Jonas Kuhn

Computerlinguistische Textanalyse in der Literaturwissenschaft? Oder: »The Importance of Being Earnest« bei quantitativen Untersuchungen

Abstract: In its first part, this article gives some illustrative insights into the spectrum of methods and model types from Computational Linguistics that one could in principle apply in the analysis of literary texts. The idea is to indicate the considerable potential that lies in a targeted refinement and extension of the analysis procedures, as they have been typically developed for newspaper texts and other everyday texts. The second part is a personal assessment of some key challenges for the integration of working practices from Computational Linguistics and Literary Studies, which ultimately leads to a plea for an approach that derives the validity of model-based empirical text analysis from the annotation of reference corpus data. This approach should make it possible, in perspective, to refine modeling techniques from Computational Linguistics in such a way that even complex hypotheses from Literary Theory can be addressed with differential, data-based experiments, which one should ideally be able to integrate into a hermeneutic *argumentation*.

Einleitung

Die Computerlinguistik und die Sprachtechnologieforschung entwickeln ihre Modelle und Methoden überwiegend für Gebrauchstexte wie Zeitungsartikel, Produktbesprechungen auf Internetseiten, Forenbeiträge in den Sozialen Medien etc. Dennoch üben literarische Texte mit ihren vielfältigen Herausforderungen an die Textanalyse eine große Anziehungskraft auf Computerlinguistinnen und -linguisten aus und in den wichtigsten Publikationsorganen, den Tagungsbänden der großen Computerlinguistikkonferenzen, erscheinen seit vielen Jahren vereinzelt, aber immer wieder Beiträge zur Erweiterung von computerlinguistischen Analysemodellen, die auf Charakteristika literarischer Texte abzielen.[1]

1 Vgl. u. a. David K. Elson, Nicholas Dames und Kathleen R. McKeown: »Extracting social networks from literary fiction«, in: *Proceedings of the 48th Annual Meeting of the Association for*

© 2018 Jonas Kuhn, publiziert von De Gruyter.
Dieses Werk ist lizenziert unter der CreativeCommons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783110523300-002

Die wachsende Aufmerksamkeit für die *Digital Humanities* – nicht zuletzt dank der Förderinitiativen der letzten Jahre im deutschsprachigen Raum – hat das Interesse in der Computerlinguistik-Community für interdisziplinäre Zusammenarbeit mit der Literaturwissenschaft weiter verstärkt. Wer sich in einer technischen und vorwiegend methodenorientierten Disziplin auf einen Analysegegenstand aus einem anderen Fachkontext einlässt, tut dies in dem Bewusstsein bzw. in der sicheren Erwartung, dass die etablierten Analysemodelle stark angepasst und erweitert werden müssen (beispielsweise um der Vielschichtigkeit eines Erzähltextes gerecht zu werden) und dass in der interdisziplinären Kooperation die methodischen Grundannahmen aus den unterschiedlichen Fächerkulturen sorgfältig herausgearbeitet und die gemeinsame Agenda entsprechend differenziert aufgesetzt werden muss. Der vorliegende Beitrag skizziert einerseits, wie die zu erwartenden Anpassungen des methodischen Vorgehens aus Sicht der Computerlinguistik aussehen, und wirft andererseits die Frage auf, ob und wie diese tatsächlich einen fruchtbaren Beitrag zu literaturwissenschaftlichen Kernfragen leisten können – oder ob die Grundannahmen zur textanalytischen Praxis so stark divergieren, dass noch grundlegendere Anpassungen erforderlich wären.

Die Computerlinguistik kann auf lange, fruchtbare Kooperationserfahrungen mit der theoretischen Linguistik zurückblicken, aus der u. a. Praktiken des quantitativ-korpuslinguistischen Arbeitens mit Werkzeugunterstützung (wie *Part-of-Speech-Tagging*, also automatische Auszeichnung von Wortarten) hervorgegangen sind. Hierfür waren und sind durchaus unterschiedliche Erkenntnisinteressen und Arbeitshypothesen abzustimmen – methodisch hat sich die Computerlinguistik in den letzten 20 bis 30 Jahren sehr weit von der Linguistik entfernt, es

---

Computational Linguistics, ACL '10. Stroudsburg, PA, USA, 2010 (Association for Computational Linguistics), S. 138–147; David Bamman, Ted Underwood und Noah A. Smith: »A Bayesian Mixed Effects Model of Literary Character«, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore 2014, S. 370–379; Justine Kao und Daniel Jurafsky: »A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry«, in: *Proceedings of the Workshop on Computational Linguistics for Literature* (*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*), Montréal 2012, S. 8–17; Hardik Vala, David Jurgens, Andrew Piper und Derek Ruths: »Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts«, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, hg. v. Association for Computational Linguistics. Lisabon September 2015; Julian Brooke, Adam Hammond und Graeme Hirst: »Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction«, in: *Digital Scholarship in the Humanities* (2016).

dominieren statistische Modelle der Sprachverarbeitung. Und so hat sich ein Bewusstsein für einen methodischen Anpassungsbedarf in Abhängigkeit von linguistischer Beschreibungsebene – Phonologie, Morphologie, Syntax, Semantik, Aspekte der Pragmatik – und theoretischem Ansatz herausgebildet. Aus computerlinguistischer Sicht erscheint es naheliegend, die Kooperation mit Linguistinnen und Linguisten als paradigmatisch für einen Dialog zwischen der geisteswissenschaftlichen Auseinandersetzung mit Sprache und Text und der komputationellen Modellierung von Textanalyseprozessen generell zu betrachten. Der Übergang zu literarischen Texten lässt aus dieser Sicht sicherlich besondere Herausforderungen an die Analysetiefe und die Abstimmung des deskriptiven Begriffsinventars erwarten, also einen intensiveren Anpassungsprozess, aber keinen grundsätzlich anders gearteten. In konkreten Überlegungen zu möglichen Kooperationen zwischen Literaturwissenschaft und Computerlinguistik erweist es sich jedoch nicht selten, dass die Herausforderungen weniger in einer schrittweisen Erweiterung der vorhandenen Analysemodelle liegen, sondern vielmehr das hermeneutisch geprägte Grundverständnis auf der einen und das stark experimentell-datenorientierte Vorgehen auf der anderen Seite selbst kooperationsfreudige Partner zunächst vor grundsätzlichere Fragen stellen. Diese Situation und ein möglicher Ansatz für die Praxis sollen in diesem Aufsatz aus dem Blickwinkel eines Computerlinguisten mit Interesse an einer fundierten Erweiterung des textanalytischen Methodeninventars diskutiert werden.

Teil 1 skizziert exemplarisch textanalytische Problemstellungen jenseits der etablierten linguistischen Analyseebenen, für die der Computerlinguistik ein Inventar an Modellierungsverfahren zur Verfügung steht, welches sich grundsätzlich um weitere Analyseebenen erweitern lässt. Das übliche Vorgehen besteht in einem Aufbrechen einer komplexeren Analyseaufgabe in Teilschritte, für die sich die jeweils beabsichtigte Kategorisierung von empirischen Texteigenschaften operationalisieren lassen, also auf Basis einer intersubjektiven Übereinstimmung festgelegt werden können. Konkret wird anhand eines Beispiels aus Mark Twains *Adventures of Tom Sawyer* illustriert, welche oberflächenorientierten Analyseschritte erforderlich sind, um in Erzähltexten wörtliche Rede den Figuren zuzuordnen.

Viele operationalisierte Analysemodelle lassen sich (i) für qualitative Fragestellungen bei der Textanalyse einsetzen (und sicherlich auch für den Abgleich von literaturtheoretischen Hypothesen gegen die Empirie, also einen einzelnen Text oder eine kleine Auswahl von Werken); mit der Möglichkeit einer Automatisierung bestimmter Teilanalysen erschließen sich jedoch – mit der nötigen methodenkritischen Reflexionsbereitschaft – vor allem auch Wege, (ii) ein größeres Korpus von Zieltexten hinsichtlich ausgewählter Eigenschaften systematisch zu

untersuchen, beispielsweise explorativ im Sinne des *Distant Reading* oder für Vergleichsstudien. Im Rahmen des vorliegenden Bandes liegt der Fokus auf (ii), also automatisierten Analyseschritten in der Aufbereitung von größeren Korpora für mögliche quantitative Fragestellungen. Eine computergestützte Identifikation und Zuordnung von Figurenrede in Mark Twains *Huckleberry Finn* soll beispielhaft verdeutlichen, wie der Einsatz von computerlinguistischen Analysemodellen es ermöglicht, ein größeres Textkorpus in einer feineren Granularität zu erschließen – hier für stilistische Untersuchungen zur Figurenrede – als dies mit gängigen quantitativen Verfahren möglich ist.

Teil 2 soll etwas ausführlicher auf die eingangs angedeutete Problematik eingehen, die im weitesten Sinn wissenschaftstheoretisch bzw. -soziologisch ist: Trotz der großen Dynamik innerhalb der Fachcommunity der *Digital Humanities*, in der aus naheliegenden Gründen ein Ausloten von korpusorientierten Modellierungsmöglichkeiten mit computerlinguistischen Verfahren methodologisch relevant ist, erscheinen Vertreter aus den »Kernbereichen« der Literaturwissenschaften (sofern eine derartige Generalisierung überhaupt zulässig ist) vielfach reserviert, wenn es um die Frage geht, ob sie einer Argumentation folgen würden, die sich teils auf computerlinguistische Analysen stützt. Teil 2 spekuliert über Gründe für diese Reserviertheit (im Anschluss an einen Beitrag zur Methodendiskussion des interdisziplinären Autorenteams Hammond/Brooke/Hirst 2013) und schließt Überlegungen an, ob und, wenn ja, wie sie auf breiterer Basis zu überwinden wäre.

Diejenigen, die sich gegenüber computergestützten Verfahren in der Literaturwissenschaft offen zeigen (und sie werden immer mehr und sind in der deutschsprachigen *Digital Humanities*-Community recht gut vernetzt), sehen sich einer – oft unübersichtlichen – Fülle von technischen Möglichkeiten gegenüber; mangels etablierter Arbeitspraktiken zur Integration von klassisch hermeneutischen Arbeitsschritten und formalisierten Analysemodellen ist zunächst unklar, wie sich geeignete Kombinationen methodenkritisch etablieren lassen und wie vermieden werden kann, dass Werkzeuge entgegen ihren Anwendungsbedingungen eingesetzt und so eine irreführende Pseudo-Objektivität erzeugt wird. Zu diesem Punkt argumentiert dieser Beitrag abschließend für sehr hohe Standards bei der Legitimation eines werkzeuggestützten Analyseschritts, wobei sich diese Standards durch eine Probe aufs Exempel etablieren lassen: Dabei wird die Analyse des Untersuchungsgegenstandes durch die Analyse eines unabhängig annotierten »Referenzkorpus« gegengeprüft – unter Beachtung der Regeln der Korpusannotationspraxis, die auch (und gerade) in den Zuständigkeitsbereich hermeneutischer Praxis fallen sollten. Das Ausfindigmachen und die sorgfältige

Aufbereitung und Annotation geeigneter Referenzdaten, die in relevanten Eigenschaften als hinreichend repräsentativ für die analytischen Fragestellungen betrachtet werden, ist zwar dem klassisch-hermeneutischen Vorgehen fremd und macht ein Umdenken notwendig. Da sich das Vorgehen jedoch sehr flexibel in die Textanalysepraxis einbinden lässt, die Optimierung computerlinguistischer Modelle rechtzeitig im Projektverlauf ermöglicht und eine kritische disziplinübergreifende Auseinandersetzung mit der Spezifikation der Analysekategorien unterstützt, mag es die Basis für eine Synthese aus den Arbeitspraktiken darstellen.

1 Textanalytisches Potenzial und Herausforderungen

Im Kern geht es der Computerlinguistik darum, Modelle und Algorithmen für die syntaktische und semantische bzw. pragmatische Analyse (oder Generierung) von sprachlichen Äußerungen zu entwickeln – also die *strukturellen* Eigenschaften von sprachlichen Äußerungen und Texten systematisch zu erfassen und die Texte, ausgehend von ihren strukturellen (und lexikalischen) Eigenschaften, in Beziehung zu setzen zu einer oder zu mehreren *inhaltlichen* Ebenen. Inhaltlich müssen (a) die wörtliche Bedeutung und (b) die pragmatisch zu erklärenden Inhalte bestimmt werden, die gleichsam mitverstanden werden und für die der situative Kommunikationskontext und der (ggf. sehr weit zu fassende) Diskurskontext zu berücksichtigen sind. In voller Allgemeinheit ist eine formal exakte und umfassende Modellierung des menschlichen Vermögens, sprachliche Äußerungen und Texte zu produzieren und im Kontext zu verstehen, offensichtlich jenseits der realistischen Möglichkeiten – müsste sie doch u. a. unsere Fähigkeit einschließen, beliebige Inferenzen aus konkurrierenden Interpretationsalternativen zu ziehen, um sie gegen den Kontext abzugleichen. Das hierfür notwendige Modell wäre dann auch in der Lage, im Prinzip jedes intelligente menschliche Verhalten nachzumodellieren – was die meisten Beteiligten für grundsätzlich unmöglich erachten.[2] Mit einem breiten Inventar von unterschiedlichen formalen

2 In der klassischen Debatte um die Grenzen der künstlichen Intelligenz wird dieses Argument gern als KI-Vollständigkeit bezeichnet. Die umfassende Lösung des Problems des Sprachverstehens wäre zugleich eine Lösung für jedes andere Problem, das sämtliche Facetten menschlicher Intelligenz erfordert.

und algorithmischen Ansätzen, die jeweils einen definierten Ausschnitt der Gesamtproblematik anhand von konkreten Sprach- und Textdaten in validierbarer Form erfasst, ist es heute jedoch möglich, belastbare Analyseergebnisse für eine Vielfalt von klar definierten Teilaufgaben zu erhalten. Beispielsweise können aus Nachrichtentexten Meldungen zu bestimmten Ereignistypen mit großer Verlässlichkeit extrahiert werden (*X hat Y für eine Funktion F bestellt* oder *in der Region A ist zum Zeitpunkt T ein Naturereignis N eingetreten*); mit der sogenannten Technik der Sentimentanalyse kann für wertende Texte einer bekannten Gattung oder Untergattung (wie z. B. Produkt- oder Filmrezensionen) die Polarität der subjektiven Wertung recht zuverlässig automatisch bestimmt werden; maschinelle Übersetzung für Textsorten, für die eine große Sammlung von »Trainingsdaten« vorliegt, ist auf einem Qualitätsniveau möglich, das vor zehn Jahren noch als völlig utopisch gegolten hätte.

Entsprechend liegen Analysemodelle vor, die auf Eigenschaften von literarischen Texten abheben oder so erweitert werden könnten, dass sie zu literaturwissenschaftlichen Fragestellungen relevante Teilanalysen in abschätzbarer Qualität auf einem größeren, verhältnismäßig homogenen Textkorpus automatisch liefern können. So lassen sich beispielsweise *Distant Reading*-Phasen in einem korpusorientierten Vorgehen unterstützen. Im Hintergrund kann dabei durchaus eine literaturtheoretische Konzeption stehen, die zusätzlich zu den linguistischen Ebenen der grammatischen Struktur, der Diskursstruktur, des wörtlich-semantischen Textinhalts und der pragmatischen, kontextbezogenen Bedeutung weitere interpretations- oder deutungsrelevante Ebenen ansetzt – etwa die Textrezeption in einer bestimmten Epoche vor dem Hintergrund eines etablierten Kanons.

In Teil 2 werden wir auf Umstände zu sprechen kommen, die es zunächst möglicherweise erschweren oder gar verhindern, dass die bestehenden Möglichkeiten zu einer Fülle von Projekten für entsprechende Erweiterungen des computerlinguistischen Analyseinventars führen. Vorher soll hier zunächst ausführlicher dargestellt werden, wie man sich solche Erweiterungen konkreter vorstellen kann. Dabei werden unterschiedliche Typen von Analysekomponenten vorgestellt, mit denen die Computerlinguistik arbeitet (ohne das Spektrum systematisch abdecken zu wollen). Ein ausführlicheres Beispiel, in dem unterschiedliche Analysekomponenten auf Texte von Mark Twain angewandt werden, wird den Teil 1 abschließen.

Zwei grundlegend verschiedene Ansatzpunkte für formalisierte Modelle der Textanalyse liegen in einer linguistisch-strukturellen vs. einer distributionellen Basis. **Der linguistisch-strukturelle Analyseansatz** geht von der sprachlichen Struktur des Textes aus und operationalisiert Kategorien von analyserelevanten

Texteinheiten (z. B. Personennamen[3] oder Zeitausdrücken), deren Verteilung im Text die Modelle dann vorhersagen. Häufig sind mehrere strukturelle Kategorien hierarchisch ineinander geschachtelt, d. h. größere Analyseeinheiten werden bei der Vorhersage auf darin enthaltene kleinere Einheiten überprüft. Eine mittlerweile etablierte Analysemethode[4] überprüft beispielsweise Textabschnitte (wie Kapitel) auf die darin verwendeten Figurennamen, bildet daraus eine Relation zwischen Figuren (*X und Y tauchen im gleichen Kapitel auf*) und kann so für ganze Korpora Figurennetzwerkkonstellationen bzw. die jeweilige Entwicklung von Relationen im Textverlauf analysieren. Durch den Einsatz von computerlinguistischen Komponenten wie Parsern, die die syntaktische Struktur analysieren (etwa: *X verdächtigt Y eines Vergehens*), ist eine Verfeinerung der automatischen Analyse auf inhaltlich ausdifferenzierte Relationen denkbar.

Der linguistisch-strukturelle Ansatz nähert sich interpretationsrelevanten Analysekategorien generell entlang eines Mehrebenenmodells, das die bedeutungstragenden Ausdrücke strukturell identifiziert und zueinander in Beziehung setzt. Algorithmisch kommen für die Umsetzung regelbasierte Komponenten ebenso in Frage wie statistische Verfahren, deren Parameter anhand von annotierten Korpusdaten trainiert werden (das sogenannte »überwachte« maschinelle Lernen). Die effektive Kombination von ebenenspezifischen Modulen und ein robustes Analyseverhalten bei Texten, die vom Standardszenario (zumeist Nachrichtentexte) abweichen, gehören zu den besonderen methodischen Herausforderungen für die Computerlinguistik. Für klar umrissene Zielkonfigurationen lassen sich die Komponenten jedoch häufig gut optimieren (im Sinne einer Maximierung der Vorhersagequalität auf vorab annotierten Testdaten).

Distributionelle Ansätze nähern sich interpretationsrelevanten Analysekategorien über Beobachtungen zur Verteilung des lexikalischen Materials (also der unterschiedlichen Wortformen) im Text – in der Regel, ohne grammatische Struk-

3 Fotis Jannidis u. a. verweisen auf die Problematik, wenn bei der Analyse literarischer Erzähltexte ausschließlich Standard-*Named Entity Recognition*-Systeme aus der Sprachtechnologie eingesetzt werden: nicht selten wird auf wichtige Figuren mit definiten Beschreibungen (wie »der Gärtner«) referiert. Vgl. Fotis Jannidis, Markus Krug, Isabella Reger, Martin Toepfer, Lukas Weimer und Frank Puppe: *Automatische Erkennung von Figuren in deutschsprachigen Romanen*. Digital Humanities im deutschsprachigen Raum (DHd) 2015, Graz, https://opus.bibliothek.uni-wuerzburg.de/files/14333/Jannidis_Figurenerkennung_Roman.pdf (31. Juli 2017).

4 David K. Elson, Nicholas Dames und Kathleen R. McKeown: »Extracting social networks from literary fiction«, in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10. Stroudsburg, PA, USA, 2010 (Association for Computational Linguistics), S. 138–147.

turen direkt zu berücksichtigen. Unter einer statistischen Betrachtung von (typischen vs. atypischen) Wort-Kookkurrenzen oder von Häufigkeitsprofilen des Vokabulars im textübergreifenden Vergleich lässt sich eine stilistische oder inhaltliche Verwandtschaft von Texten und Textpassagen häufig überraschend präzise erschließen. Distributionelle Ansätze erlauben es, die Ähnlichkeit zwischen zwei Texten abzuschätzen und zu beziffern (»Wie ähnlich ist das Häufigkeitsprofil der Wortformen bei Twains *Huckleberry Finn* im Vergleich zu Harriet Beecher Stowes *Uncle Tom's Cabin*?«). Paarweise angewandt auf alle Texte in einer größeren Sammlung, kann so ein »unüberwachtes« Clustering durchgeführt werden – etwa zur Hypothesengenerierung für Textverwandtschaften, die mit bloßem Auge schwer zu erkennen sind. Anders als der linguistisch-strukturelle Ansatz erfordern distributionelle Verfahren praktisch keine[5] sprachspezifische Teilkomponenten und können damit ohne großen Anpassungsaufwand auf beliebige Sprachen und historische Sprachstufen angewendet werden.

Ein Beispiel für eine Klasse von distributionellen Verfahren, die in der digitalen Literaturwissenschaft als große Erfolgsgeschichte zu bezeichnen sind, sind stilometrische Ähnlichkeitsmaße wie Burrows's Delta.[6] Es hat sich erwiesen, dass sich die stilistischen Eigenheiten einer Autorin oder eines Autors sehr stark in der relativen Verwendungshäufigkeit der unterschiedlichen Funktionswörter niederschlagen, so dass das Häufigkeitsprofil etwa der 100 häufigsten Wörter bereits bei kurzen Texten wie ein Fingerabdruck auf den Autor schließen lässt.[7] Ein anderer verbreiteter distributioneller Ansatz sind sogenannte *Topic*-Modelle,[8]

5 In der Praxis spielen allerdings sog. Stoppwortlisten (für die häufigsten Funktionswörter einer Sprache, d. h. Artikel, Auxiliare etc.) eine wichtige Rolle bzw. Verfahren zur Bestimmung von hochfrequenten Eigennamen in einem Text; Hintergrund ist, dass zwar generell die am häufigsten auftretenden Wortformen Funktionswörter sind, während einzelne Typen von Inhaltswörtern seltener verwendet werden. In einzelnen Texten oder in kleineren, inhaltlich zusammenhängenden Korpora treten jedoch i. d. R. bestimmte Inhaltswörter, insbesondere Eigennamen, gehäuft auf.

6 John Burrows: »›Delta‹: A Measure of Stylistic Difference and a Guide to Likely Autorship«, in: *Literary and Linguistic Computing* 17 (2002), S. 267–287; vgl. hierzu auch den Beitrag von Schöch (in diesem Band).

7 U. a. Fotis Jannidis und Gerhard Lauer: »Burrows's Delta and Its Use in German Literary History«, in: *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, hg. v. Matt Erlin und Lynne Tatlock. Rochester 2014, S. 29–54; Stefan Evert, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch und Thorsten Vitt: »Towards a better understanding of Burrows's Delta in literary authorship attribution«, in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver 2015, S. 79–88.

8 Thomas K. Landauer, Peter Foltz und Darrell Laham: »Introduction to Latent Semantic Analysis«, in: *Discourse Processes* 25 (1998), S. 259–284; David M. Blei, Andrew Y. Ng und Michael I.

durch die anhand eines relativ großen Textkorpus Cluster von (semantisch) ähnlichen Wörtern über das Vokabular der Sprache induziert werden – einzig aufgrund der angenommenen Tendenz, dass innerhalb eines Textabschnitts eher inhaltlich zusammengehörige Wörter auftreten. Die Cluster stehen im Ergebnis nicht für eine definierte Bedeutungsdimension (wie z. B. Kulinarik), nähern sich interpretierbaren semantischen Wortfeldern jedoch oft an. Allerdings schließt das statistisch induktive Verfahren nicht aus, dass ein etabliertes semantisches Feld »quer« zu den induzierten *Topic*-Clustern liegt, weshalb man eine unreflektierte Gleichsetzung der technischen *Topics* mit Themenfeldern bei der Meta-Analyse vermeiden sollte.

Topic-Modelle werden vielfältig eingesetzt, um für einen unbekannten Text eine »latente semantische Struktur« zu approximieren: ein einmal trainiertes *Topic*-Modell kann verwendet werden, um ohne einen händischen Eingriff Passagen zu trennen, in denen die Wörter stark zu unterschiedlichen Cluster-Zugehörigkeiten tendieren. Die Tatsache, dass kein überwachtes Training erforderlich ist, macht *Topic*-Modelle zu einem attraktiven Explorationswerkzeug; allerdings erweist es sich in der Praxis auch als problematisch, dass die Wahl der Modellparameter (wie der vorgegebenen Zahl der *Topic*-Cluster) i. d. R. unterdeterminiert ist und es mitunter schwer zu beurteilen ist, welche Modellvorhersagen eine systematische Basis haben. In den *Digital Humanities* wird der methodische Status von *Topic*-Modellen seit Jahren recht ausgiebig diskutiert.[9]

Innerhalb der Computerlinguistik kommen sehr weit entwickelte distributionelle Modelle für korpusbasierte Ansätze zur lexikalischen Semantik zum Einsatz (und es ist eine Frage von einiger Brisanz, welches die bestgeeignete Modellerweiterung ist, um die Semantik größerer sprachlicher Einheiten zu erfassen[10]). Gerade dank der erfolgreichen Neuauflage von Lernverfahren, die mit künstlichen neuronalen Netzen arbeiten (dem sog. »Deep Learning«), haben »neuronale« distributionelle Modelle große Verbreitung gefunden – am bekanntesten

Jordan: »Latent dirichlet allocation«, in: *Journal of machine Learning research* 3 (2003), S. 993–1022.

9 Clay Templeton: *Topic Modeling in the Humanities: An Overview. Maryland Institute for Technology in the Humanities*, 2011. http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview (28. April 2017); Megan R. Brett: »Topic Modeling: A Basic Introduction«, in: *Journal of Digital Humanities* 2012, S. 12–17.

10 U. a. Beiträge in Cécile Fabre und Alessandro Lenci: *TAL Journal: Special issue on Distributional Semantics (Traitement Automatique des Langues / Natural Language Processing)* 56.2 (2015).

ist hier das *word2vec*-Modell.[11] Jede Wortform einer Sprache wird in einem neuronalen Modell als ein Zahlenvektor mit beispielsweise 1000 Dimensionen repräsentiert, wobei die Zahlenwerte der Aktivierungsstärke bestimmter Neuronen entsprechen; ähnliche Wörter werden durch ähnliche Aktivierungsprofile über die Dimensionen hinweg repräsentiert. Das große Potenzial der Modelle rührt daher, dass sich die Aktivierungslevels für eine bestimmte Wortform in einem zyklischen Lernprozess, der ein (meist sehr großes) Textkorpus in vielen Iterationen durchläuft, selbständig »einpegeln«. Die Tendenz zweier Wortformen zur Kookkurrenz führt in einem hochgradig verflochtenen Neuronennetzwerk zur Verstärkung der Synapsen zwischen denjenigen Neuronen, die jeweils charakteristische Worteigenschaften repräsentieren. Dabei geht die datengesteuerte Induktion der »dichten« konnektionistischen Repräsentation Hand in Hand mit der Ausprägung der Synapsen – sodass im Laufe des Trainings eine kompakte Darstellung entsteht (qua »Bootstrapping«), die gerade jene Generalisierungen erfasst, die sich in beobachtbaren Mustern im Korpus niederschlagen. Für das *word2vec*-Modell basiert das Training auf einem Kontextfenster von fünf Wörtern, deren Vektor-Repräsentationen sich wechselseitig beeinflussen. Sehr viel Beachtung haben die Analogieschlüsse gefunden,[12] zu denen dieses Modell im Ergebnis in der Lage ist: man kann dasjenige Wort *X* bestimmen, dessen Vektor zu dem eines vorgegebenen Ausgangsworts (z. B. *actor*) am ehesten im gleichen Verhältnis (also *actor* : *X*) steht wie die Vektoren eines anderen Wortpaares (z. B. *king* : *queen*), und in sehr vielen Fällen führt dies zum erwarteten Ergebnis (hier *X* = *actress*) – obgleich im Training keinerlei explizite semantische Information zur Verfügung gestellt wurde: das Lernen basiert ausschließlich auf reinen Oberflächenfolgen von Wörtern in einem Korpus (welches allerdings sehr umfangreich sein sollte, um robuste Ergebnisse zu erzielen).

Die Kombination von Analysekomponenten. Gerade für anspruchsvollere analytische Fragestellungen, wie sie wohl mit den meisten literaturwissenschaftlichen Untersuchungen zu einem Text oder einem Textkorpus einhergehen (also jenseits der reinen Text- bzw. Korpus*exploration*), kann die unmittelbare Anwendung computerlinguistischer Standardmodelle und -werkzeuge zu Ergebnissen führen, die nur eingeschränkt aussagekräftig sind. Etablierte linguistisch-strukturelle Werkzeuge sind in der Regel auf kanonische linguistische Analyseebenen

11 Tomas Mikolov, Greg Corrado, Kai Chen und Jeffrey Dean: *Efficient estimation of word representations in vector space*. 2013. arXiv preprint. arXiv:1301.3781.

12 Vgl. u. a. Omer Levy und Yoav Goldberg: »Linguistic Regularities in Sparse and Explicit Word Representations«, in: *Proceedings of the Eighteenth Conference on Computational Language Learning*, 2014, S. 171–180.

(z. B. syntaktische Dependenzstruktur) und/oder typische anwendungsrelevante Kategorien (z. B. Namen von Personen, Firmen, Produkten und »geopolitischen Entitäten«) ausgerichtet und optimiert – eine literaturwissenschaftliche Untersuchung will jedoch zumeist auf eine davon abweichende Zielstruktur oder Kategorisierung hinaus.[13] Zudem wurden die verfügbaren Werkzeuge in aller Regel auf gegenwartssprachlichen Nachrichtentexten entwickelt und bedürfen der Anpassung, will man andere Sprachregister und historische Sprachstufen mit vergleichbaren Qualitätsansprüchen analysieren.

Bei unmodifizierten distributionellen Werkzeugen dürfte (trotz der Unabhängigkeit von Spezifika der Subsprache bzw. des Sprachstadiums) häufig deren »strukturelle Blindheit« zu Einschränkungen bei der Interpretierbarkeit führen – für sie manifestiert sich jede Textpassage ausschließlich in den Häufigkeiten der darin auftretenden Wortformen. Zwar können Filter angesetzt werden, die den Blick auf einen Ausschnitt des Vokabulars lenken (z. B. durch Ausschluss mittels Stoppwortliste), diese fungieren jedoch global und können kontextuelle Abhängigkeiten nicht berücksichtigen. Gut illustriert wird die Problematik durch das einfache Beispiel der Negation. Eine Passage wie *X hatte weder den Mut für die Reise, noch war er ein Kenner des Landes* könnte unter einem distributionellen Ansatz die Figur X mit Eigenschaften in Verbindung bringen, die ihr explizit abgesprochen werden (da sie im Skopus der Negation *weder … noch …* stehen). Von größerer praktischer Relevanz dürfte diese Problematik bei längeren Einbettungen sein, wie Sprüngen in der Erzählebene oder Schilderungen der Sinneswahrnehmung einer Figur. Aber bereits die Zuordnung von distributionell erfassten semantischen Eigenschaften zu Figuren, Orten o. ä. – unabhängig von einer möglichen Negation oder modalen Einbettung – lässt sich nur mit einer strukturbezogenen Erweiterung der Basismodelle präzise erfassen.

Bestehende computerlinguistische Werkzeuge und Modelle können allerdings in vielen Fällen so erweitert und/oder kombiniert werden, dass sie für weitergreifende analytische Aufgaben eingesetzt werden können. (Nicht zuletzt deshalb stellt die mögliche Einbeziehung von Fragestellungen zu literarischen Texten eine attraktive Perspektive für die Computerlinguistik dar.)

Im verbleibenden Teil dieses Abschnitts soll ein konkretes Beispiel angeführt werden, das einerseits verdeutlicht, wie stark auf bestehende Lösungen aufgebaut werden kann, andererseits aber auch zeigt, dass für tragfähige Ergebnisse

13 Vgl. auch Fotis Jannidis u. a.: *Automatische Erkennung von Figuren in deutschsprachigen Romanen*.

zusätzliche Arbeit bei der Modellentwicklung notwendig ist (ebenso für die Entwicklung von Arbeitspraktiken, die automatische Werkzeuge geeignet in ein methodenkritisches Vorgehen einbetten).

Ausgangspunkt sei – zu rein illustrativen Zwecken – eine konventionelle distributionelle Vergleichsanalyse einiger Texte von Mark Twain und einer Anzahl von möglichen Vergleichstexten, die rasch aus volltextdigitalisiert verfügbaren Quellen zusammengestellt wurde[14] – den Romanen aus Mark Twains Zyklus um Tom Sawyer und Huckleberry Finn: *The Adventures of Tom Sawyer* (1876), *Adventures of Huckleberry Finn* (1884), *Tom Sawyer Abroad* (1894), *Tom Sawyer, Detective* (1896), Twain: *The Prince and the Pauper* (1881, historischer Roman, der im 16. Jh. am englischen Königshof spielt), *Roughing It* (1872, Reiseberichte, teils autobiographisch), *Following the Equator* (1897, Reisebericht); Harriet Beecher Stowe: *Uncle Tom's Cabin* (1852, Roman, der die Sklaverei thematisiert); Thomas Bailey Aldrich: *The Story of a Bad Boy* (1870, Abenteuererzählung), Booth Tarkington: *Penrod* (1914, Abenteuererzählung); Artemus Ward: *To California and Return* (Teil 4 der gesammelten Werke, Reiseberichte).

Abbildung 1 zeigt eine einfache distributionelle Analyse, in der aufgrund der Ähnlichkeit in der Häufigkeitsverteilung des Textvokabulars ein hierarchisches Clustering über allen Texten erzeugt wurde.

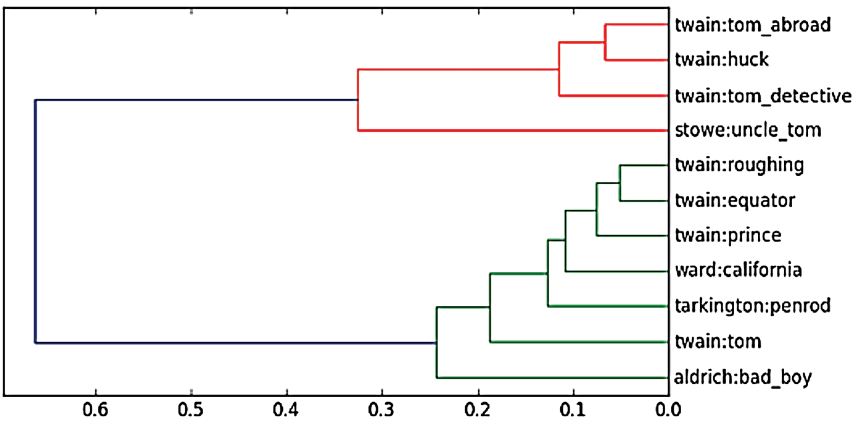

Abb. 1: Hierarchisches Clustering nach distributioneller Ähnlichkeit

14 Die Untersuchung basiert auf der Version der Texte auf gutenberg.org.

Die Baumdarstellung enthält in den feiner verzweigten Ästen jeweils die ähnlichsten Texte. Insgesamt scheint die Anordnung auf den ersten Blick die grobe, intuitive Erwartungen zu erfüllen, dass sich verwandte Handlungsorte der Texte und ähnliche gesellschaftliche Umstände stark in der Lexik niederschlagen: Twains Abenteuerromane aus dem Tom Sawyer-Zyklus, lokalisiert in der Sphäre der Südstaaten, bewegen sich im gleichen Bereich, *Uncle Tom's Cabin* findet sich in ihrer Nähe; hingegen clustert sich beispielsweise Twains Reiseliteratur eher mit Wards Reiseberichten.

Überraschend erscheint dann jedoch, dass der Vergleich von *The Adventures of Tom Sawyer* (in Abbildung 1 bezeichnet als *twain:tom*) und den anderen Tom Sawyer/Huckleberry Finn-Romanen zu einer relativ geringen Ähnlichkeit führte (wohingegen ersterer Roman erwartungsgemäß recht große Ähnlichkeiten mit den ähnlich lokalisierten Abenteuerromanen *The Story of a Bad Boy* und *Penrod* von Aldrich und Tarkington aufwies). Die detailliertere distributionelle Analyse mit einem *Topic*-Modell[15] – illustriert in Abbildung 2 auf der Folgeseite – hilft dabei, eine plausible Erklärung zu finden.[16]

In dieser Abbildung werden die Texte durch Säulen repräsentiert, die sich zu jeweils unterschiedlichen Anteilen aus Wörtern konstituieren, die das zugrundeliegende *Topic*-Modell jeweils einem von zehn verschiedenen induzierten *Topic*-Clustern zugeordnet hat. Die drei späteren Tom Sawyer/Huckleberry Finn-Romane – *Adventures of Huckleberry Finn* (6. Säule von links), *Tom Sawyer Abroad* (10.), sowie *Tom Sawyer, Detective* (11.) – enthalten jeweils einen sehr dominanten Anteil des hellblau dargestellten *Topics* #3 – welches in den ursprünglichen *Adventures of Tom Sawyer* (9. Säule) praktisch fehlt. Betrachtet man die dominanten Wortformen, die dieses *Topic* prägen, wird deutlich: es handelt sich um dialektal-umgangssprachliche Formen (»ain't, didn't, warn't«) – die drei Romane sind alle aus der Perspektive von Huckleberry Finn in Ich-Form und in der Umgangssprache des »Pike County dialect« verfasst[17] (die *Adventures of Huckleberry Finn* beginnen beispielsweise wie folgt: *You don't know about me without you have*

15 Die Analyse wurde mit dem Mallet-Toolkit (mallet.cs.umass.edu) durchgeführt. Ein Tutorium zu einfachen Analysen findet sich unter https://de.dariah.eu/tatom/topic_model_mallet.html (31. Juli 2017).

16 Die *Topic*-Analyse in Abbildung 2 enthält mit *The Rector of Veilbye* (1829) zusätzlich die englische Übersetzung einer Novelle des Dänen Steen Blicher. Es gab Debatten, ob Twain die Handlung zu *Tom Sawyer, Detective* aus dieser Erzählung übernommen habe.

17 David Carkeet: »The Dialects in Huckleberry Finn«, in: *American Literature* 51.3 (1979), S. 315–332 (zitiert nach Sieglinde Lemke: *The Vernacular Matters of American Literature*. New York 2009).

read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter). Dagegen ist *The Adventures of Tom Sawyer* in dritter Person von einem allwissenden Erzähler geschildert (der sich möglicherweise besonders stark in *Topic* #8 niederschlägt – charakteristisch sind hier Wörter wie »boy, boys, began«).

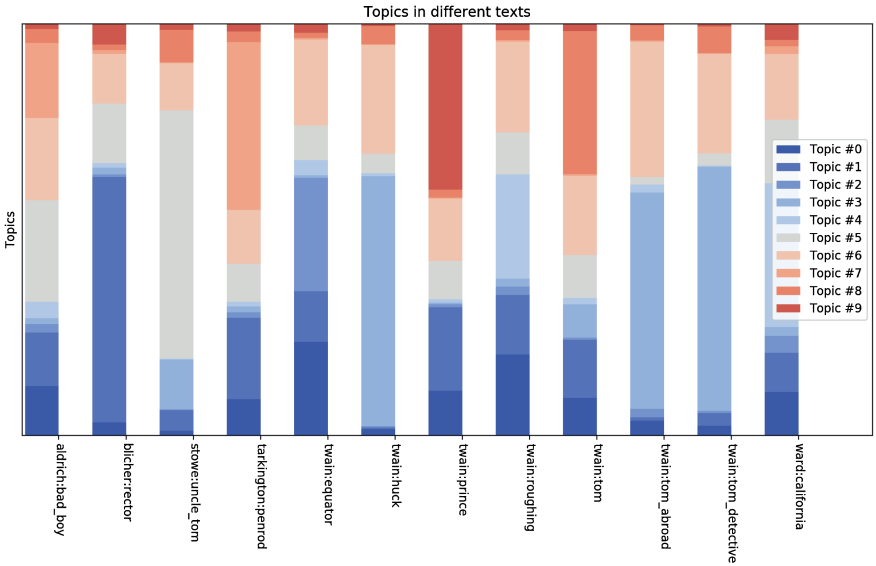

Abb. 2: Latente Topic-Analyse verschiedener Texte Mark Twains und einiger Vergleichstexte

Ein sehr textspezifisch charakteristisches *Topic* ist im übrigen #9, das fast ausschließlich in Twains *The Prince and the Pauper* (7. Säule) zum Tragen kommt. Es wird von der historischen englischen (Hof-)Sprache dominiert und enthält im Kern »thou, thy, Lord«.

Die rein distributionelle Analyse der Gesamttexte lässt bereits erahnen, dass bei Mark Twain eine Differenzierung der stilometrischen Untersuchungen nach Figurenrede (ggf. im Vergleich zu unterschiedlichen Erzählerstimmen) zu einer reicheren Grundlage für Detailanalysen führen dürfte. Eine solche Differenzierung ist nur möglich, wenn der distributionelle Ansatz mit einem linguistisch-strukturellen Vorgehen gekoppelt wird: aus dem Erzähltext muss die wörtliche Rede extrahiert werden und den unterschiedlichen Figuren zugeordnet werden, so dass beispielsweise die gesamte Figurenrede von Tom Sawyer und von Jim distributionell untersucht werden kann; ebenso müssen die verbleibenden Textpassagen extrahiert werden, die der Erzählerstimme zuzuordnen ist.

Der linguistisch-strukturelle Anteil dieser Analyseaufgabe ist klar umrissen und kann zu einem guten Teil auf bestehende Komponenten zurückgreifen. Dennoch ist die Aufgabe alles andere als trivial, wie im Folgenden kurz ausgeführt werden soll.[18] Beispiel (T1) ist eine Passage aus *The Adventures of Tom Sawyer*, Kapitel 3.

(T1)

> Tom turned homeward alone.
> As he was passing by the house where Jeff Thatcher lived, he saw a new girl in the garden.
> *[400 Wörter, ohne Verwendung des Eigennamens ›Tom‹]*
> He returned, now, and hung about the fence till nightfall, »showing off,« as before; but the girl never exhibited herself again, though Tom comforted himself a little with the hope that she had been near some window. [...]
> All through supper his spirits were so high that his aunt wondered »what had got into the child.« He took a good scolding about clodding Sid, and did not seem to mind it in the least. He tried to steal sugar under his aunt's very nose, and got his knuckles rapped for it. He said:
> »Aunt, you don't whack Sid when he takes it.«
> »Well, Sid don't torment a body the way you do. You'd be always into that sugar if I warn't watching you.«

In Bezug auf diese Passage besteht unsere Analyseaufgabe konkret Aufgabe darin, die beiden Äußerungen am Ende des Ausschnitts jeweils einer Figur zuzuordnen. Die korrekte Lösung ist: die erste Äußerung stammt von Tom, die zweite von Tante Polly. Um zu diesem Ergebnis mit einem algorithmischen Verfahren zu gelangen, sind in einer Reihe von (Standard-)Analyseschritten relevante Entscheidungen zu treffen:

(1) Tokenisierung (und Satzerkennung): dieser Standardschritt der Vorverarbeitung überführt den digitalisierten Text in eine Folge von sog. Tokens, d. h. Basisanalyseeinheiten für alle weiteren Schritte. In erster Näherung werden Leerzeichen und Zeilenumbrüche zur Trennung von Tokens herangezogen; außerdem muss jedoch bei Interpunktionssymbolen eine Entscheidung getroffen werden: ein Punkt bei einer Folge wie *Mr. Walters* ist Bestandteil eines Tokens »Mr.«, alle Punkte in der Passage (T1) markieren jedoch eine Satzgrenze. Tokenisierung und Satzgrenzenerkennung greifen also ineinander. Für die Redezuordnung stellt sich bereits in diesem Schritt eine nicht immer triviale Aufgabe: Textpassa-

18 David K. Elson und Kathleen R. McKeown: »Automatic attribution of quoted speech in literary narrative«, in: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (AAAI '10) 2010. AAAI Press, S. 1013–1019.

gen, die (hier) in doppelten Anführungszeichen eingeschlossen sind, sind Kandidaten für wörtliche Rede; (T1) enthält allerdings zwei Verwendungsbeispiele von Anführungszeichen, die keine direkte Rede signalisieren (bei *»showing of«* handelt es sich um einen modalisierenden Gebrauch; *»what had got into the child«* dürfte einen nicht ausgesprochenen Gedanken wiedergeben, der (vermutlich) nicht die gleiche Behandlung wie wörtliche Rede erfahren sollte. Als Indikatoren für wörtliche Rede wird man neben den Anführungszeichen Inquit-Formeln (*He said:*), sowie die Typographie (Zeilenumbrüche) und typische Muster (längere Dialogsequenzen, auch ohne Inquit-Formeln) heranziehen.

(2) Wortartenerkennung, einschl. Erkennung von Eigennamen: ein Analyseschritt, der mit recht hoher Analysequalität automatisch durchgeführt werden kann, ist die Zuweisung von Wortkategorien zu den Tokens (das sog. *Part-of-Speech-Tagging*). PoS-Tagger erzielen oft auch für Subsprachen oder Sprachstufen, für die eine detailliertere automatische Analyse problematisch ist, passable Ergebnisse. In der Regel werden in diesem Schritt die Bestandteile typischer Eigennamen (wie wir sie für die Zuordnung der wörtlichen Rede benötigen) erkannt; für komplexere Ausdrücke, die auf Figuren referieren (z. B. *the minister*), können evtl. Muster von Wortartenfolgen definiert werden.

Neben der Wortartenerkennung ist in der Sprachtechnologie auch die speziellere Aufgabe der Erkennung von Eigennamen etabliert (unter der etwas irreführenden Bezeichnung *Named Entity Recognition* (NER), wobei nicht wirklich Entitäten erkannt werden, sondern Namen im Text, die i. d. R. auf Entitäten referieren). Die Aufgabe beinhaltet, Beginn und Ende eines Namensausdrucks zu identifizieren. Für viele anwendungsrelevante Informationsextraktions-Aufgaben ist eine präzise und umfassende Erkennung von Entitätenbezeichnungen beispielsweise von Personen, Firmen, Produkten etc. sehr zentral, und so gibt es unabhängig von PoS-Taggern NER-Komponenten, die häufig auch für Spezialaufgaben angepasst werden können.

Wie Beispiel (T1) zeigt, ist eine Beschränkung auf Schritt (1) und (2) für die Rede-Zuordnung nicht ausreichend: würde man z. B. für die Äußerung *»Aunt, you don't whack Sid when he takes it.«* den nächstliegenden Eigennamen im Vorkontext suchen, würde man sie womöglich fälschlicher Weise Sid zuordnen. Für eine zuverlässige Analyse müssen anaphorische Pronomina wie das Personalpronomen in *He said* auf die Figuren abgebildet werden. Bevor dies ermöglicht werden kann, ist ein zusätzlicher vorbereitender Zwischenschritt erforderlich:

(3) Grammatische Analyse: Für vielfältige weitergehende Analysen ist eine Erfassung der syntaktischen Struktur der Sätze erforderlich – eines der klassischen Forschungsgebiete der Computerlinguistik. Es wird zwischen einer Dependenzanalyse (in der die grammatischen Relationen zwischen den Wörtern im

Vordergrund stehen – *he* als Subjekt von *said*) und einer Phrasenstruktur- oder Konstituentenanalyse unterschieden (die z. B. *the girl* oder auch *his aunt's very nose* als Nominalphrasen erkennt, daneben aber ebenso Ein-Wort-Nominalphrasen wie *himself* und *she*). Um die Referenzen auf Figuren zusammenzuführen, ist in unserem Zusammenhang vor allem eine Konstituentenanalyse erforderlich; viele interpretationsrelevante Analysen können jedoch robust auf einer Dependenzanalyse aufbauen. Für die automatische grammatische Analyse, das *Parsing*, gibt es unzählige Modellierungsansätze, die teils mehr, teils weniger explizites grammatisches Regelwissen voraussetzen – in den letzten Jahren durchgesetzt haben sich statistische Ansätze, in die Regelwissen indirekt durch überwachtes Training auf handannotierten Korpora (sog. Baumbanken) eingeht. Da eine vollständige Annotation von syntaktischen Strukturen sehr zeitaufwändig ist, liegen qualitativ hochwertige *Parser* nur für relativ wenige Sprachen bzw. Sprachstadien/Subsprachen vor. In letzter Zeit wird jedoch an *Parsern* gearbeitet, die sprachenübergreifend eine robuste (grobe) Analyse ermöglichen.[19]

(4) Koreferenzanalyse: In diesem Schritt werden alle Kandidaten für referierende Ausdrücke (d. h. in etwa alle Nominalphrasen) in einem Text herangezogen, und es wird entschieden, welche davon auf die gleiche (reale oder fiktionale, evtl. auch abstrakte) Entität referieren und deshalb in die gleiche Koreferenzkette eingeordnet werden.[20] Zum Einsatz kommen dabei heute zumeist maschinelle Lernverfahren, die eine Vielzahl von Indikatoren in Rechnung stellen und insofern sowohl grammatische Kriterien (wie die Genus-Kongruenz von Pronomina mit ihrem Antezendens) als auch beispielsweise Muster der lokalen Textkohärenz (bereits eingeführte stark saliente Entitäten verbleiben überwiegend in der Rolle des Subjekts) einbeziehen und zueinander gewichten.

In (T2) ist das Ergebnis einer manuellen Koreferenzanalyse für unsere Textpassage illustriert; Zahlenindices und unterschiedliche typographische Hervorhebungen verdeutlichen die entstehenden Ketten (nicht hervorgehoben sind einige referentielle Ausdrücke, die hier nur einmal auftauchen, wie *Jeff Thatcher*

19 Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev und Keith B. Hall: »Universal Dependency Annotation for Multilingual Parsing«, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), Sofia 2013, S. 92–97.

20 Anders Björkelund und Jonas Kuhn: »Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features«, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore 2014, S. 47–57; Ina Rösiger und Jonas Kuhn: »IMS HotCoref DE: A Data-Driven Co-Reference Resolver for German«, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož 2016, S. 155–160.

und *the garden*; in der Trennung der Bezüge auf wiederkehrende Figuren von Nebenfiguren oder Gegenständen liegt jedoch eine der großen Herausforderungen für die automatische Analyse).

(T2)

Tom$_1$ turned homeward alone.
As **he**$_1$ was passing by the house where Jeff Thatcher lived, **he**$_1$ saw a new girl$_2$ in the garden.
[400 Wörter, ohne Verwendung des Eigennamens ›Tom‹]
He$_1$ returned, now, and hung about the fence till nightfall, »showing off,« as before; but the girl$_2$ never exhibited herself$_2$ again, though **Tom**$_1$ comforted **himself**$_1$ a little with the hope that she$_2$ had been near some window. [...]
All through supper **his**$_1$ spirits were so high that [**his**$_1$ aunt]$_3$ wondered »what had got into **the child**$_1$.« **He**$_1$ took a good scolding about clodding *Sid*$_4$, and did not seem to mind it in the least. **He**$_1$ tried to steal sugar under [**his**$_1$ aunt]$_3$'s very nose, and got **his**$_1$ knuckles rapped for it. **He**$_1$ said:
»[Aunt]$_3$, [you]$_3$ don't whack *Sid*$_4$ when *he*$_4$ takes it.«
»Well, *Sid*$_4$ don't torment a body the way **you**$_1$ do. **You**$_1$'d be always into that sugar if [I]$_3$ warn't watching **you**$_1$.«

Anhand des Beispiels dürfte deutlich werden, dass eine Automatisierung der Analyse auf längeren Erzähltexten alles andere als trivial ist: der Eigenname *Tom* als Anker für eine Vielzahl von anaphorischen Ausdrücken taucht über lange Strecken nicht auf; beim Lesen trägt nicht selten das Inhaltsverständnis zur Auflösung von referentiellen Ambiguitäten bei. Eine vollautomatische Analyse kann daher derzeit nicht als Basis für streng quantitative Untersuchungen herangezogen werden; mit heuristischen Filtern oder einer manuellen Nachanalyse erschließen die verfügbaren Koreferenz-Werkzeuge jedoch erhebliche Textbereiche für eine Figurenanalyse, die von einem konventionellen namensbasierten Verfahren nicht berücksichtigt werden können.

Die Zuordnung der vorletzten Äußerung im Ausschnitt als wörtliche Rede Toms würde (bei einer perfekten Koreferenzanalyse) mit den dargestellten Analyseschritten (1)–(4) korrekt erfasst. Die letzte Äußerung (von Tante Polly) zeigt jedoch, dass die Schritte im allgemeinen Fall immer noch nicht eindeutig zum korrekten Ergebnis führen. Es fehlt eine Inquit-Formel. Die Leserin oder der Leser erschließt aus mehreren Indikatoren, dass es sich beim direkten Gegenüber in dem einsetzenden Dialog um Polly handeln muss: (i) im vorangegangenen Satz sind es bereits Tom und seine Tante, die miteinander interagieren (recht schmerzhaft für Tom ...); (ii) Toms Äußerung enthält eine Vokativ-Anrede der Tante; (iii) auf Sid, der als weitere Figur in der Passage salient ist, wird in beiden Äußerungen in der dritten Person referiert, so dass er vermutlich nicht unmittelbar zugegen ist. Nicht alle diese Aspekte ließen sich wohl in einem Computermodell erfassen, aber das Beispiel verdeutlicht, dass für eine verlässliche Redezuordnung

in Erzähltexten ein Schritt (5) der Dialogmodellierung angemessen wäre. Dialogmodelle werden in der Sprachtechnologie bislang hauptsächlich in interaktiven Dialogsystemen und für die Aufgabe des sogenannten *Speaker Tracking* eingesetzt. Eine Übertragung und Anpassung auf literarische Texte ist jedoch denkbar. (Zu erwarten wäre sicherlich ein hoher Grad an Genre- und Autorenabhängigkeit; der introspektive Leseeindruck ist, dass die Muster der Zuordnung teilweise stark konventionalisiert sind.[21])

Für die hier skizzierte Beispielstudie hat der Autor die Schritte (1)–(4) mit Analysewerkzeugen aus der Stanford CoreNLP-Sammlung durchgeführt.[22] Das Ergebnis der automatischen Koreferenz-Erkennung kann nicht ohne Nachbereitung verwendet werden, ist jedoch an vielen Stellen korrekt. Abbildung 3 zeigt einen Ausschnitt aus der Ausgabe, die sich für unsere Passage ergibt, visualisiert mit dem Explorationswerkzeug ICARUS.[23] Jede Koreferenzkette ist mit einem numerischen Index markiert und in einem eigenen Grünton hervorgehoben.

21 Zusätzlich verkompliziert wird die Modellierung, wenn sich in der fiktionalen Welt mehrere Wahrnehmungsebenen überlagern; so geben sich in *Huckleberry Finn* in einer Passage in Kapitel 41 Tom und Huck als Sid und Tom aus; entsprechend werden sie in der wörtlichen Rede der Dialoge angeredet, die Einbettung in die Erzählersicht (Hucks Sicht) referiert jedoch auf die tatsächlichen Identitäten – abgesehen von einigen Passagen, in denen er die Ebenen durch Referenz mit Anführungszeichen andeutet: »So away I shoved, and turned the corner, and nearly rammed my head into Uncle Silas's stomach! He says: ›Why, Tom! Where you been all this time, you rascal?‹ ›I hain't been nowheres,‹ I says, ›only just hunting for the runaway nigger – me and Sid.‹ [...] So then we went to the post-office to get ›Sid‹« (Twain: *Huckleberry Finn*, Kap. 41).

22 stanfordnlp.github.io; Werkzeuge für die Analyse von deutschen Texten sind über die CLARIN-D-Infrastruktur verfügbar (www.clarin-d.de).

23 Markus Gärtner, Anders Björkelund, Gregor Thiele, Wolfgang Seeker und Jonas Kuhn: »Visualization, Search, and Error Analysis for Coreference Annotations«, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2014; Markus Gärtner, Katrin Schweitzer, Kerstin Eckart und Jonas Kuhn: »Multi-modal Visualization and Search for Text and Prosody Annotations«, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Peking 2015, http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/icarus.html (31 Juli 2017).

39: [1][Tom][1] came up to [16][the fence][16] and leaned on [16][it][16] , grieving , and hoping [58][she][58] would tarry y
40: [58][She][58] halted [59][a moment][59] on the steps and then moved toward the door .
41: [1][Tom][1] heaved a great sigh as [58][she][58] put [58][her][58] foot on the threshold .
42: But [54][[1][his][1] face][54] lit up , right away , for [58][she][58] tossed [55][a pansy][55] over [16][the fence][16] [59][a mon
43: [52][The boy][52] ran around and stopped within [56][a foot][56] or two of the flower , and then [57][shaded [52][his
44: Presently [52][he][52] picked up a straw and began trying to balance [56][it][56] on [62][[52][his][52] nose][62] , with [52]
45: But only for a minute -- only while [52][he][52] could button [61][the flower inside [52][his][52] jacket , next [46][[52][h
46: [52][He][52] returned , now , and hung about [16][the fence][16] till nightfall , `` showing off , " as before ; but [5
47: Finally [52][he][52] strode home reluctantly , with [52][his][52] poor head full of visions .
48: All through supper [52][his][52] spirits were so high that [39][[52][his][52] aunt][39] wondered `` what had got into
49: [52][He][52] took [64][a good scolding][64] about clodding [30][Sid][30] , and did not seem to mind [64][it][64] in the le
50: [52][He][52] tried to steal sugar under [63][[52][his][52] aunt 's [62][very nose][62]][63] , and got [52][his][52] knuckles rap
51: [52][He][52] said : `` [63][Aunt][63] , [65][you][65] do n't whack [30][Sid when [30][he][30] takes [62][it][62]][30] . "
52: `` Well , [30][Sid][30] do n't torment a body [8][the way][8] [65][you][65] do .
53: [65][You][65] 'd be always into that sugar if [30][I war][30] n't watching [65][you][65] . "
54: Presently [65][she][65] stepped into the kitchen , and [30][Sid][30] , happy in [30][his][30] immunity , reached for th
55: But [30][Sid 's][30] fingers slipped and the bowl dropped and broke .
56: [1][Tom][1] was in ecstasies .
57: In such ecstasies that [1][he][1] even controlled [1][his][1] tongue and was silent .

Abb. 3: Ergebnis einer automatischen Koreferenz-Analyse mit den Stanford CoreNLP-Werkzeugen, visualisiert mit der ICARUS-Oberfläche

Der gezeigte Ausschnitt macht einige der typischen Schwierigkeiten deutlich: für jede der Hauptfiguren erzeugt das auf Nachrichten trainierte Werkzeug mehrere separate Koreferenzketten, die noch zusammengeführt werden müssten. Für *Tom* liegt in Satz 39–42 die Kette mit dem Index 1 vor, weitergeführt ab Satz 56. Dazwischen setzt das Werkzeug eine andere Kette an, die von *the boy* aufgespannt wird (Index 52, eingeführt in Satz 43). Dass es sich bei dem Jungen um Tom handelt, setzt in der Tat ein tiefes Inhaltsverständnis voraus – an anderer Stelle wird mit *the boy* selbstverständlich auf andere Figuren referiert. Mit interaktiver Nachbereitung bzw. einigen Heuristiken lassen sich aber derartige Fälle relativ robust behandeln.

Die Referenz auf den Zaun der Familie Thatcher wird in der automatischen Analyse sehr gut erfasst (Index 16: Satz 39, 42, 46), hingegen werden in Satz 44 und 50 die Referenzen auf Nasen fälschlich zusammengelegt (Index 62). Ein zusätzliches Problem liegt in deiktischen Pronomen in der wörtlichen Rede: *you* wird in Satz 51 nicht als koreferent mit dem Vokativ *Aunt* erkannt, dafür entgeht dem System der Sprecherwechsel zwischen Satz 51 und 52 (was nicht weiter verwunderlich ist, da keine explizite Repräsentation für Figurenrede erzeugt wird – was jedoch bei Weiterentwicklungen denkbar wäre).

Für die Zwecke der hier diskutierten illustrativen Studie wurden auf Basis der automatischen Koreferenz-Erkennung eine Reihe von heuristischen Regeln formuliert, welche die wörtliche Rede in den verhältnismäßig klaren Fällen via Koreferenzkette namentlich genannten Figuren zuordnen; unklare Fälle wurden herausgefiltert. Da die Figurenrede in *Adventures of Huckleberry Finn* hinsichtlich der Stilistik differenzierter ist als in den *Adventures of Tom Sawyer,* betrachten wir hier den zweiten Roman im Zyklus.

Die nach diesem Verfahren zugewiesenen Redebeiträge der sieben Figuren mit den größten Redeanteilen (Huck [»I«], der als Ich-Erzähler seine eigene Rede wiedergibt, Jim, Hucks Vater [»Pap«], Ben Rogers, Tante Pollys Schwester Sally, Tom, und der »König«) sowie der Erzählertext können nun separat stilometrisch untersucht werden. Abbildung 4 zeigt in Entsprechung zur werkübergreifenden Analyse in Abbildung 2 eine *Topic*-Analyse mit 10 *Topic*-Clustern (die hier nur auf den Figurenreden und dem Erzählertext in *Huckleberry Finn* induziert wurden). Als ein sehr charakteristisches *Topic* erweist sich #2 (dargestellt mit dem mittleren Blauton), das bei Jim wesentlich stärker als bei den anderen Figuren ausgeprägt ist und das in der Tat von Charakteristika in Twains Wiedergabe der afroamerikanischen Umgangssprache dominiert wird (»de, dat, dey«).[24]

24 Lisa Cohen Minnick präsentiert eine detaillierte linguistische Studie von Twains Charakterisierung der Sprache Jims, in der sie auch auf die Debatte um rassistische Stereotype eingeht, dies.: »Jim's language and the issue of race in Huckleberry Finn«, in: *Language and Literature* 10.2 (2001), S. 111–128.

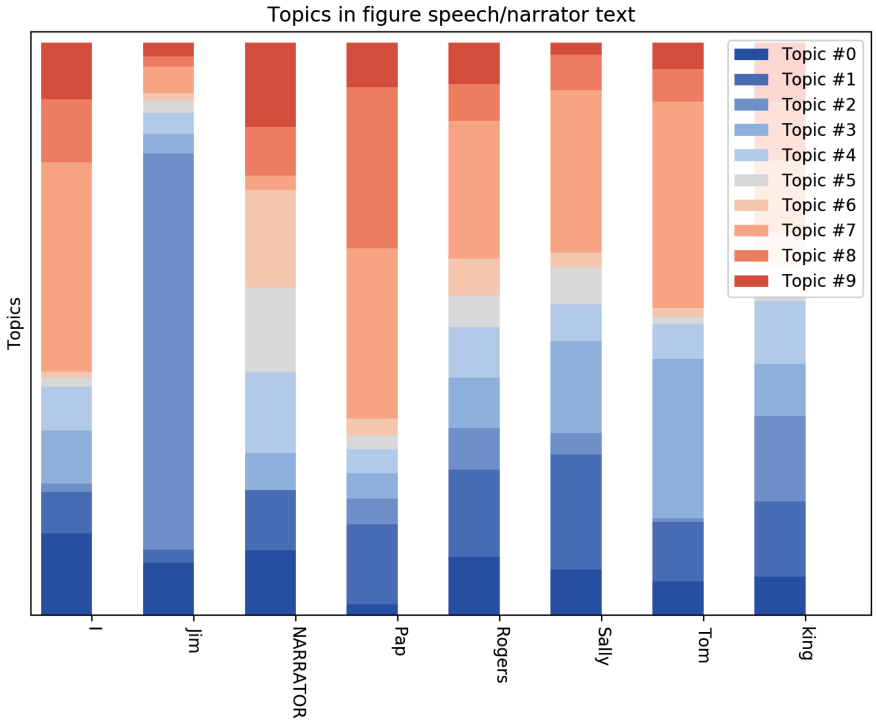

Abb. 4: Analyse der automatisch extrahierten Figurenrede in *Huckleberry Finn* mit latenten *Topics*

Die direkte Rede der übrigen Figuren (Huck [»I«], Tom, Sally, Rogers usw.) setzt sich vom Erzählertext recht deutlich durch *Topic* #7 ab (dominant sind weit verbreitete umgangssprachliche Elemente: »don't ain't ,'ll, won't«). Generell muss der Vergleich aufgrund der verhältnismäßig kleinen Sprachausschnitte und möglicher verbleibender Fehlzuweisungen mit Vorsicht genossen werden. Das Beispiel sollte vor allem die methodische Perspektive aufzeigen, die sich prinzipiell auf ganze Korpora skalieren ließe.

2 Formalisierte Textanalysemodelle und die Arbeitspraxis in den Literaturwissenschaften

Die Diskussion der computerlinguistischen Modellierungsansätze und die Analysebeispiele in Teil 1 zeigen, dass literarische Texte interessante Anknüpfungspunkte für die Anwendung von Modellen und Methoden aus der Computerlinguistik bieten. Gleichzeitig dürfte deutlich werden, dass eine unmodifizierte Anwendung von Standard-Ansätzen zwar gelegentlich möglich ist, aber meist den weitergehenden Fragestellungen nicht optimal gerecht wird – hierfür sollten die Analysemodelle angepasst und weiterentwickelt werden.

Die aus computerlinguistischer Sicht naheliegende Erwartung ist also, dass der Austausch mit Literaturwissenschaftlerinnen und -wissenschaftlern rasch dazu führt, die Grenzen der existierenden, oberflächennahen Analysemodelle systematisch zu erfassen und Wege für eine theoretisch fundierte Weiterentwicklung aufzuzeigen. Eine praktische Erwartung wäre, mit bestehenden Analysewerkzeugen bei der Exploration größerer Korpora von literarischen Texten einen Beitrag zum *Distant Reading* leisten zu können, mindestens zum Vorfiltern von Texten oder Textstellen für eine anschließende Detailanalyse durch *Close Reading*. Nach Pilotanalysen mit einigen denkbaren Analysemodellen könnte es im Dialog relativ zügig gelingen, Fragestellungen zum Text bzw. zum Korpus zu formulieren, für die eine Weiterentwicklung von bestehenden Modellierungsansätzen gleichzeitig computerlinguistisch realistisch und literaturwissenschaftlich zielführend ist. Anders formuliert liegt es aus technischer Sicht nahe, die Erfahrungen mit anderen Anwenderinnen von Sprachanalysekomponenten auf die computerunterstützte Analyse von literarischen Texten zu übertragen: so legen in der sogenannten Bio-NLP[25] biomedizinische Experten eine Begriffsontologie fest (beispielsweise Enzymbezeichnungen und relevante Prozesse, in denen die Enzyme eine Rolle spielen), annotieren in einem Korpus von Fachtexten textuelle Bezüge auf die Begriffe und schaffen so Referenzdatensätze für die Anpassung und Weiterentwicklung von computerlinguistischen Algorithmen und Modellen (mit der Informatik-Methode des sogenannten *Benchmarkings*, das Unterschiede im Modellverhalten dadurch systematisch erfasst, dass Vergleichsmodelle immer wieder auf die gleichen Testdaten angewandt werden). Nicht grundsätzlich anders funktioniert die Kooperation in der Korpuslinguistik, in der zu komplexeren

25 Kurz für *Biomedical Text Mining*, also *Natural Language Processing* für Textsammlungen der biomedizinischen Fachliteratur.

linguistischen Phänomenen (etwa auf Ebene der Semantik und Pragmatik) parallel zur Theorieentwicklung Sprachdaten in einem Referenzkorpus nach den theoretischen Kategorien annotiert werden, sodass Modelle für eine automatische Vorhersage empirisch evaluiert werden können. (Das letztere Szenario zielt häufig gar nicht auf die Entwicklung vollautomatischer Werkzeuge für reale Anwendungen ab, sondern nutzt experimentelle Vorhersagemodelle für die Überprüfung von theoretischen Hypothesen.)

Sucht man jedoch für den Analysegegenstand »literarische Texte« ganz praktisch nach einem möglichen Ausgangspunkt für entsprechende korpusbasierte Entwicklungsperspektiven, zeigt sich: abseits der noch kleinen Community der digitalen Literaturwissenschaften, auf die wir noch zurückkommen, ist eine Übertragung des Vorgehens nicht ohne weiteres möglich. Man müsste für eine nicht-triviale, jedoch auch nicht hochkomplexe Analyseaufgabe eine studienübergreifend nutzbare Operationalisierung entwickeln, anhand der ein Referenzkorpus annotiert wird. Mit diesem Korpus stünde dann der Computerlinguistik (bzw. der digitalen Literaturwissenschaft) ein Datensatz zur Verfügung, der eine empirisch kontrollierbare Anpassung, Weiterentwicklung und Optimierung von Modellen ermöglicht. Die jeweiligen Vorhersagemodelle könnten in der literaturwissenschaftlichen Forschung auf anderen Texten experimentell eingesetzt werden – möglicherweise bereits »produktiv« für explorative oder quantitative Studien, vor allen Dingen jedoch zur Hypothesenüberprüfung bei der Operationalisierung von Analysekategorien (und damit zur Theorieentwicklung). Die Gründe, weshalb ein derartiges Vorgehen (derzeit noch) weniger praktikabel ist als in anderen Analyseszenarien, sind vielfältig, und eine belastbare Beurteilung bedürfte einer umfassenden Meta-Reflexion. Als Beitrag zur Diskussion seien hier dennoch Annahmen und Vermutungen zu einigen wichtigen Gründen aufgelistet – im Bewusstsein der Einseitigkeit einer computerlinguistisch geprägten Betrachtung und ohne behaupten zu wollen, die Einschränkungen seien jeweils systematisch und unüberwindbar.

Einige Gründe liegen in der Unterschiedlichkeit der etablierten Arbeitspraktiken:

a) Der Originalitätsanspruch in literaturwissenschaftlichen Beiträgen läuft einer wiederholten Auseinandersetzung mit demselben Text und denselben Teilfragen entgegen (wie der *Benchmarking*-Ansatz es mit sich bringt). Selbst wenn theoretische Betrachtungen und die Methodenentwicklung im Vordergrund stehen, würde ein Beitrag in den Literaturwissenschaften zur exemplarischen Illustration wohl eher einen (in jüngerer Zeit) wenig untersuchten Text einsetzen als die propagierte Analysesystematik auf einen Referenztext anzuwenden, zu dem eine

Vielzahl von alternativen Ansätzen veröffentlicht ist (was genau dem gängigen computerlinguistischen Vorgehen entspräche).[26]

b) Aufgrund des verbreiteten Fokus auf einem vergleichsweise kleinen Kanon der Hochliteratur erschließt sich für viele Fachwissenschaftlerinnen nicht der Vorteil, den operationalisierte Analysemodelle für Kernfragen der Interpretation haben sollten: im günstigen Fall gelingt es, mit den Modellen bestimmte deskriptive Textanalysen analog zu dem Zugang eines professionellen Lesers zu erfassen. Wie kann dies aber zu einem Erkenntnisgewinn beitragen, den ein Spezialist nicht aufgrund seiner eigenen Lektüre mindestens ebenso gut erlangt hätte?

c) Soweit die Zielsetzung darin besteht, das Singuläre in den Werken der Hochliteratur zu erfassen, das zu ihrem epochenübergreifenden Stellenwert beiträgt, dürften über die deskriptive Analyse hinaus stets Aspekte der Deutung ins Spiel kommen, die von Fall zu Fall so spezifisch sind, dass eine generalisierende Behandlung unerreichbar erscheint.

d) Erweitert man den Gegenstandsbereich auf größere Korpora von literarischen Texten, möglicherweise unter Einschluss der populären Literatur, erweist sich eine exakt operationalisierte Charakterisierung von zentralen literaturhistorischen Beschreibungskategorien (wie bestimmten Gattungen oder Epochen) als schwierig, da in den Literaturwissenschaften zumeist kein streng empirischer Ansatz verfolgt wurde und diese Kategorien oft konzeptionell vage bleiben.

e) Die Tatsache, dass mit jedem automatisierten Analysewerkzeug eine Rest-Ungenauigkeit verbunden ist, lässt viele Literaturwissenschaftler vor dem Gedanken zurückschrecken, Interpretationen auf Werkzeugergebnisse aufzubauen – zumal ein effektiver Einsatz von Computermodellen auf Korpora sehr zeitintensiv ist und Kompetenzen erfordert, die traditionell nicht in einer geisteswissenschaftlichen Universitätsausbildung vertieft werden. Auf eine Analysekomponente, die nicht zweifelsfrei verlässlich, nicht in allen Details durchschaubar und deren Einsatz dazu noch mit großem Aufwand verbunden ist, mag manche oder mancher lieber verzichten.

f) Der konzeptionelle und zeitliche Aufwand, der mit Modellierungsexperimenten verbunden ist, hat in der Computerlinguistik und Informatik zu einer weitreichenden Spezialisierung in der Methodenentwicklung geführt. Experimente auf Korpusdaten werden wie in naturwissenschaftlichen Fachrichtungen

26 Christof Schöch widmet sich jedoch der Idee einer systematischen Wiederholung von Forschung in den digitalen Literaturwissenschaften, ders.: »Wiederholende Forschung in den digitalen Geisteswissenschaften«, DHd-Tagung 2017: *Digitale Nachhaltigkeit*, Bern.

mit stark teambasierten Laborpraktiken umgesetzt: in Arbeitsgruppen ist die Expertise zu wiederkehrenden Teilaufgaben häufig aufgeteilt; Arbeiten werden in Ko-Autorschaft veröffentlicht. Werden unterschiedliche Methoden zusammengeführt, geschieht dies oft durch eine Kooperation zwischen mehreren Arbeitsgruppen. Keiner der Beteiligten überschaut in einer solchen Situation jede Komponente des Gesamtmodells in jedem Detail; zur Absicherung eines methodischen validen Vorgehens müssen an den Schnittstellen Evaluierungen vorgenommen werden. Diese Praxis läuft der etablierten Arbeitsorganisation in den Geisteswissenschaften entgegen, nach der die Erwartung wäre, dass eine Wissenschaftlerin jede Methodik, die sie zur Anwendung bringt, eigenständig unter Kontrolle hat.

Neben diesen in der Arbeitspraxis verankerten Hürden werden immer wieder auch tiefer liegende Gründe verantwortlich gemacht:

f) Nicht selten werden grundsätzliche Bedenken vorgebracht, dass der Weg über die technisch machbaren Datenanalysen bei der Entwicklung einer These und ihrer Rechtfertigung die Unbefangenheit eines klassisch hermeneutischen Vorgehens gefährdet. Werden so nicht Fragen bevorzugt verfolgt, zu denen ein bestimmter methodischer Zugang naheliegende Antworten liefert?[27]

g) Allgemeiner weist beispielsweise das literaturwissenschaftlich-computerlinguistisch gemischte Autorenteam Adam Hammond, Julian Brooke und Graeme Hirst von der *University of Toronto* in dem Workshopbeitrag »A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together«[28] auf das sehr unterschiedliche Selbstverständnis in den beteiligten Fächerkulturen hin, wie beispielsweise C. P. Snow (1959)[29] die Situation recht drastisch charakterisiert hat mit seiner These zu den sich gegenseitig ignorierenden intellektuellen Kulturen der Geistes- und Literaturwissenschaften einerseits und der Naturwissenschaften und Technik andererseits. Hammond u. a. erkennen in der verbreiteten Skepsis unter Literaturwissenschaftlern gegenüber Computermodellen (und umgekehrt in der schwach ausgebildeten Fähigkeit unter Informatikerinnen, das literaturwissenschaftliche Vorgehen nachzuvollziehen) Auswirkungen des unterschiedlichen wissenschaftlichen Selbstverständnisses. Die Computer-

27 Reichert verweist beispielsweise im Vorwort zum Sammelband *Big Data* auf die Gefahr einer »evidenzbasierte[n] Konzentration auf das mit den Daten Mögliche«, vgl. Ramón Reichert (Hg.): *Big Data. Analysis on the digital transformation of knowledge, power and economy*. Bielefeld 2014.

28 Adam Hammond, Julian Brooke und Graeme Hirst: »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, 2013, S. 1–8.

29 Charles Percy Snow: *The Two Cultures and the Scientific Revolution*. Cambridge 1959.

linguistik arbeitet in der naturwissenschaftlichen Tradition grundsätzlich *problemorientiert*: die Forschungsagenda wird grundsätzlich so definiert, dass ein als problematisch erkannter Aspekt der vorhandenen Theorien, Modelle und Methoden überwunden wird. In der hermeneutischen Tradition liegt das übergeordnete Ziel einer Studie hingegen nicht in der Lösung eines bekannten Problems – ein bemerkenswerter Beitrag zeichnet sich vielmehr dadurch aus, dass er Fragen als relevant aufdeckt, die bislang nicht im Bewusstsein der Fachwelt lagen.

Aus der eigenen Erfahrung heraus stellen Hammond u. a. fest, dass für eine effektive Kooperation beide Seiten die »Komfortzone« ihrer disziplinären Gepflogenheit verlassen müssen – eine Beobachtung, die sicherlich die meisten Kooperationstandems unterschreiben werden.

h) Den Reibungspunkt der *Problemorientierung* konkretisieren Hammond u. a. (2013) anhand der Haltung der Disziplinen zur Ambiguität von Sprache und Texten. In der Tat steht die Ambiguitäts- bzw. Polyvalenzfrage häufig im Kern der Abstimmungsproblematik: die Computerlinguistik, in Nachfolge des klassisch linguistischen Vorgehens, sieht ihre Aufgabe darin, die allgegenwärtigen Phänomene der Ambiguität und Vagheit auf unterschiedlichsten Sprach- und Textebenen dahingehend aufzuklären, dass die Bedingungen einer kontextabhängigen Disambiguierung – soweit jeweils möglich – systematisch erfasst werden, und sie auf dieser Basis in einem algorithmischen Verfahren zu modellieren. Bei einem empirisch datenorientierten Vorgehen besteht ein naheliegender Schritt auf diesem Weg in der Erhebung des realen Disambiguierungsverhaltens von kompetenten Sprecherinnen der Sprache bzw. Leserinnen von Texten. Und genau so kann man den zentralen Schritt der Annotation von Korpusdaten verstehen: die Annotation einer Textstelle durch eine kompetente Sprecherin ist gleichsam ein empirisches Experiment zum komplexen kognitiven Interaktionsprozess von Wissensquellen (und weiteren Faktoren).

In der Literaturwissenschaft gilt die These der Polyvalenz von literarischen Texten als ein Grundkonsens über die unterschiedlichsten Strömungen hinweg: es gibt keine singuläre, »korrekte« Interpretation oder Deutung eines Texts. Gerade hochliterarische Texte zeichnen sich dadurch aus, dass sie in verschiedenen Rezeptionskontexten zu sehr unterschiedlichen Interpretationen einladen.[30] Vordergründig erscheint also der datenorientierte Ansatz der Computerlinguistik

30 Fotis Jannidis diskutiert den Status der Polyvalenzthese kritisch, die den Eindruck erwecken könnte, eine abwägende wissenschaftliche Auseinandersetzung mit konkurrierenden Interpretationsansätzen wäre unmöglich (was die Frage der Beliebigkeit aufwürfe). Er kommt zu dem Ergebnis, dass die These der Polyvalenz von Texten keinesfalls in Widerspruch zur Zielsetzung steht, unter den denkbaren Interpretationen diejenigen zu identifizieren, welche die »für einen

und Linguistik grundlegend inkompatibel: die Auszeichnung eines Texts im Sinne einer Referenzannotation (z. B. im Rahmen des *Benchmarkings*) erscheint als unzulässige Festlegung auf eine bestimmte Lesart.

In der disziplinübergreifenden Arbeitspraxis, die Hammond u. a. vorschlagen, versuchen sie die Spannung durch subjektive Mehrfachannotationen von literarischen Texten aufzulösen (zur Analysefrage der freien indirekten Rede, die sich im allgemeinen Fall nicht interpretationsunabhängig beantworten lässt). Computerlinguistische Modelle bzw. maschinelle Lernverfahren werden dann verwendet, um gerade die Indikatoren zu ermitteln, die Vieldeutigkeit erzeugen.

Auch das Hamburg-Heidelberger literaturwissenschaftliche Annotationsprojekt heureCLÉA[31] verwendet große Sorgfalt auf den Umgang mit der Frage der Polyvalenz. Die Guidelines[32] zielen darauf ab, sich bei der Annotation von narratologischen Kategorien so weit wie möglich auf deskriptive Analysen des Textinhalts zu beschränken, über die intersubjektive Übereinstimmung herrscht und für die keine interpretatorischen Schritte notwendig sind (so dass das Annotationsergebnis prinzipiell alle Interpretationsmöglichkeiten offen lassen sollte). In einem zyklischen Verfahren der Mehrfachannotation wird ein Konsens zu dieser intersubjektiven Basisanalyse hergestellt. Dieser Herangehensweise folgend könnte die Computerlinguistik bzw. Informatik ihre Aufgabe darin suchen, die intersubjektiven, prä-interpretatorischen Schritte in formalisierte Modelle zu implementieren und so mittelfristig auf größeren Korpora eine automatische quantitative Textinhaltsanalyse anzustreben, die hermeneutischen Studien als Grundlage dient (womit man einigen der oben genannten methodischen Bedenken zuvorkäme).

Möglicherweise ist es für eine gesunde Entwicklung des Methodeninventars auch förderlich, wenn zunächst »flachere« Verfahren der (deskriptiven) Textanalyse vorangetrieben werden – damit also neben distributionellen Ansätzen solche strukturelle Verfahren, die keine tieferen interpretatorischen Schritte beinhalten. Diese lassen sich robuster über Fragestellungen und Textspezifika

---

Leser durch die Lektüre eines Textes manifest gewordenen Informationen« besser als andere erfasse, ders.: »Polyvalenz – Konvention – Autonomie«, in: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*, hg. v. Fotis Jannidis, Gerhard Lauer, Matías Martínez und Simone Winko. Berlin, New York 2003, S. 3–30.

31 heureclea.de; eine Kooperation zwischen einer literaturwissenschaftlichen Arbeitsgruppe (unter Leitung von Jan Christoph Meister) und einer Informatikgruppe (geleitet von Michael Gertz).

32 Evelyn Gius und Janina Jacke: *Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets*. Version 2. Hamburg 2016.

hinweg übertragen und können so möglicherweise in größerer Breite neue Formen der Textbetrachtung inspirieren.

Dennoch soll hier betont werden, dass aus Sicht des Autors der Eindruck täuscht, ein »tiefer« (computer-)linguistischer Modellierungsansatz sei nicht mit der Polyvalenz-These vereinbar. Zwar ist es richtig, dass empirisch orientierte Zweige der modernen Linguistik und die Computerlinguistik wie oben ausgeführt bei der Korpusannotation anstreben, dass im jeweiligen Äußerungskontext die präferierte Lesart fixiert wird, um empirische Anhaltspunkte für die Interaktion von Informationsquellen zu erhalten. Entsprechend ist gern von »Gold-Standard-Annotationen« auf unterschiedlichen Analyseebenen die Rede, die beim Training von maschinellen Lernverfahren dafür sorgen, dass Disambiguierungsstrategien induziert werden können. Konzeptuell trennt die Linguistik jedoch sehr deutlich zwischen dem Teil der sprachlichen (und weitergehenden kognitiven) *Kompetenz* auf der einen Seite, die für eine gegebene Äußerung die Menge der prinzipiell möglichen Lesarten/Interpretationen erschließt, und andererseits jenen Mechanismen, die innerhalb dieser Menge die kontextuell plausibelste auswählen – unter Berücksichtigung des situativen und des Diskurskontexts und des Weltwissens usf. Eine formale Behandlung von Alternativen bei der Rezeption von literarischen Texten ist in diesem Rahmen ohne weiteres möglich.

Ein experimenteller Einsatz von Computermodellen zur Erfassung von interpretationsrelevanten Texteigenschaften bietet sich vor allem an, wenn nicht die Automatisierung der Textanalyse *per se* im Vordergrund steht (bei der mit zunehmender Analysetiefe der Grad der Verlässlichkeit in aller Regel abnimmt), sondern die Modelle für (differenzielle) Hypothesentests zu theoretischen Zusammenhängen verwendet werden. Beispielsweise könnte anhand von zwei Vergleichsmodellen, die auf unterschiedlichen Kanones für eine Zielkategorisierung trainiert werden (etwa eine Gattungs-Zuordnung), geklärt werden, welche Auswirkungen die (Nicht-) Berücksichtigung eines bestimmten Autorenwerks (das gemeinhin als einflussreich angesehen wird) für die simulierte Ausbildung der Gattungskonvention einer bestimmten Epoche hat.

Vorläufige Schlussfolgerung: »The Importance of Being Earnest« bei quantitativen Untersuchungen. Für ein abschließendes Urteil zum effektivsten Zusammenspiel zwischen Computerlinguistik und Literaturwissenschaft ist es zu früh. Sinnvoll ist sicherlich ein Weg der »zwei Geschwindigkeiten« bzw. zwei Komplexitätsstufen bei der Computermodellierung. Werkzeuge, die ohne massiven Anpassungsaufwand und mit realistischer Einarbeitungszeit in die Methodik auf neue Texte angewandt werden können, tragen stark zur Erschließung von Pfaden für eine »digital« informierte Textbetrachtung bei. Gleichzeitig bietet es sich an, die Möglichkeiten einer aufwändigen und strukturell anspruchsvollen

Modellierung von Textanalyse so zu erweitern, dass sie sich in literaturtheoretische Paradigmen einpassen lassen bzw. die konzeptuellen Lücken der etablierten computerlinguistischen Ansätze schließen.

Dank einer aktiven deutschsprachigen Community der digitalen Literaturwissenschaft und eines großen Kooperationsinteresses in Teilen der Computerlinguistik-Community sind Entwicklungen auf beiden Pfaden im Fluss.

Ganz besonders hervorzuheben ist, dass die bekannteren Aktivitäten in diesem Spannungsfeld mit einem hohen Anspruch der methodischen Reflexion umgesetzt werden[33] – gerade auch die Ansätze, die mit leicht übertragbaren, generischen Werkzeugen operieren. Methodische Reflexion ist stets von Bedeutung – und im nicht unumstrittenen Experimentierfeld der digitalen bzw. quantitativen Literaturwissenschaft sicherlich von besonders großer: Werkzeuge und Untersuchungsmethoden, die sich rein technisch von einem Textkorpus auf ein anderes übertragen lassen, sind noch keine Gewährleistung, dass eine Fragestellung aus dem ursprünglichen Kontext auch im Zielkontext sinnvoll zu beantworten ist. Da die Standard-Modelle in der Regel für zeitgenössische Nachrichtentexte entwickelt wurden, sind Qualitätsverluste bei der Analyse literarischer Texte nicht ungewöhnlich.

Besonders »tückisch« ist die Tatsache, dass die meisten Werkzeuge robust auf Abweichungen reagieren, so dass Probleme sich mitunter gar nicht in unmittelbar sichtbaren Fehlanalysen niederschlagen. Gerade bei komplexeren quantitativen Analysen können vermeintliche Bagatellprobleme zu falschen Schlussfolgerungen führen. Ein plastisches Beispiel zu sprachtechnologisch unterstützter Webanalyse führt David Jurgens an:[34] Die Erkennung, in welcher Sprache eine Kurzmitteilung verfasst ist, gilt als sprachtechnologisch gelöstes Problem. Entsprechend werden beispielsweise nach Sprache automatisch gefilterte Twitter-Nachrichten für demographische Untersuchungen ausgewertet. Es zeigt sich

33 Vgl. etwa Peer Trilcke: »Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft«, in: *Empirie in der Literaturwissenschaft*, hg. v. Philip Ajouri, Katja Mellmann und Christoph Rauen. Münster 2013, S. 201–247; Nils Reiter, Anette Frank und Oliver Hellwig: »An NLP-based Cross-Document Approach to Narrative Structure Discovery«, in: *Literary and Linguistic Computing* 29.4 (2014), S. 583–605; Jannidis u. a.: *Automatische Erkennung von Figuren in deutschsprachigen Romanen*; Thomas Bögel, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris und Jannik Strötgen: »Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative«, in: *DHCommons journal* 2015; Gius und Jacke: *Zur Annotation narratologischer Kategorien der Zeit*; Schöch: *Wiederholende Forschung in den digitalen Geisteswissenschaften*.

34 Postdoctoral Scholar, Stanford University; Workshop-Vortrag Universität Stuttgart, Oktober 2016.

allerdings, dass ein erheblicher Anteil von Kurznachrichten in afroamerikanischem Englisch nicht der Kategorie *Englisch* zugeordnet wird (was in den Standard-Testszenarien jedoch nicht ins Gewicht fällt). Nutzt man eine Analyse »aller« englischsprachigen Kurznachrichten beispielsweise für die Wahlforschung, kann es so zu systematisch falschen Vorhersagen kommen.

Eine quantitative Datenanalyse wird im Rahmen einer literaturwissenschaftlichen Studie die Erwartung einer objektiven Ergänzung der anderweitigen Argumentation wecken. Umso sorgfältiger sollte die Validierung der Methoden durchgeführt werden. Der Anschein pseudo-objektiver Analysen würde den gesamten Ansatz der *Digital Humanities* diskreditieren. So mag es noch wichtiger sein als die Entwicklung von anspruchsvollen Modellen, dass für jede Methode vor einer Verwendung die Adäquatheit in Bezug auf die Zieldaten überprüft wird und die zu erwartende Qualität der Ergebnisse abgeschätzt wird. Hierfür genügt es zumeist nicht, die Werkzeugausgabe auf einigen Eingabedaten in Augenschein zu nehmen (da Fehler so leicht übersehen werden), sondern es sollte zumindest eine kleine Sammlung von unabhängig annotierten Testdaten als *Referenzkorpus* eingesetzt werden, dessen relevante Eigenschaften repräsentativ für die tatsächlichen Zieldaten sind.[35] In aller Regel lohnt es sich sehr, einige Stunden in die werkzeugunabhängige Annotation von Testdaten zu investieren und das Werkzeug gegen diese zu evaluieren und eventuelle Parameter zu kalibrieren. Selbst wenn eine automatische Analyse nur explorativ eingesetzt wird, etwa im Rahmen eines *Distant Reading*, kann eine Fehleinschätzung zur Verlässlichkeit zu sehr irreführenden Schlussfolgerungen führen, die gerade aufgrund der Distanz zum Text auch nicht in der weiteren Betrachtung zutage treten.

Für größer angelegte Analyseaufgaben stehen zumeist alternative Modellierungsverfahren bzw. Werkzeugkombinationen zur Auswahl. Der Abgleich der erzielbaren Ergebnisse mit einer Testmenge von Referenzdaten kann entscheidend zu einem effektiven Werkzeugeinsatz beitragen. In der Kooperation zwischen Fachwissenschaftlern und Informatikerinnen machen Referenzdaten eine gezielte Modelloptimierung möglich.

In der Regel entwickeln sich in hermeneutisch geprägten Projekten die analytischen Fragestellungen erst im Zuge der Auseinandersetzung mit dem Untersuchungsgegenstand, was eine verzahnte Verfeinerung der technischen Analyse-(teil-)ziele erforderlich macht. Auch hierbei scheint die entwicklungsbegleitende Annotation von möglichst repräsentativen Referenzkorpusdaten mit

35 Um argumentative Schlussfolgerungen aus den Beispielanalysen in Teil 1 dieses Beitrags zu ziehen, wäre beispielsweise zwingend eine Validierung der automatischen Extraktionsergebnisse erforderlich.

den (jeweils vorläufigen) Zielkategorien der beste Garant für ein effektives und dennoch methodenkritisches Vorgehen.[36]

Indem auch beim hermeneutisch geprägten Vorgehen der Blick auf die Texteigenschaften gelenkt wird, die besonders starken Einfluss auf bestimmte Kategorien in der deskriptiven Analyse haben, dürften einige der oben aufgeführten Unterschiede in der Arbeitspraxis schwinden und so erscheint es perspektivisch denkbar, dass durch die empirische Verankerung in geeigneten Referenzdaten computerlinguistische Modellierungsansätze gezielt so weiterentwickelt werden, dass auch zu komplexen literaturtheoretischen oder -historischen Hypothesen differenzielle datengestützte Experimente durchgeführt werden können, die sich *idealiter* ergänzend in eine hermeneutische Argumentation einfügen lassen.

Bibliographie

Bamman, David, Ted Underwood und Noah A. Smith: »A Bayesian Mixed Effects Model of Literary Character«, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore 2014, S. 370–379.

Björkelund, Anders und Jonas Kuhn: »Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features«, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore 2014, S. 47–57.

Blei, David M, Andrew Y. Ng und Michael I. Jordan: »Latent dirichlet allocation«, in: *Journal of machine Learning research* 3 (2003), S. 993–1022.

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris und Jannik Strötgen: »Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative«, in: *DHCommons journal* 2015.

Brett, Megan R.: »Topic Modeling: A Basic Introduction«, in: *Journal of Digital Humanities* 2012, S. 12–17.

Brooke, Julian, Adam Hammond und Graeme Hirst: »Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction«, in: *Digital Scholarship in the Humanities*, 2016.

36 Wir diskutieren ein entsprechendes Vorgehensmodell in größerem Detail anhand eines Textanalyseportals zu biographischen Texten in: Jonas Kuhn und André Blessing: »Die Exploration biographischer Textsammlungen mit computerlinguistischen Werkzeugen – methodische Überlegungen zur Übertragung komplexer Analyseketten in den Digital Humanities«, in: *Europa baut auf Biographien*. Wien 2018. Das Stuttgarter *Digital Humanties*-Methodenzentrum CRETA (*Center for Reflected Text Analytics*), das seit 2016 vom BMBF gefördert wird, ist als Versuch angelegt, Arbeitspraktiken zu erkunden, die textanalytische Fragestellungen disziplinübergreifend entlang eines referenzdatenorientierten Entwicklungszyklus angehen.

Burrows, John: »›Delta‹: A Measure of Stylistic Difference and a Guide to Likely Autorship«, in: *Literary and Linguistic Computing* 17 (2002), S. 267–287.

Carkeet, David: »The Dialects in Huckleberry Finn«, in: *American Literature* 51.3 (1979), S. 315–332.

Cohen Minnick, Lisa: »Jim's language and the issue of race in Huckleberry Finn«, in: *Language and Literature* 10.2 (2001), S. 111–128.

Elson, David K. und Kathleen R. McKeown: »Automatic attribution of quoted speech in literary narrative«, in: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (AAAI '10) 2010. AAAI Press, S. 1013–1019.

Elson, David K., Nicholas Dames und Kathleen R. McKeown: »Extracting social networks from literary fiction«, in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10. Stroudsburg, PA, USA, 2010 (Association for Computational Linguistics), S. 138–147.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch und Thorsten Vitt: »Towards a better understanding of Burrows's Delta in literary authorship attribution«, in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver 2015, S. 79–88.

Fabre, Cécile und Alessandro Lenci: *TAL Journal: Special issue on Distributional Semantics (Traitement Automatique des Langues / Natural Language Processing)* 56.2 (2015).

Gärtner, Markus, Anders Björkelund, Gregor Thiele, Wolfgang Seeker und Jonas Kuhn: »Visualization, Search, and Error Analysis for Coreference Annotations«, in: P*roceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2014.

Gärtner, Markus, Katrin Schweitzer, Kerstin Eckart und Jonas Kuhn: »Multi-modal Visualization and Search for Text and Prosody Annotations«, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Peking 2015, http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/icarus.html (31. Juli 2017).

Gius, Evelyn und Janina Jacke: *Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets*. Version 2. Hamburg 2016.

Hammond, Adam, Julian Brooke und Graeme Hirst: »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, 2013, S. 1–8.

Jannidis, Fotis und Gerhard Lauer: »Burrows's Delta and Its Use in German Literary History«, in: *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, hg. v. Matt Erlin und Lynne Tatlock. Rochester 2014, S. 29–54.

Jannidis, Fotis, Gerhard Lauer, Matías Martínez und Simone Winko: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Berlin, New York 2003.

Jannidis, Fotis, Markus Krug, Isabella Reger, Martin Toepfer, Lukas Weimer und Frank Puppe: *Automatische Erkennung von Figuren in deutschsprachigen Romanen*. Digital Humanities im deutschsprachigen Raum (DHd) 2015, Graz, https://opus.bibliothek.uni-wuerzburg.de/files/14333/Jannidis_Figurenerkennung_Roman.pdf (31. Juli 2017).

Jannidis, Fotis: »Polyvalenz – Konvention – Autonomie«, in: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*, hg. v. ders., Gerhard Lauer, Matías Martínez und Simone Winko. Berlin, New York 2003, S. 3–30.

Kao, Justine und Daniel Jurafsky: »A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry«, in: *Proceedings of the Workshop on Computational Linguistics for*

Literature (Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT), Montréal 2012, S. 8–17.

Kuhn, Jonas und André Blessing: »Die Exploration biographischer Textsammlungen mit computerlinguistischen Werkzeugen – methodische Überlegungen zur Übertragung komplexer Analyseketten in den Digital Humanities«, in: *Europa baut auf Biographien*. Wien 2018.

Landauer, Thomas K., Peter Foltz und Darrell Laham: »Introduction to Latent Semantic Analysis«, in: *Discourse Processes* 25 (1998), S. 259–284.

Lemke, Sieglinde: *The Vernacular Matters of American Literature*. New York 2009.

Levy, Omer und Yoav Goldberg: »Linguistic Regularities in Sparse and Explicit Word Representations«, in: *Proceedings of the Eighteenth Conference on Computational Language Learning*, 2014, S. 171–180.

McDonald, Ryan T., Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall: »Universal Dependency Annotation for Multilingual Parsing«, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), Sofia 2013, S. 92–97.

Mikolov, Tomas, Greg Corrado, Kai Chen, und Jeffrey Dean: *Efficient estimation of word representations in vector space.* 2013. arXiv preprint. arXiv preprint arXiv:1301.3781.

Reichert, Ramón (Hg.): *Big Data. Analysis on the digital transformation of knowledge, power and economy*. Bielefeld 2014.

Reiter, Nils, Anette Frank und Oliver Hellwig: »An NLP-based Cross-Document Approach to Narrative Structure Discovery«, in: *Literary and Linguistic Computing* 29.4 (2014), S. 583–605.

Rösiger, Ina und Jonas Kuhn: »IMS HotCoref DE: A Data-Driven Co-Reference Resolver for German«, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož 2016, S. 155–160.

Schöch, Christof: »Wiederholende Forschung in den digitalen Geisteswissenschaften«, DHd-Tagung 2017: *Digitale Nachhaltigkeit*, Bern.

Snow, Charles Percy: *The Two Cultures and the Scientific Revolution*. Cambridge 1959.

Templeton, Clay: *Topic Modeling in the Humanities: An Overview. Maryland Institute for Technology in the Humanities*, 2011. http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/ (28. April 2017).

Trilcke, Peer: »Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft«, in: *Empirie in der Literaturwissenschaft*, hg. v. Philip Ajouri, Katja Mellmann und Christoph Rauen. Münster 2013, S. 201–247.

Vala, Hardik, David Jurgens, Andrew Piper und Derek Ruths: »Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts«, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, hg. v. Association for Computational Linguistics. Lisabon September 2015.

Zitierte URLs

Tutorium zum Mallet-Toolkit. https://de.dariah.eu/tatom/topic_model_mallet.html (31. Juli 2017).