Initiative zur Einrichtung eines Schwerpunktprogramms

Computational Literary Studies

Programmausschuss

Prof. Dr. Fotis Jannidis,
Julius-Maximilians-Universität Würzburg, Professor, Computerphilologie (*Koordinator*)

Dr. Evelyn Gius,
Universität Hamburg, wissenschaftliche Mitarbeiterin, Literaturwissenschaft

Prof. Dr. Jonas Kuhn,
Universität Stuttgart, Professor, Informatik/Computerlinguistik

Prof. Dr. Christof Schöch,
Universität Trier, Professor, Digital Humanities

Prof. Dr. Simone Winko,
Georg-August-Universität Göttingen, Professorin, Literaturwissenschaft

---

1 Zusammenfassung

Datenzentrierte Forschungsansätze haben in den letzten 10-20 Jahren viele Forschungsfelder stark geprägt. Diese Entwicklung beeinflusst auch, angetrieben durch die Digitalisierung großer Textsammlungen in den letzten 5-10 Jahren, die Praxis und Methodik der Literaturwissenschaften. Die Anwendung von Algorithmen auf Sammlungen digitaler literarischer Texte ermöglicht neue Erkenntnisse über deren Strukturen und Entwicklungen. Ziel des Schwerpunktprogramms ist es, systematisch das Verständnis zu vertiefen, wie formale Modelle in der literaturwissenschaftlichen Forschung genutzt werden können. Folgende Forschungsschwerpunkte stehen im Fokus:

- Ermittlung, welche Verfahren der Informatik und Computerlinguistik für die Analyse literarischer Texte relevant sind,
- Anwendung bereits existierender Algorithmen auf neue Datensätze zur Generierung neuer Erkenntnisse über kulturelle Phänomene, Veränderungen und Strukturen,
- Forschung zu bereits bestehenden Algorithmen, um diese zu erweitern, Wege zur Anpassung von Parametern zu finden sowie das Verständnis ihrer Interaktion mit literarischen Texten zu verbessern,
- Formale Modellierung literaturwissenschaftlich relevanter Konzepte, wobei die Tiefe der formalen Modellierung literarischer Phänomene skalierbar ist,
- Integration von Ergebnissen quantitativ-empirischer Forschung in den qualitativ-hermeneutischen Forschungsprozess und die Theorie- und Begriffsbildung.

Im Zentrum des SPP *Computational Literary Studies* sollen literarische Texte in einem weiten Sinn stehen. Sie sind bislang erheblich seltener quantitativ erforscht worden als nicht-literarische Texte und stellen für quantitative Verfahren eine Herausforderung dar, da sie besonders komplex sind, z.B. wegen ihrer Fiktionalität, der kreativen, ästhetisch motivierten Verwendung von Sprache und/oder des spezifischen Charakters literarischer Zeichen. Mit literarischen Texten wird häufig nicht explizit, sondern indirekt durch Geschichten und Bilder kommuniziert, und sie neigen zur Individualisierung und Hybridisierung von Stil, Themen und Formen. Aus den Eigenheiten des Untersuchungsgegenstands und dem Forschungsstand des emergenten Feldes ergeben sich eine Reihe von Problemfeldern: die Notwendigkeit der Domänenadaption von Werkzeugen zur Textanalyse; das Verhältnis von Norm und Abweichung; die Bestimmung von Gemeinsamkeiten und Unterschieden zwischen literarischen und nicht-literarischen Texten; Schwierigkeiten, einen Konsens bei der Annotation literaturwissenschaftlich relevanter Texteigenschaften zu finden; der Mangel an interpretatorischer Transparenz moderner formaler Modelle; die Notwendigkeit des Aufbaus von Referenzkorpora. Dieses Forschungsprogramm lässt sich nur durch eine eng vernetzte interdisziplinäre und damit auch risikoreiche

Forschung bewältigen, an der Forschende aus der Literaturwissenschaft, den Digital Humanities[1], der Korpuslinguistik, der Computerlinguistik und der Informatik beteiligt sind. Ziel des Schwerpunktprogramms ist es, durch die systematische Bearbeitung von Grundlagenfragen und die eng koordinierte Entwicklung von *best practices* zur Konsolidierung dieses wachsenden, internationalen Forschungsfeldes beizutragen, um auf interdisziplinär vernetzte Weise das Methodenspektrum zur Analyse literarischer Texte nachhaltig um innovative, dem Gegenstand angemessene und neue Einsichten versprechende Verfahren zu erweitern.

2 Stand der Forschung und eigene Vorarbeiten

2.1 Stand der Forschung

Die Unterscheidung von **Gattungen bzw. Textsorten** aufgrund sprachlicher Merkmale ist schon seit längerem Gegenstand von Arbeiten in der Korpuslinguistik (Biber, Conrad, Reppen 1998) und die Unterscheidung von Großformen (z.B. Komödie/Tragödie) ist recht zuverlässig. Mit automatischen Verfahren lassen sich etwa aus sehr großen Sammlungen prosaischer Texte mit über 90-prozentiger Wahrscheinlichkeit erzählende Werke extrahieren (Underwood 2015) oder fiktionale von nicht-fiktionalen Erzähltexte unterscheiden (Piper 2016). Die Unterscheidung von Untergattungen innerhalb dieser Großformen kann sehr schwierig sein (z.B. Entwicklungs- vs. Gesellschaftsromane: Hettinger et al. 2016). Auf der Basis einer breiten Auswertung von Forschungsliteratur formuliert Moretti (2005) die These, dass die Lebensspanne von (englischen) Romangattungen generationsbedingt ist. Jockers (2013, 87) kann diese These mit automatischen Verfahren zumindest für eine Reihe von kleineren Romangattungen bestätigen, allerdings hat der Autorstil einen großen Anteil daran (Jockers 2013, 81; Allison et al. 2011).

Während sich Tragödie und Komödie mit topic modeling zuverlässig unterscheiden lassen, ist das Verhältnis der Tragikomödie zu den Großformen äußerst komplex (Schöch 2017a, angenommen). Mit einem Verfahren zur Schlüsselwort-Extraktion klassifizieren Willand & Reiter (2017) die Dramen von Kleist (die der Autor selbst teilweise nur “Schauspiel” nennt) in die beiden Gattungen. Die automatische Klassifikation von literarischen Untergattungen stellt auch deshalb eine Herausforderung dar, weil sehr unterschiedliche Merkmale diese Untergattungen definieren und diese Merkmale zudem auch unterschiedlich salient sind. Zuordnungen von Einzeltexten zu Gattungen sind auch immer wieder Gegenstand literaturwissenschaftlicher Debatten, so dass ein Goldstandard schwer herzustellen ist.

Die **Stilanalyse** kann auf eine lange Geschichte zurückblicken (Grzybek 2014). Eine wichtige Einsicht der Stilometrie besteht darin, dass es keinen stilistischen ‚Fingerabdruck‘ gibt, der eine eindeutige Zuordnung von Texten erlaubt – sei es zu Autoren, Gattungen, Epochen oder Figuren. Allerdings lassen sich innerhalb eines Korpus aufgrund geeigneter Merkmale Distanzen zwischen Texten berechnen, die belastbare Aussagen über Autorschaft und andere Gruppenzugehörigkeiten erlauben. In den letzten Jahren sind zahlreiche Studien entstanden, die verschiedene Textmerkmale (z.B. Satzlänge, Interpunktionsdichte, Wortlänge, Satzstruktur, Häufigkeiten von Wörtern) als Stilmerkmale heranziehen und insbesondere für die Autorschaftszuschreibung verwenden – da die Autoren der meisten Texte bekannt sind, erlaubt dies eine einfache Evaluation der Verfahren (Love 2002). Burrows (2002) beschrieb auf der Grundlage ausschließlich der häufigsten Wörter ein einfaches Maß, Delta, das seitdem erfolgreich für viele verschiedene Gattungen und Sprachen getestet (Craig und Kinney 2009; Eder und Rybicki 2013) und weiterentwickelt wurde (Evert et al. 2017). Auch jenseits der Autorschaftszuschreibung wurden stilometrische Verfahren erfolgreich eingesetzt, da sich darüber neben dem Autorsignal Gattungs-, Epochen-, und auch Figurensignale zeigen lassen (Jockers 2013, Kap. 6; Jannidis und Lauer 2014; Hoover, 2017). Dies gilt allerdings auch für

[1] Die CLS sind ein Teilbereich der Digital Humanities, allerdings umfassen diese viele weitere Fächer, z.B. Musikwissenschaft oder Archäologie, und viele weitere Arbeitsgebiete, z.B. die Erstellung digitaler Editionen oder *Music Information Retrieval*.

inhaltliche Aspekte, wie Schöch (2017b) am Beispiel von französischen Kriminalromanen zeigen kann.

Stilometrische bzw. korpuslinguistische Verfahren konnten diachron eingesetzt auch sprachliche Veränderungen über längere Zeiträume hinweg aufdecken: Biber und Finegan (1989) zeigen für literarische Texte eine generelle Entwicklung hin zu einem direkteren, mündlichen Stil vom 17. bis zum 20. Jahrhundert, Heuser und Le-Khac (2012) zeigen den im 19. Jahrhundert zunehmenden Gebrauch von Konkreta, Underwood und Sellers (2012) beschreiben eine Zunahme der Beschreibung mündlicher Kommunikation, Egbert (2012) beschreibt ein mehrdimensionales Kategoriensystem für literarische Texte. Einen Befund aus der Dickens-Forschung zu unterbrochener Redewiedergabe überprüfen Mahlberg, Smith, & Preston (2013) durch Korrelationsanalyse zwischen repräsentierter Körper- und Verbalsprache.

Auch die **Charakterisierung von Figuren** in Erzähltexten sind mit manuellen und automatischen Verfahren untersucht worden; die Erfolge waren bei bestimmten Formen populärer Literatur deutlich besser, wohl weil diese schematischer sind (Koolen und Cranenburgh 2017). Versuche, aus sehr großen Textsammlungen Figurentypen zu ermitteln, können erste Ergebnisse vorweisen, lassen sich aber aufgrund der fehlenden Referenzmaßstäbe bislang nur schlecht evaluieren (Bamman et al. 2013, 2014). Untersuchungen zu genderabhängigen Figurenbeschreibungen machen bestehende Verfahren für die Figurenanalyse fruchtbar (Underwood und Bamman 2016; Jockers und Kiriloff 2016). Die **Rede-, Schreib- und Gedankenwiedergabe** nimmt in vielen Erzähltexten einen breiten Raum ein, ihre automatische Erkennung und Verarbeitung ist daher ein wichtiger Baustein zur Erzähltextanalyse, wenn z.B. die Kommunikationsinteraktion von Figuren analysiert werden soll. Semino und Short (2004) haben ein englisch- und Brunner (2015) ein deutschsprachiges Korpus annotiert. Die einfache (und maschinenlesbar markierte) Attribution der Figurenrede in Dramen erlaubt weitergehende und robustere Analysen, auch im Hinblick auf Kategorien von Figuren: Bullard & Ovesdotter Alm (2014) haben anhand eines kleinen skandinavischen Korpus untersucht, welche linguistischen Mittel Autoren verwenden, um Geschlecht, Alter und sozialen Status der Figuren zu markieren. Unterschiede zwischen männlichen und weiblichen Figuren zeigen sich auch in Dramen von Kleist im Bezug auf deren thematische Ausrichtung (Willand & Reiter 2017). Das korpuslinguistische Projekt “Encyclopaedia of Shakespeare's Language”[2] untersucht seit 2016 Shakespeares Dramen, u.a. zur Erkennung von *character profiles*. Generell befindet sich das Feld im Bereich der Charakterisierung von Figuren noch am Anfang: Ein textanalytisches Desideratum sind Ansätze zur umfassenden Extraktion und zielgerichteten Zusammenführung beschreibender Informationen aus Texten. Etablierte Referenzmaßstäbe würden die Entwicklung von Werkzeugen befördern.

Relationen werden auf Basis der Kommunikation der Figuren untereinander (Moretti 2011; Elson et al. 2010), der Kookkurrenz von Figurenreferenzen im Text (Park et al. 2013; Ardanuy und Sporleder 2014; Blessing et al., 2017) oder der Themen der Gespräche (Celikyilmaz et al. 2010) modelliert. Solche Daten lassen sich gut mit den Mitteln der sozialen **Netzwerkanalyse** auswerten (Trilcke 2013) und erlauben eine Modellierung von Konzepten wie Haupt- und Nebenfigur, ermöglichen den Überblick über historisch dominante Figurenkonstellationen oder lassen sich als Ähnlichkeitsdimension von Texten auffassen. Aus dramatischen Texten lassen sich Kopräsenznetzwerke vergleichsweise leicht in großem Stil extrahieren (Trilcke et al., 2015). Trilcke et al. (2016) verwenden Konzepte aus der Netzwerktheorie (small worlds) zur Gruppierung von Dramen. An diese Netzwerke schließt die Analyse spezifischer Figurenrelationen an, z.B. von Familien- (Makazhanov et al. 2014) oder Liebesbeziehungen (Karsdorp et al. 2015). Nalisnick & Baird (2013) untersuchen zudem, ob sich mit Verfahren der sentiment analysis die Wendepunkte in der Beziehung zwischen Figuren automatisch identifizieren lassen. Reichert (1965) beschreibt ein formales Verfahren zur Untersuchung der sozialen Struktur und Durchmischung innerhalb des Personals in Dramen. Für Erzähltexte und

[2] http://wp.lancs.ac.uk/shakespearelang/

Dramen ergeben sich Probleme der Interpretation von Netzwerken im Bezug auf literaturwissenschaftliche Fragestellungen. Kopräsenz oder -okkurrenz kann dabei zwar als nützliche Heuristik gelten, wird aber in vielen Fällen der Komplexität von Figureninteraktion nicht gerecht. An der vielschichtigen „Semantisierung" von Figurennetzwerken und -relationen besteht daher noch Forschungsbedarf.

Eine **Strukturierung** von Erzähltexten in z.B. Handlungsabschnitte könnte zunächst die automatische Verarbeitung solcher Texte erleichtern (etwa im Bereich der Koreferenzresolution). Narratologisch motivierte Einteilungen wurden im Projekt heureCLÉA konzeptualisiert und annotiert (Bögel et al., 2015). Die kollaborative Entwicklung von Annotationsrichtlinien für das Phänomen der Erzählebenen ist Thema eines derzeit laufenden shared tasks (Reiter et al., 2017), womit sich auch Wallace (2012) beschäftigt hat. Eine **Segmentierung** nach hierarchisch organisierten Themen wird von Kazantseva und Szpakowicz (2014) vorgeschlagen; Reiter (2015) beschreibt Experimente zur manuellen Segmentierung von Erzähltexten, die nahelegen, Segmentierung und Zusammenfassung zusammen zu denken.

Die Modellierung von **Handlung** ist komplex, da das narratologische Konzept eines Plots als Folge von Ereignissen aufgrund des problematischen Ereignisbegriffs (Meister 2003; Dunn und Schumacher 2016) nur schwer operationalisierbar ist. Eine Reihe von Vorschlägen setzt daher bei den Figuren an: Chambers und Jurafsky (2009) schlagen vor, narrative Ereignisketten – basierend auf Verben und Koreferenzketten – und die zugehörigen Rollen unüberwacht, also ohne Rückgriff auf annotierte Trainingsdaten, zu lernen. Ein anderer Vorschlag lautet, den Verlauf der im Zusammenhang mit Figuren verwendeten Emotionswörter und die mit den Figuren besonders häufig verbundenen Wörter als Indikatoren für den Plot zu nehmen und auf diese Weise Handlungsähnlichkeiten zwischen Romanen zu erfassen (Elsner 2012; Jockers 2015; Bilenko und Miyakawa 2013). Geschichten mit vergleichbaren Plots können auf ihre Ähnlichkeiten hin untersucht werden, um die ähnlichen Passagen zu alignieren (Finlayson 2009; Reiter et al. 2014). Bei einem Korpus von nicht-fiktionalen Erzählungen haben sich hierarchische Graphen als nützlich erwiesen, wobei stabile Fünf-Wort-Verbindungen als Knoten dienten, um Handlungsabschnitte und deren sprachliche Realisierung sichtbar zu machen (Bubenhofer et al. 2013). Für dramatische Texte schlagen Trilcke et al. (2017) eine Analyse des Personalaustausches über Szenengrenzen hinweg vor, um sich dem Handlungsverlauf zumindest anzunähern.

Ein Problem vieler dieser Ansätze ist der Mangel an klaren Definitionen und Operationalisierungen der einschlägigen literaturwissenschaftlichen Begriffe, was nicht zuletzt eine vergleichende Evaluation der Verfahren erschwert. Neben der Exploration automatischer Verfahren im Bezug auf die Analyse von Handlungen ist daher die Weiterentwicklung theoretischer Konzepte und deren praktischer Anwendung (etwa durch Annotationen) ein Desideratum des Schwerpunktprogramms. Eine vielversprechende Strategie ist hierbei, vielschichtige literaturwissenschaftliche Konzepte wie 'Handlung' oder 'Gattung' zu dekomponieren, d.h. in konzeptuelle Teilaspekte zu zerlegen (bei der Handlung u.a. Figuren, Orte, Zeiten, Ereignisse, Sentiment; bei der Gattung u.a. Themen, Textstruktur, Handlungsort, Figureninventar). Die Teilaspekte können dann zunächst separat für Analysen und Annotationen operationalisiert werden, bevor sie gegebenenfalls synthetisiert werden.

Die automatische Erkennung und Analyse **metrischer Strukturen** für lyrische Texte verschiedener Sprachen ist ein gut entwickelter Forschungsbereich, in dem Arbeiten zum Deutschen (Bobenhausen 2009), Mittelhochdeutschen (Estes und Hench, 2016), Russischen (Birnbaum & Thorsten 2015), Spanischen (Navarro-Colorado, Lafoz, Sanchez 2016), Französischen (Beaudoin und Yvon 1996), Englischen (Agirrezabal et al. 2013, Greene, 2010, Hammond 2014) sowie Griechischen und Lateinischen (Fusi 2009) vorgelegt wurden, die teils für spezifische Untergattungen, teils für die Verslyrik einer Sprache insgesamt funktionieren. Darüber hinaus wird die Visualisierung metrischer Strukturen (Chaturvedi et al. 2011) oder klanglicher Muster (McCurdy et al. 2016) bzw. von Lyriksammlungen (Bobenhausen Metricalizer 3.0) untersucht. Wünschenswert wäre es darüber hinaus zum einen, die bisherigen Ergebnisse so weiterzuentwickeln, dass auch unregelmäßige Fälle und Strophenformen zuverlässig erkannt

werden können. Wenn etwa freie Rhythmen zuverlässiger erkannt werden könnten, wäre von hier aus ein vielversprechender Weg zum Erkennen von Prosarhythmus möglich.

Sowohl für Prosa- als auch für Dramen-Texte wurden verschiedene Formen der **formalen Modellierung** vorgeschlagen, die zwar zumeist nicht direkt mit einer literaturwissenschaftlichen Forschungsfrage im Zusammenhang stehen, aber eine produktive Abstraktionsschicht für solche Fragen darstellen können. Die Geschichte der formalisierten Erzählforschung beginnt in den 1960er Jahren (Hockey 2000, Kap. 5), wobei allerdings bis in die Gegenwart der Einfluss von Propps Korpusstudie zu den Zaubermärchen aus den 1920er Jahren spürbar ist (Meister 2014). Das spezifische Wissen über Text- und Kontextstrukturen musste in den frühen Jahren jeweils aufwendig manuell modelliert werden und war aufgrund seiner Text- und Domänenabhängigkeit nur schlecht übertragbar. Dennoch entwickelte sich eine Forschungsgemeinschaft, die sich zumeist auf die Analyse von Einzeltexten und kleinen Sammlungen konzentrierte und auf die Anwendung einfacher statistischer Verfahren bzw. die Sichtung aller Belege aufgrund von Suchmustern (z.B. Rommel 1995). Mit der Verfügbarkeit größerer Textsammlung entwickelte sich das Feld der formalen Modellierung rasch. Im Fokus standen dabei die Zeit (Mani 2010), Figurenintention (Elson, 2012) und Namen (van Dalen-Oskam 2013). Inzwischen liegen auch Vorschläge vor, die einen Großteil der Genette'schen Kategorien abdecken (vgl. Mani 2013; Gius 2015). Die starke Strukturierung von Dramen hat zu formalisierten Ansätzen geführt: Marcus (1973) repräsentiert Dramen als Binärmatrix und definiert eine Reihe von Parametern, die sich daraus errechnen lassen. Draxler (1988) wertet ebenfalls gegebene Binärmatritzen aus und erweitert Marcus' Definitionen für Szenen und Akte. Mit einer eigenen Implementierung der Definitionen von Marcus (IDAP[3]) führt Ilsemann (z.B. 1998, 2008) verschiedene Untersuchungen auf Shakespeare-Dramen durch.

Die sprachübergreifend adäquate Modellierung und Annotation von Eigenschaften lyrischer Texte wird im Projekt POSTDATA untersucht (González-Blanco et al. 2016; Malta et.al. 2017). Ausgehend von der hoch verdichteten Sprache in Lyrik wäre es fruchtbar, systematischer die Kategorien zu nutzen, die die Rhetorik zur Verfügung stellt, weil in Gedichten gehäuft Beispiele für rhetorische Mittel vorkommen, etwa Parallelismen oder uneigentliche Rede. Viel Forschung dagegen gibt es zu **Metaphern**, die ein zentrales Merkmal literarischer Texte darstellen und insbesondere für die Analyse lyrischer Texte hochrelevant sind. Gleichzeitig sind Metaphern stark kontext- und interpretationsabhängig, was ihre Operationalisierung komplex und herausfordernd macht. Gleichwohl wurden sowohl Ansätze zur intersubjektiv stabilen, manuellen Annotation (englisch/niederländisch: Steen et al., 2010) als auch automatischen Erkennung (Shutova 2015) publiziert. Untersuchungsgegenstände sind ferner kulturelle Differenzen, die sich textuell niederschlagen (z.B. Bestimmbarkeit bzw. Bestimmung des Autorgenders: Olsen 2005; Argamon et al. 2009; Rybicki, 2015 oder der Einfluss der Kategorie Gender im Kontext von Korpuserstellung und Kanonisierung: Mandell, 2015), emotionsgeladene Wörter und Wortgruppen in literarischen Texten (Klinger et al., 2015; Buechel et al., 2017), Personifikation (Dorst, 2011) oder etwa der Einsatz von Topic Modeling für die Analyse der ekphrastischen Tradition (Rhody 2012). Daneben entstehen auch zunehmend Ressourcen, die eine wichtige Grundlage für weitere Arbeitsschritte darstellen können, z.B. Lexika für die *Sentiment Analysis* oder Verzeichnisse von narrativen Topoi, etwa die französischsprachige Datenbank SatorBase für erzählerische Topoi vor 1800 (Sinclair et al. 2000–2006).

2.2 Eigene Vorarbeiten

Fotis Jannidis beschäftigt sich seit (Jannidis & Lauer & Rapp 2006) mit Fragen der quantitativen Analyse narrativer Texte und hat sich mit deren theoretischen Grundlagen (Jannidis 2010) und allgemein mit Fragen der formalen Modellierung (Flanders & Jannidis 2016) auseinandergesetzt. In den letzten Jahren war er an der Erstellung eines Goldstandards für Figurenreferenzen in narrativen Texten beteiligt (DROC 2017). Einen weiteren Arbeitsbereich stellt die quantitative Analyse literarischer Texte dar, etwa die Analyse von Themen der Zeitschrift

[3] http://www.unics.uni-hannover.de/nhtnilse/progdoku.html

'Die Grenzboten' mittels Topic Modeling (Jannidis 2016) oder zur Klassifikation von Romangattungen (Hettinger et al. 2016) oder zum Happy End in Romanen (Jannidis et al. 2017). Die Analyse einer Textsammlung aus dem 19. Jahrhundert ('Novellenschatz') zeigt u.a., dass deren Zusammensetzung auch durch die Geschlechtertheorie der Herausgeber bestimmt ist (Jannidis 2017). In stilometrischen Arbeiten hat Jannidis die Leistungsfähigkeit von Verfahren wie Burrows Delta auch jenseits der Autorschaftszuschreibung erprobt (Jannidis & Lauer 2014), ist der Frage nachgegangen, wie sich der stilometrische Stilbegriff zum literaturwissenschaftlichen verhält (Jannidis 2014), und hat die Funktionsweise von Delta genauer untersucht (Evert et al. 2017). Vom 9.-13.10.2017 hat Jannidis das Symposium 'Digitale Literaturwissenschaft' in der Reihe der 'Literaturwissenschaftlichen DFG-Symposien' organisiert.[4]

Evelyn Gius hat für manuelle und automatische Analyse annotierte Korpora erstellt (Bögel et al. 2015, 2016; Gius 2013a, b; Gius 2015; Gius et al. 2017a, b, c), und dabei auch die Auswirkungen kollaborativer Analysemodi auf die literaturwissenschaftliche Arbeit reflektiert (Gius & Jacke 2015a, 2016). Insbesondere hat sie Guidelines (Gius & Jacke 2015b) und Evaluationskriterien für Annotationen entwickelt, die die Erstellung von Goldstandards auch unter Rückgriff auf verhältnismäßig vage literaturwissenschaftliche Konzepte ermöglichen sowie die Polyvalenz und Ambiguität literarischer Texte gegen die Bedarfe der Einstimmigkeit von kollaborativen Annotationen abwägen (Gius & Jacke 2017). Außerdem hat Gius zu literaturwissenschaftlich anschlussfähigen Anwendungen für die Textanalyse geforscht. Dafür arbeitet sie u.a. an der Textanalyse-Plattform CATMA[5] mit, die konzeptionell auf die Modellierung des literaturwissenschaftlich-hermeneutischen Textanalyseprozesses ausgerichtet ist und seit 2008 vielfach eingesetzt wird. Zudem hat sie sich mit der Frage nach der adäquaten Visualisierung geisteswissenschaftlicher Daten im Projekt 3DH[6] auseinandergesetzt (Gius & Petris 2015, Gius et al. 2017), ist Ko-Organisatorin eines shared tasks zur Modellierung und Erkennung von Erzählebenen (zusammen mit Nils Reiter, Jannik Strötgen und Marcus Willand; Reiter et al. 2017)[7], und untersucht die Rolle von manuellen wie automatischen Annotationen in hermeneutischen Textanalyseprozessen im Projekt hermA[8].

Ein Ziel der Forschung von **Jonas Kuhn** war stets, Algorithmen für Textanalyseaufgaben nicht nur möglichst effektiv zu gestalten, sondern die Wechselwirkungen mit den Einsatzfeldern, denen die übergeordneten Fragen entstammen, zu reflektieren und in Workflows einzubeziehen. Ausgehend von maschinellen Lernverfahren für Textanalyseaufgaben lag der Fokus zunächst auf dem Spannungsverhältnis zu linguistischer Theorie und Korpuslinguistik (so in Arbeiten zur Syntax, u.a. in SFB 732, etwa Seeker und Kuhn 2013; Zarrieß und Kuhn 2013). Eng verbunden mit Arbeiten zu Algorithmen und Methoden liegt ein Schwerpunkt in Kuhns Arbeitsgruppe auf Fragen der Ressourceninfrastruktur (vgl. u.a. Gärtner et al. 2013; Kuhn und Blessing, erscheint). So leitet Kuhn seit 2010 das Stuttgarter CLARIN-Zentrum. Die Vielschichtigkeit der Kontextbezüge bei der Textanalyse legte eine Erweiterung der betrachteten Fragekontexte auf andere textwissenschaftliche Disziplinen nahe, in Kooperationsprojekten zu politischen Diskursen (u.a. in dem BMBF-geförderten Projekt *e-Identity* mit Cathleen Kantner und ab 2018 dem Projekt *Modeling Argumentation Dynamics in Political Discourse* im SPP 1999 „Robust Argumentation Machines") und literaturwissenschaftlichen Untersuchungsgegenständen, so dem BMBF-Projekt *ePoetics* mit Sandra Richter und in Arbeiten zu Übersetzungsalignierung (Cap et al. 2015), Koreferenz-Erkennung (Rösiger und Kuhn 2016), sowie

[4] Zum Programm vgl. https://goo.gl/HUYC9x.
[5] http://www.catma.de/
[6] http://threedh.net/
[7] https://sharedtasksinthedh.github.io/
[8] https://www.herma.uni-hamburg.de/

Überlegungen zu robusten Analyseverfahren für interpretationsrelevante narratologischen Kategorien.[9] Zentral angelegt auf die interdisziplinäre Reflexion von Textanalyse ist das BMBF-*Digital Humanities*-Zentrum CRETA, das Kuhn federführend konzipiert hat.

Christof Schöch hat sich seit seiner annotationsbasierten Dissertation zu literarischen Techniken der Beschreibung im frz. Roman von 1750-1800 (Schöch 2010) ein wachsendes Spektrum an Methoden der quantitativen Literaturwissenschaften und der Digital Humanities angeeignet. Dabei ging es u.a. um die Definition zentraler Begriffe im Kontext digitaler Methoden (Daten: Schöch 2013; Stil: Herrmann et al. 2015) oder die Strukturierung der Digital Humanities durch eine Taxonomie (TaDiRAH: Dombrowski et al. 2016). Seit 2014 leitet Schöch die BMBF-geförderte Nachwuchsgruppe "Computergestützte literarische Gattungsstilistik" (CLiGS). Hier sind u.a. Arbeiten zu Autorschaft und Gattungszugehörigkeit aus stilometrischer Sicht (Schöch 2014) sowie aus Perspektive des Topic Modeling entstanden, sowohl zum französischen Theater (Schöch 2017a) als auch zum französischen Kriminalroman (Schöch 2017b). Ein neueres Thema sind Maße zur Extraktion distinktiver Merkmale von Textgruppen mit einem Fokus auf Burrows' “Zeta” (Schöch angenommen). Schöch hat mehrere Kapitel zum Lehrbuch *Digital Humanities* (Jannidis et al. 2017) beigetragen. Kürzlich wurde eine von Schöch beantragte COST-Action zum Thema "Distant Reading for European Literary History" bewilligt (Laufzeit Dez. 2017-Nov. 2021), die als Netzwerk von über 20 Ländern einen internationalen Resonanzboden sowie breite Kooperationsmöglichkeiten für das SPP *Computational Literary Studies* bieten wird.

Als Literaturwissenschaftlerin mit einem ausgeprägten Forschungsschwerpunkt auf der Theorie- und Begriffsbildung hat **Simone Winko** Arbeiten zur Literaturtheorie (Köppe & Winko 2013) vorgelegt, zur Begriffsbildung, etwa zum Interpretationsbegriff (Winko 2002) und zum Begriff der Literarizität (Winko 2009a), sowie zu literaturwissenschaftlichen Analyseverfahren (Winko 2009, Köppe & Winko 2010). Probleme der Begriffsbildung und Textanalyse stehen auch im Zentrum der Beiträge zur digitalen Literatur (Winko 1999, Winko 2005). Für die literaturwissenschaftliche Reflexionsebene des beantragten SPP sind der terminologische und der methodologische Schwerpunkt von Winkos Forschungen einschlägig, ebenso die Arbeiten zu den Standards literaturwissenschaftlichen Argumentierens (Winko 2015). Sie hat die Sektion 1 “Literatur(wissenschaft) unter digitalen Bedingungen” des DFG-Symposiums “Digitale Literaturwissenschaft” (2017) geleitet. Für ihr Projekt “Emotionen in Lyrik-Anthologien um 1900: Quantitative und hermeneutische Ansätze“ hat Winko nach TEI-Standards ein Korpus von 12 zeitgenössischen Gedicht-Anthologien (ca. 2.800 Gedichte) aufgebaut. Das Projekt ist als Use Case eingebunden in DARIAH-DE III, Cluster 5: Big Data in den Geisteswissenschaften: Topic Modeling.

3 Themenbezogene Publikationen der Mitglieder des Programmausschusses

3.1 Veröffentlichte Arbeiten aus Publikationsorganen mit wissenschaftlicher Qualitätssicherung, Buchveröffentlichungen sowie bereits zur Veröffentlichung angenommene, aber noch nicht veröffentlichte Arbeiten

Bögel, Thomas; Gertz, Michael; **Gius, Evelyn**; Jacke, Janina; Meister, Jan Christoph; Petris, Marco; Strötgen, Jannik, „Collaborative Text Annotation Meets Machine Learning. heureCLÉA, a Digital Heuristic of Narrative“, *DHCommons* 1 (2015). https://goo.gl/qTjTtf

Gius, Evelyn, *Erzählen über Konflikte. Ein Beitrag zur digitalen Narratologie,* Berlin/Boston 2015.

Gius, Evelyn; Jacke, Janina, „Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse“, *Sonderband Der Zeitschrift für digitale Geisteswissenschaften/*1 (2015).

Gius, Evelyn; Jacke, Janina, „The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Text Analysis“, *International Journal of Humanities and Arts Computing* 11/2 (2017), 233–254.(im Erscheinen)

Jannidis, Fotis; Flanders, Julia, „Data Modeling“, in Schreibman, Susan; Siemens, Ray; Unsworth, John (Hg.): *A New Companion to Digital Humanities* (2016), 229-237.

[9] Hierzu Kuhns Beitrag zum DFG-Symposion *Digitale Literaturwissenschaft* 2017: “Empirie – Beschreibung – Interpretation: über den Platz von Computermodellen in den hermeneutisch- historisch orientierten Literaturwissenschaften”.

Jannidis, Fotis, „Perspektiven quantitativer Untersuchungen des Novellenschatzes.“, *Zeitschrift für Literaturwissenschaft und Linguistik* 68,5 (2017), 1–21.

Evert, Stefan; **Jannidis, Fotis**; Proisl, Thomas; Reger, Isabella; Pielström, Steffen; **Schöch, Christof**; Vitt, Thorsten, „Understanding and explaining Delta measures for authorship attribution“, *Digital Scholarship Humanities* (2017).

Seeker, Wolfgang; **Kuhn, Jonas**, „The Effects of Syntactic Features in Automatic Prediction of Morphology“, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013), 333–344.

Richardson, Kyle; **Kuhn, Jonas**, Learning to Make Inferences in a Semantic Parsing Task. Transactions of the Association of Computational Linguistics (TACL), 4 (2016), 155-168.

Kuhn, Jonas, Computerlinguistische Textanalyse in der Literaturwissenschaft? – oder: »The Importance of Being Earnest« bei quantitativen Untersuchungen, in: Andrea Albrecht, Sandra Richter, Marcel Lepper, Marcus Willand, Toni Bernhart (Hrsg.), *Quantitative Verfahren in der Literaturwissenschaft. Von einer Scientia Quantitatis zu den Digital Humanities* (angenommen).

Schöch, Christof, „Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama“, *Digital Humanities Quarterly* 11/2 (2017a). https://goo.gl/W6qZ1X

Schöch, Christof, „Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie“, in: Andrea Albrecht, Sandra Richter, Marcel Lepper, Marcus Willand, Toni Bernhart (Hrsg.), *Quantitative Verfahren in der Literaturwissenschaft. Von einer Scientia Quantitatis zu den Digital Humanities* (angenommen).

Winko, Simone, „Textualitätsannahmen und die Analyse literarischer Texte“, *Zeitschrift für Germanistische Linguistik* 36/3 (2009), 427–443.

Winko, Simone, „Standards literaturwissenschaftlichen Argumentierens. Grundlagen und Forschungsfragen“, *Germanisch-Romanische Monatsschrift* 65/1 (2015), 14–29.

Winko, Simone, „Literatur und Literaturwissenschaft im digitalen Zeitalter. Ein Überblick“, *Der Deutschunterricht* LXVIII/5 (2016), 2–13.

3.2 Andere Veröffentlichungen

Jannidis, Fotis, „Der Autor ganz nah – Autorstil in Stilistik und Stilometrie“, in: Matthias Schaffrick, Marcus Willand (Hrsg.), *Theorien und Praktiken der Autorschaft,* Berlin 2014, 169–195.

Jannidis, Fotis; Lauer, Gerhard, „Burrows’s Delta and Its Use in German Literary History“, *Distant Readings. Topologies of German Culture in the Long Nineteenth Century* (2014), 29-54.

Kuhn, Jonas; Blessing, Andre, „Die Exploration biographischer Textsammlungen mit computerlinguistischen Werkzeugen – methodische Überlegungen zur Übertragung komplexer Analyseketten in den Digital Humanities“, in: Christine Gruber, Ágoston Zénó Bernád, Maximilian Kaiser (Hrsg.), *Europa baut auf Biographien. Aspekte, Bausteine, Normen und Standards für eine europäische Biographik,* Wien 2017, 215–246.

Schöch, Christof, „Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“, *Philologie im Netz* Beiheft 7 (2014), 130–157. http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf

Winko, Simone, „Auf der Suche nach der Weltformel. Literarizität und Poetizität in der neueren literaturtheoretischen Diskussion“, in: Simone Winko, Fotis Jannidis, Gerhard Lauer (Hrsg.), *Grenzen der Literatur. Zum Begriff und Phänomen des Literarischen,* Berlin, New York 2009a, 374-396.

4 Literaturverzeichnis

Agirrezabal, Manex; Arrieta, Bertol; Astigarraga, Aitzol; Hulden, Mans, „ZeuScansion. A tool for scansion of English poetry“, *11th International Conference on Finite State Methods and Natural Language Processing* (2013).

Allison, Sarah; Heuser, Ryan; Jockers, Matthew; Moretti, Franco; Witmore, Michael, „Quantitative Formalism“, *Stanford Literary Lab Pamphlet* 1 (2011).

Ardanuy, Mariona Coll; Sporleder, Caroline, „Structure-based Clustering of Novels“, *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* 2014.

Argamon, Shlomo; Jean-Baptiste, Guerlain; Russell, Horton; Olsen, Mark, „Vive la Différence! Text Mining Gender Difference in French Literature“, *Digital Humanities Quarterly* 3/2 (2009).

Bamman, David; O’Connor, Brendan; Smith, Noah A., „Learning Latent Personas of Film Characters“, *Proceedings of the 51st Annual Meeting of the ACL,* Sofia, Bulgaria 2013, 352-361.

– ; Underwood, Ted; Smith, Noah A., „A Bayesian Mixed Effects Model of Literary Character“, *Proceedings of the 52nd Annual Meeting of the ACL,* Baltimore, MD, USA 2014, 370-379.

Beaudoin, Valérie; Yvon, François, „The Metrometer: a Tool for Analysing French Verse“, *Literary and Linguistic Computing*/11 (1996), 23-32.

Biber, Douglas; Conrad, Susan; Reppen, Randi, *Corpus Linguistics. Investigating Language, Structure and Use,* Cambridge u.a. 1998.
– ; Finegan, Edward, „Drift and the evolution of English style. A history of three genres“, *Language* 65 (1989), 487-517.
Bilenko, Natalia Y; Miyakawa, Asako, *Visualization of Narrative Structure. Analysis of Sentiments and Character Interaction in Fiction* 2013, Internet: http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/images/7/7b/AMNBpaper.pdf.
Birnbaum, David j; Thorsen, Elise, „Markup and meter. Using XML tools to teach a computer to think about versification“, *Proceedings of Balisage: The Markup Conference,* Washington D.C. 2015.
Björkelund, Andres; Faleńska, Agnieszka; Seeker, Wolfgang; Kuhn, Jonas, „How to Train Dependency Parsers with Inexact Search for Joint Sentence Boundary Detection and Parsing of Entire Documents“, Proceedings of the 54th Annual Meeting of the ACL /1 (2016).
Bobenhausen, Klemens; Gehl, Günter, „Automatisches metrisches Markup deutschsprachiger Gedichte“, *Jahrbuch für Computerphilologie* 9 (2009), 61-85.
Bögel, Thomas; Gius, Evelyn; Jacke, Janina; Strötgen, Jannik, „From Order to Order Switch. Mediating between Complexity and Reproducibility in the Context of Automated Literary Annotation“, *Digital Humanities 2016: Conference Abstracts* (2016), 379-382.
Brunner, Annelen, *Automatische Erkennung von Redewiedergabe,* Berlin, New York 2015.
Bubenhofer, Noah; Müller, Nicole; Scharloth, Joachim, „Narrative Muster und Diskursanalyse. Ein datengeleiteter Ansatz“, *Zeitschrift für Semiotik* 35/3-4 (2013), 419-444.
Buechel, Sven; Hellrich, Johannes; Hahn, Udo, „The Course of Emotion in Three Centuries of German Text—A Methodological Framework“, *Abstracts for DH 2017* (2017).
Bullard, Joseph; Alm, Cecilia Ovesdotter, „Computational analysis to explore authors' depiction of characters“, *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL),* Gothenburg, Sweden 2014, 11-16.
Burrows, John, „Computers and the Idea of Authorship“, in: Fotis Jannidis, Gerhard Lauer, Matias Martinez, Simone Winko (Hrsg.), *Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs,* Berlin, New York 1999, 167-181.
– , „'Delta' - A measure of stylistic difference and a guide to likely authorship“, *Literary and Linguistic Computing* 17.3 (2002), 267-287.
Celikyilmaz, Asli; Hakkani-Tur, Dilek; He, Hua; Kondrak, Greg; Barbosa, Denilson, „The Actor Topic Model for Extracting Social Networks in Literary Narrative“, *NIPS Workshop: Machine Learning for Social Computing* 2010.
Chambers, Nathanael; Jurasky, Dan, „Unsupervised Learning of Narrative Schemas and their Participants“, *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP,* Singapore 2009, 602-610.
Chaturvedi, Manish; Gannod, Gerald; Mandell, Laura; Hodgson, Eric, *Visualization of TEI encoded texts in support of close reading* 2011.
Craig, Hugh; Kinney, Arthur F., *Shakespeare, Computers, and the Mystery of Authorship,* Cambridge 2009.
Dombrowski, Quinn; Borek, Luise; Perkins, Jody; Schöch, Christof, „TaDiRAH: a Case Study in Pragmatic Classification“, *Digital Humanities Quarterly* 10/1 (2016).
Dorst, Aletta G., „Personification in discourse. Linguistic forms, conceptual structures and communicative functions“, *Language and Literature* 20/2 (2011), 113--135.
Draxler, Christoph, *Computerunterstützte Dramenanalyse,* München 1988.
Dunn, Stuart; Schumacher, Mareike, „Explaining Events to Computers. Critical Quantification, Multiplicity and Narratives in Cultural Heritage“, *Digital Humanities Quarterly* 10/3 (2016).
Eder, Maciej; Rybicki, Jan, „Do birds of a feather really flock together, or how to choose training samples for authorship attribution“, *Literary and Linguistic Computing* 28/2 (2013), 229-236.
Egbert, Jesse, „Style in nineteenth century fiction. A Multi-Dimensional analysis“, *Scientific Study of Literature* 2/2 (2012), 167--198.
Elsner, Micha, „Character-based kernels for novelistic plot structure“, *Proceedings of the 13th Conference of the European Chapter of the ACL* 2012, 634-644.
Elson, David K., „DramaBank: Annotating Agency in Narrative Discourse.“, *Proceedings of the Eight International Conference on Language Resources and Evaluation* (2012).
– ; Dames, Nicholas; McKeown, Kathleen R., „Extracting Social Networks from Literary Fiction“, *Proceedings of the 48th Annual Meeting of the ACL* 2010, 138-147.
Estes, Alex; Hench, Christopher, „Supervised Machine Learning for Hybrid Meter“, *Proceedings of the Fifth Workshop on Computational Linguistics for Literature* (2016), 1-8.
Fusi, Daniele, *An Expert System for the Classical Languages. Metrical Analysis Components* 2009.

Gärtner, Markus; Thiele, Gregor; Seeker, Wolfgang; Björkelund, Andres; Kuhn, Jonas, „ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks“, *Proceedings of the 51st Annual Meeting of the ACL: System Demonstrations* (2013).

Gius, Evelyn, „Korpus ‚Erzählen über Konflikte‘“, 2013a. https://doi.org/10.5281/zenodo.894732.

– ; Jacke, Janina, „Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets“, 2015b. http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf.

– ; Jacke, Janina, „Kollaboratives Annotieren literarischer Texte“, *DHd 2016 Modellierung, Vernetzung, Visualisierung* (2016), 240-243.

– ; Jacke, Janina; Meister, Jan Christoph; Petris, Marco, „Heurecléa Source Documents: 1.0“, 2017a. https://doi.org/10.5281/zenodo.274962.

– ; Petris, Marco, „Die explorative Visualisierung von Texten“, *DHd2015 Von Daten zu Erkenntnissen: Book of Abstracts* (2015), 85-92.

– ; Kleymann, Rabea; Meister, Jan Christoph; Petris, Marco, „Datenvisualisierung als Aisthesis: Plädoyer für ein geisteswissenschaftliches Visualisierungsparadigma“, Dhd 2017 Digitale Nachhaltigkeit Konferenzabstracts (2017), 115-120.

González-Blanco, Elena; Riande, Gimena Del Rio; Cantón, Clara Martinez, „Linked open data to represent multilingual poetry collections“, *Proceedings of the LREC 2016 Workshop, LDL 2016 - 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources* 2016.

Greene, Erica; Bodrumlu, Tugba; Knight, Kevin, „Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation“, *Empirical Methods in NLP* (2010).

Grzybek, Peter, „The Emergence of Stylometry: Prolegomena to the History of Term and Concept“, in: Katalin Kroó, Peeter Torop (Hrsg.), *Text within Text - Culture within Culture*, Budapest 2014, 58–75.

Hammond, Michael, „Calculating syllable count automatically from fixed-meter poetry in English and Welsh“, *Literary and Linguistic Computing* 29/2 (2014), 218-233.

Herrmann, Berenike; van Dalen-Oskam, Karina; Schöch, Christof, „Revisiting Style, a Key Concept in Literary Studies“, *Journal of Literary Theory* 9/1 (2015), 25-52.

Hettinger, Lena; Reger, Isabelle; Jannidis, Fotis; Hotho, Andreas, „Classification of Literary Subgenres“, *Modellierung, Vernetzung, Visualisierung DHd 2016 Universität Leipzig. Konferenzabstracts* (2016), 160-164.

Heuser, Ryan; Le-Khac, Long, „A Quantitative Literary History of 2.958 Nineteenth-Century British Novels. The Semantic Cohort Method“, *Stanford Literary Lab Pamphlet 4* (2012).

Hockey, Susan, *Electronic Texts in the Humanities. Principles and Practice,* New York 2000.

Hover, David L., „The microanalysis of style variation“, *Digital Scholarship in the Humanities* (2017).

Ilsemann, Hartmut, *Shakespeare Disassembled. Eine quantitative Analyse der Dramen Shakespeares,* Frankfurt a.M. 1998.

– ; „More statistical observations on speech lengths in shakespeare’s plays“, *Literary and Linguistic Computing* 23/4 (2008), 397-407.

Jannidis, Fotis; Lauer, Gerhard; Rapp, Andrea, „Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des Romans in Deutschland (1500-1900)“, in: Lucas Marco Gisi, Jan Loop, Michael Stolz (Hrsg.), *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*, Zürich 2006.

– ; Hubertus, Kohle; Rehbein, Malte (Hrsg.), *Digital Humanities. Eine Einführung,* Stuttgart 2017.

– ; Reger, Isabella; Zehe, Albin; Becker, Martin; Hettinger, Lena; Hotho, Andreas, „Analyzing Features for the Detection of Happy Endings in German Novels“,(2017) https://arxiv.org/abs/1611.09028v1.

Jockers, Matthew L., *Macroanalysis. Digital Methods and Literary History,* Urbana 2013.

– , *Revealing Sentiment and Plot Arcs with the Syuzhet Package* 2015, http://www.matthewjockers.net/2015/02/02/syuzhet/.

– ; Kiriloff, Gabi, „Understanding Gender and Character Agency in the 19th Century Novel“, *Journal of Cultural Analytics* (2016).

Karsdorp, Folgert; Kestemont, Mike; Schöch, Christof; van den Bosch, Antal, „The Love Equation. Computational Modeling of Romantic Relationships in French Classical Drama“, *6th Workshop on Computational Models of Narrative* (2015), 98-107.

Kazantseva, Anna; Szpakowicz, Stan, „Hierarchical Topical Segmentation with Affinity Propagation“, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (2014), 37-47.

Kelly, David; Tonra, Justin; Reid, Lindsay Ann, *Personæ - a character-visualisation tool for dramatic texts* 2016, http://www.davidkelly.ie/projects/personae/.

Klinger, Roman; Suliya, Samat Surayya; Reiter, Nils, „Automatic Emotion Detection for Quantitative Literary Studies - A case study based on Franz Kafka's "Das Schloss" and "Amerika".", *Digital Humanities 2016: Conference Abstracts* (2015), 826-628.
Koolen, Corina; van Cranenburgh, Andreas, „Blue Eyes and Porcelain Cheeks. Computational Extraction of Physical Descriptions from Dutch Chick Lit and Literary Novels", *Digital Scholarship in the Humanities* (2017).
Köppe, Tilmann; Winko, Simone, „Methoden der analytischen Literaturwissenschaft", in: Vera Nünning, Ansgar Nünning (Hrsg.), *Methoden der literatur- und kulturwissenschaftlichen Textanalyse,* Weimar 2010.
– ; Winko, Simone, *Neuere Literaturtheorien. Eine Einführung,* 2., aktualisierte und erweiterte Auflage, Stuttgart, Weimar 2013.
Love, Harold, *Attributing Authorship. An Introduction,* Cambridge 2002.
Mahlberg, Michaela; Smith, Catherine; Preston, and Simon, „Phrases in literary contexts. Patterns and distributions of suspensions in Dickens novels", *International Journal of Corpus Linguistics* 18/1 (2013), 35-56.
Makazhanov, Aibek; Barbosa, Denilson; Kondrak, Grzegorz, *Extracting Family Relationship Networks from Novels* 2014, https://arxiv.org/pdf/1405.0603.pdf.
Malta, Mariana Curado; González-Blanco, Elena; Cantón, Clara Martínez; Rio, Gimena Del, „A Common Conceptual Model for the Study of Poetry in the Digital Humanities", *Abstracts for DH 2017* (2017).
Mandell, Laura C., „Gendering Digital Literary History. What Counts for Digital Humanities", in: Susan Schreibman, Ray Siemens, John Unsworth (Hrsg.), *New Companion to Digital Humanities,* Chichester 2015, 511-523.
Mani, Inderjeet, *The Imagined Moment. Time, Narrative, and Computation,* Lincoln, Nebraska 2010.
– , „Computational Narratology", in: Peter Hühn, John Pier, Wolf Schmid, Jörg Schönert (Hrsg.), *The living handbook of narratology,* Hamburg 2013.
Marcus, Solomon, *Mathematische Poetik,* Frankfurt a.M. 1973.
McCurdy, Nina; Lein, Julie; Coles, Katharine; Meyer, Miriah, „Poemage. Visualizing the Sonic Topology of a Poem", *IEEE Transactions on Visualization and Computer Graphics* 2016.
Meister, Jan Christoph, „Computing Action. A Narratological Approach", *Narratologia,* Berlin/New York 2003.
– , „Toward a Computational Narratology", in: Maristella Agosti, Francesca Tomasi (Hrsg.), *Collaborative Research Practices and Shared Infrastructures for Humanities Computing. 2nd Aiucd Annual Conference, Aiucd 2013,* Padua, Italy 2014, 17-38.
Moretti, Franco, *Graphs, Maps, Trees. Abstract Models for a Literary History,* New York 2005.
– , *Network Theory, Plot Analysis,* Stanford 2011, https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.
Nalisnick, Eric T; Baird, Henry S., „Character-to-character sentiment analysis in shakespeare's plays", *In Proceedings of the 51st Annual Meeting of the ACL,* Sofia, Bulgaria 2013, 479-483.
Navarro-Colorado, Borja; Lafoz, Marıa Ribes; Sanchez, Noelia, „Metrical annotation of a large corpus of Spanish sonnets. Representation, scansion and evaluation", *Workshop on Computational Linguistics for Literature,* Denver 2016.
Olsen, Mark, „Ecriture Feminine. Searching for an Indefinable Practice?", *Literary and Linguistic Computing* 20/1 (2005), 147-164.
Park, Gyeong-Mi; Sung-Hwan, Kim; Hwang, Hye-Ryeon; Hwan-Gue, Cho, „Complex System Analysis of Social Networks Extracted from Literary Fictions", *International Journal of Machine Learning and Computing* 3/1 (2013), 107-111.
Piper, Andrew, „Fictionality", *Journal of Cultural Analytics* (2016).
Reichert, Waltraud, *Informationsästhetische Untersuchungen an Dramen*, Stuttgart 1965.
Reiter, Nils; Frank, Anette; Hellwig, Oliver, „An NLP-based Cross-document Approach to Narrative Structure Discovery", *Literary and Linguistic Computing* 29/4 (2014), 583-605.
– , „Towards Annotating Narrative Segments", in: Kalliopi Zervanou, Marieke van Erp, Beatrice Alex (Hrsg.), Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Beijing, China 2015.
– ; Gius, Evelyn; Strötgen, Jannik; Willand, Marcus, „A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts", *Digital Humanities 2017: Conference Abstracts* (2017).
Rhody, Lisa M., „Topic Modeling and Figurative Language", *Journal of Digital Humanities* 2 (2012).
Roberts-Smith, Jennifer; Gabriele, Sandra; Ruecker, Stan; Sinclair, Stéfan; Bouchard, Matt; Jakacki, Shawn DeSouza-Coelho Diane; Kong, Annemarie; Lam, David; Rodriguez, Omar, „The text and the line of action. Re-conceiving watching the script", *New Knowledge Environments* 1/1 (2009).

Rommel, Thomas, *'And trace it in this poem every line'. Methoden und Verfahren computergestützter Textanalyse am Beispiel von Lord Byrons Don Juan,* Tübingen 1995.
Rösiger, Ina; Kuhn, Jonas, „IMS HotCoref De: A data-driven co-reference resolver for German", Proceedings of LREC 2016.
Rybicki, Jan, „Vive La Différence. Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies", *Digital Scholarship in the Humanities* 31/4 (2015), 746-761.
Schöch, Christof, *La Description double dans le roman français des Lumières, 1760-1800,* Paris 2011.
– , „Big? Smart? Clean? Messy? Data in the Humanities", *Journal of Digital Humanities* 2/3 (2013), 2-13.
– , „Gattungen des Kriminalromans. Ein quantitativer, topic-basierter Zugang", in: Sabine Schmitz, Corinna Koch, Sandra Lang (Hrsg.), *Dialogische Krimianalysen: Fachdidaktik und Literaturwissenschaft untersuchen aktuelle Krimiliteratur aus Belgien und Frankreich,* Frankfurt a.M. 2017b, 37-59.
Semino, Elena; Short, Mick, *Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing,* London 2004.
Shutova, Ekaterina, „Design and Evaluation of Metaphor Processing Systems", *Computatitional Linguistics* 41/4 (2015).
Sinclair, Stéfan, *Satorbase* 2000-2006, Internet: https://.satorbase.org.
Steen, Gerard; Dorst, Lettie; Herrmann, Berenike; Kaal, Anna; Krennmayr, Tina, „Metaphor in usage", *Cognitive Linguistics* 21 (2010).
Trilcke, Peer, „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft", in: Philip Ajouri, Katja Mellmann, Christoph Rauen (Hrsg.), *Empirie in der Literaturwissenschaft,* Münster 2013, 201-247.
– ; Fischer, Frank; Kampkaspar, Dario, „Digital network analysis of dramatic texts", *DH2015 Conference Abstracts,* Sydney, Australia 2015.
– ; Fischer, Frank; Göbel, Mathias; Kampkaspar, Dario, „Theatre plays as 'small worlds'? network data on the history and typology of german drama", *Digital Humanities 2016: Conference Abstracts,* Kraków 2016, 385-387.
– ; Fischer, Frank; Göbel, Mathias; Kampkaspar, Dario; Kittel, Christopher, „Netzwerkdynamik und Plotanalyse. zur Visualisierung und Berechnung der "Progressiven Strukturierung" literarischer Texte", *Book of Abstracts of DHd 2017,* Bern, Switzerland 2017, 175-180.
Underwood, Ted, „Understanding Genre in a Collection of a Million Volumes", *NEH White Papers* (2015).
– ; Bamman, David, *The instability of gender. The Stone and the Shell. Using large digital libraries to advance literary history* 2016, https://tedunderwood.com/2016/01/09/the-instability-of-gender/.
– ; Sellers, Jordan, „The Emergence of Literary Diction", *Journal of Digital Humanities* 1/2 (2012).
van Dalen-Oskam, Karina, „Names in novels. An experiment in computational stylistics", *LLC: The Journal of Digital scholarship in the Humanities* 28/2 (2013), 359-370.
Wallace, Byron C., „Multiple Narrative Disentanglement: Unraveling Infinite Jest.", *Proceedings of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies* (2012), 1-10.
Wilhelm, Thomas; Burghardt, Manuel; Wolff, Christian, „"to see or not to see" - an interactive tool for the visualization and analysis of shakespeare plays", in: Regina Franken-Wendelstorf, Elisabeth Lindinger, Jürgen Sieck (Hrsg.), *Kultur und Informatik - Visual Worlds and Interactive Spaces,* Berlin 2013, 175-185.
Willand, Marcus; Reiter, Nils, „Geschlecht und Gattung. Digitale Analysen von Kleists ›Familie Schroffenstein‹", Kleist-Jahrbuch 2017, 142–160.
Winko, Simone, „Lost in hypertext? Autorkonzepte und neue Medien", in: Fotis Jannidis et al. (Hrsg.), *Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs* 1999, 511-533.
– , „Lektüre oder Interpretation?", *Mitteilungen des Deutschen Germanistenverbandes* 49/2 (2002), 128-141.
– , „Hyper-Text-Literatur. Digitale Literatur als Herausforderung an die Literaturwissenschaft", in: Harro Segeberg, Simone Winko (Hrsg.), *Digitalität und Literalität. Zur Zukunft der Literatur im Netzzeitalter*, München 2005, 137–157.
Zarrieß, Sina; Kuhn, Jonas, „Combining Referring Expression Generation and Surface Realization: A Corpus-based Investigation of Architectures", Proceedings of ACL (2013).Sofia.

5 Inhaltliche Begründung unter Berücksichtigung der Programmziele

In den letzten rund zehn Jahren ist mit den *Computational Literary Studies* ein Forschungsfeld entstanden, in dessen Zentrum die algorithmenbasierte Analyse größerer Sammlungen literarischer Texte steht. Sie partizipieren dabei an einem Transformationsprozess (der Digitalisierung und datenbasierten Modellierung und Analyse), der eine Vielzahl von geistes- und sozialwissenschaftlichen – wie im übrigen auch naturwissenschaftlichen – Disziplinen erfasst hat. Eine Reihe von Studien konnte zeigen, dass sich mit Verfahren dieses Forschungsfeldes innerhalb der Digital Humanities Erkenntnisse bislang vor allem zu gattungsgeschichtlichen und allgemein literarhistorischen Entwicklungen gewinnen lassen, die entweder Thesen in den etablierten Literaturwissenschaften in Frage stellen (z.B. Piper 2016 zum Begriff der 'Fiktionalität', die entgegen der landläufigen literaturtheoretischen Auffassung doch abhängig von Texteigenschaften konzipiert werden sollte, oder Burrows 1999, der die wirkungsmächtige Annahme Foucaults in Frage gestellt hat, dass der Autor eine reine Konstruktion sei) oder diese aus einer anderen Perspektive bestätigen (z.B. Jockers 2013 zur Brauchbarkeit etablierter Gattungsbegriffe zum englischen Roman im 19. Jahrhundert). Vor allem aber sind ganz neue Einsichten entstanden, da die neuen Verfahren im Gegensatz zu etablierten Ansätzen fundierte Aussagen über dutzende, hunderte, wenn nicht tausende von Texten machen können (so konnten z.B. Heuser und Le Khac 2012 zwei bis dahin ungekannte Trends in der Entwicklung der Sprache des britischen Romans des 19. Jahrhunderts auf der Grundlage von knapp 3000 Romanen nachweisen: eine Abnahme abstrakter Begrifflichkeiten und eine korrespondierende Zunahme eines konkreten, physischen Vokabulars). Was den Verfahren der *Computational Literary Studies* an hermeneutischer Tiefenschärfe fehlt, gleichen sie durch ihre Reichweite aus. Sie ergänzen die vorhandenen Ansätze in den Literaturwissenschaften auf fruchtbare Weise und erweitern deren Methodenspektrum durch den Anschluss an moderne Verfahren der Datenanalyse. Dabei ergibt sich aus der Reflexion, der Validierung und der Optimierung der Methoden ein intensiver Forschungsprozess innerhalb der *Computational Literary Studies*. Zudem bildet die Notwendigkeit der formalen Modellierung literaturwissenschaftlicher Konzepte im Zuge des Forschungsprozesses eine ständige Quelle produktiver Irritation, die zur theoretischen Revision dieser Konzepte Anlass gibt. Die *Computational Literary Studies* sind zu einem *emerging field* geworden, und es liegt daher nahe, die vorhandenen Forschungsinteressen und -kapazitäten jetzt in einem Schwerpunktprogramm zu bündeln.

Wichtig für die Effektivität des Schwerpunktprogramms ist zum einen, dass der Gegenstandsbereich so offen gehalten ist, dass alle Beteiligten die gemeinsame Arbeit am Schnittmengenthema mit der eigenen Forschungsagenda abstimmen können, und zum anderen, dass eine größtmögliche Fokussierung auf den intendierten Kern des Schwerpunkts gewährleistet ist: die Methoden. Das SPP *Computational Literary Studies* ist offen, was die geographische, kulturelle oder zeitliche Dimension der untersuchten Literaturen angeht. Im Zentrum soll die Erkundung und Reflexion von datenorientierten Methoden der Untersuchung literarischer Texte stehen. Dabei wird ein weiter Literaturbegriff zugrunde gelegt, als dessen Kriterien Fiktionalität und eine besondere Art der Sprachverwendung angesetzt werden können. Den prototypischen Schwerpunkt von Literatur bilden fiktionale und sprachlich als poetisch markierte Texte (z.B. Romane, Dramen, Gedichte), und zwar nicht nur der Hochliteratur, sondern auch der Schema- wie z.B. Heftchenliteratur und deren verschiedenen Varianten sowie der Übergänge zwischen ihnen. Darüber hinaus können auch nicht-fiktionale Textsorten wie Briefe und Essays oder mit der Grenze zur Fiktion spielende Formen wie Autobiografien und Reiseliteratur in verschiedenen Formaten einbezogen werden. Der Vergleich mit Gebrauchstexten kann Besonderheiten literarischer Textsorten hervortreten lassen. In Hinsicht auf die Einbettung in einen literaturtheoretischen Kontext bestehen ebenfalls keine Vorgaben.

Die Forschung der letzten Jahrzehnte hat deutlich gemacht, dass das Lesen eines Textes ein vielschichtiger Prozess ist, in dem text- und weltwissenbasierte Impulse zusammenwirken. Die Modellierung spezifischer Aspekte von Texten funktioniert allerdings umso schlechter, je mehr diese Aspekte das vollständige Verstehen des Textes und den Aufbau eines mentalen

Modells der dargestellten Welt voraussetzen. So lassen sich etwa Formen der Redewiedergabe recht gut modellieren, insbesondere wenn sie unmittelbar durch Oberflächenaspekte erkennbar sind. Dagegen stellen z.B. abstrakte Kategorie wie ‘Plotmuster’ oder ‘Metapher’ deutlich größere Herausforderungen für die computergestützte Analyse dar. Diese Differenz zeigte sich im Stand der Forschung (siehe 2.1) darin, dass zu einigen Aspekten vergleichsweise viele Beiträge vorliegen, während andere, aus der Sicht der traditionellen Literaturwissenschaften eng verwandte Aspekte, kaum oder nur mit mäßigem Erfolg bearbeitet worden sind. Hier setzt das SPP *Computational Literary Studies* an, indem es zum Ausgleich der Defizite methodisch innovative und damit auch zugleich risikobehaftete Forschung fördern will.

Die besonderen Herausforderungen ergeben sich, wenn bei der Untersuchung literarischer Texte Algorithmen, Computermodelle und formalisierte Verfahren zur Korpusannotation zum Einsatz kommen – und dies nicht nur zur Vorverarbeitung oder zur Exploration einer Textsammlung, sondern als zentraler Bestandteil der Untersuchungsmethodik. Modellierungsansätze und Komponenten, die hierzu in Anschlag gebracht werden können, speisen sich aus der Computerphilologie, der Literaturwissenschaft, der Korpuslinguistik, der Computerlinguistik und der Informatik. Dieses interdisziplinäre Spannungsfeld hat in den letzten Jahren eine beachtliche Dynamik erfahren: Digitalisierungsinitiativen treiben die digitale Erschließung der Untersuchungsgegenstände voran, der Aufbau von Forschungsinfrastrukturen erlaubt den effizienten Austausch von Ressourcen und Ergebnissen, und die Förderung von disziplinübergreifenden Forschungsprojekten in den Digital Humanities sensibilisiert die angrenzenden Felder für das große Innovationspotenzial, das in der interdisziplinären Methodenkombination liegt. Nicht zuletzt ergeben sich aus den jüngeren Entwicklungen bei maschinellen Lernverfahren für die Analyse von großen Datenmengen Chancen und Herausforderungen; so erzielt man mit *Deep Learning* etwa oft deutlich bessere Klassifikationsergebnisse, aber anders als bei herkömmlichen maschinellen Lernverfahren ist es sehr viel schwerer zu verstehen, welche Texteigenschaften verantwortlich sind, das Verfahren bleibt eine *Black Box*.

Vor dem Hintergrund der beschriebenen Dynamik stehen im Zentrum des Schwerpunktprogramms *Computational Literary Studies*:

- die Entwicklung neuer Verfahren zur korpusgestützten Analyse literarischer Texte, z.B. zur Erkennung von Handlungseinheiten,
- die Identifikation und Auswahl derjenigen Verfahren aus der Fülle der in der Informatik und Computerlinguistik bekannten Verfahren, die für das Feld der *Computational Literary Studies* von Bedeutung sind,
- die systematische Untersuchung von Verfahren im Feld der *Computational Literary Studies*, u.a. um besser zu verstehen, für welche Textsorten und Probleme sie geeignet sind bzw. um den Zusammenhang zwischen Anwendungskontext und Parametrisierung auszuleuchten,
- die Anwendung bereits existierender Algorithmen auf neue Datensätze zur Generierung neuer Erkenntnisse über literarische Phänomene, Entwicklungen und Strukturen, wobei die Kombination aus Fragestellung und Methode neu sein sollte,
- Fragen rund um die formale Modellierung literaturwissenschaftlich relevanter Phänomene und Konzepte, z.B. durch Metadaten für ganze Texte oder Annotationen zur Beschreibung von Textsegmenten, sowie
- weiterführende Fragestellungen, etwa wie der Schritt von der Analyse zur Interpretation gestaltet werden kann oder wie Text-Kontext-Beziehungen analysiert oder modelliert werden können, also z.B. wie Sammlungen literarischer Texte in Beziehung zu anderen literarischen Texten oder anderen Phänomenen (etwa der Sozialstruktur, die wiederum datenbasiert modelliert wird) stehen.

Ein wichtiger Aspekt dieser Ansätze besteht in der Reflexion der Beziehung formaler Methoden zur Literaturwissenschaft: Wo stehen der Untersuchungsgegenstand und die methodischen Ansätze der *Computational Literary Studies* in einem Spannungsverhältnis zueinander? Dabei kann es sich um die Spannung zwischen der Individualität komplexerer Einzelwerke und dem Flächenblick quantitativer Verfahren handeln, aber auch um die Frage, wie

sich historisch einzigartige Entwicklungen mit Verfahren erfassen lassen, die wenigstens zum Teil auf der Annahme von Regelhaftigkeit basieren. Auch die Folgen, die die Arbeitsweisen der *Computational Literary Studies* für die Begriffsbildung der Literaturwissenschaft haben, gehören dazu. Hierbei ergibt sich die komplementäre Anforderung, dass Begriffe der statistischen Analyse in phänomenadäquater Weise auf die Ereignisse und Entitäten des Literatursystems bezogen und datenzentrierte Arbeitspraktiken an die spezifischen Anforderungen der Literaturwissenschaft angepasst werden müssen. Aus den oben skizzierten Besonderheiten des Untersuchungsgegenstands – literarische Texte in ihrem jeweiligen Kontext – ergeben sich spezifische Herausforderungen, die für eine adäquate Entfaltung der methodischen Möglichkeiten zu berücksichtigen sind:

Notwendigkeit der Domänenadaption. Verfügbare Textanalysewerkzeuge sind zumeist auf der Grundlage von Nachrichtentexten optimiert und unterstützen entweder eine linguistische Strukturanalyse oder einfache oberflächennahe Inhaltsrecherchen, für die sehr große Mengen von Beispieldaten verfügbar sind. Jede Übertragung auf literarische Texte und literaturwissenschaftliche Fragestellungen erfordert Anpassungen, die nicht selten sehr weit gehen müssen, wenn Sprachstadium, Stilistik und gattungsspezifische Texteigenschaften berücksichtigt werden sollen. Dabei werden die digital verfügbaren Textmengen aus technologischer Sicht stets vergleichsweise klein sein. "*Big Data*"-Techniken, die in der Lage sind, systematische Generalisierungen aus der großen Menge gleichartiger Daten zu induzieren, werden oft nicht ohne findige Anpassungen zu Erfolgen führen.

Verhältnis von Norm und Abweichung. Etablierte Modelle und Werkzeuge für die Sprach- und Textanalyse gehen in der Regel von einer verhältnismäßig homogenen Gesamtheit von "möglichen" Texten aus, sodass verfügbare Korpora – etwa für linguistische Untersuchungen zu Eigenschaften des Sprachsystems – als repräsentativ angesetzt werden. Die literarischen Texte, die für die *Computational Literary Studies* im Kern des Untersuchungsinteresses liegen, zeichnen sich jedoch zum Teil dadurch aus, dass ihre Eigenschaften in einer jeweils nicht vorhersagbaren Weise von bestehenden Erwartungen abweichen. Das Spannungsverhältnis zwischen dem regelhaft Beschreibbaren und dem Individuellen macht eine differenzierte Weiterentwicklung von korpusbasierten Standardverfahren (auf ohnehin relativ kleinen Referenzkorpora) erforderlich. Auch andere Aspekte literarischer Texte und der Umgang mit ihnen bedingen eine angemessene Umgangsweise, die die schlichte Übernahme von Verfahren problematisch macht; das gilt z.B. für Fiktionalität, das Verhältnis von Erzählstimme und Wahrnehmung usw.

Konsensuelle Annotation literaturwissenschaftlich relevanter Texteigenschaften. Maschinelle Lernverfahren erlauben es, automatische Klassifikatoren für relevante Analysekategorien zu trainieren, wenn geeignete Zielannotationen vorliegen. Mit der korpuslinguistischen Arbeitspraxis der Mehrfachannotation ist eine Methodik etabliert, die für intersubjektiv stabile Kategorien eine Abschätzung der Vorhersagegenauigkeit zulässt. Inwieweit analog zu sprachlichen Standardkategorien eine intersubjektiv stabile Annotation bestimmter Texteigenschaften – wie etwa von Redewiedergabe oder Anschaulichkeit – erstellt werden und systematisch in eine literaturwissenschaftliche Textanalyse oder gar -interpretation einfließen kann, ist allerdings kontrovers. Bei vielen interpretatorischen Fragen ergibt sich eine große Divergenz von Auffassungen aufgrund langer Inferenzprozesse im Anschluss an den Text. Das erzeugt eine Spannung zwischen methodenbedingter Anforderung an Übereinstimmung und gegenstandsbedingtem Auseinandertreten von Auffassungen, die durch innovative Forschungsstrategien bewältigt werden muss.

Transparenz der Modelle. Anders als beim typischen Einsatz maschineller Lernverfahren, wo in erster Linie die Vorhersagequalität des erzeugten Modells zählt, stellt sich bei der Anwendung solcher Verfahren in den Literaturwissenschaften die Frage einer nachvollziehbaren Interpretation von internen Modellrepräsentationen. Die hierfür notwendige Transparenz ist Voraussetzung für ein methodisch reflektiertes Vorgehen und die Vermeidung von Fehlschlüssen aufgrund von modellspezifischen Artefakten. Dieses anspruchsvolle Ziel kann am besten

erreicht werden durch disziplinübergreifende Abstimmung von Modellierungsannahmen, sorgfältige quantitative und qualitative Ergebnisanalysen und die Entwicklung von diagnostischen Werkzeugen. Von großer Relevanz ist die Diskussion der Frage, welche Rückwirkungen die Anwendung quantitativer Verfahren und deren Ergebnisse auf die Theorie und die Begriffsbildung in der Literaturwissenschaft haben. Die aufgelisteten Punkte verdeutlichen, dass eine geeignete Lösung für die *Computational Literary Studies* häufig in der reflektierten Kombination von disziplintypischen Methoden und Arbeitspraktiken liegt. Der nötige disziplinübergreifende Austausch findet seit einigen Jahren bereits punktuell statt, und in den beteiligten Disziplinen besteht generell eine große Offenheit; mit der Einrichtung eines DFG-Schwerpunktprogramms könnte der Aufbau einer hinreichend breiten aktiven Community erreicht und der wechselseitige Austausch befördert werden, der nicht zuletzt eine Voraussetzung für die Etablierung von *best practice*-Ansätzen darstellt.

Ausstrahlung. Wiederkehrende, systematische Vorgehensweisen wie auch konkrete Werkzeuglösungen für Teilprobleme lassen eine breitere Ausstrahlung auf textanalytische Fragestellungen erwarten – auch jenseits der Anwendung auf literarische Texte. Ergebnisse aus dem SPP können so zur Vertiefung des Methodeninventars für Text Mining und andere datenintensive Analyseaufgaben beitragen. So könnte eine Sentiment Analysis, die den komplexen Gefühlsformulierungen in literarischen Texten angemessen ist, auch für andere Texte mit Gewinn eingesetzt werden. In einer von der Digitalisierung geprägten Informationsgesellschaft haben textbasierte Datenanalysen bereits heute erheblichen Stellenwert, komplexere kontextuelle Abhängigkeiten bleiben jedoch häufig unberücksichtigt. Verfahren und Praktiken aus dem SPP könnten zukünftig also zentrale Bausteine für umfassendere Ansätze der reflektierten Datenanalyse darstellen. Unmittelbar profitieren würden damit auch die anderen textwissenschaftlichen Disziplinen in den Geistes- und Sozialwissenschaften, mit denen über Foren der Digital Humanities ein stetiger Austausch besteht.

Die Gegebenheiten eines typischen Analyseprojekts in den *Computational Literary Studies* stellen jedoch auch die Standardmethodik der Modell- und Werkzeugentwicklung in Informatik und Computerlinguistik vor interessante Herausforderungen – der Umfang verfügbarer (homogener) Korpusdaten ist vergleichsweise gering; Goldstandard-Daten sind für die relevante Ausprägung von Analysekomponenten oft nicht verfügbar. Dafür lässt sich die Analysefrage häufig in einem reichen Spektrum von Forschungsansätzen kontextualisieren; beim Analysedesign können aufgrund der philologischen Fachkompetenz Überlegungen angestellt werden, die jenseits der üblichen Modellierung von sachlichem Expertenwissen liegen. Diese Konstellation ist paradigmatisch für andere Spezialanwendungen von maschinellen Lernverfahren u.ä., so dass die Methodenentwicklung in die Informatik ausstrahlen dürfte. So ist etwa der Bedarf an interpretierbaren Modellrepräsentationen (besonders bei neuronalen Ansätzen) breit erkannt; die disziplinübergreifende Praxis in den *Computational Literary Studies* kann hier übertragbare Strategien entwickeln.

5.1 Originalität der wissenschaftlichen Fragestellungen unter thematischen und/oder methodischen Aspekten

Die innerhalb des *Computational Literary Studies*-Programms aufgeworfenen Forschungsfragen sind sowohl aus Perspektive der Literaturwissenschaften als auch der angewandten Informatik hoch innovativ. Die theoretischen Konzepte, welche die Literaturwissenschaft strukturieren (z.B. Autoren, Gattungen, Stil und Epochen), und zentrale Aspekte der literarischen Analyse (wie Handlung, Figurenkonstellation, Redewiedergabe, Erzählperspektive) werden unter Verwendung aktueller Methoden aus der angewandten Informatik untersucht: Die genutzten Methoden und Werkzeuge stammen hauptsächlich aus dem Natural Language Processing, der Computerlinguistik, der Korpuslinguistik, dem Text/Data Mining (einschließlich Guideline-basierter Annotation), dem Machine Learning (bis hin zu Neural Networks) und dem Information Retrieval. Die informatischen Zugänge müssen dabei an die spezifischen Anforderungen der literaturwissenschaftlichen Forschung angepasst werden, wobei eine besondere Herausforderung in der formalen Modellierung literaturwissenschaftlicher Konzepte liegt.

Damit sind die in *Computational Literary Studies* genutzten und entwickelten Forschungsmethoden notwendig interdisziplinär und tragen zum Erkenntnisgewinn in allen involvierten Disziplinen bei. Für die konkrete Forschung innerhalb des Forschungsfeldes bedeutet dies, dass die Fragestellungen von interdisziplinär ausgerichteten Gruppen oder von Forschenden mit Erfahrung in der Kombination von computergestützten Verfahren und literaturwissenschaftlichen Fragestellungen bearbeitet werden sollten.

Neben der Reflexion der Adaption automatischer Verfahren vor dem Hintergrund der literaturwissenschaftlichen Forschungspraxis stellen insbesondere manuelle Verfahren der Tiefenannotation die etablierten Begriffe und Verfahren der literaturwissenschaftlichen Textanalyse erneut auf den Prüfstand. Die Forschung in den *Computational Literary Studies* erfordert damit generell ein grundlegendes Verständnis der jeweils relevanten literaturwissenschaftlichen Methodik und trägt gleichzeitig zur Erweiterung von Methodenspektrum und Theoriebildung bei. Hinzu kommt, dass sich innerhalb der *Computational Literary Studies* bereits eine gewisse interne Struktur mit neuen Forschungsfeldern andeutet (siehe nächster Abschnitt). Diese Entwicklung ermöglicht und erfordert es, die parallel innerhalb dieser Forschungsfelder stattfindende Arbeit zu koordinieren und zu vernetzen. Denn jedes Forschungsfeld weist zwar spezifische und klar abgegrenzte Herausforderungen auf, die Fortschritte in einem einzelnen Feld können aber gleichzeitig von Nutzen für die benachbarten Felder sein und tragen damit zum Fortschritt des gesamten Feldes bei.

5.2 Eingrenzung der wissenschaftlichen Fragestellungen unter Berücksichtigung der Laufzeit eines Schwerpunktprogramms

Das beantragte Programm soll die Forschung zu wesentlichen, formal bzw. computergestützt modellierbaren Dimensionen literarischer Texte fördern. Literarische Texte sind hochkomplex und stellen vorliegende Instrumente vor spezifische Herausforderungen. Daraus ergibt sich bei der Anpassung vorhandener informatischer Anwendungen und Methoden ein Aufwand, der sowohl konzeptionell als auch in Bezug auf konkrete Arbeitsschritte erheblich ist. Obwohl bereits einige Fortschritte innerhalb des relativ jungen Forschungsfeldes erzielt wurden, ist es deshalb sinnvoll, die Forschung zunächst auf die Analyse literarischer Texte zu fokussieren, bevor sie auf andere literaturwissenschaftliche Probleme erweitert wird. Deshalb sollen während der Laufzeit des Schwerpunktprogramms Projekte gefördert werden, die primär die Analyse literarischer Texte beinhalten. Nicht berücksichtigt werden sollen vorerst darüber hinausgehende Fragestellungen wie etwa Methoden zur Modellierung einer breiten Kontextualisierung und Interpretation literarischer Texte, soziologische Perspektiven auf Publikation oder Buchgeschichte, Modelle und Methoden zur Generierung literarischer Texte – wie *digital storytelling* oder *electronic literature* oder auch Simulation –, Fragen nach der Veränderung der literarischen Kommunikation durch die Digitalisierung sowie die (primäre) Digitalisierung von Texten selbst.

Im Schwerpunktprogramm sollen Projekte gefördert werden, die literarische Texte zum Gegenstand haben und digitale Methoden der Annotation, Beschreibung und Analyse einsetzen. Für das Programm relevant sind deshalb u.a. folgende Forschungsfelder:

- Gattungsanalyse: Interne Differenzierung der Großgattungen (Epik, Drama, Lyrik), maschinell verwertbare textuelle Aspekte innerhalb von Gattungen; Modellierung prototypischer Eigenschaften; historische Veränderung von Gattungen und Effekte von Hybridisierung; Verhältnis von literarischen zu nicht-literarischen Texten.
- Stilanalyse: Identifizierung zentraler stilistischer Merkmale wie rhetorische Figuren (z.B. Parallelismus) oder bildhafte Sprache (z.B. Vergleiche und Metaphern); Analyse textueller Ähnlichkeit aufgrund von lexikalischen, morpho-syntaktischen und syntaktischen Eigenschaften oder Mustern.
- Figurendarstellung: Extraktion und Repräsentation figurenbezogener Informationen, wie Figureneigenschaften (z.B. “schön”, “gefährlich”) oder Beziehungen zwischen Figuren (z.B. Mutter-Tochter-Beziehung). Dies schließt dynamische Informationen mit ein, etwa die Veränderung von Merkmalen, Emotionen oder Beziehungen im Handlungsverlauf

(z.B. durch Hochzeit, Tod), ebenso wie die Markierung von Identität und Differenz in der Figurengestaltung.
- Figurennetzwerke: Extraktion aufgrund komplexerer Interaktionsformen, z.B. der Kommunikationsstruktur, sowie die Nutzung von Eigenschaften von Figuren und Figurenrelationen für die Informationsanreicherung von Netzwerken.
- Handlung: Modellierung von Handlungselementen (z.B. Ereignissen) und Handlungsverläufen (z.B. Happy End), insbesondere in Erzählungen und Dramen unter Verwendungen verschiedener Methoden wie z.B. *event alignment* in vergleichbaren Texten.
- Textsegmentierung: Identifizierung von Szenen- oder Episodengrenzen; Identifizierung und Klassifikation von Formen der Figurenrede, der Gedankenwiedergabe etc. (z.B. indirekte Rede); Abgrenzung von narrativen, deskriptiven und argumentativen Abschnitten.
- Metrum und Prosarhythmus: Klassifizierung und Modellierung metrischer Strukturen von Versen (in Dramen und Lyrik) sowie von Prosarhythmus (in Erzähltexten).
- Themen / Motive: Erkennung von prominenten Motiven und Themen, ihrer synchronen Verteilung und der Entwicklung, z.B. mit *Topic Modeling*.
- Literaturgeschichte: Identifikation von Langzeitentwicklungen und Trends in literarischen Textgruppen (z.B. gruppiert nach Gattungen oder Verfahren der Figurendarstellung) und quantitative Belege zur literarischen Klassifikation nach Epochen (z.B. Entwicklung von Romantik hin zu Realismus/Naturalismus)

Im Kontext eines oder mehrerer dieser Forschungsfelder können Verfahren für die Tiefenannotation der Texte zum Einsatz kommen, die neue Zugänge entwickeln und umsetzen. Dabei sollte die algorithmische Implementierung der Annotation durch Machine Learning-Verfahren angestrebt werden, ebenso wie die Rückbindung der erarbeiteten Konzepte und Verfahren an die literaturwissenschaftliche Begriffsbildung und Methodenentwicklung. Zusätzlich sollte die Nachnutzbarkeit der Verfahren und Annotationen in weiteren Projekten, insbesondere auch anderen *Computational Literary Studies*-Projekten, möglich sein.

Die genannten Forschungsfelder decken unterschiedliche Aspekte literarischer Texte und eine Vielzahl relevanter Methoden ab. Der oben erläuterte Fokus auf die Analyse literarischer Texte ermöglicht außerdem eine Vernetzung und nachhaltige Organisation der geförderten Projekte, von der Synergieeffekte für das gesamte Feld zu erwarten sind. Dazu gehört auch ein Beitrag zur Entwicklung und Bereitstellung wichtiger Ressourcen sowie zur Anreicherung von Infrastruktur, der von den geförderten Projekten erwartet wird:

- Aufbau und Kuration von Korpora und Textsammlungen zu Forschungsfragen im Kontext eines Projekts. Dies umfasst auch manuelle oder teilautomatisierte Annotation, um Trainingsdaten für Annotationsalgorithmen und weitere Tools zu erzeugen.
- Entwicklung von Algorithmen, Methoden, Werkzeugen und anderen Ressourcen, deren Nachnutzbarkeit in weiteren Kontexten relevant ist. Insbesondere geht es um Ressourcen, die zur Analyse weiterer Textsammlungen genutzt werden können oder das Potenzial haben, auch literarische Phänomene höherer Ordnung zu identifizieren (z.B. ein Werkzeug zur Annotation direkter Rede, dessen Ergebnisse zum Aufbau von Figurennetzwerken genutzt werden können).
- Aufbau von Infrastrukturkomponenten zur Entwicklung, Publikation und Nachnutzung von Datensammlungen und Werkzeugen, wobei auch die Komponenten selbst wiederverwendet und erweitert werden können sollen.

Die Bereitstellung von Korpora, Werkzeugen und anderen Ressourcen für die Forschungsgemeinschaft der *Computational Literary Studies* sorgt nicht nur für Transparenz und Reproduzierbarkeit bzw. Nachvollziehbarkeit der Forschung, sondern trägt auch zur Vermeidung unnötiger Mehrarbeit bei und beschleunigt so die Entwicklung des Forschungsfelds.

Die geförderten Projekte können auf die oben skizzierten Vorarbeiten (siehe Abschnitt 2) aufbauen und deshalb zusammengenommen in sechs Jahren substantiell zum jeweiligen Feld beitragen. Die dargelegte Eingrenzung sowie die genannten Synergieeffekte fördern die Übertragung der Erkenntnisse, Methoden und Ressourcen aus der ersten Förderphase (Jahre 1–3) in die zweite Phase (Jahre 4–6), sowohl in Form von Folgeprojekten als auch als neue

Projekte, die auf den Erkenntnissen vorhergehender Projekte aufbauen. Nicht zuletzt für den angestrebten Identifizierungsprozess von sachlichen und methodischen Fragestellungen an Schnittstellen zwischen Projekten (in der ersten Förderphase) dürfte die Möglichkeit, in besonders erfolgreichen Fällen die Gelegenheit für ein darauf abgestimmtes Folgeprojekt zu erhalten, sehr motivierend wirken.

5.3 Kohärenz der geplanten Forschungsaktivitäten

Im Rahmen des Schwerpunktprogramms bestimmen wir das Feld der *Computational Literary Studies* so, dass im Zentrum die algorithmenbasierte Textanalyse sowie die Tiefenannotation literarischer Texte etwa für Goldstandards steht. Auf diese Weise wird in doppelter Hinsicht eine Kohärenz der Forschungsaktivitäten gewährleistet: durch den Gegenstand und die Methodik. Das SPP konzentriert sich auf literarische Texte und deren Eigenheiten, die diese Texte untereinander und vor allem von Gebrauchstexten unterscheiden. Dabei soll die Konzentration auf literarische Texte eine Vielfalt der Sprachen nicht ausschließen: Projekte zu Literaturen verschiedener Sprachen sind dezidiert erwünscht, um durch den Vergleich weitere Erkenntnisse gewinnen zu können. Den Zusammenhang zwischen den Projekten sichern neben dem Gegenstand ‘Literatur’ die Methoden und die Begriffsreflexion. Gemeinsam ist den Methoden, dass sie Texte und deren Annotationen analysieren, während wir etwa Fragen nach der Simulation des Vortrags literarischer Texte oder der Videoanalyse solcher Darstellungen (Theater, Poetry Slam usw.) vorläufig ausklammern. Dafür sprechen eine Reihe von Gründen: Für die performativen Aspekte von Dramen gibt es eine eigene Disziplin, die Theaterwissenschaft, die Analyse und Simulation der Performanz stellt methodisch ganz andere Anforderungen, und sie kann noch nicht auf entsprechende Korpora zurückgreifen wie die textbasierten Analysen. Eine in jedem Projekt vorgenommene und in Treffen diskutierte Reflexion der aus verschiedenenen Disziplinen übernommenen Begriffe soll ebenfalls den Zusammenhang stärken. Die Kohärenz, die die genannte Konzentration auf literarische Texte und Textanalysemethoden erzeugt, wird auch organisatorisch bei der Gestaltung der interdisziplinären und ortsübergreifenden Zusammenarbeit bzw. Netzwerkbildung berücksichtigt. Die Projekte werden innerhalb des SPP entlang der beiden Achsen Gegenstand und Methode gruppiert, um einen koordinierten Austausch bei gleichzeitiger inhaltlicher Fokussierung zu ermöglichen (s. Abschnitt 5.4). Durch eine Gruppierung nach methodischen Aspekten können auch zwischen Projekten zu Texten verschiedener Sprachen Kooperationsmöglichkeiten aufgedeckt werden.

5.4 Konzepte zur Gestaltung der interdisziplinären und ortsübergreifenden Zusammenarbeit/Netzwerkbildung

In der Informatik sind standortübergreifende, am Forschungsbedarf orientierte Team-Kooperationen etablierte Praxis, eine SPP-übergreifende Herausforderung besteht jedoch in der Abstimmung der Informatik-Methoden mit den literaturwissenschaftlichen Forschungspraktiken. Innerhalb des literaturwissenschaftlichen Gegenstandsbereichs mag die Identifikation von “Schnittstellen” für übergreifende Teilfragen – etwa zwischen einer mediävistischen Forschungsagenda zu epischen Verstexten und einem Projekt zu zeitgenössischen englischen Erzähltexten – vielen zunächst weniger naheliegend erscheinen; erst indem jedoch Modellierungsideen und -erfolge aus *einem* Projektkontext, auch sehr punktuell, in Relation zu einem *anderen* Kontext gestellt werden, kann die Diskussion möglicher Generalisierungen eröffnet werden.

Die geplanten Maßnahmen zur Förderung der Zusammenarbeit zwischen den Einzelprojekten lassen sich in drei Kategorien zusammenfassen:

- Organisation der Teilprojekte in (i) gegenstands- und (ii) methodenbezogene Projektcluster,
- Maßnahmen zur Koordination und Vertiefung der Zusammenarbeit zwischen den Teilprojekten bzw. zur Dissemination von Methoden und Ergebnissen in die weitere Community
- Standardisierung von Abläufen und Daten, um die Nachnutzbarkeit von Methoden, Daten und Ergebnissen zu gewährleisten.

Die **Projektcluster** dienen der Förderung des Austausches zwischen Forschergruppen mit gemeinsamen Interessen. Es ist geplant, den Forschungsschwerpunkten entsprechend je drei gegenstands- und methodenbezogene Cluster zu bilden. Methodenorientiert bietet sich eine Gruppierung z.B. in Netzwerk-, Annotations- und distributionelle Semantik-Projekte an, gegenstandsbezogen lassen sich Projekte z.B. nach behandelten Textsorten, Epochen und Sprachräumen bündeln. Welche Kriterien sich als sinnvoll erweisen, wird von den aufgenommenen Projekten abhängen. Jedes Projekt wird jeweils zwei Clustern zugeordnet: einem gegenstands- und einem methodenorientierten. Angestrebt sind hierbei Cluster aus 4-5 Teilprojekten. Pro Förderphase sollen jeweils drei der sechs Cluster durch ein thematisch passendes Mercator-Fellowship verstärkt werden (siehe Abschnitt 5.6). Zur Koordination der laufenden Projekte finden im Jahresrhythmus Treffen sowohl aller Projekte als auch innerhalb der Cluster statt. Um die Terminfindung zu vereinfachen und Reisekosten zu reduzieren, werden die Clustertreffen jeweils am Vortag des Gesamttreffens organisiert.

Events spielen für die Netzwerkbildung eine zentrale Rolle. Das Schwerpunktprogramm strebt darum sowohl einen hohen Grad an Präsenz bei etablierten regelmäßigen Konferenzen der Fachcommunities als auch die Organisation eigener events an. Als Beteiligung an etablierten Veranstaltungen wird sich das Schwerpunktprogramm regelmäßig bei den jährlichen Konferenzen der ADHO, der ACL und der DHd, bei der European Summer School, sowie auf geeigneten fachwissenschaftlichen Tagungen für die Ausrichtung eigener Workshops bewerben. Vor dem Start der beiden Förderphasen wird das SPP jeweils ein für alle Interessierten offenes Treffen organisieren, auf dem Projektideen präsentiert und ausgetauscht werden. Zum Ende der beiden 3-jährigen Förderphasen wird jeweils eine Abschlusskonferenz mit allen Teilprojekten organisiert, deren Beiträge dann in einem Sonderband der *Zeitschrift für digitale Geisteswissenschaft* publiziert werden. Eine ebenfalls sehr direkte Möglichkeit zur Vernetzung mit der Zielgruppe des SPP bieten zudem die Treffen der COST Action 'Distant Reading for European Literary History', mit der das Schwerpunktprogramm personelle Überschneidungen hat. Darüber hinaus wird das SPP Shared Tasks organisieren, um die Entwicklung von Methoden von projektübergreifendem Interesse (z.B. Erkennung von Metaphern) zielgerichtet zu fördern und zusätzliche Gelegenheiten zur Vernetzung zu schaffen.

Um zusätzliche Reichweite bei der **Dissemination** von Ergebnissen und Methoden aus dem Schwerpunktprogramms zur erreichen, wird laufende Arbeit durch monatliche Beiträge auf einem gemeinsamen Blog einer breiten Öffentlichkeit zugänglich gemacht. Die Blogbeiträge werden dafür reihum von den geförderten Projekten erstellt, wobei Querverbindungen zwischen den Projekten explizit adressiert und durch eine geeignete Verschlagwortung der Beiträge verdeutlicht werden sollen. Im Sinne des *mixed methods*-Ansatzes publizieren die Projektbeteiligten außerdem in einschlägigen Organen ihrer Fachcommunity und beteiligen sich sowohl mit Beiträge zu den erzielten Ergebnissen als auch mit Workshops zur Vermittlung der entwickelten Methoden an einschlägigen Fachtagungen. So werden die Ergebnisse und Verfahren, die innerhalb des Schwerpunktprogramms erarbeitet und entwickelt werden, im Sinne der disziplinären Anschlussfähigkeit an die jeweiligen Fachcommunities zurückgebunden.

Der dritte Eckpfeiler des Konzepts zur Gestaltung der interdisziplinären und ortsübergreifenden Zusammenarbeit sind **gemeinsame Standards** für Daten, Metadaten und Annotationen. Obgleich dies ohnehin den Anforderungen an ein modernes Forschungsdatenmanagement entspricht, kommt der koordinierten Aufbereitung der Daten im Kontext des SPP eine Schlüsselstellung zu: hier soll sich das angestrebte Mitdenken von **gegenstands- und methodenorientierten Schnittstellen** manifestieren, wie zu Beginn des Abschnitts charakterisiert. Um einen sachlichen oder methodologischen Austausch zu befördern, streben die Projekte an, zu jeweils abgestimmten bzw. dynamisch abzustimmenden Teilfragen Datensammlungen für einen projektübergreifenden Austausch zu erschließen. Entsprechend sind die Forschungsgegenstände aller bewilligten Teilprojekte gemäß etablierter Standards aufzubereiten bzw. vorhandene Textsammlungen zu konvertieren. Hierbei kann einerseits auf die DARIAH- und CLARIN-Infrastruktur und das vorhandene Know-How zurückgegriffen werden. Gleichzeitig kann mit der Entwicklung neuartiger Begriffsformalisierungen, Analysetechniken, Arbeitspraktiken etc. einhergehen, dass auch beim Datenmanagement Neuland betreten werden muss.

Auch erfordert eine risikobereite Exploration von möglichen Pfaden eine Balance zwischen nachnutzungsorientierten Verpflichtungen und der notwendigen Leichtfüßigkeit im Forschungsprozess.

Der Programmausschuss hat in der technischen Unterstützung eines Schnittstellen-Denkens, das Einzelprojekte nicht einengt, gleichzeitig jedoch zur praktischen Umsetzung von übergreifenden Gedanken einlädt, eine zentrale Aufgabe erkannt, die über einen bloßen Einsatz von Infrastrukturangeboten hinausgeht. Daher soll eine **zentrale Koordinationsstelle für Datenmanagement** für alle SPP-Projekte im Umfang einer Vollzeitstelle (Postdoc) geschaffen werden. Darüber hinaus werden vor allem die Hilfskräfte, die mit der Aufarbeitung der Daten betraut sind, in zentralen Schulungen in der Anwendung der gemeinsamen Standards ausgebildet. Die ausgeprägte Vernetzungsstruktur des SPP erfordert zudem eine zentrale organisatorische Vollzeitstelle, um die 15 Projekte, die Clusteraktivitäten und die Dissemination der Ergebnisse zu koordinieren. Den hohen Organisationsaufwand sollen nicht die Forschenden leisten. Bei der **organisatorischen Koordinationsstelle** laufen alle nicht primär datenbezogenen Informationen des SPP zusammen. Die Person in dieser integrierenden Funktion muss nicht nur über organisatorische, sondern auch über sachliche Kompetenzen verfügen (Postdoc). Um das Wissensgefälle bezüglich der Anwendung von Standards unter den Forschenden in den Teilprojekten auszugleichen, werden Mittel für die Erstellung geeigneter Online-Lehrmaterialien durch Hilfskräfte beantragt.

5.5 Maßnahmen zur Förderung des wissenschaftlichen Nachwuchses, Förderung von Wissenschaftlerinnen, Angebote zur Familienfreundlichkeit

Nachwuchsförderung: Um eine nachhaltige Entwicklung für ein *emerging field*, wie die *Computational Literary Studies*, sicherzustellen, ist die frühe Einbindung und Förderung des wissenschaftlichen Nachwuchses zentral. Dies wurde bereits bei der Vorbereitung des Antrags durch die Zusammensetzung des Programmausschusses berücksichtigt. Durch die Einbindung einer Nachwuchswissenschaftlerin sind die Perspektiven und Bedarfe des wissenschaftlichen Nachwuchses fest in der Organisation und Umsetzung des Schwerpunktprogramms integriert. Darüber hinaus erhofft sich der Ausschuss, dass dies als Signal gewertet wird, dass die Beteiligung des Nachwuchses am Programm explizit erwünscht ist.

Die konkrete Nachwuchsförderung innerhalb des SPP findet durch die Organisation und Finanzierung von 12 Forschungsaufenthalten in einem der Projekte statt. Die Dauer der voll finanzierten Forschungsaufenthalte beläuft sich hierbei auf bis zu 3 Monate. Den Teilnehmenden werden so Einblicke in die Forschungspraxis ermöglicht, die zu methodischer Sicherheit und dem Ausbau der Fähigkeit, eigene Fragestellungen zu formulieren, führen sollen. Zusätzlich bietet das Programm die Chance, eigene Netzwerke innerhalb der Wissenschaftsgemeinschaft aufzubauen. Es ist ausdrücklich erwünscht, dass so Geförderte in der anschließenden Förderphase entweder einer bestehenden Forschungsgruppe beitreten oder ihr eigenes Projekt einreichen. Alle Teilnehmer/innen werden zum nächsten Jahrestreffen der Projektgruppen eingeladen. Dort können sie sich in einem eigens eingerichtetem Kolloquium über Erfahrungen und Zukunftspläne austauschen. Darüber hinaus verpflichten sich die Projekte, Kurse im Rahmen einer Summer School für ein breiteres Publikum anzubieten. Denkbar ist sowohl die Organisation einer eigenen Veranstaltung als auch der Anschluss an ein bestehendes, etabliertes Angebot. Im Vorlauf zur ersten Förderphase ist ein Treffen zur Identifizierung gemeinsamer Interessen und anschließender Gründung von Projektgruppen geplant. Um eine Teilnahme des wissenschaftlichen Nachwuchses zu erleichtern, vergibt der Programmausschuss in einem kompetitiven Verfahren 12 Reisekostenstipendien.

Förderung von Wissenschaftlerinnen: Das Schwerpunktprogramm setzt an zwei Zeitpunkten in der Karriere von Wissenschaftlerinnen an, um diese gezielt zu fördern. Die erste Gruppe sind Wissenschaftlerinnen am Anfang ihrer Laufbahn. Sie sollen dezidiert zur Teilnahme an Forschungsaufenthalten aufgefordert werden. Die Forschungsaufenthalte sollen möglichst zur Hälfte an Wissenschaftlerinnen vergeben werden. Darüber hinaus soll, wenn Bedarf besteht, in den Projekten arbeitenden jungen Wissenschaftlerinnen die Teilnahme an geeigneten Pro-

grammier-Kursen finanziert werden. Die zweite Maßnahme richtet sich an Wissenschaftlerinnen, die bereits eine Karriere begonnen haben. Diese werden zu einem einmal jährlich stattfindenden Coaching-Workshop eingeladen, um sich mit weiblichen Führungskräften mit akademischem Hintergrund aus Wissenschaft und Wirtschaft auszutauschen. Ziel ist es, eine Fortsetzung der akademischen Laufbahn zu erleichtern, indem zum einen Vorbilder geschaffen, zum anderen Karriereoptionen aufgezeigt werden.

Familienfreundlichkeit: Sämtliche vom Programmausschuss organisierte Veranstaltungen bieten die Möglichkeit einer kostenlosen Kinderbetreuung, um Wissenschaftlerinnen und Wissenschaftlern mit jungen Familien die Teilnahme zu erleichtern.

5.6 Vernetzung der geplanten Forschungsaktivitäten im int. Wissenschaftssystem

Zur Kommunikation von Ergebnissen des SPP in die beteiligten Fächer sollen vor allem deren etablierten Veranstaltungen und Publikationskanäle dienen. Hierbei spielen vor allem die jährlichen Verbandskonferenzen eine zentrale Rolle, auf denen das SPP regelmäßig mit seinen Forschungsergebnissen präsent sein wird. Für die Philologien sind das vor allem die Tagungen der Fachverbände, für die Digital Humanities die jährlichen Konferenzen der DHd und der ADHO, für die Computerlinguistik die Veranstaltungen der ACL und für die Informatik z.B. die Tagungen der Gesellschaft für Informatik oder die NIPS (Neural Information Processing Systems).

Der Programmausschuss des SPP ist bereits über diverse Organisationsformen direkt mit einigen der betreffenden Organisationen vernetzt. Dazu zählen im Bereich der Philologien die AG “Digitale Romanistik” im Deutschen Romanistenverband (DRV, Christof Schöch)[10], und die AG “Digitale Germanistik” im Germanistenverband (Fotis Jannidis). Im Bereich der Digital Humanities kommen die COST Action “Distant Reading for European Literary History”[11] (Christof Schöch und Fotis Jannidis), die die Vernetzung digital arbeitender Literaturwissenschaftler vor allem auf Europäischer Ebene fördern soll, die ADHO Special Interest Group "Digital Literary Stylistics" (SIG-DLS, Evelyn Gius, Nils Reiter, Fotis Jannidis und Christof Schöch)[12], die der Stilometrie-Community auf internationaler Ebene einen organisatorischen Rahmen gibt, die “Federation of Stylometry Labs” (FoSL, Würzburger Lehrstuhl für Computerphilologie)[13], und die DARIAH-EU Working Group ‘Text and Data Analytics’ (DARIAH-TDA, Fotis Jannidis)[14], die sich der Verbreitung von methodischem Know-How im Bereich der computergestützten Textanalyse für geisteswissenschaftliche Fragestellungen widmet. Eine Verbindung in die Fachinstitutionen der Computerlinguistik besteht über die ACL "Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities" (SIGHUM, Jonas Kuhn, Nils Reiter)[15]. Insbesondere durch die engen Verbindungen zur FoSL, zur ADHO-SIG, zur DARIAH-WG und zur COST Action bieten sich mit diesen Organisationen auch direkte Kooperationen an, z.B. in Form gemeinsam ausgerichteter Veranstaltungen oder Shared Tasks. Durch gemeinsame Veranstaltungen mit den europäischen Infrastrukturen für digitale Geisteswissenschaften, CLARIN und DARIAH, werden diese mit einbezogen.

Als weiteres Mittel zur gezielten Vertiefung der internationalen Vernetzung des SSP dienen **Mercator-Fellowships**. International renommierte Wissenschaftlerinnen und Wissenschaftler sollen innerhalb der Cluster die gegenstandsbezogene und methodische Forschung aus projektübergreifender Perspektive verstärken (siehe Abschnitt 5.4). Die Auswahl der Mercator-Fellows erfolgt durch den Programmausschuss in einem kompetitiven Verfahren auf der Basis von Vorschlägen aus den Clustern. Ausgewählt wird nach inhaltlichen Kriterien, aber mit dem erklärten Ziel, ein ausgeglichenes Geschlechterverhältnis zu erreichen. Geplant sind sechs Mercator-Fellowships für 4-6 Wochen, die auf je drei pro Förderphase aufgeteilt werden.

---

[10] http://www.deutscher-romanistenverband.de/der-drv/ag-digitale-romanistik/
[11] http://www.cost.eu/COST_Actions/ca/CA16204
[12] https://dls.hypotheses.org/
[13] http://fosl.ijp.pan.pl/
[14] http://www.dariah.eu/activities/working-groups.html
[15] https://sighum.wordpress.com

6 Abgrenzung zu anderen laufenden Programmen

Das Feld der *Computational Literary Studies* konnte in den letzten Jahren vom Aufbau nationaler und internationaler Infrastrukturen profitieren, die auf die Digitalisierung der Geistes- und Kulturwissenschaften ausgerichtet sind. Hier sind besonders die BMBF-geförderten Projekte DARIAH-DE (einschließlich Textgrid) und CLARIN-D zu nennen. Im Zentrum dieser Projekte steht die Bereitstellung von Textsammlungen, z.B. das Deutsche Textarchiv oder das Textgrid-Repository, sowie von Werkzeugen des Natural Language Processing, z.B. CLARINs Weblicht, fürs Edieren, z.B. Textgrid, oder für das allgemeine Projektmanagement. Ein zentraler Beitrag wird von beiden Projekten auch im Bereich der Dissemination von DH-Kompetenzen geleistet. Diese Infrastruktur-Elemente stehen nicht im Mittelpunkt des Schwerpunktprogramms, der dezidiert auf die Methodenentwicklung und -anwendung fokussiert ist. Dies unterscheidet das beantragte Programm auch von den BMBF-Zentren im Bereich Digital Humanities, von denen hier insbesondere CRETA (Stuttgart) und KALLIMACHOS (Würzburg) zu nennen sind. Diese Zentren sind ebenfalls vor allem darauf ausgerichtet, eine Infrastruktur aufzubauen. Das Schwerpunktprogramm kann von den hier genannten Aktivitäten nur profitieren, da auf diese Weise Texte und Werkzeuge zur Verfügung gestellt werden, die zur beschleunigten Emergenz des Feldes beitragen.

7 Qualifikation der Koordinatorin/des Koordinators

Fotis Jannidis forscht seit 20 Jahren in den Digital Humanities und seit rund sieben Jahren im Feld der quantitativen Analyse von Literatur (siehe oben, 2.2 und 3). Er hat an mehreren institutionsübergreifenden Projekten teilgenommen. Dabei hat er zum Teil leitende bzw. koordinierende Positionen ausgefüllt, namentlich die Leitung eines Arbeitspaketes in TextGrid (2006-2015), die Leitung eines Projektclusters in DARIAH-DE (seit 2011) sowie die Co-Leitung des Projekts “Digitale Faustedition” (2009-2014). Er ist im Leitungsgremium des BMBF-geförderten Projekts ‘KALLIMACHOS. Zentrum für digitale Edition und quantitative Analyse’. Zusammen mit Mike Kestemont (Antwerpen) leitet er die Arbeitsgruppe ‘Text and Data Analytics’ von DARIAH-EU (seit 2016). Im Management Committee der COST Action 16204 ‘Distant Reading for European Literary History’ ist er einer der beiden Vertreter Deutschlands.