuren einen merklichen Einfluss auf die stilometrischen Untersuchungen nimmt.

Abstract: By means of sytlometric analyses, this contribution tries to evaluate if authors create stylistically distinctive character speeches for their dramatic characters. This hypothesis seems to be plausible, as both the character speech and its linguistic realisation are subject to various poetological reflections. Examining 60 German-language dramas between 1740 and 1930, it turns out that the results depend on the chosen authors, the publication date and further poetological considerations. They also illustrate that a play's structure has a noticeable influence on the stylometric analysis of the characters' speeches.

1 Problemaufriss

Die sprachliche Ausgestaltung dramatischer Figurenrede ist seit jeher ein Fixpunkt poetologischer Reflexionen. Mit Blick auf seine Übersetzung einiger Passagen von *The Unhappy Favourite, or, the Earl of Essex* (1681), einer Tragödie des britischen Dramatikers John Banks, bemerkt etwa Gotthold Ephraim Lessing:

Anmerkung: Der Aufsatz basiert auf zwei Vorträgen und wurde für die vorliegende Fassung substantiell überarbeitet und erweitert. Vgl. Krautter 2018a und Krautter 2018b.

Benjamin Krautter, Germanistisches Seminar, Universität Heidelberg

Open Access. © 2020 Benjamin Krautter, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-013

> Wie ich Banks Elisabeth sprechen lasse, weiß ich wohl, hat noch keine Königin auf dem französischen Theater gesprochen. Den niedrigen vertraulichen Ton, in dem sie sich mit ihren Frauen unterhält, würde man in Paris kaum einer guten adlichen Landfrau angemessen finden.[1] (Lessing 1985a, S. 476)

Auf diese Übersetzung bezugnehmend urteilt Lessings Bruder Karl in einem Brief vom 03. Februar 1772 über das Trauerspiel *Emilia Galotti* (1772): „In Deiner Emilia Galotti herrscht ein Ton, den ich in keiner Tragödie, so viel ich deren gelesen, gefunden habe" (Lessing 1988, S. 343). Gemäß Karl Lessing sei die Sprache „ganz natürlich" (Lessing 1988, S. 343) und würde im Ton den übersetzten Passagen von *The Unhappy Favourite* entsprechen. Lessing reflektiert die Redeweise seiner Figuren aber nicht nur, er verlagert diese Überlegungen auch als Gestaltungsprinzip in die Figurenrede selbst und charakterisiert dadurch seine Figuren. Denn diese sind zumindest teilweise im Stande, ihren eigenen Sprachstil zu hinterfragen. In *Minna von Barnhelm, oder das Soldatenglück* (1767) glaubt die gleichnamige Titelfigur im neunten Auftritt des zweiten Aktes, ihre Rede unbewusst im Ton verändert zu haben:

> DAS FRÄULEIN Sie können; Sie müssen wissen, was in Ihrem Herzen vorgeht. – Lieben Sie mich noch, Tellheim? – Ja, oder Nein.
> V. TELLHEIM Wenn mein Herz –
> DAS FRÄULEIN Ja, oder Nein!
> V. TELLHEIM Nun, Ja!
> DAS FRÄULEIN Ja?
> V. TELLHEIM Ja, ja! – Allein –
> DAS FRÄULEIN Geduld! – Sie lieben mich noch: genug für mich. – In was für einen Ton bin ich mit Ihnen gefallen! Ein widriger, melancholischer, ansteckender Ton. – Ich nehme den meinigen wieder an. (Lessing 1985b, S. 44 f.)

Ganz ähnliche Überlegungen äußert auch Friedrich Schiller, der die Figurenrede der beiden Brüder Karl und Franz Moor – sie sind die Protagonisten seines Dramendebüts *Die Räuber* (1781) – in einem Brief an Wolfgang Heribert von Dalberg eindrücklich kontrastiert. Während Karl in seiner Figurenrede „Handlung" und „anschauliches Leben" abbilde, sei Franz „raisonierend" angelegt. Er würde dadurch zwar „den denkenden Leser befriedigen", nicht aber den Zuschauer im Theater, „der vor sich nicht philosophirt, sondern gehandelt haben will" (Schiller 1956, S. 21). Über die Sprache in Schillers *Räubern* urteilt Monika Ritzer dazu

1 Beispielhaft führt Lessing daran anschließend Ausschnitte der Figurenrede von Königin Elisabeth auf: „Ist dir nicht wohl? – Mir ist ganz wohl. Steh auf, ich bitte dich. – Nur unruhig; ein wenig unruhig bin ich. – Erzehle mir doch. – Nicht wahr, Nottingham? Tu das! Laß hören! […]." (Lessing 1985a, S. 476)

passend: Schiller vermeide „stilistische Homogenität, indem er den Sprachgestus variiert und damit die Sprache als Ausdrucksmedium zur Typologisierung der Figuren nutzt" (Ritzer 2011, S. 257).[2]

Die drei Beispiele verdeutlichen, wie sich sozialer Stand, Empathie und die Charakterisierung der Figuren als poetologische Gestaltungsgründe für die Figurenrede einsetzen lassen. Bereits das berühmte sechste Kapitel der aristotelischen *Poetik* teilt die für das Trauerspiel konstitutive „Nachahmung einer H a n d l u n g [...], die von einer bestimmten Person geschiehet", in zwei ursächliche Aspekte: „Gesinnungen und [...] Sitten" (zitiert nach Lenz 1987, S. 650).[3] Gemeinsam würden diese für Glück oder Unglück der Figuren verantwortlich zeichnen. Während Aristoteles die Sitten als Art des Handelns einer Figur begreife, seien die Gesinnungen – so zitiert Jakob Reinhold Michael Lenz die *Poetik* – die „Gemütsart und der Ausdruck derselben im Sprechen" (Lenz 1987, S. 651).[4]

Im Folgenden greife ich die schlaglichthaft skizzierten poetologischen Aussagen und analytischen Forschungszuschreibungen auf und werde diese mittels quantitativer Methoden nachzeichnen, um auf diese Weise ein zusätzliches Argument in die Forschungsdiskussion einzuführen. Dazu wird ein Korpus von 60 Bühnenstücken zwischen 1740 und 1930 – jeweils drei Stücke von 20 Autor*innen – stilometrisch untersucht. Ziel des Beitrags ist es also, anhand stilometrischer Analysen zu ergründen, ob sich die einzelnen Bühnenfiguren eines Dramas schon durch die sprachliche Gestaltung ihrer Figurenrede unterscheiden lassen. Dies setzt die Annahme voraus, dass sich der Stil beziehungsweise der Ton, in dem literarische Figuren sprechen, auch in ihrer Wortverwendung niederschlägt. Stilometrisch auszumachende Differenzen in den Wortfrequenzen könnten dann als individueller ‚Stil' der Figuren interpretiert werden.

In einem ersten Schritt werde ich die Bandbreite der umfassenden Forschungsliteratur kursorisch nachzeichnen, um daran anschließend meine Fragestellung zu konkretisieren und den daraus abgeleiteten methodischen Zuschnitt

2 Durchaus mit Ritzer vergleichbar äußert sich Mikhail Bakhtin zu den Figuren Dostojewskis. Letzterm sei es möglich, seinen Romanfiguren eine je einzigartige Stimme einzuschreiben: „*A plurality of independent and unmerged voices and consciousnesses, a genuine polyphony of fully valid voices is in fact the chief characteristic of Dostoevsky's novels*" (Bakhtin 1994, S. 6).

3 Michael Conrad Curtius nutzt in seiner Übersetzung von 1753 das Begriffspaar „Sitten und Meynungen (Grundsätze)" (Aristoteles 1973, S. 12).

4 Lenz bezieht sich in seinen *Anmerkungen übers Theater* wohl nicht auf die damals gängige Poetikübersetzung von Curtius, sondern auf eine eigene Arbeitsübersetzung (vgl. Rector 2017, S. 211 f.). In der Übersetzung von Curtius heißt es zur gleichen Passage: „die Meynungen sind es, wodurch die redenden Personen ihre Neigungen offenbaren, und ihr Gemüth entdecken" (Aristoteles 1973, S. 13). Vgl. auch die Übersetzung von Manfred Fuhrmann (Aristoteles 1982, S. 19–21).

auszuführen (2). Daran anknüpfend stelle ich das zu untersuchende Dramenkorpus genauer vor und erläutere Entscheidungen der Parametrisierung und Segmentierung (3). Der Auswertung der Analysen (4) folgt eine Diskussion der Ergebnisse, die zugleich einen Ausblick auf mögliche Evaluierungsstrategien geben soll (5).

2 Forschungsüberblick und methodischer Zuschnitt

Auch computergestützte Forschungsarbeiten legen nahe, dass versierte Autor*innen ihren literarischen Figuren sogenannte *distinctive voices* einschreiben können (vgl. Hoover 2017, S. ii17). „Authors create characters who speak in distinctive voices: imagined beings, with their own styles of utterance", postulieren etwa John Burrows und Hugh Craig gleich zu Beginn ihres 2012 veröffentlichten Aufsatzes *Authors and Characters* (Burrows und Craig 2012, S. 292).[5] Figuren könnten demzufolge also nicht nur durch körperliche Merkmale oder spezifische Charakterzüge ausgestaltet werden, sondern auch durch eine sprachlich je individuell ausgearbeitete Figurenrede, die sich zudem stilometrisch fassen und unterscheiden lasse (vgl. Hoover 2017, S. ii17 f.). Dahinter verbirgt sich die Präsupposition, dass Autor*innen tatsächlich intendierten, ihre Figuren sprechen zu lassen, „as one would expect of a character of that gender, age, social background, etc." (van Dalen-Oskam 2014, S. 443 f.). Anders als bei der Differenzierung von Autorschafts-, Gattungs- oder Epochensignalen – dem üblichen Einsatzgebiet der Stilometrie[6] – handelt es sich bei diesem stilometrischen Zugriff auf die Figurenrede um ein zuerst intratextuelles Vergleichs- bzw. Unterscheidungskriterium.

Im Fokus der Analysen stehen also vor allem die Figuren einzelner literarischer Texte, die selbst wiederum auf die Teile beschränkt bleiben, die Figurenrede beinhalten.[7] David Hoover nennt dieses Vorgehen bei der Textselektion und -aufbereitung *microanalysis* (vgl. Hoover 2017, S. ii17) und setzt sich damit einer-

5 Vgl. dazu auch die Ausführungen in Vishnubhotla et al. 2019, S. 29 f.

6 Für einen literaturwissenschaftlich motivierten Zugang zur Stilometrie vgl. etwa Schöch 2014, S. 130–157.

7 Dieser Zuschnitt des Untersuchungsgegenstands unterliegt oftmals einer weiteren Restriktion, da die Figurenrede im Kontext stilometrischer Arbeiten häufig auf direkte Rede reduziert wird (vgl. etwa Bockwinkel 2018, S. 119). Die direkte Rede ist die unmittelbarste Form der Redewiedergabe, bei der die vermittelnde Erzählerinstanz nur mehr die Verwendung der *verba dicendi* moderiert. Es entsteht also – zumindest bei Erzähltexten – ein Textmodell, das als Untersuchungs-

seits von Schlagwörtern wie *big data* oder *large scale* ab. Er betont damit andererseits aber auch die Unterschiede zu Konzepten wie dem von Franco Moretti geprägten *distant reading* (vgl. Moretti 2000, 2013) – das inzwischen mehrheitlich als Synonym für korpusbasiertes quantifizierendes Arbeiten in den Digital Humanities gebraucht wird[8] – oder dem von Matthew Jockers eingeführten Begriff *macroanalysis* (vgl. Jockers 2013, S. 3–32). Die zur Anwendung gebrachten quantitativen Methoden sind zwar durchaus vergleichbar, der epistemologische Zugriff indes ist ein anderer.

Studien zur quantitativen Differenzierung von Figurenrede finden sich sowohl für erzählende Literatur wie auch für Bühnenstücke. Im Folgenden gebe ich einen knappen Überblick, der die Forschungsliteratur zur quantitativen Unterscheidung literarischer Figuren auffächern und die spezifischen Problemstellungen verdeutlichen soll. Beginnen werde ich mit Hoovers Aufsatz „The Microanalysis of Style Variation", der gleich mehrere Romane ins Zentrum seiner Analysen rückt, vor allem aber Arthur Conan Doyles *The Hound of the Baskervilles* (1901–1902). Mit Blick auf die Protagonisten der *Sherlock-Holmes*-Romane fragt Hoover: „is it reasonable to expect Doyle to have created distinctive voices for Holmes and Watson as a way of characterizing them?" (Hoover 2017, S. ii19) Zu beantworten versucht er diese Frage anhand stilometrischer Clusteranalysen, für die er die Redeanteile der sprechenden Figuren in Segmente mit einer Länge von jeweils 1500 Wörtern teilt.

Ein anderes Vorgehen wählt Karina van Dalen-Oskam, die den Briefroman *Sara Burgerhart* (1782) betrachtet, der von den niederländischen Autorinnen Elisabeth Wolff und Agatha Deken im Kollektiv geschrieben wurde. Anhand von Zeta-Analysen[9] schließt sie, dass nur eine kleine Zahl der Hauptfiguren einen ausgeprägten Individualstil aufweisen würde (vgl. van Dalen-Oskam 2014, S. 448–450).

Eine 2017 veröffentlichte Studie von Paul J. Fields, Larry Bassist und Matt Roper nutzt ein Korpus von insgesamt acht Romanen – jeweils zwei von Jane Austen, Charles Dickens, James Fennimore Cooper und Mark Twain –, um intratextuell zwischen den Stimmen der einzelnen Figuren, aber auch zwischen Figuren und Erzählern zu unterscheiden:[10] „Using stylometric analyses we considered

gegenstand nur noch einen Bruchteil des eigentlichen literarischen Textes in die quantitative Analyse einbezieht.

8 Dazu etwa Weitin et al. 2016, S. 104: *Distant reading* sei „im Kontext der *Digital Humanities* zum geflügelten Wort für quantitative Analysen geworden[]".

9 Für eine Übersicht zu Zeta vgl. Schöch 2018, S. 77–94.

10 Da die Figuren selbst zu intradiegetischen Erzählern werden können, ist diese Binnendifferenzierung durchaus problematisch. Siehe dazu auch den Beitrag von Ketschik et al. (2020) ab S. 440 in diesem Band.

whether or not an author can create characters with different wordprint voices and found persuasive evidence that they can." (Fields et al. 2017, S. 3)[11]

Als wichtiger Wegbereiter vieler stilometrischer Arbeiten darf John Burrows einschlägige Monografie *Computation into Criticism* gelten. In seiner Untersuchung fokussiert Burrows den Idiolekt der Romanfiguren von Jane Austen, also das von ihnen durch Ausdruck und Wortschatz demonstrierte Sprechverhalten. Für ihn sind gerade die gebräuchlichsten Wörter – Pronomen, Artikel, Präpositionen – geeignete Indikatoren, um das Sprechverhalten literarischer Figuren zu unterscheiden. Burrows betont, dass es starke Hinweise – wenn nicht sogar Belege – „of stable *differentiation* between character and character, idiolect and idiolect" gebe (Burrows 1987, S. 39). Die statistische Auswertung von relativen Wortfrequenzen ermögliche also eine dem Gegenstand angemessene Unterscheidung zwischen den Figuren Jane Austens (Burrows 1987, S. 4).

Vergleichbare Überlegungen gibt es auch zur sprachlichen Gestaltung der Figurenrede in Bühnenstücken. Gerade der hohe Grad an Strukturierung lässt die quantitative Untersuchung der Figurenrede im Drama plausibel erscheinen. Die Redeanteile sind darüber hinaus weder von einem Erzähler sortiert, noch werden sie von ihm kommentiert oder in einen Rahmen gebettet. Der bereits angeführte Aufsatz „Authors and Characters" von Burrows und Craig widmet sich etwa dem Verhältnis von Figurenrede und Autor. In Rückgriff auf William Shakespeare, der besonders unverwechselbare Figuren auf die Bühne gestellt habe, leiten sie ihre Betrachtungen ein:[12] „William Shakespeare is 'myriad-minded' [...] and readers and audiences can hear cadences, expressions and turns of phrase in Cleopatra, Hamlet, Falstaff and Caliban which seem peculiar to each character, and vastly different from each other." (Burrows und Craig 2012, S. 293)[13] Der Aufsatz reagiert auf kritische Stimmen, die die (statistische) Autorschaftsattribution von Dramen grundsätzlich in Frage stellen, insbesondere bei kollektiv geschriebenen Stücken, da Dramatiker*innen ihre Figuren ganz bewusst in verschiedenen Stimmen spre-

11 Sie rekurrieren dabei auf eine Untersuchung von Hiatt und Hilton (1990, S. 52), die William Faulkners *As I Lay Dying* und dessen besondere Anlage beleuchten: „*As I Lay Dying* is told in a unique way: fifteen narrators speak from their own points of view. The language of each narrator contributes greatly to each's uniqueness. Just how well-characterized are the narrators?"

12 Vgl. dazu auch die Ausführungen von Masten (1997, S. 18), der Shakespeares Sprache als vielfältig, wandelbar und von großer Bandbreite beschreibt.

13 Siehe hierzu auch Craig und Kinney (2009, S. 15), die nach dem Verhältnis von Shakespeares Sprache und der Sprache seiner Figuren fragen: „But what, say, of Hal, Hotspur, and Falstaff? They inhabit the same play, but each has a recognizable style that sets him apart from the other two. Can there be a single, identifiable Shakespearean language that unites their three very different kinds of speech?"

chen lassen würden (vgl. McMullan 1996, S. 448–452).[14] So schließt Jeffrey Masten in seinem Aufsatz „Beaumont and/or Fletcher: Collaboration and the Interpretation of Renaissance Drama":

> a playwright im/personates another (many others) in the process of writing a play-text and this refracts the supposed singularity of the individual in language. At the same time, he often stages in language the *sense* of distinctive personae, putting 'characteristic' words in another's mouth. (Masten 1992, S. 342)[15]

Burrows und Craig versuchen hingegen zu zeigen, dass die von Masten betonte individuelle Ausgestaltung der Figurenrede und eine ausgeprägte ‚Handschrift des Autors' – das, was gemeinhin als Autorstil bezeichnet wird (vgl. Jannidis 2014, S. 173–177) – sich nicht kontradiktorisch gegenüberstehen müssen (vgl. Burrows und Craig 2012, S. 293).

Einen ähnlichen Grundgedanken, wenn auch ausschließlich auf literarische Figuren hin perspektiviert, verfolgt Hugh Craig bereits in seinem früheren Aufsatz „Contrast and Change in the Idiolects of Ben Jonson Characters". Denn einerseits kontrastiert und gruppiert Craig diejenigen 214 Figuren aus Jonsons Dramen, die über 500 Wörter sprechen, versucht andererseits aber auch, Variationen innerhalb einzelner Redesegmente der Figuren zu finden (vgl. Craig 1999, S. 221–240).

Während die bis dato angeführten Studien hauptsächlich anhand von *principle-component*-Analysen (PCA) oder Clusteranalysen argumentieren, die jeweils die Verteilung der häufigsten Wörter auswerten, bedienen sich Krishnapriya Vishnubhotla, Adam Hammond und Graeme Hirst in ihrer jüngst publizierten Arbeit „Are Fictional Voices Distinguishable?" einer Klassifikationsaufgabe.[16] Sie untersuchen dabei 63 Dramen – veröffentlicht zwischen 1880 und 1920 – von sieben Autor*innen, darunter George Bernard Shaw und Oscar Wilde. Aus ihren Ergeb-

14 Bei McMullan heißt es: „Playwrights create characters by providing them with modes of expression which distinguish them one from another; this by definition means that there cannot be any guaranteed uniformity of linguistic characteristic across a single play, never mind across an entire œuvre." (McMullan 1996, S. 451) Vgl. dazu auch die Abschlussbemerkungen von Ian Lancashire in seiner Monografie *Forgetful Muses: Reading the Author in the Text* (Lancashire 2010, S. 256).

15 Masten bezieht sich dabei auf die Arbeiten von Cyrus Hoy, der die Zusammenarbeit der beiden britischen Dramatiker John Fletcher und Francis Beaumont in einer Reihe von Aufsätzen näher zu erörtern sucht (vgl. Hoy 1956, S. 129–146).

16 Sie nutzen dafür unter anderem lexikalische und syntaktische Features (Satzlänge, Wortlänge, Type-Token-Verhältnis, Anteil an Funktionswörtern) sowie *topic models*, *SAGE models*, *distributed word vectors* und Emotionswörterbücher (vgl. Vishnubhotla et al. 2019, S. 30–31).

nissen folgern sie, dass sich die untersuchten Figuren tatsächlich mit relativ hoher Präzision unterscheiden ließen (vgl. Vishnubhotla et al. 2019, S. 31–33).[17]

Führt man das jeweilige Vorgehen und die daraus resultierenden Ergebnisse der kursorisch überblickten Forschungsliteratur zusammen, lassen sich verschiedene methodische Zugriffe identifizieren, um literarische Figuren anhand ihrer ‚Stimme', also ihrer Figurenrede, quantitativ zu differenzieren. Zugleich werden aber auch konzeptionelle Unterschiede evident. Denn das Distinktive scheint stets unterschiedlich bestimmt zu sein. Was etwa sind die unterscheidungstragenden Eigenschaften? Was sind überhaupt die Gegenstände des Vergleichs? Was also wird voneinander abgegrenzt? Was nach trivial zu beantwortenden Fragen klingt, hat für den Versuchsaufbau durchaus Gewicht. Wählt man eine intratextuelle Vergleichsfolie, vergleicht also Figuren desselben literarischen Textes, sind die Ergebnisse der Analyse anders zu interpretieren als bei einem Vergleich mehrerer Texte von potenziell unterschiedlichen Autor*innen. Die Figurenrede von Hamlet kann in ihrer sprachlichen Gestaltung beträchtlich von derjenigen Cleopatras abweichen, über das sprachstilistische Verhältnis von Hamlet und Claudius oder Hamlet und Horatio ist damit aber noch nichts ausgesagt.

Vergleichbar argumentiert auch Hoover in seiner Untersuchung zu *The Hound of the Baskervilles*: „Analyzing the voices of Holmes and Watson for distinctiveness requires comparing them with Doyle's other characters, and, because reliable results require substantial amounts of text, I focus here on the longest Holmes novel" (Hoover 2017, S. ii19). Diese Grundannahme überführt Hoover in eine theoretisch-methodische Überlegung. Um die intratextuelle Ähnlichkeit der Figurenrede nicht nur anhand der Abgrenzung zu anderen Figuren bemessen zu können – etwa daran, dass die Figurenrede von Baskerville und Watson stilometrisch ähnlicher wäre als diejenige von Holmes und Watson –, nimmt er eine zusätzliche Segmentierung der Figurenrede vor. Er teilt die Redeanteile jeder Figur in jeweils 1500 Wörter umfassende Segmente. Auf diese Weise erhofft er sich einen besseren Vergleich: Ob die ‚Stimme' einer Figur distinktiv angelegt ist, bemisst sich jetzt nicht mehr ausschließlich anhand anderer Figuren, sondern vor allem an der stilometrischen Ähnlichkeit derjenigen Redesegmente, die einer Figur zugehören. Hoovers Operationalisierung lässt sich somit in vier Schritten zusammenfassen: Seine Ausgangsfrage, ob Doyle seine Figuren durch eine sie auszeichnende einzigartige sprachliche Gestaltung charakterisiert, sucht er durch stilometrische Analysen zu beantworten. Dafür extrahiert er die direkte Fi-

17 Für die Klassifikation der Figuren aller 63 Dramen erzielen sie einen durchschnittlichen F_1-Wert von 0,561. Unklar bleibt allerdings, welche Figuren der einzelnen Stücke klassifiziert werden sollten und was im Versuchsaufbau die Zielklasse ist.

gurenrede in Doyles *The Hound of the Baskervilles*. Um einen aussagekräftigeren intratextuellen Vergleich zu gewährleisten, segmentiert er die Figurenrede. Sind sich die einzelnen Figurensegmente daran anschließend in einer stilometrischen Clusteranalyse am ähnlichsten, gruppieren sie sich also in unmittelbarer Nähe auf einem eigenen Ast im Dendrogramm, könne dies als quantitatives Argument für eine tatsächlich distinktive Figurenrede gewertet werden (vgl. Hoover 2017, S. ii18–ii20).[18] Versteht man die Stilometrie als „Anwendung quantitativer Methoden zur Erfassung und Klassifizierung stilistischer Merkmale von Texten" (Viehhauser 2015, o.S.), scheint sie grundsätzlich gut geeignet, um die sprachliche Gestaltung der Figurenrede zu untersuchen. Die Interpretation der Analyseergebnisse muss gleichwohl in Rechnung stellen, dass ein derartiges Stilverständnis vor allem auf Wortfrequenzen basiert.[19]

In weiten Teilen übernehme ich das dargelegte methodische Vorgehen von Hoover für meine eigenen Analysen. Denn für die fokussierte Fragestellung scheinen unüberwachte Clusteranalysen sinnvoller zu sein als eine überwachte Klassifikationsaufgabe. Anders als eine Klassifikation setzt das Clustering keine zumindest impliziten Annahmen über die Intention der Autor*innen voraus. Ob diese grundsätzlich intendieren sprachstilistisch einzigartige Figuren auf die Bühne zu stellen, ist zumindest bei stark typisierten Figuren zu hinterfragen. Zugleich bleibt unklar, ob die Umsetzung einer solch faktisch oder hypothetisch angenommenen Intention überhaupt immer gelingt. Clusteranalysen zeigen sich auch dahingehend voraussetzungsfrei. Sie setzen zudem – anders als eine überwachte Klassifikation – keine Referenzdaten voraus, deren Annotation im Fall stilistisch distinktiver Figurenrede einen hohen analytischen Aufwand voraussetzen würde. Anders als Hoover verzichte ich jedoch auf eine arbiträre Segmentierung der Redeanteile. Ich nutze stattdessen die gegebenen Aktgrenzen, um die Figurenrede in einzelne Segmente zu untergliedern. Hiervon erhoffe ich mir sowohl zusätzliche Kontexte für die Interpretation der Ergebnisse als auch ein größeres Potenzial für literaturwissenschaftliche Anschlussfragen. Denn eine an den Akten orientierte

18 Auch Hugh Craig nutzt Segmentierungstechniken, um die Redebeiträge von Ben Jonsons Figuren näher zu untersuchen. Anders als Hoover geht es ihm jedoch vornehmlich um die sprachlichen Veränderungen der Figurenrede im Verlauf der Stücke und weniger um die Abgrenzung von anderen Figuren (vgl. Craig 1999, S. 231–238).

19 Zur Frage, inwieweit ein solches Stilverständnis mit der literaturwissenschaftlichen Stilistik kompatibel ist, vgl. Jannidis 2014, S. 173–177. Bei Jannidis heißt es: „Stil wird häufig als Wahl bestimmt; nach dieser Auffassung wird Stil dort sichtbar, wo eine Wahl besteht. Diese Wahl kann bewusst und unbewusst sein – die meisten Stilkonzepte schließen beides ein" (Jannidis 2014, S. 174). Vgl. dazu auch die Ausführungen von Stockinger (2007, S. 69 f.) zum Stilbegriff Emil Staigers.

Segmentierung behält beispielsweise die dortigen Auf- und Abtrittsstrukturen der Figuren bei. Es lässt sich also weiterhin nachvollziehen, wann eine Figur spricht und welche anderen Figuren zu diesem Zeitpunkt auf der Bühne präsent sind. Meine Analysen fokussieren überdies nicht nur einige wenige Einzeltexte, sondern versuchen, ein größeres Dramenkorpus in den Blick zu nehmen. Dadurch zeichnen sich möglicherweise erste Trends und Muster ab, ob, wann, von wem und wo die sprachliche Gestaltung der Figurenrede zu *distinctive voices* im deutschsprachigen Drama führt.

3 Korpus und Parametrisierung

Gegenstand der Untersuchung ist ein Korpus von 60 deutschsprachigen Dramen, das einen Zeitraum von knapp 200 Jahren abdeckt – zwischen 1740 und 1930. Die Dramen entstammen dem *German Drama Corpus* (Fischer et al. 2019).[20] Abbildung 1 zeigt, wie sich die einzelnen Stücke auf die verschiedenen Jahrzehnte ihrer Entstehung verteilen. Das Korpus wurde so gewählt, dass jedes Jahrzehnt zumindest durch ein Stück repräsentiert wird. Dennoch ergibt sich eine Ungleichverteilung: insbesondere der Zeitraum zwischen 1780 und 1790 zeigt sich mit acht Dramen, etwa Goethes *Torquato Tasso* (1790) oder Schillers *Don Karlos* (1787), überrepräsentiert.[21] Diese Verteilung ist auch Ursache des Korpusdesigns. Jede*r im Korpus vertretene Autor*in steuert genau drei Bühnenstücke bei.[22] Für die Produktion der 60 Dramen zeichnen also genau 20 Autor*innen verantwortlich, darunter hochkanonische wie Gotthold Ephraim Lessing, Heinrich von Kleist oder Arthur Schnitzler. Der Aufbau erlaubt den analytischen Vergleich verschiedener Autor*innen, führt aber zugleich dazu, dass die Dramen einiger Autor*innen nicht berücksichtigt werden konnten. So enthält das *German Drama Corpus* von Georg Büchner momentan nur die beiden vollendeten Bühnenstücke *Leonce und Lena* (1838) sowie *Dantons Tod* (1835), nicht aber das Dramenfragment *Woyzeck*. Dies wurde aufgrund der editionsphilologischen Schwierigkeiten – in welcher Reihen-

20 Das Dramenkorpus umfasst inzwischen 496 Dramen (Stand 20.05.2020).

21 Das *German Drama Corpus* führt Metadaten für die Veröffentlichungszeitpunkte der Dramen. Da Erstaufführung und Erstpublikation zeitlich recht weit auseinander liegen können und zudem nicht für alle Dramen beide Datumsangaben gelistet sind, wurde die jeweils frühere der beiden Angaben für die Zuordnung genutzt.

22 Ich orientiere mich dabei an Jannidis et al. (2015). Die drei dort genutzten Korpora, die jeweils 75 auf Deutsch, Französisch und Englisch geschriebene Romane von 25 Autor*innen beinhalten, wurden später in weiteren Studien verwendet (vgl. Evert, Jannidis et al. 2016 und Evert, Proisl et al. 2017).

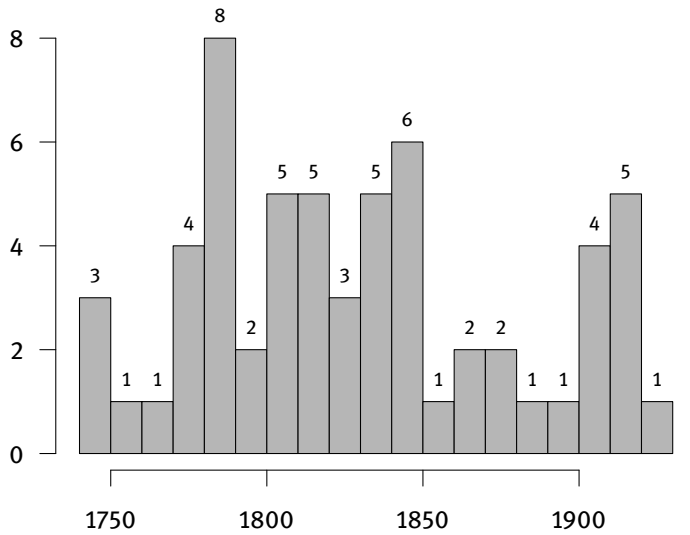


Abb. 1: Korpusübersicht: Verteilung der Dramen in Intervallen von 10 Jahren.

folge etwa sollen die losen Szenen angeordnet werden? – nicht aufgenommen. Ferner schränken auch die methodischen Vorentscheidungen die Zusammenstellung des Korpus ein. Um die geplante Segmentierung realisieren zu können, ist für jedes Drama eine Mindestzahl an Figuren nötig, deren Redebeiträge sich zu einer gewissen Länge summieren müssen. Andernfalls fehlen die intratextuellen Vergleichsobjekte. Christian Dietrich Grabbe stellt in seinem Historiendrama *Napoleon oder Die hundert Tage* (1831) etwa fast 200 Figuren oder Figurengruppen auf die Bühne. Mit einem Redeanteil von 6289 Tokens ist Napoleon jedoch die einzige Figur, die mehr als 2000 Tokens äußert. Im Durchschnitt beschränkt sich die Zahl der geäußerten Tokens pro Figur auf 218 (Standardabweichung: 566). Würde man die Redeanteile der Figuren zusätzlich segmentieren, ließen sich letztlich nur mehr die Segmente Napoleons vergleichen.

Wie lang ein Textsegment indes sein muss, um ein zuverlässiges Funktionieren stilometrischer Analysen gewährleisten zu können, ist nicht pauschal zu beantworten, sondern von der Forschungsfrage und dem untersuchten literarischen Korpus abhängig. Auch die in der Forschung genannten Zahlen weichen recht stark voneinander ab und reichen von weniger als 500 (vgl. Stamatatos et al. 2014, S. 881, 886, 889 und Rebora et al. 2018, S. 602 f.) bis zu mindestens 5000 Wörtern

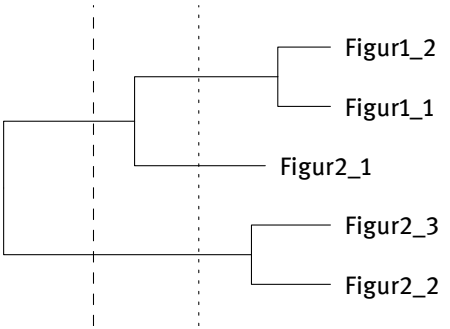


Abb. 2: Dendrogramm, das in verschiedene Cluster geteilt wird.

(vgl. Eder 2015, S. 180).[23] Prinzipiell dürfte aber gelten, dass stilometrische Analysen mit zunehmender Textlänge auch zuverlässigere Vorhersagen ermöglichen (vgl. Eder 2015, S. 167–182). Für den vorliegenden Fall bedeutet das, einen Kompromiss zu finden zwischen einer angemessenen Länge der Redesegmente auf der einen und einer möglichst großen Zahl an Vergleichsobjekten, also Segmenten der Figurenrede, auf der anderen Seite. Da die Evaluation potenziell distinktiver Figurenrede keinesfalls trivial ist – die Ergebnisse lassen sich nicht als besser oder schlechter, sondern höchstens als plausibel oder weniger plausibel taxieren –, ist ein alternatives Verfahren nötig, um den Einfluss der Redelänge auf die Clusterergebnisse zu bemessen. Ein recht simpler Versuchsaufbau leistet hierbei Abhilfe, auch wenn es sich lediglich um eine Annäherung handeln kann. Indem die Evaluation auf die Zuordnung eines Redesegments zu dem ihm zugehörigen Drama ausgelagert wird, lassen sich die Ergebnisse anhand der anders perspektivierten Fragestellung näherungsweise beurteilen. Dazu wird aus Gründen der Transparenz die sogenannte *cluster purity* berechnet, die den Prozentsatz der Mehrheitsklasse in jedem Cluster angibt. Die Zahl der zu betrachtenden Cluster wird vorab bestimmt (vgl. Manning et al. 2009, S. 356–360). Abbildung 2 zeigt eine hypothetische Clusteranalyse, die das Vorgehen anhand von fünf Figurensegmenten veranschaulicht. Die gestrichelte Linie teilt das Dendrogramm in zwei Cluster. Der ‚Figur2_1' benannte Datenpunkt wäre in diesem Szenario dem Cluster der Klasse

23 Siehe dazu auch die folgenden Aussagen: „For stylometric reliability the minimum sample size allowed is 1000 words" (Holmes et al. 2001, S. 406) und „[i]t is evident that, with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship" (Burrows 2002, S. 276).

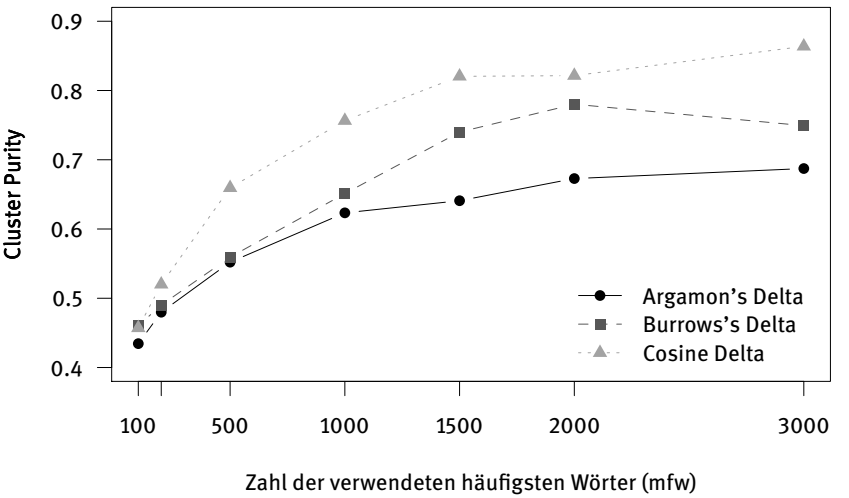


Abb. 3: *Cluster purity* für verschiedene Distanzmaße. Segmentlänge: 500 Wörter, Ward.D2 Clustering, 100–3000 mfw, 60 Clusters.

‚Figur1' zugehörig. Durch die fehlerhafte Zuordnung sinkt die *cluster purity* auf 80 Prozent. Die gepunktete Linie würde das Dendrogramm hingegen in drei Cluster gliedern, wodurch ‚Figur2_1' ein eigenes Cluster ausbildet.

Im Folgenden vergleiche ich die Performanz verschiedener stilometrischer Analysemodelle, um anhand der Ergebnisse sowohl die Wahl der Segmentgröße als auch die der Analyseparameter möglichst reflektiert gestalten zu können. Abbildung 3 stellt dabei die Leistungsfähigkeit dreier bekannter Distanzmaße gegenüber:[24] Burrows's Delta (vgl. Burrows 2002, S. 267–287), Argamon's Delta (vgl. Argamon 2008, S. 131–147) und Cosine Delta (vgl. Evert, Proisl et al. 2017, S. ii4–ii16).[25] Die Textgrundlage bilden 969 Redesegmente mit einer Mindestlänge von 500 Wörtern. Getestet werden nicht nur die verschiedenen Distanzmaße, sondern auch verschiedene Wortumfänge, die bei der Analyse berücksichtigt werden: Diese reichen von den 100 häufigsten bis zu den 3000 häufigsten Wörtern (mfw). Die abgebildeten Werte beruhen auf der *cluster purity*, für deren Berech-

24 Diese und die folgenden Analyseergebnisse wurden mit Hilfe des R-Pakets *Stylo* berechnet (Eder et al. 2016).

25 Burrows's Delta basiert auf der *manhatten distance*, Argamon's Delta auf der *euclidean distance* und Cosine Delta auf der *cosine distance*. Argamon's Delta bezieht sich hier auf das von ihm vorgeschlagene Maß *quadratic Delta* (vgl. Argamon 2008, S. 135 f.).Für eine Einführung in die stilometrisch genutzten Distanzmaße vgl. Jannidis 2014, S. 178–189.

Tab. 1: Verhältnis von Segmentlänge und Zahl der Redesegmente im Korpus.

Segmentlänge in Wörtern	Zahl der Redesegmente
500	972
600	842
700	718
800	602
900	517
1000	442

nung 60 Cluster zugrunde gelegt wurden: Das entspricht der Zahl der untersuchten Dramen, also dem schwerstmöglichen Szenario. Je höher die Werte ausfallen, desto einheitlicher sind die einzelnen Klassen auf die Cluster verteilt, desto zuverlässiger ist also die Voraussage, ob die Redesegmente dem gleichen Dramentext entstammen. Anhand der Abbildung wird nun zweierlei deutlich. Cosine Delta erzielt durchgängig die höchsten Werte, ist für die gewählte Fragestellung also performanter als Burrows's Delta und Argamon's Delta. Zugleich steigt die Genauigkeit der Analysen mit der Zahl der häufigsten Wörter, die in die Berechnungen einbezogen werden, kontinuierlich an.[26]

In einem zweiten Schritt soll nun der Einfluss der Segmentlänge auf die Clusterergebnisse näher ergründet werden. Dazu wurden verschieden gewählte Mindestumfänge der Redesegmente zwischen 500 und 1000 Wörtern veranschlagt und aus den 60 Dramentexten extrahiert. Tabelle 1 gibt das Verhältnis von Segmentlänge und Anzahl der Redesegmente wieder. Durch die schrittweise Verdoppelung des Redeumfangs reduziert sich die Zahl der Segmente um mehr als die Hälfte.

In Abbildung 4 steht dementsprechend nicht mehr die Performanz der einzelnen Distanzmaße im Vordergrund – alle hier veranschaulichten Werte wurden mittels Cosine Delta berechnet –, sondern die Abhängigkeit der Analyseergebnisse von der Länge der Redesegmente. Die dargestellten Werte sind jedoch weniger eindeutig zu interpretieren als diejenigen aus Abbildung 3. Zwar steigt die *cluster purity* mit zunehmender Segmentlänge tendenziell an, unklar bleibt an dieser Stelle allerdings, ob der dadurch bedingte Wegfall von Redesegmenten zu Erkenntniseinbußen bei der eigentlich fokussierten Fragestellung führt. Den besten Kompromiss scheint deshalb die Segmentlänge von mindestens 600 Wörtern zu bieten. Hier bleibt einerseits die Mehrzahl der Redesegmente erhalten (842)

26 Diese Ergebnisse decken sich mit der Evaluationsstudie von Evert, Proisl et al. (2017).

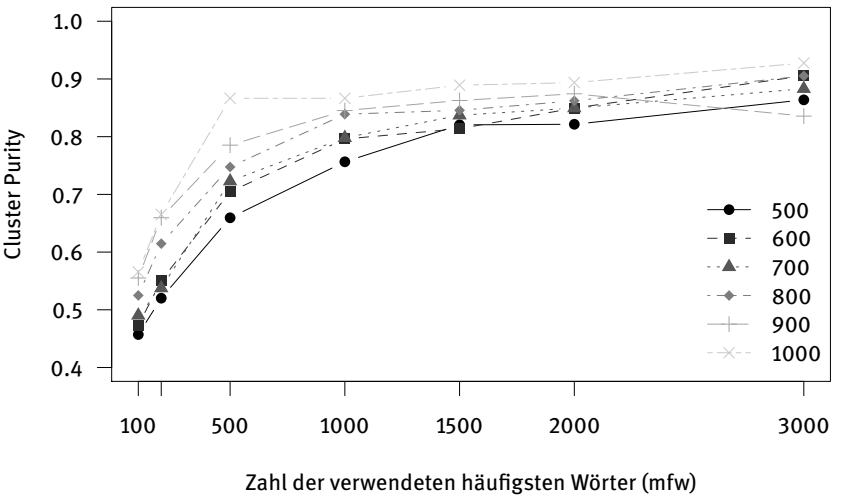


Abb. 4: *Cluster purity* für verschiedene Segmentlängen (in Wörtern): Cosine Delta, Ward.D2 Clustering, 100–3000 mfw, 60 Cluster.

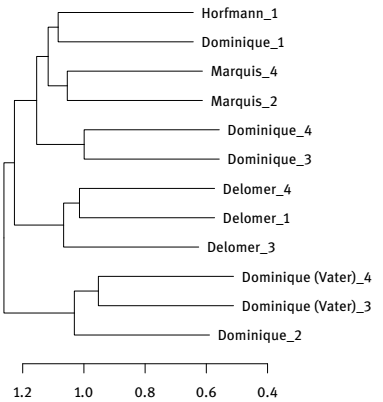
und andererseits entsprechen die Werte der *cluster purity* (2000 mfw: 0,87, 3000 mfw: 0,91) nahezu denjenigen der zumindest 1000 Wörter langen Segmente.[27]

Für die weiteren Analysen ergeben sich somit die folgenden Rahmenbedingungen: Die Segmentlänge der untersuchten Redeanteile wird auf mindestens 600 Wörter festgelegt. Die Berechnung der Distanzen erfolgt mittels Cosine Delta und die Zahl der untersuchten Wörter wird auf die 2000 (3000, 1000) häufigsten festgelegt.

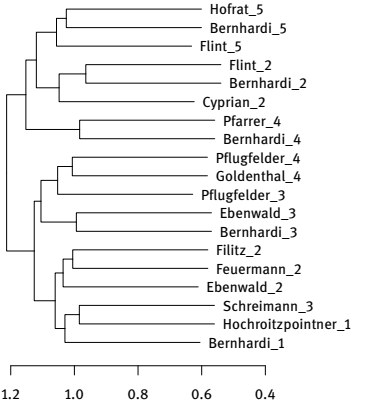
4 Analyseergebnisse

Lassen sich die dramatischen Figuren nun also anhand der sprachlichen Gestaltung ihrer Figurenrede distinguieren? Sollte es Dramatiker*innen tatsächlich gelingen, die ‚Stimmen' ihrer Figuren jeweils distinktiv auszuarbeiten, müsste dies – gemäß der von Hoover gewählten Operationalisierung – in stilometrischen Clusteranalysen augenscheinlich werden. Die entstehenden Dendrogramme sollten

27 Verschiedene Tests mit moderatem *culling*, also der Beschränkung des Wortmaterials auf die Wörter, die in einer gewissen Zahl an Vergleichstexten enthalten sind, führten zu keiner Verbesserung der Resultate. Im Gegenteil: die *cluster purity* nahm sogar leicht ab.



(a) Dendrogramm von Ifflands *Das Erbtheil des Vaters* (1802): Cosine Delta, 1000 mfw, Ward.D2 Clustering.



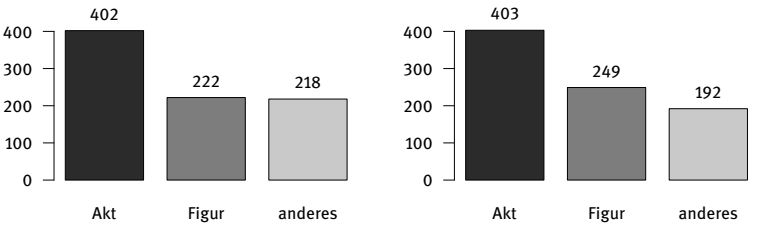
(b) Dendrogramm von Schnitzlers *Professor Bernhardi* (1912): Cosine Delta, 3000 mfw, Ward.D2 Clustering.

Abb. 5: Vergleich zweier Clusteranalysen.

dann mit demjenigen vergleichbar sein, das in Abbildung 5a exemplarisch aufgeführt wird. Denn die Figuren in August Wilhelm Ifflands *Das Erbtheil des Vaters* (1802) gruppieren sich dort, von nur wenigen Ausnahmen abgesehen, anhand ihrer Redesegmente, d. h.ihren Redeanteilen pro Akt, indiziert durch die Zahl nach dem Unterstrich. Verglichen mit dem übrigen Bühnenpersonal sind sich also die Redesegmente stilometrisch betrachtet am ähnlichsten, die derselben Figur zugehörig sind. Im Besonderen trifft das auf Delomer zu, dessen Segmente einen gänzlich eigenen Ast besetzen. Das Dendrogramm sekundiert damit die Beobachtungen Hoovers weitestgehend.

Im Korpus finden sich, wie Abbildung 5b veranschaulicht, jedoch auch Dramen, die einem anderen stilometrischen Ordnungsprinzip zu unterliegen scheinen. In Arthur Schnitzlers paratextuell als Komödie ausgezeichnetem Stück *Professor Bernhardi* sind nicht die zusammengehörigen Redesegmente einzelner Figuren als strukturgebend zu erkennen, sondern eher die Aktgrenzen, anhand derer die Figurenrede segmentiert wurde. So lassen sich mehrere Gruppierungen erkennen, insbesondere für die Akte fünf und zwei, die die Redesegmente gemäß ihrer Aktzugehörigkeit anordnen.

Um die Analyseergebnisse nicht nur auf einzelne Beispiele zu beschränken, habe ich die Distanztabellen, auf denen die Dendrogramme basieren, systematisch ausgewertet. Für jedes Redesegment der 60 Dramen wurde der sogenannte *nearest neighbor* bestimmt, also dasjenige andere Segment mit der geringsten stilometrischen Distanz, oder anders formuliert: der größten Ähnlichkeit. In Abhän-



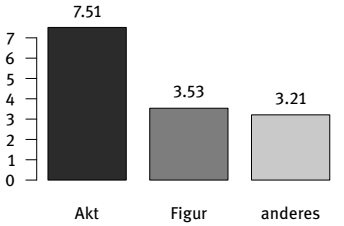
(a) 2000 mfw. **(b)** 1000 mfw.

Abb. 6: Auswertung der Distanztabellen von 60 Dramen und Einteilung in die Klassen *Akt*, *Figur* und *anderes*.

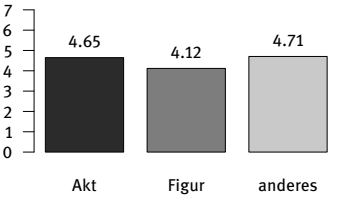
gigkeit von den Eigenschaften des *nearest neighbor* erfolgte eine Einteilung in drei Klassen: *Figur*, *Akt*, *anderes*. Sind das in Frage stehende Redesegment und dessen *nearest neighbor* der gleichen dramatischen Figur zugehörig, wird es der Klasse *Figur* zugeordnet. Entstammen die beiden Segmente dagegen dem gleichen Akt, wird es folgerichtig der Klasse *Akt* zugerechnet. Trifft keiner der beiden genannten Fälle zu, wird das Redesegment unter *anderes* verbucht. Abbildung 6 zeigt die Resultate für Berechnungen mit den 1000 und 2000 häufigsten Wörtern.

Aus den beiden Balkendiagrammen (Abbildung 5) wird ersichtlich, dass die Klasse *Akt* die meisten Redesegmente beinhaltet. Es sind jeweils knapp über 400. Die Klasse schließt damit erheblich mehr Redesegmente ein als die beiden anderen Klassen *Figur* und *anderes*. Die Diagramme zeigen auch, dass die Ergebnisse in den beiden Frequenzbereichen tendenziell stabil bleiben. Nutzt man für die stilometrischen Analysen lediglich die 1000 häufigsten Wörter, steigt die Zahl derjenigen Segmente leicht an, die einem Segment derselben Figur am ähnlichsten sind (249), die Klasse *anderes* verliert indes Zuschreibungen in fast identischem Umfang. Was bedeuten diese Ergebnisse aber für die Frage nach der sprachstilistischen Gestaltung der Figurenrede? Sprechen die einzelnen Figuren tatsächlich stilometrisch distinktiv? Die hier angeführten Analysen legen nahe, dass vor allem die Kopräsenz von Figuren stilometrische Ähnlichkeit erzeugt. Redesegmente, die im selben Akt geäußert werden – für die Figuren heißt das im Umkehrschluss, dass sie in diesem Akt (gemeinsam) auf der Bühne stehen –, scheinen durch vergleichbare Wortfrequenzen charakterisiert zu sein. Ursächlich für diese Ähnlichkeit könnten gemeinsame Themen sein, die in den Dialogen der kopräsenten Figuren diskutiert werden (vgl. Krautter 2018a, S. 227).

Abbildung 7 versucht, die gerade dargestellten Ergebnisse zu spezifizieren, indem die Werte für versifizierte und in Prosa verfasste Dramen unterschieden werden. Da die Zahl der versifizierten Dramen weit geringer ausfällt – nur 17 Dramen



(a) Dramen in Prosa. Normalisiert anhand der Zahl der Dramen.

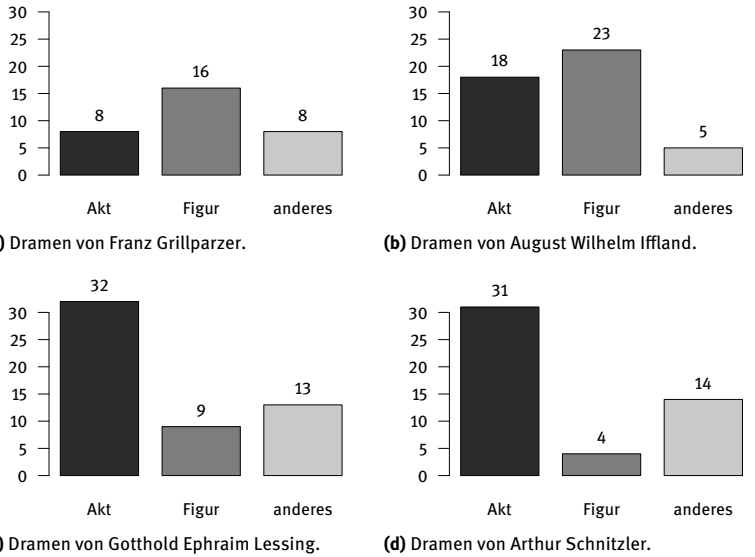


(b) Versifizierte Dramen. Normalisiert anhand der Zahl der Dramen.

Abb. 7: Auswertung der Distanztabellen von 17 versifizierten und 43 prosaischen Dramen und Einteilung in die Klassen *Akt*, *Figur* und *anderes*. 2000 mfw.

im Korpus sind in Versen abgefasst – als diejenige der in Prosa gehaltenen, wurden die Werte normalisiert. Es handelt sich also um Mittelwerte. Diese sind jedoch durchaus aufschlussreich, zeigen sie doch unterschiedliche Muster. So scheint die in Abbildung 6 angezeigte Diskrepanz der Klassen *Figur* und *Akt* vor allem auf nicht-versifizierte Dramen zurückführbar zu sein. Eine tiefergreifende Interpretation der Datenwerte hätte allerdings eine präzisere funktionsgeschichtliche Kontextualisierung verschiedener Dramengattungen zur Voraussetzung. Denn Gattungskonventionen sind sowohl durch „*normbildende Werke*" als auch „durch die wechselseitige *Komplementarität von Gattungserwartungen und Werkantworten*" beeinflusst (vgl. Voßkamp 1977, S. 30). Während Johann Christoph Gottsched in der ersten Hälfte des 18. Jahrhunderts – angelehnt an die französische Regelpoetik – für „regelmåßige Tragǒdien in Versen" zur „Verbesserung der deutschen Schaubǔhne" wirbt (Gottsched 1970, S. 9), favorisieren die Stürmer und Dränger, die im späteren 18. Jahrhundert mit Blick auf Shakespeare an einer möglichst „natürlichen Figurengestaltung" interessiert sind (Reiter und Willand 2018, S. 67), prosaische Tragödien. Sprachliche Unterschiede hängen somit auch von der konzeptionellen Zusammensetzung des Bühnenpersonals ab.[28] In seinen *Gedanken zur Aufnahme des dänischen Theaters* (1764 [1747]) bestimmt Johann Elias Schlegel die Tragödie etwa als „Handlungen hoher Personen, welche die Leidenschaften erregen." (Schlegel 2003, S. 90) Die sprachliche Ausarbeitung der Figurenrede sei dabei direkt an die Anlage der Dramenfiguren geknüpft, so dass „der geringste Fehler im Ausdrucke" einem „Fehler im Charakter" gleich komme (Schlegel 2003, S. 104). Von einem ‚hohen' Bühnenpersonal ist hier also auch eine ‚hohe' in Ver-

28 An dieser Stelle sei nochmals auf Schillers Drama *Die Räuber* verwiesen, in dem eine ausgeprägte „Grobheit der Räuber-Sprache" wahrnehmbar sei (Ritzer 2011, S. 257).



(a) Dramen von Franz Grillparzer.

(b) Dramen von August Wilhelm Iffland.

(c) Dramen von Gotthold Ephraim Lessing.

(d) Dramen von Arthur Schnitzler.

Abb. 8: Auswertung der Distanztabellen von zwölf Dramen, die sich auf vier Autoren verteilen. 2000 mfw.

sen gehaltene Sprache zu erwarten. Auch für die Versifizierung selbst ergeben sich Anschlussfragen: In welchem Versmaß sind die Dramen gehalten? Liegt eine durchgängige Reimstruktur vor?

Zuletzt drängt sich die Frage auf, ob denn nicht auch bei unterschiedlichen Autor*innen auseinandergehende Muster zu entdecken sein könnten. Abbildung 8 vergleicht die Dramen von vier exemplarisch gewählten Autoren: Gotthold Ephraim Lessing (*Miß Sara Sampson* [1755], *Minna von Barnhelm, oder das Soldatenglück* [1767], *Nathan der Weise* [1779]), August Wilhelm Iffland (*Die Jäger* [1785], *Figaro in Deutschland* [1790], *Das Erbtheil des Vaters* [1802]), Franz Grillparzer (*Die Ahnfrau* [1817], *Sappho* [1818], *Des Meeres und der Liebe Wellen* [1831]) und Arthur Schnitzler (*Der einsame Weg* [1904], *Das weite Land* [1910], *Professor Bernhardi* [1912]).

Die hier angeführten Autoren wurden gezielt ausgewählt, um zwei verschiedene Verteilungsmuster deutlich zu machen. Während bei Grillparzer und Iffland in der Tendenz eine stilometrische Ähnlichkeit der Redesegmente einer Figur wahrzunehmen ist – hier also *distinctive voices* vorliegen könnten –, zeichnet die Dramen von Lessing und Schnitzler eine umgekehrte Verteilung der Werte aus. Prävalent sind hier Redesegmente, die sich entsprechend des zugehörigen Aktes grup-

pieren. Instruktiv könnte dahingehend auch die nähere Betrachtung von Heinrich Laubes Stück *Die Karlsschüler* (1846) sein, das den jungen Friedrich Schiller in den Mittelpunkt der Handlung rückt. Dabei verspricht insbesondere die gezielte Verbindung qualitativer und quantitativer Methoden einen fruchtbaren Zugriff.[29] Denn gerade die Dichterfigur Schiller, so zeigen es die Analysen, scheint im Drama die einzige Figur zu sein, deren Sprache sich stilometrisch von den anderen Bühnenfiguren abgrenzen lässt. Laube selbst gibt in der Einleitung des Schauspiels erste Analysehinweise, wenn er erläutert, den „dreiundzwanzigjährigen Jüngling herausschälen" zu wollen und nicht „den großen Poeten Schiller zum Helden" zu machen. Somit müsste er sich auch nicht an einem Stil ausrichten, „wie er im Wallenstein und den ähnlichen Werken Schillers herrscht", sondern könnte stattdessen seiner Idealvorstellung eines „natürlichen Stil[s]" folgen (Laube 1909, S. 161).

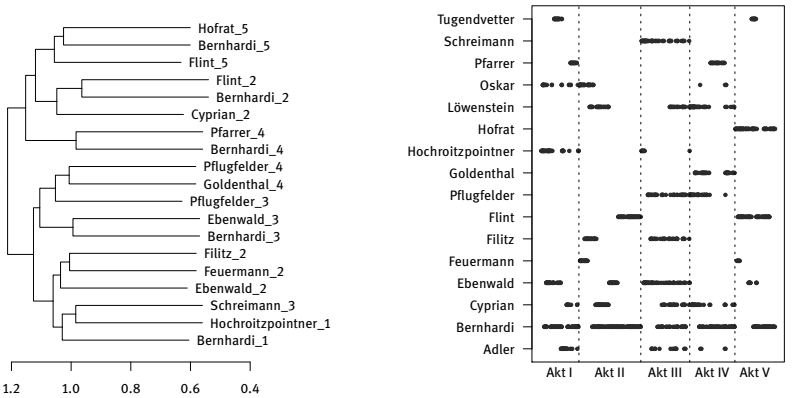
5 Diskussion und Ausblick

Neben den von Burrows, Craig, Hoover usw. betonten spezifischen Idiolekten einer Figur, also den sogenannten *distinctive voices*, scheint in den betrachteten Dramen ein zweites, vielleicht sogar stärkeres stilometrisches Signal deutlich zu werden: die gemeinsame Bühnenpräsenz der Figuren innerhalb eines Aktes. Eine genauere Untersuchung dieses Zusammenhangs steht allerdings noch aus.[30]

Vergleicht man die bereits gezeigte Clusteranalyse von Schnitzlers *Professor Bernhardi* mit einer Darstellung der Bühnenpräsenz, werden erste Zusammenhänge besser greifbar. So verdeutlicht Abbildung 9, dass das stilometrische Cluster aus Hofrat, Bernhardi und Flint – die drei Redesegmente sind dem fünften Akt entnommen – genau in dieser Konfiguration im fünften Akt auf der Bühne steht. Die Präsenzdarstellung listet die 16 meistsprechenden Figuren des Stücks. Jede Redeäußerungen einer Figur wird im Verlauf des Dramas von links nach rechts durch einen Punkt markiert. Ähnlich markant wie die Figurenkonfiguration im fünften Akt scheint die Interaktion zwischen Bernhardi und dem Pfarrer in Akt IV zu sein. Mit Auftritt des Pfarrers verlassen ausgenommen von Professor Bernhardi alle übrigen Figuren die Bühne. Die gemeinsame Kopräsenz von Pfarrer und Bernhardi spiegelt sich sodann in ihrer stilometrisch berechneten Nähe wider. Die beiden Redesegmente gruppieren sich auf einem eigenen Ast im Dendrogramm. Künftige

29 Für diese Verbindung prägte Martin Mueller den Begriff *scalable reading*. Siehe dazu Mueller 2014, S. 27–38 und Weitin 2017, S. 1–6.

30 Erste Überlegungen dazu finden sich in Krautter 2018a, S. 225–228.



(a) Clusteranalyse von Schnitzlers *Professor Bernhardi* (1912). Cosine Delta, 3000 mfw, Ward.D2 Clusterung.

(b) Bühnenpräsenz der 16 meistsprechenden Figuren in *Professor Bernhardi* (1912).

Abb. 9: Zusammenhang zwischen stilometrischer Ähnlichkeit und Bühnenpräsenz der Figuren

Analysen, die sich intensiver mit den Mikrostrukturen einzelner Dramen auseinandersetzen, haben zur Aufgabe, die Gründe dieser stilometrischen Ähnlichkeit umfassender zu erörtern. Wird hier der Redestil abgebildet, also die rhetorische Gestaltung der Figuren, oder handelt es sich eher um die thematische Gestaltung der dramatischen Handlung, die in der Segmentierung der Figurenrede augenscheinlich wird?

Der hier skizzierte explorative Zugang ließe sich in Zukunft weiter vertiefen, etwa durch Metriken der sozialen Netzwerkanalyse. So könnte in größeren Korpora ausgewertet werden, ob sich die stilometrisch ähnlichsten Figuren auch in Netzwerkgraphen durch eine starke Verbindung – beispielsweise anhand eines hohen *weighted degree*-Werts – auszeichnen.

Welche Folgerungen lassen die vorgestellten Befunde nun zu? Sollte die Vorstellung, dass Dramatiker*innen sprachlich individuell gestaltete Figuren entwerfen, verabschiedet werden? Zwar deuten die Ergebnisse der Analysen vordergründig nicht auf übermäßig viele stilometrisch distinktiv angelegte Figuren hin, es ist aber auch nicht anzunehmen, dass Autor*innen jeder (wichtigen) Dramenfigur eine sie spezifisch auszeichnende Stimme zu verleihen intendieren, etwa wenn ganz gezielt holzschnittartige Typen vorgeführt werden sollen. Abhängig von Produktions- und Rezeptionsbedingungen, man denke an die in der Theaterpraxis des 18. Jahrhunderts dominanten Rollenfächer (vgl. Harris 1992, S. 221–235), haben regelpoetische Setzungen oder poetologische Überlegungen starken Einfluss auf die Gestaltung der Dramen. Gemessen an den *cluster-purity*-Werten, zeigen sich die Analysen der einzelnen Dramen zudem ein wenig volatil.

Das mag auch an der Segmentierung der Figurenrede liegen, die vor allem die literaturwissenschaftliche Interpretierbarkeit der Analyseergebnisse als Hintergedanken trägt. Arbiträre oder sogar randomisierte Segmentierungen müssten den hier illustrierten Ergebnissen in künftigen Untersuchungen als Vergleichsfolie dienen.

Letztlich müssen die durchgeführten Analysen vor allem als eine erste Standortbestimmung bewertet werden, die nach einer noch systematischeren Untersuchung verlangt. Für eine abschließende Beurteilung bleiben zu viele Einflussvariablen offen. Denn eine individuell gestaltete Figurensprache erschöpft sich nicht in Wortfrequenzen und den dadurch implizit einbezogenen syntaktischen Strukturen. Welchen Einfluss üben unterschiedliche Satz- und Replikenlängen aus? Wie äußern sich längere Erzählpassagen, also Botenberichte oder Teichoskopien? Und in welchem Verhältnis stehen Redesemantiken und Sprachstil? Obwohl viele dieser Fragen unbeantwortet bleiben müssen, bietet der Beitrag einen differenzierteren Blick auf die stilometrisch ergründete Gestaltung der dramatischen Figurenrede. Er will dadurch die Einsicht in einen für stilometrische Betrachtungen bislang kaum erörterten Gegenstand öffnen: die Strukturierung des Dramas durch die Autor*innen, wobei insbesondere die Kopräsenz von Figuren relevant zu sein scheint.

Funding: Dieser Beitrag entstand im Rahmen des des von der VolkswagenStiftung geförderten *mixed-methods*-Projekts „Quantitative Drama Analytics“ (QuaDramA).

Verwendete Dramen

Autor*in	Titel
Anzengruber, L.	Der Pfarrer von Kirchfeld (1870)
Anzengruber, L.	Der Meineidbauer (1871)
Anzengruber, L.	Der Gwissenswurm (1874)
Birch-Pfeiffer, Ch.	Johannes Gutenberg (1835)
Birch-Pfeiffer, Ch.	Vatersorgen (1849)
Birch-Pfeiffer, Ch.	In der Heimath (1865)
Goethe, J. W.	Egmont (1788)
Goethe, J. W.	Torquato Tasso (1790)
Goethe, J. W.	Der Großkophta (1792)
Grabbe, Chr. D.	Herzog Theodor von Gothland (1827)
Grabbe, Chr. D.	Scherz, Satire, Ironie und tiefere Bedeutung (1827)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Autor*in	Titel
Grabbe, Chr. D.	Don Juan und Faust (1829)
Grillparzer, F.	Die Ahnfrau (1817)
Grillparzer, F.	Sappho (1818)
Grillparzer, F.	Des Meeres und der Liebe Wellen (1831)
Hauptmann, C.	Ephraims Breite (1900)
Hauptmann, C.	Gaukler, Tod und Juwelier (1917)
Hauptmann, C.	Tobias Buntschuh (1920)
Hebbel, F.	Genoveva (1843)
Hebbel, F.	Herodes und Mariamne (1849)
Hebbel, F.	Agnes Bernauer (1852)
Hofmannsthal, H. v.	Ödipus und die Sphinx (1906)
Hofmannsthal, H. v.	Der Schwierige (1917)
Hofmannsthal, H. v.	Der Turm (1924)
Holz, A.	Die Familie Selicke (1890)
Holz, A.	Sonnenfinsternis (1908)
Holz, A.	Ignorabimus (1914)
Iffland, A. W.	Die Jäger (1785)
Iffland, A. W.	Figaro in Deutschland (1790)
Iffland, A. W.	Das Erbtheil des Vaters (1802)
Immermann, K.	Das Gericht von St. Petersburg (1832)
Immermann, K.	Merlin (1832)
Immermann, K.	Andreas Hofer, der Sandwirt von Passeyer (1835)
Kleist, H. v.	Die Familie Schroffenstein (1803)
Kleist, H. v.	Das Käthchen von Heilbronn oder die Feuerprobe (1810)
Kleist, H. v.	Prinz Friedrich von Homburg (1810)
Klinger, F. M.	Die neue Arria (1776)
Klinger, F. M.	Die Zwillinge (1776)
Klinger, F. M.	Simsone Grisaldo (1776)
Kotzebue, A. v.	Die Indianer in England (1789)
Kotzebue, A. v.	Menschenhaß und Reue (1790)
Kotzebue, A. v.	Die beiden Klingsberg (1801)
Laube, H.	Monaldeschi (1841)
Laube, H.	Gottsched und Gellert (1845)
Laube, H.	Die Karlsschüler (1846)
Lessing, G. E.	Miß Sara Sampson (1755)
Lessing, G. E.	Minna von Barnhelm, oder das Soldatenglück (1767)
Lessing, G. E.	Nathan der Weise (1779)
Schiller, F.	Die Räuber (1781)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Autor*in	Titel
Schiller, F.	Don Karlos, Infant von Spanien (1787)
Schiller, F.	Maria Stuart (1800)
Schlegel, J. E.	Der geschäftige Müßiggänger (1743)
Schlegel, J. E.	Canut (1746)
Schlegel, J. E.	Der Triumph der guten Frauen (1748)
Schnitzler, A.	Der einsame Weg (1904)
Schnitzler, A.	Das weite Land (1910)
Schnitzler, A.	Professor Bernhardi (1912)
Weißenthurn, J. v.	Johann, Herzog von Finnland (1811)
Weißenthurn, J. v.	Welche ist die Braut! (1813)
Weißenthurn, J. v.	Das Manuscript (1817)

Literatur

Argamon, Shlomo (2008). „Interpreting Burrows's Delta: Geometric and Probabilistic Foundations". In: *Literary and Linguistic Computing* 23.2, S. 131–147. DOI: 10.1093/llc/fqn003.

Aristoteles (1973). *Dichtkunst. Ins Deutsche übersetzt, mit Anmerkungen und besonderen Abhandlungen versehen von Michael Conrad Curtius [1753]*. Übers. von Michael Conrad Curtius. Nachdruck der Ausgabe von 1753. Hildesheim und New York [NY]: Georg Olms Verlag.

Aristoteles (1982). *Poetik. Griechisch/Deutsch*. Hrsg. und übers. von Manfred Fuhrmann. Stuttgart: Reclam.

Bakhtin, Mikhail (1994). *The Problems of Dostoevsky's Poetics*. Hrsg. und übers. von Caryl Emerson. 6. Auflage. Minneapolis [MN] und London: University of Minnesota Press.

Bockwinkel, Peggy (2018). „Wie anders ist Figurenrede? Die Rolle der direkten Rede in quantitativen Erzähltextanalysen". In: *Digital Humanities: Perspektiven der Praxis*. Hrsg. von Peggy Bockwinkel, Beatrice Nickel und Gabriel Viehhauser. Berlin: Frank & Timme, S. 117–148.

Burrows, John (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Burrows, John (2002). „'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3, S. 267–287. DOI: 10.1093/llc/17.3.267.

Burrows, John und Hugh Craig (2012). „Authors and Characters". In: *English Studies* 93.3, S. 292–309. DOI: 10.1080/0013838X.2012.668786.

Craig, Hugh (1999). „Contrast and Change in the Idiolects of Ben Jonson Characters". In: *Computers and the Humanities* 33.3, S. 221–240.

Craig, Hugh und Arthur F. Kinney (2009). *Shakespeare, Computers and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, Maciej (2015). „Does size matter? Authorship attribution, small samples, big problem". In: *Digital Scholarship in the Humanities* 30.2, S. 167–182. DOI: 10.1093/llc/fqt066.

Eder, Maciej, Jan Rybicki und Mike Kestemont (2016). „Stylometry with R: A Package for Computational Text Analysis". In: *The R Journal* 8.1, S. 107–121. URL: https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf (besucht am 1. Juni 2020).

Evert, Stefan, Fotis Jannidis, Thomas Proisl, Thorsten Vitt, Christof Schöch, Steffen Pielström und Isabella Reger (2016). „Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution". In: *Digital Humanities 2016 Conference Abstracts*. Hrsg. von Maciej Eder und Jan Rybicki. Krakau: Jagiellonian University & Pedagogical University, S. 188–191. URL: https://dh2016.adho.org/abstracts/static/dh2016_abstracts.pdf (besucht am 1. Juni 2020).

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch und Thorsten Vitt (2017). „Understanding and Explaining Delta Measures for Authorship Attribution". In: *Digital Scholarship in the Humanities* 32. Supplement 2, S. ii4–ii16. DOI: 10.1093/llc/fqx023.

Fields, Paul J., Larry Bassist und Matt Roper (2017). „Characters in 19th Century Novels Display Distinctive Voices as Seen by Stylometric Analysis". In: *Digital Humanities 2017 Conference Abstracts*. Montreal. URL: https://dh2017.adho.org/abstracts/494/494.pdf (besucht am 1. Juni 2020).

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling und Peer Trilcke (2019). „Programmable Corpora: Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor". In: *DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt a.M. und Mainz: Digital Humanities im deutschsprachigen Raum e.V., S. 194–197. DOI: 10.5281/zenodo.2596095.

Gottsched, Johann Christoph (1970). „Der sterbende Cato. Des Herrn Verfassers Vorrede, zur ersten Ausgabe 1732". In: *Johann Christoph Gottsched: Ausgewählte Werke*. Hrsg. von Joachim Birke. Bd. 2. Berlin: De Gruyter, S. 3–18.

Harris, Edward P. (1992). „Lessing und das Rollenfachsystem. Überlegungen zur praktischen Charakterologie im 18. Jahrhundert". In: *Schauspielkunst im 18. Jahrhundert. Grundlagen, Praxis, Autoren*. Hrsg. von Wolfgang F. Bender. Stuttgart: Steiner, S. 221–235.

Hiatt, Tim und John Hilton (1990). „Can Authors Alter Their Wordprints? Faulkner's Narrators in As I Lay Dying". In: *Desert Language and Linguistic Society Symposium* 16.1, S. 52–62.

Holmes, David I., Lesley J. Gordon und Christine Wilson (2001). „A Widow and her Soldier: Stylometry and the American Civil War". In: *Literary and Linguistic Computing* 16.4, S. 403–420.

Hoover, David L. (2017). „The Microanalysis of Style Variation". In: *Digital Scholarship in the Humanities* 32. Supplement 2, S. ii17–ii30. DOI: 10.1093/llc/fqx022.

Hoy, Cyrus (1956). „The Shares of Fletcher and His Collaborators in the Beaumont and Fletcher Canon (I)". In: *Studies in Bibliography* 8, S. 129–146.

Jannidis, Fotis (2014). „Der Autor ganz nah: Autorstil in Stilistik und Stilometrie". In: *Theorien und Praktiken der Autorschaft*. Hrsg. von Matthias Schaffrick und Marcus Willand. Berlin und Boston [MA]: De Gruyter, S. 169–195.

Jannidis, Fotis, Steffen Pielström, Christof Schöch und Thorsten Vitt (2015). „Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures". In: *Digital Humanities 2015 Conference Abstracts*. Sydney, S. 1–10.

Jockers, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana [IL], Chicago [IL] und Springfield [IL]: University of Illinois Press.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Krautter, Benjamin (2018a). „Quantitative Microanalysis? Different Methods of Digital Drama Analysis in Comparison". In: *Digital Humanities 2018 Conference Abstracts*. Mexiko Stadt, S. 225–228. URL: https://dh2018.adho.org/en/quantitative-microanalysis-different-methods-of-digital-drama-analysis-in-comparison/ (besucht am 1. Juni 2020).

Krautter, Benjamin (2018b). „Quantitatives ‚close reading'? Vier mikroanalytische Methoden der digitalen Dramenanalyse im Vergleich". In: *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts*. Köln: Digital Humanities im deutschsprachigen Raum e.V., S. 295–300. URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf (besucht am 1. Juni 2020).

Lancashire, Ian (2010). *Forgetful Muses: Reading the Author in the Text*. Toronto, Buffalo [NY] und London: University of Toronto Press.

Laube, Heinrich (1909). „Die Karlsschüler. Einleitung des Verfassers [1846]". In: *Heinrich Laube: Gesammelte Werke in fünfzig Bänden*. Hrsg. von Heinrich Hubert Houben. Bd. 25. Leipzig: Max Hesses Verlag, S. 149–181.

Lenz, Jakob Michael Reinhold (1987). „Anmerkungen übers Theater [1774]". In: *Jakob Michael Reinhold Lenz: Werke und Briefe in drei Bänden*. Hrsg. von Sigrid Damm. Bd. 2. München und Wien: Carl Hanser Verlag, S. 641–671.

Lessing, Gotthold Ephraim (1985a). „Hamburgische Dramaturgie [1767/1769]". In: *Gotthold Ephraim Lessing: Werke und Briefe in 12 Bänden*. Hrsg. von Wilfried Barner. Bd. 6. Frankfurt a.M.: Deutscher Klassiker Verlag, S. 181–778.

Lessing, Gotthold Ephraim (1985b). „Minna von Barnhelm, oder das Soldatenglück [1767]". In: *Gotthold Ephraim Lessing: Werke und Briefe in 12 Bänden*. Hrsg. von Wilfried Barner. Bd. 6. Frankfurt a.M.: Deutscher Klassiker Verlag, S. 9–110.

Lessing, Gotthold Ephraim (1988). „Briefe von und an Lessing 1770-1776". In: *Gotthold Ephraim Lessing: Werke und Briefe in 12 Bänden*. Hrsg. von Wilfried Barner. Bd. 11/2. Frankfurt a.M.: Deutscher Klassiker Verlag.

Manning, Christopher D., Prabhakar Raghavan und Hinrich Schütze (2009). *An Introduction to Information Retrieval*. Online Edition. Cambridge: Cambridge University Press. URL: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf (besucht am 1. Juni 2020).

Masten, Jeffrey (1992). „Beaumont and/or Fletcher: Collaboration and the Interpretation of Renaissance Drama". In: *ELH* 59, S. 337–356.

Masten, Jeffrey (1997). *Textual Intercourse: Collaboration, Authorship and Sexualities in Renaissance Drama*. Cambridge: Cambridge University Press.

McMullan, Gordon (1996). „‚Our Whole Life is Like a Play': Collaboration and the Problem of Editing". In: *Textus* 9, S. 437–460.

Moretti, Franco (2000). „Conjectures on World Literature: My Mission: to Say It More Simply than I Understand It". In: *New Left Review* 1, S. 54–68.

Moretti, Franco (2013). *Distant Reading*. London: Verso.

Mueller, Martin (2014). „Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment". In: *Digital Humanities Quarterly* 8.3, S. 1–41. URL: http://digitalhumanities.org:8081/dhq/vol/8/3/000183/000183.html (besucht am 1. Juni 2020).

Rebora, Simone, J. Berenike Herrmann, Gerhard Lauer und Massimo Salgaro (2018). „Robert Musil, a War Journal, and Atylometry: Tackling the Issue of Short Texts in Authorship Attribution". In: *Digital Scholarship in the Humanities* 34.3, S. 582–605. DOI: 10.1093/llc/fqy055.

Rector, Martin (2017). „Theoretische Schriften". In: *J.M.R. Lenz Handbuch*. Hrsg. von Julia Freytag, Inge Stephan und Hans-Gerd Winter. Berlin und Boston [MA]: De Gruyter, S. 186–241.

Reiter, Nils und Marcus Willand (2018). „Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Hrsg. von Toni Bernhart, Marcus Willand, Sandra Richter und Andrea Albrecht. Berlin und Boston [MA]: De Gruyter, S. 45–76. DOI: 10.1515/9783110523300-003.

Ritzer, Monika (2011). „Schillers dramatischer Stil". In: *Schiller-Handbuch*. Hrsg. von Helmut Koopmann. 2. Aufl. Stuttgart: J.B. Metzler, S. 254–284.

Schiller, Friedrich (1956). „Briefwechsel. Schillers Briefe 1772–1785". In: *Schillers Werke: Nationalausgabe*. Hrsg. von Walter Müller-Seidel. Bd. 23. Weimar: Böhlau.

Schlegel, Johann Elias (2003). „Gedanken zur Aufnahme des dänischen Theaters [1764]". In: *Canut. Ein Trauerspiel*. Hrsg. von Horst Steinmetz. Stuttgart: Reclam, S. 75–111.

Schöch, Christof (2014). „Corneille, Molière et les autres: Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik". In: *Literaturwissenschaft im digitalen Medienwandel*. Hrsg. von Christof Schöch und Lars Schneider. Beihefte zu Philologie im Netz. PhiN, S. 130–157. URL: https://hal.archives-ouvertes.fr/hal-00957091/document (besucht am 1. Juni 2020).

Schöch, Christof (2018). „Zeta für die kontrastive Analyse literarischer Texte Theorie, Implementierung, Fallstudie". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Hrsg. von Toni Bernhart, Marcus Willand, Andrea Albrecht und Sandra Richter. Berlin und Boston [MA]: De Gruyter, S. 77–94. DOI: 10.1515/9783110523300-004.

Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez und Alberto Barrón-Cedeño (2014). „Overview of the Author Identification Task at PAN 2014". In: *Proceedings of PAN/CLEF 2014*, S. 877–897.

Stockinger, Claudia (2007). „‚Lektüre'? ‚Stil'?. Zur Aktualität der Werkimmanenz". In: *1955–2005: Emil Staiger und Die Kunst der Interpretation heute*. Hrsg. von Joachim Rickes, Volker Ladenthin und Michael Baum. Bern: Peter Lang, S. 61–85.

Van Dalen-Oskam, Karina (2014). „Epistolary Voices: The Case of Elisabeth Wolff and Agatha Deken". In: *Literary and Linguistic Computing* 29.3, S. 443–451. DOI: 10.1093/llc/fqu023.

Viehhauser, Gabriel (2015). „Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte". In: *Grenzen und Möglichkeiten der Digital Humanities*. Hrsg. von Constanze Baum und Thomas Stäcker. Sonderband der Zeitschrift für digitale Geiesteswissenschaften. URL: http://www.zfdg.de/sb001_009 (besucht am 1. Juni 2020).

Vishnubhotla, Krishnapriya, Adam Hammond und Graeme Hirst (2019). „Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama". In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis [MN]: Association for Computational Linguistics, S. 29–34. DOI: 10.18653/v1/W19-2504.

Voßkamp, Wilhelm (1977). „Gattungen als literarisch-soziale Institutionen. Zu Problemen sozial- und funktionsgeschichtlich orientierter Gattungstheorie und -historie". In: *Textsortenlehre – Gattungsgeschichte*. Hrsg. von Walter Hinck. Heidelberg: Quelle & Meyer, S. 27–44.

Weitin, Herget (2017). „Falkentopics. Über einige Probleme beim Topic Modeling literarischer Texte". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, S. 29–48. DOI: 10.1007/s41244-017-0049-3.

Weitin, Thomas, Thomas Gilli und Nico Kunkel (2016). „Auslegen und Ausrechnen“. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 46.1, S. 103–115. DOI: 10.1007/s41244-016-0004-8.

Axel Pichler, André Blessing, Nils Reiter und Mirco Schönfeld

Algorithmische Mikrolektüren philosophischer Texte

Ein auf der digitalen Netzwerkanalyse basierender Vergleich der ‚Begriffsnetzwerke' bei Adorno und Carnap

Zusammenfassung: Der Beitrag untersucht die Frage, inwiefern ausgewählte Texte von Theodor W. Adorno das von ihm propagierte Konzept eines konstellativen Begriffsgebrauchs tatsächlich realisieren, indem die Begriffsverknüpfungen in Adornos Texten mit denjenigen in ausgewählten Texten von Rudolf Carnap verglichen werden. Zur Durchführung dieses Vergleiches wird ein Konzept von Begrifflichkeit entwickelt, das es erlaubt, Begriffsreferenzen und deren Verhältnis zueinander auf der Basis von linguistischen Kriterien zu operationalisieren, in Netzwerke zu überführen und schließlich netzwerkanalytisch miteinander zu vergleichen. Die netzwerkanalytischen Vergleiche zeigen, dass beide Autoren ihren Idealvorstellungen der Begriffsverknüpfung nicht vollständig entsprechen, Adorno jedoch die seinige weitaus konsequenter umsetzt als Carnap.

Abstract: This contribution is devoted to the question of the extent to which selected texts by Theodor W. Adorno actually realize the concept of a constellative use of terms propagated by him by comparing the links between concepts in Adorno's texts with those in selected texts by Rudolf Carnap. In order to carry out this comparison, a model of conceptuality is developed which allows to operationalize term references and their relationship to one another on the basis of linguistic criteria and transfers these relations into networks which are finally compared with one another by means of network analysis. The network analytical comparisons show that both authors do not fully correspond to their ideal of concept-linkage, but that Adorno implements his own much more consistently than Carnap.

Axel Pichler, Stuttgart Research Center for Text Studies, Universität Stuttgart
André Blessing, Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Mirco Schönfeld, Bayrische Hochschule für Politik an der Technischen Universität München

Open Access. © 2020 Axel Pichler, André Blessing, Nils Reiter und Mirco Schönfeld; publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-014

1 Einleitung

Die computergestützte Textanalyse wird spätestens seit Franco Morettis *Distant Reading* (cf. Moretti 2013) mit *big data* assoziiert und als Alternative oder Ergänzung zur vermeintlich dominierenden Praxis des *close readings* einiger weniger kanonischer Texte in den traditionellen Geisteswissenschaften erachtet. Werk- bzw. textimmanente Fragen, die auf die Mikrostruktur(en) von Texten abzielen und daher mithilfe von für die jeweilige Fachdisziplin gängigen Verfahren auch textimmanent beantwortet werden zu können scheinen, zählen dieser Auffassung entsprechend nicht zu den paradigmatischen Fragen der computergestützten Textanalyse. Wir möchten im Folgenden an einem konkreten Fallbeispiel demonstrieren, inwiefern algorithmische Methoden zur Beantwortung dieses Typus von textwissenschaftlichen Fragen – d. s. Fragen, welche die Struktur oder Poetologie einzelner Texte oder kleinerer Textgruppen betreffen – beitragen können.[1] Das Fallbeispiel stammt aus der jüngeren Philosophiegeschichte und dreht sich um die Frage, inwiefern philosophische Texte die von ihnen propagierten textuellen Verfahrensweisen tatsächlich realisieren. Konkret geht es um Theodor W. Adornos Konzept eines konstellativen Begriffsgebrauchs: Folgt man Adornos Selbstbeschreibungen seiner philosophischen Verfahrensweise (siehe dazu die Abschnitte 2 und 3), unterscheidet sich diese von anderen – z. B. stärker szientistisch ausgerichteten – Ansätzen dadurch, dass sie Begriffe nicht qua Definition und Inferenz in ein hierarchisches Verhältnis zueinander setzt, sondern assoziativ miteinander verknüpft. Eine derartige Selbstbeschreibung wirft unter anderem folgende zwei Fragen auf:

1. Entsprechen die in Adornos Texten realisierten Begriffsrelationen tatsächlich Adornos Beschreibungen seines eigenen philosophischen Begriffsgebrauchs, wie die Adornoforschung beinahe einhellig annimmt (cf. Lehr 2000; Hogh 2015; Seel 2019)?
2. Unterscheiden sich die Begriffsrelationen in Adornos Texten von denjenigen in Texten alternativer zeitgenössischer philosophischer Strömungen und wenn ja, inwiefern?

1 Kenner der DH wissen, dass auf derartige Fragestellungen fokussierte Arbeiten in den digitalen Geisteswissenschaften weitaus stärker verbreitet sind, als auf Grund des medial verbreiteten *big picture* zumeist angenommen wird. Man denke nur an John F. Burrows' wegbereitende Studie zu Jane Austin, cf. Burrows 1987.

Diese beiden Fragen sollen im Folgenden durch die Analyse einer repräsentativen Auswahl von Texten Adornos und Rudolf Carnaps beantwortet werden. Die Entscheidung, Texte von Carnap als Vergleichskorpus heranzuziehen, ist der Tatsache geschuldet, dass Carnaps Philosophie gemeinhin als Inbegriff eines durch definitorische Klarheit und argumentative Stringenz gekennzeichneten Philosophierens erachtet wird; eine Charakteristik, die sich in pejorativer Form auch in Adornos Auseinandersetzung mit Carnaps Philosophie findet.

Durch die Beantwortung dieser beiden Fragen möchte der vorliegende Aufsatz erstens einen Beitrag zur Adornoforschung liefern, die sich bis dato auf abstrakte Rekonstruktionen von Adornos Ideal einer konstellativen Begriffsverknüpfung beschränkt hat, jedoch die Umsetzung dieser Verfahrensweise – zumindest unseres Wissens nach – noch nicht an der Textoberfläche von Adornos philosophischen Schriften empirisch überprüft hat.[2] Zweitens soll derartig gezeigt werden, inwiefern algorithmische Textanalyseverfahren auch für die Beantwortung von im weitesten Sinne textimmanenten Fragen, die sich auf kleine Textkorpora (‚*small data*') beziehen, fruchtbar sein können.

Dabei gehen wir folgendermaßen vor: Abschnitt 2 bietet eine kurze Darstellung von Adornos Vorstellung, wie Begriffe in der Philosophie miteinander zu verknüpfen sind: als Konstellation[3]. Im Anschluss daran werden Adornos und Carnaps Vorstellungen, wie Begriffe in der Philosophie zu verwenden und zusammenzuführen sind, rekonstruiert (Abschnitt 3) und in (Netzwerk-)Modelle überführt, die diese Vorstellungen realisieren (Abschnitt 4). Diese Modelle erfüllen drei Funktionen: Sie sollen Adornos und Carnaps Konzepte der Begriffsrelation anschaulich und empirisch überprüfbar machen, bereiten derartig deren Netzwerkanalyse vor und dienen so im weiteren Verlauf als Vergleichs-Hypothesen. Ferner ist es notwendig, die Art und Weise, wie in Texten auf Begriffe referiert wird, zu bestimmen. Dieser Bestimmung widmet sich ebenfalls der Abschnitt 4. Auf Basis dieser Begriffsoperationalisierung wird ein computerlinguistisches Modell zur Darstellung der Begriffsrelationen in den Texten von Adorno und

2 Um etwaigen Missverständnissen vorzubeugen: Was der vorliegende Aufsatz nicht liefern wird, ist eine philosophisch-epistemologische Diskussion des Gehalts und der Valenz von Adornos Verfahrensweise der Konstellation. Ebenso wird im Folgenden auf eine detaillierte Rekonstruktion von Adornos Selbstaussagen bezüglich seiner Verfahrensweise verzichtet, sondern in Betreff derselben an die Adornoforschung angeknüpft.
Zum Begrifflichen: ‚Konstellation' ist der von Adorno selbst eingeführte Ausdruck zur Bezeichnung seines Ideals einer philosophischen Darstellungsform bzw. Verfahrensweise – die beiden letztgenannten Ausdrücke werden im vorliegenden Aufsatz synonym verwendet. Innerhalb dieses Ideals kommt der Verknüpfung von Begriffen eine zentrale Rolle zu.

3 Einen guten Einstieg in Adornos Verständnis dieser Verfahrensweise bieten die 19. und 20. Vorlesung in Adorno 2010 und Adorno 2003b.

Carnap entwickelt: Die Begriffswörter stehen auf der Textoberfläche in einem komplexen Verhältnis linguistischer Dependenzen, die automatisch annotiert und auf Basis dieser Annotationen regelbasiert in Netzwerke überführt werden, was in Abschnitt 5 paradigmatisch anhand von zwei Absätzen aus Texten von Adorno und Carnap vorgeführt wird. Die solcherart gewonnenen Daten bilden die Grundlage eines netzwerkanalytischen Vergleichs, in welchem ein Verfahren zum Clustering von Netzwerken vorgestellt wird (Abschnitt 6). Im Mittelpunkt dieses Clustering-Verfahrens steht der Vergleich von ego-zentrierten Netzwerken mit den in Abschnitt 3 entwickelten Hypothesen zu den zu erwartenden ‚Begriffsnetzwerken' in den Texten von Adorno und Carnap (Abschnitt 7). Eine Auswertung der solcherart gewonnenen Befunde beschließt den Beitrag (Abschnitt 8).

2 Voraussetzungen: Adornos Begriffsverständnis und seine Konsequenzen für die philosophische Darstellung

Adornos Begriffsverständnis ist schwer zu fassen, da es aus Adornos produktivkritischer Auseinandersetzung mit der klassischen deutschen Philosophie – insbesondere Kant und Hegel (Klein et al. 2019, S. 377–391) – hervorging, Adorno jedoch auf eine explizite Bestimmung dessen, was unter einem Begriff zu verstehen ist, verzichtet hat. Obwohl Adorno immer wieder vom Begriff als der „Merkmalseinheit des darunter Befaßten" (Adorno und Horkheimer 2003, S. 32) schreibt, ist es an zahlreichen Stellen seiner Schriften nicht klar, ob er mit dem Ausdruck ‚Begriff' tatsächlich nur auf generelle Termini (z. B. ‚Kunstwerk') Bezug nimmt oder seine Verwendung auch singuläre Termini (z. B. Eigennamen) miteinschließt.[4] In Folge dieser Praxis ist die Vermutung geäußert worden, dass in Adornos Texten „Worte der Umgangssprache [...] gleichrangig behandelt [werden] wie Begriffe" (Ritter 2008).

Trotz dieser Unschärfen in Hinblick auf die Bestimmung dessen, was ein Begriff ist, hat Adorno sehr konkrete Vorstellungen davon, wie Begriffe philosophisch zu verwenden sind. Exemplarisch sei hier ein Zitat aus einer seiner Vorlesungen angeführt:

4 Zu dieser Unterscheidung und ihrer Bedeutung in der Gegenwartsphilosophie siehe auch den Philosophieabschnitt in Ketschik et al. 2020, S. 227 ff. dieses Bandes.

> Die Kunst nun oder die Aufgabe, vor welche der Gebrauch der Begriffe die dialektische Methode stellt, ist nun, d[as] in jedem Begriff Enthaltene zu bewahren, es also nicht abzuschneiden, nicht durch willkürliche Setzungen oder Festlegungen zu verdecken, aber es gleichzeitig so zum Bewußtsein zu erheben, daß es doch eben aus der Sphäre der Zweideutigkeit oder der schlechten Vagheit eigentlich heraustritt. Das geschieht nun aber nicht durch die Definition, sondern statt dessen durch die Konstellation, in welche die Begriffe treten. (Adorno 2010, S. 281)

Die hier dargestellte Form des Begriffsgebrauchs steht – wie Adorno selbst seit seiner Antrittsvorlesung betonte – in Opposition zur Verfahrensweise einer philosophischen Richtung, die Adorno zeit seines Lebens als positivistisch bezeichnet hat. Zu dieser philosophischen Richtung zählten für Adorno neben den Mitgliedern des Wiener Kreises auch Karl Popper und Hans Albert.

Die positivistische Philosophie ist nach Adorno insbesondere durch zwei Merkmale gekennzeichnet: Einerseits propagiere sie einen Primat der formalen Logik, das andererseits durch einen sensualistischen Empirismus ergänzt werde, was in einen Widerspruch führe, der das ganze Projekt der „Logistik" – so Adornos Bezeichnung für die formale Logik – infrage stelle.[5]

Dem vermeintlichen Zwangscharakter der positivistischen Logik entkommt nach Adorno nur ein dialektisches Denken, da es den Gegenstand nicht bereits vorweg begrifflich identifiziere:

> In gewissem Betracht ist die dialektische Logik positivistischer als der Positivismus, der sie ächtet: sie respektiert, als Denken, das zu Denkende, den Gegenstand auch dort, wo er den Denkregeln nicht willfahrt. Seine Analyse tangiert die Denkregeln. Denken braucht nicht an seiner eigenen Gesetzlichkeit sich genug sein zu lassen; es vermag gegen sich selbst zu denken, ohne sich preiszugeben; wäre eine Definition von Dialektik möglich, so wäre das als eine solche vorzuschlagen. (Adorno 2003d, S. 144)

Ein solches Denken realisiert sich nach Adorno in Konstellationen.

5 Das hat Adorno bereits im Rahmen seiner Korrespondenz mit Horkheimer über dessen 1937 publizierten Aufsatz „Der neueste Angriff auf die Metaphysik" betont: „Die prinzipielle Unmöglichkeit, ihre beiden Grundoperationen, Experiment und Kalkül, in Übereinstimmung zu bringen, ist die Ausgangsantinomie der Logistik d. h. der Beweis, daß es ihr nicht gelingt, eben jene einheitliche Interpretation zu geben, die sie beansprucht; weil nämlich die Wirklichkeit ihr widerspricht, und weil sie selber brüchig ist." (Brief an Horkheimer vom 28.11.1936; zit. nach Dahms 1994, S. 88)

3 Hypothesenbildung: Wie realisiert sich eine konstellative Darstellung im Text?

Im Folgenden sollen Adornos leitende Vorstellung des philosophischen Begriffsgebrauches und der damit einhergehenden Darstellungsform sowie ihr Verhältnis zu denjenigen eines der zentralen Repräsentanten des Wiener Kreises, Rudolf Carnap, kurz skizziert und im Anschluss daran Arbeitshypothesen aufgestellt werden, wie sich ein solcher Begriffsgebrauch auf der Textoberfläche realisiert und von demjenigen Carnaps unterscheidet. Die Bildung derartiger Hypothesen ist notwendig, da, wie bereits das obige Zitat belegt, Adorno bei der Beschreibung seines Verfahrens, Begriffe zueinander ins Verhältnis zu setzen, nicht nur durchgehend auf konkrete Beispiele verzichtet, sondern bei ihrer Kennzeichnung regelmäßig auf Metaphern zurückgreift.

Im Zentrum steht dabei der Ausdruck ‚Konstellation'. Wie die jüngere Forschung gezeigt hat, dient das Konzept der ‚Konstellation' in den Sozial- und Geisteswissenschaften des 20. Jahrhunderts „zur Bezeichnung einer mehrstelligen Beziehungsstruktur, das heißt eines Ensembles differenter [...] Positionen und Faktoren, die [...] einen dynamischen, veränderbaren Wirkungszusammenhang bilden und auch nur aus diesem relationalen Zusammenhang heraus angemessen erklärt oder verstanden werden können" (Albrecht 2010, S. 107).

Die solcherart gegebene relationale Struktur von Begriffen in philosophischen Texten rechtfertigen es, die in diesem Aufsatz untersuchten Begriffsrelationen mithilfe netzwerkanalytischer Verfahren in den Blick zu nehmen. Dafür ist es notwendig, einen Weg zu finden, Adornos und Carnaps Selbstreflexionen ihres Begriffsgebrauches in Modelle zu überführen, die sich mithilfe netzwerkanalytischer Metriken beschreiben lassen.

Wagt man sich an die Zusammenschau der in Adornos Schriften seit den dreißiger Jahren zu findenden Passagen zu seinem Begriffsgebrauch bzw. zu seinem Ideal der philosophischen Darstellung, stößt man insbesondere auf folgende Kennzeichnungen derselben:

1. An die Stelle der Definition tritt die Konstellation: Die Bedeutung eines Begriffes sowie sein Verhältnis zu anderen Begriffen wird nicht qua Nominaldefinition, sondern implizit, d. h. durch seine Verknüpfung mit anderen Begriffen im konkreten sprachlichen Gebrauch/Textvollzug bestimmt.[6]

6 Siehe diesbezüglich zum Beispiel Adorno (2003d, S. 164 f.): „Das einigende Moment überlebt, ohne Negation der Negation, doch auch ohne der Abstraktion als oberstem Prinzip sich zu über-

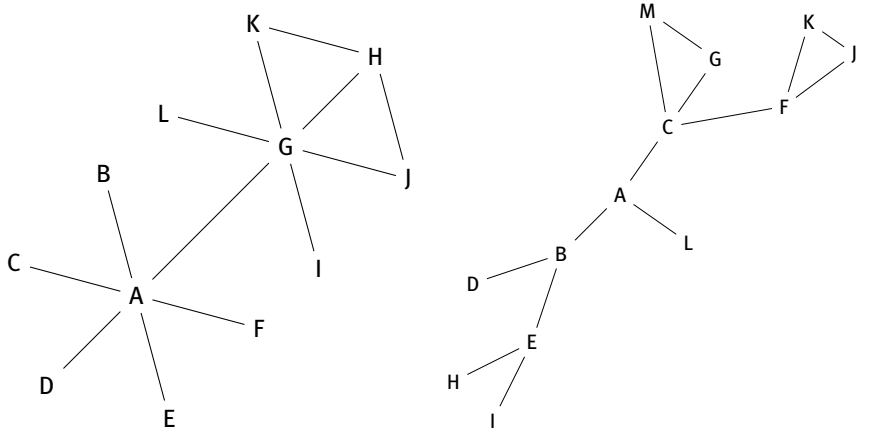
2. An die Stelle der Hierarchisierung von Aussagen (und Begriffen) tritt die Parataxe: Darunter versteht Adorno „kunstvolle Störungen […], welche der logischen Hierarchie subordinierender Syntax ausweichen“ (Adorno 2003e, S. 471), was soviel heißt wie, dass aufeinanderfolgende Sätze nicht in einem Ableitungsverhältnis stehen.

Linguistisch sollte sich eine solche Form der Begriffsverknüpfung auf der Textoberfläche derartig realisieren, dass in Adornos Texten einige wenige Begriffe in sehr vielen Sätzen auftauchen, während die restlichen Begriffe das nur sehr selten tun, da sie primär dazu dienen, die Hauptbegriffe näher zu bestimmen.

Die Programmschrift des Wiener Kreises propagiert ein Adornos Vorstellungen diametral entgegengesetztes Vorgehen. Angestrebt wird unter anderem „ein[] Gesamtsystem der Begriffe“ (Neurath et al. [1929] 2006, S. 16). Innerhalb desselben soll „der Sinn eines jeden Begriffs, zu welchem Wissenschaftszweige er immer gehören mag, […] durch eine schrittweise Rückführung auf andere Begriffe, bis hinab zu den Begriffen niederster Stufe, die sich auf das Gegebene selbst beziehen“ (Neurath et al. [1929] 2006, S. 15), angegeben werden. Diesem Selbstverständnis entsprechend ist zu erwarten, dass in den Schriften der Autoren des Wiener Kreises

1. die Begriffe regelmäßig definiert und
2. die Begriffe zueinander in ein hierarchisches Verhältnis gesetzt werden.

antworten, dadurch, daß nicht von den Begriffen im Stufengang zum allgemeineren Oberbegriff fortgeschritten wird, sondern sie in Konstellation treten. Diese belichtet das Spezifische des Gegenstands, das dem klassifikatorischen Verfahren gleichgültig ist oder zur Last. Modell dafür ist das Verhalten der Sprache. Sie bietet kein bloßes Zeichensystem für Erkenntnisfunktionen. Wo sie wesentlich als Sprache auftritt, Darstellung wird, definiert sie nicht ihre Begriffe. Ihre Objektivität verschafft sie ihnen durch das Verhältnis, in das sie die Begriffe, zentriert um eine Sache, setzt. Damit dient sie der Intention des Begriffs, das Gemeinte ganz auszudrücken. Konstellationen allein repräsentieren, von außen, was der Begriff im Innern weggeschnitten hat, das Mehr, das er sein will so sehr, wie er es nicht sein kann. Indem die Begriffe um die zu erkennende Sache sich versammeln, bestimmen sie potenziell deren Inneres, erreichen denkend, was Denken notwendig aus sich ausmerzte.“



(a) Modell der Begriffsverknüpfung bei Adorno **(b)** Modell der Begriffsverknüpfung bei Carnap

Abb. 1: Modellnetzwerke für Adorno und Carnap

4 Operationalisierung: Vom Text über den Begriff zum ‚Begriffsnetzwerk'

In Anbetracht besagter Leitvorstellungen lassen sich – stark vereinfachend – folgende Hypothesen über die Begriffsverknüpfung bei den genannten Autoren aufstellen: Es ist zu erwarten, dass in Adornos Texten die zentralen Begriffe direkt mit einer großen Anzahl an anderen Begriffen verknüpft werden, letztere jedoch zueinander nicht zwingend in einem Verhältnis stehen. Ein Modell dieses Verhältnisses könnte folgendermaßen aussehen: A – B; A – C; A – D; A – F; A – G; G – L; G – K; G – H; G – J; G – I; K – H; H – J u. s. f. (Abb. 1a)

In Hinblick auf die Texte des Wiener Kreises ist dagegen zu erwarten, dass die Oberbegriffe indirekt mit der Mehrheit der ihnen untergeordneten Begriffe, direkt jedoch nur mit den ihnen unmittelbar untergeordneten Begriffen verknüpft sind. Daraus ergibt sich folgendes Modell: L – A; A – B; A – C; B – D; B – E; E– H; E– I u. s. f. (Abb. 1b)

Carnaps Selbstbeschreibungen seiner Darstellungsweise entsprechend ließe sich der Modellgraph seines Begriffsnetzwerkes in eine Begriffspyramide überführen. Der Modellgraph für Adornos Begriffsnetzwerk kennzeichnet sich hingegen dadurch, dass zwei vermeintlich zentrale Begriffe – im Modell die Begriffe A und G – durch eine Vielzahl anderer Begriffe näher bestimmt werden, diese Begriffe jedoch nicht vollständig weiter bestimmt werden.

Um die soeben entwickelten Modelle der zu erwartenden Begriffsrelationen bei Adorno und Carnap an deren Texten überprüfen zu können, ist es notwendig festzulegen, was im Folgenden unter ‚Begriff' verstanden wird und wie sich ein solcher auf der Textoberfläche realisiert bzw. wie von dieser auf ihn referiert wird. Was ein ‚Begriff' ist, ist in der Philosophie seit der Antike umstritten. Die diesbezüglichen erkenntnistheoretischen Positionen reichen vom Begriffsrealismus, der davon ausgeht, dass Begriffe geist- und sprachunabhängige Entitäten sind, über den Konzeptualismus, nach dem es Begriffe in der Wirklichkeit nicht gibt, sondern nur im Geist, zum Nominalismus, nach dem Begriffe abstrakte Objekte seien, die nur als Zeichen existieren.

In Anbetracht der Vielfalt an philosophischen Begriffskonzeptionen sowie der Konsequenzen, die die Bevorzugung einer dieser Konzeptionen für die auf Basis derselben gewonnenen Befunde mit sich bringt,[7] folgt der vorliegende Aufsatz einem Verständnis von Begriff, der eine lange Tradition hat, auch in der Philosophie kaum umstritten und zudem in der Linguistik anerkannt ist, wie unter anderem das *Metzler Lexikon Sprache* belegt. Nach diesem handelt es sich bei einem ‚Begriff' um das „Aggregat kategorialer oder relationaler Merkmale, das die Gegenstände, Zustände, Prozesse etc., denen die Merkmale zukommen, zu einer Klasse zusammenfasst und das mit einem kommunizierbaren, i. d. R. verbalen Ausdruck verknüpft ist" (Clément und Rödel 2016, S. 93). Der verbale Ausdruck, der Begriffe bezeichnet, wird auch ‚Begriffswort' genannt. Während seit dem 20. Jahrhundert die formale Logik Begriffe als Prädikate bzw. prädikative Funktoren versteht, für die innerhalb der formalen Logik eine eigene Notationsform existiert, dominieren im Alltagsgebrauch Nomen als Begriffsworte.[8]

In Übereinstimmung mit dieser Praxis werden wir im Folgenden sämtliche in den untersuchten Korpora vorhandenen Nomen als Begriffsworte verstehen. Diese Gleichsetzung mag für so manchen auf den ersten Blick kontraintuitiv erscheinen. Für sie sprechen neben den genannten noch folgende drei Gründe: Erstens

7 Zum Zusammenhang von Operationalisierung, impliziten Vorannahmen und Deutung von Befunden im Rahmen einer reflektierten algorithmischen Textanalyse siehe Pichler und Reiter 2020 ab Seite 43 in diesem Band.

8 Dies reicht soweit, dass Willard von Orman Quine im Zuge der Erläuterung des Unterschiedes von singulären und generellen Termini sowie von konkreten und abstrakten Termini im Rahmen seiner *Grundzüge der Logik* fast ausschließlich auf Nomen als Beispiele zurückgreift: „Konkrete [Termini] sollen auf Individuen, physikalische Objekte, Ereignisse verweisen, abstrakte Termini auf abstrakte Objekte, z. B. auf Zahlen, Klassen, Attribute. So sind manche singulären Termini konkret, z. B. ‚Sokrates', ‚Zerberus', ‚Erde', ‚der Autor von Waverly', andere sind dagegen abstrakt, z. B. ‚7', ‚3+4', ‚Frömmigkeit'. Ebenso sind manche allgemeinen Termini konkret, wie ‚Mensch', ‚Haus', ‚rotes Haus' [...]; andere dagegen wie ‚Primzahl', ‚zoologische Gattung', ‚Tugend', sind abstrakt" (Quine 1974, S. 262).

Tab. 1: Hochfrequente Wörter aus dem Adorno-Korpus. Gezählt wurden Lemmata, Zahlen zeigen den Anteil der verschiedenen Formen an der Gesamtzahl der Wörter in Prozent.

Werk	Häufigste Wörter
Der Essay als Form	Essay (1,05), Begriff (0,53), Gegenstand (0,42), Form (0,37), Sache (0,32), Wissenschaft (0,29), Gedanke (0,28), Wahrheit (0,27), Sein (0,23), Geist (0,22)
Die Kunst und die Künste	Kunst (1,58), Kunstwerk (0,39), Musik (0,37), Sinn (0,27), Moment (0,24), Geist (0,24), Gattung (0,24), Begriff (0,22), Kunstgattung (0,18), Zusammenhang (0,16)
ND Einleitung	Philosophie (0,64), Begriff (0,54), Denken (0,39), Gedanke (0,31), Erfahrung (0,27), Subjekt (0,27), Dialekt (0,27), Erkenntnis (0,24), Bewusstsein (0,24), System (0,22)
Stimmigkeit und Sinn	Kunst (0,98), Kunstwerk (0,82), Form (0,74), Sinn (0,44), Werk (0,31), Moment (0,25), Einheit (0,24), Inhalt (0,21), Gebilde (0,20), Intention (0,19)
Zur Theorie des Kunstwerks	Kunstwerk (1,20), Kunst (0,93), Werk (0,65), Einheit (0,35), Begriff (0,34), Erhabene (0,32), Moment (0,30), Geist (0,30), Bewusstsein (0,22), Geschichte (0,19)

verwendet Adorno selbst mehrheitlich Nomina als Beispiele für Begriffe. Zweitens handelt es sich bei den von Adorno und Carnap verwendeten Nomina fast ausschließlich um Abstrakta – in den seltenen Fällen, in denen sie Konkreta verwenden, handelt es sich zudem meist um Appelativa –, wie ein Blick in Wortfrequenzlisten zeigt (Tabelle 1). Drittens beschränkt sich Adorno in seiner Beschreibung der Konstellation von Begriffen nicht auf prädikative Aussagesätze, sondern ‚spricht' ganz allgemein vom Verhältnis zwischen den Begriffen. Die Analyse hat sich daher nicht auf diejenigen Nomina zu beschränken, bei denen es sich sowohl grammatisch als auch logisch-semantisch um Subjekte und Prädikate handelt, sondern ist auf sämtliche in den Texten verwendeten Nomina auszuweiten.[9] Sämtliche Nomina in den Texten von Adorno und Carnap werden also im Folgenden als Begriffsworte, d. h. als Referenzen auf Begriffe im oben definierten Sinne verstanden.

Die technische Modellierung der oben entwickelten Begriffsrelations-Modelle folgt den Grundsätzen der computerwissenschaftlichen Netzwerkanalyse. Dabei besteht ein Graph G aus einer Menge Knoten (*vertices*) V und einer Menge Kanten (*edges*) E. Kanten wiederum werden als Tupel aus zwei Knoten sowie eventueller

9 Neben den hier genannten Gründen haben auch CRETA-spezifische arbeitspraktische Gründe zu der hier dargelegten Operationalisierung von ‚Begriffen' geführt. Siehe dazu den Abschnitt zur Philosophie in Ketschik et al. 2020, S. 227 ff. dieses Bandes.

Werkzeug	Modus	Referenz	Ergebnis
OCR (Acrobat Reader Pro)	A		maschinenlesbarer Text
Nachkorrektur	M	–	maschinenlesbarer Text
Tokenizer[11]	A		Satz- und Tokengrenzen
TreeTagger2013[12]	A	Schmid (1994)	Wortarten und Lemmata
Stuttgart Dependency Parser[13]	A	Bohnet (2010)	syntaktische Dependenzstruktur
CorefAnnotator[14]	M	Reiter (2018)	Koreferenzen

Tab. 2: Vorverarbeitungsschritte für das Adorno/Carnap-Korpus. Modus beschreibt, ob der Schritt automatisch (A) oder manuell (M) durchgeführt wurde.

Gewichte bzw. Bezeichnungen modelliert. Für die vorliegende Untersuchung ist also zu klären, wie die auf den Texten basierenden Netzwerke gebildet werden, so dass die Modi der Begriffsverknüpfung von Adorno und Carnap untersucht werden können. Diese Transformation erfolgt in zwei Schritten: Einer linguistischen Anreicherung der Texte sowie der Netzwerkextraktion.

Die einzelnen Werkzeuge, die für die Digitalisierung und linguistische Anreicherung verwendet wurden, sind in Tabelle 2 aufgeführt. Zu beachten ist dabei, dass einige Schritte manuell, andere automatisch durchgeführt wurden. Ausgangspunkt der Arbeit waren gedruckt vorliegende Texte.[10] Diese wurden mittels *optical character recognition* (OCR) automatisch digitalisiert und anschließend manuell nachkorrigiert. Die Texte lagen somit schließlich als *plain-text*-Dateien vor, bevor sie linguistisch aufbereitet wurden.

Der digitalisierte Text wurde dann zunächst tokenisiert, also in Sätze und einzelne Wortformen (sog. *tokens*) zerlegt. Für jede Wortform wurden mithilfe des *tree tagger*s die Wortart sowie das Lemma erkannt. Anschließend wurden syntaktische Abhängigkeiten zwischen Wortformen (Dependenzen) ebenfalls automatisch erkannt.

(1) Die wissenschaftliche Weltauffassung kennt keine unlösbaren Rätsel.

Tabelle 3 zeigt als Beispiel die linguistischen Annotationen für den in (1) wiedergegebenen Satz. Abbildung 2 zeigt die gewonnenen Dependenzrelationen visuell.

10 Dabei handelte es sich um folgende Ausgaben: Adorno 2003b, Adorno 2003d, Adorno 2003a, Adorno 2003c sowie Carnap 1931b, Carnap 1931a, Carnap 1930, Carnap 1932 – 1933.

11 http://hdl.handle.net/11858/00-247C-0000-0007-3736-B

12 http://hdl.handle.net/11858/00-247C-0000-0022-D906-1

13 http://hdl.handle.net/11022/1007-0000-0000-8DEE-6

14 https://github.com/nilsreiter/CorefAnnotator/

Tab. 3: Tabellarische Darstellung einer computerlinguistischen Analyse des Satzes in (1).

N	Token	Wortart	Lemma	*Governor*	Text
1	Die	ART	der	3	Die
2	wissenschaftliche	ADJA	wissenschaftlich	3	wissenschaftliche
3	Weltauffassung	NN	Weltauffassung	4	Weltauffassung
4	kennt	VVFIN	kennen	–	kennt
5	keine	PIAT	kein	7	keine
6	unlösbaren	ADJA	unlösbar	7	unlösbaren
7	Rätsel	NN	Rätsel	4	Rätsel
8	.	$.	–	8	.

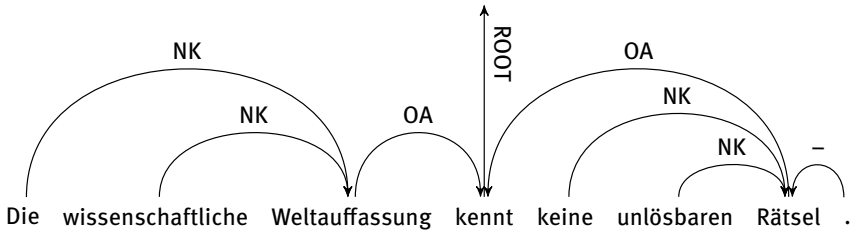


Abb. 2: Visuelle Darstellung der in Satz (1) enthaltenen Dependenzrelationen.

Beide Darstellungen können über die CLARIN-Webservices abgerufen werden, die auch die technische Grundlage für die Verarbeitungs-*pipeline* bildete.

Auch wenn eine vollautomatische Verarbeitung wünschenswert gewesen wäre, zeigten sich in der Praxis Probleme, die nur mittels manueller Eingriffe gelöst werden konnten. Eine manuelle strichprobenartige Inspektion ergab zunächst, dass die Erkennungsrate der automatischen Texterkennung nicht hoch genug war, was eine zeitintensive manuelle Nachkorrektur erforderlich machte. Basierend auf korrekt digitalisierten Texten konnten dann einige linguistische Werkzeuge vollautomatisch angewendet werden.

Die Auflösung von Koreferenzen (also die korrekte Zuordnung von Pronomen und anderen Verweisen auf Entitäten, siehe Ketschik et al. 2020, S. 205) ist zwar seit vielen Jahren Forschungsgegenstand der Computerlinguistik, allerdings mit Fokus auf englischsprachige Zeitungstexte und selbst da mit vergleichsweise mäßigen Resultaten.[15] Koreferenzen wurden daher vollständig manuell annotiert,

15 Joshi et al. (2019) berichten von 76,9 % durchschnittlichem F_1-score. Der *state of the art* für das Deutsche liegt nach Rösiger (2019, S. 101) bei etwa 48,61 % durchschnittlichem F_1-score.

wobei im gleichen Schritt auch fälschlicher Weise erkannte Nomina entfernt wurden.

Basierend auf der linguistisch angereicherten Textrepräsentation wurde in einem nächsten Schritt für jeden Text ein Netzwerk erstellt. Im Falle der vorliegenden Fragestellung entsprechen die Knoten den Begriffen in den untersuchten Texten, die Kanten den Relationen zwischen denselben. Dabei verstehen wir, wie oben erläutert, unter ‚Begriff' sämtliche in den Texten vorkommenden Nomina.

Für die Bestimmung des Verhältnisses zwischen den so bestimmten Begriffen greifen wir auf ein Modell zurück, das die Relation zwischen ihnen auf einer grammatisch-linguistischen Basis bestimmt. Als miteinander verbunden werden diejenigen Nomina erachtet, die entweder direkt vom Subjekt derselben Verbphrase oder direkt voneinander abhängen.

(2) Denn es ist bloßer Aberglaube der aufbereitenden Wissenschaft, die Begriffe wären an sich unbestimmt, würden bestimmt erst durch ihre Definition.

Aus Beispiel (1) würde sich damit eine Relation zwischen den Begriffen ‚Weltauffassung' und ‚Rätsel' ergeben, da beide syntaktisch von der Verbform ‚kennt' regiert werden. Aus Beispiel (2) würde (u. a.) eine Relation zwischen den Begriffen ‚Aberglaube' und ‚Wissenschaft' abgeleitet, da letzteres vom ersteren syntaktisch regiert wird.

Auf diese Weise wird zunächst eine Menge an Begriffspaaren je Satz erzeugt. Diese werden, für alle Sätze eines Textes, zu einem Gesamtnetzwerk aggregiert, mehrfach auftretende Kanten höher gewichtet und die verbundenen Knoten zusammengefasst. Durch die manuelle Annotation der Koreferenzen können die Knoten nicht nur an der Textoberfläche zusammengefasst werden. Stattdessen werden Knoten synonymer Begriffe ebenfalls zusammengefasst. Das Netzwerk beinhaltet also gewichtete, ungerichtete Kanten. Die Richtung der syntaktischen Abhängigkeit wird also im Netzwerk nicht (mehr) beachtet.

5 Illustration: Von der Modellierung über den Text zum Netzwerk

Zur Illustration des soeben vorgestellten Modellierungsansatzes soll dieser im Folgenden anhand von Ausschnitten aus zwei der untersuchten Texte ‚textnah' vorgestellt werden. Um dabei das Datenmodell zu veranschaulichen, werden die Begriffsverhältnisse der ausgewählten Passagen in Netzwerkgraphen visualisiert.

Dies erlaubt eine erste Annäherung an die Hypothesen, die in Abschnitt 3 entwickelt wurden.

Bei den ausgewählten Passagen handelt es sich jeweils um einen Abschnitt aus Adornos „Der Essay als Form" und Carnaps „Überwindung der Metaphysik durch logische Analyse der Sprache". Die beiden Passagen wurden nicht nur aufgrund ihres paradigmatischen Charakters für die Darstellungsform der beiden Autoren, sondern auch aus inhaltlichen Gründen ausgewählt: Sie behandeln Topoi, die selbst in unmittelbarem Zusammenhang mit der hier erhobenen Fragestellung – den Modi der Begriffsverknüpfung bei Adorno und Carnap – stehen. So sollen die folgenden Beispiele nicht nur zur Illustration der Modellierung und somit zur Vorbereitung der algorithmischen Analyse, sondern auch zur Vertiefung des Verständnisses der beiden Autoren beitragen.

Den Anfang macht dabei der Abschnitt aus Adornos „Der Essay als Form". Der Text gilt der Forschung als zentrale Metareflexion von Adornos Verfahrensweise (cf. Sonderegger 2019). Zudem wurde und wird die These vertreten, dass Adorno hier die im Text reflektierte Verfahrensweise bereits selbst realisiert. Bei dem ausgewählten Abschnitt handelt es sich um einen Passus am Ende des ersten Teils des Essays:

> Wie er [= der Essay; A.P.] Urgegebenheiten verweigert, so verweigert er die Definition seiner Begriffe. Deren volle Kritik ist von der Philosophie unter den divergentesten Aspekten erreicht worden; bei Kant, bei Hegel, bei Nietzsche. Aber die Wissenschaft hat solche Kritik niemals sich zugeeignet. Während die mit Kant anhebende Bewegung, als eine gegen die scholastischen Residuen im modernen Denken, anstelle der Verbaldefinitionen das Begreifen der Begriffe aus dem Prozeß rückt, in dem sie gezeitigt werden, verharren die Einzelwissenschaften, um der ungestörten Sicherheit ihres Operierens willen, bei der vorkritischen Verpflichtung zu definieren; darin stimmen die Neopositivisten, denen die wissenschaftliche Methode Philosophie heißt, mit der Scholastik überein. Der Essay dafür nimmt den antisystematischen Impuls ins eigene Verfahren auf und führt Begriffe umstandslos, „unmittelbar" so ein, wie er sie empfängt. Präzisiert werden sie erst durch ihr Verhältnis zueinander. Dabei jedoch hat er eine Stütze an den Begriffen selber. Denn es ist bloßer Aberglaube der aufbereitenden Wissenschaft, die Begriffe wären an sich unbestimmt, würden bestimmt erst durch ihre Definition. Der Vorstellung des Begriffs als einer tabula rasa bedarf die Wissenschaft, um ihren Herrschaftsanspruch zu festigen; als den der Macht, welche einzig den Tisch besetzt. In Wahrheit sind alle Begriffe implizit schon konkretisiert durch die Sprache, in der sie stehen. Mit solchen Bedeutungen hebt der Essay an und treibt sie, selbst wesentlich Sprache, weiter; er möchte dieser in ihrem Verhältnis zu den Begriffen helfen, sie reflektierend so nehmen, wie sie bewußtlos in der Sprache schon genannt sind. Das ahnt das Verfahren der Bedeutungsanalyse in der Phänomenologie, nur daß es die Beziehung der Begriffe auf die Sprache zum Fetisch macht. Dazu steht der Essay ebenso skeptisch wie zu ihrer Definition. Er zieht ohne Apologie den Einwand auf sich, man wisse nicht über allem Zweifel, was man unter den Begriffen sich vorzustellen habe. Denn er durchschaut, daß das Verlangen nach strikten Definitionen längst dazu herhält, durch festsetzende Manipulationen der Begriffsbedeutungen das Irritierende und Gefährliche der Sachen wegzuschaffen,

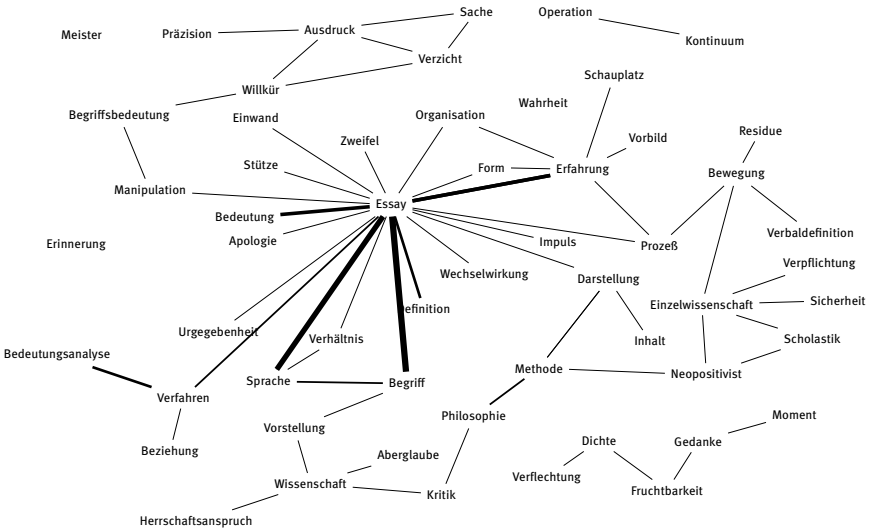


Abb. 3: Netzwerkgraph der ausgewählten Passage aus „Der Essay als Form“. Das Layout des Graphen ist auf Lesbarkeit optimiert, Distanzen also nicht direkt interpretierbar.

> die in den Begriffen leben. Dabei jedoch kommt er weder ohne allgemeine Begriffe aus – auch die Sprache, die den Begriff nicht fetischisiert, kann seiner nicht entraten – noch geht er mit ihnen nach Belieben um. Die Darstellung nimmt er darum schwerer als die Methode und Sache sondernden, der Darstellung ihres vergegenständlichten Inhalts gegenüber gleichgültigen Verfahrensweisen. Das Wie des Ausdrucks soll an Präzision erretten, was der Verzicht aufs Umreißen opfert, ohne doch die gemeinte Sache an die Willkür einmal dekretierter Begriffsbedeutungen zu verraten. Darin war Benjamin der unerreichte Meister. Solche Präzision kann jedoch nicht atomistisch bleiben. Weniger nicht, sondern mehr als das definitorische Verfahren urgiert der Essay die Wechselwirkung seiner Begriffe im Prozeß geistiger Erfahrung. In ihr bilden jene kein Kontinuum der Operationen, der Gedanke schreitet nicht einsinnig fort, sondern die Momente verflechten sich teppichhaft. Von der Dichte dieser Verflechtung hängt die Fruchtbarkeit von Gedanken ab. Eigentlich denkt der Denkende gar nicht, sondern macht sich zum Schauplatz geistiger Erfahrung, ohne sie aufzudröseln. Während aus ihr auch dem traditionellen Denken seine Impulse zuwachsen, eliminiert es seiner Form nach die Erinnerung daran. Der Essay aber wählt sie als Vorbild, ohne sie, als reflektierte Form, einfach nachzuahmen; er vermittelt sie durch seine eigene begriffliche Organisation; er verfährt, wenn man will, methodisch unmethodisch. (Adorno 2003b, S. 19 f.)

Die Passage widmet sich dem Begriffsgebrauch im Essay und den sich daraus ergebenden Unterschieden zur zeitgenössischen wissenschaftlichen Prosa der 50er-Jahre. Sie setzt ein mit der Feststellung, dass der Essay seine Begriffe nicht definiere und rechtfertigt diese Praxis der Begriffsverwendung durch den Verweis auf begriffskritische Positionen der Philosophiegeschichte. Im Anschluss

daran wird der Begriffsverwendung im Essay diejenige der zeitgenössischen Wissenschaft kontrastierend gegenübergestellt: Selbiger fehle das begriffskritische Moment, das den essayistischen Begriffsgebrauch auszeichne. Dieser wird in einem nächsten Schritt unter Rückbindung an Adornos Sprachphilosophie weiter ausdifferenziert und mündet in der aus dem zuvor Entwickelten folgenden Feststellung, dass „man [nicht über allem Zweifel] wisse […], was man unter den Begriffen sich vorzustellen habe". Abschließend werden die Konsequenzen dieser nicht-bestimmenden Bestimmung des Begriffs für die Darstellung im Essay ausgeführt.

Der Passus realisiert so die in ihm verhandelte Problematik: Er bestimmt den Essay näher, indem er ihn zu bestimmten Formen des Begriffsgebrauchs und deren Kritik ins Verhältnis setzt, ohne dabei die Begriffe, die zur Bestimmung des Essays herangezogen werden, selbst zu definieren. Kontexte – wie die Begriffskritik bei Kant, Hegel und Nietzsche – werden bloß angedeutet, nicht jedoch ausgeführt. Dasselbe gilt für jene Form der Wissenschaft, die auf Basis eines ebenfalls nur angedeuteten Sprachverständnisses, kritisiert wird.

Diese – in hohem Grad assoziative – Praxis der Bestimmung des Essays spiegelt auch das ‚Begriffsnetzwerk' des Passus wider, das in Abbildung 3 dargestellt ist. Im Zentrum des Netzwerks steht der Begriff ‚Essay', dem die Mehrheit der restlichen Begriffe direkt zugeordnet sind. Nur wenige dieser Begriffe werden selbst durch weitere Begriffe bestimmt. Es gibt nur sieben Äste in dem Graphen, die direkt an den ‚Begriff' anschließen und mehr als einen Knoten besitzen. Dort, wo dies jedoch der Fall ist, finden sich jene Differenzierungen, die den Text auszeichnen – so zum Beispiel die Unterscheidung der Methoden in Wissenschaft und Philosophie.

So ergeben sich sowohl strukturelle Ähnlichkeiten als auch Unterschiede zum im Abschnitt 3 entwickelten Modell des Begriffsgebrauchs beim konstellativen Schreiben. Dabei überwiegen jedoch die Übereinstimmungen: Sowohl im Modell-Graphen als auch in demjenigen, der aus der soeben behandelten Passage aus „Der Essay als Form" extrahiert wurde, stehen einer oder einige wenige ‚Haupt'-Begriffe, die durch eine Vielzahl anderer Begriffe näher bestimmt werden, die selbst nur sehr selten weiter ausdifferenziert werden – im Falle der Passage aus „Der Essay als Form" ist es jedoch nur ein Begriff, während es im Modell zwei Begriffe sind, die durch die restlichen Begriffe näher bestimmt werden. Ob sich das bei allen der im Folgenden untersuchten Texte so verhält, wird sich zeigen.

Als nächstes soll ein Blick auf einen Abschnitt aus Carnaps „Überwindung der Metaphysik durch logische Analyse der Sprache" geworfen werden. Dabei handelt es sich um die Eröffnung des zweiten Kapitels des Aufsatzes, in dem Carnap seine Bedeutungstheorie in Grundzügen entwickelt:

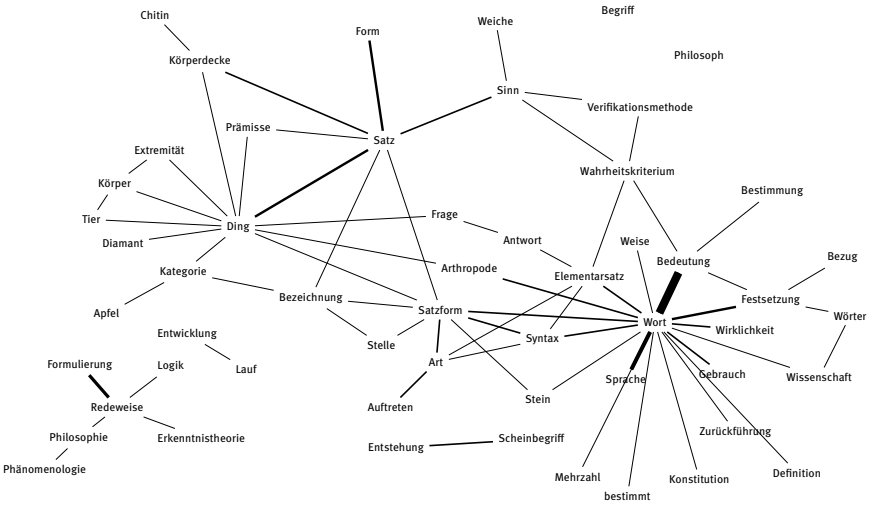


Abb. 4: Netzwerkgraph der ausgewählten Passage aus „Überwindung der Metaphysik durch logische Analyse der Sprache". Das Layout des Graphen ist auf Lesbarkeit optimiert, Distanzen also nicht direkt interpretierbar.

Hat ein Wort (innerhalb einer bestimmten Sprache) eine Bedeutung, so pflegt man auch zu sagen, es bezeichne einen „Begriff"; sieht es nur so aus, als habe das Wort eine Bedeutung, während es in Wirklichkeit keine hat, so sprechen wir von einem „Scheinbegriff". Wie ist die Entstehung eines solchen zu erklären? Ist nicht jedes Wort nur deshalb in die Sprache eingeführt worden, um etwas Bestimmtes auszudrücken, so daß es von seinem ersten Gebrauch an eine bestimmte Bedeutung hat? Wie kann es da in der traditionellen Sprache bedeutungslose Wörter geben? Ursprünglich hat allerdings jedes Wort (abgesehen von seltenen Ausnahmen, für die wir später ein Beispiel geben werden) eine Bedeutung. Im Lauf der geschichtlichen Entwicklung ändert ein Wort häufig seine Bedeutung. Und nun kommt es zuweilen auch vor, daß ein Wort seine alte Bedeutung verliert, ohne eine neue zu bekommen. Dadurch entsteht dann ein Scheinbegriff. Worin besteht nun die Bedeutung eines Wortes? Welche Festsetzungen müssen in Bezug auf ein Wort getroffen sein, damit es eine Bedeutung hat? (Ob diese Festsetzungen ausdrücklich ausgesprochen sind, wie bei einigen Wörtern und Symbolen der modernen Wissenschaft, oder stillschweigend vereinbart sind, wie es bei den meisten Wörtern der traditionellen Sprache zu sein pflegt, darauf kommt es für unsere Überlegungen nicht an.) Erstens muß die Syntax des Wortes festliegen, d. h. die Art seines Auftretens in der einfachsten Satzform, in der es vorkommen kann; wir nennen diese Satzform seinen Elementarsatz. Die elementare Satzform für das Wort „Stein" ist z. B. „x ist ein Stein"; in Sätzen dieser Form steht an Stelle von „x" irgendeine Bezeichnung aus der Kategorie der Dinge, z. B. „dieser Diamant:", „dieser Apfel". Zweitens muß für den Elementarsatz S des betreffenden Wortes die Antwort auf folgende Frage gegeben sein, die wir in verschiedener Weise formulieren können: 1. Aus was für Sätzen ist S ableitbar, und welche Sätze sind aus S ableitbar? 2. Unter welchen Bedingungen soll S wahr, unter welchen falsch sein? 3. Wie ist S zu verifizieren? 4. Weichen Sinn hat S? (1) ist die korrekte Formulierung;

> die Formulierung (2) paßt sich der Redeweise der Logik an, (3) der Redeweise der Erkenntnistheorie, (4) der der Philosophie (Phänomenologie). Daß das, was die Philosophen mit (4) meinen, durch (2) erfaßt wird, hat Wittgenstein ausgesprochen: der Sinn eines Satzes liegt in seinem Wahrheitskriterium. [(1) ist die „metalogische" Formulierung; eine ausführliche Darstellung der Metalogik als Theorie der Syntax und des Sinnes, d. h. der Ableitungsbeziehungen, soll später an anderer Stelle gegeben werden.] Bei vielen Wörtern, und zwar bei der überwiegenden Mehrzahl aller Wörter der Wissenschaft, ist es möglich, die Bedeutung durch Zurückführung auf andere Wörter („Konstitution", Definition) anzugeben. Z. B.: „‚Arthropoden' sind Tiere mit gegliedertem Körper, gegliederten Extremitäten und einer Körperdecke aus Chitin." Hierdurch ist für die elementare Satzform des Wortes „Arthropode", nämlich für die Satzform „das Ding x ist ein Arthropode", die vorhin genannte Frage beantwortet; es ist bestimmt, daß ein Satz dieser Form ableitbar sein soll aus Prämissen von der Form „x ist ein Tier", „x hat einen gegliederten Körper", „x hat gegliederte Extremitäten", „x hat eine Körperdecke aus Chitin", und daß umgekehrt jeder dieser Sätze aus jenem Satz ableitbar sein soll. Durch diese Bestimmungen über Ableitbarkeit (in anderer Ausdrucksweise: über das Wahrheitskriterium, die Verifikationsmethode, den Sinn) des Elementarsatzes über „Arthropode" ist die Bedeutung des Wortes „Arthropode" festgelegt. In dieser Weise wird jedes Wort der Sprache auf andere Wörter und schließlich auf die in den sog. „Beobachtungssätzen" oder „Protokollsätzen" vorkommenden Wörter zurückgeführt. Durch diese Zurückführung erhält das Wort seine Bedeutung. (Carnap 1931b, S. 221 f.)

Auffällig an dem Abschnitt ist, dass er seine Fragestellung – entgegen den Erwartungen in Bezug auf Carnaps Texte – nicht rein argumentativ entwickelt, sondern auch auf alternative Verfahrensweisen zurückgreift. So folgen auf die Unterscheidung zwischen ‚Begriff' und ‚Scheinbegriff' eine Reihe von Fragen, deren Beantwortung zu einer weiteren Ausdifferenzierung des zuvor eingeführten Begriffspaares führt, die dann im weiteren Verlauf des Absatzes nicht mehr explizit aufgegriffen wird. Stattdessen widmet sich der Text der Ausdifferenzierung des Bedeutungsbegriffes, spricht dabei aber nicht mehr, wie aufgrund der eingangs erfolgten Ausdifferenzierung zu erwarten wäre, von ‚Begriff', sondern von der Bedeutung eines ‚Wortes'. Diese wird bestimmt über eine Reihe von Kriterien, die ein Wort zu erfüllen hat, um als bedeutungstragend erachtet zu werden. Auch bei der Ausbuchstabierung dieser Kriterien arbeitet Carnap mit Fragen.

Nachdem Carnap auf diese Weise die zentralen Bedeutungskriterien – Syntax des Wortes und Bestimmungen des Elementarsatzes – bestimmt hat, widmet er sich einem alternativen Modus der Bedeutungsbestimmung: der ‚Konstitution' bzw. Definition. Bei deren Explikation kommen zahlreiche Beispiele zum Einsatz.

In Summa wird in dem Text ein Vorgehen realisiert, das zwar zu einer präzisen Bestimmung des Verhältnisses der unterschiedlichen Begriffe zueinander beiträgt, dabei aber nicht derartig strikt hierarchisch vorgeht, wie in Anbetracht von Carnaps Äußerungen zum begrifflichen Vorgehen der Einheitswissenschaft zu erwarten wäre. Letztlich beschränkt sich der Abschnitt auf die Unterscheidung zwischen Begriffen und Scheinbegriffen und erläutert, welchen Kriterien die zen-

trale Bedingung für Begrifflichkeit – die Tatsache, dass ein Wort Bedeutung besitzt – entsprechen muss, um als erfüllt erachtet zu werden.

Das bestätigt auch das aus dem Absatz abstrahierte Netzwerk, das in Abbildung 4 dargestellt ist. In diesem findet sich keine Begriffspyramide, sondern ein Netzwerk mit zwei zentralen Knoten, die lose miteinander verbunden sind. Dabei kommt einem Knoten, demjenigen von ‚Wort', eine weitaus größere Bedeutung zu – er hat fast in allen netzwerkanalytischen Kategorien die höchsten Werte. Der zweite zentrale Knoten, ‚Ding', veranschaulicht die Beispiellastigkeit der Explikation des dritten Bedeutungskriteriums, das im Text über Sätze der Form ‚Das Ding x ist y' exemplifiziert wird.

Der zentrale Unterschied zwischen diesem Begriffsnetzwerk und demjenigen des zuvor behandelten Absatzes Adornos besteht darin, dass bei Carnap sowohl die unterschiedlichen ‚Haupt'- bzw. Oberbegriffe weitaus inniger miteinander verwoben sind, als das bei Adorno der Fall war, als auch darin, dass die zur Bestimmung der zentralen Begriffe herangezogenen weiteren Begriffe häufig über zusätzliche Begriffe weiter ausdifferenziert werden. So entsteht ein Graph, der zwar mit dem Modell hierarchischer Begriffsverknüpfung nicht vollständig übereinstimmt, dabei jedoch die von Carnap selbst gesetzte Vorgabe erfüllt, die Begriffe in ein klares Verhältnis zueinander zu setzen.

6 Vergleich von Netzwerken

Die im Abschnitt 3 entwickelten und im vorausgehenden Abschnitt präzisierten Netzwerkmodelle basieren auf der Annahme, dass sich die für die Autoren typische Verwendung von Begriffen in charakteristischen Netzwerken manifestiert. Zur Beschreibung dieser Charakteristika wird jedes Netzwerk auf ausgewählte Merkmale abgebildet, die schließlich den Vergleich der Netzwerke erlauben. Dieses Vorgehen orientiert sich an erprobten Verfahren der Netzwerkanalyse.[16]

16 Prinzipiell existieren drei Herangehensweisen für den Vergleich von Netzwerken. Diese unterscheiden sich dadurch, dass sie auf unterschiedlichen Ebenen agieren: Der Vergleich auf der *Makroebene* erfolgt hinsichtlich globaler Eigenschaften. Üblicherweise werden dazu etwa durchschnittliche Knotengrade, Dichte, Durchmesser oder Transitivität der Netzwerke bestimmt. Über den Vergleich dieser globalen Kennzahlen werden strukturelle Unterschiede ersichtlich.
Verfahren der *Mezzoebene* lassen sich auf Basis ihrer Vorgehensweise in folgende drei Kategorien unterteilen: a.) der Vergleich über strukturelle Gruppen (sog. *communities*) und die Zuordnung der Knoten zu diesen Gruppen. Ein Beispiel für diese Vorgehensweise ist die wegweisende Arbeit von Rosvall und Bergstrom (2008). In diesem Paper wird zunächst ein Algorithmus zur Erkennung von Gruppierungen angewandt und die Zugehörigkeiten der Knoten zu den erkannten

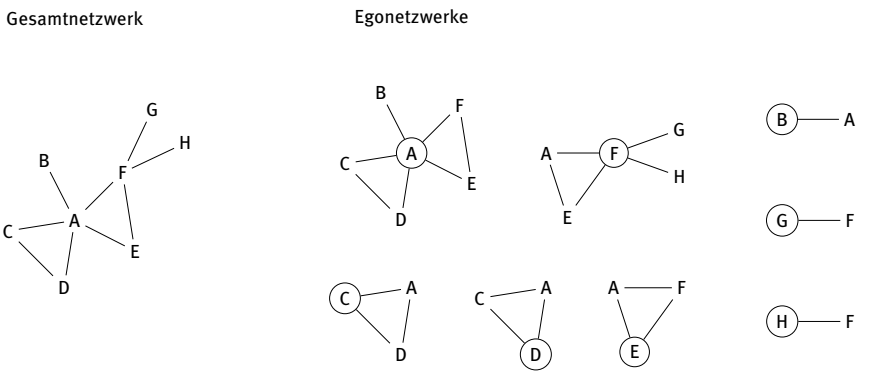


Abb. 5: Beispielnetzwerk (links) und sich daraus ergebende Egonetzwerke (rechts). Der Fokalknoten ist jeweils eingekreist.

Eine Herausforderung beim Vergleich von Netzwerken sind deren potenziell unterschiedlichen Größen, da solche Größenunterschiede bei vielen Metriken zu unerwünschten Verzerrungen führen. Um derartige Verzerrungen zu vermeiden, verwenden wir im Folgenden ein Verfahren, das globale Netzwerkmaße mit denjengen von sogenannten Egonetzwerken kombiniert. Zur Bildung von Egonetzwerken wird aus einem gegebenen Netzwerk eine (mglw. große) Menge an Teilnetzwerken gebildet, die einen (fokalen) Knoten sowie seine direkten Nach-

Gruppierungen ermittelt. Darauf aufbauend werden unterschiedliche Aggregationen der Zugehörigkeiten der Knoten und ihrer unmittelbaren Nachbarn evaluiert, die die zu vergleichenden Netzwerke möglichst eindeutig charakterisieren sollen.

b.) Verfahren, die ein Netzwerk in ein anderes zu überführen versuchen und die Anzahl der Modifikationsoperationen vergleichen (z. B. auf Basis der sogenannten *graph edit distance*); GED ist eine Metrik, die die Unterschiedlichkeit von Netzwerken in der Anzahl der Operationen quantifiziert, die einen Graphen in einen anderen überführen würden (cf. Sanfeliu und Fu 1983; Koutra et al. o.D.). Zulässige Operationen sind das Einfügen und Löschen von Knoten und Kanten sowie die Veränderung der Start- und Endpunkte von Kanten. Ein großer Nachteil dieser Verfahren liegt darin, dass eine grundsätzliche Übereinstimmung zwischen den Knoten der beiden Netzwerke gefunden werden muss – ein NP-hartes Problem, das diese Verfahren in der Regel äußerst rechenintensiv macht (Garey und Johnson 2002).

c.) Kernel-basierte Methoden (z. B. Borgwardt und Kriegel 2005; Shervashidze et al. 2011; Kondor und Pan 2016; Nikolentzos et al. 2017). Diese Klasse von Verfahren transformiert Netzwerke implizit in einen Merkmalsraum, um sie dann in diesem Merkmalsraum zu vergleichen. Hier sei beispielhaft das Verfahren von Borgwardt und Kriegel genannt, wonach die Anzahl der übereinstimmenden kürzesten Pfade verwendet wird, um Netzwerke miteinander zu vergleichen. Verfahren der *Mikroebene* beschreiben Netzwerke schließlich hinsichtlich lokaler Eigenschaften, die aus der unmittelbaren Umgebung einzelner Knoten abgeleitet werden (s. o.).

barn enthalten (und alle Verbindungen unter diesen). Abbildung 5 zeigt für ein fiktives Netzwerk die sich daraus ergebenden Egnonetzwerke. Wir knüpfen damit an eine Arbeit von Berlingerio et al. (2013) an, in der ein statistisches Framework vorgestellt wird, das die Merkmale eines Netzwerks aus dem Egonetzwerk jedes Knotens ermittelt. Zu den Merkmalen zählen etwa die Anzahl der Kanten der Egonetzwerke sowie der durchschnittliche Knotengrad in den Egonetzwerken. Der Vergleich zweier Netzwerke gelingt, indem die Merkmale der Egonetzwerke zu Netzwerksignaturen aggregiert und diese Signaturen anschließend verglichen werden.[17] Jedes der Egonetzwerke beschreiben wir in unterschiedlichen Aspekten – analog zu dem Vorgehen von Berlingerio et al. – mit sechs Merkmalen aus dem Egonetzwerk jedes Knotens. Da diese Merkmale für jedes Egonetzwerk extrahiert werden, für ein Netzwerk also eine Vielzahl an Einzelwerten extrahiert wird, werden diese nach einer Normalisierung mit *L2*-Norm durch Mittelwert, Median, Standardabweichung, Schiefe und Kurtosis aggregiert.[18] Tabelle 4 zeigt eine Übersicht über die Merkmale, die wir im Folgenden noch einzeln diskutieren.

17 Ein ähnliches Verfahren stellen Bonner et al. (2016) vor. Sie extrahieren jedoch eine deutlich größere Anzahl an Merkmalen, die auch die globalen Eigenschaften der Netzwerke wie etwa hinsichtlich der Anzahl der Komponenten beschreiben. In einer Arbeit von Zhang et al. (2018) wird die relative Entropie hinsichtlich der Knotengrade der Egonetzwerke verglichen, um die strukturelle Äquivalenz einzelner Knoten zu ermitteln.

18 Schiefe (*skewness*) und Kurtosis (Wölbung) werden über eine Häufigkeitsverteilung bestimmt und beschreiben die Abweichungen von einer Normalverteilung. Die *Schiefe* beschreibt, wie sehr der Gipfel der Verteilung nach links (dann ist die Schiefe negativ) oder nach rechts (dann ist sie positiv) geneigt ist. Bezogen auf unsere Netzwerke gibt die Schiefe an, ob die Mehrheit der Knoten eher große oder kleine Werte aufweisen. Die *Kurtosis* hingegen beschreibt die Höhe des Gipfels: Bei einer negativen Kurtosis sind alle Werte gleichmäßiger verteilt als eine Normalverteilung, bei einer positiven sind einzelne Werte häufiger als in einer Normalverteilung.
Bei mehreren Merkmalen spielt die Tatsache eine Rolle, dass in Adornos Texten zwei Arten von Begriffen verwendet werden: Zentrale Begriffe, die insgesamt häufig vorkommen und im Zentrum des Textes stehen, und andere Begriffe, die verwendet werden um die ‚Haupt'-Begriffe zu definieren. Durch diese ungleichmäßige Verteilung ist in allen Merkmalen eine höhere Kurtosis in den Netzwerken zu erwarten, die aus Adornos Texten extrahiert wurden.

Tab. 4: Übersicht über die extrahierten Merkmale und ihre Wertebereiche

Nr	Merkmal	Bezug	Formel	Wertebereich
1	*Closeness*	Gesamt	$\frac{\lvert V\rvert-1}{\sum_{j\in V} d(i,j)}$	$[\epsilon; 1]$
2	Knotengrad	Gesamt	$\frac{1}{\max_{j\in V}\deg(j)} \sum_{j\in \mathrm{nb}(i)} \deg(j)$	$[1; \infty]$
3	Anzahl Kanten	Ego_1	$\lvert E_{\mathrm{ego}_1(i)}\rvert$	$[1; \infty]$
4	Lokale Transitivität	Ego_1	$\frac{\lvert E_{\mathrm{ego}_1(i)}\rvert}{\lvert \mathrm{nb}(i)\rvert\times\lvert(\mathrm{nb}(i)-1)\rvert}$	$[\epsilon; 1]$
5	Clustering-Koeffizient	Ego_1	$\frac{1}{\deg(i)} \sum_{j\in \mathrm{nb}(i)} m_4(j)$	$[\epsilon; 1]$
6	Nachbarsnachbarn	Ego_2	$\lvert \mathrm{ego}_2(i)\rvert - \deg(i)$	$[0; \infty]$

Wir verwenden dafür die folgende Notation:

$G = (V, E)$	Graph
V	Menge der Knoten, $\lvert V\rvert$ Anzahl der Knoten im Netzwerk
$E \subseteq V^2$	Menge der Kanten, wobei eine Kante als Paar aus Knoten definiert ist
$d(i, j) : V^2 \to \mathbb{N}$	Anzahl an Kanten zwischen *i* und *j* auf kürzestem Weg
$\mathrm{nb}(i) : V \to V^n$	Menge der direkten Nachbarn von *i*, $0 \le n \le \lvert V\rvert$
$\mathrm{ego}_k(i) : V \to V^n$	Egonetzwerk mit Knoten bis zu *k* Kanten entfernt, $1 \le n \le \lvert V\rvert$
$\deg(i) : V \to N$	Anzahl der Kanten an *i* (= Grad)

Für jedes Merkmal werden dessen Werte in den Modellnetzwerken, die Formel zur Berechnung, eine Intuition und eine Interpretationshypothese angegeben. Die Hypothesen finden sich auch im Anschluss gesammelt in Tabelle 5. Diese basieren – darauf sei hier noch einmal explizit hingewiesen – auf den in Abschnitt 3 entwickelten Modellnetzwerken und nicht auf den im Abschnitt 5 aus einzelnen Absätzen von Adorno und Carnap extrahierten Netzwerkgraphen, da Erstere auf den expliziten Idealvorstellungen der beiden Philosophen zur Begriffsverknüpfung basieren, während die Repäsentativität von Letzteren (noch) nicht eindeutig bestimmt werden kann. Die in Abschnitt 5 gewonnenen Einsichten werden jedoch bei der finalen Auswertung der Befunde in Abschnitt 8 berücksichtigt.

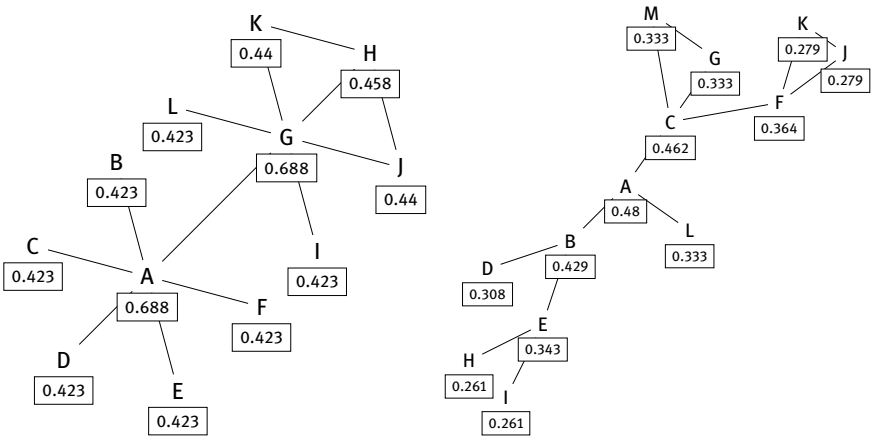


Abb. 6: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: *closeness*-Werte (Merkmal 1).

Merkmal 1: *Closeness*

Die *closeness* oder *closeness centrality* ist ein nähebasiertes Zentralitätsmaß, nach welchem diejenigen Knoten eines Netzwerkes zentral sind, welche von den restlichen Knoten desselben Netzwerkes auf möglichst kurzer Strecke erreichbar sind. Wir arbeiten im Folgenden mit der durchschnittlichen *closeness* zu allen anderen Knoten im gesamten Netzwerk, wobei mit $|V|$ die Anzahl der Knoten im Netzwerk bezeichnet wird und *d(i,j)* die kürzeste Pfaddistanz zwischen den Knoten *i* und *j* repräsentiert. Es handelt sich hier also um ein Merkmal, für dessen Berechnung keine Egonetzwerke verwendet werden.

$$m_1(i) = \frac{|V| - 1}{\sum_{j \in V} d(i,j)} \qquad i \neq j \tag{1}$$

Ausgehend von den im Abschnitt 3 entwickelten Netzwerkmodellen ist zu erwarten, dass die Pfaddistanzen bei Adorno kürzer ausfallen als bei Carnap, da davon auszugehen ist, dass in den aus Adornos Texten extrahierten Netzwerken sich die Mehrheit der Begriffe unmittelbar um einige wenige zentrale Begriffe anordnet – im Modellnetzwerk sind das die Begriffe *A* und *G*. Es ist daher mit einem höheren arithmetischen Mittelwert und Median bei Adorno zu rechnen.

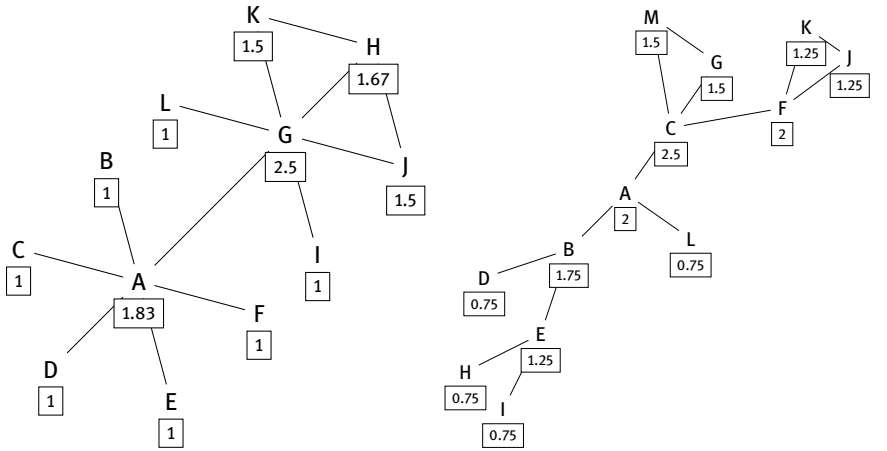


Abb. 7: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: Knotengrad-Werte (Merkmal 2).

Merkmal 2: Knotengrad der Nachbarn

Normalisierter summierter Knotengrad der Nachbarn von *i*. Die Normalisierung erfolgt hier über den maximalen Knotengrad im Gesamtnetzwerk, so dass die Werte netzwerkübergreifend vergleichbar sind. Hier werden die Nachbarn von Knoten *i* mit *nb(i)* und der Knotengrad eines Knotens *i* mit *deg(i)* bezeichnet.

$$m_2(i) = \frac{1}{\max_{j \in V} \deg(j)} \sum_{j \in \mathrm{nb}(i)} \deg(j) \tag{2}$$

Den Knotengrad der Nachbarn von *i* zu verwenden, basiert auf der Theorie des sogenannten sozialen Kapitals von Coleman (1966). In Hinblick auf dieses Merkmal ist zu vermuten, dass die Knoten in den ‚Begriffsnetzwerken' Carnaps tendenziell gleichmäßiger vernetzt sind, da in Carnaps Texten den übergeordneten Begriffen stets nur eine begrenzte Anzahl an abhängigen Begriffen untergeordnet ist. Das ‚soziale Kapital' der Begriffe sollte daher relativ gleichmäßig verteilt sein. Demgegenüber ist anzunehmen, dass die Verteilung des ‚Kapitals' bei Adorno unregelmäßiger sein wird, da es nur eine geringe Anzahl an Knoten geben sollte, die eine hohe Anzahl an Nachbarn haben. Die Kurtosis sollte damit bei Adorno- höher ausfallen als bei Carnap-Texten.

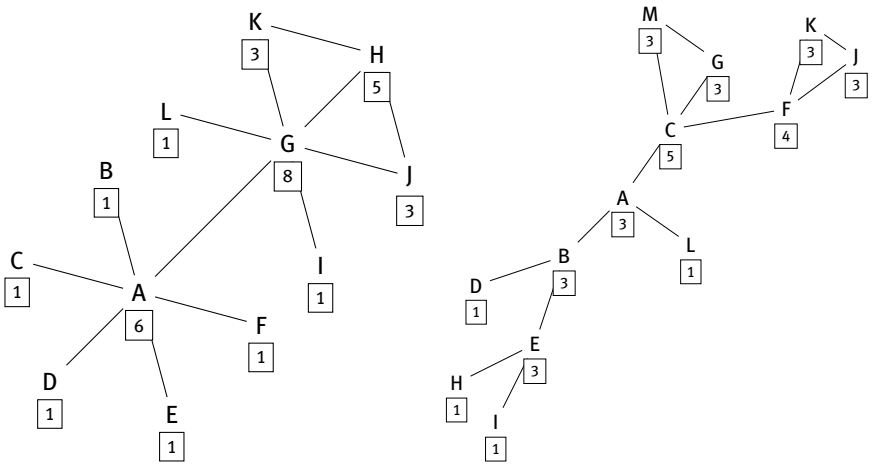


Abb. 8: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: Anzahl der Kanten im jeweiligen Egonetzwerk (Merkmal 3).

Merkmal 3: Anzahl Kanten

Anzahl der Kanten im Egonetzwerk von *i*; $ego_k(i)$ bezeichnet hier das Egonetzwerk von *i*, das Knoten bis zu einer Entfernung von *k* beinhaltet.

$$m_3(i) = |E_{\mathrm{ego}_1(i)}| \tag{3}$$

Die Anzahl der Kanten in den Egonetzwerken erfasst zusätzlich zur Anzahl der Nachbarknoten auch deren Vernetzung untereinander. Während dieses Merkmal für beide Autoren im Mittel ähnlich ausgeprägt sein sollte, ist bei Adorno erneut mit deutlicheren Ausreißern zu rechnen. Das liegt wiederum an den zentralen Begriffen, die bei Adorno mit einer großen Anzahl anderer Begriffe verbunden sind, wodurch die Egonetzwerke dieser Begriffe zunächst eine außergewöhnlich hohe Anzahl an Nachbarn enthalten. Werden dann zusätzlich die Verbindungen der Nachbarn untereinander berücksichtigt, verstärkt das die Ausprägung dieses Charakteristikums abermals. Im Modellnetzwerk ist es in diesem Fall die Merkmalsausprägung von Knoten *G*, die auffällt. Da nur wenige solcher Ausreißer zu erwarten sind, sollten Adornos Texte insbesondere höhere Kurtosis-Werte aufweisen. Gleichzeitig sollte auch die Standardabweichung bei den Texten Adornos höher sein als bei denen von Carnap.

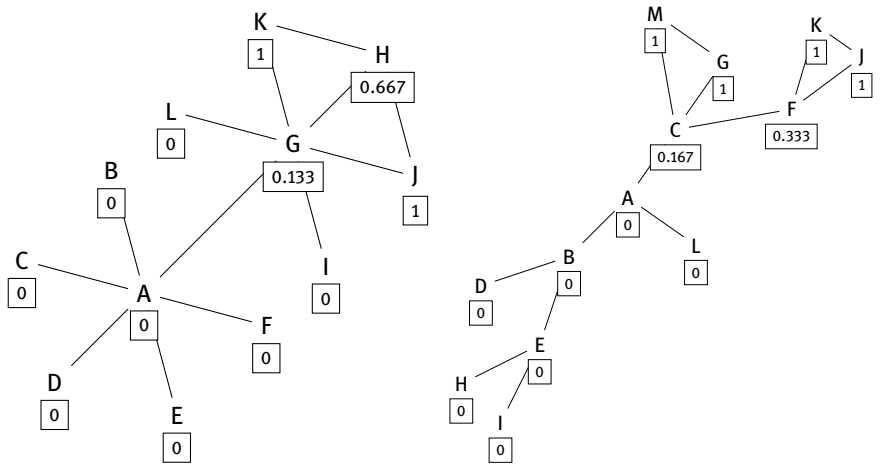


Abb. 9: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: Lokale Transitivität (Merkmal 4).

Merkmal 4: Transitivität
Lokaler Clustering-Koeffizient (Transitivität) von *i*.

$$m_4(i) = \frac{|E_{\text{ego}_1(i)}|}{|\,\text{nb}(i)| \times |(\text{nb}(i) - 1)|} \tag{4}$$

Der lokale Clustering-Koeffizient beschreibt, wieviele der insgesamt möglichen Kanten im (Ego-)Netzwerk auch tatsächlich realisiert werden. Damit soll erfasst werden, ob sich in den Netzwerken Triaden ausbilden, also Verknüpfungen unter den Nachbarn von *i*. Das wird eher in den Netzwerken zu Carnaps Texten erwartet, da seinem Ideal der Begriffsverknüpfung entsprechend nicht nur das Verhältnis von Unter- zu Oberbegriffen, sondern potenziell auch das der Unterbegriffe zueinander ausbuchstabiert wird. Carnaps Texte sollten daher einen höheren Mittelwert, jedoch weniger Ausreißer, d. h. niedrigere Kurtosis-Werte, als die Texte von Adorno aufweisen.

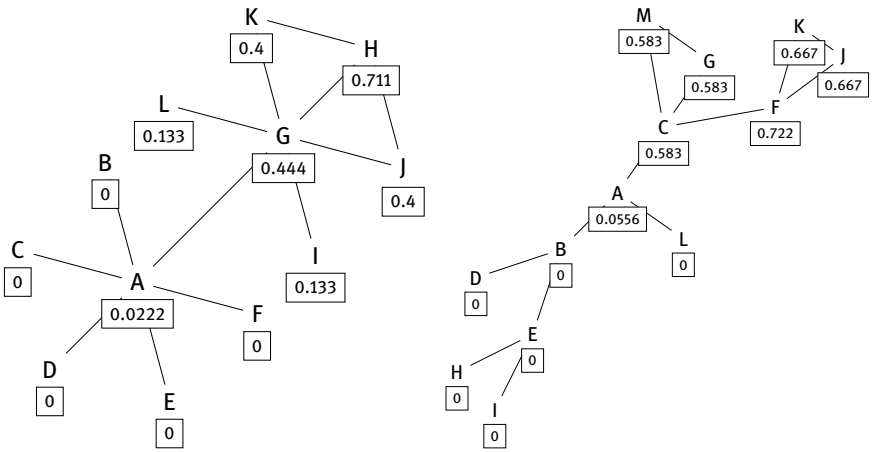


Abb. 10: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: Clustering-Koeffizienten (Merkmal 5).

Merkmal 5: Clustering-Koeffizient (= Transitivität der Nachbarn)
Normalisierter Clustering-Koeffizient der lokalen Clustering-Koeffizienten der Nachbarn von *i*.

$$m_5(i) = \frac{1}{\deg(i)} \sum_{j \in \mathrm{nb}(i)} m_4(j) \tag{5}$$

Neben der lokalen Transitivität (Merkmal 4) wird auch die Transitivität der Nachbarn berücksichtigt. Damit soll die Existenz von Begriffen einfließen, die den zentralen Begriffen beigeordnet sind. Dieses Merkmal nimmt einen hohen Wert an, wenn ein beispielhafter Knoten *i* eine Verbindung zu dem stark vernetzten Egonetzwerk von Knoten *j* hat, wie das zum Beispiel bei der Verbindung des Knoten *H* mit dem Knoten *G* im Modellnetzwerk von Adorno der Fall ist. Erwartet wird, dass dieses Phänomen häufiger bei Carnap auftritt (höheres arithmetisches Mittel). Gleichzeitig wird erwartet, dass in den Texten Adornos – wie im Modell angedeutet – eine Vielzahl anderer Begriffe den zentralen Begriffen beigeordnet wird, die selber aber nicht unbedingt näher bestimmt werden. Auch hier kann mit starken Ausreißern in Adornos Texten gerechnet werden (höhere Kurtosis). Für Adorno ist zudem eine rechtsschiefe Verteilung zu erwarten, also ein negativer Schiefe-Wert (viele der Begriffe weisen einen geringen Merkmalswert auf).

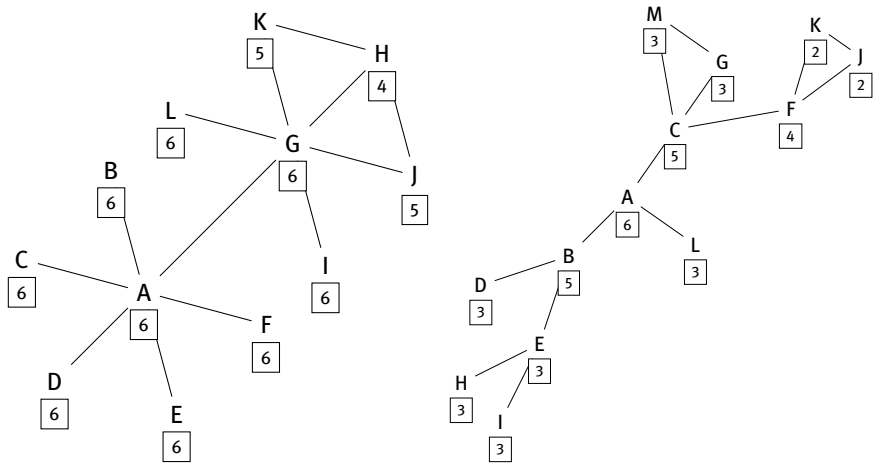


Abb. 11: Modellnetzwerke zu Adorno (links) und Carnap (rechts). In Kästen: Anzahl der Nachbarsnachbarn (Merkmal 6).

Merkmal 6: Nachbarsnachbarn
Anzahl der Nachbarsnachbarn von Knoten *i*

$$m_6(i) = |\,\mathrm{ego}_2(i) \setminus \mathrm{nb}(i)| = |\,\mathrm{ego}_2(i)| - \deg(i) \tag{6}$$

Der Wert dieses Merkmals wird auf Basis eines Ego_2-Netzwerks bestimmt, also eines Netzwerks, das Knoten bis zu einer Entfernung von 2 beinhaltet. Ermittelt wird die Anzahl der Knoten des Ego_2-Netzwerks, ohne die direkten Nachbarn von *i* zu berücksichtigen, also die Nachbarsnachbarn ohne die direkten Nachbarn, allerdings inklusive des Knotens *i* (da jeder Knoten auch ein Nachbarsnachbar von sich selbst ist). Damit sollen die bereits angesprochenen Triaden auch in einem größeren Kontext ermittelt werden. Da die Knoten in Adorno-Netzwerken vergleichsweise dicht zusammenhängen, ist hier auch mit einer höheren durchschnittlichen Anzahl an Nachbarsnachbarn sowie einer geringeren Standardabweichung zu rechnen.

Tab. 5: Übersicht über die formulierten Hypothesen für die einzelnen Merkmale. *A*: Adorno, *C*: Carnap.

Merkmal	Mittelwert	Standardabweichung	Kurtosis
1: *Closeness*	$A > C$		
2: Knotengrad			$A > C$
3: Anzahl Kanten	$A \approx C$	$A > C$	$A > C$
4: Lokale Transitivität	$A < C$		$A > C$
5: Clustering-Koeffizient	$A < C$		$A > C$
6: Nachbarsnachbarn	$A > C$	$A < C$	$A > C$

7 Digitale Analyse: Egonetzwerke

7.1 Clustering der Netzwerke

Im Folgenden sollen die aus den Texten von Adono und Carnap konstituierten Begriffsnetzwerke miteinander verglichen werden. Jedes dieser Begriffsnetzwerke wird durch einen Signaturvektor beschrieben. Dieser Signaturvektor wird auf folgende Art und Weise gebildet: Aus den Egonetzwerken der Begriffe werden die oben erläuterten sechs Merkmale extrahiert. Dazu werden die Spalten der N×6-Matrix zunächst einzeln mit ihrer *L2*-Norm normalisiert – praktisch bewegen sich dadurch alle Einträge eines Vektors im Intervall [0;1] – und anschließend jeweils mit Mittelwert, Median, Standardabweichung, Schiefe (*skewness*) und Kurtosis (Wölbung) beschrieben. Aus der Konkatenation der fünf Aggregatoren pro Spalte ergibt sich ein 30-elementiger Signaturvektor pro Netzwerk. Die Unähnlichkeit zweier Netzwerke kann so über die Canberra-Distanz in Adkins-Form quantifiziert werden, die zwischen den dazugehörigen Signaturvektoren bestimmt wird (Lance und Williams 1967):[19]

$$d\left(p,q\right) = \frac{1}{n-Z}\sum_{i=1}^{n}\frac{|p_i - q_i|}{|p_i| + |q_i|} \tag{7}$$

Im Folgenden wird der Wert dieser Distanzberechnung synonym als Unähnlichkeit bezeichnet und meint die Unähnlichkeit zwischen den dazugehörigen

19 Darin sind $p=(p_1,p_1,...,p_n)$ und $q=(q_1,q_1,...,q_n)$ zwei n-elementige Vektoren und Z die Anzahl der Elemente in p und q, die Null sind. Da die Canberra-Distanz in der angegebenen Form die Dreiecksungleichung erfüllt, also eine Metrik im engeren Sinne ist, kann sie in einem Clustering-Verfahren eingesetzt werden. Indem die Merkmale stets für alle Knoten der Netzwerke bestimmt und dann die Verteilungen der Merkmalsausprägungen analysiert werden, enthält der für den Vergleich wichtige Signaturvektor zudem einen hohen Informationsgehalt.

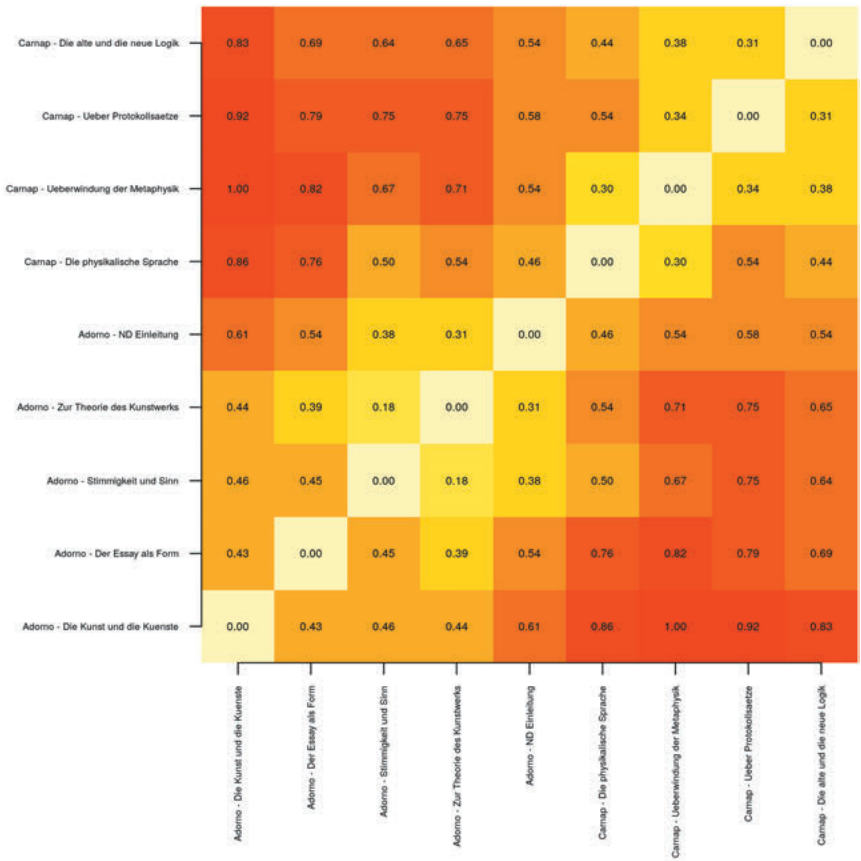


Abb. 12: Distanzen zwischen allen Paaren von Texten

Netzwerken. Das Ergebnis aller paarweisen Vergleiche der normierten Signaturvektoren ist eine Distanzmatrix. Die Abbildung 12 zeigt ein Wärmebild der Vergleiche (*heatmap*), also eine farblich kodierte Darstellung der Distanzmatrix, in dem ein zunehmend roter Farbton für eine zunehmende Unähnlichkeit steht. Für diese Darstellung wurden die Distanzen normiert, sodass das Wärmebild relative Unähnlichkeiten zeigt.

In der Abbildung 12 sieht man, dass Carnaps „Überwindung der Metaphysik“ und Adornos „Die Kunst und die Künste“ den größten Unterschied zueinander in diesem Datensatz aufweisen. In der Abbildung ist auch eine Tendenz zur Bildung zweier Gruppen erkennbar, anhand derer die Autoren unterschieden werden können: Adornos Texte sind sich untereinander ähnlicher als zu denen Carnaps und

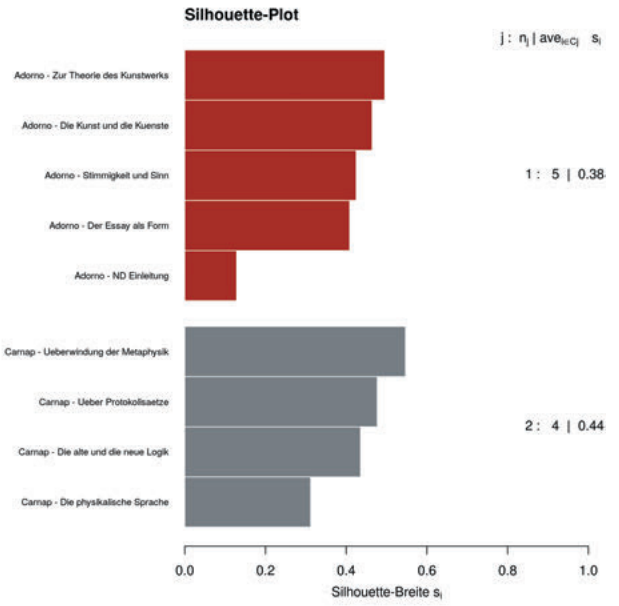


Abb. 13: Silhouette des Clusterings

umgekehrt. Eine Ausnahme bilden die „Einleitung" von Adornos *Negativer Dialektik* und „Die physikalische Sprache als Universalsprache der Wissenschaft" von Carnap. Die „Einleitung" weist zu „Der Essay als Form" und zu „Die Kunst und die Künste" vergleichsweise hohe relative Unähnlichkeiten von >0,5 auf, die in etwa den Unähnlichkeiten zu Carnaps Texten entsprechen (ebenfalls >0,5). Ähnliches gilt für „Die physikalische Sprache als Universalsprache der Wissenschaft", die insbesondere zu Carnaps „Über Protokollsätze" eine stark ausgeprägte relative Unähnlichkeit zeigt. Trotz der Tendenz zur Trennung der Autoren sind sich die Texte eines Autors untereinander vergleichsweise unähnlich. So weisen die Texte von Adorno untereinander relative Distanzen zwischen 0,18 und 0,61 auf und die von Carnap Distanzwerte zwischen 0,30 und 0,58.

Die Tendenz zur Gruppenbildung wurde zusätzlich mit Hilfe einer hierarchischen Cluster-Analyse mittels *complete-linkage*-Fusionierungsstrategie evaluiert: Zu Beginn dieses Algorithmus bilden zunächst alle Elemente ihr eigenes Cluster. Diese Cluster werden dann sukzessive in ein einziges Cluster zusammengefasst. Währenddessen wird die größtmögliche Unähnlichkeit zwischen den Elementen der zusammenfallenden Cluster notiert. Diese Cluster-Analyse ist also besonders für eine explorative Evaluation geeignet.

Das Ergebnis dieser Cluster-Analyse ist in Abbildung 13 dargestellt. Die Abbildung zeigt eine Silhouette der Cluster-Analyse (Rousseeuw 1987). Darin repräsentieren die Balken die gruppierten Texte. Die Breite eines Balkens gibt das Verhältnis der Abstände zu allen Instanzen derselben Gruppe zu den Abständen der Instanzen aller anderen Gruppen an: Je näher der Balken am maximalen Wert 1,0 liegt, desto eindeutiger ist die Zuordnung einer Instanz zur entsprechenden Gruppe.

Anhand der Abbildung lässt sich zunächst die Trennung der beiden Autoren erkennen, wie sie sich in Abbildung 12 bereits angedeutet hat. In dieser Darstellung fällt zudem die „Einleitung" zur *Negativen Dialektik* erneut ins Auge, die zwar der Gruppe von Adornos Texten zugeordnet wird, insgesamt aber nah an den Instanzen des Carnap-Clusters liegt. Zudem ist die durchschnittliche Breite mit 0,47 nur mittelmäßig ausgeprägt. An dieser Stelle lässt sich dennoch festhalten, dass die gewählte Vergleichsmethode grundsätzlich geeignet ist, die unterschiedliche Verwendung von Begriffen, wie Adorno und Carnap sie praktizieren, zu erfassen. Dennoch scheinen die Merkmale die Trennung nur knapp zu erreichen – der geringe durchschnittliche Silhouette-Koeffizient belegt dies.

Im Folgenden werden sämtliche Merkmalsausprägungen im Detail betrachtet und hinsichtlich der Hypothesen über die relativen Unterschiede untersucht. Das bedeutet, dass alle $N\times6$-Merkmalsmatrizen herangezogen werden, deren Konstruktion im Abschnitt 6 beschrieben ist. Die Spalten dieser Matrizen werden durch jeweils fünf (ebd. genannte) statistische Aggregatoren beschrieben, also den Median, das arithmetische Mittel, Standardabweichung, Schiefe und Kurtosis. Um die Unterschiede der Merkmalsausprägungen bei den Autoren besser ersichtlich zu machen, wird eine grafische Darstellung der aggregierten Merkmalsmatrizen gewählt. Es folgt also zu jedem Merkmal eine Grafik, in der die Texte der Autoren hinsichtlich ihrer statistischen Kennzahlen verglichen werden.

7.2 Analyse der Merkmale im Einzelnen

Im Folgenden werden die sechs Merkmale einzeln diskutiert und die gemessenen Werte mit den Hypothesen abgeglichen. Dabei ist zu beachten, dass sich die vorhergesagten Verteilungen auf die Verteilungen der Merkmalswerte innerhalb eines Netzwerks, d. h. eines Textes, beziehen. Tabelle 6 liefert eine kondensierte Darstellung der Hypothesen und ihrer Bestätigung bzw. Ablehnung.

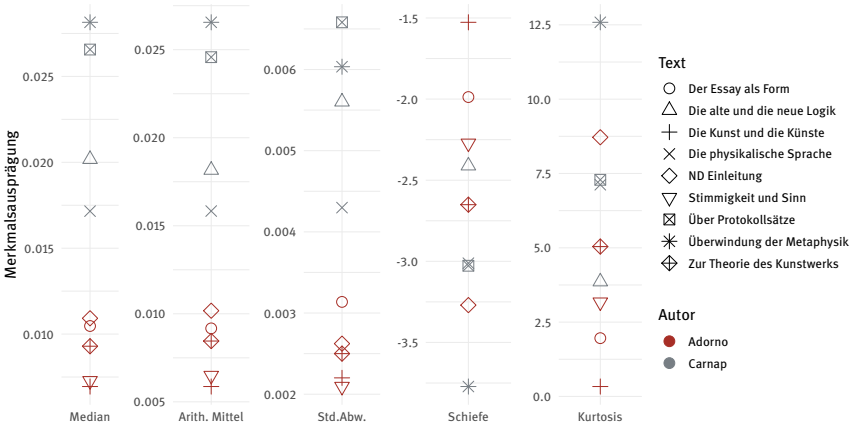


Abb. 14: Merkmal 1: Durchschnittliche *closeness*

Merkmal 1: *Closeness*

Hypothese

Die Pfaddistanzen sollten bei Adorno kürzer ausfallen. Daher war hier mit einer insgesamt kompakteren Verteilung mit einigen wenigen Ausreißern zu rechnen. Im Vergleich zu Carnap sollte daher der Mittelwert größer sein.

Auswertung

In Adornos Textnetzwerken haben die Begriffe entgegen den Erwartungen eine geringere durchschnittliche *closeness* zueinander als diejenigen in den Texten Carnaps, und das bei sehr geringen Werten. Trotzdem diskriminiert das Merkmal die Texte Adornos klar von denen Carnaps, was ebenfalls für den Median und die Standardabweichung zutrifft. Bei Kurtosis und Schiefe hingegen ist keine klare Trennung der Autoren zu erkennen. Dabei zeigen die negativen Schiefe-Werte, dass die *closeness* zwischen den Knoten relativ gleichmäßig verteilt ist. Zudem deutet die positive Kurtosis darauf hin, dass sowohl in den Texten von Carnap als auch in denjenigen Adornos Ausreißer zu finden sind.

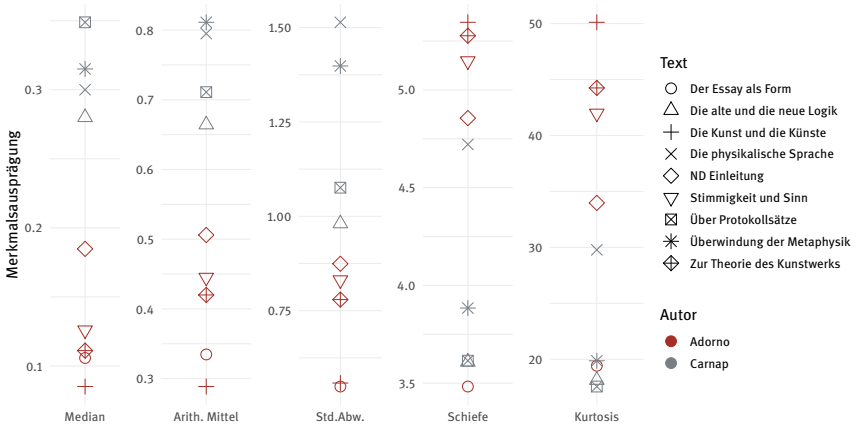


Abb. 15: Merkmal 2: Knotengrad der Nachbarn

Merkmal 2: Knotengrad der Nachbarn

Hypothese

Bei diesem Merkmal sollten die Verteilungen in den Texten Adornos unregelmäßiger sein und stärkere Ausreißer aufweisen als in den Texten von Carnap. Erwartet wird daher eine höhere Kurtosis.

Auswertung

Die Texte werden insgesamt gut voneinander getrennt. Die Werte bestätigen fast durchgehend die Erwartungen – mit einer Ausnahme: Entgegen der Hypothese nehmen in Adornos „Der Essay als Form" Schiefe und Kurtosis sehr geringe Werte an. Analog verhält es sich mit „Die physikalische Sprache als Universalsprache der Wissenschaft" von Carnap. Dieser Text weist hinsichtlich der Kurtosis unerwartet hohe Werte auf. Beachtenswert ist zudem die teilweise große Streuung bei Carnaps Texten, auch hier insbesondere bei „Die physikalische Sprache als Universalsprache der Wissenschaft". Die positive Schiefe deutet hier, wie auch bei den folgenden Merkmalen, bei beiden Autoren darauf hin, dass viele Begriffe in Betreff des Merkmals eher geringe Werte aufweisen.

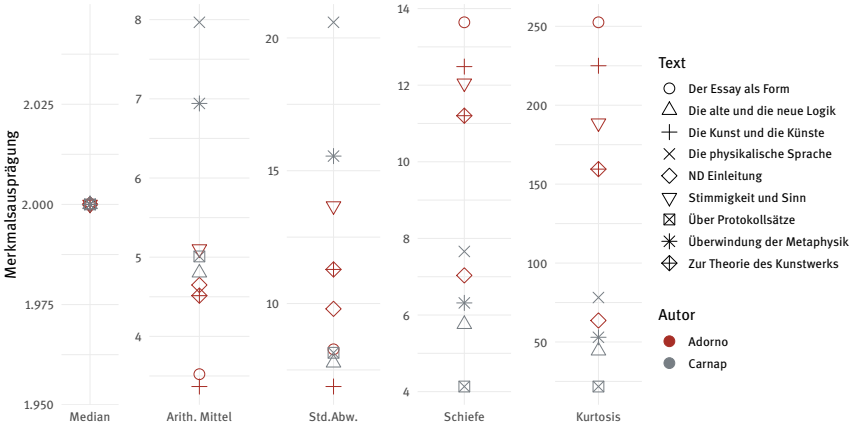


Abb. 16: Merkmal 3: Anzahl der Kanten im Egonetzwerk

Merkmal 3: Anzahl Kanten

Hypothese

Bezüglich dieses Merkmals sollten Adorno und Carnap im Durchschnitt nahe beieinander liegen. Bei Adorno wird jedoch mit deutlicheren Ausreißern, d. h. mit höheren Kurtosis-Werten und einer größeren Standardabweichung gerechnet.

Auswertung

In Hinblick auf das Merkmal 3 liegen beide Autoren im Mittel nahe beieinander – im Median decken sie sich sogar im Wert 2. Wie vermutet sind jedoch Unterschiede in Bezug auf die Kurtosis feststellbar. Einzig die „Einleitung" zur *Negativen Dialektik* von Adorno weicht hier von der Hypothese ab. In diesem Text sind wesentlich schwächere Ausreißer auszumachen, was sich in der vergleichsweise geringen Kurtosis widerspiegelt. Eine höhere Standardabweichung für Adornos Texte kann hingegen nicht festgestellt werden, da sie sich, wie manche Texte von Carnap, im unteren Teil des Spektrums sammeln.

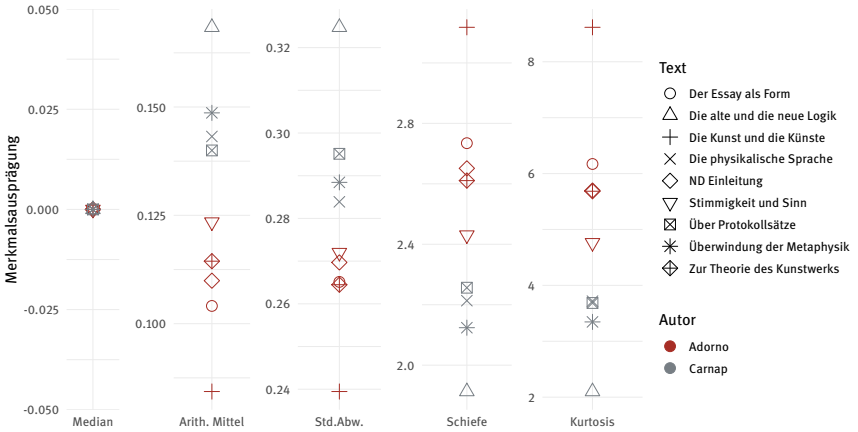


Abb. 17: Merkmal 4: Lokale Transitivität

Merkmal 4: Transitivität

Hypothese

Hier wird in den Netzwerken zu Adornos Texten mit einem niedrigeren arithmetischen Mittelwert und einer höheren Kurtosis gerechnet.

Auswertung

In Merkmal 4 bestätigt sich die oben formulierte Hypothese. Im Median gleichen sich zwar alle Texte (am Wert 0). Dennoch ist in den Texten Carnaps eine stärkere mittlere Cluster-Bildung erkennbar. Bei Adorno sind es demgegenüber nur wenige Egonetzwerke, in denen eine für seine Texte ungewöhnlich hohe Anzahl an Triaden geschlossen ist, wie die Kurtosis-Werte seiner Texte belegen. Dass sich diese nicht nennenswert auf die Standardabweichung auswirken, unterstreicht deren geringe Anzahl.

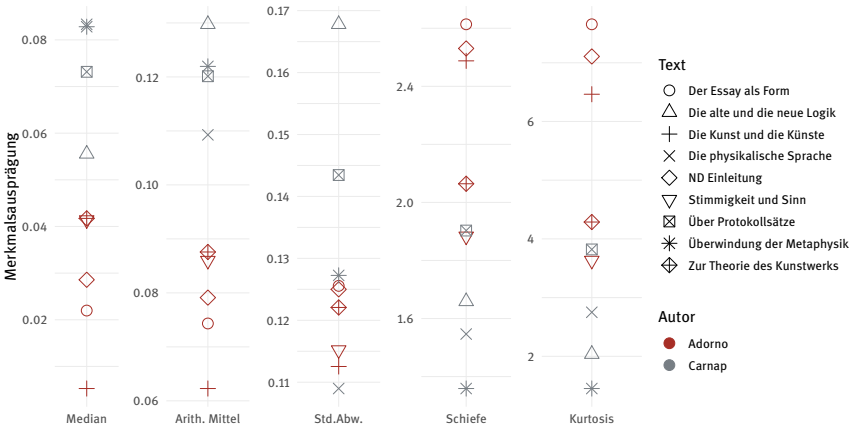


Abb. 18: Merkmal 5: Transitivität der Nachbarn

Merkmal 5: Clustering-Koeffizient

Hypothese

Hier sollten Adornos Texte einen geringeren Mittelwert und höhere Kurtosis aufweisen.

Auswertung

Die Erwartungen werden hier fast durchgehend erfüllt: Die angesprochenen Merkmale diskriminieren die beiden Autoren klar, lediglich Carnaps Text „Über Protokollsätze“ fällt mit seiner Kurtosis etwas aus dem Rahmen. Daraus kann gefolgert werden, dass es in den Texten von Adorno deutlich mehr beigeordnete Begriffe gibt als bei Carnap.

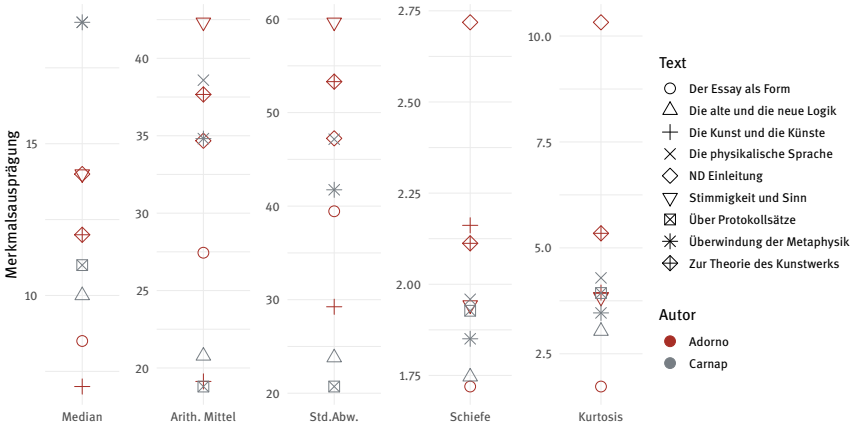


Abb. 19: Merkmal 6: Anzahl der Nachbarsnachbarn

Merkmal 6: Nachbarsnachbarn

Hypothese

Auch bei diesem Merkmal wird mit starken Ausreißern bei Adorno, sowie einem höheren durchschnittlichen Wert und einer geringeren Standardabweichung gerechnet.

Auswertung

Keine der Metriken vermag die Texte von Adorno und Carnap sauber zu trennen. An den Nachbarsnachbarn scheint sich die unterschiedliche Verfahrensweise nicht zu manifestieren. Gleichwohl treffen die Hypothesen zur Kurtosis in Hinblick auf Adorno für drei seiner fünf Texte zu. „Der Essay als Form“ und „Stimmigkeit und Sinn“ sind hier nicht von den Texten Carnaps zu unterscheiden. Es scheint in diesen beiden Texten weniger Begriffe mit ungewöhnlich hohen Knotengraden zu geben, die die Merkmalsverteilung in ihrer Kurtosis beeinflussen würden.

Tab. 6: Übersicht über die formulierten Hypothesen für die einzelnen Merkmale. *A*: Adorno, *C*: Carnap. Die Symbole ✔ und ✖ markieren, ob eine Hypothese bestätigt werden konnte.

Merkmal	Mittelwert	Standardabweichung	Kurtosis
1: *Closeness*	$A > C$ ✖		
2: Knotengrad			$A > C$ ✔
3: Anzahl Kanten	$A \approx C$ ✔	$A > C$ ✖	$A > C$ ✔
4: Lokale Transitivität	$A < C$ ✔		$A > C$ ✖
5: Clustering-Koeffizient	$A < C$ ✔		$A > C$ ✔
6: Nachbarsnachbarn	$A > C$ ✖	$A < C$ ✖	$A > C$ ✔

8 Zusammenfassung und Interpretation der algorithmischen Analyse

Fasst man die Resultate der digitalen Analyse zusammen und versucht sich im Zuge dessen an ihrer Deutung, ergibt sich folgendes Bild: Erstens hat der Gesamtvergleich ausgewählter Texte Adornos und Carnaps bestätigt, dass die beiden Autoren bei der Begriffsverknüpfung unterschiedlich verfahren (siehe die Abbildungen 12 und 13 oben). Die dabei qua Clustering erfolgte Bildung von in ihren Eigenschaften ähnlichen Textgruppen (siehe die Silhouette in Abbildung 13) legen nahe, dass beide Autoren voneinander abweichende Modi der Begriffsverknüpfung praktizieren, die sich jedoch weitaus weniger von einander unterscheiden als in Anbetracht der Idealvorstellungen von Adorno und Carnap in Hinblick auf den Begriffsgebrauch zu erwarten wäre.

Die Gruppenbildung von Adornos Texten (Abbildung 13) kann durch die Zuhilfenahme geisteswissenschaftlicher Kontexte erläutert werden. Im Falle der Texte von Adorno entspricht sie der Gattungszugehörigkeit der jeweiligen Texte: Eine Gruppe umfasst die Essays „Der Essay als Form“ und „Die Kunst und die Künste“, eine weitere die „Einleitung“ zur *Negativen Dialektik*, und eine dritte zwei Kapitel aus der Leseausgabe der *Ästhetischen Theorie*. Bei Carnap scheint nicht die Gattungszugehörigkeit für die Gruppenbildung verantwortlich zu sein – alle vier Texte sind wissenschaftliche Aufsätze –, sondern der dominierende Modus derselben: Während „Die physikalische Sprache als Universalsprache der Wissenschaft“ und „Die Überwindung der Metaphysik durch logische Analyse der Sprache“ vorwiegend argumentativ-definitorisch verfahren, herrscht in den anderen beiden Texten von Carnap ein narrativer Modus vor. Dieser könnte der Tatsache geschuldet sein, dass die beiden Texte den Leser in neue Formen des Philosophierens einführen, diese jedoch nicht selbst praktizieren.

Zweitens hat sich im Zuge der digitalen Analyse herausgestellt, dass Adornos und Carnaps Texte in der Art und Weise variieren, wie in ihnen Begriffe miteinander verknüpft werden. Das bedeutet, dass also nicht nur zwischen den Texten der beiden Autoren, sondern auch zwischen den Texten desselben Autors Unterschiede in der Begriffsverknüpfung bestehen, was insbesondere bei Adorno dazu führt, dass einer seiner Texte – die „Einleitung" zur *Negativen Dialektik* – nur eine relativ geringe Übereinstimmung mit den restlichen Texten des Korpus aufweist (Abbildungen 12 bis 13). Auch hierfür scheint erneut die Gattungszugehörigkeit der „Einleitung" verantwortlich zu sein. Wie unter anderem Axel Honneth gezeigt hat, kennzeichnet sich dieser Text dadurch, dass Adorno in ihm sämtliche zentralen Topoi der *Negativen Dialektik* kurz vorstellt, sich also nicht auf eine einzelne Fragestellung konzentriert (cf. Honneth 2004). Die anderen Texte fokussieren dagegen stets auf einen zentralen Topos bzw. ‚Begriff', sei es der Essay in „Der Essay als Form" oder die Kunst in „Die Kunst und die Künste".[20]

Drittens hat der Vergleich der sechs netzwerkanalytischen Merkmalsausprägungen in den ausgewählten Texten von Adorno und Carnap die eingangs entwickelten Hypothesen zu deren Begriffsgebrauch mit nur sehr geringen Einschränkungen bestätigt: Bei vier der sechs Merkmale (soziales Kapital, Anzahl der Kanten, lokale Transitivität, Transitivität der Nachbarn) haben sich die zuvor entwickelten Hypothesen fast durchgehend erfüllt. Nur in Hinblick auf zwei Merkmale (*closeness*, Anzahl der Nachbarsnachbarn) haben sich die Hypothesen nicht erfüllt. Bei einem dieser Merkmale, der Anzahl der Nachbarsnachbarn, ist es Adornos „Der Essay als Form", der letztendlich zur Falsifikation der Hypothese geführt hat. Dieser Text weicht auch bei anderen Merkmalen – z. B. dem Knotengrad der Nachbarn – von den restlichen Texten Adornos ab, was Zweifel aufkommen lässt in Hinblick auf die weitverbreitete Forschungsmeinung, dass es sich bei diesem Text nicht nur um die zentrale Metareflexion von Adornos Verfahrensweise, sondern zugleich deren paradigmatische Umsetzung handelt. Die Werte, welche die einzelnen statistischen Maße in Hinblick auf die *closeness* annehmen, zeigen dagegen, dass eine der Vorannahmen in Bezug auf die Begriffsverknüpfung bei Adorno zu präzisieren ist, nämlich diejenige, wie sich diejenigen Begriffe Adornos, die zur Bestimmung der wenigen zentralen Begriffe herangezogen werden, zueinander vehalten: Das geringe arithmetische Mittel in Kombination mit der ebenfalls sehr geringen Standardabweichung bei der *closeness* sprechen – noch zusätzlich

20 Offensichtlich wird dies durch einen Vergleich der Ausreißer im Merkmal 3 – der Anzahl der Kanten im Egonetzwerk: In „Der Essay als Form" besitzt in diesem Merkmal nur der Begriff „Essay" einen Wert über 100 (exakt: 168), der Begriff mit dem nächsthöheren Wert ist ‚Gedanke' mit 44. In der Einleitung gibt es dagegen zwei Begriffe – ‚Philosophie' und ‚Begriff' –, die einen Wert über 100 besitzen.

unterstützt durch die Werte des Merkmals 4 (Lokale Transitivität) – dafür, dass besagte Begriffe untereinander nur in seltenen Fällen miteinander verknüpft sind. Die diesbezüglich höheren Werte bei den Texten von Carnap legen nahe, dass es bei Letzterem in Hinblick auf die *closeness* zu einer wesentlich kompakteren Verteilung kommt. Dies spricht dafür, dass Carnap – im Unterscheid zu Adorno – weitaus häufiger auch nicht im Mittelpunkt seiner Texte stehende Begriffe zueinander ins Verhältnis setzt; eine Vermutung, die von der Lektüre der Texte der beiden Autoren sowie den aus einzelnen Absätzen in Abschnitt 5 extrahierten Netzwerkgraphen gestützt wird.

Des Weiteren kann man aus den vier erfüllten Merkmalsausprägungen in Hinblick auf die Modi der Begriffsverknüpfungen bei Adorno und Carnap noch folgende Schlüsse ziehen: Die ausgewählten Texte Adornos zeichnen sich dadurch aus, dass in ihnen einige wenige zentrale Begriffe durch eine Vielzahl anderer Begriffe ‚bestimmt' werden, die nur selten auch zueinander in einem direkten Verhältnis stehen (Indikatoren: soziales Kapital, Anzahl der Kanten). Nicht die hierarchische Gliederung der Begriffe sowie die Bestimmung ihres logischen Verhältnisses zueinander, sondern die Ausschöpfung des semantischen Potenzials einiger weniger Begriffe stehen im Zentrum von Adornos Essay (Indikator: lokale Transitivität, Transitivität der Nachbarn). Adorno scheint also tatsächlich in Opposition zum traditionellen definitorischen Verfahren ‚das Spezifische des Gegenstandes', auf das er mit seinen zentralen Begriffen zielt, dadurch zu fassen zu versuchen, dass er das semantische Potenzial dieser Begriffe aktiviert, indem er sie mit einer Vielzahl anderer Begriffe zusammenführt.

Der Sonderstatus der Einleitung zur *Negativen Dialektik* und deren daraus sich ergebende relative Ähnlichkeit mit Carnaps „Die physikalische Sprache als Universalsprache der Wissenschaft" lässt sich abermals durch Rückgriff auf fachspezifisches Domänenwissen erklären. Es könnte eine Folge der Tatsache sein, dass Adorno in der „Einleitung" das semantische Potenzial seiner zentralen Begriffe noch nicht vollkommen ausschöpft – das ist Aufgabe der folgenden Abschnitte des Buches –, sondern den Problemhorizont des Buches und damit die zentralen Begriffe nur kurz skizziert. Letztere werden dementsprechend noch nicht vollständig in ihrem Facettenreichtum dargestellt wie in den darauffolgenden Kapiteln.

Carnap hingen verfährt tatsächlich anders als Adorno: Wie die Werte des arithmetischen Mittels zum sozialen Kapital (= Knotengrad der Nachbarn) der Texte Carnaps belegen, ähnelt sich das Verhältnis der Begriffe zumindest in dieser Hinsicht in sämtlichen der ausgewählten Texte: Nicht einzelne Hauptbegriffe, sondern die präzise Bestimmung des Verhältnisses sämtlicher Begriffe zueinander charakterisiert den Text. Dafür sprechen auch die höheren Werte der lokalen Transitivität sowie der Transitivität der Nachbarn in Carnaps Texten: Carnap hier-

archisiert seine Begriffe nicht nur, sondern differenziert auch diejenigen, die sich auf derselben Ebene der Begriffspyramide befinden, indem er auch sie zueinander ins Verhältnis setzt und so voneinander abgrenzt. Dies führt zu stärkerer Triadenbildung und somit zu den genannten höheren Werten. Die ausgewählten Texte erfüllen so die im *Manifest des Wiener Kreises* ausformulierten und eingangs vorgestellten metaphilosophischen Vorgaben zur Begriffsgestaltung.

Dies gilt – viertens – auch für Adornos Konzept der Konstellation, das sich auf Basis der digitalen Analyse näher bestimmen lässt. So handelt es sich bei denjenigen Begriffen der ausgewählten Texte Adornos, die für die Merkmale 1, 2, 3, und 6 die höchsten Werte besitzen, jeweils um die zentralen Begriffe der Texte: den ‚Essay' in „Der Essay als Form", die ‚Kunst' in „Die Kunst und die Künste" sowie ‚Philosophie' in der Einleitung der *Negativen Dialektik*.

In Summa bedeutet dies, dass die beiden am Anfang dieses Aufsatzes ausformulierten Fragen in Hinblick auf die für sie entwickelten Netzwerkmodelle mit einem ‚ja' beantwortet werden können: Im Rahmen der im Zuge dieses Aufsatzes entwickelten Modelle für die Referenzen auf Begriffe von der Textoberfläche sowie der damit einhergehenden Modellierung der Begriffsverknüpfung entspricht Adornos Begriffsgebrauch seinen eigenen Vorgaben und unterscheidet sich von demjenigen in den Texten von Rudolf Carnap. Diese Unterschiede sind jedoch – zumindest im Rahmen des im Zuge dieser Arbeit realisierten Modells – weitaus geringer als erwartet. Adorno scheint seine diesbezüglichen Vorgaben konsequenter zu realisieren als Carnap. Dies liegt eventuell daran, dass Carnap – wohl aus stilistischen Gründen – sich nicht auf eine rein deduktive Ausdifferenzierung seiner Begriffe beschränkt, sondern zu deren Bestimmung auch andere Stilmittel wie zum Beispiel Fragen heranzieht, was durch die Nähe der netzwerkanalytischen Metriken von Texten wie „Die physikalische Sprache als Universalsprache der Wissenschaft" zu denjenigen von Adorno bestätigt wird. Trotz einer partiellen Nähe zu Adornos Verfahrensweisen haben sowohl die Netzwerkanalysen (siehe insbesondere das Merkmal zur lokalen Transitivität) als auch die Absatz-Lektüre gezeigt, dass Carnap seinen eigenen metaphilosophischen Forderungen entsprechend an einer hierarchieartigen Ausdifferenzierung der von ihm behandelten Begriffe arbeitet. In der konkreten textuellen Realisierung manifestiert sich eine solche jedoch weitaus weniger rigoros als in Anbetracht von Carnaps Metaphilosophie zu erwarten wäre.

Zum Schluss des Artikels möchten wir auf die eingangs aufgeworfene Frage, inwiefern algorithmische Verfahren der Textanalyse auch in Szenarien mit verhältnismäßig kleinen Daten fruchtbar sein können, zurückkommen. Mit einer Korpusgröße von neun Texten ist diese Arbeit weit weg von dem, was man gemeinhin als ‚*big data*' bezeichnet. Zudem widmet sie sich einer Fragestellung, die – zumindest auf den ersten Blick – ihre Beantwortung mithilfe traditioneller textwis-

senschaftlicher Verfahren nahelegt. Gleichwohl beinhaltet die hier gezeigte Analyse, würde sie von einem Menschen durchgeführt, eine große Menge an Einzelentscheidungen und arbeitsaufwendigen sowie fehleranfälligen Auszählungen: Alleine die (korrekte) Zählung von Ko-Okkurrenzen von Begriffen erscheint zwar nicht als unmöglich, erfordert aber einen ernstzunehmenden Zeit- und Energieeinsatz – mit einer vorherigen syntaktischen Analyse der Sätze im Einzelnen hätte man da noch gar nicht angefangen. Menschliche Leserinnen und Leser können natürlich, das ist auch klar, bei der Analyse auf ‚höherem' Niveau einsteigen: Eine operationalisierte Definition davon, was unter einem Begriff zu verstehen ist, kann bei philosophisch vorgebildeten Lesenden vielleicht entfallen – sofern dies nicht zu Lasten der Exaktheit geht. Die konsistente Anwendung eines exakt definierten Begriffs-Verständnisses ist jedoch keine Tätigkeit, die Menschen besonders leicht fällt: Nach der Lektüre von Aufsätzen, die wie die Texte von Adorno in einer sehr metaphernreichen Sprache von Begriffen sprechen, wird das Verständnis von Begrifflichkeit ein anderes sein, was sich wahrscheinlich in inkonsistenten Annotation niederschlagen würde. Ein konsistente Auszeichnung von Referenzen auf diese Begriffe von der Textoberfläche ist so kaum zu leisten. Ein quantitativer Vergleich ist allerdings nur dann sinnvoll, wenn überhaupt die Vorrausetzungen für Vergleichbarkeit gegeben sind. Dazu braucht es konsistente Anwendungen von Definitionen, die – auch bei kleinen Datenmengen – algorithmisch zuverlässiger realisiert werden können.

Danksagung: Für die manuelle Nachkorrektur der (computer-)linguistischen Annotationen danken wir den Hilfskräften Lars Amann und Fabian Schan. Für die Unterstützung bei der Koreferenzauflösung bedanken wir uns bei Hanna Winter.

Primärliteratur

Adorno, Theodor W. (2003a). *Ästhetische Theorie*. Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 7. Gesammelte Schriften. Frankfurt am Main: Suhrkamp.

Adorno, Theodor W. (2003b). „Der Essay als Form". In: *Noten zur Literatur*. Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 11. Gesammelte Schriften. Frankfurt am Main: Suhrkamp, S. 9–33.

Adorno, Theodor W. (2003c). „Die Kunst und die Künste". In: *Kulturkritik und Gesellschaft*. Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 10. Gesammelte Schriften. Frankfurt am Main: Suhrkamp, S. 432–453.

Adorno, Theodor W. (2003d). „Negative Dialektik". In: Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 6. Gesammelte Schriften. Frankfurt am Main: Suhrkamp, S. 7–412.

Carnap, Rudolf (1930). „Die alte und die neue Logik“. In: *Erkenntnis* 1, S. 12–26.
Carnap, Rudolf (1931a). „Die physikalische Sprache als Universalsprache der Wissenschaft“. In: *Erkenntnis* 2, S. 432–465.
Carnap, Rudolf (1931b). „Überwindung der Metaphysik durch logische Analyse der Sprache“. In: *Erkenntnis* 2, S. 219–241.
Carnap, Rudolf (1932 – 1933). „Über Protokollsätze“. In: *Erkenntnis* 3, S. 215–228.

Sekundärliteratur

Adorno, Theodor W. (2003e). „Parataxis“. In: *Noten zur Literatur*. Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 11. Gesammelte Schriften. Frankfurt am Main: Suhrkamp, S. 447–494.
Adorno, Theodor W. (2010). *Einführung in die Dialektik*. Hrsg. von Christoph Ziermann. Frankfurt am Main: Suhrkamp.
Adorno, Theodor W. und Max Horkheimer (2003). *Dialektik der Aufklärung*. Hrsg. von Rolf Tiedemann, Gretel Adorno, Susan Buck-Morss und Klaus Schultz. Bd. 3. Gesammelte Schriften. Frankfurt am Main: Suhrkamp.
Albrecht, Andrea (2010). „Konstellationen. Zur kulturwissenschaftlichen Karriere eines astrologisch-astronomischen Konzepts bei Henrich Rickert, Max Weber, Alfred Weber und Karl Mannheim“. In: *Scientia Poetica* 14, S. 104–149.
Berlingerio, Michele, Danai Koutra, Tina Eliassi-Rad und Christos Faloutsos (2013). „Network Similarity via Multiple Social Theories“. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '13. Niagara, Ontario, Canada: ACM, S. 1439–1440. DOI: 10.1145/2492517.2492582.
Bohnet, Bernd (2010). „Top Accuracy and Fast Dependency Parsing is not a Contradiction“. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, S. 89–97. URL: https://www.aclweb.org/anthology/C10-1011 (besucht am 1. Juni 2020).
Bonner, S., J. Brennan, G. Theodoropoulos, I. Kureshi und A.S. McGough (2016). „Efficient comparison of massive graphs through the use of 'graph fingerprints'.“ URL: http://dro.dur.ac.uk/19773/ (besucht am 1. Juni 2020).
Borgwardt, K. M. und H. P. Kriegel (2005). „Shortest-path kernels on graphs“. In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. DOI: 10.1109/ICDM.2005.132.
Burrows, John (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
Clément, Danièle und Michael Rödel (2016). „Begriff“. In: *Metzler Lexikon Sprache*. Hrsg. von Helmut Glück und Michael Rödel. 5. Aufl. Stuttgart: Metzler, S. 93–95.
Coleman, James S. (1966). „Individual interests and collective action“. In: *Papers on Non-market Decision Making* 1.1, S. 49–62. DOI: 10.1007/BF01718988.
Dahms, Hans-Joachim (1994). *Positivismusstreit*. Frankfurt am Main: suhrkamp.
Garey, Michael R und David S Johnson (2002). *Computers and intractability*. Bd. 29. WH Freeman.
Hogh, Philip (2015). *Kommunikation und Ausdruck. Sprachphilosophie nach Adorno*. Weilerswist: Velbrück Wissenschaft.

Honneth, Axel (2004). „Einleitung. Zum Begriff der Philosophie". In: *Theodor W. Adorno, Negative Dialektik*. Hrsg. von Axel Honneth und Christoph Menke. Klassiker Auslegen. Berlin: Akademie Verlag, S. 11–27.

Joshi, Mandar, Omer Levy, Daniel S. Weld und Luke Zettlemoyer (2019). *BERT for Coreference Resolution: Baselines and Analysis*.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Klein, Richard, Johann Kreuzer und Stefan Müller-Dohm, Hrsg. (2019). *Adorno-Handbuch. Leben – Werk - Wirkung*. 2. Aufl. Stuttgart: Metzler.

Kondor, Risi und Horace Pan (2016). „The Multiscale Laplacian Graph Kernel". In: *Advances in Neural Information Processing Systems 29*. Hrsg. von D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon und R. Garnett. Curran Associates, Inc., S. 2990–2998. URL: http://papers.nips.cc/paper/6135-the-multiscale-laplacian-graph-kernel.pdf (besucht am 1. Juni 2020).

Koutra, Danai, Joshua T. Vogelstein und Christos Faloutsos (o.D.). „D<sc>elta</sc>C<sc>on</sc>: A Principled Massive-Graph Similarity Function". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*, S. 162–170. DOI: 10.1137/1.9781611972832.18.

Lance, Godfrey N und William T Williams (1967). „Mixed-Data Classificatory Programs I - Agglomerative Systems". In: *Australian Computer Journal* 1.1, S. 15–20.

Lehr, Andreas (2000). „Kleine Formen: Adornos Kombinationen: Konstellation/Konfiguration, Montage und Essay". Diss. Universität Freiburg. URL: https://freidok.uni-freiburg.de/data/27 (besucht am 1. Juni 2020).

Moretti, Franco (2013). *Distant Reading*. London: Verso.

Neurath, Otto, Rudolf Carnap und Hans Hahn ([1929] 2006). „Wissenschaftliche Weltauffassung. Der Wiener Kreis". In: *Wiener Kreis. Texte zur wissenschnschaftlichen Weltauffassung*. Hrsg. von Michael Stöltzner und Thomas Übel. Hamburg: Meiner, S. 3–29.

Nikolentzos, Giannis, Polykarpos Meladianos und Michalis Vazirgiannis (2017). „Matching Node Embeddings for Graph Similarity". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, S. 2429–2435.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Quine, Willard Van Orman (1974). *Grundzüge der Logik*. Frankfurt am Main: Suhrkamp.

Reiter, Nils (2018). „CorefAnnotator – A New Annotation Tool for Entity References". In: *Abstracts of EADH: Data in the Digital Humanities*. Galway, Ireland. DOI: 10.18419/opus-10144.

Ritter, Henning (2008). „Wenn Adorno spricht". In: *Frankfurter Allgemeine Zeitung*. URL: https://www.faz.net/aktuell/feuilleton/bilder-und-zeiten-1/adornos-stil-wenn-adorno-spricht-1712550.html (besucht am 1. Juni 2020).

Rösiger, Ina (2019). „Computational modelling of coreference and bridging resolution". Diss. Stuttgart University. DOI: 10.18419/opus-10346.

Rosvall, Martin und Carl T. Bergstrom (2008). „Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4, S. 1118–1123. DOI: 10.1073/pnas.0706851105.

Rousseeuw, Peter (1987). „Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis“. In: *J. Comput. Appl. Math.* 20.1, S. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

Sanfeliu, A. und K. Fu (1983). „A distance measure between attributed relational graphs for pattern recognition“. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.3, S. 353–362. DOI: 10.1109/TSMC.1983.6313167.

Schmid, Helmut (1994). „Probabilistic part-of-speech tagging using decision trees“. In: *Proceedings of the conference on New Methods in Language Processing* 12.

Seel, Martin (2019). „‚Philosophie wäre erst zu komponieren‘“. In: *Eros und Erkennntnis. 50 Jahre „Ästhetische Theorie“*. Hrsg. von Martin Endres, Axel Pichler und Claus Zittel. Berlin: De Gruyter, S. 167–171.

Shervashidze, Nino, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn und Karsten M. Borgwardt (2011). „Weisfeiler-Lehman Graph Kernels“. In: *J. Mach. Learn. Res.* 12, S. 2539–2561.

Sonderegger, Ruth (2019). „Essay und System“. In: *Adorno-Handbuch*. Hrsg. von Richard Klein, Johann Kreuzer und Stefan Müller-Doohm. 2. Aufl. Stuttgart: Metzler, S. 534–536.

Zhang, Qi, Meizhu Li und Yong Deng (2018). „Measure the structure similarity of nodes in complex networks based on relative entropy“. In: *Physica A: Statistical Mechanics and its Applications* 491, S. 749–763. DOI: 10.1016/j.physa.2017.09.042.

Gabriel Viehhauser

Zur Erkennung von Raum in narrativen Texten

Spatial frames und Raumsemantik als Modelle für eine digitale Narratologie des Raums

Zusammenfassung: Der Beitrag erprobt, inwieweit das von Ruth Ronen eingeführte Konzept des *spatial frames* für eine digitale Narratologie des Raums fruchtbar gemacht werden kann. Dabei stehen insbesondere zwei Fragestellungen im Fokus: 1. Inwieweit lässt sich vom Auftreten expliziter Raummarker an der Textoberfläche auf dahinterliegende implizite Raummodelle schließen? Sowie 2. Kann das Konzept des *spatial frames* dabei helfen, die Raumsemantik eines Textes (im Sinne Juri Lotmans) mit Hilfe von digitalen Mitteln zu explorieren? Als exemplarischer Testfall wird dazu die Raumkonstellation in einem relativ einfach strukturierten Text, nämlich dem Märchen von Hänsel und Gretel, ausgewertet.

Abstract: This article explores to what extent the concept of spatial frames as introduced by Ruth Ronen can be made fruitful for a digital narratology of space. The focus is on two questions in particular: 1. To what extent can the appearance of explicit spatial markers on the text surface be used to infer implicit spatial models behind them? And 2. can the concept of the spatial frame help to explore the spatial semantics of a text (in the sense of Juri Lotman) by digital means? As an exemplary test case, the spatial constellation in a relatively simply structured text, namely the fairy tale of Hansel and Gretel, will be evaluated.

1 Einleitung

1.1 Raum als narratologische Kategorie

Trotz seiner Bedeutung als Grundkonstituens von Erzählungen hat der Raum in der Narratologie weit weniger Aufmerksamkeit gefunden als andere Phänomene wie etwa Figuren oder Handlungsstrukturen. Dies mag unter anderem daran liegen, dass beim Raum die Kluft zwischen den beiden klassischen Analyseebenen der Narratologie (Genette 1998), der *histoire*-Ebene (was wird erzählt?) und der *dis-*

Gabriel Viehhauser, Abteilung Digital Humanities, Institut für Literaturwissenschaft, Universität Stuttgart

Open Access. © 2020 Gabriel Viehhauser, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-015

cours-Ebene (wie wird erzählt?), besonders groß erscheint: Raum ist zwar immer Voraussetzung einer Erzählung und damit unabdinglich Teil der *histoire*, muss aber nicht immer detailliert auserzählt und damit auf der *discours*-Ebene ausführlich präsentiert werden.

Oftmals ist Raum bei einer Erzählung einfach implizit mitgedacht. Dies geht so weit, dass schon reine Tätigkeit- oder Professionsbezeichnungen implizit auf Raumstrukturen verweisen können, die der Leser oder die Leserin dann in seiner oder ihrer Vorstellung als mentales Modell zusammenfügen. So verweist etwa ein Erzählanfang wie ‚Ich erwachte und blickte der Ärztin ins Gesicht' nicht nur bereits auf einen (wie auch immer gearteten) Schlafplatz, sondern mutmaßlich auch auf ein Krankenzimmer und ein (normalerweise in einer Stadt situiertes) Krankenhaus oder eine ähnliche medizinische Einrichtung, ohne dass der Raum explizit beschrieben worden wäre.[1]

Auch im Vergleich mit der Kategorie der Zeit zeigt sich, dass der Raum besondere Herausforderungen für eine digitale Modellierung mit sich bringt: Erzählungen sind per se linear und zeitlich, während Raum dies gerade nicht ist (Zoran 1984). Zeit vergeht in Erzählungen daher sozusagen auf natürliche Weise, während das Setting einer Geschichte zwar andauernd vorhanden ist, aber gerade deshalb nicht andauernd beschrieben oder konstruiert werden muss; der Raum läuft im Hintergrund der *histoire* immer mit und verschwindet gerade dadurch auf der *discours*-Ebene. Damit ergibt sich ein großer Abstand zwischen dem, was auf der Textoberfläche steht, und dem, was die Leserin oder der Leser als mentales Modell konstruiert. Erschwert diese Kluft schon konventionelle Zugänge zur Narratologie des Raumes erheblich, so stellt sich die Situation beim Versuch einer digitalen Modellierung von Raum in Erzähltexten noch erheblich schwieriger dar: Denn der Computer ist ja zunächst vor allem auf die Auswertung positiver Daten auf der Textoberfläche angewiesen und kann bloß implizierte Informationen nicht ohne Weiteres erkennen.

1 Welche Vorstellungen von der Textwelt dabei genau die Grundlagen dieser Rekonstruktion bilden, lässt sich mit dem *principle of minimal departure* (Ryan 1980) beschreiben: Die Leserin oder der Leser geht davon aus, dass die erzählte Welt der für sie oder ihn als real empfundenen Welt so weit wie möglich gleicht, bis im Text Hinweise darauf gegeben werden, dass die beschriebene Welt von dieser abweicht. Was als ‚normale' Welt angesetzt wird, hängt nicht zuletzt von Genrekonventionen ab, so würde etwa im obigen Beispiel die Rekonstruktion ‚Krankenhaus' vermutlich ausfallen, wenn die Leserin oder der Leser wüsste, dass es sich um den Beginn eines historischen Romans handelt, der in der Antike spielt.

1.2 *Spatial frames*

Die Schwierigkeiten, die die Analyse des Raums in Erzähltexten bereitet, ist in den bislang spärlichen, aber höchst substanziellen Arbeiten zur Narratologie des Raums bereits scharfsinnig beschrieben worden, insbesondere etwa von Ruth Ronen in ihrer Studie „Space in Fiction" (Ronen 1986). In ihrem Aufsatz entwickelt Ronen das Konzept so genannter ‚*spatial frames*', indem sie Raum als semantisches Konstrukt beschreibt, das aus dem im Text vorhandenen sprachlichen Strukturen zusammengesetzt ist. *Spatial frames* lassen sich mithin als eine Art mentaler Aggregationen beschreiben, die die Hinweise im Text auf räumliche Begebenheiten zu übergeordneten Raumentitäten bündeln:

> A frame, as defined here, is a strictly spatial concept, designating the location of various fictional entities. Various expressions in the text construct different types of frames which compose the global structure of the space of a story. It is, of course, possible for a frame to have only one textual manifestation; yet, regardless of the number of expressions manifesting it, a frame, being a constructional concept, exceeds its own linguistic manifestations. (Ronen 1986, S. 421–422)

Ein *spatial frame* kann also explizit benannt bzw. beschrieben, aber auch implizit konstruiert werden und damit mehr sein als das, was aus den bloßen Informationen auf der Textoberfläche für sich genommen ersichtlich wird. Zu Beginn von Goethes *Wilhelm Meister* etwa lässt sich der *spatial frame* von Marianes Haus, das der Schauplatz des ersten Abschnitts ist, nach und nach aus Textinformationen erschließen, ohne dass das Haus zunächst benannt wird:

> Das Schauspiel dauerte sehr lange. Die alte Barbara trat einigemal ans Fenster und horchte, ob die Kutschen nicht rasseln wollten. Sie erwartete Marianen, ihre schöne Gebieterin, die heute im Nachspiele, als junger Officier gekleidet, das Publikum entzückte, mit größerer Ungedult, als sonst, wenn sie ihr nur ein mäßiges Abendessen vorzusetzen hatte (Goethe 1795, Bd. 1, 3)

Obwohl in diesen Anfangssätzen bloß ein Raummarker (das Fenster, näher positioniert durch die Präposition „ans") explizit erwähnt wird, ist die Leserin bzw. der Leser vermutlich schon nach diesem kurzen Textstück in der Lage, aus den dort gegebenen Angaben (Kutschen vor dem Fenster, Barbara ist offensichtlich Bedienstete von Mariane und erwartet sie mit einem Abendessen, was die Vorstellung eines bürgerlichen Haushalts evoziert) ein rudimentäres Setting zu rekonstruieren, dem in der Folge immer mehr Informationen zugeordnet und das zu anderen *frames* in Relation gesetzt werden kann (da Mariane das Abendessen nach der Aufführung erreichen kann, wobei zumindest der letzte Teil des Weges mittels Kutschenfahrt absolviert wird, befindet sich das Haus wohl in relativer

Nähe bzw. sehr wahrscheinlich in derselben Stadt wie das Theater, in dem Mariane spielt; die Leserin oder der Leser wird daher vermutlich das Theater und das Haus als zwei *frames* im übergeordneten *frame* der Stadt konstruieren). Zur Orientierung im räumlichen Setting der Erzählung ist also eine Aggregationsleistung nötig, die die isolierten Informationen zu *spatial frames* zusammenfügt.

2 *Spatial frames* als Muster für eine digitale Modellierung von Raum?

2.1 Ein Beispiel: *Hänsel und Gretel*

Wenngleich Ronens Konzept eine ausgeprägte kognitionswissenschaftliche Komponente aufweist, die im Digitalen schwer nachmodelliert werden kann, könnte es sich aufgrund dieses Aggregationsmechanismus doch dazu eignen, die Kluft zwischen Raummarkern auf der Textoberfläche und zugrunde liegenden Raumentitäten zu überbrücken, und damit als Folie für digitale Zugänge zur Narratologie des Raums dienen. Lässt man die mentale Konstruktion der Semantik von *spatial frames* beiseite, so könnte man diese nämlich gleichsam als latente Konzepte fassen, die im Hintergrund der Textmarker stehen und über die sich diese Konzepte dann wiederum erschließen lassen sollten.

Doch bleibt eine solche Konzeptualisierung auch in der Praxis tragfähig? Um dies zu beantworten, möchte ich im Folgenden anhand eines relativ einfachen und überschaubaren Falls darstellen, wie sich *spatial frames* und Raummarker im Text zueinander verhalten und damit sozusagen experimentell überprüfen, ob der Ansatz sich für eine digitale Auswertung von Raumkonstellationen in Erzählungen als praktikabel erweisen könnte. Mein Beispiel hierfür ist das bekannte grimm'sche Märchen von *Hänsel und Gretel*, hier in der englischen Übersetzung.[2]

Abbildung 1 zeigt eine mögliche Skizze der räumlichen Struktur der Erzählung, die ich basierend auf meiner Lektüre des Märchens erstellt habe. Die Kreise in der Abbildung stellen die *spatial frames* dar, die sich im Fortlauf der Handlung ergeben: Zu Beginn bildet das väterliche Haus von Hänsel und Gretel den zentralen räumlichen *frame*. Da die Familie dort lebensbedrohlichen Hunger leidet und die Eltern nicht mehr wissen, wie sie sich versorgen sollen, beschließen sie des Nachts im elterlichen Schlafzimmer, die Kinder am nächsten Tag in den Wald zu

2 Der Rückgriff auf die englische Version erfolgt aus arbeitstechnischen Gründen, da das Experiment Teil umfassenderer vergleichender Vorstudien zum Thema war. Der digitale Text wurde bezogen von https://www.grimmstories.com/en/grimm_fairy-tales/hansel_and_gretel.

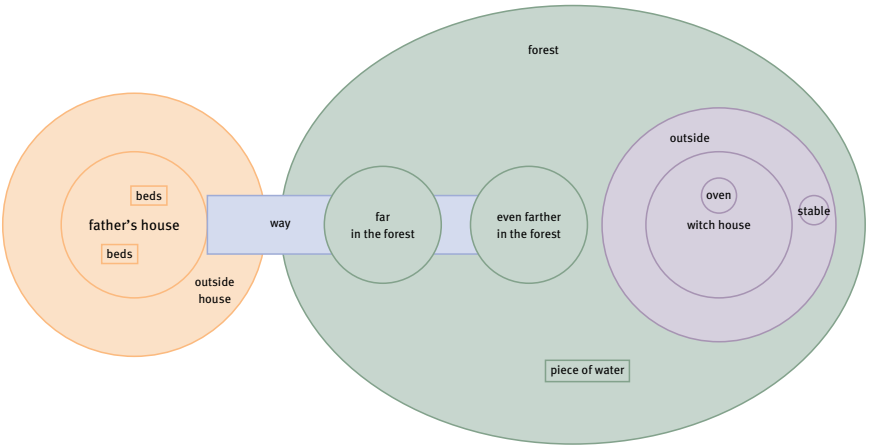


Abb. 1: *Spatial Frames* in *Hänsel und Gretel*. Der orange Kreis repräsentiert den *frame* des elterlichen Hauses, der grüne Kreis den Wald, violett ist der *frame* des Hexenhauses eingezeichnet. In Blau ist der Weg in den Wald dargestellt, der eine Art Übergangsbereich darstellt.

bringen und dort zurückzulassen. Treibende Kraft ist dabei die böse Mutter der Kinder. Doch Hänsel überhört ihr Gespräch und sammelt daraufhin Kieselsteine außerhalb des Hauses auf (äußeres Kreissegment im orangen *frame*-Bereich des Hauses). Am nächsten Tag werden die Geschwister in den Wald gebracht. In der Skizze ist der Wald als großer grüner Kreis rechts dargestellt, innerhalb dieses *frames* ist ein weiterer eingebettet, nämlich ein unbestimmter Ort tief im Wald, an dem die Eltern ihre Kinder zurücklassen. Unter Umständen ließe sich auch der Weg in den Wald als *frame* auffassen, in der Skizze ist er daher als blaues Viereck dargestellt.

Aufgrund der von Hänsel gesammelten und auf dem Weg ausgestreuten Kieselsteine finden die Kinder den Weg zurück und die Geschichte beginnt sozusagen von Neuem und wiederholt sich: Nach einer weiteren Hungerperiode beschließen die Eltern ein zweites Mal, die Kinder auszusetzen, und bringen sie in den Wald, diesmal zu einem Ort noch tiefer im Wald (zweiter kleinerer Kreis im Wald-*frame*). Weil die böse Mutter Hänsels Intentionen durchschaut und ihn daran hindert, Kieselsteine zu sammeln, kann Hänsel diesmal bloß Brotkrumen ausstreuen, die jedoch von Tieren gefressen werden, so dass die Kinder diesmal wirklich im Wald die Orientierung verlieren. Es bleibt ihnen nichts über, als durch die Wälder zu streifen, wo sie schließlich auf das Hexenhaus treffen (violetter Kreis). Dort sperrt die Hexe Hänsel in den Stall, um ihn zu füttern, und versucht, Gretel in den Ofen zu werfen. Doch Gretel dreht den Spieß um und lockt ihrerseits die Hexe in den Ofen und verbrennt sie dort. In einer typisch märchenhaften Wendung

finden die Kinder daraufhin im Haus immense Reichtümer und – aus rationaler Sicht erstaunlicherweise – so gut wie ohne weitere Probleme den Weg zurück nach Hause. Das einzige Hindernis dorthin ist ein Gewässer im Wald (als positionell unbestimmtes Viereck im grünen Wald-Kreis eingetragen), das sie relativ einfach, wenngleich auch auf märchenhaft-magische Art auf dem Rücken einer Ente überqueren können.[3] Danach kehren sie nach Hause zurück, wo die böse Mutter in der Zwischenzeit gestorben ist.

Die Zuordnung der *spatial frames* im Märchen von *Hänsel und Gretel* erscheint relativ einfach. Eventuelle Unklarheiten bestehen lediglich zum einem in Bezug auf ihre Granularität, da manche *frames* in andere eingebettet sind und sich die Frage stellt, ab wann ein Bereich in einer übergeordneten Struktur einen eigenen *frame* bildet (etwa der Weg in den Wald). Zum anderen lässt sich der genaue Punkt, an dem ein *frame* wechselt, nicht immer mit letzter Sicherheit angeben, was insbesondere dann der Fall ist, wenn sich die Erzählung entlang der Bewegungen von Figuren entwickelt. So ließe sich etwa der folgende Satz, mit dem die erste Rückkehr der Kinder aus dem Wald erzählt wird, noch dem *frame* ‚Weg' zuordnen, an seinem Ende ist jedoch bereits das väterliche Haus erreicht: „They walked on the whole night through, and at the break of day they came to their father's house."

Aus dem väterlichen Haus, das zunächst noch als Markierungspunkt im *frame* ‚Weg' fungiert, wird sozusagen auf engstem Raum ein eigener *spatial frame*. Auch solche Übergangspassagen verweisen auf ausfransende Grenzen der *spatial frames* an ihren Rändern bzw. auf die dynamische Funktion von Raummarkern. Lässt man diese Randphänomene jedoch beiseite, so kann die Frage, wo die Erzählung spielt, in jeder Passage des Märchens ziemlich genau beantwortet werden.

2.2 Von Markern zu *frames*?

Um die Brauchbarkeit des Modells für eine digitale Auswertung zu überprüfen, möchte ich als erstes die Frage stellen, ob es eine einfache Möglichkeit gibt, von Raummarkern auf der Textoberfläche zu den Aggregationen der *frames*, die ich in die Skizze eingetragen habe, zurück zu gelangen. Wie groß ist etwa die Möglichkeit einzuschätzen, aufgrund des Auftretens bestimmter Wörter wie ‚Bett' oder ‚Zimmer' an einer Textstelle eine Vorhersage zu treffen, dass sich diese Stelle im *frame* ‚Elterliches Haus' abspielt?

3 Diese Episode begegnet im grimmschen Text noch nicht in der Erstfassung von 1812 (Grimm 1812), sondern ist erst ab der Zweitfassung von 1819 (Grimm 1819) in den Text aufgenommen.

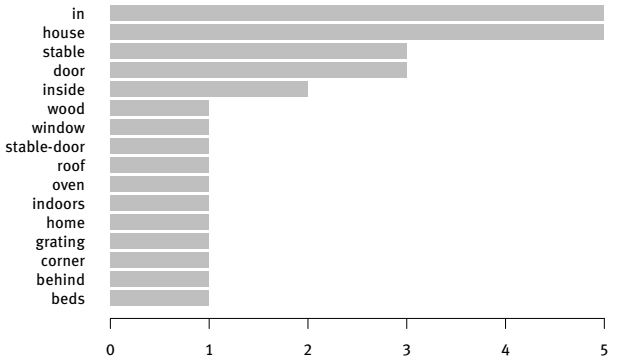


Abb. 2: Häufigste Raummarker in Textpassagen, die im Hexenhaus spielen

Zu diesem Zweck habe ich den Text nach den vier großen *frames* in die Textabschnitte ‚Elterliches Haus', ‚Weg', ‚Wald' und ‚Hexenhaus' aufgeteilt und darin manuell Raummarker annotiert. Unter ‚Raummarkern' wurden dabei im Wesentlichen Objekte mit Lokalisierungsfunktion (z. B. ‚bed', ‚way', ‚forest', ‚fire' etc.) sowie Adverbien (z. B. ‚in', ‚out', ‚inside') verstanden. Die Klassifizierung eines Wortes als Marker erfolgte für die hier skizzierte erste experimentelle Annäherung zunächst nach meiner persönlichen Einschätzung. Deren Stichhaltigkeit müsste für detailliertere Studien freilich in Hinblick auf mögliches *inter-annotator agreement* (Pustejovsky und Stubbs 2013) mehrerer Annotator/inn/en überprüft werden.[4] Im Anschluss habe ich schlicht ausgezählt, welche räumlichen Marker in den vier Passagen erscheinen.

Für manche der *frames* scheinen die Ergebnisse in Hinblick auf die Einsetzbarkeit des Modells ermutigend, zum Beispiel für das Hexenhaus (vgl. hierzu Abbildung 2). In den Passagen, die in diesem *frame* situiert sind, begegnen überwiegend auf das Haus bezogene Marker wie ‚house', ‚door', ‚window' und der ‚stable', in dem Hänsel eingeschlossen wird. Ein Rückschluss von der Textoberfläche auf das zugrunde liegende *frame* scheint unter diesen Voraussetzungen möglich. Andere *frames* hingegen zeigen aber weniger eindeutige Befunde, etwa die Passagen, die im ‚elterlichen Haus' spielen (vgl. Abbildung 3). In ihnen treten am häufigsten die Raummarker ‚wood' und ‚forest' auf, die eigentlich außerhalb des *frames* liegen. Dies lässt sich vor allem darauf zurückführen, dass die Eltern in diesen Passagen häufig davon reden, die Kinder in dem Wald auszusetzen. In der Figu-

4 Die Annotationen wurden mit Hilfe des von Nils Reiter in CRETA entwickelten CorefAnnotators vorgenommen (Reiter 2018). Die annotierten Daten sind unter https://github.com/Gabvie/SpatialFrames zur Verfügung gestellt.

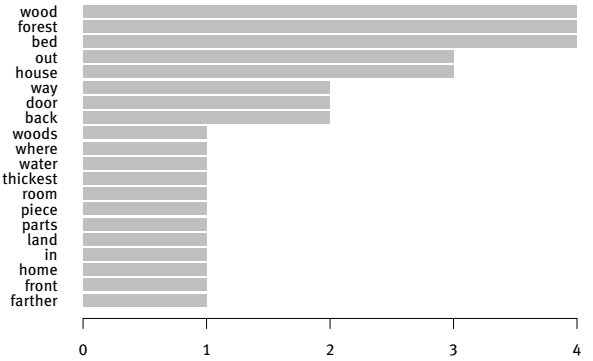


Abb. 3: Häufigste Raummarker in Textpassagen, die im elterlichen Haus spielen

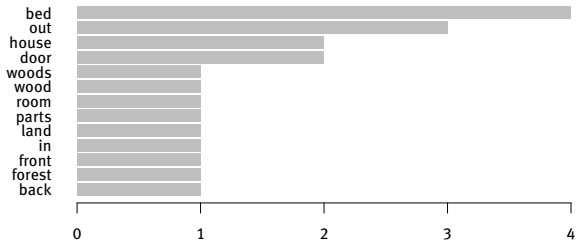


Abb. 4: Häufigste Raummarker in Textpassagen, die im elterlichen Haus spielen, ohne direkte Rede

renrede wird also der räumliche *frame* der folgenden Szene gleichsam proleptisch vorweggenommen.

Es liegt daher nahe, in einem nächsten Schritt direkte Rede aus dem Text zu entfernen. Das Bild wird dadurch tatsächlich etwas deutlicher (vgl. Abbildung 4), ‚bed' ist der häufigste Marker, ‚door' und ‚house' erscheinen ebenfalls mehrfach. Allerdings ist noch immer von ‚woods' und ‚wood' die Rede, auch von ‚out', da die Umgebung des Hauses und der bevorstehende Aufbruch auch in der Erzählrede thematisiert wird. Allgemein sind die Fallzahlen aufgrund der Kürze des Textes natürlich gering. Es wäre daher erst an unterschiedlichen und vor allem auch an längeren Texten zu überprüfen, wie sehr etwa die Annahme generalisierbar ist, dass Passagen mit direkter Rede für vorausdeutende Schilderungen räumlicher Gegebenheiten sozusagen anfälliger sind und daher für deren Analyse eher ausgeschlossen werden sollten. Jedenfalls zeigt aber bereits das Beispiel der räumlich gesehen sicher noch ziemlich einfach gestrickten Geschichte von Hänsel und Gretel, dass sich ein völlig direkter Schluss von der Textoberfläche auf dahinter-

liegende *frames* nicht ohne Weiteres erreichen lässt. Bei komplexeren Texten sind dementsprechend noch größere Verwerfungen zu erwarten.

3 *Spatial frames* als Grundbausteine der Raumsemantik

3.1 Lotmans Raumsemantik

Bei der Analyse räumlicher Konstellationen ist die Frage, an welchem Schauplatz die Szene einer Erzählung spielt, nicht nur bloßer Selbstzweck. Es ist in der Literaturwissenschaft schon lange bemerkt worden, dass sich an narrative Räume auch bestimmte semantische Gehalte anlagern können, die den Raum zum symbolischen Austragungsort der konzeptionellen Spannungen machen können, die eine Erzählung durchziehen. Die prominenteste Ausformulierung hat dieser Gedanke in Juri Lotmans Konzept der Raumsemantik gefunden (Lotman 1977). Im Zentrum von Lotmans Konzept steht die Verbindung der Kategorien von Raum, Handlung und Figuren. Gemäß Lotman bilden Erzählungen nur dann ein sogenanntes *sujet* aus (also so etwas wie eine ereignishafte Handlung), wenn in ihnen eine Figur die Grenzen zwischen zwei semantisch aufgeladenen räumlichen Sphären überschreitet. Das Konzept beruht also auf der Idee, dass Erzählungen mit ihrer räumlichen Struktur ein semantisches Feld aufbauen, das in (zumindest) zwei gegensätzlich semantisierte Räume aufgeteilt ist. Eine klassische räumliche Opposition wäre zum Beispiel jene zwischen der Sphäre der Zivilisation bzw. der Kultur und der Sphäre der Wildnis bzw. der Natur, insbesondere im Märchen. Eine solche Aufteilung repräsentiert die Norm, die normalerweise nicht überschritten werden kann. Nur hervorgehobene Figuren (wie etwa der Held) sind fähig, diese Grenze (auch im buchstäblich räumlichen Sinn) zu überschreiten und damit die ereignishafte Handlung auszulösen. Solche exorbitanten Figuren werden auch als bewegliche Figuren bezeichnet, die sich von den unbeweglichen Figuren abheben, die bloß an einem Ort verbleiben (wie etwa die stets im Wald verharrende Hexe).

Im Rahmen von CRETA haben wir einige Versuche unternommen, Lotmans Konzept für die Beschreibung von Raumkonstellation, aber auch für die Unterscheidung von beweglichen und unbeweglichen Figuren fruchtbar zu machen (Barth und Viehhauser 2017; Viehhauser und Barth 2017; Viehhauser, Kirstein et al. 2018). Dazu haben wir insbesondere bimodale Kollokations-Netzwerkvisualisierungen zum Einsatz gebracht, mithin also Netzwerkdarstellungen, die aus zwei Arten von Knoten gebildet sind, welche durch eine Relation verbunden werden, wenn sie im selben sprachlichen Zusammenhang auftreten. Dabei haben wir Fi-

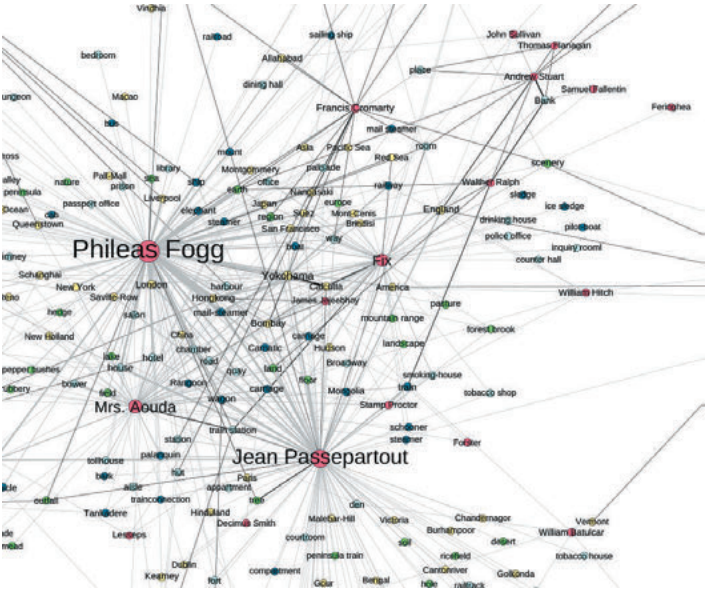


Abb. 5: Bewegliche Figuren in Jules Vernes *Reise um die Erde in 80 Tagen* in einer Netzwerkdarstellung (Ausschnitt)

guren als die eine Art von Knoten repräsentiert, Raummarker als die andere. Relationen bzw. Netzwerkkanten zwischen Figuren und Räumen werden immer dann hergestellt, wenn diese im selben Satz vorkommen. Auf diese Weise lassen sich Netzwerke konstruieren, bei denen der Netzwerk-Grad der einzelnen Figurenknoten (also die Anzahl der Verbindungen zu unterschiedlichen Raummarkern, die eine Figur aufweist) deren Beweglichkeit anzeigt (vgl. Abbildung 5).

Eine Schwierigkeit, die sich dabei gezeigt hat, betrifft nun wieder die Aggregation von Raummarkern: Denn anders als bei Lotman werden in unseren Netzwerkdarstellungen nicht ganze semantische Felder (etwa der Wald) berücksichtigt, sondern nur einzelne Raummarker (also z. B. jede unterschiedliche Benennung eines Details des Waldes). Dadurch wäre es zum Beispiel möglich, dass in einem Text, der einen an sich statischen Raum in vielfältigen Ausdrücken beschreibt, eine Figur öfter mit unterschiedlichen Raummarkern verbunden (und damit beweglicher) erscheint als andere, die sich zwar nicht so lange im selben Raum aufhalten, dafür aber mehrere Räume durchstreifen. Zudem ist für Lotman, anders als in unserer Darstellung, auch nicht jeder beliebige Raumwechsel von

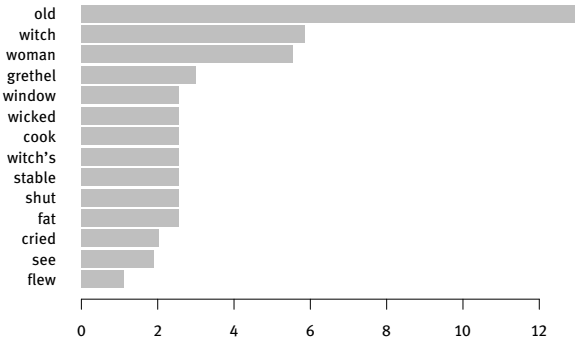


Abb. 6: Distinkte Wörter im *spatial frame* ‚Hexenhaus'

Bedeutung, sondern nur derjenige zwischen semantisch unterschiedlich aufgeladenen Sphären.[5]

3.2 Semantischer Gehalt von *spatial frames*

Auch hier könnte also das Modell der *spatial frames* helfen, einzelne Raummarker zu größeren Konzepten zusammenzuschließen, die dann mit Lotmans semantischen Feldern zusammengedacht werden können. Um die Praktikabilität eines solchen Modells abzuschätzen, habe ich am *Hänsel-und-Gretel*-Text ein weiteres Experiment vorgenommen, das zeigen soll, ob und inwieweit die Passagen der unterschiedlichen *spatial frames* auch unterschiedlich semantisch aufgeladen sind.

Für eine erste Annäherung habe ich den Wortbestand der einzelnen Textpassagen mit dem Distinktivitätsmaß der *likelihood ratio* ausgewertet (Dunning 1993), mit dessen Hilfe überdurchschnittlich häufige Wörter des Abschnitts im Vergleich zum Wortgebrauch der anderen Abschnitte berechnet werden können.[6]

Wie sich zeigt, ist eine rein auf der Textoberfläche basierende Suche nach semantischen Feldern jedoch mit Schwierigkeiten behaftet.[7]

Abbildung 6 zeigt die im Vergleich zu den anderen drei Abschnitten distinktesten Wörter im *spatial frame* ‚Hexenhaus'. Eine klare Opposition etwa entlang

5 „Eine Verschiebung des Helden *innerhalb* des ihm zugewiesenen Raumes ist kein Ereignis" (Lotman 1977, S. 338).

6 Die Berechnung wurde mit Hilfe der Funktion `textstat_keyness()` des Quanteda-Packages in R (Benoit et al. 2018) durchgeführt. Stoppwörter wurden entfernt.

7 Und dies wohl nicht nur aufgrund der methodischen Probleme, die das angewandte Distinktivitätsmaß der *likelihood ratio* mit sich bringt, vgl. hierzu Kilgariff 2001.

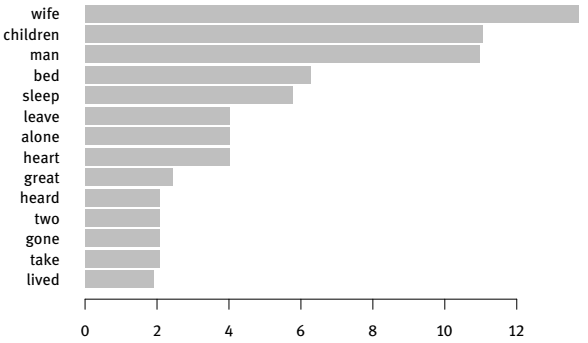


Abb. 7: Distinkte Wörter im *spatial frame* ‚Elternhaus'

der Achse Zivilisation vs. Natur bzw. Anti-Zivilisation lässt sich hier nicht ablesen. Dafür führt die Auswertung allerdings auf andere, erwägenswerte Oppositionspaare, die beim ersten Lesen des Märchens vielleicht gar nicht so klar ins Bewusstsein treten: Die überdurchschnittlich häufige Verwendung von ‚old' verweist darauf, dass das Hexenhaus als Domäne der alten Hexe einen Gegenpol zur Jugendlichkeit von Hänsel und Gretel darstellt. Mit ‚witch', ‚woman' und ‚Grethel' ließe sich der *frame* zudem als Sphäre der Weiblichkeit charakterisieren – und in der Tat ist es auch beim konventionellen Lesen auffällig, dass gerade hier im Hexenhaus Gretel eine besonders aktive Rolle spielt: Während es zuvor vor allem Hänsel war, der etwa mit dem Sammeln der Kieselsteine aktiv auf sein Schicksal einwirkt, und Gretel passiv bleibt, erscheinen hier die Verhältnisse umgekehrt: die Hexe wird geradezu im Alleingang von dem Mädchen überwunden, während Hänsel handlungsunfähig im Stall eingesperrt ist. Der Gegenort des Hexenhauses bringt also auch eine Umkehrung des gewöhnlichen Geschlechterverhältnisses mit sich und dies stellt einen vielleicht gar nicht so unwesentlichen Aspekt der im Märchen ausgestellten Normüberschreitung im Sinne Lotmans dar.

Auch die distinktesten Wörter für die Abschnitte, die im elterlichen Haus spielen (vgl. Abbildung 7), lassen sich nicht klar einer Sphäre der Zivilisation zuordnen: Zwar ist der *frame* durch ‚wife', ‚children' und ‚man' als Domäne der Familie, und durch ‚bed' und ‚sleep' geradezu als Ruhepol ausgewiesen, doch weist das im Ranking folgende Wort ‚alone' schon darauf hin, dass die Sphäre nicht als ungestörter Ausgangspunkt betrachtet werden darf. Überblickt man das ganze Märchen, so wird auch evident, dass eine klare Aufteilung der Erzählung in die klischeehafte Opposition von Zivilisation und Anti-Zivilisation überhaupt nicht greift. Denn gerade das elterliche Haus ist ja durch die Bösartigkeit der Mutter

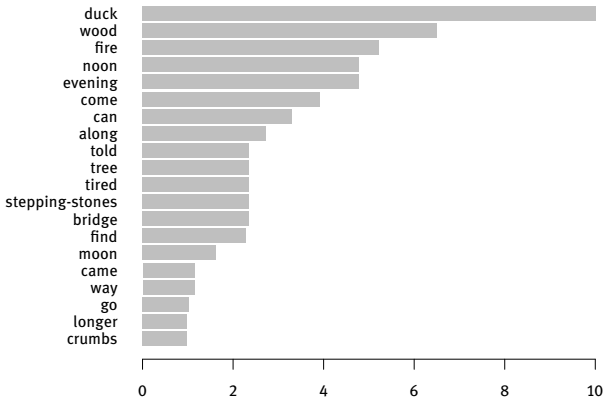


Abb. 8: Distinkte Wörter im *spatial frame* ‚Wald'

schon von Beginn an ein kaum weniger feindlicher Ort als das Hexenhaus im fernen Wald.

Schließlich ließe sich noch darauf hinweisen, dass die Raumstruktur in *Hänsel und Gretel* schon von ihrer grundlegenden Struktur her nicht binär, sondern eher mehrstellig ist. Zwar könnte man die Sphäre des elterlichen Hauses als den einen Pol verstehen, der dem Wald gegenübersteht, doch wird mit dem Hexenhaus eine dritte Komponente eingeführt, die ebenfalls als Gegenpol zum elterlichen Haus konstruiert ist und in einem unklaren Verhältnis zum Wald bleibt: Denn zum einem ist es Teil des Waldes und damit der Gegenwelt, zum anderen hebt es sich aber durch diverse zivilisatorische Einrichtungen wie Ofen und Stall auch wieder von der völligen Unwirtlichkeit des Waldes ab. Eine alternative Möglichkeit wäre es, den Wald, zumindest soweit er von den ja eigentlich unbeweglichen Figuren der Eltern betreten werden kann, als eine Art Grenzbereich anzusehen, der erst in Richtung auf die Märchenwelt überschritten wird. Doch wäre gerade in diesem Fall die Topographie nicht mehr trennscharf auf die semantischen Sphären abbildbar, da der Wald sozusagen unmerklich in die Märchenwelt übergeht.[8]

8 Lotman selbst hat bereits in der Grundkonzeption seines Raummodells durchaus mit der Möglichkeit solcher mehrstufigen Gliederungen gerechnet, die das binäre Grundmuster seines Konzepts weiter differenzieren. Neben dem „grundlegende[n] und wichtigste[n]" Fall, bei dem „der Raum des Textes von einer Grenze in zwei Teile geteilt wird und jede Figur zu einem dieser Teile gehört", seien durchaus auch „kompliziertere Fälle möglich", (Lotman 1977, S. 328). Die strenge Dichotomie der Raumsemantik wird schließlich mit dem Konzept der *Semiosphäre* weiter aufgebrochen, das Lotman in einem späteren Aufsatz entwickelt. Im Konzept der Semiosphäre wird

Die häufigsten Begriffe in den Wald-Abschnitten selbst sind wenig aussagekräftig und könnten tatsächlich am ehesten die These vom Mischcharakter des Waldes stützen (vgl. Abbildung 8): In diesem Bereich wird nämlich etwas überraschend die Ente als distinktivstes Wort gegenüber den anderen Textteilen ausgewiesen. Die Ente tritt nur gegen Ende der Erzählung, nach dem schon überstandenen Abenteuer im Hexenhaus, auf. Die Hervorhebung dieses Wortes ist sicherlich auch der angewandten *likelihood-ratio*-Testmethode geschuldet, bei der in begrenzten Abschnitten häufig vorkommende Wörter überbewertet werden. Sie verweist aber auch auf eine genauere Fassung von Lotmans Konzept: Denn eigentlich ist der Wald, den Hänsel und Gretel nach der Tötung der Hexe betreten, ein ganz anderer als zuvor. Nach der Grenz- und damit Normüberschreitung durch die ‚Helden' Hänsel und Gretel löst sich das Sujet der Geschichte in dem Moment auf, in dem die Hexe überwunden wird, wodurch auch der semantische Gegensatz der Sphären zusammenbricht. Der Weg durch den Wald kann daher unproblematisch wieder gefunden werden, die einstige Grenze zwischen den Sphären hat ihre Bedeutung verloren. Dennoch, und hier ergibt sich wieder eine Abweichung von Lotmans Grundmodell, bleibt der Wald durch die Enten-Episode weiter märchenhaft aufgeladen. Fast scheint es, als wäre es geradezu die Funktion dieser zunächst überflüssig erscheinenden Szene, dafür zu sorgen, dass der Märchencharakter des Settings weiter aufrecht erhalten bleibt. Methodisch ergibt sich aus dem Befund jedenfalls die Konsequenz, dass damit zu rechnen ist, dass sich die semantische Aufladung von Sphären im Verlauf der Erzählung ändern kann. Da also Verwerfungen zwischen Topographie und Semantik möglich sind, kann sich auch hier der Schluss von der Textoberfläche auf zugrundeliegende Konzepte als trügerisch erweisen.

4 Fazit

Im vorliegenden Beitrag sollte anhand eines relativ leicht zu überblickenden Beispielfalls, nämlich des Märchens von Hänsel und Gretel, erprobt werden, ob sich das in der Narratologie entwickelte Modell der *spatial frames* als Grundlage für eine digitale Analyse von Raum in Erzählungen eignen kann.

Obwohl die hier durchgeführten Experimente einige konzeptionelle Probleme aufgezeigt haben, erscheint mir eine weiterführende Beschäftigung mit dem

---

die Möglichkeit eines diskontinuierlichen, von Grenzbereichen umgebenen Raums eingeräumt, vgl. Lotman 1990.

Konzept sinnvoll. Als künftige Forschungsaufgaben ergeben sich dabei folgende Punkte:

Der Schluß von Raummarkern auf der Textoberfläche auf dahinterliegende strukturelle Konzepte ist nicht trivial, es müssen daher entsprechende Instrumentarien entwickelt werden. Insbesondere dürfte es dabei von Bedeutung sein, Raummarker, die nicht auf das zum Zeitpunkt der Erzählung aktuelle Setting verweisen, von solchen zu trennen, die aussagekräftig sind.

Die Semantisierung von räumlichen Sphären ist mit digitalen Mitteln schwer nachzuzeichnen. Dies ergibt sich aus der nicht vollständigen Explizitheit von Texten (gerade bei Märchen wie *Hänsel und Gretel* ist eine gewisse Nüchternheit des Erzählstils durchaus auffällig, die eben nicht alles deutlich ausspricht), darüber hinaus aber auch daraus, dass die Semantisierung der Räume in sehr individuellen Formen sich ausdrücken kann. Gerade weil es bei literarischen Texten schwer möglich ist, über größere Datenmengen zu aggregieren, wird das Bild der digitalen Analyse durch Idiosynkrasien sehr beeinträchtig. Es wäre zu überlegen, ob es räumliche Sphären gibt, die so allgemein sind (der Wald z. B.), dass man Semantisierungen auch über große Textmengen nachvollziehen kann.

Allerdings, so hat sich etwa bei der Analyse der Sphäre des Hexenhauses gezeigt, ließen sich durch engmaschige Betrachtungen eventuell auch überraschende Entdeckungen von Strukturmustern machen, die beim ‚konventionellen' Lesen nicht sofort ins Auge springen bzw. durch die Strukturanalyse deutlicher hervortreten (z. B. die stark weibliche und damit gegenweltliche Prägung der Hexenhaus-Szene). Inwieweit solche ‚Entdeckungen' systematisiert werden können, bleibt allerdings eine offene Frage.

Danksagung: Mein Überlegungen zum Raum sind durch die interdisziplinäre Zusammenarbeit deutlich befördert worden. Insbesondere Florian Barth, Andreas Pairamidis und Roman Klinger haben im CRETA-Projekt die Arbeit an einer digitalen Narratologie des Raums vorangebracht. Ich danke zudem meinen Korrekturlesern Fabian Mauch und Nils Reiter für wertvolle Hinweise.

Primärliteratur

Goethe, Johann Wolfgang von (1795). *Wilhelm Meisters Lehrjahre*. Berlin: Unger.
Grimm, Brüder (1812). *Kinder- und Hausmärchen*. 1. Auflage. Berlin: Realschulbuchhandlung.
Grimm, Brüder (1819). *Kinder- und Hausmärchen*. 2. Auflage. München: G. Reimer.
Hansel and Gretel. A fairy tale by the brothers Grimm (2020). URL: https://www.grimmstories.com/en/grimm_fairy-tales/hansel_and_gretel (besucht am 24. März 2020).

Sekundärliteratur

Barth, Florian und Gabriel Viehhauser (2017). „Digitale Modellierung literarischen Raums“. In: *Abstracts der DHd: Digitale Nachhaltigkeit*. Bern: Digital Humanities im deutschsprachigen Raum e.V., S. 128–132.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller und Akitaka Matsuo (2018). „quanteda: An R package for the quantitative analysis of textual data“. In: *Journal of Open Source Software* 3.30, S. 774. DOI: 10.21105/joss.00774. (Besucht am 1. Juni 2020).

Dunning, Ted (1993). „Accurate Methods for the Statistics of Surprise and Coincidence“. In: *Computational Linguistics* 19, S. 61–74.

Genette, Gérard (1998). *Die Erzählung*. 2. Auflage. München: Fink.

Kilgariff, Adam (2001). „Comparing Corpora“. In: *International Journal of Corpus Linguistics* 6:1, S. 97–133.

Lotman, Jurij M. (1977). *Die Struktur literarischer Texte*. München: Fink.

Lotman, Jurij M. (1990). „Über die Semiosphäre“. In: *Zeitschrift für Semiotik* 12.4, S. 287–305.

Pustejovsky, James und Amber Stubbs (2013). *Natural Language Annotation for Machine Learning*. Sebastopol: O’Reilly.

Reiter, Nils (2018). „CorefAnnotator – A New Annotation Tool for Entity References“. In: *Abstracts of EADH: Data in the Digital Humanities*, S. 128–132.

Ronen, Ruth (1986). „Space in Fiction“. In: *Poetics Today* 7.3, S. 421–438.

Ryan, Marie-Laure (1980). „Fiction, Non-Factuals, and the Principle of Minimal Departure“. In: *Poetics* 9, S. 403–422.

Viehhauser, Gabriel und Florian Barth (2017). „Towards a Digital Narratology of Space“. In: *Digital Humanities 2017: Conference Abstracts*. Montreal. URL: https://dh2017.adho.org/abstracts/413/413.pdf (besucht am 1. Juni 2020).

Viehhauser, Gabriel, Robert Kirstein, Florian Barth und Andreas Pairamidis (2018). „Cadmus and the Cow: A Digital Narratology of Space in Ovid’s Metamorphoses“. In: *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*. Cham: Springer, S. 293–301.

Zoran, Gabriel (1984). „Towards a Theory of Space in Narrative“. In: *Poetics Today* 5(2), S. 309–335.

Marcus Willand, Evelyn Gius und Nils Reiter

SANTA: Idee und Durchführung

Zusammenfassung: In diesem Kapitel werden die grundsätzlichen Überlegungen, das Evaluationsschema sowie die Ergebnisse des ersten *shared tasks* in den und für die Digital Humanities vorgestellt. Der *shared task* hat die Erstellung von Annotationsrichtlinien für Erzählebenen zum Ziel und fand im Jahr 2018 mit acht teilnehmenden Teams statt. Das Evaluationsschema, das bei *shared tasks* eine zentrale Rolle einnimmt, wurde spezifisch für diesen *shared task* entwickelt und kombiniert die Dimensionen ‚Theorieabdeckung', ‚Anwendbarkeit' und ‚Nützlichkeit' mit der Messung von *inter-annotator agreement* und konnte herausarbeiten, dass es möglich ist, komplexe geisteswissenschaftliche Konzepte durch Annotationsrichtlinien intersubjektiv anwendbar zu machen. Die Gewinner-Richtlinie des *shared tasks* zeichnet sich dadurch aus, dass sie eine gute Balance zwischen den verschiedenen Zielen und Anforderungen findet.

Abstract: This chapter presents the basic considerations, the evaluation scheme and the results of the first *shared task* in and for digital humanities. The *shared task* aims to create annotation guidelines for narrative levels and has been able to attract eight participating teams in 2018. The evaluation scheme combines the dimensions 'conceptual coverage', 'applicability' and 'usefulness' with the measurement of inter-annotator agreement and was able to show that it is possible to make complex humanities concepts intersubjectively applicable through annotation guidelines. The winning guideline of the shared task is characterised by finding a good balance between the different goals and requirements.

1 Einleitung

In diesem Beitrag wird mit dem *shared task* SANTA (*Systematic Analysis of Narrative Texts through Annotation*) eine Aktivität vorgestellt, in die viele CRETA-Ideen eingeflossen sind und an der auch eine Reihe von CRETA-Mitgliedern teilgenommen hat. Ziel war es, die Methodenentwicklung für die automatische Erkennung

Anmerkung: Das Kapitel ist eine deutsche Übersetzung und Überarbeitung von: Evelyn Gius, Nils Reiter und Marcus Willand, Hrsg. (2019). *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*.

Marcus Willand, Germanistisches Seminar, Universität Heidelberg
Evelyn Gius, Institut für Sprach- und Literaturwissenschaft, TU Darmstadt
Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Marcus Willand, Evelyn Gius und Nils Reiter; publiziert von De Gruyter BY-NC-ND
Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-016

von Erzählebenen mittels eines *shared task* anzustoßen. Gleichzeitig soll der Annotationsprozess genutzt werden, um Erkenntnisse über die annotierten Phänomene zu sammeln und in die theoretische Ebene zurückzuspiegeln.

Da bisher für Erzählebenen keine nutzbaren Annotationen vorliegen, müssen diese Annotationen zunächst erstellt werden, wozu es wiederum Annotationsrichtlinien braucht.[1] Das Erstellen von Annotationsrichtlinien steht also zunächst im Fokus, wobei hierfür das Konzept eines *shared task* in Abwandlung verwendet wurde.

2 Motivation

Mit diesem *shared task* werden zwei für die Digital Humanities und computationelle Literaturwissenschaft zentrale Herausforderungen angegangen: Die Arbeitsteilung in Digital-Humanities-(DH)-Projekten inklusive der Schnittstellen zwischen den Arbeitsbereichen und die intersubjektive manuelle und robuste automatische Erkennung von Erzählebenen in narrativen Texten.

2.1 Aufteilung von Arbeit, Kompetenzen und Teilaufgaben

Beim derzeitigen Stand der automatischen Analyse von narrativen Texten müssen Digital-Humanities-Projekte, die inhaltsbezogene Aspekte solcher Texte in großem Maßstab analysieren wollen, technisch-methodischen Fortschritt erzielen, da viele relevante Phänomene mit bestehenden Verfahren noch nicht – oder nur unzuverlässig – gefunden werden können. Viele solcher Projekte sind also Kooperationsprojekte zwischen Forscher*innen aus den Computerwissenschaften (inklusive der Computerlinguistik) und den Literatur- oder Kulturwissenschaften. Auch wenn die Zahl der Tutorials, angebotenen Sommerschulen und Handbücher in den letzten Jahren massiv gestiegen ist, bleibt die tägliche, konkrete Organisation solcher Projekte eine Herausforderung:[2]

Eine der ersten Aufgaben, die ein interdisziplinäres Digital-Humanities-Projekt zu lösen hat, ist die Entwicklung einer gemeinsamen Sprache und eines gemeinsamen Verständnisses für den Untersuchungsgegenstand. Denn manch-

1 Vgl. die Beiträge von Pagel et al. (2020), Pichler und Reiter (2020) und Reiter (2020) in diesem Band.

2 Einige davon werden auch im vorliegenden Band explizit diskutiert, etwa im Beitrag von Reiter et al. (2020) ab Seite 467.

mal interessieren sich die beteiligten Computerwissenschaftler*innen ausschließlich für den methodischen Anteil, ohne die erzielten Ergebnisse in irgendeiner Art zu interpretieren. Geisteswissenschaftler*innen fokussieren hingegen viel stärker auf konzeptuelle oder begriffliche Fragen oder auf die Interpretation der Resultate, ohne die Methode explizit zu reflektieren. Innerhalb eines Projektes verfolgen die Beteiligten also oft unterschiedliche Ziele.

Wir sind überzeugt, dass *shared tasks* insbesondere in interdisziplinären Bereichen wie den Digital Humanities sehr viel Potenzial haben, eine gute Schnittstelle und Mittlerposition zwischen den unterschiedlichen Interessen einzunehmen. In einem *shared task* können sich die Beteiligten auf das konzentrieren, was sie am besten können. Literaturwissenschaftler*innen konzentrieren sich auf das literarische Phänomen, das sie beschäftigt und mit dem sie sich auskennen.[3] Durch ihre breite Textkenntnis und Interpetationsexpertise sind sie bestens dafür qualifiziert, das Phänomen zu definieren und zu exemplifizieren, ohne sich über Implementierungsdetails oder überhaupt die Implementierbarkeit eines Finde-Tools Gedanken machen zu müssen. Das Phänomen kann dann so komplex und kontextabhängig definiert sein, wie es nötig ist – die Frage der Möglichkeiten und Grenzen potenzieller Implementierungen kann hingegen bei der Konzeptarbeit ignoriert werden. Trotzdem können die Definitionen nicht beliebig komplex werden, da sie immer noch intersubjektiv anwendbar sein müssen – allerdings von Menschen und nicht von Computern. Sind die Konzepte so operationalisiert, und wurde ein Korpus mit dem Phänomen annotiert, kann jede*r Computerwissenschaftler*in an der Entwicklung von Methoden für die Erkennung des Phänomens arbeiten. Dazu sind dann keinerlei Kenntnisse in Literatur oder narrativer Texte nötig, da eine für die Implementierung geeignete ‚Wahrheit' in Form von Annotationen existiert, gegen die automatische Erkenner evaluiert werden können. Der Komplexität oder Transparenz der Verfahren (z. B. des maschinellen Lernens) sind dann keine Grenzen gesetzt: Da ein vertrauenswürdiger Goldstandard existiert, können alle Verfahren an ihm getestet und überprüft werden. Die in Digital-Humanities-Zugängen normalerweise wichtige Abwägung zwischen Vorhersagekraft und Transparenz/Interpretierbarkeit wäre dann nicht mehr zentral, da die Werkzeuge zuverlässig empirisch getestet werden können. Es können also zur Erkennung die Methoden an den Start gebracht werden, die nötig sind, ohne dass auf die Transparenz oder Erklärbarkeit Rücksicht genommen werden muss.

3 Neben den in diesem Beitrag behandelten Erzählebenen können das z. B. sein: Fokalisierung, Aspekte der Geschichte wie Zeitverlauf, Textelemente wie ‚Exposition' oder ‚Katastrophe', aber auch interpretative Zuschreibungen, etwa zur Identifikation zentraler Sätze eines Textes.

Die beiden von uns organisierten *shared tasks* decken diese beiden Seiten eines Digital-Humanities-Analyseprojekts ab und haben daher auch unterschiedliche primäre Zielgruppen. Der erste *shared task*, der die Entwicklung von Annotationsrichtlinien zum Ziel hat, wird dabei den Grundstein legen, so dass im zweiten *shared task* eine (annotierte) Textbasis für eine unabhängige und fachwissenschaftlich valide Evaluation der Ergebnisse zur Verfügung steht. Ein Modell, das im zweiten Task gut abschneidet, kann dann verlässlich für neue Texte eingesetzt werden (sofern die Texte den Testdaten ähnlich genug sind).

Durch die Durchführung von zwei *shared tasks* ist die technische von der konzeptuellen Arbeit entkoppelt. So können auch Forscher*innen gewissermaßen miteinander kooperieren, ohne im gleichen Projekt zu sein. Damit werden nicht zuletzt auch die Eintrittshürden in das Feld der Digital Humanities gesenkt.

2.2 Annotationsrichtlinien für Erzählebenen

Das im beschriebenen Vorhaben behandelte Phänomen ist das der Erzählebene. Die Identifikation von Erzählebenen oder zumindest von kohärenten Textsegmenten ist ein notwendiger Schritt für jede Analyse narrativer Texte, die sich nicht nur auf die sprachliche Oberflächenform bezieht. Damit sind z. B. Analysen der Handlung, der Figuren oder der erzählten Welt gemeint. Zum Beispiel muss einer Untersuchung der Interaktionen von Figuren die Erkennung von Erzählebenen vorausgehen – sonst würden etwa Figuren als interagierend gezählt, die einander zwar textlich nah sind, sich aber auf der Handlungsebene nie begegnen. Eine quantitative Untersuchung von solchen Interaktionen ist damit auf die automatische Erkennung von Ebenen angewiesen, die eine Schlüsselkomponente in der computationellen Literaturwissenschaft sind bzw. sein sollten.

Daneben können Erzählebenen auch als eine Art Mediator zwischen einer hermeneutischen und einer automatischen Textanalyse fungieren. Denn ihre Komplexität ist aus literaturwissenschaftlicher Sicht vergleichsweise gering, aus Sicht der natürlichen Sprachverarbeitung hingegen vergleichsweise hoch. Außerdem sind sie potenziell für Textanalysen aller Art relevant. Darüber hinaus sind Erzählebenen im Vergleich zu anderen Phänomenen ein in der Literaturwissenschaft eher wenig umstrittenes Phänomen, schließlich basieren die Definitionen von Erzählebenen in der Regel auf textuellen und narrativen Merkmalen. So können beispielsweise *verba dicendi*, also Verben der Äußerung, und die anschließende direkte Rede textuelle Signale für narrative Ebenen sein oder das Vorhandensein einer anderen ‚Geschichtenwelt' kann durch die Analyse des Raumes oder anderer narrativer Phänomene identifiziert werden. Narrative Ebenen sind nicht zuletzt für die Analyse von Texten nützlich, die eine Abweichung zwischen ihrer

textlichen und inhaltlichen Struktur aufweisen, also etwa zwischen Erzähl- und Handlungsreihenfolge.

Erzählebenen sind also eine gute Wahl für ein Phänomen des ersten *shared task* in den Digital Humanities. Ihre wichtigste Eigenschaft ist, dass sie eine Brücke zwischen den theoretischen Diskussionen innerhalb der Narratologie und Anwendungen im *text mining* schlagen können. Moderne automatische Textverarbeitung basiert zu einem großen Teil auf maschinellen Lernverfahren. Angesichts der Verschränktheit von Erzählebenen mit anderen Phänomenen sowie mit Eigenschaften des textuellen Kontexts sind maschinelle Lernverfahren der naheliegendste Ansatz zur automatischen Erkennung von Erzählebenen. Die Qualität dieser Verfahren hängt allerdings von der Verfügbarkeit ausreichender Trainings- und Testdaten ab, die annotiert – also mit Markierungen für Wechsel zwischen Erzählebenen versehen – sind. Sind solche Daten verfügbar, können Modelle trainiert werden, die anschließend zur Erkennung von Ebenen in neuen, nicht annotierten Texten verwendet werden können.

Um diese Annotationen für die Implementierung der Erkennung zu erzeugen, sind Annotationsrichtlinien nötig.[4] Die Richtlinien stellen nicht nur die Kohärenz der Annotationen sicher, sondern definieren auch, wie mit unklaren Fällen umgegangen werden soll, und erlauben es Nicht-Expert*innen, an der Annotation mitzuwirken. Da Annotationsprozesse teuer und zeitaufwändig sind, ist es allerdings wenig realistisch, für die vielen Varianten eines Konzeptes annotierte Korpora zu erstellen. Daher ist es sinnvoll, einen gewissen Konsens in Bezug auf das theoretische Konzept herzustellen, bevor die Annotationsrichtlinien erstellt werden. Dieser Konsens muss sich dabei nicht einem einzelnen theoretischen Ansatz verschreiben – idealerweise wird versucht, möglichst unterschiedliche Ansätze zu berücksichtigen, so dass verschiedene Anschlussfragen auf Basis des implementierten Konzepts möglich sind. Während die Automatisierbarkeit bei der Entwicklung der Richtlinien keine Rolle spielen sollte, ist die Annotierbarkeit ein wichtiges, aber nicht das einzige Kriterium (s. u.).

Damit die in den Annotationsrichtlinien definierten Konzepte den literaturwissenschaftlichen Ansprüchen an die Ergebnisse der Automatisierung genügen, ist es also ideal, wenn Expert*innen aus der Narratologie sich bereits bei der Entwicklung von Richtlinien einbringen, etwa im Rahmen eines *shared task*.

4 Siehe hierzu auch die Beiträge von Reiter (2020) und Ketschik, Blessing et al. (2020) im vorliegenden Band.

3 Der *Shared Task*

3.1 *Shared tasks* in der Computerlinguistik

Shared tasks sind ein etabliertes Forschungsparadigma in der Computerlinguistik. Kernidee eines *shared task* ist, dass verschiedene Teilnehmende an der gleichen, von einem Organisationsteam gestellten Aufgabe arbeiten, also der automatischen Erkennung eines Textphänomens wie z. B. Wortarten. Die verschiedenen Lösungen, d. h. die entwickelten Systeme, werden dann *auf dem gleichen Datensatz* und *mit der gleichen Evaluationsmetrik* evaluiert und sind damit direkt vergleichbar.

Typischerweise läuft ein *shared task* folgendermaßen ab: Das Organisationsteam veröffentlicht einen *call for participation*, in dem die Aufgabe und die bereitgestellten Daten beschrieben werden. Zeitgleich oder kurz danach wird ein annotierter Datensatz zur Verfügung gestellt, den die Teilnehmenden zur Entwicklung ihrer Systeme nutzen können. Der Datensatz ist annotiert, damit die Teilnehmenden ihre Systeme trainieren und intern testen können. Nach einigen Monaten werden die Testdaten publiziert, wobei es sich um einen nicht annotierten Datensatz handelt. Die Teilnehmenden wenden ihre bis dahin entwickelten Systeme auf diesen Datensatz an und schicken die Vorhersagen ihres Systems an das Organisationsteam, das sie dann gesammelt vergleicht und in eine Rangfolge bringt. Am Ende des *shared task* steht meistens ein Workshop, in dem die Systeme diskutiert und mögliche nächste Schritte identifiziert werden.

Historische Entwicklung

Innerhalb der Computerlinguistik haben *shared tasks* ihren Ursprung in den *message understanding conferences* (MUC Sundheim 1993). Ziel dabei war, automatisch Informationen aus Ausschnitten von Zeitungsberichten zu extrahieren. Sundheim und Chinchor (1993) kategorisieren die Fortschritte durch *shared tasks* wie folgt: (i) Als Fortschrittsmessung beschreiben sie, wie gut verschiedene Systeme eine Aufgabe lösen, wenn standardisierte Performanzmetriken angewendet werden. Dadurch wird der aktuelle Forschungsstand direkt ermittelt und transparent, also das System identifiziert, welches zu einem bestimmten Zeitpunkt die besten Ergebnisse liefert. (ii) In der Frage, welche Metrik eigentlich geeignet ist, um die Performanz zu messen, haben *shared tasks* ebenfalls weitergeholfen. Dadurch, dass verschiedene Metriken angewendet und Teil des wissenschaftlichen Diskurses wurden, konnten Stärken und Schwächen verschiedener Metriken offengelegt werden. (iii) Nicht zuletzt generieren die *MUC-shared-tasks* auch Erkenntnisse darüber, warum welche Systeme besser oder schlechter abschneiden.

Durch die Teilnahme und die anschließende Analyse können z. B. bestimmte Lücken identifiziert werden, die ein System nicht schließen kann. Alle drei Aspekte sind in allen computerlinguistischen *shared tasks* mehr oder weniger explizit präsent.

Seit dem Jahr 2000 bietet die *Conference on Natural Language Learning* (CoNLL) ein Dach für eine Reihe von *shared tasks* zu verschiedenen computerlinguistischen Themen: *chunking* (Sang und Buchholz 2000), *clause identification* (Sang und Déjean 2001), sprachunabhängige Erkennung von Eigennamen (Tjong Kim Sang und De Meulder 2003), verschiedene Varianten syntaktischer Analyse, entweder eingeschränkt auf bestimmte Sprachen oder sprach-agnostisch (Buchholz und Marsi 2006; Nivre et al. 2007; Kübler 2008) sowie *semantic role labeling* (Carreras und Màrquez 2004; Carreras und Màrquez 2005). Andere Konferenzen haben das Konzept ebenfalls aufgenommen, z. B. für die Aufgabe, textuelle Inferenzen zu identifizieren (*recognizing textual entailment*, Dagan et al. 2006), die bis 2013 lief. Unter der Bezeichnung ‚SensEval' startete im Jahr 2000 eine *shared-task*-Initiative (Kilgarriff und Rosenzweig 2000). Diese wird aber mittlerweile etwas allgemeiner ‚SemEval' genannt um anzuzeigen, dass man sich generell semantischen Phänomenen widmet. ‚SemEval' ist heute vor allem ein Dach für die Organisation von *shared tasks*, unter dem allein im Jahr 2018 zwölf verschiedene tasks organisiert wurden.[5]

Als Gründe für die anhaltende Beliebtheit von *shared tasks* in der Computerlinguistik nennen Escartín et al. (2017) neben dem direkten Wettbewerb die Möglichkeit, Systeme direkt vergleichen zu können. Nicht zu unterschätzen ist auch, dass die *shared tasks* eine Reihe von De-Facto-Standards etabliert haben (z. B. das breit verwendete CoNLL-Format für annotierte Daten). Auch die Publikation annotierter Daten, die im Rahmen von *shared tasks* entstanden, haben die Forschungsgemeinschaft vorangebracht, da diese nachnutzbar sind. Direkter Wettbewerb und hohe Sichtbarkeit fördern jedoch auch unethisches Verhalten, wie Escartín et al. (2017) ausführen: Forschungsdaten werden womöglich zurückgehalten oder Teilnehmende sind weniger offen, Details zu ihren Arbeiten zu teilen, um anderen voraus zu sein. Escartín et al. geben auch eine Reihe von Empfehlungen für *shared tasks*, um das Risiko unethischen Verhaltens zu minimieren,

5 Affect in Tweets, Multilingual Emoji Prediction, Irony Detection in English Tweets, Character Identification on Multiparty Dialogues, Counting Events and Participants within Highly Ambiguous Data covering a very long tail, Parsing Time Normalizations, Semantic Relation Extraction and Classification in Scientific Papers, Semantic Extraction from CybersecUrity REports using Natural Language Processing (SecureNLP), Hypernym Discovery, Capturing Discriminative Attributes, Machine Comprehension using Commonsense Knowledge, Argument Reasoning Comprehension Task. Siehe auch: http://alt.qcri.org/semeval2018/index.php?id=tasks.

etwa dass explizit von Anfang an geklärt wird, dass die Ergebnisse der teilnehmenden Systeme unter bestimmten Lizenzen zu veröffentlichen sind.

3.2 Zwei verbundene *shared tasks*

Da sich die Forschungscommunities, -praktiken und auch -ziele in der Computerlinguistik und den Literaturwissenschaften stark unterscheiden, konnte das computerlinguistische Modell für *shared tasks* nicht direkt für unser Vorhaben übernommen werden. Wir haben entsprechend mehrere Anpassungen am Verfahren vorgenommen, und es in zwei *shared tasks* aufgeteilt. Die beiden *shared tasks* haben unterschiedliche Ziele, Daten und Zielgruppen, beschäftigen sich aber beide mit dem Phänomen der Erzählebene. Das Ziel des ersten *shared task* ist es, Annotationsrichtlinien für Erzählebenen zu etablieren. Diese werden dann auf ein größeres Korpus angewendet, das im zweiten *shared task* Verwendung als Trainings- und Testkorpus findet. In diesem besteht dann das Ziel in der automatischen Erkennung von Erzählebenen. Damit ist der zweite *shared task* ein ‚normaler' *shared task*, der so abläuft, wie oben beschrieben. Im Folgenden wird ein Meilenstein des ersten *shared task* dokumentiert: Die Teilnehmenden haben ihre Annotationsrichtlinien eingereicht und diese wurden in einem dreitägigen Workshop vergleichend evaluiert.

3.2.1 *Shared Task* 1: Systematic Analysis of Narrative Texts through Annotation (SANTA)

Der erste *shared task* stellt die Herausforderungen bei der Konzeptualisierung und Definition von Erzählebenen ins Zentrum, wobei auch die manuelle Anwendung in Texten Berücksichtigung findet. Die Aufgabe für die teilnehmenden Teams war es, Annotationsrichtlinien für das Phänomen der Erzählebene zu entwickeln. Dafür wurden keine theoretischen Vorgaben gemacht. Allerdings wurde eine Liste möglicher Referenzen als Startpunkt für eigene Recherchen bereitgestellt, nebst einer Anleitung zur Entwicklung von Richtlinien. Eine von Reiter (2020) weiter ausgearbeitete Fassung davon findet sich ab Seite 193 in diesem Band.

Damit die entwickelten Richtlinien sich nicht auf zu korpusspezifische Merkmale konzentrieren, wurden sie anschließend auf einem den Teilnehmenden *unbekannten* Korpus getestet. Es wurde ein Entwicklungskorpus publiziert und angekündigt, dass das finale Korpus ähnlich sein würde. Die Teilnehmenden waren also gezwungen, ihre Richtlinien so zu schreiben, dass sie möglichst breit anwendbar sind. Dieses Vorgehen ist von der Aufteilung in Trainings- und Test-

daten, die bei maschinellen Lernverfahren verwendet wird, inspiriert (cf. Witten und Frank 2005, S. 144 f.). Hier wie dort soll dadurch sichergestellt werden, dass die entwickelten Systeme bzw. Richtlinien so generisch wie möglich sind und von den konkreten Daten abstrahieren.

Korpus

Das Korpus wurde so zusammengestellt, dass es möglichst viele Aspekte des Phänomens abdeckt. Repräsentativität war kein Entscheidungskriterium. Die Texte im Korpus unterscheiden sich in Bezug auf Genre, Publikationszeitraum und Länge.[6] Alle Texte wurden auf Deutsch und Englisch bereitgestellt, teilweise handelt es sich um Übersetzungen aus einer dritten Sprache.

Wir haben festgelegt, dass kein Text länger als 2000 Wörter sein sollte. Damit sich daraufhin nicht nur sehr kurze Texte im Korpus befanden (was sich potenziell auf die Erzählebenenstruktur hätte auswirken können), wurden auch längere Texte in gekürzter Form aufgenommen. Bei den Kürzungen wurde darauf geachtet, dass die Ebenenstruktur davon nicht betroffen war, wobei die Konzeptionen, die den Teilnehmenden in Referenzen zur Verfügung gestellt worden waren, als Entscheidungsgrundlage dienten. Insgesamt wurden 17 Texte als Teil des Entwicklungskorpus zur Verfügung gestellt, das Testkorpus bestand aus acht Texten. Die Tabellen 1 und 2 zeigen die Texte sowie Metadaten. Alle Texte sind frei verfügbar und können über das GitHub repository[7] bezogen werden.

Parallelannotationen

Die Messung des sog. *inter-annotator agreement*s (IAA) ist ein gängiger Weg, um die intersubjektive Anwendbarkeit von Annotationsrichtlinien zu beurteilen. Um das IAA zu berechnen, muss der Text von mehreren Annotierenden unabhängig voneinander, aber mit der gleichen Richtlinie, annotiert werden.

Im Rahmen des *shared task* wurde dies umgesetzt, indem jedes teilnehmende Team gebeten wurde, das Testkorpus nach der eigenen und nach einer von dem Organisationsteam zugewiesenen fremden Richtlinie zu annotieren. Zusätzlich wurde eine Gruppe studentischer Hilfskräfte dafür bezahlt, die Richtlinien ebenfalls auf die gleichen Texte anzuwenden. Insgesamt wurde also jeder Text dreimal nach der gleichen Richtlinie annotiert. Tabelle 3 zeigt eine Übersicht der verschiedenen Annotationstypen.

---

6 Genres: Anekdote, Fabel, Märchen, Kunstmärchen, Roman, Novelle, Erzählung, Kurzgeschichte. Entstehungszeitpunkte: 19. und 20. Jahrhundert. Textlänge: maximal 2000 Wörter.

7 https://github.com/SharedTasksInTheDH

Tab. 1: Das Entwicklungskorpus

Titel (orig.)	Autor*in	Titel (englisch)	Genre	Jahr	Sprache (orig.)
	Aesop	The Wolf and the Lamb	fable	600 v.u.Z.	
Rosen-Alfen	Andersen, Hans-Christian	The Elf of the Rose	folktale	1839	dänisch
Kjærestefolkene [toppen og bolden]	Andersen, Hans Christian	The Top and Ball	folktale	1862	dänisch
Se una notte d'inverno un viaggiatore	Calvino, Italo	If on a Winter's Night a Traveller	novel	1979	italienisch
Мститель	Čechov, Anton Pavlovič	An Avenger	short story	1887	russisch
The Child's Story	Dickens, Charles	The Child's Story	short story	1852	englisch
Die drei Federn	Grimm, Brüder	Feathers	folktale	1819	deutsch
Das wohlfeile Mittagessen	Hebel, Johann Peter	The Cheap Meal	anecdote	1804	deutsch
Der geheilte Patient	Hebel, Johann Peter	The Cured Patient	anecdote	1811	deutsch
Hills Like White Elephants	Hemingway, Ernest	Hills Like White Elephants	short story	1920	englisch
How the Leopard got his Spots	Kipling, Rudyard	How the Leopard got his Spots	short story	1901	englisch
Beyond the Pale	Kipling, Rudyard	Beyond the Pale	short story	1888	englisch
Unwahrscheinliche Wahrhaftigkeiten	Kleist, Heinrich von	Improbable Veracities	anecdote	1810	deutsch
	Lagerlöf, Selma	Among the Climbing Roses	narration	1894	schwedisch
The Cask of Amontillado	Poe, Edgar Allen	The Cask of Amontillado	short story	1846	englisch
Frankenstein or The Modern Prometheus	Shelley, Mary	Frankenstein or The Modern Prometheus	novel	1818	englisch
A Haunted House	Woolf, Virginia	A Haunted House	short story	1921	englisch

Tab. 2: Das Testkorpus

Titel (orig.)	Autor*in	Titel (englisch)	Genre	Jahr	Sprache (orig.)
Lenz	Büchner, Georg	Lenz	novella	1839	deutsch
Выигрышный билет	Čechov, Anton Pavlovič	The Lottery Ticket	short story	1887	russisch
The Gift of the Magi	Henry, O.	The Gift of the Magi	short story	1905	englisch
Kleine Fabel	Kafka, Franz	A Little Fable	fable	1831	deutsch
Der blonde Eckbert	Tieck, Ludwig	The White Egbert	literary fairy tale	1797	deutsch
Der Schimmelreiter	Storm, Theodor	The Rider of the White Horse	novella	1888	deutsch
Anekdote aus dem letzten preußischen Kriege	Kleist, Heinrich von	Anecdote from the Last Prussian War	anecdote	1810	deutsch
Herr Arnes penningar	Lagerlöf, Selma	The Treasure	narration	1904	schwedisch

Bezeichnung	Beschreibung
Own	Annotation durch die Autorinnen/Autoren einer Richtlinie
Foreign	Annotation durch die Autorinnen/Autoren einer *anderen* Richtlinie
Student	Annotation durch die studentischen Hilfskräfte

Tab. 3: Übersicht über die Annotationstypen

Workshop

Als Meilenstein des ersten *shared task* wurden alle Teilnehmenden nach Hamburg zu einem dreitägigen Workshop eingeladen. Von den acht teilnehmenden Teams konnten sieben physisch anwesend sein, das achte Team konnte zeitweise per Videokonferenz teilnehmen. Am ersten Tag der dreitägigen Veranstaltung lernten die Teilnehmenden sich gegenseitig kennen und stellten ihre Richtlinien vor. Dazu wurden kurze Präsentationen gehalten und danach bereits erste Gemeinsamkeiten und Unterschiede benannt. Am zweiten Tag wurden die Evaluationskriterien (s. u.) vorgestellt und diskutiert und im Anschluss mithilfe eines Fragebogens mit einer Bewertungsskala auf die Richtlinien angewendet. Im Anschluss begründeten die Teilnehmenden ihre Punktevergabe. Am letzten Tag wurden die kumulierten Evaluationsergebnisse zusammen mit dem *inter-annotator agreement* präsentiert und von der gesamten Gruppe diskutiert. Zum Abschluss wurden die nächsten Schritte besprochen.

3.2.2 Ausblick: *shared task* 2 – Automatische Erkennung von Erzählebenen

Der zweite *shared task* soll als ‚normaler', also computerlinguistischer *shared task* abgehalten werden und richtet sich daher primär an Wissenschaftler*innen aus der Computerlinguistik und verwandten Gebieten. Es ist geplant, dass der *shared task* im Sommer 2021 stattfindet. Das bis dahin annotierte Korpus wird dafür in ein Entwicklungs-, Trainings-, und Testkorpus aufgeteilt und zu bestimmten Zeitpunkten im Laufe des *tasks* zur Verfügung gestellt. Es ist anvisiert, dass der *shared task* im Rahmen der SemEval-Initiative stattfindet. Das erwartete Ergebnis des zweiten *shared task* sind eine Reihe automatischer Systeme, die Erzählebenen erkennen.

4 Evaluation von Annotationsrichtlinien

Die Evaluation von Annotationsrichtlinien im Rahmen eines *shared task* ist eine Herausforderung, die so noch nicht angegangen wurde. Das hier vorgestellte Evaluationsprozedere wurde von Grund auf neu entwickelt und ist, so unsere Überzeugung, auch jenseits der Annotation von Textphänomenen anwendbar.

4.1 Grundlagen und Herausforderungen

Unser grundsätzliches Ziel besteht in der Etablierung eines Verfahrens, das den disziplinären Ansprüchen sowohl der Literaturwissenschaft als auch der Computerlinguistik gerecht wird. Das heißt konkret, dass die folgenden vier Bedingungen erfüllt sein müssen:

Etablierung einer Rangfolge: Die Methode muss offensichtlich in der Lage sein, eine Rangfolge zu etablieren. Diese Rangfolge sollte so klar wie möglich sein und Gleichstände, soweit es geht, vermeiden.

Explizit- und Klarheit: Da *shared tasks* als Wettbewerb angelegt sind, müssen die Richtlinien nach einer Zielfunktion sortiert werden. Die Zielfunktion muss dabei vorher so genau wie möglich festgelegt werden, damit die Teilnehmenden vorher wissen, worauf sie sich einlassen und möglichst wenig Raum für Missverständnisse entsteht.

Praktikabilität: Die Evaluation muss innerhalb der praktischen Beschränkungen eines *shared task* durchführbar sein. Bei uns hieß das, dass es möglich sein musste, innerhalb eines dreitägigen Workshops ein Ergebnis zu erzielen.

Umsetzung der Evaluationskriterien: Jede Evaluationsmethode muss die Kriterien möglichst passgenau umsetzen, so dass eine Richtlinie, die bestimmte Aspekte besser löst als eine andere, auch eine höhere Bewertung bekommt. Welche Kriterien genau angelegt werden, ist eine Setzung, die das Organisationsteam vornehmen muss. Im Sinne der Fairness sollten sie so früh wie möglich kommuniziert werden.

Diese Anforderungen ergeben sich daraus, dass Annotationsrichtlinien in einem *shared task* erstellt werden. Während in computerlinguistischen *shared tasks* ein Goldstandard existiert, der so gut wie möglich reproduziert werden muss, gibt es für Annotationsrichtlinien bisher keine etablierte *ground truth*, anhand derer sie bewertet werden könnten. Selbst das Messen von *inter-annotator agreement*, das gelegentlich als Qualitätsmaßstab für Annotationsrichtlinien herangezogen wird, könnte in diesem Fall in die Irre führen, da mit nicht eindeutig auflösba-

ren Ambiguitäten zu rechnen ist. Annotationsunterschiede sollten nämlich nicht als Mangel der Richtlinien dargestellt werden, wenn das Annotationsphänomen stark ambig ist.

Zusätzlich zu den genannten allgemeinen Kriterien, die jedes Evaluationsverfahren erfüllen sollte, bringt unser *shared task* einige spezifische Herausforderungen mit sich:

Da es sich bei diesem *shared task* um eine interdisziplinäre Unternehmung handelt, muss mit einem heterogenen Kreis aus Teilnehmer*innen gerechnet werden. Das Konzept Annotation spielt, in verschiedenen Benennungen und Ausformungen, verschiedene Rollen in den beteiligten Disziplinen, wodurch sich eine diverse Sammlung an best practices, Regeln und Traditionen etabliert hat. Literaturwissenschaftliche Annotation z. B. wird typischerweise als Unterstreichung von Textpassagen oder als *note-taking*-Annotationen verstanden, also als Anfertigen von Notizen am Seitenrand, ohne feste Kategorien oder explizite Verbindung zur auslösenden Textstelle. In der Computerlinguistik hingegen werden Annotationen typischerweise parallel von mehreren Annotator*innen vorgenommen und eine hohe Übereinstimmung ist oberstes Ziel. Letzteres ist bei subjektiven Annotationen mit dem Ziel, eine Materialsammlung für eine Interpretation zu liefern, schlicht nicht relevant. Die Teilnehmenden haben also unterschiedliche Vorerfahrungen und Erwartungen an den Annotationsprozess – gleichwohl muss das Ergebnis der Evaluation für alle Beteiligten nachvollziehbar, funktional und auch nutzbringend sein.

Eine weitere Herausforderung liegt in der für die Literaturwissenschaft durchaus typischen Vagheit der genutzten Konzepte. Die Narratologie stellt innerhalb der Digital Humanities ein beliebtes Betätigungsfeld dar, was vermutlich auf ihren strukturalistischen Hintergrund und die damit einhergehende, vergleichsweise einfachere Operationalisierung zurückzuführen ist. Zudem vereint die Narratologie und die Digital Humanities die Auffassung, dass eine strukturelle Analyse von Texten *überhaupt* interessante textbezogene Beobachtungen zutage fördert. Aus Annotationssicht bleiben die in der Narratologie diskutierten Konzepte jedoch zu vage, um sie direkt anzuwenden. Neben den Primärtexten selbst geben auch die narratologischen Konzepte einen gewissen Interpretationsspielraum. Wir haben daher im Vorfeld beschlossen, dass die Evaluation möglichst viele verschiedene, auch unterschiedliche narratologische Perspektiven berücksichtigen sollte. Vagheiten und komplexe Abhängigkeiten der Konzepte sollten also von den Richtlinien nicht zwingend aufgelöst werden. Trotzdem sollte idealerweise ein Umgang damit beschrieben werden.

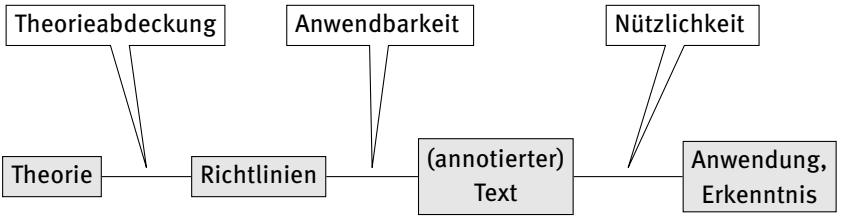


Abb. 1: Die drei Evaluationsdimensionen (oben) verbinden Schritte in der computationellen Textanalyse in den Digital Humanities (unten)

4.2 Evaluationskonzept

Das im *shared task* verwendete Evaluationskonzept beruht auf drei Dimensionen, mit denen unterschiedliche Eigenschaften der Richtlinien beurteilt werden können. Abbildung 1 zeigt schematisch, wo sich die Dimensionen im Verhältnis zu Forschungsaktivitäten im Kontext computationeller Textanalysen befinden. Damit verteilen sich die Dimensionen auf den gesamten Forschungsprozess, von der narratologischen Theorie über die Erstellung von Richtlinien und die Annotation von Texten bis hin zu den Erkenntnissen, die sich aus den annotierten Daten, seien es einzelne Texte oder ganze Korpora, ergeben.

Die Dimension **Theorieabdeckung** reflektiert, wie viel von einer theoretischen Basis in den auf ihr aufbauenden Richtlinien enthalten ist. Wenn etwa eine Richtlinie explizit auf einer narratologischen Theorie aufbaut, könnte sie darauf abzielen, jede einzelne Definition, Regel und Ausnahme in die Richtlinie aufzunehmen. Eine andere Richtlinie, die auf derselben Theorie basiert, könnte hingegen Aspekte weglassen oder stark verändern. Die Dimension **Theorieabdeckung** verbindet also Richtlinien und Theorie.

Anwendbarkeit betrachtet die Relation zwischen Richtlinien und Text. Diese Dimension repräsentiert, wie gut eine Richtlinie die Annotator*innen darauf vorbereitet, die Annotationen tatsächlich durchzuführen. Die Anwendbarkeit einer Richtlinie kann z. B. durch gut durchdachte und instruktive Beispiele erhöht werden, aber auch durch eine klare Struktur und/oder einen sorgsamen Gebrauch der Terminologie. *Inter-annotator agreement* und Systematizität der erfolgten Annotationen fallen ebenfalls in diese Dimension.

Die Dimension der **Nützlichkeit** verbindet schließlich den annotierten Text mit Anwendungen und Erkenntnissen, wobei beide Begriffe hier relativ weit gefasst sind. Anwendungen decken hier z. B. auch automatische Analyseverfahren ab, die auf die Erzählebenen aufbauen, aber auch die Nützlichkeit der Ergebnisse von Analysen großer Datenmengen für die anschließende (menschliche) Interpre-

tation. In dieser Dimension wird also auf das Ergebnis, also die erfolgten Annotationen als solche, abgezielt und nicht so sehr auf den Annotationsprozess.[8]

Die drei beschriebenen Dimensionen erlauben eine balancierte Evaluation von Richtlinien mit heterogenen disziplinären Hintergründen, Zielen und Annahmen über theoretische Konzepte. Eine Richtlinie, die sich nur auf eine Dimension konzentriert, kann zwar in dieser gut abschneiden, läuft aber Gefahr, genau dadurch die anderen Dimensionen zu schwächen. So könnte eine Richtlinie ausschließlich auf narratologische Kategorien und Begriffe setzen und würde damit zwar eine hohe Theorieabdeckung erzielen, aber, wegen der oben beschriebenen Vagheit, keine sehr hohe Anwendbarkeit. Wird eine Richtlinie hingegen rein auf ihre Anwendbarkeit optimiert, ließen sich zwar schnell und zuverlässig Annotationen gewinnen – die aber am Ende niemandem nützen (weil sie im Extremfall definieren, dass ein Text aus einer Erzählebene besteht). Die reine Optimierung auf Nützlichkeit führt womöglich zu Richtlinien, die keinerlei Bezug mehr zu theoretischen Konzepten haben. Die Herausforderung für die Teilnehmenden besteht also in der Balancierung unterschiedlicher Anforderungen, die sich aus den Dimensionen ergeben.

Selbstverständlich kann auch eine Richtlinie, die nicht alle drei Dimensionen gleich gewichtet, sinnvolle Anwendungen ermöglichen. Richtlinien ohne Theoriebezug könnten etwa auch neue theoretisch interessante Phänomene aufspüren helfen. Es ist also in der Praxis und außerhalb von *shared tasks* sicher nicht nötig, auf alle drei Dimensionen abzuzielen. Doch auch in diesem Fall können die Dimensionen eine bewusste Entscheidung unterstützen.

4.3 Umsetzung

Um das oben beschriebene, dreidimensionale Evaluationsmodell praktisch einzusetzen und zu erproben, wurde jede Dimension mit einer Reihe von konkreten Fragen unterfüttert, die verschiedene Aspekte aus der Dimension abfragen. Die Fragen sind weiter unten aufgelistet. Da sie im Original auf Englisch verwendet wurden, sind sie hier ebenfalls auf Englisch abgedruckt.

Die Fragen wurden den Teilnehmenden vor der Einreichung ihrer Richtlinien zur Verfügung gestellt. Während des Evaluationsworkshops wurden sie auf zwei-

8 Vergleiche zu den drei Dimensionen auch die in der empirischen Sozialwissenschaft etablierten Begriffe der Validität und Reliabilität. Auch wenn sie nicht exakt deckungsgleich sind, lassen sich Bezüge herstellen: Die Dimension Anwendbarkeit ist dabei verwandt mit der Reliabilität, während die externe Validität von der Dimension Nützlichkeit und die interne von der Dimension Theorieabdeckung abgedeckt werden.

erlei Weisen verwendet. Zunächst bildeten sie den Leitfaden für eine qualitative Diskussion der Richtlinien. Durch den Fragebogen wurde sichergestellt, dass alle Aspekte in der Diskussion zur Sprache kamen und dass an alle Richtlinien die gleichen Maßstäbe angelegt wurden. Diese Diskussion wurde von den Teilnehmenden als intensiv, aber sehr produktiv wahrgenommen und gab bereits Anlass für Verbesserungen.

Außerdem wurden die Fragen quantitativ beantwortet. Dazu wurde jeder Frage eine Likert-Skala mit vier Werten zugeordnet. Die teilnehmenden Teams wurden dann gebeten, jeder (anderen) Richtlinie in jeder Frage Punkte zuzuweisen, wobei mehr Punkte eine bessere Bewertung bedeuteten.

Unser Fragebogen definiert vier Fragen für die Dimensionen Theorieabdeckung und Nützlichkeit und zwei Fragen für die Dimension Anwendbarkeit. Damit alle Dimensionen gleichwertig waren, wurde die Dimension Anwendbarkeit ergänzt durch die Ergebnisse des skalierten *inter-annotator agreement* (gemessen mit ɣ, Mathet et al. 2015). Damit wurden jeder Richtlinie in jeder Dimension vier Werte zugeordnet, die erst als Dimensionswerte und dann als Gesamtwert aufaddiert wurden.

4.3.1 Fragebogen

Der Fragebogen wird hier im Original wiedergegeben.

Theorieabdeckung

1. Is the narrative level concept explicitly described?
   Explanation: Narrative levels can be described or defined. This depends on the narratology used; some of them are structuralist, others are post-structuralist. Regardless of the mode, is the description/definition understandable and clear?
   - 1: I did not understand what the guideline describes as 'narrative level'.
   - 4: I fully understood the concept described in the guideline.
2. Is the narrative level concept based on existing concepts?
   Explanation: The level concepts can be self-designed, oriented on existing narratologies or copied from an existing level definition
   - 1: The theory relation of the used level concept is not clear.
   - 4: It is clearly mentioned whether the level concept is made up or (partially) based on a theory.
3. How comprehensive is the guideline with respect to aspects of the theory? Does it omit something? Explanation: If the guideline is based on a theory

or multiple theories, does it include the whole theory or only parts of it? Are there reasons mentioned why aspects are in-/excluded?
   - 1: The guideline does not clearly state the extension of its dependence on theory/ies.
   - 4: The guideline unambiguously states the scope of its theory-dependance.
4. How adequately is the narrative level concept implemented by this guideline in respect to narrative levels?
   Explanation: Narratologies differ in their complexity. Firstly, you have to decide whether complexity or simplicity (in relation to x) is desirable, then you have to answer:
   - 1: The guideline is too simple or too complex for narrative levels and thus not adequate.
   - 4: The guideline's complexity is adequate.

Anwendbarkeit

1. How easy is it to apply the guideline for researchers with a narratological background?
   Explanation: The question asks for an assessment of the ease of use of the guideline for an annotator with some narratological background. Indicators can be: Complexity of the concepts, length of the guideline, clarity of examples, clear structure, difficulty of finding special cases, etc.
   - 1: Even as a narratology expert, I needed to read the guideline multiple times and/or read additional literature.
   - 4: The guideline is very easy to apply, and I always knew what to do.
2. How easy is it to apply the guideline for researchers without a narratological background?
   Explanation: The question asks for an assessment of the ease of use of the guideline if we assume an annotator who doesn't have a narratological background (e.g., an undergraduate student). Indicators can be: Complexity of the concepts, length of the guideline, use of terminology, clarity of examples, reference to examples only by citation, clear structure, difficulty of finding special cases, etc.
   - 1: Non-experts have no chance to use this guideline.
   - 4: The guideline is very easy to apply, and non-experts can use them straight away.

Nützlichkeit

1. Thought experiment: Assuming that the narrative levels defined in the annotation guideline can be detected automatically on a huge corpus. How helpful are these narrative levels for an interesting corpus analysis?
   Explanation: This question focuses on the relevance of the narrative level annotations for textual analysis of large amounts of texts, e.g., for the analysis of developments over time with regard to narrative levels or a classification of texts with regards to genre, based on narrative levels.
   - 1: The narrative levels annotations are irrelevant for corpus analysis.
   - 4: The annotations provide interesting data for corpus analysis.
2. How helpful are they as an input layer for subsequent corpus or single text analysis steps (that depend on narrative levels)?
   Explanation: The analysis of some other textual phenomena depends on narrative levels, e.g., chronology should be analyzed within each narrative level before analyzing it for the whole text. This question asks whether the analysis of such phenomena is possible or even better when based on the narrative level annotations.
   - 1: The usage of the narrative levels annotations makes no difference for subsequent analyses.
   - 4: Subsequent analyses are possible only because of the narrative level annotations.
3. Do you gain new insights about narrative levels in texts by applying the foreign guideline, compared to the application of your own guideline?
   Explanation: In most cases annotating a text in accordance to a guideline changes the evaluation of textual phenomena in the text, e.g., the quality (or quantity) of narrative levels in the text.
   - 1: It doesn't make a difference—I get no additional insights with the foreign guideline.
   - 4: I gain a lot of new insights about narrative levels in texts based on this guideline.
4. Does the application of this guideline influence your interpretation of a text?
   Explanation: Interpretations are normally based on the analysis of a text and thus on the observation of the presence (or absence) of certain textual phenomena. Therefore, the application of the guidelines may result in annotations that are relevant for your interpretation, e.g. the detection of a narrative level of a certain type may influence your interpretation of the reliability of a narrator.
   - 1: My interpretation is independent from the annotations based on the guideline.

– 4: My interpretation is based primarily on the annotations based on the guideline.

4.3.2 Integration der Evaluationsergebnisse

Mithilfe des Fragebogens wurden für jede Richtlinie sieben Bewertungen eingeholt (von den sieben jeweils anderen Teams). Aus diesen Bewertungen wurde eine Gesamtpunktzahl wie folgt berechnet:

1. Für jede der zehn Fragen wird das arithmetische Mittel über alle Antworten berechnet. Daraus ergeben sich zehn Mittelwerte, vier in der ersten und dritten und zwei in der zweiten Dimension.
2. Die IAA-Ergebnisse werden in das Intervall [1;4] skaliert und zweimal als ‚virtuelle Fragen' in der Dimension der Anwendbarkeit addiert. Damit existieren vier Werte je Dimension für jede Richtlinie.
3. Die Werte jeder Dimension werden dann addiert, woraus sich für jede Richtlinie und jede Dimension ein Wert im Intervall [4;16] ergibt.
4. Als Gesamtpunktzahl werden diese Werte addiert. Die Gesamtpunktzahl liegt dann im Intervall [12;48].

5 Ergebnisse des *shared task*

5.1 Allgemeine Beobachtungen

Wie zu erwarten, waren die Einreichungen für den *shared task* in Bezug auf Disziplinen und geographische Herkunft divers. Tabelle 4 zeigt einige Eckdaten der teilnehmenden acht Teams. Die Teams unterscheiden sich in Bezug auf Alter, Geschlecht, Größe, akademischen Status und disziplinären Hintergrund. Diese Heterogenität zeigt sich auch in den Richtlinien, deren Länge von einer bis fünfzig Seiten reicht und unter denen sich sowohl theoretische Essays als auch praxisorientierte Anleitungen finden. Zwei der eingereichten Richtlinien wurden von Mitarbeiter*innen aus CRETA verfasst und sind daher auch in diesem Buch abgedruckt: Richtlinie IV (Ketschik, Murr et al. 2020) ab Seite 440 und Richtlinie V (Barth 2020) ab Seite 423.

9 https://culturalanalytics.org/

Tab. 4: Teilnehmende Teams. Alle Richtlinien wurden in einer Sonderausgabe der Zeitschrift *Cultural Analytics*[9] publiziert. Die mit * markierten Richtlinien sind in diesem Band abgedruckt als Barth (2020) und Ketschik, Murr et al. (2020). Disziplinärer Hintergrund basiert auf einer Selbstbeschreibung der Teilnehmenden: Computerlinguistik (CL), Englische Literatur (EL), Data science (DS), Literaturwissenschaft (LW), Digital Humanities (DH). Herkunft: Vereinigte Staaten von Amerika (US), Deutschland (DE), Schweden (SWE), Irland (IRL), Kanada (CA). Größe beschreibt die Zahl der Autorinnen und Autoren einer Richtlinie, L markiert Richtlinien, die im Rahmen einer Lehrveranstaltung entstanden sind.

Richtlinie	Referenz	Disziplin	Herkunft	Größe
I	Eisenberg und Finlayson (2019)	CL	US	2
II	Kearns (2019)	EL/DS	IRL	1
III	–	EL	DE	2L
IV	Ketschik, Krautter et al. (2019)*	LW/DH	DE	4
V	Barth (2019)*	DH/LW	DE	1
VI	Bauer und Lahrsow (2020)	EL	DE	2L
VII	Wirén et al. (2020)	CL	SWE	3
VIII	Hammond (2020)	LW	CA	1L

5.1.1 Richtlinien und Erzählebenen

Da es den Teilnehmenden überlassen war, ob und ggf. an welchen theoretischen Werken sie sich orientierten, stellen wir die theoretischen Grundlagen der Richtlinien auf Basis der in ihnen enthaltenen Referenzen und Erwähnungen dar. Tabelle 5 zeigt eine Übersicht über die narratologischen Werke, während Tabelle 6 die erwähnten Konzepte zeigt.[10]

Es ist nicht überraschend, dass Genettes *Narrative Discourse* am häufigsten referenziert wurde, da das Konzept der Erzählebene hier prominent diskutiert wird und die meisten anderen Ansätze in der einen oder anderen Form darauf aufbauen. Andere Werke, die relativ prominent sind (in mehr als drei Richtlinien referenziert), sind die Einführungstexte von Jahn („N2.4. Narrative Levels“) und Pier („Narrative Levels“) sowie der Ansatz von Ryan (*Possible worlds, artificial intelligence, and narrative theory*).

10 Man beachte jedoch, dass hier die Referenz auf Werke bzw. Konzepte gesammelt wurde. Nicht beachtet wurde hingegen, ob die Interpretation des Werkes bzw. die Nutzung des Konzeptes kompatibel sind.

11 Die folgenden Arbeiten wurden zusätzlich in einzelnen Richtlinien zitiert: Richtline II: McHale (2014); Rickert (1923). Richtline III: Nünning (2004); Ryan (1986). Richtline IV: Abrams (1999); Pier (2016); Fludernik (2009); Genette (1988); Genette (1997); Margolin (2014); Ryan (2001); Schmid (2014). Richtline V: Duyfhuizen (2005); Genette (1988); Gius (2015); Rimmon-Kenan (2005); Ryan

Tab. 5: Narratologische Arbeiten, die in den Richtlinien operationalisiert wurden. Die Zuweisung basiert auf Referenzen in den Richtlinien und/oder Einleitungen.

	Richtlinie	I	II	III	IV	V	VI	VII	VIII
Narratologische Arbeiten	Genette 1980		x	x	x	x	x	x	
	Jahn (2017)		x		x	x		x	x
	Lahn und Meister (2008)				x	x			
	Lämmert (1955)					x			
	Mani (2012)								
	Martínez und Scheffel (2003)				x				
	Nelles (1997)			x		x	x		
	Nelles (2005)					x			
	Neumann und Nünning (2012)					x			
	Pier (2014)			x		x	x		
	Pier (2012)				x				
	Rimmon-Kenan (2004)								
	Romberg (1962)								
	Ryan (1991)			x	x	x			
	zusätzliche Referenzen[11]		x	x	x	x	x	x	
	eigene Beiträge	x							x

Tab. 6: Narratologische Konzepte, die in den Richtlinien operationalisiert wurden. Gesammelt sind explizite Referenzen.

	Richtlinie	I	II	III	IV	V	VI	VII	VIII
Narratologisches Konzept	Definition einer ‚Erzählung'	x			x	x	x	x	x
	Erzählebene	x	x	x	x	x	x		x
	Diskursebene							x	
	Erzähler (Identität, Verhältnis zum Text)	x	x	x	x	x	x	x	x
	Rezipient			x			x	x	
	Struktur der Ebenengrenze	x		x	x	x	x	x	
	Art der Ebenengrenze (z. B. Metalepsis)			x	x	x	x	x	
	Sprecher (illokutionäre Grenze)			x	x	x		x	
	Veränderung der Welt (ontol. Grenze)	x		x		x	x		
	Fokalisierung			x			x	x	
	Analepse/Prolepse	x	x						
	Bewusstseinsstrom/free indirect discourse		x						
	Dehnung/Raffung		x						

Tabelle 6 zeigt eine Übersicht über die Konzepte, die von den Teilnehmenden für die Identifikation von Erzählebenen als relevant erachtet wurden. Die aufgeführten Konzepte können danach unterschieden werden, ob sie unmittelbar mit Erzählebenen zusammenhängen (z. B. Art der Ebenengrenze) oder häufig mit Erzählebenen gemeinsam auftreten (z. B. Fokalisierung). Während erstere direkt für die Operationalisierung nützlich sein können, sind letztere eher für weitere Analysen relevant, da sie häufig innerhalb von bzw. zusammen mit Erzählebenen auftreten. Diese Mischung der Konzepte kann einerseits durch die theoretische Offenheit und andererseits durch die unterschiedlichen Ziele der Teilnehmenden erklärt werden.

6 Ergebnisse der Evaluation

Im Folgenden werden die Evaluationsergebnisse der *ersten Runde* des *shared task* präsentiert. Die Gesamtergebnisse finden sich in Tabelle 7. Danach folgen die Ergebnisse in den Dimensionen Theorieabdeckung (Tabelle 8), Anwendbarkeit (Tabelle 9) und Nützlichkeit (Tabelle 10). Die dimensionsspezifischen Tabellen zeigen Mittelwerte und Standardabweichungen gemäß den Fragebögen. Die Fragen selbst sind in Abschnitt 4.3.1 wiedergegeben.

Die Ergebnisse in Tabelle 8 basieren auf vier Fragen. Richtlinie IV (S. 440) hat hier das beste Ergebnis erzielt. Dabei handelt es sich um eine Richtlinie, die auf eine tiefe narratologische Beschreibung setzt und von narratologischen Be-

Tab. 7: Erzielte Ergebnisse der Guidelines nach Dimensionen

Richtlinie	Theorieabdeckung	Anwendbarkeit	Nützlichkeit	Summe	IAA
V	14,14	12,09	**12,88**	**39,1**	0,25
II	11,17	11,89	12,57	35,63	0,24
VI	12,33	11,01	11,37	34,71	0,21
IV	**14,43**	7,71	11,26	33,4	0,05
VIII	8,1	**14,14**	9,12	31,36	**0,3**
VII	11,6	9,82	9,77	31,18	0,23
I	7,83	10,39	10	28,22	0,18
III	10,29	6,48	10,95	27,72	0,07

(2002). Richtline VI: Füredy (1989); Nünning (2004); Ryan (1986). Richtline VII: Doležel (1973); Niederhoff (2013); Rimmon-Kenan (2005); Todorov (1966). Richtline VIII: Abbott (2002). Richtline III wurde nicht veröffentlicht, da sie zurückgezogen wurde.

Tab. 8: Evaluationsergebnisse in der Dimension Theorieabdeckung. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

Richtlinie	Q1	Q2	Q3	Q4
I	3 ±1,1	1 ±0	1 ±0	2,83±0,98
II	2,67±1,03	3 ±1,1	2,83±0,75	2,67±1,21
III	2,43±0,98	2,86±1,07	2,71±1,38	2,29±0,76
IV	3,71±0,49	4 ±0	3,71±0,49	3 ±0,82
V	3,71±0,49	4 ±0	3,43±0,53	3 ±1
VI	3,33±0,82	3,33±0,82	2,67±0,52	3 ±0,89
VII	2,71±1,11	3,33±0,82	2,71±0,95	2,83±1,33
VIII	2,29±1,11	1 ±0	1,67±1,21	3,14±0,69

Richtlinie	Q1	Q2
I	2,6 ±0,55	2,67±0,52
II	3,17±0,41	2,17±0,98
III	2,57±1,13	1,43±0,53
IV	3,57±0,53	2,14±0,38
V	3 ±1,15	2,29±0,95
VI	3,17±1,17	2 ±0,89
VII	2,33±0,52	1,17±0,41
VIII	2,71±1,38	3,43±0,79

Tab. 9: Evaluationsergebnisse in der Dimension Anwendbarkeit. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

griffen intensiv Gebrauch macht. Am anderen Ende des Spektrums findet sich Richtlinie I, die die wenigsten Punkte erzielt hat. Dies deckt sich mit der Selbstbeschreibung und dem primären Forschungsziel der Einreichenden, das auf das „computational unterstanding of stories“ abzielt (Eisenberg und Finlayson 2019, S. 1). Damit stehen in Richtlinie I klar Anwendungen im Vordergrund, eine Anbindung an narratologische Theorie ist eher zweitrangig. Außerdem werden einige nicht-erzählende Textbeispiele genannt (z. B. Drehbücher), die aus einem Annotationsprojekt der Richtlinien-Autoren stammen. Dadurch dass die narratologisch motivierten Teilnehmenden in der Überzahl waren, wurden diese Abweichungen relativ stark ‚bestraft‘, was in einem geringen Punktwert resultierte.

Die Punktzahl in der Anwendbarkeits-Dimension basiert auf der Berechnung des *inter-annotator agreement* und zwei Fragen aus dem Fragebogen (Tabelle 9). Die erste der beiden Fragen zielt darauf ab, wie gut Expert*innen für narratologische Textanalyse die Richtlinie anwenden können, die zweite stellt die gleiche Frage für Laien. Richtlinie VIII (Hammond 2020) erzielt hier die meisten Punkte, während Richtlinie III die wenigsten erhält. Für eine Interpretation dieser Ergebnisse ist es hilfreich, die einzelnen Fragen und das IAA separat zu betrach-

Tab. 10: Evaluationsergebnisse in der Dimension Nützlichkeit. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

Richtlinie	Q1	Q2	Q3	Q4
I	3,17 ±0,75	3 ±0,71	2,17 ±0,75	1,67±0,52
II	3,5 ±0,55	3,4 ±0,89	3 ±0,89	2,67±0,52
III	3,33±0,82	3,17 ±0,98	2,29±0,76	2,17 ±0,75
IV	3,29±0,76	2,83±0,98	2,71 ±0,76	2,43±0,98
V	3,5 ±0,55	3,4 ±0,55	3,14 ±0,69	2,83±0,75
VI	3,4 ±0,55	3,4 ±0,89	2,17 ±0,75	2,4 ±0,55
VII	3 ±0,63	2,6 ±0,55	2,17 ±0,98	2 ±0,89
VIII	3,2 ±0,84	2,8 ±1,1	1,29±0,76	1,83±0,75

ten. Dass Richtlinie VIII hier die höchste Punktzahl erzielte, wirft die Frage nach dem Verhältnis aus Einfachheit des Konzeptes und Anwendbarkeit auf. Richtlinie VIII erzielt das höchste IAA, scheint also zumindest am robustesten anwendbar zu sein. Gleichzeitig hat Richtlinie VIII nur eine Position im Mittelfeld für Expert*innen erzielt und lediglich für Laien eine sehr gute Anwendbarkeit. In der Praxis scheint ein einfaches Ebenenkonzept für Laien gut anwendbar zu sein, dies scheint jedoch nicht auch (automatisch) für Expert*innen zu gelten.

Die gute Bewertung von Richtlinie V in der Dimension der Nützlichkeit lässt sich wahrscheinlich auf a) die vielen Beispiele, die die Benutzung der Kategorien illustrieren, und b) die klare Beschreibung der Forschungsziele durch den Richtlinienautor zurückführen. Ähnlich lässt sich die gute Bewertung von Richtlinie II erklären, die hier den zweiten Platz erzielt hat. Die Richtlinie erklärt, sie sei „designed for annotating analepsis, prolepsis, stream-of-consciousness, free indirect discourse, and narrative levels, with facility also for annotating instances of extended or compressed time, and for encoding the identity of the narrator“ (Kearns 2019, S. 3). Da die Dimension Nützlichkeit eine Reihe von Einsatzzwecken und letztlich das Potenzial der Richtlinie abdeckt, ist es interessant zu beobachten, dass die Angaben über konkrete Einsatzszenarien die Einschätzung der Nützlichkeit insgesamt zu erhöhen scheint.

Zum Abschluss noch ein Kommentar zur Gewinnerrichtlinie (Richtlinie V Barth 2019), für diesen Band überarbeitet und abgedruckt als Barth 2020 ab Seite 423. Richtlinie V hat keineswegs in allen Dimensionen gewonnen. Allerdings hat sie eine gute Balance zwischen den drei Evaluationsdimensionen geschaffen. Sowohl die quantitativen Ergebnisse der Evaluation als auch die Diskussion während des Workshops bestätigten, dass die Richtlinie Erzählebenen genau und präzise beschreibt. Die Richtlinie unterscheidet Erzählebenen von Erzählakten und

verwendet auch andere narratologische Konzepte (z. B. Erzähler) als Hilfsmittel zur Erkennung von Ebenen. Abstrakte Beispiele in der Form von Diagrammen und Tabellen machen die Richtlinie verständlich. Die Richtlinie enthält zudem konkrete Handlungsanweisungen und stellt heraus, wozu die Ebenenannotationen nützlich sein können.

7 Fazit

Da der *shared task* noch läuft, ist es für ein abschließendes Fazit zu früh. Einige erste Einsichten, gerade zum ersten Teil des *shared task*, lassen sich jedoch bereits formulieren.

inter-annotator agreement

Die Werte des *inter-annotator agreement* basieren auf einer relativ kleinen Zahl an Annotationen, die von Annotierenden ohne systematisches Training durchgeführt wurden. Die Annotationen wurden daher auf unterschiedlichen Niveaus durchgeführt (cf. Abbildung 3): Die studentischen Annotierenden waren mehrheitlich untrainiert (nur zwei von acht Studierenden hatten Erfahrung in anderen Annotationsaufgaben) und hatten keinerlei Vorkenntnisse in narratologischen Fragen. Die Annotierenden für die ‚foreign'-Annotationen waren naturgemäß auf ihre eigenen Richtlinien trainiert und einige hatten offensichtliche Schwierigkeiten, sich von ihnen zu lösen. Die Teilnehmenden waren unterschiedlich zufrieden mit den Annotationen, welche ihnen schon vor dem Workshop zur Verfügung gestellt worden waren. Einige Teilnehmende empfanden gerade die ‚foreign'-Annotationen als nicht adäquat, etwa bei narratologisch komplexen Richtlinien, die von nicht-narratologischen Teilnehmenden angewendet werden mussten. Teilweise ist dies sicher eine Folge der Interdisziplinarität des Vorhabens – für die Teilnehmenden war es schwer einzuschätzen, welche Kenntnisse sie bei den Annotierenden voraussetzen konnten.

Eine daran anschließende Beobachtung ist, dass in einigen Fällen das Agreement zwischen den ‚own'- und ‚student'-Annotationen höher war als zwischen den ‚own'- und ‚foreign'-Annotationen. Die Autor*innen der Richtlinien stimmten also eher mit den studentischen Annotierenden überein als mit den anderen Teilnehmenden am *shared task*. Es wurde daher diskutiert, ob nicht die disziplinäre Herkunft bei der Zuteilung hätte berücksichtigt werden müssen, was auf Organisationsseite wiederum nicht praktikabel erschien. Nichtsdestoweniger sind Alternativen wünschenswert, die ohne ‚foreign'-Annotationen auskommen. Letzten Endes wurde der beschriebene Annotationsmodus gewählt, um überhaupt ein *inter-*

annotator agreement ermitteln zu können. Bei einer geeigneten Finanzierung von Annotationen unter kontrollierteren Bedingungen könnte darauf verzichtet werden.

Fragebogen

Während des Workshops wurde zudem der Fragebogen in mehreren Hinsichten diskutiert:

- Fragen zur Dimension der Theorieabdeckung zu beantworten, erfordert breites narratologisches Wissen, das nicht alle Teilnehmenden mitbrachten. Dadurch konnten letztlich nicht alle Richtlinien nach gleichen Maßstäben bewertet werden.
- In der Dimension der Anwendbarkeit wird nach der Anwendbarkeit für Laien bzw. Expert*innen. Da die teilnehmenden Gruppen sich aber aus Laien *oder* Expert*innen zusammensetzten, konnte jede Gruppe nur eine der beiden Fragen aus eigener Anschauung beantworten, während für die andere eine Reihe von Annahmen getroffen werden musste.
- Insbesondere Literaturwissenschaftler*innen äußerten Besorgnis, dass das Ergebnis der komplexen Evaluation am Ende wiederum nur Zahlen sein würden. Diese Unzufriedenheit bezog sich nicht auf die Messung des IAA oder auf den Fragebogen, sondern gründete in einer grundsätzlichen methodischen Skepsis, ob Annotationsrichtlinien auf diese Weise bewertet werden können. Dies wurde von den Teilnehmenden mit technischem Hintergrund weniger als Problematik wahrgenommen.
- Zuletzt waren auch die Fragen in der Dimension der Nützlichkeit im Einzelnen schwer zu beantworten. Die Nützlichkeit einer Richtlinie kann am besten beurteilt werden, nachdem tatsächlich ein Nutzungsversuch stattgefunden hat. Dies war aber im Rahmen des *shared task* nicht machbar, weswegen in den Antworten zum Fragebogen ein gewisses Maß an Spekulation enthalten ist.

Trotz dieser Schwierigkeiten geben die Antworten die inhaltliche, qualitative Diskussion des Workshops aus unserer Sicht angemessen wieder und zeigen eine erstaunliche Homogenität (geringe Standardabweichung, s. o.).

Evaluationsmodell

Das von uns entworfene, dreidimensionale Bewertungsmodell wurde so gestaltet, dass jede Dimension auf unterschiedliche Weise Merkmale begünstigt, die mit der disziplinären Herkunft der Teilnehmenden zusammenhängen. Die Idee

war, dadurch die disziplinär bedingte Bevorzugung bestimmter Aspekte abzumildern. Entsprechend wurden die Dimensionen und ihre Kombination so gestaltet, dass sie disziplinär bedingte Verzerrungen möglichst aufheben. Die Ergebnisse der Evaluation sprechen dafür, dass dies gelungen ist. Beispielsweise erreichte die Richtlinie IV, deren Autor*innen alle einen literaturwissenschaftlichen Hintergrund haben, die erste Position in Theorieabdeckung, eine Mittelfeldposition in Nützlichkeit und die vorletzte Position in Anwendbarkeit. Richtlinie I, die von Computerlinguisten verfasst wurde, lag bei der Theorieabdeckung an letzter Stelle, erhielt aber in den anderen beiden Dimensionen durchschnittliche Bewertungen. Dies gibt uns Grund zu der Annahme, dass disziplinäre Vor- und Nachteile durch unseren Bewertungsansatz tatsächlich ausgeglichen werden. Auch die Tatsache, dass die Richtlinien IV und VIII in den Dimensionen 1 und 2 invertierte Ränge erreichten, deutet darauf hin, dass die Dimensionen die disziplinären Vorteile neutralisieren. Die insgesamt am höchsten bewertete Richtlinie erhielt in allen Dimensionen hohe Werte, erzielte aber nur in einer Dimension den ersten Platz. Dies deutet darauf hin, dass man, um generell erfolgreich zu sein, eine gewisse Balance zwischen den Dimensionen erreichen sollte. Dies ist der Effekt, der bei der Gestaltung des Evaluationsschemas angestrebt wurde.

Funding: Der *shared task* wurde freundlicherweise von der VolkswagenStiftung unterstützt, die sowohl die Reisekosten als auch einen Teil der Kosten für die Arbeiten der Hilfskräfte übernommen hat.

Danksagung: Wir bedanken uns bei den teilnehmenden Teams für die Experimentierfreude und das Vertrauen.

Literatur

Abbott, H. Porter (2002). *The Cambridge Introduction to Narrative*. Cambridge: Cambridge University Press.

Abrams, Meyer Howard (1999). „Narrative and Narratology". In: *A Glossary of Literary Terms*. Hrsg. von Meyer Howard Abrams. 7. Aufl. Boston: Heinle & Heinle.

Barth, Florian (2019). „Annotation Guideline No. 5: Annotation Guidelines for Narrative Levels and Narrative Acts". In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.056.

Barth, Florian (2020). „Annotation narrativer Ebenen und narrativer Akte". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 423–438.

Bauer, Matthias und Miriam Lahrsow (2020). „Annotation Guideline No. 6: SANTA 6 Collaborative Annotation as a Teaching Tool Between Theory and Practice“. In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.059.

Buchholz, Sabine und Erwin Marsi (2006). „CoNLL-X Shared Task on Multilingual Dependency Parsing“. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City: Association for Computational Linguistics, S. 149–164. URL: http://www.aclweb.org/anthology/W/W06/W06-2920 (besucht am 1. Juni 2020).

Carreras, Xavier und Lluís Màrquez (2004). „Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling“. In: *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA: Association for Computational Linguistics, S. 89–97.

Carreras, Xavier und Lluís Màrquez (2005). „Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling“. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ann Arbor, Michigan: Association for Computational Linguistics, S. 152–164. URL: http://www.aclweb.org/anthology/W/W05/W05-0620 (besucht am 1. Juni 2020).

Dagan, Ido, Oren Glickman und Bernardo Magnini (2006). „The PASCAL Recognising Textual Entailment Challenge“. In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Berlin/Heidelberg: Springer Berlin Heidelberg, S. 177–190.

Doležel, Lubomír (1973). *Narrative Modes in Czech Literature*. Toronto: University of Toronto Press.

Duyfhuizen, Bernard (2005). „Framed Narrative“. In: *The Routledge Encyclopedia of Narrative Theory*. Hrsg. von David Herman. London/New York: Routledge, S. 186–188.

Eisenberg, Joshua und Mark Finlayson (2019). „Annotation Guideline No. 1: Cover Sheet for Narrative Boundaries Annotation Guide“. In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.051.

Escartín, Carla Parra, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way und Chao-Hong Liu (2017). „Ethical Considerations in NLP Shared Tasks“. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, S. 66–73. DOI: 10.18653/v1/W17-1608. (Besucht am 1. Juni 2020).

Fludernik, Monika (2009). *An Introduction to Narratology*. London: Taylor & Francis.

Füredy, Viveca (1989). „A Structural Model of Phenomena with Embedding in Literature and Other Arts“. In: *Poetics Today* 10.4, S. 745–769.

Genette, Gérard (1980). *Narrative Discourse – An Essay in Method*. Übers. von Jane E. Lewin. Ithaca, New York: Cornell University Press.

Genette, Gérard (1988). *Narrative Discourse Revisited*. Ithaca, New York: Cornell University Press.

Genette, Gérard (1997). *Paratexts*. Cambridge: Cambridge University Press.

Gius, Evelyn (2015). *Erzählen über Konflikte – Ein Beitrag zur digitalen Narratologie*. Narratologia. Berlin: De Gruyter.

Gius, Evelyn, Nils Reiter und Marcus Willand, Hrsg. (2019). *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*.

Hammond, Adam (2020). „Annotation Guideline No. 8: Annotation Guidelines for Narrative Levels“. In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/001c.11773.

Jahn, Manfred (2017). „N2.4. Narrative Levels". In: *Narratology: A Guide to the Theory of Narrative*. Hrsg. von Manfred Jahn. Universität zu Köln.

Kearns, Edward (2019). „Annotation Guideline No. 2: For Annotating Anachronies and Narrative Levels in Fiction". In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.052.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Ketschik, Nora, Benjamin Krautter, Sandra Murr und Yvonne Zimmermann (2019). „Annotation Guideline No. 4: Annotating Narrative Levels in Literature". In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.055.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Kilgarriff, Adam und Joseph Rosenzweig (2000). „Framework and Results for English SENSEVAL". In: *Computers and the Humanities* 34.1, S. 15–48. DOI: 10.1023/A:1002693207386. (Besucht am 1. Juni 2020).

Kübler, Sandra (2008). „The PaGe 2008 Shared Task on Parsing German". In: *Proceedings of the Workshop on Parsing German*. Columbus, Ohio: Association for Computational Linguistics, S. 55–63. URL: http://www.aclweb.org/anthology/W/W08/W08-1008 (besucht am 1. Juni 2020).

Lahn, Silke und Jan Christoph Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart, Germany: Metzler.

Lämmert, Eberhard (1955). „Das Gefüge der Handlungsstränge". In: *Bauformen des Erzählens*. Stuttgart: Metzler, S. 43–67.

Mani, Inderjeet (2012). *Computational Modeling of Narrative*. Hrsg. von Graeme Hirst. Bd. 18. Synthesis Lectures on Human Language Technologies. Bristol: Morgan & Claypool. DOI: 10.2200/S00459ED1V01Y201212HLT018.

Margolin, Uri (2014). „Narrator". In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

Martínez, Matias und Michael Scheffel (2003). *Einführung in die Erzähltheorie*. 8. Aufl. München: C.H.Beck, S. 75–80.

Mathet, Yann, Antoine Widlöcher und Jean-Philippe Métivier (2015). „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3, S. 437–479. DOI: 10.1162/COLI_a_00227.

McHale, Brian (2014). „Speech Representation". In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. 2014. Aufl. Hamburg: Hamburg University Press.

Nelles, William (1997). *Frameworks: Narrative Levels and Embedded Narrative*. New York: P. Lang.

Nelles, William (2005). „Embedding". In: *Routledge Encyclopedia of Narrative Theory*. Hrsg. von David Herman, Manfred Jahn und Marie-Laure Ryan. London/New York: Routledge.

Neumann, Birgit und Ansgar Nünning (2012). „Metanarration and Metafiction". In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

Niederhoff, Burkhard (2013). „Focalization“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel und Deniz Yuret (2007). „The CoNLL 2007 Shared Task on Dependency Parsing“. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, Czech Republic: Association for Computational Linguistics, S. 915–932. URL: http://www.aclweb.org/anthology/D/D07/D07-1096 (besucht am 1. Juni 2020).

Nünning, Ansgar (2004). „On Metanarrative: Towards a Definition, a Typology and an Outline of the Functions of Metanarrative Commentary“. In: *The Dynamics of Narrative Form – Studies in Anglo-American Narratology*. Hrsg. von John Pier. Narratologia. Berlin: De Gruyter, S. 11–58.

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Pier, John (2012). „Metalepsis“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

Pier, John (2014). „The Living Handbook of Narratology“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Revised version. Hamburg: Hamburg University Press.

Pier, John (2016). „Narrative Levels“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Revised version. Hamburg: Hamburg University Press.

Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.

Reiter, Nils, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler (2020). „*Reaching out*: Interdisziplinäre Kommunikation und Dissemination“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 467–484.

Rickert, Edith (1923). „Some Straws in Contemporary Literature: Fiction in England and America“. In: *The English Journal* 12.8, S. 599–604.

Rimmon-Kenan, Shlomith (2004). „Narration: Levels and Voices“. In: *Narrative fiction*. Hrsg. von Shlomith Rimmon-Kenan. 2. Aufl. London/New York: Routledge, S. 87–106.

Rimmon-Kenan, Shlomith (2005). *Narrative Fiction: Contemporary Poetics*. London/New York: Routledge.

Romberg, Bertil (1962). *Studies in the Narrative Technique of the First-Person Novel*. Stockholm: Almqvist & Wiksell.

Ryan, Marie-Laure (1986). „Embedded Narratives and Tellability“. In: *Style* 20.3, S. 319–340.

Ryan, Marie-Laure (1991). *Possible worlds, artificial intelligence, and narrative theory*. Bloomington, Indiana: Indiana University Press.

Ryan, Marie-Laure (2001). „The Narratorial Functions: Breaking down a Theoretical Primitive“. In: *Narrative* 9.2, S. 146–152.

Ryan, Marie-Laure (2002). „Stacks, Frames, and Boundaries". In: *Narrative dynamics: essays on time, plot, closure, and frames*. Columbus, Ohio: Ohio State University Press, S. 366–385.

Sang, Erik F. Tjong Kim und Sabine Buchholz (2000). „Introduction to the CoNLL-2000 Shared Task Chunking". In: *Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, S. 127–132. URL: http://www.aclweb.org/anthology/W00-0726 (besucht am 1. Juni 2020).

Sang, Erik F. Tjong Kim und Hervé Déjean (2001). „Introduction to the CoNLL-2001 shared task: clause identification". In: *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*. Toulouse, Frankreich. URL: http://www.aclweb.org/anthology/W01-0708 (besucht am 1. Juni 2020).

Schmid, Wolf (2014). *Elemente der Narratologie*. 3. Aufl. Berlin: De Gruyter.

Sundheim, Beth M. (1993). „The Message Understanding Conferences". In: *Proceedings of the TIPSTER Text Program: Phase I*. Fredericksburg, Virginia, USA: Association for Computational Linguistics, S. 5. DOI: 10.3115/1119149.1119153. (Besucht am 1. Juni 2020).

Sundheim, Beth M. und Nancy A. Chinchor (1993). „Survey of the Message Understanding Conferences". In: *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, S. 56–60. URL: http://www.aclweb.org/anthology/H93-1011 (besucht am 1. Juni 2020).

Tjong Kim Sang, Erik F. und Fien De Meulder (2003). „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, S. 142–147. URL: http://www.aclweb.org/anthology/W03-0419 (besucht am 1. Juni 2020).

Todorov, Tzvetan (1966). „Les catégories du récit littéraire". In: *Communications* 8, S. 125–151.

Wirén, Mats, Adam Ek und Anna Kasaty (2020). „Annotation Guideline No. 7: Guidelines forannotation of narrative structure". In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.060.

Witten, I. H. und Eibe Frank (2005). *Data Mining*. 2. Aufl. Practical Machine Learning Tools and Techniques. Burlington/San Francisco: Morgan Kaufmann Publishers.

Florian Barth

Annotation narrativer Ebenen und narrativer Akte

Zusammenfassung: Der von Genette eingeführte Begriff der *narrativen Ebene* basiert auf der Erzählung einer eingebetteten Geschichte in Verbindung mit einem Erzählerwechsel. Zur Formalisierung des Konzepts hin zu einer Annotation werden in diesem Beitrag Bedingungen und Attribute für Erzählebenen vorgestellt und das korrespondierende Konzept des Narrativs unter dem Begriff *narrativer Akt* eingeordnet. Dabei wird ein besonderes Augenmerk auf die Grenzen zwischen narrativen Ebenen gelegt, die mit der Berücksichtigung eines ontologischen Ebenenwechsels auch über Genettes Konzept hinausgehen.

Abstract: The concept of *narrative levels* introduced by Genette is based on the presence of an embedded story in conjunction with a change of the narrator. To formalize the concept towards an annotation, this paper introduces conditions and attributes for narrative levels and sets it into relation with the corresponding concept of the narrative under the term *narrative act*. Special attention is paid to the boundaries between narrative levels, which, with the consideration of an ontological level change, also go beyond Genette's concept.

1 Vorbemerkung

Im vorliegenden Text werden wesentliche Aspekte der Konzeptualisierung des Begriffs der Erzählebene und des Narrativs (bzw. narrativen Aktes) sowie die damit verbunden Annotationskategorien behandelt. Auf konkrete Annotationsanweisungen sowie Textauszüge wurde verzichtet, weil diese schon bei wenigen Beispielen deutlich mehr Raum einnehmen würden als der Inhalt des Beitrages.

Anmerkung: Dieser Beitrag ist eine Re-Publikation des narratologisch-theoretischen Teils der Annotationsrichtlinien für narrative Ebenen und narrative Akte, die in einer ersten und einer zweiten überarbeiteten Fassung in englischer Sprache vorliegen. Die erste Version der Guidelines wurde bereits in der Zeitschrift *Cultural Analytics* publiziert (Barth 2019), die Publikation der überarbeiteten Fassung erfolgt demnächst.

Florian Barth, Abteilung Digital Humanities, Institut für Literaturwissenschaft, Universität Stuttgart

Open Access. © 2020 Florian Barth, publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-017

2 Konzept

Der von Gérard Genette eingeführte Begriff der *narrativen Ebene* zielt darauf ab, die Beziehungen zwischen einer *eingebetteten Erzählung* (*embedding*)[1] und der Diegese zu beschreiben (Genette 1983, S. 227–231, Pier und Coste 2014) sowie eine klare hierarchische Struktur zwischen diegetischen Ebenen zu etablieren. Genette erklärt explizit seine Absicht, den bestehenden Begriff des *embeddings* zu systematisieren, bei dem seiner Meinung nach eine klare Grenze zur eingeschobenen Erzählung ebenso fehlt wie die Möglichkeit, eine zweite Diegese innerhalb der ersten Diegese hierarchisch zu strukturieren (Genette 1988, S. 88).

In diesem Beitrag werden die direkt auf Erzählebenen bezogenen Begriffe *embedding* und *framed narrative* (*Rahmenerzählung*) unter dem Begriff *narrativer Akt* gruppiert.[2] Da narrative Akte nicht allein in Verbindung mit vertikalen Ebenen[3] auftreten können, trennen wir klar zwischen:

- narrativen Ebenen (kurz: *Ebene*)
- narrativen Akten (kurz: *Narrativ*)

In der vertikalen Struktur kann jede Ebene eine unbegrenzte Anzahl von narrativen Akten enthalten:

- Ebene 1
  - Narrativ 1
  - Narrativ 2
  - ...
  - Narrativ *n*
- Ebene 2
  - Narrativ 1
  - Narrativ 2
  - ...
  - Narrativ *n*
- ...

1 Im deutschsprachigen Raum wird insbesondere auch der Begriff *Binnenerzählung* verwendet (Lahn und Meister 2008, S. 79, Pier und Coste 2014).

2 Gemeinsamkeiten und Abgrenzungen zwischen *embeddings* und *framed narratives* sind in Abschnitt 3.2 im entsprechenden Unterabschnitt angegeben.

3 Nelles beschreibt die Möglichkeit von horizontal angeordneten *embeddings*, vgl. Abschnitt 3.2 und Nelles (1997, S. 132).

– Ebene *n*

Um eine singuläre *Kategorie* für die Annotation bereitzustellen, wurden beide Konzepte kombiniert:

– Ebene 1 Narrativ 1
– Ebene 1 Narrativ 2
– ...
– Ebene 2 Narrativ 1
– Ebene 2 Narrativ 2
– ...
– Ebene *n* Narrativ *n*

Ergänzend zur Kategorie werden *Attribute* vergeben, die zusätzliche Informationen über die Art der Ebenengrenze, die Identität des Erzählers und seine Präsenz in der Geschichte sowie metafiktionale Elemente oder metaleptische Eingriffe erfassen.

Wenn der Erzähler einer bestimmten Ebene Teil der Geschichte ist, kann das Attribut für seine ‚diegetische Präsenz' beispielsweise auf ‚homodiegetisch' gesetzt werden – wenn er nicht Teil der Geschichte ist, sollte es ‚heterodiegetisch' sein:

– Ebene 1 Narrativ 1 | diegetische Präsenz des Erzählers: heterodiegetisch
– Ebene 1 Narrativ 2 | diegetische Präsenz des Erzählers: heterodiegetisch
– ...
– Ebene 2 Narrativ 1 | diegetische Präsenz des Erzählers: homodiegetisch
– Ebene 2 Narrativ 2 | diegetische Präsenz des Erzählers: homodiegetisch
– ...
– Ebene *n* Narrativ *n*

3 Kategorie der Annotation

Obwohl Ebenen und Narrative in einer Annotationskategorie zusammengefasst sind, beschreiben wir den theoretischen Hintergrund in den getrennten Unterabschnitten 3.1 und 3.2, während in 3.3 das Verhältnis zwischen beiden näher beleuchtet wird. Abschnitt 4 geht auf die Attribute ein, die der Kategorie zugeordnet sind.

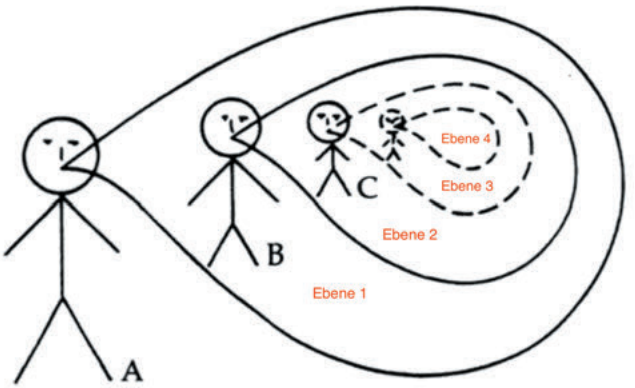


Abb. 1: Narrative Ebenen in Verbindung mit Sprechakten nach Genette (2010, S. 226)

3.1 Narrative Ebenen

In der Regel entstehen Erzählebenen, wenn eine Figur innerhalb einer Erzählung beginnt, eine eigene Geschichte zu berichten, wodurch ein narrativer Akt innerhalb eines narrativen Aktes erzeugt wird (Lahn und Meister 2008, S. 83, Jahn 2005). Der Wechsel des *Erzählers* ist das grundlegendste Merkmal narrativer Ebenen und in Genettes Theorie obligatorisch, wobei für jeden narrativen Akt auf einer bestimmten Ebene ein anderer Erzähler auftritt (vgl. Jahn 2005). In Genettes Konzept ist die narrative Instanz einer ersten Ebene (Erzähler A in Abbildung 1) „per definitionem extradiegetisch" (Genette 1998) und seine Geschichte auf Ebene 1 diegetisch. Ein intradiegetischer Erzähler (B) erzählt dann eine metadiegetische Geschichte (Ebene 2), ein metadiegetischer Erzähler (C) eine metametadiegetische Erzählung (Ebene 3) usw. Wie oben gesehen, bezeichnen wir die Ebene nur durch eine Nummer und für die Identität des Erzählers setzen wir eine eindeutige ID (vgl. Abschnitt 4, „Erzähler: Identität").

Ein häufig herangezogenes Beispiel für einen Erzählerwechsel ist Joseph Conrads Roman *Herz der Finsternis*, in dem ein namenloser Seemann seine Eindrücke einer Bootsfahrt auf der Themse in Richtung London schildert. Innerhalb dieser kurzen Erzählung auf erster Ebene schildert die Figur des Marlow auf einer zweiten Ebene die Geschichte einer Reise auf dem Kongo-Fluss und erst am Ende des Romans kehrt die Erzählung in einigen kurzen Sätzen zur Perspektive des unbenannten Seemanns zurück. Diese Struktur ist in Abbildung 2 dargestellt.

Genette verwendet den Begriff ‚Erzähler' für seine Vorstellung der *Stimme* einer Erzählung (Jahn 2005). Andere, wie Marie-Laure Ryan (Ryan 1991) oder Manfred Jahn (Jahn 2005), nutzen zusätzlich den Begriff *Sprecher* für die erzählerische

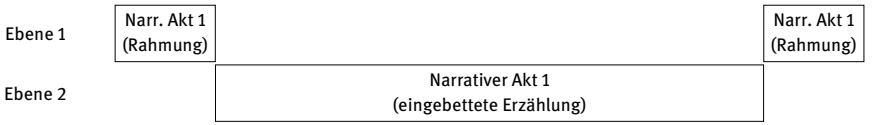


Abb. 2: Narrative Akte auf zwei Ebenen in Joseph Conrads *Herz der Finsternis*

Instanz, welche die Sprecher-Adressaten-Beziehung innerhalb eines Kommunikationsmodells zwischen dem Erzähler und dem Leser oder der Leserin eines literarischen Textes akzentuiert (Jakobson 1960; Banfield 1973). Da auch Ryan in einer späteren Studie die Bedeutung und spezielle Verwendung des Begriffs ‚Erzähler' hervorhebt (Ryan 2001), bleiben wir in unserem Konzept bei diesem Begriff und verwenden ‚Sprecher' nur dann, wenn in spezifischen narratologischen Theorien vom letzteren Terminus Gebrauch gemacht wird.

3.1.1 Grenzen zwischen narrativen Ebenen

Marie-Laure Ryan konzentriert sich in ihrem Ebenenkonzept auf die Art der Grenze zwischen den Erzählebenen, weshalb sie den Wechsel der Sprecher (Erzähler) als eine *illokutionäre Grenze* beschreibt. Diese Grenze kann *tatsächlich überschritten* werden, wenn eine neue Stimme, z. B. eine Figur, eine Geschichte auf der zweiten Ebene innerhalb eines direkten Sprechaktes wie in *Herz der Finsternis* berichtet (Ryan 1991, S. 176).

Darüber hinaus können Ebenengrenzen *virtuell überschritten* werden, wenn Äußerungen von Figuren durch den Erzähler im indirekten Diskurs präsentiert werden – wie etwa bei Gedanken einer Figur oder bei indirekter Rede (Ryan 1991, S. 176).

Weiterhin betont Ryan, dass Ebenen nicht nur durch den Wechsel der Sprecher (Erzähler) entstehen, sondern auch, wenn ein neues ‚Realitäts-System' eingeführt wird, wie in Lewis Carrolls *Alice im Wunderland*, wo sich der Wechsel der primären Realität der Alltagswelt in die Traumwelt des Wunderlandes in einem kontinuierlichen Sprechakt vollzieht (Ryan 1991, S. 177). Damit wird ein völlig neues Konzept für die Konstitution von Erzählebenen etabliert, das Ryan als Überschreitung einer *ontologischen Grenze* definiert.

Während *Alice im Wunderland* die tatsächliche Überschreitung einer ontologischen Grenze beinhaltet (die fiktiven Charaktere betreten tatsächlich eine andere Realitätsebene), erfolgt eine virtuelle Überschreitung ontologischer Grenzen, wenn die zweite Realität in der primären ‚verankert' ist, z. B. bei der Beschreibung der Handlung eines Films aus der Perspektive der primären Erzählebene

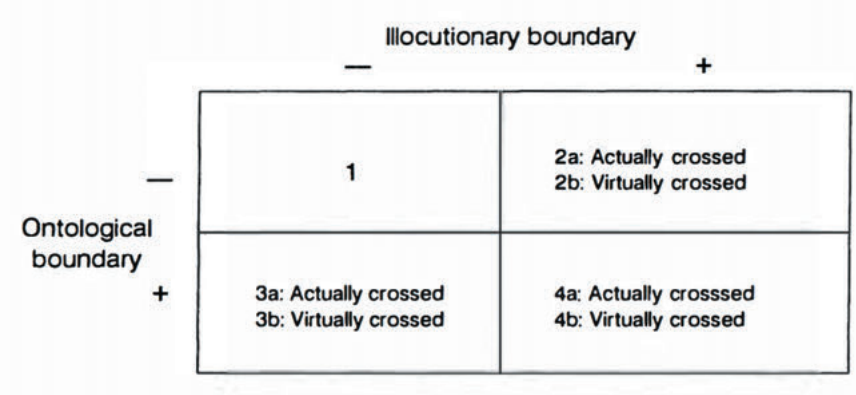


Abb. 3: Grenzen zwischen narrativen Ebenen nach Ryan (1991, S. 176)

(Ryan 1991, S. 177). Eine ontologische Grenze wird ebenfalls virtuell überschritten, wenn der Erzähler der ersten Ebene einen existierenden fiktionalen Text zitiert, wie in *Stiller* von Max Frisch, worin Washington Irvings Erzählung *Rip van Winkle* vorgetragen wird. Illokutionäre und ontologische Grenzen können kombiniert auftreten und Ryan liefert eine tabellarische Darstellung für sechs mögliche Kombinationen (Abbildung 3).

Ryan bestimmt eine tatsächlich überschrittene illokutionäre und ontologische Grenze (4a in Abbildung 3) als „fiction within a fiction" und nennt als Beispiele die Geschichten, welche die intradiegetische Erzählerin Scheherazade in *Tausendundeine Nacht* berichtet (Ryan 1991, S. 177).

Ein weiteres Beispiel für diese doppelte Grenzüberschreitung ist der Prolog eines fiktiven Herausgebers, der im narrativen Akt zweiter Ebene nicht vorkommt. Dies geschieht in Boccaccios *Dekameron*, wo ein Prolog den Rahmen für die 100 Geschichten bildet, die von sieben Frauen und drei Männern erzählt werden. Eine solche Präsentationstechnik kommt insbesondere in Briefromanen zum Einsatz und wird in der deutschen Literaturwissenschaft als *Herausgeberfiktion* bezeichnet (Wirth 2008; auch Ryan 2001, S. 151). Für die Kombination beider Grenzen (illokutionär und ontologisch) ist es notwendig, dass der Herausgeber nicht Teil der Geschichte ist. Ansonsten handelt es sich lediglich um eine illokutionäre Grenze wie in Goethes *Die Leiden des jungen Werthers*, wo der Herausgeber zumindest behauptet, die Hauptfigur zu kennen und ihre Handlungen in der diegetischen Welt beobachtet zu haben.

Eine virtuelle Überschreitung beider Grenzen (4b) bezieht sich dagegen auf eine Beschreibung einer metafiktionalen Geschichte aus der Perspektive des Erzählers der ersten Ebene, bei der ein Erzählerwechsel auf zweiter Ebene angedeutet,

aber nicht tatsächlich vollzogen wird (Ryan 1991, S. 177). Diese seltene Konstellation tritt in *Thema vom Verräter und vom Helden* von Jorge Luis Borges auf, wo der primäre Erzähler seinen Plan darlegt, eine Geschichte zu schreiben, deren Erzähler ‚Ryan' sein wird, wobei der Erzähler der ersten Ebene aber nie tatsächlich als ‚Ryan' spricht (Ryan 1991, S. 177).

Marie-Laure Ryan weist auch darauf hin, dass jede Äußerung einer neuen Stimme ein eigenes semantisches Universum kreiert, das möglicherweise von der primären Realität der Erzählung abweicht und potenziell eine neue Erzählebene etablieren kann (Ryan 1991, S. 176–177). Aus unserer Sicht stellt dies jedoch eine Mischung zwischen beiden Grenz-Typen dar, was für eine eindeutige definitorische Darlegung ungeeignet erscheint. Stattdessen unterscheiden wir illokutionäre und ontologische Grenzen klar voneinander, auch wenn sie kombiniert auftreten können. Ein neuer Sprechakt markiert dagegen nur dann eine illokutionäre Grenze, die zu einer untergeordneten Ebene führt, wenn er tatsächlich einen neuen narrativen Akt erzeugt. Das heißt, nur wenn eine Figur auf einer bestimmten Ebene direkt (tatsächliche Grenzüberschreitung) oder indirekt (virtuelle Grenzüberschreitung) eine neue Geschichte erzählt, eröffnen ihre Äußerungen eine neue Ebene.

3.2 Narrative Akte

Wie Didier Coste in einem neueren Beitrag feststellt, war der Begriff des ‚Narrativs' stets umstritten (Coste 2017, S. 3). Er hebt die großen Unterschiede literaturwissenschaftlicher Ansätze hervor, unter denen die vermeintlich ‚klassisch'-formalistischen und strukturalistischen Konzepte zunächst durch eine Diversifizierung in einer sogenannten ‚post-klassischen' Phase und darüber hinaus durch den Aufstieg kognitiver Theorien und den Einfluss der Neurowissenschaften abgelöst wurden (Coste 2017, S. 3). Während die Definitionen des Narrativs innerhalb der Literaturwissenschaft bereits variieren, bleiben sie zudem oft unvereinbar mit anderen Wissensbereichen wie der Linguistik (Coste 2017, S. 4). Coste weist darauf hin, dass die Theoriebildung zum Narrativ stark von historischen Umständen, bestimmten ideologischen Positionen und vorab getroffenen theoretischen Festlegungen beeinflusst wird (Coste 2017, S. 4). Gerade in Hinblick auf letztere müssen wir uns bewusst sein, dass jegliche definitorische Eingrenzung eines Narrativs (bzw. narrativen Aktes) in diesem Beitrag durch den übergeordneten Zweck, der Beschreibung und Erfassung narrativer Ebenen, bestimmt wird.

Vor diesem Hintergrund streben wir eine Definition von narrativen Akten an, die einerseits textuell überprüfbare Elemente enthält und andererseits die Intui-

tion dafür schärft, was ein Narrativ vor dem Hintergrund einer Ebenenstruktur darstellt.

Eberhard Lämmert hält in einem frühen Ansatz fest, dass zur Eröffnung eines übergeordneten oder horizontal angeordneten narrativen Aktes ein Wechsel von Zeit, Schauplatz oder Charakteren vorliegen müsse (Lämmert 1955, S. 44). Beispiele hierfür haben wir bereits diskutiert: Wenn Marlow in *Herz der Finsternis* seine Erzählung auf einer zweiten Ebene beginnt, bleibt nur dieser Charakter konstant – andere Figuren sowie Zeit und Handlungsort haben sich geändert. In Kleists Kurzgeschichte *Unwahrscheinliche Wahrhaftigkeiten* unterscheiden sich die drei untergeordneten Geschichten (Abbildung 4), die von einem Erzähler der ersten Ebene berichtet werden, in Zeit, Handlungsort und den Charakteren (mit Ausnahme des intradiegetischen Erzählers). Dies verdeutlicht den Status der drei Erzählungen als eigenständige narrative Akte auf Ebene 2 (vgl. dazu auch den nächsten Unterabschnitt „Embedding und Framing").

Es ist zu ergänzen, dass auch Ereignisse auf einen neuen narrativen Akt hinweisen können. Peter Hühn betont, die narratologische Kategorie des *Ereignisses* korrespondiere mit bestimmten Definitionen des Narrativs, insbesondere mit einer Vorstellung des Erzählvorgangs als Zustandsveränderung unterschiedlicher Ausprägung (Hühn 2009, S. 80). Daher wurde bei der Annotation auf unerwartete Veränderungen geachtet, die mit einem Ereignis korrespondieren, das zur Etablierung eines narrativen Aktes beiträgt. Wenn Alice beispielsweise bemerkt, dass ein Kaninchen sprechen kann und eine Uhr aus der Westentasche nimmt, markiert dies ein ungewöhnliches, metaleptisches Ereignis, das den Beginn eines darauf folgenden untergeordneten narrativen Aktes einläutet, der von einer ontologischen Grenze umschlossen ist. Bei der Identifizierung solcher ungewöhnlicher Ereignisse handelt es sich bereits um einen Interpretationsprozess, der über das Erkennen von textuellen Artefakten hinausgeht.[4]

Beide Aspekte, sowohl Entitäten wie Zeit, Raum oder Charaktere, als auch interpretative Komponenten, wie Ereignisse, sind Indizien für einen neuen narrativen Akt, aber keine Voraussetzung dafür. Bei der Annotation wurde generell auf erzählerische Veränderungen geachtet: Ein narrativer Akt muss sich von dem übergeordneten bzw. horizontal beigeordneten Akt in Bezug auf seinen Erzähler oder die diegetische Welt unterscheiden und es muss zwischen diesen verschiedenen narrativen Akten eine punktuelle verifizierbare Grenze geben, wie sie von Ryan definiert wird.

4 Hühn beschreibt ungewöhnliche und unerwartete Veränderungen innerhalb seines Begriffs *Ereignis II*, während *Ereignis I* jede Zustandsänderung beinhaltet, die explizit oder implizit in einem Text dargestellt wird (Hühn 2009, S. 80). Zur Erkennung der Grenzen zwischen narrativen Akten ist insbesondere *Ereignis II* relevant.

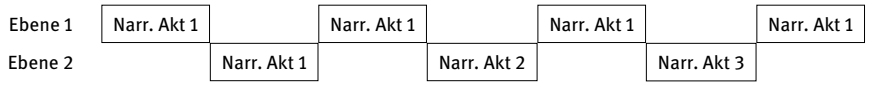


Abb. 4: Multiple *embeddings* unabhängiger narrativer Akte in Kleists *Unwahrscheinliche Wahrhaftigkeiten*

Embedding und *framing* von narrativen Akten

Wie zu Beginn festgehalten, beinhaltet der Begriff des narrativen Aktes sowohl *embeddings* (eingebettete Erzählungen) als auch *framed narratives* (Rahmenerzählungen). *Framing* stellt eher eine Präsentationstechnik dar, bei der die meist kurze Rahmenerzählung eine umfassendere untergeordnete Geschichte einschließt (Pier und Coste 2014). Ein Beispiel ist der oben genannte Roman *Herz der Finsternis*, für welchen die Rahmenhandlung und eingebettete Erzählung in Abbildung 2 dargestellt sind.

Dagegen können *embeddings* als kleinere Einfügungen innerhalb einer größeren Einheit angesehen werden (Pier und Coste 2014), z. B. erzählt in der kurz zuvor erwähnten Kurzgeschichte *Unwahrscheinliche Wahrhaftigkeiten* ein Offizier drei Geschichten, die als eigenständige Erzählakte auf der zweiten Ebene angesiedelt sind. In Abbildung 4 ist zu sehen, dass für jede Geschichte auf Ebene 2 ein neuer narrativer Akt eröffnet wird. In unserem Konzept zählen wir die narrativen Akte auf jeder Ebene separat. Die narrativen Akte 1 bis 3 auf der zweiten Ebene stellen die eingebetteten Geschichten dar, während der narrative Akt 1 auf der ersten Ebene die Rahmenhandlung markiert, in welcher der Offizier diese Geschichten erzählt.

In der Praxis dominiert oftmals weder eine eingebettete Erzählung noch eine Rahmenerzählung, weshalb die Annotation nicht darauf abzielte, *framing*- oder *embedding*-Techniken, deren spezifische Funktion (Rimmon-Kenan 2005, S. 95; Lahn und Meister 2008, S. 87–90) oder eine bestimmte ‚Haupterzählung' innerhalb mehrerer verschachtelter narrativer Akte zu identifizieren (Gius 2015, S. 164).

Horizontal eingebettete narrative Akte ohne Ebenenwechsel

Im Gegensatz zur vertikalen Anordnung von narrativen Akten zwischen Erzählebenen beschreibt Nelles die Existenz von horizontal eingebetteten Narrativen auf

Ebene 1	Narr. Akt 1 Brief der Figur ‚Herz‘	Narr. Akt 2 Brief der Figur ‚Schatouilleuse‘	Narr. Akt 1 Brief der Figur ‚Herz‘	Narr. Akt 2 Brief der Figur ‚Schatouilleuse‘	Narr. Akt 3 Brief der Figur ‚Rothe‘

Abb. 5: Die ersten fünf Briefe aus dem Briefroman *Der Waldbruder*

der ersten Ebene (Nelles 1997, S. 132).[5] Dies geschieht, wenn narrative Akte verschiedener Sprecher nebeneinander ohne einen übergeordneten Erzähler präsentiert werden. In J. M. R. Lenz' Briefroman *Der Waldbruder* werden beispielsweise mehrere Briefe durch wechselnde Charaktere auf der gleichen diegetischen Ebene präsentiert. Da diese Figuren nicht direkt miteinander korrespondieren, kann jeder Brief als Teil eines eigenständigen narrativen Aktes angesehen werden, in dem die spezifische Perspektive der jeweiligen Figur auf das erzählte Geschehen entwickelt wird (siehe Abbildung 5).

Pierre und Coste erfassen zudem gesammelte Geschichten ohne übergeordneten Erzähler als horizontal eingebettete narrative Akte ohne Ebenenwechsel (Pier und Coste 2014). Ein Beispiel dafür wäre Günter Grass' Roman *Mein Jahrhundert*, der jeweils eine Kurzgeschichte für jedes Jahr des Jahrhunderts enthält, aber im Gegensatz zum *Dekameron* gibt es keinen übergeordneten Erzähler. Dennoch sind Grass' Kurzgeschichten miteinander verbunden, z. B. bleibt in den Jahren beider Weltkriege der Erzähler konstant, berichtet aber verschiedene Episoden. Und darüber hinaus scheint das ganze Werk durch eine übergeordnete Instanz arrangiert, die am ehesten mit einem impliziten Autor zu vergleichen ist (Booth 2010).

Weiterhin beschreiben Pier und Coste *Abschweifung* (*digression*) als eine Form der Einbettung ohne Wechsel der Ebene (Pier und Coste 2014). Dazu gehört auch ein *Exkurs*, z. B. wenn der Erzähler den Leser direkt anspricht. Abschweifungstechniken finden sich ebenfalls auf der gleichen Ebene wie der narrative Akt, in den sie eingeschoben sind.

Zudem korrespondiert ein Exkurs oft mit Metanarration und Metafiktion, die jeweils mit einem zusätzlichen Attribut erfasst werden (siehe „Metanarration und Metafiktion“ in Abschnitt 4). Ein treffendes Beispiel hierfür ist Laurence Sternes Roman *Leben und Ansichten von Tristram Shandy, Gentleman*, wo der Erzähler nicht nur Exkurse unterschiedlicher Art unternimmt, sondern diese auch in metafiktionalen Textpassagen reflektiert.

5 Nelles definiert auch den Begriff des *modal embedding* für Traumwelten (Nelles 2010). Im Gegensatz zu Ryan sieht er darin keinen Ebenenwechsel, obwohl er eine Verschiebung der ‚Realität' der fiktiven Welt feststellt. Für unsere Annotation wurde jedoch die Annahme einer untergeordneten Ebene bei Überschreitung einer ontologischen Grenze beibehalten.

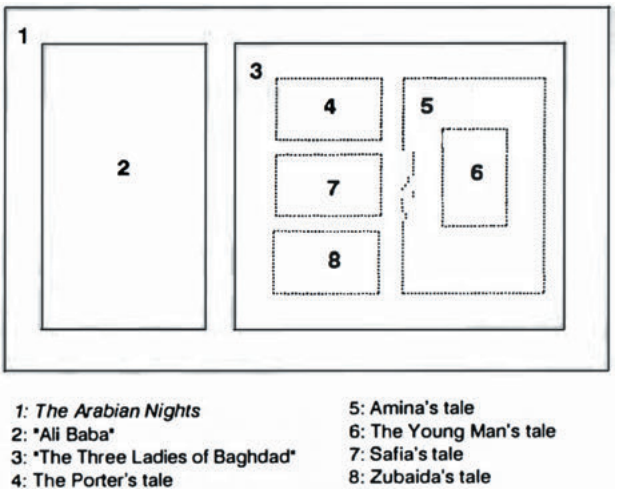


Abb. 6: Eingebettete Erzählungen in *Tausendundeine Nacht* nach Ryan (1991, S. 178)

3.3 Verhältnis von narrativen Ebenen und narrativen Akten

Das Hauptaugenmerk der Annotation lag auf der Bestimmung des Verhältnisses zwischen den narrativen Akten innerhalb ihrer spezifischen, vertikalen Ebenenstruktur bzw. einer horizontalen Anordnung. Dabei können narrative Akte mehrfach eingebettet sein und auf jeder Ebene sind mehrere unabhängige narrative Akte möglich.

Zusätzlich zur Darstellung der Ebenenstruktur in Abschnitt 2 veranschaulicht Ryan die Einbettung und Stapelung von narrativen Akten anhand von *Tausendundeine Nacht*, wobei der rahmende narrative Akt mit Scheherazade und dem Sultan direkt die Geschichten und gleichnamigen Unterkapitel von „Ali Baba“ und „Der Träger und die drei Damen“ (vgl. die englische Kapitelüberschrift „The Three Ladies of Baghdad“ in Abbildung 6) einschließt, die von Scheherazade auf Ebene 2 erzählt werden (Ryan 1991, S. 178; Ryan 2002, S. 880). Darüber hinaus enthält letztere Geschichte mehrere unabhängige narrative Akte auf Ebene 3 (Geschichten 4, 5, 7, 8 in Abbildung 6) darunter die Erzählung Aminas (Amina's tale), welche auch die Geschichte des jungen Mannes (The young Man's Tale) auf Ebene 4 umfasst (vgl. Abbildung 7; Ryan 1991, S. 178; Ryan 2002, S. 880).

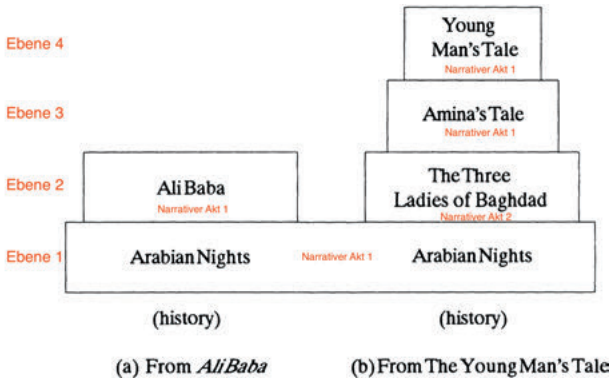


Abb. 7: Stapelung narrativer Ebenen in *Tausendundeine Nacht* nach Ryan (1991, S. 182)

4 Attribute der Annotation

Wie eingangs erwähnt, erfassen die Attribute zusätzliche Informationen über die Annotationskategorie. Dies hilft den Annotatoren, ihre Klassifizierungsentscheidung zu reflektieren, z. B. durch die Bestimmung des Typs der Ebenengrenze oder durch die Bestimmung der diegetischen Präsenz des Erzählers. Zudem liefern die Attribute eine Datengrundlage für die Entwicklung von Features zur automatischen Klassifizierung von Erzählebenen und narrativen Akten.

Erzähler: Identität

Da gestapelte Narrative mehrere Erzähler haben können, annotieren wir die Identität jedes einzelnen. Dies geschieht durch alphabetische IDs für jede Erzähler-Identität:[6]

- Erzähler-Identität a
- Erzähler-Identität b
- Erzähler-Identität c
- ...
- Erzähler-Identität n

6 Wir verwenden nicht Genettes Terminologie (extradiegetisch, intradiegetisch, metadiegetisch), da sie nur die Ebene eines Erzählers bezeichnet, nicht seine Identität.

Zum Beispiel wechselt in Mary Shelleys *Frankenstein oder Der moderne Prometheus* der Erzähler auf fast jeder Ebene: Robert Walton erzählt in seinem Tagebuch von der Begegnung mit Victor Frankenstein und zitiert die mündliche Erzählung Frankensteins, der die metadiegetische Erzählung seiner Kreatur zitiert (Duyfhuizen 2005, S. 187).

Wenn der Erzähler dagegen zwischen den Ebenen konstant bleibt, wird er als die gleiche Erzähler-Identität annotiert, z. B. berichtet die Kreatur innerhalb ihrer metadiegetischen Erzählung auf einer untergeordneten Ebene die Geschichte der Familie De Lacey.

Erzähler: Diegetische Präsenz

Wie in Abschnitt 2 beschrieben, erfasst dieses Attribut, ob ein Sprecher in der Geschichte anwesend ist oder nicht. Wir verwenden die von Genette definierten Begriffe:

- homodiegetisch: Der Erzähler ist Teil der Diegese.
- heterodiegetisch: Der Erzähler ist nicht Teil der Diegese.

Text-Typ

Für jeden narrativen Akt wurde eine Textart angegeben. Vordefiniert sind neun Möglichkeiten, die Annotatoren konnten aber fehlende Typen hinzuzufügen.

- undefiniert (dies trifft auf die meisten extradiegetischen Erzähler auf Ebene 1 zu)
- direkte Rede (wie in *Herz der Finsternis* in Abbildung 2)
- indirekte Äußerungen (z. B. Gedanken einer Figur oder eine transkribierte Rede)
- Zitat eines literarischen Werkes (wie der Vortrag der Erzählung *Rip van Winkle* in Max Frischs *Stiller*)
- Brief (z. B. die Briefe in *Der Waldbruder*, siehe Abbildung 5
- Notizen (dies umfasst Notizen oder Artikel ohne eindeutigen Adressaten wie in den Aufzeichnungen Stillers im gleichnamigen Roman)
- Prolog
- Apolog (z. B. ein Gleichnis am Ende einer Fabel)

Ebenengrenze: Typ

Dieses Attribut kennzeichnet die Art der Ebenengrenze zwischen einem narrativen Akt und der übergeordneten Erzählung nach Ryan (vgl. Abbildung 3):

- illokutionäre Grenze, tatsächlich überschritten
- illokutionäre Grenze, virtuell überschritten
- ontologische Grenze, tatsächlich überschritten
- ontologische Grenze, virtuell überschritten

Ebenengrenze: Übergeordnetes Narrativ

Zur eindeutigen Zuordnung des übergeordneten narrativen Aktes wird dieser als Attribut festgehalten. Zum Beispiel ist in *Tausendundeine Nacht* die Geschichte „Der Träger und die drei Damen" („The Three Ladies of Baghdad" in Abbildung 7) der Erzählung von Amina (Amina's Tale) übergeordnet und stellt den narrativen Akt 2 auf Ebene 2 dar (siehe Abbildung 7).

Metanarration und Metafiktion

Sowohl Metanarration als auch Metafiktion beinhalten selbstreflexive Äußerungen des Erzählers. Während Metanarration Reflexionen über den Prozess des Erzählens umfasst, stellen metafiktionale Passagen eher Kommentare zur Fiktionalität und/oder Konstruiertheit der Erzählung dar (Neumann und Nünning 2015).

Wie in Abschnitt 3.2 festgestellt, sind Metanarration und Metafiktion auf der gleichen Erzählebene angesiedelt, in die sie eingeschoben sind – jedoch stellen sie einen eigenen narrativen Akt dar. Diese narrativen Akte erhalten das entsprechende Attribut (‚Metanarration' oder ‚Metafiktion').

Metanarration kommt beispielsweise in Italo Calvinos *Wenn ein Reisender in einer Winternacht* vor, wo der Erzähler den Leseprozess in der zweiten Person beschreibt. Jedes Kapitel enthält eine andere Version davon, wie der Roman hätte geschrieben werden können, aber keine dieser Geschichten wird beendet.

Metaleptischer Einfluss

Wir erfassen metaleptische Einflüsse durch die über- oder untergeordnete narrative Ebene (Lahn und Meister 2008, S. 90). Wenn ein Charakter aus Ebene 2 in

einem Erzählakt auf Ebene 1 auftaucht, indem er ontologische Grenzen verletzt, kennzeichnen wir die Ebene und den narrativen Akt, von welcher der metaleptische Einfluss ausgeht.[7]

Eine Metalepse wurde bereits anhand des Beispiels aus *Alice im Wunderland* in Verbindung mit der ontologischen Ebenengrenze (Abschnitt 3.1) und der Diskussion ungewöhnlicher Ereignisse (Abschnitt 3.2) aufgezeigt: Wenn Alice ein sprechendes Kaninchen trifft, markiert dies ein metaleptisches Eindringen einer Figur aus der zweiten narrativen Ebene, die einen anderen ontologischen Status hat und von der primären Erzählebene abweichenden Gesetzmäßigkeiten folgt.

Literatur

Banfield, Ann (1973). „Narrative Style and the Grammar of Direct and Indirect Speech". In: *Foundations of Language* 10.1, S. 1–39.

Barth, Florian (2019). „Annotation Guideline No. 5: Annotation Guidelines for Narrative Levels and Narrative Acts". In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.056.

Booth, Wayne C (2010). *The rhetoric of fiction*. University of Chicago Press: Chicago.

Coste, Didier (2017). *Narrative Theory and Aesthetics in Literature*. Oxford. DOI: 10 . 1093 / acrefore/9780190201098.013.116.

Duyfhuizen, Bernard (2005). „Framed narrative". In: *Routledge encyclopedia of narrative theory*, S. 186–188.

Genette, Gérard (1983). *Narrative discourse: An essay in method*. Ithaca, New York: Cornell University Press.

Genette, Gérard (1988). *Narrative Discourse Revisited, translated by Jane E. Lewin*. Ithaca, New York.

Genette, Gérard (1998). *Die Erzählung*. München: W. Fink.

Genette, Gérard (2010). *Die Erzählung. 3. durchgesehene und korrigierte Auflage*. Paderborn: W. Fink.

Gius, Evelyn (2015). *Erzählen über Konflikte: Ein Beitrag zur digitalen Narratologie*. Bd. 46. Walter de Gruyter GmbH & Co KG.

Hühn, Peter (2009). „Event and eventfulness". In: *Handbook of narratology*. Hrsg. von Peter Hühn, Pier, John, Wolf Schmid und Jörg Schönert. Bd. 19. Berlin: De Gruyter, S. 80.

Jahn, Manfred (2005). „Narratology: A guide to the theory of narrative". Webpage. English Department, University of Cologne.

7 Ursprünglich beinhaltet Genettes Konzept der Metalepse jegliches Eindringen eines extradiegetischen Erzählers in die diegetische Welt (Genette 1983, S. 234–235). Wenn zum Beispiel zwei intradiegietische Figuren auf Ebene 1 über den Erzähler sprechen, der die Geschichte schreibt (wie in Flann O'Brians *Auf Schwimmen-zwei-Vögel*), bezieht sich dies auf den extradiegetischen Standpunkt des Erzählers und wird in unseren Richtlinien durch das Setzen des Attributwertes ‚Metanarration' erfasst (siehe oben).

Jakobson, Roman (1960). „Linguistics and poetics“. In: *Style in language*. Cambridge, Massachusetts: MIT Press, S. 350–377.

Lahn, Silke und Jan Christoph Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart: Metzler.

Lämmert, Eberhard (1955). *Bauformen des Erzählens*. Stuttgart: Metzler.

Nelles, William (1997). *Frameworks: Narrative levels and embedded narrative*. Bd. 33. New York: P. Lang.

Nelles, William (2010). „Embedding“. In: *Routledge Encyclopedia of Narrative Theory*, S. 134–135.

Neumann, Birgit und Ansgar Nünning (2015). „Metanarration and metafiction“. In: *Handbook of Narratology*, S. 204–211.

Pier, John und Didier Coste (2014). „Narrative Levels (revised version)“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

Rimmon-Kenan, Shlomith (2005). *Narrative fiction: Contemporary poetics*. London: Taylor & Francis.

Ryan, Marie-Laure (1991). *Possible worlds, artificial intelligence, and narrative theory*. Bloomington, Indiana: Indiana University Press.

Ryan, Marie-Laure (2001). *Narrative as virtual reality: Immersion and interactivity in literature and electronic media*. Baltimore, Maryland: Johns Hopkins University Press.

Ryan, Marie-Laure (2002). „Stacks, Frames, and Boundaries“. In: *Narrative Dynamics: Essays on Time, Plot, Closure, and Frames*. Hrsg. von Brian Richardson. Columbus, Ohio: Ohio State University Press, S. 366.

Wirth, Uwe (2008). *Die Geburt des Autors aus dem Geist der Herausgeberfiktion: Editoriale Rahmung im Roman um 1800: Wieland, Goethe, Brentano, Jean Paul und ETA Hoffmann*. München: Wilhelm Fink Verlag.

Nora Ketschik, Benjamin Krautter, Sandra Murr und
Yvonne Zimmermann

Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext

Zusammenfassung: Der vorliegende Beitrag ist im Rahmen eines Shared Task zur Analyse von Erzählebenen durch Annotation *(SANTA)* entstanden und beschäftigt sich mit der formalen Erfassung des narratologischen Konzepts ‚Erzählebene'. Der Beitrag setzt sich zunächst mit der Theorie von Erzählebenen in der Literaturwissenschaft auseinander, leitet daraus Merkmale für die Erkennung von Erzählebenen ab und entwickelt schließlich Richtlinien für ihre Annotation. Eine wesentliche Erkenntnis der definitorischen Arbeit liegt in der Verknüpfung des Konzepts ‚Erzählebene' mit der Rolle des Erzählers. Indem unterschiedliche Erzählertypen identifiziert werden, können verschiedene Szenarien für die Entstehung neuer Erzählebenen aufgezeigt und kategorisiert werden. Hierbei werden nicht nur prototypische Fälle behandelt, sondern auch seltene und problematische Fälle berücksichtigt. Unser Ziel besteht darin, sowohl eine reflektierte theoretische Auseinandersetzung mit Erzählebenen zu leisten als auch eine akkurate Erfassung des Phänomens zu ermöglichen. Unser methodisches Vorgehen – die Annäherung an ein Phänomen durch Annotation – hat sich dafür als äußert ertragreich erwiesen.

Abstract: This contribution was written in the context of the *Shared Task on the Analysis of Narrative Levels through Annotation* (SANTA) and deals with the formalization of the narratological concept of 'narrative level'. Firstly, we discuss the theory of narrative levels in literary studies. Secondly, we derive features for the identification of narrative levels and finally, we develop guidelines for their annotation. An essential finding of the theoretical work lies in connecting the concept of 'narrative level' to the narrator. By identifying different types of narrators, we are able to enumerate and categorize different scenarios for the emergence of new levels in narrative texts. Hereby, the article does not remain restricted to prototyp-

Anmerkung: Bei dem hier vorliegenden Artikel handelt es sich um eine Überarbeitung und Übersetzung der zweiten Version unserer Guidelines, mit denen wir am *Shared Task on Analysis of Narrative Levels through Annotation* (SANTA) teilgenommen haben. Sie wird derzeit für eine Publikation im Journal *Cultural Analytics* vorbereitet.

Nora Ketschik, Yvonne Zimmermann, Institut für Literaturwissenschaft, Universität Stuttgart
Sandra Murr, Stuttgart Research Center for Text Studies, Universität Stuttgart
Benjamin Krautter, Germanistisches Seminar, Universität Heidelberg

Open Access. © 2020 Nora Ketschik, Benjamin Krautter, Sandra Murr und Yvonne Zimmermann; publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-018

ical examples, but also deals with rare and problematic cases. Overall, our goal is both to provide a theoretical reflection on narrative levels and to create accurate guidelines for its recognition. Our methodological approach, i. e. addressing the phenomenon through annotation, has proven to be extremely fruitful.

1 Vorbemerkung

Im Rahmen einer Ausschreibung zur digitalen Annotation von Erzählebenen, die im Frühjahr 2018 unter dem Titel *Shared Task on the Analysis of Narrative Levels Through Annotation* veröffentlicht wurde, haben wir uns intensiv mit der theoretischen und praktischen Erfassung von Erzählebenen beschäftigt. Die Teilnahme am *Shared Task* zielte darauf ab, das Konzept der Erzählebene innerhalb von Richtlinien für eine zunächst händische, letztlich aber computergestützte Annotation aufzubereiten. Die Erstellung der Annotationsrichtlinien förderte die Einsicht zutage, dass es sich bei dem narratologischen Konzept ‚Erzählebene' um ein überaus komplexes Phänomen handelt. Die für die Annotation nötige exakte Bestimmung des Phänomens hat zu Erkenntnissen geführt, die in der üblichen literaturwissenschaftlichen Praxis nur selten bedacht werden. Auf diese Weise haben wir – angeleitet durch die interdisziplinäre Perspektive – neue Einsichten in ein vermeintlich vertrautes Phänomen gewinnen können.

Im Zuge der Ausschreibung haben wir zwei Annotationsrichtlinien erstellt. Die erste Version wurde im Rahmen eines Workshops im September 2018 diskutiert und evaluiert (Ketschik et al. 2019). Im Anschluss an diesen bestand die Möglichkeit, eine revidierte Version der Richtlinien zu erstellen, in der das Feedback aus dem Workshop berücksichtigt werden konnte. In unserem Fall wurden insbesondere literaturwissenschaftliche Begriffe genauer definiert, die Rolle des Erzählers präzisiert und ein möglicher Zusammenhang mit verwandten narratologischen Konzepten diskutiert, vornehmlich der Stellung des Erzählers und Formen der Anachronie.

Wir sind davon überzeugt, dass die Bedeutung des Erzählers und dessen Vermittlungsakt der Schlüssel zu einer adäquaten Beschreibung von Erzählebenen ist. In der Überarbeitung unserer Richtlinien haben wir deshalb nicht nur die Relevanz des Erzählers hervorgehoben, wir haben uns auch dazu entschieden, narratologische Analysen etwa zur Fokalisierung, zum fiktiven Zuhörer des Erzählakts oder zu Parallelphänomenen wie Imaginationen oder Träumen so weit wie möglich auszublenden. Wo notwendig, grenzen wir in den folgenden Ausführungen das Konzept der Erzählebene von anderen Phänomenen ab. Darüber hinaus haben wir die zweite Version unserer Annotationsrichtlinien neu strukturiert. Ziel war es dabei, eindeutiger zwischen unserer theoretischen Konzeption, der Iden-

tifikation neuer Erzählebenen und der konkreten Annotationsanleitung zu unterscheiden. Um dem Rezipienten das Verständnis unseres theoretischen Ansatzes zu erleichtern, sind ferner zusätzliche Beispiele für verschiedene (Standard- sowie Sonder-)Fälle von Erzählebenen eingearbeitet worden. Der vorliegende Text entspricht in weiten Teilen der zweiten Version unserer Guidelines.

Unser Hauptbestreben besteht darin, das narratologische Konzept von Erzählebenen vollständig zu erfassen und zu formalisieren. In diesem Sinne lehnen wir es ab, das Phänomen zu vereinfachen, nur um ein höheres Maß an intersubjektiver Übereinstimmung zu erreichen (*inter-annotator agreement*, vgl. dazu den Beitrag von Reiter (2020) ab S. 193 in diesem Band). Stattdessen möchten wir mit den folgenden Ausführungen einen Beitrag zur theoretischen Auseinandersetzung und zu einer möglichst akkuraten Erfassung von Erzählebenen leisten.

2 Theoretische Einführung und Anwendungsfälle

Die Narratologie bildet seit den frühen 1960er-Jahren ein zentrales Beschäftigungsfeld von Literaturwissenschaftlerinnen (vgl. Martínez und Scheffel 2009, S. 7). Sie befasst sich mit den Techniken des Erzählens und richtet ihren Fokus auf die systematische Beschreibung unterschiedlicher Typen, Strukturen und Funktionen narrativer Phänomene. Während ihre Begriffe längst als heuristische Instrumente der Textanalyse Anwendung finden, diskutieren Narratologen noch immer, ob die Narratologie eine Methode, eine Theorie oder eine eigenständige Disziplin sei (vgl. Meister 2014, S. 623).

Die Gegenstände der Narratologie umfassen eine große Bandbreite an Phänomenen – man denke an die unterschiedlichen Parameter zur Bestimmung des Erzählers, etwa sein Verhältnis zur erzählten Welt, die von ihm eingenommene Perspektive bzw. Fokalisierung, seine Darstellungsform oder die Zuverlässigkeit seiner Aussagen. Aber auch strukturelle Textelemente, wie der Einsatz von Erzählebenen oder Anachronien, sind Gegenstand narratologischer Analysen. Untersucht werden unter anderem die Ordnung, Dauer und Frequenz des Erzählten. Narratologen widmen sich somit vornehmlich der Struktur und dem Aufbau literarischer Texte. Um sowohl den zeitlichen Handlungsverlauf als auch seine Anordnung und Darstellung in einem literarischen Text zu beschreiben, hat Gérard Genette die Konzepte *discours* und *histoire* systematisch voneinander abgegrenzt (vgl. Genette 1994, S. 199–201). Damit trennt er die Frage nach dem ‚Was' der Erzählung von der Frage danach, ‚wie' etwas erzählt wird. Während die *histoire* „die Gesamtheit der erzählten Ereignisse" (Genette 1994, S. 199) subsumiert, meint der

discours die tatsächliche Umsetzung der *histoire* im jeweiligen Erzählakt, sei er mündlich oder schriftlich überliefert.

Um Erzählebenen voneinander zu unterscheiden, schlägt Genette eine Klassifizierung verschiedener Erzählertypen vor. Hierfür verwendet er die Begriffe ‚extradiegetisch', ‚intradiegetisch' und ‚metadiegetisch'. Der extradiegetische Erzähler führt in die „primäre Erzählung mitsamt ihrer Diegese" (Genette 1994, S. 249), sprich ihrer Erzählwelt, ein. In dieser Erzählung kann sich wiederum ein intradiegetischer Erzähler befinden – meist eine Figur der ersten Erzählebene, die innerhalb dieser eine eigenständige Erzählung anstößt. Dieses Schema lässt sich beliebig erweitern. Üblicherweise wird diese Struktur metaphorisch beschrieben, indem etwa von einer Rahmung (*framing*) oder Verschachtelung (*embedding*) gesprochen oder das Phänomen mit chinesischen Schachteln, Tablettständern oder Matruschkas verglichen wird (vgl. Ryan 1991, S. 178–182, vgl. Pier 2014, S. 547).

In unseren Annotationsrichtlinien folgen wir Genettes grundlegender Annahme, dass eine Erzählebene eine ausreichend erkennbare „*Schwelle* zwischen den einzelnen Diegesen" (Genette 1994, S. 249) benötigt. Um diese Schwelle zu definieren, knüpft Genette die Existenz einer neuen Erzählebene an die Einführung eines neuen Erzählers. Die Erzählebene entsteht also, weil ein neuer, in der fiktiven Welt der Diegesis situierter Sprecher eine Geschichte erzählt, die in einer neuen Welt angesiedelt ist bzw. neue Figuren an einem neuen Ort zu einer anderen Zeit zum Gegenstand hat oder haben kann (vgl. Ryan 1991, S. 175–177). Während dieser Sprecher seine Geschichte erzählt, wird er zu einem intradiegetischen Erzähler. Der Erzählerwechsel ist damit ursächlich für den Wechsel der Erzählebene. Der orientalische Märchenzyklus *Tausendundeine Nacht* etwa wird schon bei Genette als Beispiel für diese Art des Erzählerwechsels herangezogen.

Grundsätzlich wollen wir an dieser Annahme Genettes festhalten. Da die Literaturgeschichte jedoch Beispiele hervorgebracht hat, die auch ohne prototypischen Erzählerwechsel deutliche Anzeichen einer neuen Erzählebene aufweisen (vgl. Lahn und Meister 2013, S. 83), werden wir sie dennoch erweitern. Solche Fälle gilt es, akkurat zu beschreiben: Sie müssen einerseits eine eindeutig abgrenzbare Erzählwelt aufweisen, sind andererseits aber sowohl von Formen der Anachronie als auch von figuralen Sprechakten eindeutig zu unterscheiden. Zwar können Sprechakte ein Indikator für die Einführung einer neuen Erzählebene sein (vgl. Ryan 1991, S. 175–177), alleine stellen sie hierfür allerdings kein ausreichendes Kriterium dar. Um einen Wechsel der Erzählebene herbeizuführen, muss ein eingeführter Sprecher nach unserem Verständnis eine Geschichte ‚erzählen' und damit zum Erzähler werden. Tatsächlich kommt dies in Sprechakten aber eher selten vor.

Die Annotation von Erzählebenen kann ohne ein umfassendes Verständnis erzähltexttheoretischer Theoreme letztlich nicht gelingen. Dieses ist notwendig,

um die Struktur eines Textes zu analysieren und damit Einsichten in das Verhältnis von Form und Inhalt zu erhalten.[1] Dass für die Analyse von Erzähltexten die Erfassung von Erzählebenen von hoher Relevanz ist, erschließt sich letztlich auch daraus, dass diese den Aufbau und Inhalt des Texts ebenso wie den Einsatz unterschiedlicher Erzähler aufschlüsseln.

Mögliche Forschungsfragen, die sich an die Analyse von Erzählebenen anschließen, konzentrieren sich vorwiegend auf die strukturelle Untersuchung eines Texts. Man denke etwa an die Unterscheidung der vorkommenden Erzähler, das Verhältnis von Rahmen- und Binnenerzählung oder die Bedeutung der einzelnen Erzählungen für das Gesamtgefüge. Forschungsfragen können sich aber natürlich auch dem Inhalt der Erzählung(en) widmen. Da die Erzählebenen funktional aufeinander bezogen werden können, ist es wichtig, die Informationsverteilung der Figur/en oder der/s Erzähler/s unter Berücksichtigung solcher Interdependenzen zu interpretieren. Erzählebenen können im Textganzen unterschiedliche Funktionen einnehmen. Von einer explikativen Funktion wird gesprochen, wenn die eingeschobene Erzählung Erklärungen für Elemente der Rahmenerzählung bereitstellt. Eine aktionale Funktion hingegen liegt vor, ist die eingebettete Erzählung für die Haupterzählung handlungskonstitutiv. Und eine thematische Funktion findet sich, wenn es zwischen den beiden Erzählungen auf struktureller Ebene Analogien, Korrespondenzen oder Kontrastbeziehungen gibt (vgl. Lahn und Meister 2013, S. 83 f.). Darüber hinaus scheint uns eine systematischere Beschreibung der Übergänge von einer Erzählebene zur anderen für das Erkennen solcher Phänomene instruktiv zu sein. Die detaillierte Erfassung hilft, das Phänomen zu verstehen und damit die Ebenen besser identifizieren zu können. Mit Blick auf die Literaturgeschichte als Ganzes wäre es ebenfalls aufschlussreich, genauer zu untersuchen, ob sich für Anzahl, Länge und Funktion von Erzählebenen bestimmte Muster hinsichtlich literarischer Gattungen, literarischer Perioden, des Geschlechts oder der Herkunft des Autors ablesen lassen.

3 Narratologische Begriffe und Konzepte

Im Folgenden wollen wir einige grundlegende narratologische Begriffe, die wir in unseren Grundannahmen (Abschnitt 4) und den Annotationsrichtlinien (Abschnitt 5) verwenden, kursorisch erläutern. Dies soll zu einem besseren Verständnis unserer Richtlinien und der ihnen zugrundeliegenden literaturwissenschaftlichen Konzeption beitragen.

1 Vgl. etwa das ‚Gehalt-Gestalt-Gefüge', dazu kursorisch Klausnitzer 2013, S. 126 f.

3.1 Erzählung

Obwohl der Begriff der Narrativität inzwischen als umstritten gilt (vgl. Abbot 2014, S. 587), werden in der Literaturwissenschaft immer noch zwei unterschiedliche Konzepte verwendet: Während Narrativität in der klassischen Erzähltheorie „an die Gegenwart einer vermittelnden Instanz, des ‚Erzählers', gebunden" ist (Schmid 2014, S. 1), fokussiert der Strukturalismus vor allem die „temporale Struktur und [...] *Veränderungen* eines Zustands" (Schmid 2014, S. 2). Wir stimmen mit Wolf Schmid überein, dass Narrativität in „practical literary theory" am besten durch eine Kombination beider Konzepte erklärt werden kann (Schmid 2003, S. 17). Die Erzählung ist somit kommunikationstheoretisch betrachtet mit einem Sprechakt verbunden, da sie Teil eines Vermittlungsaktes ist. Der Erzähler adressiert seine Geschichte stets an einen fiktiven Zuhörer, egal ob dieser explizit erwähnt wird oder nicht. Unsere Minimaldefinition einer in einem Erzählakt präsentierten Geschichte[2] lautet, dass eine Handlung einen Zustandswechsel von einem Zustand A zu einem Zustand B umfasst, dass dieser Zustandswechsel motiviert und kausal verknüpft ist und dass er von einem Erzähler vermittelt wird (vgl. Martínez und Scheffel 2009, S. 109–110 und Schmid 2018, S. 312 f.).

3.2 Der Erzähler

Die Analyse des Erzählers gehört zu den zentralen Aufgaben der Narratologie (vgl. Igl 2018, S. 127).[3] Um unterschiedliche Erzähler voneinander abgrenzen und damit auch den Wechsel des Erzählers identifizieren zu können, ist es sinnvoll, die Stellung des Erzählers in Bezug auf die von ihm erzählte Geschichte zu bestimmen. Daher ist grundsätzlich zu klären, ob er Teil der von ihm erzählten Welt ist oder nicht (vgl. Genette 1994, S. 174–176).

2 Genettes Begriff *histoire* wird im Deutschen als ‚Geschichte' übersetzt. Wir meinen damit nicht nur den gesamten Text, sondern auch die Geschichten einzelner Erzählebenen.

3 Tatsächlich gibt es in der Narratologie unterschiedliche Positionen hinsichtlich der Bedeutung des Erzählers: Es wird immer noch diskutiert, ob der Erzähler „the highest-level speech position" in einem literarischen Text innehat und damit eine „strictly textual category" ist, die deutlich vom Autor zu unterscheiden sein sollte (vgl. Margolin 2014, S. 646), oder ob narrative Texte ohne Erzähler denkbar sind (vgl. Igl 2018, S. 128 f.).

3.2.1 Der homodiegetische Erzähler

Ein homodiegetischer Erzähler ist als Figur in der von ihm erzählten Welt Teil seiner eigenen Geschichte: sei es als stiller Beobachter, Nebenfigur oder Hauptfigur.[4] In retrospektiv erzählten Geschichten spaltet sich somit das Ich des homodiegetischen Erzählers in ein ‚erzählendes' und ein ‚erlebendes Ich' auf. Während das erzählende Ich auf der primären Erzählebene in der Exegesis, ansonsten aber auf einer übergeordneten Erzählebene situiert ist und von dieser Warte aus einen Überblick über den gesamten Handlungsverlauf hat, sind die Gedanken und Gefühle des erlebenden Ichs an die aktuelle Handlungssituation gebunden und vom gerade Erlebten geprägt (vgl. Lahn und Meister 2013, S. 70; Fludernik 2006, S. 103 f.). Normalerweise ist eine klare Trennung von erzählendem und erlebendem Ich wahrzunehmen, da ein Unterschied existiert zwischen dem Zeitpunkt, zu dem erzählt wird, und dem Moment, in dem erlebt wird. Es finden sich aber auch Texte, in denen die Distanz zwischen den beiden Ichs scheinbar aufgehoben wird (vgl. etwa im inneren Monolog in Arthur Schnitzlers *Leutnant Gustl*, Lahn und Meister 2013, S. 72).

3.2.2 Der heterodiegetische Erzähler

Im Gegensatz zum homodiegetischen Erzähler ist der heterodiegetische Erzähler nicht Teil der von ihm erzählten Welt. In einer Geschichte, die von einem heterodiegetischen Erzähler vermittelt wird, gibt es daher stets eine klare Unterscheidung zwischen dem Standpunkt des Erzählers in der Exegesis und der Welt der Figuren in der Diegesis. Da er nicht Teil der erzählten Welt ist, gibt es hier auch keine Aufspaltung des Erzählers in ein erzählendes und erlebendes Ich.

3.3 Erzählebenen

Um das Phänomen von Erzählebenen zu beschreiben, haben sich in der Literaturwissenschaft viele teils sehr unterschiedliche Begrifflichkeiten ausgebildet (vgl. Pier 2014, S. 549–558). Wir fokussieren uns darauf, jede neue Erzählung eines Erzähltexts zu identifizieren und sie als neue Erzählebene zu definieren. Eine neue Erzählung wird nach unserem Verständnis von einem Erzählerwechsel ausgelöst.

4 Der Begriff des autodiegetischen Erzählers hat sich für einen homodiegetischen Erzähler durchgesetzt, der die Hauptfigur seiner Geschichte ist. Für unsere Analysen ist diese Unterscheidung nicht relevant.

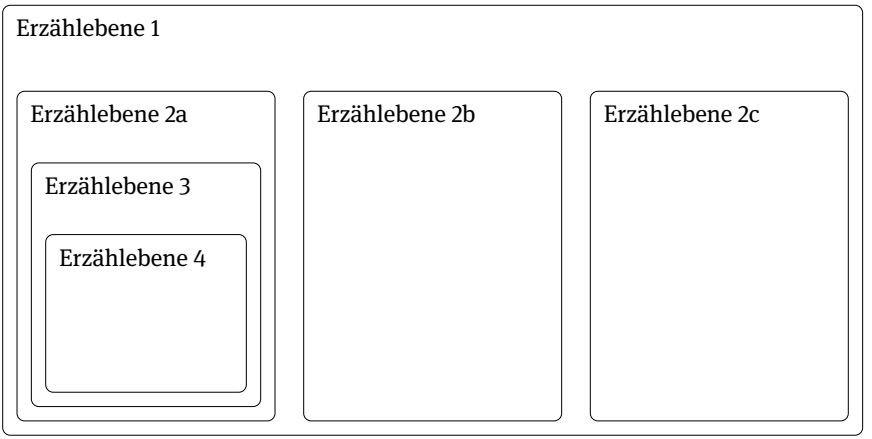


Abb. 1: Ineinander verschachtelte und sequenziell angeordnete Erzählebenen in einem literarischen Text

In den meisten Fällen ist dieser Wechsel eindeutig ersichtlich. Es gibt allerdings durchaus Erzählerwechsel, die subtiler angelegt und daher schwieriger zu greifen sind (dies wird in Abschnitt 4.2 genauer beschrieben).

Erzählebenen können ineinander verschachtelt (s. Abbildung 1: Ebenen 1 bis 4) oder sequenziell verknüpft sein (ebd.: 2a, 2b, 2c) oder aber eine Kombination aus beidem darstellen. In eine Rahmenerzählung bzw. eine übergeordnete Erzählebene können mehrere Erzählungen eingebettet sein. Da eine eingebettete Erzählung wiederum selbst für eine andere eingebettete Erzählung zur Rahmenerzählung werden kann, sprechen wir in unseren Guidelines von primären, sekundären, tertiären etc. Erzählebenen. Dadurch wird auch die Ambiguität von Genettes Terminologie (extra-, intra-, meta-, metametadiegetisch etc.) vermieden (vgl. Jahn 2017). Werden Erzählebenen sequenziell angeordnet, befinden sie sich grundsätzlich auf derselben Ebene. Sequenzielle Erzählebenen sind normalerweise immer schon auf einer zweiten, dritten, vierten etc. Ebene angeordnet und werden nur sehr selten ohne Rahmung auf der ersten Ebene aneinandergereiht (ein Beispiel hierfür wäre etwa Orhan Pamuks *Rot ist mein Name*).

3.4 Exegesis und Diegesis

„[D]iegesis designates the level of the narrated world, and exegesis the level of the narrating", konstatieren Coste und Pier (2009, S. 301). Dementsprechend gehört ein homodiegetischer Erzähler der ersten Erzählebene immer schon zu beiden Be-

reichen: In seiner Funktion als Erzähler gehört er der Exegesis an; da er aber eine Geschichte erzählt, an der er selbst teilhat, agiert er auch auf der Ebene der Diegesis (vgl. Genette 1994, S. 249). Ein heterodiegetischer Erzähler hingegen ist nur in der Exegesis situiert, da er nicht Teil der erzählten Welt ist.

Die Exegesis, also der Standort des Erzählers in einer primären Erzählsituation, bildet keine eigenständige Erzählebene. In diesem Sinne verweisen Aphorismen, Motti, Kommentare, Wertungen und Adressatenanreden[5] nicht auf eine eigenständige Erzählebene (vgl. Schmid 2014, S. 7), sondern sind Teil der aktuellen Erzählebene. Dies gilt gleichfalls für Aussagen von sekundären, tertiären etc. Erzählern, soweit sie nicht Elemente der übergeordneten Erzählebene adressieren (s. Abschnitt 5.8).

3.5 Illokutionäre und ontologische Grenzen

In ihrer Untersuchung von Erzählebenen schlägt Marie-Laure Ryan vor, sich auf den Anfang und das Ende von Erzählsträngen zu konzentrieren. Dabei unterscheidet sie illokutionäre von ontologischen Grenzen, die beide sowohl tatsächlich als auch virtuell überschritten werden können (vgl. Ryan 1991, S. 175–177).[6] Nach unserem Verständnis reicht allerdings weder ein Sprecherwechsel, also die Überschreitung der illokutionären Grenze, noch die Präsentation eines Sprechakts durch den Erzähler allein aus, um eine neue Erzählebene zu begründen. Es handelt sich bei diesen Fällen zuallererst um Sprechakte, die lediglich unter bestimmten Bedingungen (vgl. hierzu Abschnitte 4.1 und 4.2) zu einem Erzählerwechsel führen können und damit eine neue Erzählebene einleiten. Ontologische Grenzen führen nur in ganz bestimmten Fällen zu neuen Erzählebenen, nämlich

5 Im fiktionalen Text sind damit stets Anreden des fiktiven Zuhörers bzw. Lesers gemeint.

6 Ryan schreibt: „Illocutionary boundaries delimit speech acts within a text or a conversation, and their crossing introduces a new speaker or a new narrator“ (Ryan 1991, S. 175). Dabei unterscheidet sie vier Formen der Grenzüberschreitung, zwei Formen der illokutionären und zwei Formen der ontologischen Grenzüberschreitung. Eine tatsächlich überschrittene illokutionäre Grenze werde durch einen neuen Sprecher eingeführt, was sich auf der Mikroebene durch zitierte Figurenrede und auf der Makroebene zum Beispiel durch „narratives of personal experience“ zeige (Ryan 1991, S. 176). Virtuell überschrittene illokutionäre Grenzen hingegen werden durch den Erzähler eingeleitet, der den Sprechakt der Figur etwa in einer indirekten Rede vermittelt (vgl. Ryan 1991, S. 177). Eine ontologische Grenze „transports the reader to a new system of reality“ (Ryan 1991, S. 177). Wird die Grenze tatsächlich überschritten, wird das neue System als Ort der fiktionalen Realität beschrieben – zumindest vorübergehend. Das neue Realitätssystem wird also aus seinen eigenen Grenzen heraus bestimmt. Wird die Grenze hingegen virtuell überschritten, wird das neue System an einer „external perspective“ verankert, wobei das System „with repeated reminders of its [...] status in the primary reality“ verstanden wird (Ryan 1991).

dann, wenn mit ihnen ein neuer Erzähler eingeführt wird. Deshalb eröffnet beispielsweise Harry Potters Sprung durch die Mauer am Bahnhof King's Cross nach unserem Verständnis auch keine neue Erzählebene.

4 Grundannahmen: Wie findet man Erzählebenen?

Mit den folgenden Grundannahmen lassen sich alle Erzählebenen eines literarischen Textes auffinden. Generell gehen wir davon aus, dass in jedem Text mindestens eine Erzählebene vorliegt.

4.1 Neue Erzählebene durch neue Geschichte

Eine neue Erzählebene liegt vor, wenn eine neue Geschichte erzählt wird, die in einem neuen Erzählakt präsentiert wird. Sie wird durch einen Erzählerwechsel kenntlich gemacht. Den Begriff des Erzählerwechsels nutzen wir in einem weiten Sinn, der auch die Funktion des Erzählers miteinschließt. Deshalb liegt auch dann ein Erzählerwechsel vor, wenn sich die Stellung des Erzählers zum Erzählten ändert (s. Abschnitt 3.2). Eine Geschichte zeichnet sich prinzipiell durch folgende Kriterien aus:

1. Eine Geschichte ist eine in sich geschlossene Handlung, deren Ereignisse und Geschehnisse kausal miteinander verbunden sind und eine Zustandsveränderung bewirken.
2. Geschichten werden durch einen Erzähler vermittelt und als der „mündliche[] oder schriftliche[] Diskurs" präsentiert, „der von einem Ereignis oder einer Reihe von Ereignissen" berichtet (Genette 1994, S. 17).
3. Das Tempus von Erzählungen ist vornehmlich das Präteritum. In neuer Zeit gibt es einige Ausnahmen, etwa die Erzählung *Faserland* von Christian Kracht, in der ein homodiegetischer Erzähler im Präsens erzählt.

4.2 Fälle von Erzählerwechsel

Ein Erzählerwechsel führt immer zu einer neuen Geschichte, womit zugleich eine neue Erzählebene beginnt. Die Komplexität literarischer Texte zwingt uns allerdings, Genettes Annahmen vom Erzählerwechsel auszuweiten. Im Folgenden erläutern wir konkrete Szenarien, die einen Erzählwechsel und den Beginn einer neuen Erzählebene einleiten:

4.2.1 Der prototypische Fall: eine neue Erzählfigur

Üblicherweise ergibt sich ein Erzählerwechsel dadurch, dass eine Figur der erzählten Welt eine, wenn nicht gar seine eigene Geschichte erzählt. Damit ist der Erzähler zweiter Stufe Teil der Diegesis, die vom übergeordneten Erzähler vermittelt wird. In einer neuen Erzählung dieser Art findet sich somit ein neues erzählendes Ich.

Diese Variante entspricht Genettes Verständnis von Erzählebenen (vgl. Genette 1994, S. 250). Sein Beispiel aus dem von Abbé Prévost verfassten Roman *Histoire du Chevalier Des Grieux et de Manon Lescaut* (dt. *Manon Lescaut*) verweist auf den prototypischen Erzählerwechsel (vgl. Genette 1994, S. 162–164). Der Roman beginnt mit den Memoiren des Monsieur de Renoncour. Er ist der primäre Erzähler, bei dem es sich – da er seine eigene Geschichte erzählt – um einen homodiegetischen Erzähler handelt. Der Roman setzt wie folgt ein:

> Ich muss meine Leser in die Zeit meines Lebens zurückversetzen, zu der ich dem Chevalier des Grieux zum ersten Mal begegnet bin. Es war ungefähr ein halbes Jahr vor meiner Abreise nach Spanien. Obschon ich meine Abgeschiedenheit nur selten verließ, unternahm ich doch meiner Tochter zuliebe ab und zu eine kleine Reise, die ich aber tunlichst abkürzte. (Abbé Prévost 2008, S. 5)

Die in einer Analepse erläuterte Erklärung de Renoncours, die darlegt, wie er einst des Grieux traf, nimmt mehrere Seiten ein. Daran anschließend werden zwei Jahre der erzählten Zeit übersprungen. Eines Tages treffen sich de Renoncour und des Grieux in Calais wieder, wo des Grieux von seiner kürzlich begangenen Reise nach Amerika, seinen Unglücksfällen und seinem Versagen berichtet. De Renoncour führt diese Geschichte folgendermaßen ein:

> Ich muss meine Leser hier darauf aufmerksam machen, dass ich seine Geschichte fast unmittelbar nachher niedergeschrieben habe, als ich sie von ihm vernommen hatte. Ich kann also versichern, dass dieser Bericht ganz genau und fast wortgetreu ist. Ich meine wortgetreu in der Wiedergabe der Gedanken und Gefühle, die der junge Abenteurer mit dem edelsten Anstand zum Ausdruck brachte. Im Folgenden werde ich demnach seinem Bericht gar nichts hinzufügen, was nicht von ihm selbst stammt.
> Ich war siebzehn Jahre alt und vollendete meine Philosophiestudien in Amiens, wo meine Eltern, die einem der vornehmsten Häuser von P... angehörten, mich hingeschickt hatten. Ich führte ein so geordnetes, sittsames Leben, dass meine Lehrer mich der ganzen Schule als Vorbild hinstellten. Ich gab mir zwar nicht sonderlich Mühe, dieses Lob zu verdienen, aber ich bin von Natur sanft und ruhig veranlagt. Ich lag aus Neigung meinen Studien mit großem Eifer ob, und man rechnete mir eine gewisse Abneigung gegen Ausschweifungen als Tugend an. Meine Herkunft, meine erfolgreichen Studien und einige äußere Vorzüge, über die ich verfügte, hatten mir die Bekanntschaft und die Wertschätzung aller achtbaren Leute in der Stadt verschafft. (Abbé Prévost 2008, S. 13)

Als eine Figur auf der ersten Erzählebene wird aus des Grieux ein intradiegetischer Erzähler bzw. nach unserer Terminologie ein sekundärer Erzähler. Sein Sprechakt wird zu einem Erzählakt, womit ein neues erzählendes Ich entsteht. Die direkte Rede in der ersten Erzählebene indiziert einen Sprecherwechsel. Erst indem des Grieux seine Geschichte erzählt, die sich durch in sich geschlossene Handlungen auszeichnet, deren Ereignisse und Geschehnisse kausal miteinander verbunden sind und die eine Zustandsänderung bewirken, wird er vom einfachen Sprecher zum wirklichen Erzähler. Irrelevant bleibt dabei, dass seine Geschichte gleichzeitig eine ontologische Grenze überschreitet, eben weil die Reise nach Amerika eine Welt beschreibt, die sich von der Welt de Renoncours unterscheidet.

In der Literaturgeschichte kommen ähnliche Textstrukturen recht häufig vor. Heinrich von Kleists Erzählung *Unwahrscheinliche Wahrhaftigkeiten* ist eines von vielen weiteren Beispielen, das an dieser Stelle angeführt werden kann. In Kleists Anekdote liegt auf der ersten Erzählebene ein heterodiegetischer Erzähler vor. Er beginnt seine Geschichte wie folgt:

> „Drei Geschichten", sagte ein alter Offizier in einer Gesellschaft, „sind von der Art, daß ich ihnen zwar selbst vollkommen Glauben beimesse, gleichwohl aber Gefahr liefe, für einen Windbeutel gehalten zu werden, wenn ich sie erzählen wollte. Denn die Leute fordern, als erste Bedingung, von der Wahrheit, daß sie wahrscheinlich sei; und doch ist die Wahrscheinlichkeit, wie die Erfahrung lehrt, nicht immer auf Seiten der Wahrheit."
> Erzählen Sie, riefen einige Mitglieder, erzählen Sie! – denn man kannte den Offizier als einen heitern und schätzenswürdigen Mann, der sich der Lüge niemals schuldig machte.
> Der Offizier sagte lachend, er wolle der Gesellschaft den Gefallen tun; erklärte aber noch einmal im Voraus, daß er auf den Glauben derselben, in diesem besonderen Fall, keinen Anspruch mache.
> Die Gesellschaft dagegen sagte ihm denselben im Voraus zu; sie forderte ihn nur auf, zu reden, und horchte.
> „Auf einem Marsch 1792 in der Rheinkampagne", begann der Offizier, „bemerkte ich, nach einem Gefecht, das wir mit dem Feinde gehabt hatten, einen Soldaten, der stramm, mit Gewehr und Gepäck, in Reih' und Glied ging, obschon er einen Schuß mitten durch die Brust hatte; wenigstens sah man das Loch vorn im Riemen der Patrontasche, wo die Kugel eingeschlagen hatte, und hinten ein anderes im Rock, wo sie wieder herausgegangen war [...]." (Kleist 1990, S. 376)

Der Erzählerwechsel ist deutlich erkennbar, wenn der Offizier mit den Worten „Auf einem Marsch 1792 in der Rheinkampagne" einsetzt. Hier beginnt der Offizier als homodiegetischer Erzähler, eine Geschichte zu erzählen, die er selbst erlebt hat. Mit Ausnahme des *verbum dicendi* „begann der Offizier" gehört die folgende Passage zur zweiten Erzählebene.

4.2.2 Wechsel der Erzählposition

Obgleich die durch eine erzählende Figur eingeläutete Binnenerzählung der Standardfall eines Erzählebenenwechsels ist, gibt es literarische Texte, die von diesem prototypischen Szenario abweichen. In selteneren Fällen kann ein Erzählerwechsel vorliegen, ohne dass die Erzählfigur wechselt. So etwa, wenn ein extradiegetisch-homodiegetischer Erzähler plötzlich eine Geschichte erzählt, an der er entweder nicht teilhat oder die in seiner erzählten Welt selbst fiktional ist. In diesem Fall bleibt das erzählende Ich zwar dasselbe, seine Stellung zum Erzählten ändert sich jedoch.

Wir schlagen vor, auch in diesem Fall von einem Erzählerwechsel zu sprechen und unterscheiden damit zwischen dem Erzähler als Figur und seiner analytischen Funktion als Erzähler. Während ein neuer homodiegetischer Erzähler durch eine erzählende Figur eingeführt wird, ist ein neuer heterodiegetischer Erzähler auf zweiter, dritter etc. Stufe meist durch seine veränderte Stellung zur erzählten Welt auszumachen. Ein Beispiel für den zweitgenannten Fall findet sich in Max Frischs Roman *Stiller*. James White alias Anatol Ludwig Stiller erscheint über weite Teile des Romans als homodiegetischer Erzähler. Während seiner Gefangenschaft erzählt er seinem Verteidiger das Märchen von ‚Rip van Winkle':

> Dazu (was wichtig ist) hielt ich sein silbernes Feuerzeug mit Flämmchen, ohne jedoch die duftende Zigarre, diese immerhin einzige Wollust in meiner Untersuchungshaft, anzuzünden, nein, aller Begierde zum Trotz wiederholte ich meine Frage:
> „Sie kennen es nicht?"
> „Was?"
> „Das Märchen von Rip van Winkle?"
> Nur mit diesem Kniff, nämlich mit dem Feuerzeug in der Hand, das ich nach jedem Verlöschen wieder entzündete, dazu mit der Zigarre in der andern Hand, unablässig im Begriff, die schöne Zigarre endlich anzustecken, ja, einmal schon mit der ersten Glut an der Zigarre, so daß ich bloß hätte ziehen müssen, im letzten Augenblick doch jedesmal wieder verhindert – durch Rip van Winkle, dessen Märchen offensichtlich sogar akuter war als meine Zigarre – nur so konnte ich meinen geschäftigen Verteidiger überhaupt zum Zuhören, zum aufmerksamen Zuhören nötigen.
> Das Märchen lautet etwa folgendermaßen:
> Rip van Winkle, ein Nachkomme jener unerschrockenen van Winkles, die unter Hendrik Hudson dereinst das amerikanische Land erschlossen hatten, war ein geborener Faulenzer, dabei, wie es scheint, ein herzensguter Kerl, der nicht um der Fische willen fischte, sondern um zu träumen, denn sein Kopf war voll sogenannter Gedanken, die mit seiner Wirklichkeit wenig zu tun hatten. (Frisch 1954, S. 70 f.)

Der Abschnitt verdeutlicht, wie Stiller seine Erzählposition verändert: Nachdem er zu Beginn als homodiegetischer Erzähler auftritt, vermittelt er die Geschichte von Rip van Winkle nun als heterodiegetischer Erzähler. Sobald er beginnt,

das Märchen zu erzählen, verändert sich seine Stellung zur erzählten Welt. Somit liegt hier ein Wechsel der Erzählebene vor, ohne dass ein neuer Sprecher bzw. eine neue erzählende Figur eingeführt wird. Daraus folgt, dass ein Wechsel von einer homodiegetischen zu einer heterodiegetischen Erzählposition – ohne Wechsel der Sprecherposition – ausreicht, um eine neue Erzählebene zu bilden.

4.2.3 Transferierende Erzähler: Chronisten, Sammler, Herausgeber

Manche Erzähler signalisieren eine distanzierte Haltung zu der von ihnen erzählten Geschichte, weil sie die Geschichte explizit als überliefert bezeichnen. Für diese Formen führen wir den Oberbegriff ‚transferierende Erzähler' ein. Häufig äußert ein solcher transferierender Erzähler, dass er zufällig Zugang zu einer Geschichte erhalten hat, etwa indem er ein Manuskript gefunden oder eine Chronik gelesen hat. Für die Gesamterzählung hat dieser Erzähler eine explizit authentifizierende Funktion.

1. **Herausgeberfiktion.** Als ein Sonderfall narrativer Texte gilt die Herausgeberfiktion: „an agent whose sole involvement with the text is its material dissemination" (Ryan 2001, S. 151). Die Herausgeberfunktion bedingt immer eine eigene Erzählebene, auch wenn diese Erzählebene nur aus einem einzigen Satz besteht. Nicht selten kommt die Herausgeberfiktion erst am Ende eines Textes zum Vorschein. Obgleich bis zu diesem Zeitpunkt keine Indizien auf eine Rahmenhandlung hinweisen, macht die Existenz des Herausgebers den bisherigen Text als zweite Erzählebene bzw. Binnenerzählung kenntlich. Aus diesem Grund kann das Inklusionsschema (s. Abschnitt 3.3) grundsätzlich erst erstellt werden, wenn der komplette Text bekannt ist.
   Eduard Mörikes Erzählung *Lucie Gelmeroth* ist ein gutes Beispiel für die Herausgeberfiktion. Der Text setzt mit einem homodiegetischen Erzähler ein, der von seiner Urlaubsreise nach Göttingen erzählt.

   > Ich wollte – so erzählt ein deutscher Gelehrter in seinen noch ungedruckten Denkwürdigkeiten – als Göttinger Student auf einer Ferienreise auch meine Geburtsstadt einmal wieder besuchen, die ich seit lange nicht gesehen hatte. (Mörike 2005b, S. 13)

   Obwohl der erste Satz bereits auf die Präsenz einer Herausgeberfiktion hinweist, indem die „ungedruckten Denkwürdigkeiten" erwähnt werden, klärt erst das Ende der Erzählung die Verschachtelung der Erzählebenen gänzlich auf. Nach Abschluss der Erzählung ist dort zu lesen:

   > Hier bricht die Handschrift des Erzählers ab. Wir haben vergeblich unter seinen Papieren gesucht, vom Schicksal jenes flüchtigen Kaufmanns noch Etwas zu erfahren. Auch

> mit Erkundigungen anderwärts sind wir nicht glücklicher gewesen. (Mörike 2005b, S. 29)

Indem von gefundenen „Papieren" berichtet wird, zeigt sich, dass der Göttinger Student im Grunde ein sekundärer Erzähler ist. Der primäre Erzähler dagegen ist der erst am Ende sichtbar werdende Herausgeber dieser Blätter.

2. **Der Erzähler als Berichterstatter.** Im Gegensatz zu Herausgebern, die einen Erzähler einführen, der seine Geschichte selbst vermittelt, gibt es Formen der Berichterstattung, in denen ein Erzähler eine aufgefundene Geschichte erzählt. Meist handelt es sich um sogenannte Chronisten.[7] In Eduard Mörikes *Das Stuttgarter Hutzelmännlein* erzählt der etwa 1820 positionierte Erzähler die im Jahr 1220 situierte Geschichte von Seppe und Vrone und gibt sich dabei als Chronist zu erkennen. Während Seppe sich in dieser Erzählung mit einem Kutscher unterhält, berichtet der Erzähler von der „Historie der schönen Lau", die er etwa 100 Jahre vor der Seppe-Erzählung einordnet.

> Du aber, wohlgeneigter Leser, lasse dich, derweil die Beiden so zusammen discuriren, auch etlicher Dinge besonders berichten, die, ob sie sich zwar lang vor Seppes Zeit begeben, nichts desto minder zu dieser Geschichte gehören. Vernimm hiernach die wahre und anmuthige
>
> Historie von der schönen Lau.
>
> Der Blautopf ist der große runde Kessel eines wundersamen Quells bei einer jähen Felswand gleich hinter dem Kloster [...]. (Mörike 2005a, S. 130)

In diesem Fall wird weder ein neuer Erzähler eingeführt, noch ändert der Erzähler seine Stellung zur erzählten Welt. Er bleibt heterodiegetisch. Ausmachen lässt sich hingegen etwas anderes: Der Erzähler selbst weist explizit darauf hin, dass eine neue Geschichte, hier Historie, erzählt wird. Diese Geschichte ist nicht mehr als Bestandteil der übergeordneten Erzählung zu betrachten, die von Seppe und Vrone handelt. Vielmehr wird sie als Wissen einer alten Chronik ausgegeben, zu der allein der Erzähler Zugang hat. Die Vermittlung dieses Chronikwissens eröffnet eine neue Erzählebene.

Ähnliche Fälle liegen vor, wenn aus Zeitungsberichten, Briefen, Fernsehsendungen oder ähnlichen Medien Inhalte vermittelt werden, wobei stets eine Vermittlung einer Geschichte vorliegen muss (s. Abschnitt 3.1), um diese als neue Erzählebene auszuzeichnen.

---

7 Ein Chronist muss, „wie jeder gute Historiker, zumindest die Zuverlässigkeit seiner Quellen und Gewährsleute nachweisen" (Genette 1994, S. 279). Bei diesen Erzählern ist es nicht immer leicht zu entscheiden, ob es sich um homodiegetische oder heterodiegetische Erzähler handelt (vgl. Genette 1994, S. 262.). Obwohl sie Teil der diegetischen Welt sind, gibt es eine zeitlich unüberwindbare Distanz zwischen ihnen und der erzählten Welt.

4.3 Abgrenzung von verwandten Phänomenen

4.3.1 Anachronien

Für ein präzises Verständnis von Erzählebenen ist es wichtig, sie von Anachronien, sprich Analepsen und Prolepsen, zu unterscheiden. Während eine Erzählebene durch einen Erzählerwechsel – oder den Wechsel der Erzählposition – initiiert wird, bleiben Erzähler und Erzählposition in Anachronien konstant. Sie rücken ausschließlich die chronologische Ordnung des Erzählten im *discours* in den Fokus der Analyse, weshalb anachronistische Passagen die jeweilige Erzählebene nicht verlassen.

> Her father was becoming old lately, she noticed; he would miss her. Sometimes he could be very nice. Not long before, when she had been laid up for a day, he had read her out a ghost story and made toast for her at the fire. Another day, when their mother was alive, they had all gone for a picnic to the Hill of Howth. [...]
> Her time was running out but she continued to sit by the window, leaning her head against the window curtain, inhaling the odour of dusty cretonne. (Joyce 1967, S. 41)

Der heterodiegetische Erzähler in James Joyces *Eveline* erzählt die Geschichte von Eveline kurz vor ihrer geplanten Auswanderung. Der Erzähler nutzt mehrere Analepsen, um die Hauptfigur über ihre Kindheit nachdenken zu lassen. Im abgedruckten Abschnitt erinnert sie sich an eine Geistergeschichte ihres Vaters sowie an ein gemeinsames Familienpicknick. Die chronologisch früher erlebten Erinnerungen grenzen sich eindeutig durch ihr Tempus ab. Da aber weder ein Erzählerwechsel vorliegt noch der Erzähler seine Stellung zur erzählten Welt verändert oder gar als transferierender Vermittler auftritt, liegt kein Wechsel der Erzählebene vor.

4.3.2 Formen der Imagination

Bestimmte Erzählstrategien setzen figurale Gedanken, Träume, Visionen, Fantasien u. a. so ein, dass ein Wechsel der Erzählebene plausibel erscheinen mag (vgl. Pier 2014, S. 550). In diesen Fällen ist es wichtig, zwischen einem wirklichen Erzählebenenwechsel und einer starken Form interner Fokalisierung zu unterscheiden. In Anton Tschechows *Das Gewinnlos* stellt sich Ivan Dmitrich, eine der beiden Hauptfiguren, vor, wie das Leben sein könnte, wenn er in der Lotterie gewinnen würde:

> Iwan Dmitritsch malt sich den Herbst aus mit seinem endlosen Regen, den kalten Abenden, dem Altweibersommer. In dieser Jahreszeit ist es angenehm, ausgedehnte Spaziergänge zu

> machen im Garten, auf dem Felde, am Ufer des Flusses, um ordentlich durchzufrieren, dann ein großes Gläschen Wodka zu genehmigen und schnell einen eingemachten Reizker nachzuessen oder eine Dillgurke und … ein zweites Gläschen! Die Kinder kommen aus dem Gemüsegarten gelaufen und schleppen Mohrrüben und Rettich herbei, die noch ganz nach frischer Erde duften … Dann streckt man sich auf den Diwan, betrachtet in aller Gemütsruhe irgendein illustriertes Journal, breitet zu guter Letzt die Zeitung übers Gesicht, knöpft die Weste auf und verfällt in süßen Schlummer …(Tschechow 1968, S. 75)

Zwar sind die figuralen Gedanken deutlich erkennbar, allerdings bleiben sie durch den primären Erzähler vermittelt. Wo Ryan – Ivans Vorstellungen als eine neue Welt begreifend – vermutlich für eine ontologische Grenzüberschreitung plädieren würde, reicht dies nach unserem Verständnis nicht für einen Erzählebenenwechsel aus. Denn der heterodiegetische Erzähler vermittelt Ivans Gedanken. Tatsächlich sind seit der Moderne noch radikalere Formen von Gedankenrepräsentationen beobachtet worden, etwa der innere Monolog oder der *stream of consciousness* (vgl. Martínez und Scheffel 2009, S. 61–63). Erzählformen dieser Art werden vorgeblich nicht mehr durch den übergeordneten Erzähler vermittelt, sondern scheinen direkt von der Figur erlebt und erzählt zu sein (vgl. Igl 2018, S. 127). Prinzipiell ist es vorstellbar, dass auch solche Passagen als neue Ebene des Erzählens gestaltet werden. Die erlebende Figur würde dann jedoch selbst zum Erzähler werden und sich deutlich von der übergeordneten Erzählinstanz unterscheiden.

4.3.3 Metalepsen

In einigen wenigen Fällen verwehrt ein Text die Bestimmung von Erzählebenen. Dies ist insbesondere dann der Fall, wenn Erzählebenen erzähllogisch nicht mehr voneinander abgegrenzt werden können, etwa im Falle von Metalepsen (vgl. Pier 2016). In Italo Calvinos *Wenn ein Reisender in einer Winternacht* ist die Welt des Erzählers bzw. der Leser (die Exegesis) so eng mit der erzählten Welt (der Diegesis) verwoben, dass einzelne Erzählebenen nicht voneinander unterschieden werden können. In solchen Fällen sehen wir von der Annotation jeglicher Erzählebene ab.

> Ich bin der Mann, der da zwischen Café und Telefonzelle hin- und herläuft. Oder besser gesagt, dieser Mann heißt hier „ich“, und sonst weißt du nichts von ihm, wie auch dieser Bahnhof nur einfach „Bahnhof“ heißt, und außer ihm gibt es nichts als das unbeantwortete Läuten eines Telefons in einem dunklen Zimmer in einer fernen Stadt. Ich hänge den Hörer ein, warte auf das Scheppern der Münzen durch den metallenen Schlund, drehe mich um, drücke die Glastür auf und strebe wieder den Tassen zu, die sich zum Trocknen in einer Dampfwolke türmen.

> Die Espressomaschinen in Bahnhofcafés zeigen unübersehbar ihre Verwandtschaft mit den Lokomotiven, die Espressomaschinen von gestern und heute mit den Dampf- und E-Loks von gestern und heute. Ich mag hin- und herlaufen, mag mich drehen und wenden, soviel ich will: Ich sitze in einer Falle, in der zeitlosen Falle, die einem unweigerlich jeder Bahnhof stellt. Immer noch hängt ein feiner Kohlenstaub in der Luft, obwohl längst alle Strecken elektrifiziert worden sind, und ein Roman, der von Zügen und Bahnhöfen handelt, kann nicht umhin, diesen Rauchgeruch wiederzugeben. Schon mehrere Seiten hast du dich jetzt vorangelesen, es wäre mithin an der Zeit, daß dir klar gesagt wird, ob dieser Bahnhof, an dem ich ausgestiegen bin aus einem verspäteten Zug, ein Bahnhof von früher ist oder von heute; doch die Sätze bewegen sich weiter im ungewissen, im Grau, in einer Art Niemandsland der auf den kleinsten gemeinsamen Nenner verkürzten Erfahrung. Paß auf, das ist bestimmt ein Trick, um dich langsam einzufangen, dich in die Handlung hineinzuziehen, ohne daß du es merkst: eine Falle. (Calvino 1983, S. 16 f.)

5 Annotationsrichtlinien

Die nun folgenden Richtlinien dienen als praktische Arbeitsanleitung zur Annotation von Erzählebenen. Sie resultieren aus den explizierten Grundannahmen (s. Abschnitt 4). Bevor der Annotationsvorgang begonnen wird, sollte der zu annotierende literarische Text einmal komplett gelesen werden. Im Anschluss werden Erzählebenen wie folgt annotiert:

5.1 Auszeichnung

Alle Erzählebenen werden mit eckigen Klammern annotiert, wobei jede Ebene durch eine öffnende Klammer am Anfang und eine schließende Klammer am Ende gekennzeichnet wird.[8]

5.2 Nummerierung

Die unterschiedlichen Erzählebenen werden zunächst mit Nummern versehen (1, 2, 3 etc.). Abgesehen vom seltenen Sonderfall der Metalepse (s. Abschnitt 4.3.3) hat jeder Text mindestens eine Erzählebene, die mit der Nummer [1] gekennzeichnet wird. Sollte zusätzliche eine sequenzielle Reihe von Erzählebenen auftreten,

8 Die Annotation kann auch auf andere Weise erfolgen (z. B. mit unterschiedlichen Farben, die die Zugehörigkeit zu einer bestimmten Erzählebene markieren), je nachdem, welches Annotationswerkzeug verwendet wird.

werden Buchstaben genutzt, um die nebeneinander angeordnete Ebenen zu benennen.

1. Die Nummer gibt den Grad der Erzählebene im Inklusionsschema an. Damit ist Erzählebene 2 eine Erzählung zweiter Stufe, die in eine übergeordnete Erzählung (Ebene 1) eingebettet ist. Es handelt sich bei der Erzählung zweiter Stufe um eine Binnenerzählung. So ist etwa das Märchen von ‚Rip van Winkle' in Max Frischs Roman *Stiller* eine Binnenerzählung, die in die Rahmenerzählung von Stillers Leben eingebettet ist.
2. Erzählungen, die sich auf derselben Erzählebene befinden, aber sequenziell angeordnet sind, werden mit Buchstaben ausgezeichnet (a, b, c). In Boccaccios berühmtem Novellenzyklus *Il Decamerone* finden sich einhundert sequenziell angeordnete Erzählungen, die alle auf der zweiten Ebene in die Rahmenerzählung eingebettet sind. Sie werden mit der Nummerierung 2a, 2b, 2c etc. versehen; in einem Extremfall wie dem von Boccaccios Zyklus müsste man sich mit doppelter und dreifacher Buchstabennennung aushelfen (... 2y, 2z, 2aa, 2ab).

5.3 Verwendung der Klammern

Die eckigen Klammern werden wie folgt platziert:

1. Sowohl die öffnende als auch die schließende Klammer sind mit der Nummer und ggf. einem Buchstaben gekennzeichnet: $[_1 ...]_1$
2. Die übergeordnete Erzählebene beginnt stets vor der eingebetteten Erzählung. Ebenso schließen die Klammern der übergeordneten Erzählung erst, wenn die Klammern der eingebetteten Erzählebenen geschlossen sind: $[_1 ... [_2 ...]_2 ...]_1$.
3. Bei sequenziellen Anordnungen schließen die Klammern der ersten sequenziell angeordneten Erzählung (z. B. 2a), bevor die der zweiten (z. B. 2b) sich öffnen: $[_1 ... [_{2a} ...]_{2a} ... [_{2b} ...]_{2b} ...]_1$.
4. Jegliche Interpunktion wird nicht vom vorangehenden Wort bzw. Satz getrennt.

(1) $[_1$ ...$[_2$ „Auf einem Marsch 1792 in der Rheinkampagne", $]_2$ begann der Offizier, $[_2$ „bemerkte ich, nach einem Gefecht, das wir mit dem Feinde gehabt hatten, einen Soldaten ..." $]_2$ $]_1$ (Kleist 1990, S. 376)

5.4 Eingebettete Erzählebenen und Herausgeberfiktion

Jeder Text beginnt mit einer ersten Erzählebene, in die weitere Erzählebenen eingebettet sein können.

(2) [$_1$ Zum kommenden Sonntag war ich von den Paulsenschen Eheleuten auf den Abend eingeladen, um ihnen ihren Hochzeitstag feiern zu helfen. Es war im Spätsommer, und da ich mich frühzeitig auf den Weg gemacht und die Hausfrau noch in der Küche zu wirtschaften hatte, so ging Paulsen mit mir in den Garten, wo wir uns zusammen unter der großen Linde auf die Bank setzten. Mir war das „Pole Poppenspäler" wieder eingefallen, und es ging mir so im Kopf herum, daß ich kaum auf seine Reden Antwort gab; endlich, da er mich fast ein wenig ernst wegen meiner Zerstreuung zurecht gewiesen hatte, fragte ich ihn gradezu, was jener Beiname zu bedeuten habe.
Er wurde sehr zornig. „Wer hat dich das dumme Wort gelehrt?" rief er, indem er von seinem Sitze aufsprang. Aber bevor ich noch zu antworten vermochte, saß er schon wieder neben mir. „Laß, laß!" sagte er sich besinnend; „es bedeutet ja eigentlich das Beste, was das Leben mir gegeben hat. – Ich will es dir erzählen; wir haben wohl noch Zeit dazu. –
In diesem Haus und Garten bin ich aufgewachsen, meine braven Eltern wohnten hier, und hoffentlich wird einst mein Sohn hier wohnen! – Daß ich ein Knabe war, ist nun schon lange her; aber gewisse Dinge aus jener Zeit stehen noch, wie mit farbigem Stift gezeichnet, vor meinen Augen.
[$_2$ Neben unserer Haustür stand damals eine kleine weiße Bank mit grünen Stäben in den Rück- und Seitenlehnen, von der man nach der einen Seite die lange Straße hinab bis an die Kirche, nach der andern aus der Stadt hinaus bis in die Felder sehen konnte. An Sommerabenden saßen meine Eltern hier, der Ruhe nach der Arbeit pflegend; in den Stunden vorher aber pflegte ich sie in Beschlag zu nehmen und hier in der freien Luft und unter erquickendem Ausblick nach Ost und West meine Schularbeiten anzufertigen."]$_2$]$_1$ (Storm 1987, S. 166 f.)

Eine Erweiterung dieser Regel gibt es bei literarischen Texten, bei denen eine Herausgeberfiktion eine eigenständige Erzählebene evoziert. Sie wird als Rahmenerzählung auf Ebene 1 angegeben, auch wenn die Herausgeberfiktion erst am Ende der Erzählung sichtbar wird.

(3) DECEMBER 6.
[$_1$ [$_2$ Wie mich die Gestalt verfolgt! Wachend und träumend füllt sie meine ganze Seele! Hier, wenn ich die Augen schließe, hier in meiner Stirne, wo die Sehkraft sich vereinigt, stehen ihre schwarzen Augen. Hier! ich kann dir es nicht ausdrücken. Mache ich meine Augen zu, so sind sie da; wie ein Meer, wie ein Abgrund ruhen sie vor mir, in mir, füllen die Sinne meiner Stirn.
Was ist der Mensch, der gepriesene Halbgott! Ermangeln ihm nicht eben da die Kräfte, wo er sie am nöthigsten braucht? Und wenn er in Freude sich aufschwingt, oder im Leiden versinkt, wird er nicht in beiden eben da aufgehalten, eben da zu dem stumpfen kalten Bewußtsein wieder zurückgebracht, da er sich in der Fülle des Unendlichen zu verlieren sehnte?]$_2$]$_1$

Der Herausgeber an den Leser.
[$_1$ Wie sehr wünscht' ich, daß uns von den letzten merkwürdigen Tagen unsers Freundes so viel eigenhändige Zeugnisse übrig geblieben wären, daß ich nicht nöthig hätte, die Folge seiner hinterlassenen Briefe durch Erzählung zu unterbrechen. **]$_1$** (Goethe 1899, S. 139–141)

5.5 Paratexte

Wie schon in Beispiel (3) ersichtlich, werden Paratexte (vgl. Genette 1989) wie Buchtitel, Vorworte, Kapitelüberschriften und Gattungsangaben nicht annotiert. Wenn sich die Erzählebene nicht ändert, werden die Klammern am Ende des vorangehenden Kapitels geschlossen und erst nach der Überschrift wieder geöffnet.

(4) **[$_1$** By reason of these things, then, the whaling voyage was welcome; the great floodgates of the wonder-world swung open, and in the wild conceits that swayed me to my purpose, two and two there floated into my inmost soul, endless processions of the whale, and, midmost of them all, one grand hooded phantom, like a snow hill in the air. **]$_1$**
CHAPTER 2
The Carpet-Bag
[$_1$ I stuffed a shirt or two into my old carpet-bag, tucked it under my arm, and started for Cape Horn and the Pacific. Quitting the good city of old Manhatto, I duly arrived in New Bedford. It was on a Saturday night in December. Much was I disappointed upon learning that the little packet for Nantucket had already sailed, and that no way of reaching that place would offer, till the following Monday. **]$_1$** (Melville 2002, S. 22 f.)

5.6 Syntaktisch eingebundene Überschriften

Überschriften, die jedoch semantisch und syntaktisch zur aktuellen Erzählebene gehören, bilden hierzu eine Ausnahme. Sie werden zur zugehörigen Erzählebene geordnet.

(5) **[$_1$** ... **[$_n$** Vernimm hiernach die wahre und anmuthige
Historie von der schönen Lau. **]$_n$**[9]
[$_2$ Der Blautopf ist der große runde Kessel eines wundersamen Quells bei einer jähen Felswand gleich hinter dem Kloster ...**]$_2$** **]$_1$** (Mörike 2005a, S. 130)

9 Die Abkürzung n steht für ‚nicht erzählend' und wird in Abschnitt 5.8 erklärt.

5.7 Unterbrechung von Erzählebenen

Erzählebenen können durch andere Erzählebenen unterbrochen werden. So können im Erzählakt eines sekundären Erzählers etwa Aussagen vorkommen, die auf der primären Erzählebene stattfinden. Zwei Fälle solcher Unterbrechungen sind möglich:

1. eingefügte *verba dicendi*, die auf die spezifische Erzählsituation verweisen und damit noch zur übergeordneten Erzählebene gehören, und
2. eingeschobene Sprechakte, die nicht Teil der Erzählung sind, sondern sich etwa an die Figuren auf der übergeordneten Erzählebene richten.

In solchen Fällen wird die eingeschobene Erzählebene geschlossen, wenn der Einschub beginnt, und wieder geöffnet, wenn er endet. In Beispiel (6) werden drei sequenziell angeordnete Geschichten erzählt (2a, 2b, 2c), wobei hier nur die dritte zu sehen ist. Diese wird durch Einschübe auf der ersten Erzählebene unterbrochen. Dabei handelt es sich einerseits um einen Einschub mit *verbum dicendi* („ergänzte der Offizier"), andererseits um eine Ansprache des Adressatenkreises („meine Herren"). Beide Formen der Unterbrechung können auch gemeinsam auftreten, müssen in diesem Fall aber nicht separat annotiert werden (Beispiel (7): „‚Nun also', sagte er."). Einschübe dieser Art können im Übrigen auch nur aus einem Wort bestehen, ebenso wie sie mehrere Sätze umfassen können.

(6) **[$_1$** Der Landedelmann meinte, daß er die Geschichten, die seinen Satz belegen sollten, gut zu wählen wüßte.
[$_{2c}$ „Die dritte Geschichte," **]$_{2c}$** fuhr der Offizier fort, **[$_{2c}$** „trug sich zu, im Freiheitskriege der Niederländer, bei der Belagerung von Antwerpen durch den Herzog von Parma. Der Herzog hatte die Schelde, vermittelst einer Schiffsbrücke, gesperrt, und die Antwerpner arbeiteten ihrerseits, unter Anleitung eines geschickten Italieners, daran, dieselbe durch Brandner, die sie gegen die Brücke losließen, in die Luft zu sprengen. In dem Augenblick," **]$_{2c}$** meine Herren, **[$_{2c}$** da die Fahrzeuge die Schelde herab, gegen die Brücke anschwimmen, steht, das merken sie wohl, ein Fahnenjunker, auf dem linken Ufer der Schelde, dicht neben dem Herzog von Parma; jetzt, **]$_{2c}$** verstehen sie, **[$_{2c}$** jetzt geschieht die Explosion; und der Junker, Haut und Haar, samt Fahne und Gepäck, und ohne daß ihm das Mindeste auf dieser Reise zugestoßen, steht auf dem Rechten. **]$_{2c}$**
Und die Schelde ist hier, wie Sie wissen werden, einen kleinen Kanonenschuß breit.
„Haben Sie verstanden?"
Himmel, Tod und Teufel! rief der Landedelmann ...**]$_1$** (Kleist 1990, S. 378 f.)

(7) **[$_1$ [$_2$** ...Der Alte sah mich mit einem verständnisvollen Lächeln an: „Nun also!" sagte er.
[$_3$ „In der Mitte des vorigen Jahrhunderts, oder vielmehr, um genauer zu bestimmen, vor und nach derselben, gab es hier einen Deichgrafen, der von Deich- und Sielsachen

mehr verstand, als Bauern und Hofbesitzer sonst zu verstehen pflegen; aber es reichte doch wohl kaum; denn was die studierten Fachleute darüber niedergeschrieben, davon hatte er wenig gelesen; sein Wissen hatte er sich, wenn auch von Kindesbeinen an, nur selber ausgesonnen." $]_3$ Ihr hörtet wohl schon, Herr, die Friesen rechnen gut, und habet auch wohl schon über unseren Hans Mommsen von Fahretoft reden hören, der ein Bauer war und doch Boussolen und Seeuhren, Teleskopen und Orgeln machen konnte. Nun, $[_3$ ein Stück von solch einem Manne war auch der Vater des nachherigen Deichgrafen gewesen; freilich wohl nur ein kleines ...$]_3$ ...$]_2$ $]_1$ (Storm 1988, S. 639)

5.8 Nicht-erzählende Passagen

Es gibt Situationen, in denen der Erzähler seine Geschichte unterbricht, um diese zu kommentieren – allerdings ohne eine Figur einer anderen Erzählebene in der fiktionalen Welt anzusprechen (s. Abschnitt 3.4). Solche Formen von Aphorismen, Motti, Kommentare, Urteile, Gedanken und – auf primärer Erzählebene – Formen der Ansprache an den fiktionalen Adressaten (vgl. Schmid 2014, S. 7) werden als Teil der aktuellen Erzählebene verstanden und nicht als eigenständige Erzählebene betrachtet. Da es in einigen Fällen – etwa um das Urteil des Erzählers mit der Handlung zu vergleichen – aufschlussreich sein kann, annotieren wir diese Ausdrücke als ‚nicht erzählend'. Hierfür verwenden wir eckige Klammern und den Buchstaben n, womit angedeutet wird, dass solche Textstellen weder eine eigenständige Erzählebene bilden noch zur übergeordneten Erzählebene gehören (s. Abschnitt 5.7). Mit öffnenden Klammern wird der Beginn und mit schließenden Klammern das Ende des Ausdrucks markiert. Die Beispiele (8), (9) und (10) enthalten Elemente, die einen fiktiven Empfänger ansprechen und daher als nicht erzählter Teil betrachtet werden können.

(8) $[_1$ That puzzled the Leopard and the Ethiopian, but they set off to look for the aboriginal Flora, and presently, after ever so many days, they saw a great, high, tall forest full of tree trunks all 'sclusively speckled and sprottled and spottled, dotted and splashed and slashed and hatched and cross-hatched with shadows. $[_n$ (Say that quickly aloud, and you will see how very shadowy the forest must have been.)$]_n$ ...$]_1$(Kipling 1968, S. 47)

(9) $[_1$ Das war die erste lange Trennung, fast auf zwölf Stunden. $[_n$ Arme Effi! $]_n$ Wie sollte sie den Abend verbringen? Früh zu Bett, das war gefährlich, dann wachte sie auf und konnte nicht wieder einschlafen und horchte auf alles. Nein, erst recht müde werden und dann ein fester Schlaf, das war das Beste. Sie schrieb einen Brief an die Mama und ging dann zu Frau Kruse, deren gemütskranker Zustand – sie hatte das schwarze Huhn

oft bis in die Nacht hinein auf ihrem Schoß – ihr Teilnahme einflößte. $]_1$ (Fontane 1998, S. 79)[10]

(10) $[_1$ In the days when everybody started fair, $[_n$ Best Beloved $]_n$, the Leopard lived in a place called the High Veldt. $[_n$, Member $]_n$ it wasn't the Low Veldt, or the Bush Veldt, or the Sour Veldt, but ...$]_1$ (Kipling 1968, S. 41)

Primärliteratur

Abbé Prévost (2008). *Manon Lescaut*. Übers. von Walter Widmer. Zürich: Diogenes.

Calvino, Italo (1983). *Wenn ein Reisender in einer Winternacht*. Wien/München: Carl Hanser Verlag.

Fontane, Theodor (1998). *Effi Briest*. Hrsg. von Christine Hehle. Bd. 8. Große Brandenburger Ausgabe. Berlin: Aufbau-Verlag.

Frisch, Max (1954). *Stiller*. Frankfurt am Main: Suhrkamp.

Goethe, Johann Wolfgang (1899). *Die Leiden des jungen Werther*. i. A .d. Großherzogin Sophie von Sachsen. Weimar: Hermann Böhlaus Nachfolger, S. 1–192.

Joyce, James (1967). „Eveline". In: *Dubliners*. Hrsg. von Robert Scholes. Bungay/Suffolk: The Chaucer Press, S. 37–44.

Kipling, Rudyard (1968). *Just So Stories. For Little Children*. London/Toronto/Melbourne: Macmillan, S. 41–59.

Kleist, Heinrich von (1990). *Unwahrscheinliche Wahrhaftigkeiten*. Hrsg. von Klaus Müller-Salget. Bd. 3: Erzählungen, Anekdoten, Gedichte, Schriften. Sämtliche Werke und Briefe in vier Bänden. Frankfurt am Main: Deutscher Klassiker Verlag, S. 376–379.

Melville, Herman (2002). *Moby-Dick*. Hrsg. von Herhel Parker und Harrison Hayford. New York/London: Norton & Company.

Mörike, Eduard (2005a). *Das Stuttgarter Hutzelmännlein*. Hrsg. von Matthias Mayer. Bd. 6.1: Erzählungen, Erster Teil. Werke und Briefe. Stuttgart: Klett-Cotta, S. 119–168.

Mörike, Eduard (2005b). *Lucie Gelmeroth*. Hrsg. von Matthias Mayer. Bd. 6.1: Erzählungen, Erster Teil. Werke und Briefe. Stuttgart: Klett-Cotta, S. 11–29.

Storm, Theodor (1987). *Pole Poppenspäler*. Hrsg. von Karl Ernst Laage und Dieter Lohmeier. Bd. 2: Novellen 1867–1880. Sämtliche Werke in vier Bänden. Frankfurt am Main: Deutscher Klassiker Verlag, S. 164–221.

Storm, Theodor (1988). *Der Schimmelreiter*. Hrsg. von Karl Ernst Laage. Bd. 3: Novellen 1881–1888. Sämtliche Werke in vier Bänden. Frankfurt am Main: Deutscher Klassiker Verlag, S. 634–756.

Tschechow, Anton Pawlowitsch (1968). „Tschechows Erzählungen". In: *Das Gewinnlos*. Übers. von Leo Borchard und L. Flachs-Fokschneanu. Harenberg, S. 63–67.

10 Erläuterung zu Beispiel 9: Die Sätze „Früh zu Bett, das war gefährlich, dann wachte sie auf und konnte nicht wieder einschlafen und horchte auf alles. Nein, erst recht müde werden und dann ein fester Schlaf, das war das Beste" sind erlebte Rede und daher kein wertender Kommentar des Erzählers wie „Arme Effi!".

Sekundärliteratur

Abbot, H. Porter (2014). „Narrativity". In: *Handbook of Narratology*. Hrsg. von John Pier und Wolf Schmid. Berlin: De Gruyter, S. 587–607.

Coste, Didier und John Pier (2009). „Narrative Levels". In: *Handbook of Narratology*. Hrsg. von John Pier, Wolf Schmid und Jörg Schönert. Berlin: De Gruyter, S. 295–309.

Fludernik, Monika (2006). *Einführung in die Erzähltheorie*. Darmstadt: WBG.

Genette, Gérard (1989). *Paratexte. Das Buch vom Beiwerk des Buches*. Frankfurt am Main/New York: Campus Verlag.

Genette, Gérard (1994). *Die Erzählung*. Übers. von Andreas Knop. München: Fink.

Igl, Natalie (2018). „Erzähler und Erzählerstimme". In: *Grundthemen der Literaturwissenschaft: Erzählen*. Hrsg. von Martin Huber und Wolf Schmid. Berlin: De Gruyter, S. 127–149.

Jahn, Manfred (2017). *Narratology. A Guide to the Theory of Narrative*. Köln: Universität Köln. URL: http://www.uni-koeln.de/%7Eame02/pppn.htm (besucht am 1. Juni 2020).

Ketschik, Nora, Benjamin Krautter, Sandra Murr und Yvonne Zimmermann (2019). „Annotation Guideline No. 4: Annotating Narrative Levels in Literature". In: *Cultural Analytics*: *A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.055.

Klausnitzer, Ralf (2013). „Institutionalisierung und Modernisierung der Literaturwissenschaft seit dem 19. Jahrhundert". In: *Handbuch Literaturwissenschaft*. Hrsg. von Thomas Anz. Bd. 3. Metzler, S. 70–147.

Lahn, Silke und Jan-Christoph Meister (2013). *Einführung in die Erzähltextanalyse*. 2. Aufl. Stuttgart: Metzler.

Margolin, Uri (2014). „Narrator". In: *Handbook of Narratology*. Hrsg. von Peter Hühn, Jan Christoph Meister, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press. URL: https://www.lhn.uni-hamburg.de/node/44/revisions/296/view.html (besucht am 1. Juni 2020).

Martínez, Matías und Michael Scheffel (2009). *Einführung in die Erzähltheorie*. München: C.H.Beck.

Meister, Jan Christoph (2014). „Narratology". In: *Handbook of Narratology*. Hrsg. von Peter Hühn, Jan Christoph Meister, John Pier, Wolf Schmid und Jörg Schönert. Berlin: De Gruyter, S. 623–645.

Pier, John (2014). „Narrative Levels". In: Hrsg. von Peter Hühn, Jan Christoph Meister, John Pier, Wolf Schmid und Jörg Schönert. Berlin: De Gruyter, S. 547–563.

Pier, John (2016). „Metalepsis". In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press. URL: https://www.lhn.uni-hamburg.de/node/51.html (besucht am 1. Juni 2020).

Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.

Ryan, Marie-Laure (1991). *Possible Worlds, Artificial Intelligence and Narrative Theory*. Bloomington, Indiana: Indiana University Press.

Ryan, Marie-Laure (2001). „The Narratorial Functions. Breaking Down a Theoretical Primitive". In: *Contemporary Narratology* 9.2, S. 146–152.

Schmid, Wolf (2003). „Narrativity and Eventfullness". In: *What is Narratology. Questions and Answers Regarding the Status of a Theory*. Hrsg. von Tom Kindt und Hans-Harald Müller. Berlin: De Gruyter, S. 17–35.

Schmid, Wolf (2014). *Elemente der Narratologie*. 3. Aufl. Berlin: De Gruyter.
Schmid, Wolf (2018). „Ereignis". In: *Grundthemen der Literaturwissenschaft*. Hrsg. von Martin Huber und Wolf Schmid. Berlin: De Gruyter, S. 312–333.

Nils Reiter, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler

Reaching out: Interdisziplinäre Kommunikation und Dissemination

Ein CRETA-Erfahrungsbericht

Zusammenfassung: In diesem Kapitel diskutieren wir Aktivitäten, die wir im Rahmen von CRETA etabliert haben, um einerseits die interdisziplinäre Kommunikation zu verbessern und andererseits CRETA-Erkenntnisse nach außen zu tragen. Konkret stellen wir ein mehrfach durchgeführtes *hackatorial* (Workshop „Maschinelles Lernen lernen"), einen Workshop zur Operationalisierung als Kernaufgabe für die Digital Humanities, sowie das *CRETA-Coaching* vor. Zu allen Aktivitäten sammeln wir unsere Ergebnisse und Erfahrungen in einem Fazit.

Abstract: This chapter presents various activities related to internal and external communication, including activities related to the dissemination of ideas developed in CRETA. Specifically, we present the 'hackatorial' (workshop "Learning machine learning"), a 'workshop on operationalization' as a core task for the digital humanities, and the 'CRETA coaching'. For all activities we collect our results and experiences in a conclusion.

1 Einleitung

Neben die inhaltlichen und fachlichen Herausforderungen treten im Bereich der Digital Humanities die, die sich aus der Interdisziplinarität der Projektteams ergeben. Da es sich auch bei CRETA um ein interdisziplinär zusammengesetztes Zentrum handelt, waren die Themen Kommunikation und Weiterbildung von Anfang an präsent.[1] Eine funktionierende interdisziplinäre Kommunikation ist dabei eine der größten Herausforderungen. Suditsch (2017) zufolge besteht das Problem im Kern darin, dass in interdisziplinären Projekten zwar Expertinnen und Experten

1 Aus diesem Grund findet sich auch eine stilisierte Sprechblase im CRETA-Logo: .

Nils Reiter, Gerhard Kremer, Kerstin Jung, Janis Pagel, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Axel Pichler, Stuttgart Research Center for Text Studies, Universität Stuttgart
Benjamin Krautter, Germanistisches Seminar, Universität Heidelberg

Open Access. © 2020 Nils Reiter, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler; publiziert von De Gruyter BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-019

zusammenarbeiten, diese aber in unterschiedlichen Fächern Expertinnen und Experten sind. Gerade die „[Elemente] (z. B. gemeinsamer Wissenshintergrund, gemeinsame Fachsprache), die innerhalb einer Disziplin als gesetzt gelten, müssen in interdisziplinärer Zusammenarbeit neu ausdiskutiert werden“ (Suditsch 2017, S. 2). Dies führt zum sogenannten Fachsprachen-Paradox:

> Fachsprachen sind paradox, da sie innerhalb von Fachgemeinschaften explizite und effiziente Verständigung ermöglichen, während sie interdisziplinäre Kommunikation anfällig für Missverständnisse machen und ineffizient sind. (Suditsch 2017, S. 41)

Fachsprachen sind deswegen anfällig für Missverständnisse, weil Ausdrücke oft mehrfach belegt sind: Mit dem Wort ‚Modell‘ z. B. verbinden sowohl Forschende aus der Computerlinguistik, Literaturwissenschaft und Philosophie etwas – aber nicht das gleiche.[2] Wenn ein Wort je nach Disziplin verschiedene Bedeutungen annimmt, ist das meistens nicht von Anfang an offensichtlich, sondern muss erst freigelegt werden. Dieses Freilegen ist ein mühsamer Prozess, da Fachsprachen von ihren Anwendenden stark internalisiert sind und wenig bewusst verwendet werden. Letzten Endes braucht es in solchen Projekten zumindest ein paar Mitarbeitende, die die jeweils andere Disziplin zu verstehen versuchen, und immer wieder nachbohren. Für die Verständigung auf gemeinsame Begriffe und ihre Bedeutung (‚Entwicklung einer gemeinsamen Sprache‘) ist der informelle Austausch von großer Wichtigkeit: „Interdisziplinäre Projekte sind für die meisten Beteiligten immer noch eine neue Erfahrung; die daraus entstehenden Unsicherheiten lassen sich auf informeller Ebene am schnellsten ausräumen“ (Lengwiler 2005, S. 54).

In CRETA werden diese Herausforderungen auf mehreren Ebenen adressiert. In regelmäßigem, bisher halbjährlichem Abstand treffen sich alle an CRETA Beteiligten zur sog. **CRETA-Werkstatt**. Bei diesen zwei- bis dreitägigen Veranstaltungen wird einerseits in Vorträgen aus aktuellen Arbeiten der Mitarbeitenden von CRETA berichtet. Andererseits wird das Treffen von der Arbeit in Kleingruppen bereichert, das etwa die andere Hälfte der Zeit einnimmt. Thema und Format der Gruppenarbeit können dabei sehr unterschiedlich sein. So kann es Gruppen geben, die an konkreten Annotationsaufgaben (oder -richtlinien) arbeiten, während andere aktuelle Forschungsthemen oder -publikationen aus ihrem Bereich diskutieren und wieder andere die nächsten Arbeitsschritte in ihren Projekten planen. Der Vorteil der Arbeit in Kleingruppen liegt nicht nur darin, dass so im Rahmen der Werkstatt häufig inhaltliche Ergebnisse produziert werden, son-

2 Siehe hierzu auch den Abschnitt „Hintergrund: Modelle und Modellierung in der algorithmischen Textanalyse“ in der Einleitung des Bandes auf Seite 12.

dern auch darin, dass sich auch diejenigen CRETA-Beteiligten aktiv in die CRETA-Werkstatt einbringen können, die keinen Vortrag halten. Der kleinere Rahmen einer Arbeitsgruppe macht es auch leichter, Unsicherheiten und Unklarheiten auszuräumen. Es wird dabei auf die Zugänglichkeit und Offenheit für den wissenschaftlichen Nachwuchs geachtet sowie auf eine möglichst diverse Zusammensetzung der Arbeitsgruppen (in Bezug auf Disziplinen, Hierarchiestufen und Geschlechter). Nicht zuletzt weil die CRETA-Beteiligten in Stuttgart über mehrere Standorte verteilt sind, fand auch – zusätzlich zu den Werkstätten – ein regelmäßiger Stammtisch statt, bei dem ein informeller Austausch möglich war.

Als projektinterne **Weiterbildungen** wurden zunächst einige Themen von CRETA-Beteiligten an andere CRETA-Beteiligte vermittelt (z. B. Versionsverwaltung mit git, Was ist eigentlich Hermeneutik?, Einführung in die Statistik, Maschinelles Lernen lernen, …). Den Lehrenden führten diese Aktivitäten regelmäßig vor Augen, welche Vorannahmen sie implizit treffen, die im interdisziplinären Kontext neu ausgehandelt werden müssen. Für die Lernenden bot sich die Gelegenheit, eben jene Vorannahmen aus ihnen fremden Disziplinen besser zu verstehen, und so auch fremd-disziplinäre Forschungspraktiken und deren Resultate besser einordnen zu können. Aufgrund der positiven Erfahrungen dieser Formate wurden sie dann ausgeweitet und auch CRETA-extern im Rahmen von Workshops und Tutorials angeboten.

In diesem Kapitel stellen wir eine Auswahl dieser über die Jahre weiterentwickelten Aktivitäten dar (Tabelle 1 zeigt eine Übersicht). In Abschnitt 2 präsentieren wir das von uns entwickelte ‚*hackatorial*'-Konzept, das wir CRETA-intern sowie im Rahmen mehrerer Digital-Humanities-Konferenzen erfolgreich durchgeführt haben. Ziel dabei war es, die Teilnehmenden mit Grundlagen maschineller Lernverfahren vertraut zu machen und ihnen ein *know how* zu vermitteln, mit Hilfe dessen sie a) sich auf informierte Weise den eigenen Aufgaben widmen und b) Ergebnisse maschineller Lernverfahren besser einschätzen können. In Abschnitt 3 besprechen wir ein darauf aufbauendes, ergänzendes Workshop-Konzept, das im Jahr 2020 zum ersten Mal durchgeführt wurde. Der Fokus lag dabei auf der Operationalisierung geisteswissenschaftlicher Forschungsfragen und den verschiedenen Möglichkeiten, sich diesem Problem zu nähern. Während die beiden Workshops sich an ein Publikum von 20 bis 35 Teilnehmenden richteten, haben wir 2019 auch ein stark individualisiertes Coaching-Programm angeboten, an dem vier Nachwuchswissenschaftlerinnen teilgenommen haben (Abschnitt 4). Dabei haben wir unsere – auch in diesem Buch gesammelte – Kompetenz bei der Operationalisierung gemeinsam mit den zu Coachenden auf neue Fragen bzw. Probleme angewendet, um ihnen konkrete Handlungsempfehlungen zu geben. Neben der praktischen Wissenserweiterung der Teilnehmenden an unseren Disseminationsprojekten konnte bei deren Entwicklung und Durchführung wiederum das

Tab. 1: Tutorials und Workshops zu CRETA-Themen. DH: Jahrestagung der *Alliance of Digital Humanities Organizations*, DHd: Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum, ESU: *European Summer University in Digital Humanities*, LSS-ML: *Late Summer School on Machine Learning for Language Analysis*, HCH: *Heidelberg Computational Humanities School*.

Datum	Ort	Titel	Verantwortliche
08.08.2017	DH, Montreal	From Texts to Networks: Combining Entity and Segment Annotations in the Analysis of Large Text Corpora	Nils Reiter, Maximilian Overbeck, Sandra Murr
26.02.2018	DHd, Köln	Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse	Nils Reiter, Nora Ketschik, Gerhard Kremer, Sarah Schulz
23.–27.07.2018	ESU, Leipzig	Reflektierte Textanalyse in den Digital Humanities	Nils Reiter, Sarah Schulz
26.–27.09.2018	LSS-ML, Köln	Learning Machine Learning	Nils Reiter
26.03.2019	DHd, Mainz	Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse	Gerhard Kremer, Kerstin Jung
15.07.2019	HCH, Heidelberg	Quantitative Drama Analytics	Benjamin Krautter, Janis Pagel, Nils Reiter, Marcus Willand
09.–17.09.2019	Stuttgart	CRETA-Coaching	Nils Reiter, Axel Pichler
09.–20.09.2019	Stuttgart	Herbstschule der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft	Sebastian Padó, Gerhard Kremer, Jonas Kuhn, Sybille Laderer, Nils Reiter, Sabine Schulte im Walde
WS 2019/20	TU Darmstadt	Seminar Deep Learning & Digital Humanities	Steffen Eger, Thomas Haider
03.03.2020	DHd, Paderborn	Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse	Gerhard Kremer, Kerstin Jung
03.03.2020	DHd, Paderborn	Vom Phänomen zur Analyse. Ein CRETA-Workshop zur reflektierten Operationalisierung in den DH	Nora Ketschik, Benjamin Krautter, Sandra Murr, Janis Pagel, Nils Reiter

CRETA-Team Erfahrungen sammeln, die zur Verfeinerung der eigenen Methodenarbeit beigetragen haben.

2 Hackatorial: Maschinelles Lernen lernen

2.1 Einleitung

Ein mehrmals durchgeführter, halbtägiger Workshop („Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse") machte Teilnehmende praxisbezogen mit dem Thema des maschinellen Lernens vertraut. Der Titel nutzt die Kurzform *hackatorial*, angelehnt an sogenannte *hackathons* (aus engl. *to hack* und *marathon*), bei denen große Gruppen in einem kurzen Zeitrahmen Softwareentwicklung betreiben. Unser Ansatz bei „Maschinelles Lernen lernen" ist es, dieses Prinzip mit einem Tutorial, also einer Anleitung zu verbinden.

Ziel dieses *hackatorials* war, den Teilnehmenden konkrete und praktische Einblicke in einen Standardfall automatischer Textanalyse zu geben. Am Beispiel der automatischen Erkennung von Entitätenreferenzen (siehe hierzu auch den Beitrag von Ketschik, Blessing et al. (2020), ab Seite 204 in diesem Band) ging das CRETA-Team auf allgemeine Annahmen, Verfahrensweisen und methodische Standards bei maschinellen Lernverfahren ein. Die Teilnehmenden konnten beim Bearbeiten von lauffähigem Programmiercode den Entscheidungsraum solcher Verfahren ausleuchten und bekamen einen zusammenhängenden Überblick von der manuellen Annotation ausgewählter Texte über die Feinjustierung der Lernverfahren bis zur Evaluation der Ergebnisse. Es wurden dabei keinerlei Vorkenntnisse über maschinelles Lernen oder Programmierkenntnisse vorausgesetzt.

Neben einem Python-Programm für das automatische Annotieren von Entitätenreferenzen, mit und an dem während des Tutorials gearbeitet wurde, stellten wir ein heterogenes, manuell annotiertes Korpus (bestehend aus verschiedenen, separaten Teilkorpora) sowie Routinen zur Evaluation und zum Vergleich von Annotationen zu Verfügung. Das Korpus enthält Entitätenreferenzen, die in CRETA annotiert wurden und deckt Texte verschiedener Disziplinen und Sprachstufen ab.

Durch Einblick in die technische Umsetzung bekamen die Teilnehmenden ein Verständnis für die Grenzen und Möglichkeiten der Automatisierung, das sie dazu befähigt, zum einen das Potenzial solcher Verfahren für eigene Vorhaben realistisch(er) einzuschätzen, zum anderen aber auch Ergebnisse, die auf Basis

Tab. 2: Korpora die im *hackatorial* verwendet wurden. Detaillierte Beschreibungen zu den Kategorien, Korpora und Annotationen können in Ketschik, Blessing et al. 2020 nachgeschlagen werden.

Korpus	Sprache	Tokens	Entitätenreferenzen
Werther	Neuhochdeutsch	41 505	331
Parzival	Mittelhochdeutsch	30 491	2001
Bundestagsdebatten	Neuhochdeutsch	6371	488

solcher Verfahren erzielt wurden, angemessen hinterfragen und interpretieren zu können.

2.2 Entitätenreferenzen und Korpora

Als Ausgangspunkt des Hackatorials diente uns das Konzept der Entität und ihrer Referenz, das in Ketschik, Blessing et al. 2020, S. 205 ff., ausgeführt und hier nur kurz wiederholt wird. Es wurde im Rahmen von CRETA bewusst weit gefasst und damit anschlussfähig für verschiedene Forschungsfragen aus den geistes- und sozialwissenschaftlichen Disziplinen. Insgesamt wurden fünf verschiedene Entitätenklassen bestimmt: PER (Personen/Figuren), LOC (Orte), ORG (Organisationen), abstrakte Konzepte (CNCs) und Werke (WRK). ‚Entität' steht folglich für ein reales, fiktives oder mögliches Objekt, auf das von der Textoberfläche mittels eines einzelnen Wortes oder einer abgegrenzten Wortfolge referiert werden kann. Unter Entitätenreferenzen verstehen wir Ausdrücke, die auf eine Entität in der realen oder fiktiven Welt referieren. Das sind zum einen Eigennamen (*named entities*, z. B. „Peter"), zum anderen Gattungsnamen (z. B. „der Bauer"), sofern diese sich auf eine konkrete Instanz der Gattung beziehen. In CRETA wurden vier Korpora mit Entitätenreferenzen annotiert, von denen drei im *hackatorial* Verwendung fanden (Tabelle 2 zeigt eine Übersicht über die Charakteristika der Korpora).

2.3 Ablauf

Der Ablauf des Tutorials orientierte sich an sog. *shared tasks* aus der Computerlinguistik (vgl. Willand et al. 2020, S. 396 f.), wobei der Aspekt des Wettbewerbs im Tutorial vor allem spielerischen Charakter hatte. Bei einem traditionellen *shared task* arbeiten die teilnehmenden Teams, oft auf Basis gleicher Daten, an Lösungen für eine einzelne gestellte Aufgabe. Solch eine definierte Aufgabe kann z. B. *part-of-speech-tagging*, also das automatische Erkennen von Wortarten, sein. Durch ei-

ne zeitgleiche Evaluation auf demselben Goldstandard können die entwickelten Systeme direkt verglichen werden. In unserem Tutorial setzten wir dieses Konzept live und vor Ort um.

Die Teilnehmenden versuchten nach einer Einführung in die Annotationen und die Annotationspraxis, selbständig und unabhängig voneinander die optimale Kombination aus (a) maschinellem Lernverfahren und (b) Text- und Wort-Merkmalen (z. B. Groß- und Kleinschreibung oder Wortlänge) für das Training eines Modells zur Erkennung von Enitätenreferenzen zu wählen. Dabei war das Ziel, auf einem für die Teilnehmenden neuen Datensatz zu den Ergebnissen zu kommen, die dem Goldstandard der manuellen Annotation am ähnlichsten waren.

Das bedeutet, dass der Einfluss von berücksichtigten Merkmalen auf das Erkennen von Entitätenreferenzen empirisch getestet werden konnte. Dabei waren Intuitionen über die Daten und das annotierte Phänomen hilfreich, da simplem Durchprobieren aller möglichen Kombinationen zeitliche Grenzen gesetzt waren. Zusätzlich wurden bei jedem Testlauf Informationen über die Entscheidungen protokolliert, um die Erklärbarkeit der Ergebnisse zu unterstützen.

Wir verzichteten bewusst auf eine grafische Benutzerschnittstelle, um den Teilnehmenden den ‚Blick hinter den Vorhang' zu ermöglichen. Viele grafische Benutzeroberflächen verstecken die Komplexität der beim *machine learning* nötigen Entscheidungsprozesse (vgl. Reiter, Kuhn et al. 2017). Ziel des *hackatorial* war es ja, gerade diese sichtbar zu machen. Die Teilnehmenden bearbeiteten das (Python-)Programm also nach einer Einführung und unter Aufsicht direkt in einem Texteditor. Vorkenntnisse in Python waren dabei nicht nötig: Das von uns zur Verfügung gestellte Programm war so aufgebaut, dass auch Neulinge relativ schnell die zu bearbeitenden Teile verstehen und damit experimentieren konnten, indem Features zu- und abgeschaltet werden konnten (im einfachsten Fall durch das Setzen von Kommentarzeichen). Teilnehmende mit Erfahrungen im Programmieren konnten komplexere Merkmale verwenden, indem sie eigene Python-Funktionen implementierten.

3 Workshop: Operationalisierung in den Digital Humanities

3.1 Einleitung

Spätestens seit Franco Morettis 2013 veröffentlichtem Literary-Lab-Pamphlet *‚Operationalizing': or, the Function of Measurement in Modern Literary Theory* (Moretti 2013) ist die Operationalisierung geisteswissenschaftlicher Kategorien und

Konzepte zentrales Thema im Bereich der Digital Humanities (cf. Jacke 2014; Jannidis et al. 2016; Gius 2019; Pichler und Reiter 2020). Mit Moretti gesprochen gilt es, eine Brücke zu schlagen, und zwar von „concepts to measurement, and then to the world“ (Moretti 2013, S. 1). Der auf der DHd 2020 erstmalig durchgeführte Workshop adressierte diese spezifische Herausforderung der Messbarmachung theoretischer Konzepte anhand von drei Anwendungsbeispielen, die literatur- und sozialwissenschaftliche Phänomene umfassten.

Ziel des Workshops war es einerseits, die Schnittstelle zwischen computergestützten Methoden und geisteswissenschaftlichen Kategorien als wichtigen Teil der Forschungsarbeit herauszustreichen. Dadurch sollte ein Bewusstsein für die disziplinären Unterschiede und Herausforderungen ausgebildet werden, die es bei der Verschränkung von Informationstechnik und Geisteswissenschaften zu überwinden gilt. Andererseits wurden typische Problemstellungen auch in der tatsächlichen Arbeitspraxis adressiert und Lösungsmöglichkeiten aufgezeigt. Dazu nutzten wir drei im Rahmen von CRETA umfangreich bearbeitete und beforschte Anwendungsszenarien: die Extraktion von Entitäten und Entitätenreferenzen (Ketschik, Blessing et al. 2020), die Segmentierung von Erzähltexten in verschiedene Erzählebenen (Barth 2020; Ketschik, Murr et al. 2020; Willand et al. 2020), sowie das holistische Textphänomen der sogenannten ‚Wertherness‘ (Richter 2020).[3] Die theoretische und praktische Auseinandersetzung mit diesen Phänomenen hatte zum Ziel, die einzelnen Schritte auf dem Weg zu einer angemessenen und reflektierten Operationalisierung geisteswissenschaftlicher Konzepte aufzuzeigen und mit Blick auf die Fragestellung und zu erwartenden Ergebnisse immer wieder kritisch zu hinterfragen. Denn nur eine durchdachte Operationalisierung erlaubt im Anschluss an die quantitative Auswertung eine adäquate und auch für die Fachdisziplinen relevante Interpretation der Ergebnisse.

3.2 Verschiedene Ansätze für verschiedene Phänomene

Die von uns gewählten Anwendungsszenarien waren als ‚Prototypen-Repertoire‘ gedacht, das verschiedene Verfahrensweisen der Operationalisierung vorstellt, diskutiert und reflektiert, zugleich aber auch Möglichkeiten der Übertragung auf andere Anwendungsfälle exemplarisch aufzeigt. Die Phänomene und ihre damit zusammenhängenden Analyseeinheiten sind in Tabelle 3 zusammengefasst.

3 Alle hier referenzierten Arbeiten sind Teile des vorliegenden Bandes.

Tab. 3: Phänomene, die beim Workshop zur Operationalisierung verwendet wurden

Phänomen	Einheit	Datenquelle/Referenz
Entitätenreferenzen	Wörter/Wortgruppen	Ketschik, Blessing et al. (2020)
Erzählebenen	Textsegmente	Teil IV: Ketschik, Murr et al. (2020) und Willand et al. (2020)
Wertherness	Ganze Texte	Richter (2020)

Der Workshop konzentrierte sich auf zwei verschiedene Ansätze der Operationalisierung, die sich – abhängig von Forschungsfrage und -phase – gegenseitig ergänzen können.[4] Der erste Ansatz stellt die Definition von Konzepten oder Kategorien mittels Annotationen in den Mittelpunkt. (Manuelle) Annotationen dienen hierbei zur Schärfung der untersuchten Konzepte. Probleme, Unklarheiten oder nicht beachtete Teilphänomene, die der Annotationsprozess aufdecken kann, werden in die Definition zurückgespielt, die Konzepte und Kategorien geschärft. Dies hilft einerseits eine größere intersubjektive Übereinstimmung der Annotationen zu erzielen, kann aber auch in die Theoriediskussion einfließen und diese bereichern (vgl. Gius und Jacke 2017 oder Pagel et al. 2020, ab Seite 125 in diesem Band). Die intensive Auseinandersetzung mit dem Material und den annotierten Instanzen gibt zudem Impulse für die computergestützte Operationalisierung.

Der zweite Ansatz schlägt eine indirekte Operationalisierung der betrachteten Phänomene vor. Dabei werden mehrere messbare Eigenschaften betrachtet, die mit dem zu operationalisierenden Konzept zwar verwandt sind, es aber jeweils einzeln nicht vollständig abdecken (vgl. Sack 2011; Reiter und Willand 2018). Bei dieser indirekten Annäherung ist vor allem die Gesamtschau der verschiedenen Einflussfaktoren aufschlussreich. Dabei lassen sich strukturelle, holistische und linguistische Eigenschaften von Texten zusammenführen.

Wie schon im *hackatorial* verwendeten wir die in CRETA annotierten **Entitätenreferenzen** in mittelhochdeutschen Artusromane und Bundestagsdebatten (s. o.). Sie dienten hier als Beispiel für die Operationalisierung von lokalen Textphänomenen.

Als zweites Phänomen beschäftigten wir uns mit der narratologischen Bestimmung von **Erzählebenen**, die Erzähltexte seriell oder ineinander verschachtelt segmentieren. Erzählebenen stehen im Zentrum des in Teil IV beschriebenen *shared tasks*, (Willand et al. 2020, ab S. 391 in diesem Band). Grundlage der im

4 Die Ansätze spiegeln die beiden Pfade des von Pichler und Reiter (2020) ab Seite 43 in diesem Band diskutierten Workflows.

Workshop verwendeten Daten waren die Richtlinien, die in Ketschik, Murr et al. 2020 ab S. 391 in diesem Band beschrieben werden. Im Gegensatz zu Entitätenreferenzen handelt es sich um ein stark variierendes Phänomen – eine Erzählebene kann eine einzelne Äußerung einer Figur umfassen, oder den Gutteil eines Textes. Außerdem können Erzählebenen in andere Erzählebenen eingebettet werden, sie sind also ggf. verschachtelt.

Das letzte Anwendungsbeispiel rückte ein holistisches Textphänomen, die sogenannte **‚Wertherness'**, ins Zentrum der Überlegungen. Die ‚Wertherness' bezeichnet eine Menge an Texteigenschaften, die Texte im Anschluss an Goethes 1774 veröffentlichten Briefroman *Die Leiden des jungen Werthers* als literarische Adaptationen, also ‚Wertheriaden', identifizieren. Ursächlich dafür sind verschiedene Bezugnahmen auf den Ursprungstext, die formaler (etwa Briefroman) wie inhaltlicher (etwa Rolle der Natur, Verhältnis Subjekt-Gesellschaft, Dreiecksbeziehung) Natur sein können. Für eine computergestützte Analyse der ‚Wertheriaden' bietet sich hier eine indirekte Operationalisierung an – welche messbaren Eigenschaften deuten auf eine Wertheriade hin? Welche der Texteigenschaften sind messbar? Neben der Identifikation der verschiedenen Eigenschaften, gilt es zusätzlich, ihre Kombination in bekannten ‚Wertheriaden' auszuloten.

3.3 Ablauf des Workshops

Der Ablauf des Workshops teilte sich in Theorie- und Praxisphasen. Anhand der drei genannten Phänomene wurden zu Beginn Herausforderungen und Problemstellungen der Operationalisierung geisteswissenschaftlicher Konzepte reflektiert. Im Anschluss daran konnten die Teilnehmenden eigenen Interessen folgend eines der Beispiele auswählen und praktisch bearbeiten.

Am Anfang stand die manuelle Annotation: Anhand eines Textauszugs und vorbereiteter Annotationsrichtlinien annotierten sie ihr gewähltes Textphänomen und präzisierten und erweiterten dabei die Richtlinien. In einer ersten Diskussionsrunde wurden die gesammelten Erfahrungen und Ergebnisse sowie der Annotationsprozess besprochen.

Für die anschließende Erprobung der indirekten Operationalisierung stellten wir einen ‚Baukasten' aus verschiedenen Python- und R-Skripten web-basiert in einem Jupyter-Notebook[5] zur Verfügung. Die Skripte waren auf das gewählte Phänomen zugeschnitten und boten den Teilnehmenden die Möglichkeit einer computergestützten Annäherung daran. Die Teilnehmenden konnten in Kleingrup-

5 https://jupyter.org/

pen arbeiten und dabei auf verschiedene Methoden der Datenanalyse zurückgreifen, Parameter justieren, manuelle Eigenschaften an- oder abwählen und die Ergebnisse einer maschinellen Klassifikation visualisieren. Hierbei konnten die Teilnehmenden auch das in der praktischen Annotationsrunde generierte Vorwissen einbringen. Die Verwendung von Jupyter-Notebooks erlaubte eine iterative Vorgehensweise, so dass die Teilnehmenden die benötigten Schritte einzeln und bei Bedarf erneut ausführen konnten, um die Ergebnisse zu aktualisieren. Außerdem erlaubte die Visualisierung der Ergebnisse ein direktes Feedback zur Parametrisierung, wodurch nachvollziehbar wurde, ob das gewählte Phänomen sinnvoll mit den gewählten Einstellungen zu fassen war. Eine abschließende Diskussion wertete die gesammelten Ergebnisse aus und erörterte, inwieweit sich die Anwendungsfälle angemessen modellieren ließen.

4 Coaching

4.1 Ziele und Motivation

Eine Erfahrung, die wir sowohl CRETA-intern als auch bei den Workshops gemacht haben, war, dass die Teilnehmenden oft sehr spezifische Probleme zu lösen versuchten und dass die Übertragung des Gelernten auf die konkreten Fragen, Probleme und Daten eine Herausforderung darstellte. Durch den Fokus auf einzelne methodische Bausteine konnte die Einbettung in den Gesamtzusammenhang einer vollständigen Forschungsarbeit im Rahmen eines Workshops nur lückenhaft und prototypisch behandelt werden. Daraus ergab sich der Bedarf für ein weiteres Schulungsangebot, bei dem wir auf die individuellen Herausforderungen der Forschungsarbeiten und Rahmenbedingungen für die Forschenden eingehen wollten.

Grundidee des Coachings war, mit den Gecoachten (im Folgenden ‚Coachees') in individuellen Gesprächsrunden einen konkreten Plan zu erarbeiten, wie sie sich ihrer textanalytischen Kernfrage nähern können. Dazu gehörte etwa die Einteilung in geeignete Teilfragen, deren konkrete Operationalisierung, die Identifikation von relevanten Ressourcen (Tools/Methoden/Korpora), sowie ggf. Publikationsstrategien für Teilergebnisse. Ziel des Coachings war also, einen konkreten Arbeitsplan zu entwickeln, auf Basis dessen die Coachees dann (weiter-)arbeiten konnten – die Arbeit am Projekt sollte also noch relativ am Anfang stehen. Zielgruppe waren dementsprechend Promovierende, die einen geisteswissenschaftlichen Hintergrund hatten und eine grundsätzlich geeignete textanalytische Frage-

stellung verfolgten oder verfolgen wollten.[6] Interessenten mussten sich mit einem Exposé zu ihrem Projekt, einem kurzen Lebenslauf sowie ggf. bereits erfolgten einschlägigen Publikationen bewerben.

Die Coaches wurden so gewählt, dass eine Bandbreite an Themen und Disziplinen von ihnen vertreten werden konnte. Neben Beteiligten in CRETA konnte auch eine externe Kooperationspartnerin gewonnen werden, mit der wir in der Vergangenheit bereits in verschiedenen Formen kooperiert hatten. Als Coaches standen bereit: Evelyn Gius (Digital Philology/Neuere Deutsche Literaturwissenschaft, TU Darmstadt), Gerhard Kremer (Computerlinguistik, Universität Stuttgart), Jonas Kuhn (Computerlinguistik, Universität Stuttgart), Janis Pagel (Computerlinguistik, Universität Stuttgart), Axel Pichler (Philosophie/Literaturwissenschaft, Universität Stuttgart), Nils Reiter (Computerlinguistik/Digital Humanities, Universität Stuttgart), Gabriel Viehhauser (Digital Humanities, Universität Stuttgart).

4.2 Ablauf

Das CRETA-Coaching bestand konkret aus fünf Terminen, die über einen Zeitraum von etwa zehn Tagen gestreut waren:

- Einem einwöchigen Kurs zur reflektierten Textanalyse, der auch im Rahmen der DGfS-CL-Herbstschule 2019[7] angeboten wurde (‚Reflected Text Analytics beyond Linguistics', Nils Reiter). Darin wurden einige methodische Grundlagen, vor allem zu Annotation, maschinellem Lernen und Evaluation, gelegt. Ausgehend von etablierten textanalytischen Verfahren aus der Computerlinguistik wurde dabei auch erarbeitet, wie diese Verfahren für andere textwissenschaftliche Fragestellungen eingesetzt werden können. Im Kurs wurde auch das oben beschriebene *hackatorial* eingesetzt.
- Im Fokus der zweitägigen CRETA-Werkstatt (s. o.) stand der Austausch über konkrete Arbeiten und Daten in größerer Runde. Neben Arbeiten von CRETA-Beteiligten konnten die Coachees ihre Projekte vorstellen und im Anschluss in themenspezifischen Arbeitsgruppen darüber diskutieren.
- In drei individuellen Coaching-Terminen trafen die Coachees mit jeweils zwei Coaches zusammen, die sich auf Basis der vorab eingereichten Exposés auch

6 Eine frühzeitige Einbeziehung der Betreuenden wurde den Promovierenden empfohlen und hat auch stattgefunden.

7 https://dgfs-clschool19.github.io

Tab. 4: Zeitliche Struktur des Coachings

	Montag - Freitag	**Montag**	**Dienstag**
Vormittag	Zwei individuelle Coaching-Termine	CRETA-Werkstatt	CRETA-Werkstatt
Nachmittag	Kurs „Reflected Text Analytics beyond Linguistics“	CRETA-Werkstatt	Abschluss-Coaching

bereits mit den Projekten auseinandergesetzt hatten. Die Coaches wechselten sich ab, so dass eine Vielfalt an Perspektiven berücksichtigt werden konnte.

Tabelle 4 gibt einen Überblick über die zeitliche Struktur des Coachings. Der Abstand zwischen einzelnen Coaching-Terminen betrug einige Tage, damit a) die Coachees genug Zeit hatten, die Gespräche zu reflektieren, und b) um die Möglichkeit zu haben, kleinere Experimente oder Analysen zwischen den Terminen durchzuführen (z. B. eine Statistik über die Wortformen zu extrahieren).

Zur Vorbereitung und Strukturierung der Coachings wurden den Coaches im Vorfeld Leitfragen ausgehändigt, die im Gespräch relevant sein könnten (Abbildung 1). Die Leitfragen deckten verschiedene Aspekte sowie wiederkehrende Probleme bei DH-Forschungen ab, die je nach Forschungsfrage unterschiedlich relevant waren.

4.3 Erfahrungen

Aus CRETA-Perspektive war das Coaching in mehrfacher Hinsicht ertragreich. Aufgrund der unterschiedlichen disziplinären Kontexte, aus denen die Coachees stammten, sowie der Tatsache, dass sich die Coachings bereits stark an konkreten Tasks orientierten sollten bzw. diese zu entwickeln suchten, war es insbesondere notwendig, eingangs den fachspezifischen Rahmen der jeweiligen Projekte abzustecken. Dazu standen zwar erste Informationen zu den von den Coachees als relevant erachteten Methoden und Kontexten in Form der eingereichten Exposées und Lebensläufe zur Verfügung. Nicht absehbar waren jedoch die arbeitspraktischen Konsequenzen im Umgang mit diesen fachspezifischen Voraussetzungen und ihr Verhältnis zu dem in CRETA entwickelten Workflow (Pichler und Reiter 2020).

Erstens galt es daher zu klären, ob und wie sich die von den jeweiligen Projekten vorgegebenen Fragestellungen operationalisieren ließen. D. h. es war zu überprüfen, ob sich die Leitfragen in Teilfragen unterteilen ließen, die weiter in prak-

- Task
  - Was genau ist der Task?
  - Ist er verwandt mit existierenden Tasks, z. B. aus der Computerlinguistik oder den Digital Humanities?
  - Gibt es ein Evaluationsszenario für den Task?
  - Wie generisch ist der Task? Wäre er auch auf anderen Daten/Epochen/Gattungen/Textsorten interessant?
  - Existieren bereits Annotationsrichtlinien oder -schemata?

- Bezug zu CRETA
  - Gibt es Anknüpfungspunkte zu CRETA?
  - Berührt die Fragestellung das Forschungsgebiet eines/einer CRETA-Angehörigen?
  - Gibt es Möglichkeiten für Kooperationen?

- Rahmenbedingungen
  - Was sind die vorliegenden Rahmenbedingungen für das Projekt?
  - Wie ist das Projekt/Zentrum/Institut zusammengesetzt, an dem das Projekt durchgeführt werden soll?
  - Gibt es vor Ort Kooperationen mit der Informatik?

- Daten
  - Woraus genau bestehen die Daten, wie sind sie zusammengesetzt?
  - Liegen die Daten in digitaler Form vor? Wie groß ist die Datenmenge?

- Annotationen
  - Wurden bereits Beispieldaten annotiert?
  - Könnten testweise (z. B. zum zweiten Termin) Annotationen erzeugt werden, die dann gemeinsam diskutiert werden können?
  - Wie schwer ist die Annotationsaufgabe? Wie hoch ist das *inter-annotator agreement*?
  - Kann die Annotation von Laien oder Experten durchgeführt werden?

- Automatische Erkennung
  - Welche (linguistischen) Vorverarbeitungsschritte würden bei einer Operationalisierung helfen? Welche Tools existieren dafür, sind verfügbar, für welche Sprache(n)?
  - Gibt es Möglichkeiten, annotierte Daten indirekt zu generieren? Sind Fehler in diesen Daten schlimm?
  - Wie könnten Visualisierungen des Ergebnisses aussehen?
  - Was sind erwartete Ergebnisse, wie sieht ein mögliches Ergebnis aus?

- Verschiedenes
  - Was sind mögliche Publikationsorte für (Zwischen-)Ergebnisse?
  - Welche Vorkenntnisse liegen seitens der Coachees vor?

Abb. 1: Leitfragen, die den Coaches im Vorfeld des Coachings ausgehändigt wurden

tisch umsetzbare Aufgaben auf Basis konkreter und präziser Handlungsanweisungen überführt werden konnten. Dabei war nicht nur die Operationalisierbarkeit selbst zu berücksichtigen, sondern auch zu klären, wie eine derartig in Teilaufgaben unterteilte Fragestellung mit den fachspezifischen Konventionen zu vereinbaren war. Zu letzteren zählen nicht nur die bereits angesprochenen Kontexte und fachspezifischen Methoden, sondern auch soziale Konventionen inklusive ihrer Auswirkungen auf die Karriereplanung. Damit diese Fragen in drei Coaching-Sitzungen adäquat behandelt werden konnten, war eine eingehende Auseinandersetzung mit den Projekten und der Austausch unter den Coaches im Vorfeld nötig.

Zweitens stellte das Coaching eine Möglichkeit dar, den CRETA-Workflow auf im Entstehen begriffene Forschungsprojekte anzuwenden. Die Coachings stellten somit – nach der CRETA-internen Anwendung – einen zweiten Praxistest für das besagte Verfahren dar. Dabei hat der CRETA-Workflow sowohl dazu beigetragen, die Gespräche zu strukturieren als auch die Zusammenhänge der einzelnen Arbeitsschritte zu verdeutlichen. Nicht zuletzt haben die im Coaching diskutierten Fragestellungen wesentlich zu einer Schärfung des in CRETA entwickelten Workflow (Pichler und Reiter 2020) beigetragen.

Wie die im folgenden partiell abgedruckten Erfahrungsberichte von zwei Teilnehmerinnen, die ursprünglich im Weblog des DHd-Verbandes[8] veröffentlicht wurden, belegen, ist es im Zuge der Coachings gelungen, die beiden oben genannten Punkte zur Zufriedenheit der Teilnehmerinnen zusammenzuführen:

> Das Coaching war – metaphorisch gesagt – ein Schlaraffenland für junge WissenschaftlerInnen. [...]
> Die Teilnahme an der CRETA-Werkstatt stellte sich somit als krönender Abschluss heraus, bei dem wir Coachees von der komprimierten Expertise der Beteiligten profitieren konnten, mit kritischen Fragen konfrontiert wurden, aber auch konstruktives Feedback und konkrete Vorschläge zur Umsetzbarkeit unserer jeweiligen Projekte sowie Ideen für Ausweitungen erhielten und neue Forschungskontakte knüpfen konnten. (Guhr 2019)

> Durch den engen Austausch mit unterschiedlichen FachvertreterInnen, die allesamt interdisziplinäres Arbeiten gewohnt sind und so gute Tipps und Einschätzungen geben konnten, entstand eine äußerst produktive und mehrdimensionale Auseinandersetzung mit der Forschungsfrage meiner Dissertation. Die Arbeit eines Promovenden – also die intensive Textarbeit, Ausarbeitung von Hypothesen und Instrumentarien –, die ja oft nur im stillen Kämmerlein stattfindet, wurde so einmal ganz offen gelegt und konstruktiv erörtert. Es entstand ein mehr als anregender Austausch, bei dem verschiedene Handlungsoptionen und Operationalisierungen durchgespielt wurden und am Ende eine Pipeline für konkrete nächs-

8 https://dhd-blog.org

> te Schritte entstand. Dabei wurden die spezifischen Fragen der Literaturwissenschaft stets im Blick behalten und mit Methodiken der Computerlinguistik flankiert. (Schmitt 2019)

Zum Erfolg trug insbesondere die Tatsache bei, dass an jedem individuellen Coachingtermin Coaches aus der Computerlinguistik und den traditionellen Geisteswissenschaften teilnahmen, die im Rahmen von CRETA bereits zusammengearbeitet hatten. Auf diese Weise war eine zentrale Grundlage für den interdisziplinären Dialog mit den Coachees gegeben, dessen Bedeutung nicht unterschätzt werden darf: ein gemeinsames Vokabular bzw. das Wissen darum, wo – je nach Disziplin – Begriffe eine abweichende Bedeutung haben. Die Beherrschung dieses interdisziplinären ‚Sprachspiels' durch die Coaches machte die Gespräche erst konstruktiv.

5 Fazit

Obwohl Disseminationsarbeit in vielen Fällen zunächst als zusätzlicher Aufwand empfunden wird und daher Gefahr läuft, hinter den ‚wirklich wichtigen' Dingen (z. B. der Arbeit an der eigenen Dissertation) zurückzutreten, haben wir mit den hier dargestellten Aktivitäten positive Erfahrungen gemacht: Bedacht gewählte Disseminationsvorhaben entfalten nicht nur eine Wahrnehmbarkeit nach außen, indem sie anderen ermöglichen, an den eigenen Projektergebnissen teilzuhaben und auf sie aufzubauen. Dies umfasst insbesondere auch methodische, pragmatische und organisatorische Erkenntnisse und Erfahrungen, die in wissenschaftlichen Publikationen nur schlecht oder gar nicht Platz finden.

Auch intern können diese Aktivitäten eine positive Wirkung entfalten. Zunächst dienen sie dem Bündeln und zielgerichteten Ausrichten der internen Arbeit. Dadurch, dass relativ klare Anforderungen innerhalb eines definierten Zeitgerüstes herrschen, fokussieren sich die Beteiligten in anderer Form auf die anstehenden Arbeiten: Die Aufbereitung von intern erarbeiteten Inhalten für Workshops und Lehre zwingt zur Präzision und zur Konzentration auf das Wesentliche, wovon alle Beteiligten profitieren. Die Vorbereitung eines Workshops verpflichtet dazu, sehr genau über den Workshop zu reden.

Weiterhin sind derartige Aktivitäten eine Gelegenheit für das Projektteam, die unterschiedlichen fachlichen Perspektiven auf das Gebiet der Digital Humanities kennenzulernen und trotz ungewohnter Denkweisen die Gemeinsamkeiten zu finden, die das Team verbinden und die Zusammenarbeit stärken. Das begünstigt insgesamt die Aufgeschlossenheit aller Beteiligten gegenüber ungewohnten Ansätzen oder Methoden.

Auch die Aktivitäten an sich und der Austausch mit den Teilnehmenden (von Workshops bzw. Coaching) haben die Arbeiten in CRETA vorangebracht. Dadurch, dass die Teilnehmenden versuchen, die erlernten Methoden auf ihre eigenen Fragestellungen anzuwenden, werden diese Methoden auch mit Situationen konfrontiert, die innerhalb des Zentrums nicht vorkommen (etwa weil niemand mit einer bestimmten Art Daten arbeitet). Selbst wenn das Ergebnis dieser Konfrontation dann mitunter ist, dass bestimmte Methoden auf bestimmte Daten nicht sinnvoll anwendbar sind, ist dies ein methodischer Fortschritt, der ohne die Konfrontation nicht erfolgt wäre. Insofern stellt selbst dieser Fall noch einen Erkenntnisgewinn dar.

Die Erfahrungen in CRETA lehren dementsprechend, dass es ein interdisziplinäres Forschungsvorhaben bedeutend voranbringen kann, frühzeitig über den Tellerrand des eigenen Projektes zu blicken und nach außen gerichtete Aktivitäten durchzuführen. Gerade im vergleichsweise jungen Bereich der Digital Humanities, in dem noch dazu ein bisweilen unübersichtlicher Methodenimport aus verschiedensten Disziplinen stattfindet, sind diese Angebote nicht nur beliebt und hoch frequentiert, sondern wirken sich auch innerhalb des eigenen Projekts positiv aus.

Literatur

Barth, Florian (2020). „Annotation narrativer Ebenen und narrativer Akte“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 423–438.

Gius, Evelyn (2019). „Computationelle Textanalysen als fünfdimensionales Problem: Ein Modell zur Beschreibung von Komplexität“. Pamphlet 8. Digital Humanities Cooperation. URL: https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/12/pamphlet_gius_2.0.pdf (besucht am 1. Juni 2020).

Gius, Evelyn und Janina Jacke (2017). „The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis“. In: *International Journal of Humanities and Arts Computing* 11.2, S. 233–254.

Guhr, Svenja (2019). *Erfahrungsbericht über ein Schlaraffenland für NachwuchswissenschaftlerInnen*. URL: https://dhd-blog.org/?p=12265 (besucht am 1. Juni 2020).

Jacke, Janina (2014). „Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology“. In: *Journal of Literary Theory* 8, S. 118–139.

Jannidis, Fotis, Isabella Reger, Markus Krug, Lukas Weimer, Luisa Macharowsky und Frank Puppe (2016). „Comparison of Methods for the Identification of Main Characters in German Novels“. In: *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland: Jagiellonian University & Pedagogical University, S. 578–582.

Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen“. In: *Reflektierte Algorithmische*

Textanalyse. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.

Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.

Lengwiler, Martin (2005). „Erfolgreich Inter-Disziplinieren". In: *WZB-Mitteilungen* 107, S. 52–55.

Moretti, Franco (2013). „‚Operationalizing': or, the Function of Measurement in Modern Literary Theory". Pamphlet 6. Stanford Literary Lab, S. 1–13. URL: https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf (besucht am 1. Juni 2020).

Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.

Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.

Reiter, Nils, Jonas Kuhn und Marcus Willand (2017). „To GUI or not to GUI?" In: *INFORMATIK 2017*. Bd. 275. Lecture Notes in Informatics (LNI). Chemnitz, Germany: Gesellschaft für Informatik e. V., S. 1179–1184. DOI: 10.18420/in2017_119.

Reiter, Nils und Marcus Willand (2018). „Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse Shakespeares natürliche Figuren im deutschen Drama des 18. Jahrhunderts". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Hrsg. von Toni Bernhart, Marcus Willand, Andrea Albrecht und Sandra Richter. Berlin: De Gruyter, S. 45–76. DOI: 10.1515/9783110523300-003.

Richter, Sandra (2020). „Reading with the workflow: Arbeitsprozesse in den Computational Literary Studies – Beiträge zur Empirisierung literaturwissenschaftlicher Verfahren". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 143–168.

Sack, Graham Alexander (2011). „Simulating Plot: Towards a Generative Model of Narrative Structure". In: *AAAI Symposium*, S. 127–136.

Schmitt, Anna (2019). *Das Creta-Coaching an der Universität Stuttgart*. URL: https://dhd-blog.org/?p=12551 (besucht am 1. Juni 2020).

Suditsch, Isabel (2017). „Interdisziplinärer Fachsprachenkontakt – Fallstudie Arena2036". Diss. Stuttgart University.

Willand, Marcus, Evelyn Gius und Nils Reiter (2020). „SANTA: Idee und Durchführung". In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 391–422.

Nils Reiter und Axel Pichler

CRETA, ein erstes Fazit

Herausforderungen, Erfolge und Empfehlungen

Das Stuttgarter *Center for Reflected Text Analytics* (CRETA) entwickelt textanalytische Methoden für die digitalen Geisteswissenschaften. Wie die Beiträge dieses Bandes zeigen, führt diese Zielsetzung weder zur Entwicklung einer explizierten, elaborierten und logisch konsistenten Interpretationstheorie – wie man sie aus den traditionellen Geisteswissenschaften, insbesondere jedoch aus den Literaturwissenschaften kennt, – noch zur Entwicklung neuer Textanalysetools.[1] Kernergebnis von CRETA ist stattdessen die Etablierung und theoretische Reflexion disziplinenunabhängiger regelgeleiteter Arbeitsablaufpraktiken (= Methoden).[2] Am Ende dieses Buches wollen wir die Herausforderungen, die mit einer solchen Zielsetzung einhergehen, und die Resultate, die unser Umgang mit diesen Herausforderungen gezeitigt haben, resümieren. Damit sollen zukünftigen Projekten, die CRETA in ihrem disziplinären und institutionellen Aufbau ähnlich sind oder vergleichbare Zielsetzungen verfolgen, hilfreiche Strategien auf den Weg mitgegeben werden.

Eine zu erwartende, jedoch trotzdem in ihrer Bedeutung nicht zu unterschätzende Herausforderung stellte für CRETA die interdisziplinäre Zusammenarbeit dar (siehe dazu auch die Einleitung des Bandes sowie den Beitrag zur Interdisziplinarität ab Seite 467). Deren Spezifika – insbesondere die Notwendigkeit der Entwicklung einer gemeinsamen Sprache oder zumindest eines Bewusstseins für mögliche ‚falsche Freunde' – wurden von Anfang an bei der Planung berücksichtigt, was sich unmittelbar auf die konkrete Forschungsarbeit auswirkte. So wurden zum Beispiel die zweiwöchigen Arbeitstreffen im ersten Förderjahr von CRE-

1 Solche wurden zwar ebenfalls im Rahmen von CRETA entwickelt, siehe zum Beispiel das Annotationstool CRETAnno (S. 207), allerdings eher als *proof of concept* denn als Bereitstellung digitaler Infrastruktur.

2 CRETA folgt dabei einem Verständnis von Algorithmus, nach dem es sich bei einem solchen um eine Handlungsvorschrift zur Beantwortung einer Frage bzw. zur Lösung eines Problems handelt. Die Handlungsvorschriften zur Lösung textanalytischer Probleme können sowohl in Code implementiert sein als auch ‚manuell', d. h. von Menschen, realisiert werden (siehe hierzu auch den Beitrag von Pichler und Reiter in diesem Band).

Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Axel Pichler, Stuttgart Research Center for Text Studies, Universität Stuttgart

Open Access. © 2020 Nils Reiter und Axel Pichler, publiziert von De Gruyter CC BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
https://doi.org/10.1515/9783110693973-020

TA zur gemeinsamen Arbeit an für alle Beteiligten relevanten Fragestellungen genutzt, was wesentlich zur Etablierung einer gemeinsamen Sprache beitrug.[3] Daneben wurden weitere Strategien entwickelt und umgesetzt, die den Dialog zwischen den CRETA-Mitarbeiter*innen förderten und vertieften, die wir weiter unten noch ausführen. Eine Konsequenz dieses Vorgehens war, dass es insbesondere am Anfang des Förderzeitraums zu einer ‚Entschleunigung' der individuellen Forschungsarbeit kam: Nicht die zügige individuelle Beantwortung von fachspezifischen Fragen standen im Mittelpunkt der interdisziplinären Aktivitäten, sondern die Erarbeitung eines gemeinsamen Fundaments. Darauf aufbauend konnten dann die fachspezifischen Fragen in Fragen überführt werden, die dem aktuellen Stand der digitalen Geisteswissenschaften entsprechen. Derartige Aktivitäten nehmen viel Zeit in Anspruch. Die temporäre ‚Entschleunigung' wird aber durch ein gemeinsam erarbeitetes interdisziplinäres Wissen ausgeglichen, was letztendlich – so unser Eindruck – zu einem schnelleren Arbeiten nach der Einarbeitungsphase beiträgt. Gleichzeitig wird dabei dem ‚*scheduling conflict*' entgegengewirkt, der in DH-Projekten leicht auftritt: Während der Entwicklung von Textanalyse-Tools herrscht auf geisteswissenschaftlicher Seite viel Leerlauf – und wenn die Tools fertig sind, ist häufig ein Großteil der Projektlaufzeit vorbei (siehe dazu auch die Einleitung dieses Bandes). Durch eine frühzeitige enge Kooperation von Computer- und Geisteswissenschaften verkürzen sich diese Entwicklungszyklen.[4]

Auswirkungen hat das interdisziplinäre Arbeiten auch auf die Publikationspraxis der Beteiligten. Dies betrifft sowohl die Publikationsarten als auch die Publikationsfrequenz. In Hinblick auf den Modus sind zum Beispiel in den Geisteswissenschaften Beiträge mit mehr als einer*m Verfasser*in sehr selten, in den Computerwissenschaften sind sie hingegen üblich. Ebenfalls unüblich ist es in den Geisteswissenschaften, Beiträge zu publizieren, die sich ausschließlich mit der Entwicklung von Methoden beschäftigen,[5] in vielen Computerwissenschaften hingegen wird ausschließlich über Methoden publiziert und auch in den Digital Humanities sind derartige Beiträge zunehmend verbreitet.[6] Hinzu kommt, dass es prinzipiell nicht üblich ist Zwischenresultate von Qualifikationsprojekten zu

3 Resultat dieser Zusammenarbeit sind unter anderem die CRETA-Richtlinien zur Entitätenauszeichnung, die in diesem Band ab Seite 204 besprochen werden.

4 Siehe hierzu auch Seite 81 ff.

5 Jenseits der Wissenschaftssoziologie und -theorie werden Methoden in den Geisteswissenschaften zumeist im Zusammenhang mit den ihnen zugrunde liegenden Theorien verhandelt, nur sehr selten jedoch *per se*.

6 Die Tatsache dass Digital-Humanities-Beiträge sich mehr und mehr methodisch orientieren, liegt sicher auch daran, dass die Methodik bzw. das ‚Digitale' das ist, was die DH verbindet.

publizieren, viele der CRETA-Mitwirkenden jedoch an solchen arbeite(te)n. Bereits diese wenigen Beispiele zeigen, dass auch das Publizieren von ‚Ergebnissen' interdisziplinärer Forschendengruppen eine Herausforderung darstellt, unter anderem auch weil in jedem Fach unterschiedliche Vorstellungen darüber herrschen, was denn ein Forschungsergebnis ist. Erschwerend kommt hinzu, dass es im deutschsprachigen Raum kaum wissenschaftlich etablierte Fachzeitschriften für die digitalen Geisteswissenschaften gibt,[7] eine Publikation von DH-Beiträgen in den fachspezifischen Organen jedoch selten versucht wird. Dies mag Folge einer Erfahrung sein, die auch ein Großteil der CRETA-Mitarbeitenden machen musste: Die Präsentation von Erkenntnissen, die mithilfe digitaler Tools gewonnenen wurden, wird in nicht DH-affinen, traditionelleren fachspezifischen Kontexten – wenn überhaupt – nur sehr kritisch zur Kenntnis genommen. Wie eine diesbezügliche Umfrage im CRETA-Kontext gezeigt hat, reichen die diesbezüglichen Kommentare von „Das ist interessant, aber uns als Vertreter*innen des Faches interessiert eigentlich etwas anderes." über „Das wissen wir doch schon alles." bis zu Fragen wie „Ist das nicht trivial?". Auch der vergleichsweise große Fokus auf die Methoden – die unseres Erachtens in jedem Fall Teil solcher Beiträge sein müssen, wie wir z. B. ab Seite 43 dargelegt haben – führt seitens der Herausgebenden und Gutachtenden häufig zu Irritationen.

Trotz dieser Herausforderungen, vor denen viele DH-Projekte stehen, sehen wir CRETA als Erfolg an. Ein Indikator dafür mag das reichhaltige ‚Ökosystem' sein, das sich um CRETA herum gebildet hat: Der interdisziplinäre Austausch in CRETA hat zu einer Reihe von weiteren Projekten geführt, die inhaltlich und strukturell von CRETA profitier(t)en, und für deren Durchführung teilweise auch Mittel eingeworben werden konnten. Wir stellen hier eine Auswahl vor:[8]

- Schwerpunktprogramm *Computational Literary Studies* (**SPP CLS**). Gefördert durch die Deutsche Forschungsgemeinschaft (DFG). Projektleitung: Fotis Jannidis (Computerphilologie, Universität Würzburg), Evelyn Gius (Digital Philology, TU Darmstadt), Jonas Kuhn (Computerlinguistik), Nils Reiter (Computerlinguistik), Christof Schöch (Digital Humanities, Universität Trier), Simone Winko (Literaturwissenschaft, Universität Göttingen).
  URL: https://dfg-spp-cls.github.io
- Modelling Argumentation Dynamics in Political Discourse (**MARDY**). Gefördert durch die DFG (Schwerpunktprogramm RATIO). Projektleitung: Jonas Kuhn (Computerlinguistik), Sebastian Padó (Computerlinguistik), Sebastian

7 Eine Ausnahme stellt die *Zeitschrift für digitale Geisteswissenschaften* (ZfdG), dar: http://www.zfdg.de.

8 Personen ohne Nennung einer Affiliation sind Angehörige der Universität Stuttgart.

Haunss (Sozialwissenschaften, Universität Bremen).
URL: https://sites.google.com/view/mardy

- Tracing Global Information Networks In Historical Newspaper Repositories (**Oceanic Exchanges**). Gefördert durch die DFG („Digging into Data Challenge"). Projektleitung: Ryan Cordell (Northeastern University), Marc Priewe (Amerikanistik), Sebastian Padó (Computerlinguistik), Steffen Koch (Visualisierung).
  URL: https://oceanicexchanges.org
- Quantitative Drama Analytics (**QuaDramA**). Gefördert durch die Volkswagen-Stiftung. Projektleitung: Nils Reiter (Computerlinguistik), Marcus Willand (Neuere Deutsche Literaturwissenschaft).
  URL: https://quadrama.github.io
- Quantitative Drama Analytics: Tracking Character Knowledge (**Q:TRACK**). Gefördert durch die DFG (Schwerpunktprogramm ‚Computational Literary Studies'). Projektleitung: Nils Reiter (Computerlinguistik), Marcus Willand (Neuere Deutsche Literaturwissenschaft, Universität Heidelberg).
  URL: https://quadrama.github.io
- Umfassende Modellierung von Redebeiträgen in Prosatexten (**QUOTE**). Gefördert durch die DFG. Projektleitung: Sebastian Padó (Computerlinguistik), Roman Klinger (Computerlinguistik).
  URL: https://www.ims.uni-stuttgart.de/forschung/projekte/quote/
- Dhimmis and Muslims – analysing multi-religious spaces in the Medieval Muslim World (**Dhimmis and Muslims**). Gefördert durch die Volkswagen-Stiftung. Projektleitung: Steffen Koch (Visualisierung), Dorothea Weltecke (Mittelalterliche Geschichte, Universität Frankfurt/Main).
  URL: https://www.vis.uni-stuttgart.de/en/projects/dhimmis-and-muslims/
- Structured Multi-Domain Emotion Analysis from Text (**SEAT**). Gefördert durch die DFG. Projektleitung: Roman Klinger (Computerlinguistik).
  URL: https://www.ims.uni-stuttgart.de/forschung/projekte/seat/
- Science Data Center für Literatur (**SDC4Lit**). Gefördert durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg. Projektleitung: Roland S. Kamzelak (Deutsches Literaturarchiv Marbach), Thomas Bönisch (Höchstleistungsrechenzentrum Stuttgart), Gabriel Viehhauser (Digital Humanities), Jonas Kuhn (Computerlinguistik).
  URL: http://www.sdc4lit.org
- Merkmale ästhetischer Reflexionsfiguren: Systematische Annotation und quantitative Analyse. Gefördert durch die DFG (Sonderforschungsbereich 1391, ‚Andere Ästhetik'). Projektleitung: Nils Reiter (Computerlinguistik), Angelika Zirker (Anglistik, Universität Tübingen).
  URL: https://uni-tuebingen.de/de/160783

- Realisierung einer Plattform und begleitender Dienste zum Forschungsdatenmanagement für die Fachcommunity Digital Humanities (**RePlay-DH**). Gefördert durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg. Projektleitung: Jonas Kuhn (Computerlinguistik), Helge Steenweg (Universitätsbibliothek Stuttgart), Stefan Wesner (KIZ Ulm).
  URL: https://www.ims.uni-stuttgart.de/forschung/projekte/replay-dh/
- Automatic Fact Checking for Biomedical Information in Social Media and Scientific Literature (**FIBISS**). Gefördert durch die DFG. Projektleitung: Roman Klinger (Computerlinguistik).
- Text Mining auf geschützten Werken durch Auszüge transparent erschließen (**Xsample**). Gefördert durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg im Rahmen des Programms „Wissenschaftliche Bibliotheken gestalten den digitalen Wandel (BW-BigDIWA)". Projektleitung: Helge Steenweg (Universitätsbibliothek Stuttgart), Jonas Kuhn (Computerlinguistik), Thomas Dreier, (Zentrum für Angewandte Rechtswissenschaft, KIT Karlsruhe).
- Intertextuelle Referenz bei Nietzsche. Projektleitung: Axel Pichler (Philosophie/Literaturwissenschaft), Nils Reiter (Computerlinguistik).

Neben diesem ‚Ökosystem' und den in diesem Buch dargestellten inhaltlichen und methodischen Ergebnissen ist die Ausbildung und Schulung der in CRETA Mitarbeitenden ein wesentliches Ergebnis unserer Bemühungen der letzten Jahre. Um diese Kompetenzen weiterzuentwickeln, ist geplant, CRETA als Methoden-Netzwerk weiterzuführen und auch CRETA-Werkstätten weiterhin durchzuführen.

Abschließend möchten wir unsere in den letzten Jahren gesammelten Erfahrungen in Form von sechs Empfehlungen verdichten, um so die enge und interdisziplinäre Zusammenarbeit bei der Entwicklung textanalytischer Methoden für die digitalen Geisteswissenschaften auch zukünftigen interdisziplinären DH-Projekten zu vermitteln:

Externe Veranstaltungen. Ein frühzeitiges Commitment zu externen Events fördert das Zusammenwachsen in Arbeitsgruppen und dient auch – durch die Arbeit an gemeinsamen Materialien wie Vortragsfolien, Handouts oder Übungsaufgaben – als Katalysator für Begriffsklärungen. Nicht zuletzt gibt eine erfolgreiche Veranstaltung einen wichtigen Referenzpunkt für gemeinsame Arbeit und liefert dafür auch ein Erfolgserlebnis.

Förderung des persönlichen Austauschs. Kaum zu unterschätzen ist der persönliche Austausch, gerade in der Anfangsphase eines Projektes. Unsicherheiten und Nichtwissen zuzugeben, fällt wissenschaftlich Arbeitenden oft schwer. Diese lassen sich daher am leichtesten im informellen Rahmen klä-

ren. Die Schaffung von Gelegenheiten für informellen Austausch (etwa durch Stammtische oder gemeinsame Mensagänge) hat die Arbeit in CRETA erleichtert und für die Beteiligten auch angenehmer gestaltet.

Keine frontalen Projekttreffen. Projekttreffen, gerade in größeren Verbünden, sollten nicht nur als Vortragsreihe gestaltet werden. Raum für spontane Aktivitäten, bilaterale Absprachen oder praktische Arbeit hat zum Beispiel die CRETA-Werkstätten (siehe dazu die Seiten 28 und 468 in diesem Band) zu den produktiven und positiven Terminen gemacht, als die die Beteiligten sie wahrgenommen haben. Die Bandbreite möglicher Aktivitäten ist dabei recht groß und reicht vom gemeinschaftlichen Annotieren über die interaktive gemeinsame Datenanalyse bis hin zu Diskussionen in Kleingruppen. Gerade eine Aufteilung von großen Projektgruppen in kleinere Einheiten erleichtert es auch dem wissenschaftlichen Nachwuchs, sich einzubringen, und sorgt generell dafür, dass mehr Teilnehmende sich beteiligen können.

Freiraum für die Arbeitsebene. Projekte in den Digital Humanities lassen sich oft nicht bis ins letzte Detail planen, da sie letztlich von empirischen Ergebnissen abhängen. Es ist daher einerseits hilfreich, sich nicht zu fest an die Projektpläne zu klammern. Andererseits ist es in DH-Projekten oft so, dass praktische und pragmatische Aspekte einen größeren Stellenwert einnehmen, als sie das sonst in geisteswissenschaftlicher Forschung tun. Projektleitende sollten sich ihre disziplinären Grenzen und Vorannahmen bewusst machen, und – gerade beim interdisziplinären Zusammenarbeiten – ihr Engagement dementsprechend ausrichten.

Gegenseitige Weiterbildung. Ein großer Vorteil von DH-Projekten ist, dass die Beteiligten aus unterschiedlichen Disziplinen kommen, und daher Expertinnen und Experten für unterschiedliche Dinge sind. Eine gegenseitige Weiterbildung, bei der Mitarbeitende aus den Computerwissenschaften z. B. Programmierkurse und Mitarbeitende aus den Literaturwissenschaften einen Einblick in z. B. Interpretationstheorien geben, kann die Projektbeteiligten motivieren und ebenfalls zusammenführen. Im besten Falle führen die erlernten Grundlagen zu einem besseren Verständnis der jeweils ‚anderen Seite' und damit zu produktiverer Zusammenarbeit.

Technische Kenntnisse. Eine häufig gestellte Frage ist die, ob Forschende aus den Geistes- und Sozialwissenschaften, die sich für Digital-Humanities-Methoden interessieren bestimmte Technologien lernen sollten, insbesondere jedoch, ob sie programmieren können sollten. Basierend auf den Erfahrungen in CRETA lässt sich sagen, dass gewisse Programmierkenntnisse für einen mündigen und selbstbestimmten Einsatz von DH-Methoden nötig sind. Aufgrund der komplexen und wandelbaren Natur einzelner Module der reflektierten Textanalyse, sind sowohl eine graphische Benutzeroberfläche

als auch eine kompatible Sammlung von Tools, die den gesamten *workflow* unterstützen, nicht zu erwarten. Dieser (und andere) *workflows* werden auf absehbare Zeit aus einer Vielzahl von Einzel-Werkzeugen zusammengesetzt, die unterschiedliche Ein- und Ausgabeformate erfordern. Eine Programmierkenntnis, die aus und in verschiedene/n Dateiformaten konvertieren kann, scheint uns eine – nicht sehr hohe – Mindestanforderung zu sein.

Trotz der von uns und anderen Projektgruppen erzielten Erfolge in den letzten Jahren bleibt die Integration von geistes- und sozialwissenschaftlichen Fragestellungen, Annahmen und Vorgehensweisen mit computerwissenschaftlichen Methoden, Tools und Formaten eine Herausforderung. Spricht hieraus Resignation? Das glauben wir nicht. „Niemals setzt sich die Wissenschaft das Phantom zum Ziel, endgültige Antworten zu geben oder auch nur wahrscheinlich zu machen; sondern ihr Weg wird bestimmt durch ihre unendliche, aber keineswegs unlösbare Aufgabe, immer wieder neue, vertiefte und verallgemeinerte Fragen aufzufinden und die immer nur vorläufigen Antworten immer von neuem und immer strenger zu prüfen.“ (Popper, *Logik der Forschung*)

Danksagungen

Die Liste derer, die einen Beitrag zum Zustandekommen und zum Funktionieren von CRETA, dem Stuttgarter *Center for Reflected Text Analytics*, geleistet haben, ist lang. Allen gebührt großer Dank. Hinzu kommen weitere Personen, denen wir speziell für die Mitwirkung an diesem Band, der über interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt berichtet, dankbar sind. Einige Individuen und Institutionen sollen hier besonders hervorgehoben werden.

Wir danken zunächst dem Bundesministerium für Bildung und Forschung dafür, dass es mit der Einrichtung einer Förderlinie für Digital-Humanities-Zentren den Anschub für die Einreichung von CRETA gegeben und mit seiner Projektfinanzierung die in diesem Band versammelten Arbeiten ermöglicht hat. Wir danken auch dem Projektträger, dem Deutschen Zentrum für Luft- und Raumfahrt e. V. (DLR), und insbesondere Frau Dr. Maria Böhme für die hervorragende Unterstützung bei der Umsetzung des Projekts.

Ein ganz großes Dankeschön geht an Frau Dr. Sabine Mohr, die zu jedem Zeitpunkt die administrativen Aspekte der CRETA-Planung und -Umsetzung im Blick hatte und den Forscherinnen und Forschern damit über die letzten Jahren den Rücken frei gehalten hat, tief in die inhaltlichen und methodischen Fragen einzudringen. Wir möchten uns auch sehr herzlich bei Fabian Mauch für das umsichtige Lektorat der Beiträge in diesem Buch bedanken.

Es würde CRETA nicht geben, hätte nicht die Leitung der Universität Stuttgart starke Signale der strukturellen Unterstützung für das Zentrum ausgesendet und umgesetzt. Dafür und für die fortdauernde Unterstützung interdisziplinärer Forschung und ganz speziell des Bereichs Digital Humanities sind wir Herrn Rektor Prof. Dr. Ressel und seinem Rekoratsteam, insbesondere Frau Ora Bukoshi, in hohem Maße dankbar. Das Strategieziel des „Stuttgarter Wegs: Vernetzte Disziplinen“ bildet einen hervorragenden Rahmen für eine Initiative wie CRETA.

CRETA war von vornherein angelegt als ein Rahmen, in dem sich zusätzliche Projektinitiativen herausbilden können. Und in der Tat sind in den vergangenen Jahren spezifische Projekte in den Digital Humanities und *Computational Social Sciences* hinzugekommen, die den CRETA-Kern bereichern und verstärken. Wir danken den Förderinstitutionen dieser Projekte sehr, namentlich der Deutschen Forschungsgemeinschaft, der VolkswagenStiftung und dem Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Württemberg.

Wissenschaft ist nichts ohne die Individuen, die die Forschung betreiben. Daher gilt unser herzlicher Dank den Mitarbeiterinnen und Mitarbeitern sowie Antragstellerinnen und Antragstellern in bzw. von CRETA, die nicht nur exzellente inhaltliche Arbeit geleistet haben, sondern mit denen zusammenzuarbeiten es im-

https://doi.org/10.1515/9783110693973-021

mer eine große Freude war. Nicht zuletzt sind wir äußerst dankbar für die produktiven und bereichernden Kooperationen innerhalb von Deutschland und darüber hinaus, die wir im Laufe der Jahre auf- und ausbauen konnten. Ein sichtbares Zeichen dieser Kooperationen ist das Geleitwort zu diesem Band, das mit der historischen Dimension der Digital Humanities diesen Band auch inhaltlich abrundet. Für das Geleitwort bedanken wir uns sehr herzlich bei Herrn Prof. Dr. Jan Christoph Meister.

Die Herausgeber, aus dem Homeoffice, im Juli 2020

Begriffsregister

https://doi.org/10.1515/9783110693973-022

recall Anteil der korrekt erkannten Instanzen an allen in den Daten vorhandenen Instanzen. *52, 72, 73, 89–91, 99, 100, 114, 214, 215, 248, 253, 256–258, 261, 494,* siehe *Evaluationsmetriken*

Reflektierte Textanalyse Reflektierte algorithmische Textanalyse bezeichnet Praktiken der computergestützten Textanalyse, die sich durch ihre interdisziplinär verzahnte Modularisierung kennzeichnen. Bei diesen Modulen handelt es sich um miteinander verknüpfte, manuelle und automatische Arbeitsschritte, die sich auf Begriffe oder Textphänomene beziehen. Die Aufteilung der Module sowie die Interpretation von deren Ergebnissen erfolgt unter Berücksichtigung des gegenstandsbezogenen Vorwissens, der Operationalisierbarkeit der Module sowie deren empirischer Validierung. *24, 45, 48, 57, 231, 232, 470, 478, 490, 494*

Reliabilität ist ein Qualitätskritierium sozialwissenschaftlicher empirischer Forschung. Ein Messinstrument ist reliabel, wenn es a) stabil ist (das Messinstrument bei wiederholter Anwendung die gleichen Werte produziert), b) reproduzierbar ist (das Messinstrument auch in anderen Situationen die gleichen Ergebnisse produziert) und c) wenn es akkurat (genau) misst. Ein reliables Messinstrument, kann trotzdem das Kriterium der Validität verletzen (z. B. ein schlecht geeichtes Thermometer). *71–73, 77, 83, 114, 186, 250, 263, 306, 406, 494*

shared task Bei einem *shared task* arbeiten verschiedene teilnehmende Teams an der gleichen, von einem Organisationsteam gestellten Aufgabe, der automatischen Erkennung eines Textphänomens. Die verschiedenen Lösungen werden dann *auf dem gleichen Datensatz* und *mit der gleichen Evaluationsmetrik* evaluiert und sind damit direkt vergleichbar. *Shared tasks* sind insbesondere innerhalb der Computerlinguistik erfolgreich und mitverantwortlich für die Fortschritte der letzten Jahre. *135, 392–399, 402–406, 410, 413, 416, 417, 472, 475, 494*

Validität ist ein Qualitätskriterium empirischer Sozialforschung, anhand dessen beurteilt wird, ob eine Messung das misst, was der Anspruch der theoretischen Konzeption der Studie verspricht. Eine Studie kann *externe* V. für sich beanspruchen, wenn ihre Ergebnisse über die konkreten analysierten Objekte hinaus generalisierbar sind, wie tw. bei quantitativen Studien mit großen Datenmengen. Eine *interne* V. liegt vor, wenn deren Ergebnisse eindeutig interpretierbar sind, alternative Erklärungen wenig plausibel sind und sich die Befunde in Einklang mit Forschungsergebnissen bringen lassen, die mit anderen Methoden gewonnen wurden. Dies ist oft bei qualitativen Studien mit

kleinen Textmengen der Fall. *48, 50, 81, 170, 171, 175, 176, 182, 184, 186, 205, 207, 406, 494*

Visualisierung In dieser Informatikdisziplin werden Ansätze konzipiert, implementiert und evaluiert, die zur automatischen Generierung von interaktiven, visuellen Repräsentationen abstrakter Daten dienen. Zweck dieser Repräsentationen ist es, die menschliche kognitive Kapazität zu erweitern, um große Datenmengen effektiv zu explorieren (cf. Card et al. 1999; Munzner 2014). *11, 50, 138, 154, 219, 225, 477, 494*

Überwachtes maschinelles Lernen (*supervised machine learning*) Eine Reihe von Methoden, um in annotierten Daten einen Zusammenhang zwischen extrahierbaren Merkmalen und Zielkategorien zu finden. *51, 68, 73, 75, 76, 79, 81, 83, 85, 87, 88, 97, 98, 126, 128, 494*

Literatur

Artstein, Ron (2017). „Inter-annotator Agreement". In: *Handbook of Linguistic Annotation*. Hrsg. von Nancy Ide und James Pustejovsky. Dordrecht, Niederlande: Springer, S. 297–313. DOI: 10.1007/978-94-024-0881-2_11.

Card, Stuart K., Jock D. Mackinlay und Ben Shneiderman (1999). „Information Visualization". In: *Readings in Information Visualization*. Hrsg. von Stuart K. Card, Jock D. Mackinlay und Ben Shneiderman. Morgan Kaufmann, S. 1–34.

Diestel, Reinhard (2010). *Graph Theory*. Electronic Edition. Bd. 173. Graduate Texts in Mathematics. Springer Berlin / Heidelberg.

Fricke, Harald (2007). „Poetik". In: *Reallexikon der deutschen Literaturwissenschaft*. Hrsg. von Harald Fricke. De Gruyter, S. 100–103.

Manning, Christopher D. und Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts und London, England: MIT Press.

Munzner, Tamara (2014). *Visualization Analysis & Design*. CRC Press.

Nünning, Ansgar (2007). „Erzähltheorie". In: *Reallexikon der deutschen Literaturwissenschaft*. Hrsg. von Harald Fricke. De Gruyter, S. 513–517.

Autorinnen und Autoren

Florian Barth
Abteilung Digital Humanities, Institut für Literaturwissenschaft, Universität Stuttgart,
florianbarth@icloud.com

Florian Barth studierte Allgemeine und Vergleichende Literaturwissenschaft sowie Filmwissenschaft in Berlin und absolvierte den Master Digital Humanities in Stuttgart, den er mit einer Arbeit zur Klassifikation von Raumkategorien in literarischen Texten abschloss. Im Zusammenhang seiner Auseinandersetzung mit der digitalen Modellierung narratologischer Konzepte erfolgte auch die Teilnahme am Shared Task SANTA. Während seines Studiums arbeitete er als Hilfskraft u. a. im CRETA-Forschungsprojekt zu Goethes Werther und am Institut für Digital Humanities. Derzeit betreut er die digitale Edition der Krankenjournale Samuel Hahnemanns am Institut für Geschichte der Medizin der Robert Bosch Stiftung.

Martin Baumann
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart,
martin.baumann@vis.uni-stuttgart.de

Martin Baumann studierte Mathematik an der ETH Zürich (Diplom) mit Schwerpunkten in Stochastik und Finanzmathematik, sowie deutsche und amerikanische Sprach- und Literaturwissenschaft an der Universität Tübingen (Magister) mit Studienaufenthalten an der Universität Wien und der University of Washington, Seattle. Seit 2016 forscht er am Institut für Visualisierung und Interaktive Systeme im Rahmen des CRETA Projektes. Seine aktuellen Forschungsinteressen liegen im Bereich der interaktiven Visualisierung von (insbesondere literarischen) Texten und von Annotationsdaten.

André Blessing
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
andre.blessing@ims.uni-stuttgart.de

André Blessing hat an der Universität Stuttgart 2004 sein Informatikstudium als Diplom-Informatiker abgeschlossen. Seitdem ist er in verschiedenen Projekten am Institut für Maschinelle Sprachverarbeitung tätig. Seine Forschungsschwerpunkte liegen dabei in den Bereichen Information Extraction, Argumentation Mining, supervised and semi-supervised learning und Digital Humanities. 2014 hat er zum Thema „Information Extraction for the Geospatial Domain“ promoviert.

Thomas Ertl
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart,
thomas.ertl@vis.uni-stuttgart.de

Thomas Ertl ist Professor für Informatik am Institut für Visualisierung und Interaktive Systeme (VIS) und Direktor des Visualisierungsinstituts der Universität Stuttgart (VISUS). Er studierte Physik und Informatik an der Universität Erlangen, erwarb einen Master of Science in Computer Science von der University of Colorado in Boulder und promovierte in Theoretischer Astrophysik an der Universität Tübingen. Seine Forschungsinteressen liegen in den Bereichen Visualisierung, Computergraphik und Mensch-Computer-Interaktion, insbesondere Volumen-, Strömungs- und Partikel-Visualisierung, hierarchische Datenstrukturen und adaptive Verfahren, parallele und Hardware-beschleunigte Algorithmen für die interaktive Exploration großer Datensätze sowie visuelle Analyse von Texten in Dokumenten und Sozialen Medien.

https://doi.org/10.1515/9783110693973-023

Dominik Gerstorfer
Institut für Philosophie, Universität Stuttgart,
dominik.gerstorfer@philo.uni-stuttgart.de

Dominik Gerstorfer studierte Philosophie, Soziologie und Politikwissenschaften in Berlin und Tübingen. Seit 2015 ist er wissenschaftlicher Mitarbeiter am Lehrstuhl für Wissenschaftstheorie und Technikphilosophie des Instituts für Philosophie an der Universität Stuttgart. Im Rahmen des CRETA-Projekts promoviert er zum Thema „Wissenschaftsphilosophische Fragen der Digital Humanities".

Sein Forschungsinteressen liegen im Bereich der der allgemeinen Wissenschaftstheorie, Theorie der Geisteswissenschaften und der Philosophie der Computerwissenschaften. Ihn interessieren u. a. die methodologischen Grundlagen der Digital Humanities und die gemeinsamen Voraussetzungen der beteiligten Fachdisziplinen.

Evelyn Gius
Institut für Sprach- und Literaturwissenschaft, TU Darmstadt, **gius@linglit.tu-darmstadt.de**

Evelyn Gius ist Professorin für Digitale Philologie und Neuere Deutsche Literatur an der Technischen Universität Darmstadt. Sie hat Germanistik, Philosophie und Informatik in Hamburg und Neapel studiert. In ihrem Promotionsprojekt entwickelte sie einen annotationsbasierten Ansatz zur Erzählstruktur von Konflikterzählungen (Erzählen über Konflikte. Ein Beitrag zur digitalen Narratologie, Berlin 2015). Evelyn Gius arbeitet seit mehr als zehn Jahren im Bereich der Digital Humanities, ihre Forschungsinteressen umfassen literarische Annotationen, die Erzählstruktur literarischer Texte und die Automatisierung literarischer Analysen. Ihre aktuellen Projekte konzentrieren sich auf die Segmentierung literarischer Texte sowie auf die Interaktion zwischen Literaturwissenschaft und Informatik aus methodischer Sicht.

Markus John
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart,
markus.john@vis.uni-stuttgart.de

Markus John studierte an der Technischen Hochschule Ulm und hat seinen B.Sc. in Nachrichtentechnik und M.Sc. in Informationssysteme erhalten. Seit 2013 forscht er als wissenschaftlicher Mitarbeiter am Institut für Visualisierung und Interaktive Systeme (VIS) an der Universität Stuttgart. Seine Forschungsschwerpunkte sind die Informationsvisualisierung, Visuelle Analytik und die Digital Humanities.

Kerstin Jung
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
kerstin.jung@ims.uni-stuttgart.de

Kerstin Jung studierte Informatik an der Universität Stuttgart und promovierte am Institut für Maschinelle Sprachverarbeitung zum Thema der aufgabenbezogenen Kombination von automatisch erstellten Syntaxanalysen. Ihre Forschungsinteressen liegen im Bereich der Nachhaltigkeit von (computer)linguistischen Ressourcen und Abläufen sowie der Verlässlichkeitsbeschreibung von automatisch erzeugten Annotationen. Dabei interessiert sie sich auch für Standards auf der Ebene von Sprachressourcen und arbeitet an der Schnittstelle zwischen Computerlinguistik und anderen sprach- und textverarbeitenden Disziplinen.

Cathleen Kantner
Institut für Sozialwissenschaften, Universität Stuttgart,
cathleen.kantner@sowi.uni-stuttgart.de

Cathleen Kantner studierte Sozialwissenschaften an der Humboldt Universität zu Berlin und der University of Connecticut (UCONN). An der Humboldt Universität promovierte sie über das

Demokratiedefizit der EU und transnationale Öffentlichkeiten. Als Post-Doc arbeitete sie als Marie-Curie-Fellow am Europäischen Hochschulinstitut EUI in Florenz und an der Freien Universität Berlin und leitete Forschungsprojekte, in denen umfangreiche mehrsprachige Medieninhaltsanalysen mit qualitativen und quantitativen Methoden durchgeführt wurden. In diesem Kontext begann sie früh mit Computerlinguistinnen und Informatikerinnen zusammenzuarbeiten.

Nach ihrer Habilitation wurde sie 2010 als Professorin für Internationale Beziehungen und Europäische Integration am Institut für Sozialwissenschaften an die Universität Stuttgart berufen. Sie leitet die CLARIN-D Facharbeitsgemeinschaft für die Sozialwissenschaften. Cathleen Kantners Forschungsinteressen umfassen die Theorien der Internationale Beziehungen, transnationale politische Kommunikation, europäische Öffentlichkeit und Identität, Außen-, Sicherheits- und Verteidigungspolitik, Institutionentheorie sowie korpuslinguistische Methoden der Textanalyse.

Nora Ketschik
Institut für Literaturwissenschaft, Universität Stuttgart, **nora.ketschik@ilw.uni-stuttgart.de**

Nora Ketschik studierte Germanistik, Romanistik und Erziehungswissenschaften an der Universität Stuttgart. Seit 2016 ist sie Promotionsstudentin in der Abteilung der Germanistischen Mediävistik der Universität Stuttgart und Wissenschaftliche Mitarbeiterin im DH-Zentrum CRETA. Im Rahmen ihrer Promotion führt sie Netzwerkanalysen zu mittelhochdeutschen Romanen durch und setzt sich kritisch mit der Operationalisierung literaturwissenschaftlicher Fragestellung für computergestützte Methoden auseinander. 2019/2020 verbrachte sie als DAAD-Stipendiatin einen einjährigen Forschungsaufenthalt am Illinois Institute of Technology (IIT) in Chicago.

Evgeny Kim
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
evgeny.kim@ims.uni-stuttgart.de

Evgeny Kim ist Doktorand an der Universität Stuttgart. Er besitzt einen Magisterabschluss in Linguistik von der Staatlichen Universität Omsk und einen Masterabschluss in Computerlinguistik von der Indiana University. Er forscht vor allem an der Beziehung von Emotionen zu literarischen Gattungen und Figuren.

Roman Klinger
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
roman.klinger@ims.uni-stuttgart.de

Roman Klinger studierte Informatik und Psychologie in Dortmund und leitet an dem Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart eine Arbeitsgruppe zur automatischen Informationsextraktion und Informationsinterpretation auf Basis von Text. Er bearbeitete und leitete unter anderem Projekte zu Therapieverfahren zu Rückenmarksverletzungen (Uni Bielefeld), Emotionsanalyse (DFG-Projekt SEAT, BMBF-Projekt CRETA, Stuttgart) und politischer Entscheidungsfindung (EU-Projekt +Spaces, Fraunhofer SCAI), wobei Methoden des maschinellen Lernens und der künstlichen Intelligenz zum Einsatz kamen. Er verbrachte Forschungsaufenthalte an der University of Massachussetts Amherst (College of Computer Science) und am Institut für Linguistik der Universität Malta.

Steffen Koch
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart,
steffen.koch@vis.uni-stuttgart.de

Steffen Koch studierte Informatik an der Universität Stuttgart. 2012 promovierte er dort am

Institut für Visualisierung und Interaktive Systeme (VIS). Seit 2015 ist er, mit Unterbrechung für eine halbjährige Vertretungsprofessur an der Universität Ulm (2017/2018), am VIS festangestellt. Er leitet und koordiniert Projekte in den Bereichen Visual Analytics und Digital Humanities der Abteilung Grafisch Interaktive Systeme des VIS.

Steffen Kochs Forschungsinteresse liegt auf den Gebieten Informationsvisualisierung und Visual Analytics. Ein Forschungsschwerpunkt ist die Entwicklung von interaktiven, visuellen Verfahren für die Analyse von Dokumentsammlungen und hochdimensionalen Daten für verschiedene Anwendungsgebiete. Letztere umfassen neben dem Bereich Digital Humanities die Analyse von Nachrichten aus Sozialen Medien, Computernetzwerkanalyse, Analysen für den Test von Integrierten Schaltkreise sowie etliche weitere Fragestellungen, die sich den Gebieten Information Retrieval und Data Mining zuordnen lassen.

Benjamin Krautter
Germanistisches Seminar, Universität Heidelberg,
benjamin.krautter@gs.uni-heidelberg.de

Benjamin Krautter studierte Germanistik, Politikwissenschaft und Geschichte in Stuttgart, Konstanz und Seoul. Seit 2017 ist er Mitarbeiter und Promotionsstudent im Projekt *Quantitative Drama Analytics* (QuaDramA). Dort arbeitet er u. a. an der Operationalisierung literaturwissenschaftlicher Kategorien für die quantitative Dramenanalyse. Im Zentrum seines Forschungsinteresses steht dabei die mögliche Verbindung quantitativer und qualitativer Methoden für die Analyse und Interpretation literarischer Texte (*scalable reading*).

Gerhard Kremer
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
gerhard.kremer@ims.uni-stuttgart.de

Gerhard Kremer wurde nach seinem Studium der Computerlinguistik an der Universität Stuttgart mit seiner Dissertationsarbeit im interdisziplinären Forschungsgebiet zwischen Computerlinguistik und Psycholinguistik zum Thema distributionelle Semantik an der Università di Trento/Italien promoviert. Er führte Forschungen in diesem Bereich an der Universität Heidelberg weiter bevor er an der Universität Stuttgart anfing, mit seiner interdisziplinären Ausrichtung beim Projekt CRETA in den Digital Humanities mitzuwirken.

Der Interessenschwerpunkt Gerhard Kremers ist der reflektierte Einsatz von Werkzeugen der Computerlinguistik für geistes- und sozialwissenschaftliche Fragestellungen. Daneben engagiert er sich bei der Vermittlung von Programmierwissen für Anfänger in den Digital Humanities und der nachhaltigen Archivierung von Ressourcen.

Jonas Kuhn
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
jonas.kuhn@ims.uni-stuttgart.de

Jonas Kuhn studierte in Stuttgart und Edinburgh Computerlinguistik und wurde 2001 an der Universität Stuttgart in diesem Fach promoviert. Ein Postdoc-Aufenthalt an der *Stanford University* schloss sich an, 2003–05 war Kuhn als Assistant Professor an der *University of Texas at Austin*, baute dann in Saarbrücken eine Emmy Noether-Gruppe auf, bevor er 2006 an die Universität Potsdam berufen wurde. 2010 folgte Kuhn einem Ruf zurück nach Stuttgart, wo er seither die Professur für Grundlagen der Computerlinguistik am Institut für Maschinelle Sprachverarbeitung innehat.

Die Forschung Kuhns und seiner Arbeitsgruppe beschäftigt sich im Kern mit algorithmi-

schen Verfahren für eine linguistisch fundierte strukturelle Analyse von Sprache und Text. Seit Beginn der 2010er-Jahre liegt ein erweiterter Schwerpunkt auf der Entwicklung von disziplinübergreifenden Vorgehensmodellen für die Textanalyse in den Geistes- und Sozialwissenschaften, der nicht zuletzt in die Gründung des Digital Humanities-Zentrums CRETA mündete.

Sandra Murr
Stuttgart Research Center for Text Studies, Universität Stuttgart,
sandra.murr@ts.uni-stuttgart.de

Sandra Murr studierte Germanistik und Kunstgeschichte an der Universität Stuttgart. Seit 2016 ist sie im interdisziplinären Forschungsprojekt CRETA als wissenschaftliche Mitarbeiterin beschäftigt und wirkt hierin an der digitalen Analyse der literarischen Rezeption von Goethes *Werther* mit. Ihr Schwerpunkt liegt auf der Operationalisierung von literaturwissenschaftlichen Konzepten, um die Wertheriaden, die in Folge der Veröffentlichung des Briefromans erschienen sind, computergestützt analysierbar und vergleichbar zu machen. Weitere Forschungsinteressen bilden die Auseinandersetzung mit Autoren der Wiener Moderne, Theorien und Methoden der Narratologie und Themen der Computational Literary Studies.

Maximilian Overbeck
Department of Communication and Journalism, The Hebrew University of Jerusalem, **m.overbeck@mail.huji.ac.il**

Maximilian Overbeck studierte Politikwissenschaft am Institut d'Études Politiques in Bordeaux und Sozialwissenschaften an der Universität Stuttgart. In seiner Dissertation, die er 2019 an der Universität Stuttgart verteidigt hat, untersuchte er die steigende politische Bedeutung religiöser Identitäten in westlichen politischen Debatten über bewaffnete Konflikte nach dem Ende des Kalten Krieges. Aktuell arbeitet er als Postdoc an der Hebrew University in Jerusalem in einem ERC Projekt, das sich mit der Bedeutung politischer Prognosen beschäftigt. In seiner Forschung untersucht er theoretisch anspruchsvolle und politikwissenschaftlich relevante Konzepte anhand qualitativer und quantitativer Forschungstechniken und Methoden. Ein besonderer Fokus seiner Forschung liegt dabei auf der theoretischen und empirischen Auseinandersetzung mit religiösen Identitäten und religiösen Überzeugungen im politischen Raum.

Sebastian Padó
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
pado@ims.uni-stuttgart.de

Sebastian Padó studierte Computerlinguistik, Informatik und Kognitionswissenschaft und wurde 2007 an der Universität des Saarlandes mit einer computerlinguistischen Arbeit zu mehrsprachiger Bedeutungsanalyse promoviert. Er war 2007–2009 Postdoc an der Stanford University, 2010–2013 Professor an der Universität Heidelberg, und ist seit 2013 Professor für theoretische Computerlinguistik an der Universität Stuttgart.

Seine Forschungsinteressen liegen vor allem im Bereich der Repräsentation und Verarbeitung lexikalisch-semantischer Phänomene, mit Schwerpunkten in den Bereichen der linguistischen Analyse distributionaler Modelle und der empirischen Modellierung theoretisch-linguistischer Konzepte wie Kompositionalität und Polysemie. Außerdem forscht er an multilingualen Aspekten der Computerlinguistik und der Nutzung paralleler und vergleichbarer Korpora als Wissensquellen.

Janis Pagel
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
janis.pagel@ims.uni-stuttgart.de

Janis Pagel studierte an der Ruhr-Universität Bochum Linguistik und Germanistik im Bachelor of Arts (2012–2015), sowie Computerlinguistik im Master of Science an der Universität Stuttgart und der Universität von Amsterdam (2015–2018). Seit April 2018 ist er als Mitarbeiter und Doktorand in CRETA, sowie im Forschungsprojekt QuaDramA (Quantitative Drama Analytics) tätig und arbeitet dort zu Koreferenzauflösung und Strukturerkennung dialogischer Texte und der automatischen Erkennung von Figurentypen in deutschsprachigen Dramen.

Axel Pichler
Stuttgart Research Center for Text Studies, Universität Stuttgart,
axel.pichler@ts.uni-stuttgart.de

Axel Pichler studierte Germanistik und Philosophie in Graz und Wien und promovierte 2009 an der Karl-Franzens-Universität Graz mit einer Arbeit über Friedrich Nietzsche (*Nietzsche, die Orchestikologie und das dissipative Denken*, Wien: Passagen 2010). 2013 forschte er als Fritz-Thyssen-Stipendiat an der FU Berlin zur Bedeutung der Textualität für die Philosophie (*Philosophie als Text – zur Darstellungsform der „Götzen-Dämmerung“*, Berlin/Boston: De Gruyter 2014) Seit 2014 ist er Post-Doc am Stuttgart Research Centre for Text Studies, wo er unter anderem im Rahmen von CRETA an einer Hybridedition und der digitalen Interpretation ausgewählter Überlieferungsträger von Theodor W. Adornos *Ästhetischer Theorie* arbeitet und die Möglichkeiten und Grenzen reflektierter Textanalyse auslotet.

Nils Reiter
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
nils.reiter@ims.uni-stuttgart.de

Nils Reiter hat Computerlinguistik und Informatik an der Universität des Saarlandes studiert. Promoviert wurde er an der Universität Heidelberg, wo er in einem Kooperationsprojekt zwischen Computerlinguistik und klassischer Indologie gearbeitet hatte. Seit 2014 ist er Postdoc am Institut für Maschinelle Sprachverarbeitung, und hat, neben Lehrtätigkeit im Studiengang Digital Humanities, die wissenschaftlichen Arbeiten in CRETA koordiniert. Seit 2017 ist er – gemeinsam mit M. Willand – auch Projektleiter im Kooperationsprojekt QuaDramA, in dem Dramen quantitativ untersucht werden. Im Studienjahr 2019/20 vertritt er die Professur für Sprachliche Informationsverarbeitung/Digital Humanities an der Universität zu Köln. Seine Forschungsinteressen bestehen in der adäquaten Anwendung computerlinguistischer Verfahren in den Digital Humanities, sowie theoretischen und praktischen Aspekten der Operationalisierung komplexer Phänomene.

Sandra Richter
Abteilung Neuere deutsche Literatur I, Institut für Literaturwissenschaft, Universität Stuttgart, **Sandra.Richter@dla-marbach.de**

Sandra Richter, Professorin für Neuere deutsche Literatur an der Universität Stuttgart und Direktorin des Deutschen Literaturarchivs Marbach, Arbeitsschwerpunkte: Poetik und Ästhetik, Methodologie, Empirische Literaturwissenschaft, Wissenschaftsgeschichte und Wissensformen in der Literatur; Publikationen u. a.: Eine Weltgeschichte der deutschen Literatur. München: C. Bertelsmann 2017 ([2]2018, [3]2019), [mit Toni Bernhard, Marcus Willand, Marcel Lepper, Andrea Albrecht] Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspekti-

ven. Berlin, Boston: de Gruyter 2019, „Mere exposure effects“: eine Komponente ästhetischer Erfahrung beim Lesen von Literatur, in: Die Wiederkehr des Erlebnisses in der Geisteswissenschaft, hg. v. Mathis Lessau und Nora Zügel. Baden-Baden: Ergon 2019, S. 43–58.

Ina Rösiger
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
ina.roesiger@ims.uni-stuttgart.de

Ina Rösiger hat einen Hintergrund in Computerlinguistik und Informatik und verteidigte im Januar 2019 ihre Dissertation über Koreferenz- und Bridgingauflösung im Bereich Automatisches Textverstehen an der Universität Stuttgart. Seitdem arbeitet sie in der IT Vorausentwicklung der Firma Bosch am prototypischen Einsatz von NLP-Komponenten.

Mirco Schönfeld
Bayrische Hochschule für Politik an der Technischen Universität München,
mirco.schoenfeld@tum.de

Mirco Schönfeld ist seit 2017 als post-doctoral researcher an der Professur für Computational Social Sciences an der Hochschule für Politik der Technischen Universität München tätig. Zuvor arbeitete er als wissenschaftlicher Mitarbeiter an der LMU München. Dort schloss er 2016 seine Promotion zum Thema „Kontextbezug und Authentizität in Sozialen Netzen“ am Institut für Informatik ab. Seine Forschungen konzentrieren sich unter anderem auf Algorithmen zur Netzwerkanalyse, zu Data Mining und zur Integration von Kontextinformationen. Anwendung finden seine Arbeiten vor allem in interdisziplinären Kooperationen mit Forscherinnen und Forschern der Geistes- und Sozialwissenschaften.

Sarah Schulz
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
sarah.schulz@ims.uni-stuttgart.de

Sarah Schulz ist Computerlinguistin mit einem Hintergrund in den Geisteswissenschaften. Ihr Bachelorstudium absolvierte sie in Theater- und Medienwissenschaften sowie Germanistik. Anschließend studierte sie Computerlinguistik in Tübingen. Ihre Dissertation über NLP-Methoden in den Digital Humanities verteidigte sie im Dez. 2017 an der Universität Stuttgart. Ihr Wissen über die Verarbeitung von Nicht-Standardtexten wendet sie seither im Bereich der biomedizinischen Textverarbeitung an.

Gabriel Viehhauser
Abteilung Digital Humanities, Institut für Literaturwissenschaft, Universität Stuttgart,
viehhauser@ilw.uni-stuttgart.de

Gabriel Viehhauser ist Professor für Digital Humanities am Institut für Literaturwissenschaft der Universität Stuttgart. Er hat deutsche Philologie und Philosophie, Psychologie und Pädagogik an der Universität Wien studiert. Im Anschluss war er wissenschaftlicher Mitarbeiter beim vom Schweizerischen Nationalfond (SNF) geförderten Projekt „Wolfram von Eschenbach Parzival. Eine überlieferungskritische Ausgabe in elektronischer Form“ an den Universitäten Basel und Bern sowie Assistent am Lehrstuhl für Germanistische Mediävistik der Universität Bern (Prof. Stolz). Nach einer Dissertation zur Parzival-Überlieferung und einem dreijährigen SNF-Stipendium für fortgeschrittene Forschende an den Universitäten Göttingen und München hat er die neueingerichtete Professur für Digital Humanities in Stuttgart übernommen. Die Forschungsinteressen von Gabriel Viehhauser liegen im Bereich der digitalen Editiorik und der digitalen Textanalyse, in deren Rahmen er sich insbesondere mit digitalen Zugänge zu einer Narratologie des Raums befasst hat.

Marcus Willand
Germanistisches Seminar, Universität Heidelberg,
marcus.willand@gs.uni-heidelberg.de

Studium (2002–08) der Sprach- und Literaturwissenschaften, Psychologie und Soziologie in Darmstadt, Berlin und Turku (Finnland). Promotion (HU-Berlin, 2009–2014) zu ‚Lesermodelle[n] und Lesertheorien', ehem. Mitglied des PhD-Net: ‚Das Wissen der Literatur' und Stipendiat der Promotionsförderung der Studienstiftung des deutschen Volkes. Aufenthalt als *visiting research scholar* an der Princeton University (NJ), USA, im WS 2009, gefördert durch ein Auslandsstipendium des DAAD. Seit 2013 wissenschaftlicher Mitarbeiter von A. Albrecht erst in Stuttgart, dann Heidelberg, 2014–2018 Redakteur der Scientia Poetica. Projektleiter (mit Nils Reiter) von ‚QuaDramA' (VolkswagenStiftung) und ‚Q:TRACK' (Schwerpunktprogramm ‚Computational Literary Studies', DFG).

Yvonne Zimmermann
Institut für Literaturwissenschaft, Universität Stuttgart,
yvonne.zimmermann@ilw.uni-stuttgart.de

Yvonne Zimmermann studierte Germanistik, Geschichte und Romanistik in Freiburg, Grenoble und Stuttgart und promovierte 2015 an der Universität Stuttgart und am King's College London mit einer Arbeit über Rudolf Alexander Schröder (*Geschichte, Politik und Poetik im Werk Rudolf Alexander Schröders*, Frankfurt a. M. 2016.). Seit 2010 ist sie Wissenschaftliche Mitarbeiterin in der Abteilung Neuere deutsche Literatur I der Universität Stuttgart. Ihre Forschungsinteressen liegen u. a. in intertextuellen, poetologischen und literatursoziologischen Fragen.