Marcus Willand, Evelyn Gius und Nils Reiter

SANTA: Idee und Durchführung

Zusammenfassung: In diesem Kapitel werden die grundsätzlichen Überlegungen, das Evaluationsschema sowie die Ergebnisse des ersten *shared tasks* in den und für die Digital Humanities vorgestellt. Der *shared task* hat die Erstellung von Annotationsrichtlinien für Erzählebenen zum Ziel und fand im Jahr 2018 mit acht teilnehmenden Teams statt. Das Evaluationsschema, das bei *shared tasks* eine zentrale Rolle einnimmt, wurde spezifisch für diesen *shared task* entwickelt und kombiniert die Dimensionen ‚Theorieabdeckung‘, ‚Anwendbarkeit‘ und ‚Nützlichkeit‘ mit der Messung von *inter-annotator agreement* und konnte herausarbeiten, dass es möglich ist, komplexe geisteswissenschaftliche Konzepte durch Annotationsrichtlinien intersubjektiv anwendbar zu machen. Die Gewinner-Richtlinie des *shared tasks* zeichnet sich dadurch aus, dass sie eine gute Balance zwischen den verschiedenen Zielen und Anforderungen findet.

Abstract: This chapter presents the basic considerations, the evaluation scheme and the results of the first *shared task* in and for digital humanities. The *shared task* aims to create annotation guidelines for narrative levels and has been able to attract eight participating teams in 2018. The evaluation scheme combines the dimensions ‘conceptual coverage’, ‘applicability’ and ‘usefulness’ with the measurement of inter-annotator agreement and was able to show that it is possible to make complex humanities concepts intersubjectively applicable through annotation guidelines. The winning guideline of the shared task is characterised by finding a good balance between the different goals and requirements.

1 Einleitung

In diesem Beitrag wird mit dem *shared task* SANTA (*Systematic Analysis of Narrative Texts through Annotation*) eine Aktivität vorgestellt, in die viele CRETA-Ideen eingeflossen sind und an der auch eine Reihe von CRETA-Mitgliedern teilgenommen hat. Ziel war es, die Methodenentwicklung für die automatische Erkennung

Anmerkung: Das Kapitel ist eine deutsche Übersetzung und Überarbeitung von: Evelyn Gius, Nils Reiter und Marcus Willand, Hrsg. (2019). *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*.

Marcus Willand, Germanistisches Seminar, Universität Heidelberg

Evelyn Gius, Institut für Sprach- und Literaturwissenschaft, TU Darmstadt

Nils Reiter, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Open Access. © 2020 Marcus Willand, Evelyn Gius und Nils Reiter; publiziert von De Gruyter Diesem Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Lizenz.
<https://doi.org/10.1515/9783110693973-016>

von Erzählebenen mittels eines *shared task* anzustoßen. Gleichzeitig soll der Annotationsprozess genutzt werden, um Erkenntnisse über die annotierten Phänomene zu sammeln und in die theoretische Ebene zurückzuspiegeln.

Da bisher für Erzählebenen keine nutzbaren Annotationen vorliegen, müssen diese Annotationen zunächst erstellt werden, wozu es wiederum Annotationsrichtlinien braucht.¹ Das Erstellen von Annotationsrichtlinien steht also zunächst im Fokus, wobei hierfür das Konzept eines *shared task* in Abwandlung verwendet wurde.

2 Motivation

Mit diesem *shared task* werden zwei für die Digital Humanities und computationally Literaturwissenschaft zentrale Herausforderungen angegangen: Die Arbeitsteilung in Digital-Humanities-(DH)-Projekten inklusive der Schnittstellen zwischen den Arbeitsbereichen und die intersubjektive manuelle und robuste automatische Erkennung von Erzählebenen in narrativen Texten.

2.1 Aufteilung von Arbeit, Kompetenzen und Teilaufgaben

Beim derzeitigen Stand der automatischen Analyse von narrativen Texten müssen Digital-Humanities-Projekte, die inhaltsbezogene Aspekte solcher Texte in großem Maßstab analysieren wollen, technisch-methodischen Fortschritt erzielen, da viele relevante Phänomene mit bestehenden Verfahren noch nicht – oder nur unzuverlässig – gefunden werden können. Viele solcher Projekte sind also Kooperationsprojekte zwischen Forscher*innen aus den Computerwissenschaften (inklusive der Computerlinguistik) und den Literatur- oder Kulturwissenschaften. Auch wenn die Zahl der Tutorials, angebotenen Sommerschulen und Handbücher in den letzten Jahren massiv gestiegen ist, bleibt die tägliche, konkrete Organisation solcher Projekte eine Herausforderung:²

Eine der ersten Aufgaben, die ein interdisziplinäres Digital-Humanities-Projekt zu lösen hat, ist die Entwicklung einer gemeinsamen Sprache und eines gemeinsamen Verständnisses für den Untersuchungsgegenstand. Denn manch-

¹ Vgl. die Beiträge von Pagel et al. (2020), Pichler und Reiter (2020) und Reiter (2020) in diesem Band.

² Einige davon werden auch im vorliegenden Band explizit diskutiert, etwa im Beitrag von Reiter et al. (2020) ab Seite 467.

mal interessieren sich die beteiligten Computerwissenschaftler*innen ausschließlich für den methodischen Anteil, ohne die erzielten Ergebnisse in irgendeiner Art zu interpretieren. Geisteswissenschaftler*innen fokussieren hingegen viel stärker auf konzeptuelle oder begriffliche Fragen oder auf die Interpretation der Resultate, ohne die Methode explizit zu reflektieren. Innerhalb eines Projektes verfolgen die Beteiligten also oft unterschiedliche Ziele.

Wir sind überzeugt, dass *shared tasks* insbesondere in interdisziplinären Bereichen wie den Digital Humanities sehr viel Potenzial haben, eine gute Schnittstelle und Mittlerposition zwischen den unterschiedlichen Interessen einzunehmen. In einem *shared task* können sich die Beteiligten auf das konzentrieren, was sie am besten können. Literaturwissenschaftler*innen konzentrieren sich auf das literarische Phänomen, das sie beschäftigt und mit dem sie sich auskennen.³ Durch ihre breite Textkenntnis und Interpretationsexpertise sind sie bestens dafür qualifiziert, das Phänomen zu definieren und zu exemplifizieren, ohne sich über Implementierungsdetails oder überhaupt die Implementierbarkeit eines Find-Tools Gedanken machen zu müssen. Das Phänomen kann dann so komplex und kontextabhängig definiert sein, wie es nötig ist – die Frage der Möglichkeiten und Grenzen potenzieller Implementierungen kann hingegen bei der Konzeptarbeit ignoriert werden. Trotzdem können die Definitionen nicht beliebig komplex werden, da sie immer noch intersubjektiv anwendbar sein müssen – allerdings von Menschen und nicht von Computern. Sind die Konzepte so operationalisiert, und wurde ein Korpus mit dem Phänomen annotiert, kann jede*r Computerwissenschaftler*in an der Entwicklung von Methoden für die Erkennung des Phänomens arbeiten. Dazu sind dann keinerlei Kenntnisse in Literatur oder narrativer Texte nötig, da eine für die Implementierung geeignete ‚Wahrheit‘ in Form von Annotationen existiert, gegen die automatische Erkenner evaluiert werden können. Der Komplexität oder Transparenz der Verfahren (z. B. des maschinellen Lernens) sind dann keine Grenzen gesetzt: Da ein vertrauenswürdiger Goldstandard existiert, können alle Verfahren an ihm getestet und überprüft werden. Die in Digital-Humanities-Zugängen normalerweise wichtige Abwägung zwischen Vorhersagekraft und Transparenz/Interpretierbarkeit wäre dann nicht mehr zentral, da die Werkzeuge zuverlässig empirisch getestet werden können. Es können also zur Erkennung die Methoden an den Start gebracht werden, die nötig sind, ohne dass auf die Transparenz oder Erklärbarkeit Rücksicht genommen werden muss.

³ Neben den in diesem Beitrag behandelten Erzählebenen können das z. B. sein: Fokalisierung, Aspekte der Geschichte wie Zeitverlauf, Textelemente wie ‚Exposition‘ oder ‚Katastrophe‘, aber auch interpretative Zuschreibungen, etwa zur Identifikation zentraler Sätze eines Textes.

Die beiden von uns organisierten *shared tasks* decken diese beiden Seiten eines Digital-Humanities-Analyseprojekts ab und haben daher auch unterschiedliche primäre Zielgruppen. Der erste *shared task*, der die Entwicklung von Annotationsrichtlinien zum Ziel hat, wird dabei den Grundstein legen, so dass im zweiten *shared task* eine (annotierte) Textbasis für eine unabhängige und fachwissenschaftlich valide Evaluation der Ergebnisse zur Verfügung steht. Ein Modell, das im zweiten Task gut abschneidet, kann dann verlässlich für neue Texte eingesetzt werden (sofern die Texte den Testdaten ähnlich genug sind).

Durch die Durchführung von zwei *shared tasks* ist die technische von der konzeptuellen Arbeit entkoppelt. So können auch Forscher*innen gewissermaßen miteinander kooperieren, ohne im gleichen Projekt zu sein. Damit werden nicht zuletzt auch die Eintrittshürden in das Feld der Digital Humanities gesenkt.

2.2 Annotationsrichtlinien für Erzählebenen

Das im beschriebenen Vorhaben behandelte Phänomen ist das der Erzählebene. Die Identifikation von Erzählebenen oder zumindest von kohärenten Textsegmenten ist ein notwendiger Schritt für jede Analyse narrativer Texte, die sich nicht nur auf die sprachliche Oberflächenform bezieht. Damit sind z. B. Analysen der Handlung, der Figuren oder der erzählten Welt gemeint. Zum Beispiel muss einer Untersuchung der Interaktionen von Figuren die Erkennung von Erzählebenen vorausgehen – sonst würden etwa Figuren als interagierend gezählt, die einander zwar textlich nah sind, sich aber auf der Handlungsebene nie begegnen. Eine quantitative Untersuchung von solchen Interaktionen ist damit auf die automatische Erkennung von Ebenen angewiesen, die eine Schlüsselkomponente in der computationellen Literaturwissenschaft sind bzw. sein sollten.

Daneben können Erzählebenen auch als eine Art Mediator zwischen einer hermeneutischen und einer automatischen Textanalyse fungieren. Denn ihre Komplexität ist aus literaturwissenschaftlicher Sicht vergleichsweise gering, aus Sicht der natürlichen Sprachverarbeitung hingegen vergleichsweise hoch. Außerdem sind sie potenziell für Textanalysen aller Art relevant. Darüber hinaus sind Erzählebenen im Vergleich zu anderen Phänomenen ein in der Literaturwissenschaft eher wenig umstrittenes Phänomen, schließlich basieren die Definitionen von Erzählebenen in der Regel auf textuellen und narrativen Merkmalen. So können beispielsweise *verba dicendi*, also Verben der Äußerung, und die anschließende direkte Rede textuelle Signale für narrative Ebenen sein oder das Vorhandensein einer anderen ‚Geschichtenwelt‘ kann durch die Analyse des Raumes oder anderer narrativer Phänomene identifiziert werden. Narrative Ebenen sind nicht zuletzt für die Analyse von Texten nützlich, die eine Abweichung zwischen ihrer

textlichen und inhaltlichen Struktur aufweisen, also etwa zwischen Erzähl- und Handlungsreihenfolge.

Erzählebenen sind also eine gute Wahl für ein Phänomen des ersten *shared task* in den Digital Humanities. Ihre wichtigste Eigenschaft ist, dass sie eine Brücke zwischen den theoretischen Diskussionen innerhalb der Narratologie und Anwendungen im *text mining* schlagen können. Moderne automatische Textverarbeitung basiert zu einem großen Teil auf maschinellen Lernverfahren. Angesichts der Verschränktheit von Erzählebenen mit anderen Phänomenen sowie mit Eigenschaften des textuellen Kontexts sind maschinelle Lernverfahren der naheliegendste Ansatz zur automatischen Erkennung von Erzählebenen. Die Qualität dieser Verfahren hängt allerdings von der Verfügbarkeit ausreichender Trainings- und Testdaten ab, die annotiert – also mit Markierungen für Wechsel zwischen Erzählebenen versehen – sind. Sind solche Daten verfügbar, können Modelle trainiert werden, die anschließend zur Erkennung von Ebenen in neuen, nicht annotierten Texten verwendet werden können.

Um diese Annotationen für die Implementierung der Erkennung zu erzeugen, sind Annotationsrichtlinien nötig.⁴ Die Richtlinien stellen nicht nur die Kohärenz der Annotationen sicher, sondern definieren auch, wie mit unklaren Fällen umgegangen werden soll, und erlauben es Nicht-Expert*innen, an der Annotation mitzuwirken. Da Annotationsprozesse teuer und zeitaufwändig sind, ist es allerdings wenig realistisch, für die vielen Varianten eines Konzeptes annotierte Korpora zu erstellen. Daher ist es sinnvoll, einen gewissen Konsens in Bezug auf das theoretische Konzept herzustellen, bevor die Annotationsrichtlinien erstellt werden. Dieser Konsens muss sich dabei nicht einem einzelnen theoretischen Ansatz verschreiben – idealerweise wird versucht, möglichst unterschiedliche Ansätze zu berücksichtigen, so dass verschiedene Anschlussfragen auf Basis des implementierten Konzepts möglich sind. Während die Automatisierbarkeit bei der Entwicklung der Richtlinien keine Rolle spielen sollte, ist die Annotierbarkeit ein wichtiges, aber nicht das einzige Kriterium (s. u.).

Damit die in den Annotationsrichtlinien definierten Konzepte den literaturwissenschaftlichen Ansprüchen an die Ergebnisse der Automatisierung genügen, ist es also ideal, wenn Expert*innen aus der Narratologie sich bereits bei der Entwicklung von Richtlinien einbringen, etwa im Rahmen eines *shared task*.

⁴ Siehe hierzu auch die Beiträge von Reiter (2020) und Ketschik, Blessing et al. (2020) im vorliegenden Band.

3 Der *Shared Task*

3.1 *Shared tasks* in der Computerlinguistik

Shared tasks sind ein etabliertes Forschungsparadigma in der Computerlinguistik. Kernidee eines *shared task* ist, dass verschiedene Teilnehmende an der gleichen, von einem Organisationsteam gestellten Aufgabe arbeiten, also der automatischen Erkennung eines Textphänomens wie z. B. Wortarten. Die verschiedenen Lösungen, d. h. die entwickelten Systeme, werden dann *auf dem gleichen Datensatz* und *mit der gleichen Evaluationsmetrik* evaluiert und sind damit direkt vergleichbar.

Typischerweise läuft ein *shared task* folgendermaßen ab: Das Organisationsteam veröffentlicht einen *call for participation*, in dem die Aufgabe und die bereitgestellten Daten beschrieben werden. Zeitgleich oder kurz danach wird ein annotierter Datensatz zur Verfügung gestellt, den die Teilnehmenden zur Entwicklung ihrer Systeme nutzen können. Der Datensatz ist annotiert, damit die Teilnehmenden ihre Systeme trainieren und intern testen können. Nach einigen Monaten werden die Testdaten publiziert, wobei es sich um einen nicht annotierten Datensatz handelt. Die Teilnehmenden wenden ihre bis dahin entwickelten Systeme auf diesen Datensatz an und schicken die Vorhersagen ihres Systems an das Organisationsteam, das sie dann gesammelt vergleicht und in eine Rangfolge bringt. Am Ende des *shared task* steht meistens ein Workshop, in dem die Systeme diskutiert und mögliche nächste Schritte identifiziert werden.

Historische Entwicklung

Innerhalb der Computerlinguistik haben *shared tasks* ihren Ursprung in den *message understanding conferences* (MUC Sundheim 1993). Ziel dabei war, automatisch Informationen aus Ausschnitten von Zeitungsberichten zu extrahieren. Sundheim und Chinchor (1993) kategorisieren die Fortschritte durch *shared tasks* wie folgt: (i) Als Fortschrittsmessung beschreiben sie, wie gut verschiedene Systeme eine Aufgabe lösen, wenn standardisierte Performanzmetriken angewendet werden. Dadurch wird der aktuelle Forschungsstand direkt ermittelt und transparent, also das System identifiziert, welches zu einem bestimmten Zeitpunkt die besten Ergebnisse liefert. (ii) In der Frage, welche Metrik eigentlich geeignet ist, um die Performanz zu messen, haben *shared tasks* ebenfalls weitergeholfen. Dadurch, dass verschiedene Metriken angewendet und Teil des wissenschaftlichen Diskurses wurden, konnten Stärken und Schwächen verschiedener Metriken offengelegt werden. (iii) Nicht zuletzt generieren die *MUC-shared-tasks* auch Erkenntnisse darüber, warum welche Systeme besser oder schlechter abschneiden.

Durch die Teilnahme und die anschließende Analyse können z. B. bestimmte Lücken identifiziert werden, die ein System nicht schließen kann. Alle drei Aspekte sind in allen computerlinguistischen *shared tasks* mehr oder weniger explizit präsent.

Seit dem Jahr 2000 bietet die *Conference on Natural Language Learning* (CoNLL) ein Dach für eine Reihe von *shared tasks* zu verschiedenen computerlinguistischen Themen: *chunking* (Sang und Buchholz 2000), *clause identification* (Sang und Déjean 2001), sprachunabhängige Erkennung von Eigennamen (Tjong Kim Sang und De Meulder 2003), verschiedene Varianten syntaktischer Analyse, entweder eingeschränkt auf bestimmte Sprachen oder sprach-agnostisch (Buchholz und Marsi 2006; Nivre et al. 2007; Kübler 2008) sowie *semantic role labeling* (Carreras und Màrquez 2004; Carreras und Màrquez 2005). Andere Konferenzen haben das Konzept ebenfalls aufgenommen, z. B. für die Aufgabe, textuelle Inferenzen zu identifizieren (*recognizing textual entailment*, Dagan et al. 2006), die bis 2013 lief. Unter der Bezeichnung ‚SensEval‘ startete im Jahr 2000 eine *shared-task*-Initiative (Kilgariff und Rosenzweig 2000). Diese wird aber mittlerweile etwas allgemeiner ‚SemEval‘ genannt um anzuzeigen, dass man sich generell semantischen Phänomenen widmet. ‚SemEval‘ ist heute vor allem ein Dach für die Organisation von *shared tasks*, unter dem allein im Jahr 2018 zwölf verschiedene *tasks* organisiert wurden.⁵

Als Gründe für die anhaltende Beliebtheit von *shared tasks* in der Computerlinguistik nennen Escartín et al. (2017) neben dem direkten Wettbewerb die Möglichkeit, Systeme direkt vergleichen zu können. Nicht zu unterschätzen ist auch, dass die *shared tasks* eine Reihe von De-Facto-Standards etabliert haben (z. B. das breit verwendete CoNLL-Format für annotierte Daten). Auch die Publikation annotierter Daten, die im Rahmen von *shared tasks* entstanden, haben die Forschungsgemeinschaft vorangebracht, da diese nachnutzbar sind. Direkter Wettbewerb und hohe Sichtbarkeit fördern jedoch auch unethisches Verhalten, wie Escartín et al. (2017) ausführen: Forschungsdaten werden womöglich zurückgehalten oder Teilnehmende sind weniger offen, Details zu ihren Arbeiten zu teilen, um anderen voraus zu sein. Escartín et al. geben auch eine Reihe von Empfehlungen für *shared tasks*, um das Risiko unethischen Verhaltens zu minimieren,

⁵ Affect in Tweets, Multilingual Emoji Prediction, Irony Detection in English Tweets, Character Identification on Multiparty Dialogues, Counting Events and Participants within Highly Ambiguous Data covering a very long tail, Parsing Time Normalizations, Semantic Relation Extraction and Classification in Scientific Papers, Semantic Extraction from Cybersecurity Reports using Natural Language Processing (SecureNLP), Hypernym Discovery, Capturing Discriminative Attributes, Machine Comprehension using Commonsense Knowledge, Argument Reasoning Comprehension Task. Siehe auch: <http://alt.qcri.org/semeval2018/index.php?id=tasks>.

etwa dass explizit von Anfang an geklärt wird, dass die Ergebnisse der teilnehmenden Systeme unter bestimmten Lizenzen zu veröffentlichen sind.

3.2 Zwei verbundene *shared tasks*

Da sich die Forschungscommunities, -praktiken und auch -ziele in der Computerlinguistik und den Literaturwissenschaften stark unterscheiden, konnte das computerlinguistische Modell für *shared tasks* nicht direkt für unser Vorhaben übernommen werden. Wir haben entsprechend mehrere Anpassungen am Verfahren vorgenommen, und es in zwei *shared tasks* aufgeteilt. Die beiden *shared tasks* haben unterschiedliche Ziele, Daten und Zielgruppen, beschäftigen sich aber beide mit dem Phänomen der Erzählebene. Das Ziel des ersten *shared task* ist es, Annotationsrichtlinien für Erzählebenen zu etablieren. Diese werden dann auf ein größeres Korpus angewendet, das im zweiten *shared task* Verwendung als Trainings- und Testkorpus findet. In diesem besteht dann das Ziel in der automatischen Erkennung von Erzählebenen. Damit ist der zweite *shared task* ein ‚normaler‘ *shared task*, der so abläuft, wie oben beschrieben. Im Folgenden wird ein Meilenstein des ersten *shared task* dokumentiert: Die Teilnehmenden haben ihre Annotationsrichtlinien eingereicht und diese wurden in einem dreitägigen Workshop vergleichend evaluiert.

3.2.1 *Shared Task 1: Systematic Analysis of Narrative Texts through Annotation (SANTA)*

Der erste *shared task* stellt die Herausforderungen bei der Konzeptualisierung und Definition von Erzählebenen ins Zentrum, wobei auch die manuelle Anwendung in Texten Berücksichtigung findet. Die Aufgabe für die teilnehmenden Teams war es, Annotationsrichtlinien für das Phänomen der Erzählebene zu entwickeln. Dafür wurden keine theoretischen Vorgaben gemacht. Allerdings wurde eine Liste möglicher Referenzen als Startpunkt für eigene Recherchen bereitgestellt, nebst einer Anleitung zur Entwicklung von Richtlinien. Eine von Reiter (2020) weiter ausgearbeitete Fassung davon findet sich ab Seite 193 in diesem Band.

Damit die entwickelten Richtlinien sich nicht auf zu korpuspezifische Merkmale konzentrieren, wurden sie anschließend auf einem den Teilnehmenden *unbekannten* Korpus getestet. Es wurde ein Entwicklungskorpus publiziert und angekündigt, dass das finale Korpus ähnlich sein würde. Die Teilnehmenden waren also gezwungen, ihre Richtlinien so zu schreiben, dass sie möglichst breit anwendbar sind. Dieses Vorgehen ist von der Aufteilung in Trainings- und Test-

daten, die bei maschinellen Lernverfahren verwendet wird, inspiriert (cf. Witten und Frank 2005, S. 144 f.). Hier wie dort soll dadurch sichergestellt werden, dass die entwickelten Systeme bzw. Richtlinien so generisch wie möglich sind und von den konkreten Daten abstrahieren.

Korpus

Das Korpus wurde so zusammengestellt, dass es möglichst viele Aspekte des Phänomens abdeckt. Repräsentativität war kein Entscheidungskriterium. Die Texte im Korpus unterscheiden sich in Bezug auf Genre, Publikationszeitraum und Länge.⁶ Alle Texte wurden auf Deutsch und Englisch bereitgestellt, teilweise handelt es sich um Übersetzungen aus einer dritten Sprache.

Wir haben festgelegt, dass kein Text länger als 2000 Wörter sein sollte. Damit sich daraufhin nicht nur sehr kurze Texte im Korpus befanden (was sich potenziell auf die Erzählebenenstruktur hätte auswirken können), wurden auch längere Texte in gekürzter Form aufgenommen. Bei den Kürzungen wurde darauf geachtet, dass die Ebenenstruktur davon nicht betroffen war, wobei die Konzeptionen, die den Teilnehmenden in Referenzen zur Verfügung gestellt worden waren, als Entscheidungsgrundlage dienten. Insgesamt wurden 17 Texte als Teil des Entwicklungskorpus zur Verfügung gestellt, das Testkorpus bestand aus acht Texten. Die Tabellen 1 und 2 zeigen die Texte sowie Metadaten. Alle Texte sind frei verfügbar und können über das GitHub repository⁷ bezogen werden.

Parallelannotationen

Die Messung des sog. *inter-annotator agreements* (IAA) ist ein gängiger Weg, um die intersubjektive Anwendbarkeit von Annotationsrichtlinien zu beurteilen. Um das IAA zu berechnen, muss der Text von mehreren Annotierenden unabhängig voneinander, aber mit der gleichen Richtlinie, annotiert werden.

Im Rahmen des *shared task* wurde dies umgesetzt, indem jedes teilnehmende Team gebeten wurde, das Testkorpus nach der eigenen und nach einer von dem Organisationsteam zugewiesenen fremden Richtlinie zu annotieren. Zusätzlich wurde eine Gruppe studentischer Hilfskräfte dafür bezahlt, die Richtlinien ebenfalls auf die gleichen Texte anzuwenden. Insgesamt wurde also jeder Text dreimal nach der gleichen Richtlinie annotiert. Tabelle 3 zeigt eine Übersicht der verschiedenen Annotationstypen.

⁶ Genres: Anekdote, Fabel, Märchen, Kunstmärchen, Roman, Novelle, Erzählung, Kurzgeschichte. Entstehungszeitpunkte: 19. und 20. Jahrhundert. Textlänge: maximal 2000 Wörter.

⁷ <https://github.com/SharedTasksInTheDH>

Tab. 1: Das Entwicklungskorpus

Titel (orig.)	Autor*in	Titel (englisch)	Genre	Jahr	Sprache (orig.)
Rosen-Alfen	Aesop	The Wolf and the Lamb	fable	600 v.u.Z.	dänisch
Kjærestefolkene [toppen og bolden]	Andersen, Hans-Christian	The Elf of the Rose	folktale	1839	dänisch
Se una notte d'inverno un viaggiatore	Andersen, Hans Christian	The Top and Ball	folktale	1862	dänisch
Мститель	Calvino, Italo	If on a Winter's Night a Traveller	novel	1979	italienisch
The Child's Story	Čechov, Anton Pavlovič	An Avenger	short story	1887	russisch
Die drei Federn	Dickens, Charles	The Child's Story	short story	1852	englisch
Das wohlfeile Mittagessen	Grimm, Brüder	Feathers	folktale	1819	deutsch
Der geheilte Patient	Hebel, Johann Peter	The Cheap Meal	anecdote	1804	deutsch
Hills Like White Elephants	Hebel, Johann Peter	The Cured Patient	anecdote	1811	deutsch
Beyond the Pale	Hemingway, Ernest	Hills Like White Elephants	short story	1920	englisch
Unwahrscheinliche Wahrhaftigkeiten	Kipling, Rudyard	How the Leopard got his Spots	short story	1901	englisch
	Kipling, Rudyard	Beyond the Pale	short story	1888	englisch
	Kleist, Heinrich von	Improbable Veracities	anecdote	1810	deutsch
The Cask of Amontillado	Lagerlöf, Selma	Among the Climbing Roses	narration	1894	schwedisch
Frankenstein or The Modern Prometheus	Poe, Edgar Allen	The Cask of Amontillado	short story	1846	englisch
A Haunted House	Shelley, Mary	Frankenstein or The Modern Prometheus	novel	1818	englisch
	Woolf, Virginia	A Haunted House	short story	1921	englisch

Tab. 2: Das Testkorpus

Titel (orig.)	Autor*in	Titel (englisch)	Genre	Jahr	Sprache (orig.)
Lenz	Büchner, Georg	Lenz	novella	1839	deutsch
Выигрышный билет	Čechov, Anton Pavlovič	The Lottery Ticket	short story	1887	russisch
The Gift of the Magi	Henry, O.	The Gift of the Magi	short story	1905	englisch
Kleine Fabel	Kafka, Franz	A Little Fable	fable	1831	deutsch
Der blonde Eckbert	Tieck, Ludwig	The White Egbert	literary fairy tale	1797	deutsch
Der Schimmelreiter	Storm, Theodor	The Rider of the White Horse	novella	1888	deutsch
Anekdoten aus dem letzten preußischen Kriege	Kleist, Heinrich von	Anecdote from the Last Prussian War	anecdote	1810	deutsch
Herr Ames penningar	Lagerlöf, Selma	The Treasure	narration	1904	schwedisch

Bezeichnung	Beschreibung
Own	Annotation durch die Autorinnen/Autoren einer Richtlinie
Foreign	Annotation durch die Autorinnen/Autoren einer <i>anderen</i> Richtlinie
Student	Annotation durch die studentischen Hilfskräfte

Tab. 3: Übersicht über die Annotationstypen

Workshop

Als Meilenstein des ersten *shared task* wurden alle Teilnehmenden nach Hamburg zu einem dreitägigen Workshop eingeladen. Von den acht teilnehmenden Teams konnten sieben physisch anwesend sein, das achte Team konnte zeitweise per Videokonferenz teilnehmen. Am ersten Tag der dreitägigen Veranstaltung lernten die Teilnehmenden sich gegenseitig kennen und stellten ihre Richtlinien vor. Dazu wurden kurze Präsentationen gehalten und danach bereits erste Gemeinsamkeiten und Unterschiede benannt. Am zweiten Tag wurden die Evaluationskriterien (s. u.) vorgestellt und diskutiert und im Anschluss mithilfe eines Fragebogens mit einer Bewertungsskala auf die Richtlinien angewendet. Im Anschluss begründeten die Teilnehmenden ihre Punktevergabe. Am letzten Tag wurden die kumulierten Evaluationsergebnisse zusammen mit dem *inter-annotator agreement* präsentiert und von der gesamten Gruppe diskutiert. Zum Abschluss wurden die nächsten Schritte besprochen.

3.2.2 Ausblick: *shared task 2* – Automatische Erkennung von Erzählebenen

Der zweite *shared task* soll als ‚normaler‘, also computerlinguistischer *shared task* abgehalten werden und richtet sich daher primär an Wissenschaftler*innen aus der Computerlinguistik und verwandten Gebieten. Es ist geplant, dass der *shared task* im Sommer 2021 stattfindet. Das bis dahin annotierte Korpus wird dafür in ein Entwicklungs-, Trainings-, und Testkorpus aufgeteilt und zu bestimmten Zeitpunkten im Laufe des *tasks* zur Verfügung gestellt. Es ist anvisiert, dass der *shared task* im Rahmen der SemEval-Initiative stattfindet. Das erwartete Ergebnis des zweiten *shared task* sind eine Reihe automatischer Systeme, die Erzählebenen erkennen.

4 Evaluation von Annotationsrichtlinien

Die Evaluation von Annotationsrichtlinien im Rahmen eines *shared task* ist eine Herausforderung, die so noch nicht angegangen wurde. Das hier vorgestellte Evaluationsprozedere wurde von Grund auf neu entwickelt und ist, so unsere Überzeugung, auch jenseits der Annotation von Textphänomenen anwendbar.

4.1 Grundlagen und Herausforderungen

Unser grundsätzliches Ziel besteht in der Etablierung eines Verfahrens, das den disziplinären Ansprüchen sowohl der Literaturwissenschaft als auch der Computerlinguistik gerecht wird. Das heißt konkret, dass die folgenden vier Bedingungen erfüllt sein müssen:

Etablierung einer Rangfolge: Die Methode muss offensichtlich in der Lage sein, eine Rangfolge zu etablieren. Diese Rangfolge sollte so klar wie möglich sein und Gleichstände, soweit es geht, vermeiden.

Explizit- und Klarheit: Da *shared tasks* als Wettbewerb angelegt sind, müssen die Richtlinien nach einer Zielfunktion sortiert werden. Die Zielfunktion muss dabei vorher so genau wie möglich festgelegt werden, damit die Teilnehmenden vorher wissen, worauf sie sich einlassen und möglichst wenig Raum für Missverständnisse entsteht.

Praktikabilität: Die Evaluation muss innerhalb der praktischen Beschränkungen eines *shared task* durchführbar sein. Bei uns hieß das, dass es möglich sein musste, innerhalb eines dreitägigen Workshops ein Ergebnis zu erzielen.

Umsetzung der Evaluationskriterien: Jede Evaluationsmethode muss die Kriterien möglichst passgenau umsetzen, so dass eine Richtlinie, die bestimmte Aspekte besser löst als eine andere, auch eine höhere Bewertung bekommt. Welche Kriterien genau angelegt werden, ist eine Setzung, die das Organisationsteam vornehmen muss. Im Sinne der Fairness sollten sie so früh wie möglich kommuniziert werden.

Diese Anforderungen ergeben sich daraus, dass Annotationsrichtlinien in einem *shared task* erstellt werden. Während in computerlinguistischen *shared tasks* ein Goldstandard existiert, der so gut wie möglich reproduziert werden muss, gibt es für Annotationsrichtlinien bisher keine etablierte *ground truth*, anhand derer sie bewertet werden könnten. Selbst das Messen von *inter-annotator agreement*, das gelegentlich als Qualitätsmaßstab für Annotationsrichtlinien herangezogen wird, könnte in diesem Fall in die Irre führen, da mit nicht eindeutig auflösba-

ren Ambiguitäten zu rechnen ist. Annotationsunterschiede sollten nämlich nicht als Mangel der Richtlinien dargestellt werden, wenn das Annotationsphänomen stark ambig ist.

Zusätzlich zu den genannten allgemeinen Kriterien, die jedes Evaluationsverfahren erfüllen sollte, bringt unser *shared task* einige spezifische Herausforderungen mit sich:

Da es sich bei diesem *shared task* um eine interdisziplinäre Unternehmung handelt, muss mit einem heterogenen Kreis aus Teilnehmer*innen gerechnet werden. Das Konzept Annotation spielt, in verschiedenen Benennungen und Ausformungen, verschiedene Rollen in den beteiligten Disziplinen, wodurch sich eine diverse Sammlung an best practices, Regeln und Traditionen etabliert hat. Literaturwissenschaftliche Annotation z. B. wird typischerweise als Unterstreichungen von Textpassagen oder als *note-taking*-Annotationen verstanden, also als Anfertigen von Notizen am Seitenrand, ohne feste Kategorien oder explizite Verbindung zur auslösenden Textstelle. In der Computerlinguistik hingegen werden Annotationen typischerweise parallel von mehreren Annotator*innen vorgenommen und eine hohe Übereinstimmung ist oberstes Ziel. Letzteres ist bei subjektiven Annotationen mit dem Ziel, eine Materialsammlung für eine Interpretation zu liefern, schlicht nicht relevant. Die Teilnehmenden haben also unterschiedliche Vorerfahrungen und Erwartungen an den Annotationsprozess – gleichwohl muss das Ergebnis der Evaluation für alle Beteiligten nachvollziehbar, funktional und auch nutzbringend sein.

Eine weitere Herausforderung liegt in der für die Literaturwissenschaft durchaus typischen Vagheit der genutzten Konzepte. Die Narratologie stellt innerhalb der Digital Humanities ein beliebtes Betätigungsfeld dar, was vermutlich auf ihren strukturalistischen Hintergrund und die damit einhergehende, vergleichsweise einfachere Operationalisierung zurückzuführen ist. Zudem vereint die Narratologie und die Digital Humanities die Auffassung, dass eine strukturelle Analyse von Texten *überhaupt* interessante textbezogene Beobachtungen zutage fördert. Aus Annotationssicht bleiben die in der Narratologie diskutierten Konzepte jedoch zu vage, um sie direkt anzuwenden. Neben den Primärtexten selbst geben auch die narratologischen Konzepte einen gewissen Interpretationsspielraum. Wir haben daher im Vorfeld beschlossen, dass die Evaluation möglichst viele verschiedene, auch unterschiedliche narratologische Perspektiven berücksichtigen sollte. Vagheiten und komplexe Abhängigkeiten der Konzepte sollten also von den Richtlinien nicht zwingend aufgelöst werden. Trotzdem sollte idealerweise ein Umgang damit beschrieben werden.

Abb. 1: Die drei Evaluationsdimensionen (oben) verbinden Schritte in der computationalen Textanalyse in den Digital Humanities (unten)

4.2 Evaluationskonzept

Das im *shared task* verwendete Evaluationskonzept beruht auf drei Dimensionen, mit denen unterschiedliche Eigenschaften der Richtlinien beurteilt werden können. Abbildung 1 zeigt schematisch, wo sich die Dimensionen im Verhältnis zu Forschungsaktivitäten im Kontext computationaler Textanalysen befinden. Damit verteilen sich die Dimensionen auf den gesamten Forschungsprozess, von der narratologischen Theorie über die Erstellung von Richtlinien und die Annotation von Texten bis hin zu den Erkenntnissen, die sich aus den annotierten Daten, seien es einzelne Texte oder ganze Korpora, ergeben.

Die Dimension **Theorieabdeckung** reflektiert, wie viel von einer theoretischen Basis in den auf ihr aufbauenden Richtlinien enthalten ist. Wenn etwa eine Richtlinie explizit auf einer narratologischen Theorie aufbaut, könnte sie darauf abzielen, jede einzelne Definition, Regel und Ausnahme in die Richtlinie aufzunehmen. Eine andere Richtlinie, die auf derselben Theorie basiert, könnte hingegen Aspekte weglassen oder stark verändern. Die Dimension **Theorieabdeckung** verbindet also Richtlinien und Theorie.

Anwendbarkeit betrachtet die Relation zwischen Richtlinien und Text. Diese Dimension repräsentiert, wie gut eine Richtlinie die Annotator*innen darauf vorbereitet, die Annotationen tatsächlich durchzuführen. Die Anwendbarkeit einer Richtlinie kann z. B. durch gut durchdachte und instruktive Beispiele erhöht werden, aber auch durch eine klare Struktur und/oder einen sorgsam Gebrauch der Terminologie. *Inter-annotator agreement* und Systematizität der erfolgten Annotationen fallen ebenfalls in diese Dimension.

Die Dimension der **Nützlichkeit** verbindet schließlich den annotierten Text mit Anwendungen und Erkenntnissen, wobei beide Begriffe hier relativ weit gefasst sind. Anwendungen decken hier z. B. auch automatische Analyseverfahren ab, die auf die Erzählebenen aufbauen, aber auch die Nützlichkeit der Ergebnisse von Analysen großer Datenmengen für die anschließende (menschliche) Interpre-

tation. In dieser Dimension wird also auf das Ergebnis, also die erfolgten Annotationen als solche, abgezielt und nicht so sehr auf den Annotationsprozess.⁸

Die drei beschriebenen Dimensionen erlauben eine balancierte Evaluation von Richtlinien mit heterogenen disziplinären Hintergründen, Zielen und Annahmen über theoretische Konzepte. Eine Richtlinie, die sich nur auf eine Dimension konzentriert, kann zwar in dieser gut abschneiden, läuft aber Gefahr, genau dadurch die anderen Dimensionen zu schwächen. So könnte eine Richtlinie ausschließlich auf narratologische Kategorien und Begriffe setzen und würde damit zwar eine hohe Theorieabdeckung erzielen, aber, wegen der oben beschriebenen Vagheit, keine sehr hohe Anwendbarkeit. Wird eine Richtlinie hingegen rein auf ihre Anwendbarkeit optimiert, ließen sich zwar schnell und zuverlässig Annotationen gewinnen – die aber am Ende niemandem nützen (weil sie im Extremfall definieren, dass ein Text aus einer Erzählebene besteht). Die reine Optimierung auf Nützlichkeit führt womöglich zu Richtlinien, die keinerlei Bezug mehr zu theoretischen Konzepten haben. Die Herausforderung für die Teilnehmenden besteht also in der Balancierung unterschiedlicher Anforderungen, die sich aus den Dimensionen ergeben.

Selbstverständlich kann auch eine Richtlinie, die nicht alle drei Dimensionen gleich gewichtet, sinnvolle Anwendungen ermöglichen. Richtlinien ohne Theoriebezug könnten etwa auch neue theoretisch interessante Phänomene aufspüren helfen. Es ist also in der Praxis und außerhalb von *shared tasks* sicher nicht nötig, auf alle drei Dimensionen abzu zielen. Doch auch in diesem Fall können die Dimensionen eine bewusste Entscheidung unterstützen.

4.3 Umsetzung

Um das oben beschriebene, dreidimensionale Evaluationsmodell praktisch einzusetzen und zu erproben, wurde jede Dimension mit einer Reihe von konkreten Fragen unterfüttert, die verschiedene Aspekte aus der Dimension abfragen. Die Fragen sind weiter unten aufgelistet. Da sie im Original auf Englisch verwendet wurden, sind sie hier ebenfalls auf Englisch abgedruckt.

Die Fragen wurden den Teilnehmenden vor der Einreichung ihrer Richtlinien zur Verfügung gestellt. Während des Evaluationsworkshops wurden sie auf zwei-

⁸ Vergleiche zu den drei Dimensionen auch die in der empirischen Sozialwissenschaft etablierten Begriffe der Validität und Reliabilität. Auch wenn sie nicht exakt deckungsgleich sind, lassen sich Bezüge herstellen: Die Dimension Anwendbarkeit ist dabei verwandt mit der Reliabilität, während die externe Validität von der Dimension Nützlichkeit und die interne von der Dimension Theorieabdeckung abgedeckt werden.

erlei Weisen verwendet. Zunächst bildeten sie den Leitfaden für eine qualitative Diskussion der Richtlinien. Durch den Fragebogen wurde sichergestellt, dass alle Aspekte in der Diskussion zur Sprache kamen und dass an alle Richtlinien die gleichen Maßstäbe angelegt wurden. Diese Diskussion wurde von den Teilnehmenden als intensiv, aber sehr produktiv wahrgenommen und gab bereits Anlass für Verbesserungen.

Außerdem wurden die Fragen quantitativ beantwortet. Dazu wurde jeder Frage eine Likert-Skala mit vier Werten zugeordnet. Die teilnehmenden Teams wurden dann gebeten, jeder (anderen) Richtlinie in jeder Frage Punkte zuzuweisen, wobei mehr Punkte eine bessere Bewertung bedeuteten.

Unser Fragebogen definiert vier Fragen für die Dimensionen Theorieabdeckung und Nützlichkeit und zwei Fragen für die Dimension Anwendbarkeit. Damit alle Dimensionen gleichwertig waren, wurde die Dimension Anwendbarkeit ergänzt durch die Ergebnisse des skalierten *inter-annotator agreement* (gemessen mit γ , Mathet et al. 2015). Damit wurden jeder Richtlinie in jeder Dimension vier Werte zugeordnet, die erst als Dimensionswerte und dann als Gesamtwert aufaddiert wurden.

4.3.1 Fragebogen

Der Fragebogen wird hier im Original wiedergegeben.

Theorieabdeckung

1. Is the narrative level concept explicitly described?

Explanation: Narrative levels can be described or defined. This depends on the narratology used; some of them are structuralist, others are post-structuralist. Regardless of the mode, is the description/definition understandable and clear?

- 1: I did not understand what the guideline describes as ‘narrative level’.
- 4: I fully understood the concept described in the guideline.

2. Is the narrative level concept based on existing concepts?

Explanation: The level concepts can be self-designed, oriented on existing narratologies or copied from an existing level definition

- 1: The theory relation of the used level concept is not clear.
- 4: It is clearly mentioned whether the level concept is made up or (partially) based on a theory.

3. How comprehensive is the guideline with respect to aspects of the theory?

Does it omit something? Explanation: If the guideline is based on a theory

or multiple theories, does it include the whole theory or only parts of it? Are there reasons mentioned why aspects are in-/excluded?

- 1: The guideline does not clearly state the extension of its dependence on theory/ies.
 - 4: The guideline unambiguously states the scope of its theory-dependence.
4. How adequately is the narrative level concept implemented by this guideline in respect to narrative levels?

Explanation: Narratologies differ in their complexity. Firstly, you have to decide whether complexity or simplicity (in relation to x) is desirable, then you have to answer:

- 1: The guideline is too simple or too complex for narrative levels and thus not adequate.
- 4: The guideline's complexity is adequate.

Anwendbarkeit

1. How easy is it to apply the guideline for researchers with a narratological background?

Explanation: The question asks for an assessment of the ease of use of the guideline for an annotator with some narratological background. Indicators can be: Complexity of the concepts, length of the guideline, clarity of examples, clear structure, difficulty of finding special cases, etc.

- 1: Even as a narratology expert, I needed to read the guideline multiple times and/or read additional literature.
 - 4: The guideline is very easy to apply, and I always knew what to do.
2. How easy is it to apply the guideline for researchers without a narratological background?

Explanation: The question asks for an assessment of the ease of use of the guideline if we assume an annotator who doesn't have a narratological background (e.g., an undergraduate student). Indicators can be: Complexity of the concepts, length of the guideline, use of terminology, clarity of examples, reference to examples only by citation, clear structure, difficulty of finding special cases, etc.

- 1: Non-experts have no chance to use this guideline.
- 4: The guideline is very easy to apply, and non-experts can use them straight away.

Nützlichkeit

1. Thought experiment: Assuming that the narrative levels defined in the annotation guideline can be detected automatically on a huge corpus. How helpful are these narrative levels for an interesting corpus analysis?

Explanation: This question focuses on the relevance of the narrative level annotations for textual analysis of large amounts of texts, e.g., for the analysis of developments over time with regard to narrative levels or a classification of texts with regards to genre, based on narrative levels.

- 1: The narrative levels annotations are irrelevant for corpus analysis.
- 4: The annotations provide interesting data for corpus analysis.

2. How helpful are they as an input layer for subsequent corpus or single text analysis steps (that depend on narrative levels)?

Explanation: The analysis of some other textual phenomena depends on narrative levels, e.g., chronology should be analyzed within each narrative level before analyzing it for the whole text. This question asks whether the analysis of such phenomena is possible or even better when based on the narrative level annotations.

- 1: The usage of the narrative levels annotations makes no difference for subsequent analyses.
- 4: Subsequent analyses are possible only because of the narrative level annotations.

3. Do you gain new insights about narrative levels in texts by applying the foreign guideline, compared to the application of your own guideline?

Explanation: In most cases annotating a text in accordance to a guideline changes the evaluation of textual phenomena in the text, e.g., the quality (or quantity) of narrative levels in the text.

- 1: It doesn't make a difference—I get no additional insights with the foreign guideline.
- 4: I gain a lot of new insights about narrative levels in texts based on this guideline.

4. Does the application of this guideline influence your interpretation of a text?

Explanation: Interpretations are normally based on the analysis of a text and thus on the observation of the presence (or absence) of certain textual phenomena. Therefore, the application of the guidelines may result in annotations that are relevant for your interpretation, e.g. the detection of a narrative level of a certain type may influence your interpretation of the reliability of a narrator.

- 1: My interpretation is independent from the annotations based on the guideline.

- 4: My interpretation is based primarily on the annotations based on the guideline.

4.3.2 Integration der Evaluationsergebnisse

Mithilfe des Fragebogens wurden für jede Richtlinie sieben Bewertungen eingeholt (von den sieben jeweils anderen Teams). Aus diesen Bewertungen wurde eine Gesamtpunktzahl wie folgt berechnet:

1. Für jede der zehn Fragen wird das arithmetische Mittel über alle Antworten berechnet. Daraus ergeben sich zehn Mittelwerte, vier in der ersten und dritten und zwei in der zweiten Dimension.
2. Die IAA-Ergebnisse werden in das Intervall [1;4] skaliert und zweimal als ‚virtuelle Fragen‘ in der Dimension der Anwendbarkeit addiert. Damit existieren vier Werte je Dimension für jede Richtlinie.
3. Die Werte jeder Dimension werden dann addiert, woraus sich für jede Richtlinie und jede Dimension ein Wert im Intervall [4;16] ergibt.
4. Als Gesamtpunktzahl werden diese Werte addiert. Die Gesamtpunktzahl liegt dann im Intervall [12;48].

5 Ergebnisse des *shared task*

5.1 Allgemeine Beobachtungen

Wie zu erwarten, waren die Einreichungen für den *shared task* in Bezug auf Disziplinen und geographische Herkunft divers. Tabelle 4 zeigt einige Eckdaten der teilnehmenden acht Teams. Die Teams unterscheiden sich in Bezug auf Alter, Geschlecht, Größe, akademischen Status und disziplinären Hintergrund. Diese Heterogenität zeigt sich auch in den Richtlinien, deren Länge von einer bis fünfzig Seiten reicht und unter denen sich sowohl theoretische Essays als auch praxisorientierte Anleitungen finden. Zwei der eingereichten Richtlinien wurden von Mitarbeiter*innen aus CRETA verfasst und sind daher auch in diesem Buch abgedruckt: Richtlinie IV (Ketschik, Murr et al. 2020) ab Seite 440 und Richtlinie V (Barth 2020) ab Seite 423.

Tab. 4: Teilnehmende Teams. Alle Richtlinien wurden in einer Sonderausgabe der Zeitschrift *Cultural Analytics*⁹ publiziert. Die mit * markierten Richtlinien sind in diesem Band abgedruckt als Barth (2020) und Ketschik, Murr et al. (2020). Disziplinärer Hintergrund basiert auf einer Selbstbeschreibung der Teilnehmenden: Computerlinguistik (CL), Englische Literatur (EL), Data science (DS), Literaturwissenschaft (LW), Digital Humanities (DH). Herkunft: Vereinigte Staaten von Amerika (US), Deutschland (DE), Schweden (SWE), Irland (IRL), Kanada (CA). Größe beschreibt die Zahl der Autorinnen und Autoren einer Richtlinie, L markiert Richtlinien, die im Rahmen einer Lehrveranstaltung entstanden sind.

Richtlinie	Referenz	Disziplin	Herkunft	Größe
I	Eisenberg und Finlayson (2019)	CL	US	2
II	Kearns (2019)	EL/DS	IRL	1
III	–	EL	DE	2L
IV	Ketschik, Krautter et al. (2019)*	LW/DH	DE	4
V	Barth (2019)*	DH/LW	DE	1
VI	Bauer und Lahrsoy (2020)	EL	DE	2L
VII	Wirén et al. (2020)	CL	SWE	3
VIII	Hammond (2020)	LW	CA	1L

5.1.1 Richtlinien und Erzählebenen

Da es den Teilnehmenden überlassen war, ob und ggf. an welchen theoretischen Werken sie sich orientierten, stellen wir die theoretischen Grundlagen der Richtlinien auf Basis der in ihnen enthaltenen Referenzen und Erwähnungen dar. Tabelle 5 zeigt eine Übersicht über die narratologischen Werke, während Tabelle 6 die erwähnten Konzepte zeigt.¹⁰

Es ist nicht überraschend, dass Genettes *Narrative Discourse* am häufigsten referenziert wurde, da das Konzept der Erzählebene hier prominent diskutiert wird und die meisten anderen Ansätze in der einen oder anderen Form darauf aufbauen. Andere Werke, die relativ prominent sind (in mehr als drei Richtlinien referenziert), sind die Einführungstexte von Jahn („N2.4. Narrative Levels“) und Pier („Narrative Levels“) sowie der Ansatz von Ryan (*Possible worlds, artificial intelligence, and narrative theory*).

¹⁰ Man beachte jedoch, dass hier die Referenz auf Werke bzw. Konzepte gesammelt wurde. Nicht beachtet wurde hingegen, ob die Interpretation des Werkes bzw. die Nutzung des Konzeptes kompatibel sind.

¹¹ Die folgenden Arbeiten wurden zusätzlich in einzelnen Richtlinien zitiert: Richtlinie II: McHale (2014); Rickert (1923). Richtlinie III: Nünning (2004); Ryan (1986). Richtlinie IV: Abrams (1999); Pier (2016); Fludernik (2009); Genette (1988); Genette (1997); Margolin (2014); Ryan (2001); Schmid (2014). Richtlinie V: Duyfhuizen (2005); Genette (1988); Gius (2015); Rimmon-Kenan (2005); Ryan

Tab. 5: Narratologische Arbeiten, die in den Richtlinien operationalisiert wurden. Die Zuweisung basiert auf Referenzen in den Richtlinien und/oder Einleitungen.

Richtlinie	I	II	III	IV	V	VI	VII	VIII
Genette 1980		x	x	x	x	x	x	
Jahn (2017)		x		x	x		x	x
Lahn und Meister (2008)				x	x			
Lämmert (1955)					x			
Mani (2012)								
Martínez und Scheffel (2003)				x				
Nelles (1997)			x		x	x		
Nelles (2005)					x			
Neumann und Nünning (2012)					x			
Pier (2014)			x		x	x		
Pier (2012)				x				
Rimmon-Kenan (2004)								
Romberg (1962)								
Ryan (1991)			x	x	x			
zusätzliche Referenzen ¹¹		x	x	x	x	x	x	
eigene Beiträge	x							x

Tab. 6: Narratologische Konzepte, die in den Richtlinien operationalisiert wurden. Gesammelt sind explizite Referenzen.

Richtlinie	I	II	III	IV	V	VI	VII	VIII
Definition einer ‚Erzählung‘	x				x	x	x	x
Erzählebene	x	x	x	x	x	x		x
Diskursebene							x	
Erzähler (Identität, Verhältnis zum Text)	x	x	x	x	x	x	x	x
Rezipient			x			x	x	
Struktur der Ebenengrenze	x		x	x	x	x	x	
Art der Ebenengrenze (z. B. Metalepsis)			x	x	x	x	x	
Sprecher (illokutionäre Grenze)			x	x	x		x	
Veränderung der Welt (ontol. Grenze)	x		x		x	x		
Fokalisierung			x			x	x	
Analepse/Prolepse		x	x					
Bewusstseinsstrom/free indirect discourse			x					
Dehnung/Raffung			x					

Tabelle 6 zeigt eine Übersicht über die Konzepte, die von den Teilnehmenden für die Identifikation von Erzählebenen als relevant erachtet wurden. Die aufgeführten Konzepte können danach unterschieden werden, ob sie unmittelbar mit Erzählebenen zusammenhängen (z. B. Art der Ebenengrenze) oder häufig mit Erzählebenen gemeinsam auftreten (z. B. Fokalisierung). Während erstere direkt für die Operationalisierung nützlich sein können, sind letztere eher für weitere Analysen relevant, da sie häufig innerhalb von bzw. zusammen mit Erzählebenen auftreten. Diese Mischung der Konzepte kann einerseits durch die theoretische Offenheit und andererseits durch die unterschiedlichen Ziele der Teilnehmenden erklärt werden.

6 Ergebnisse der Evaluation

Im Folgenden werden die Evaluationsergebnisse der *ersten Runde* des *shared task* präsentiert. Die Gesamtergebnisse finden sich in Tabelle 7. Danach folgen die Ergebnisse in den Dimensionen Theorieabdeckung (Tabelle 8), Anwendbarkeit (Tabelle 9) und Nützlichkeit (Tabelle 10). Die dimensionsspezifischen Tabellen zeigen Mittelwerte und Standardabweichungen gemäß den Fragebögen. Die Fragen selbst sind in Abschnitt 4.3.1 wiedergegeben.

Die Ergebnisse in Tabelle 8 basieren auf vier Fragen. Richtlinie IV (S. 440) hat hier das beste Ergebnis erzielt. Dabei handelt es sich um eine Richtlinie, die auf eine tiefe narratologische Beschreibung setzt und von narratologischen Be-

Tab. 7: Erzielte Ergebnisse der Guidelines nach Dimensionen

Richtlinie	Theorieabdeckung	Anwendbarkeit	Nützlichkeit	Summe	IAA
V	14,14	12,09	12,88	39,1	0,25
II	11,17	11,89	12,57	35,63	0,24
VI	12,33	11,01	11,37	34,71	0,21
IV	14,43	7,71	11,26	33,4	0,05
VIII	8,1	14,14	9,12	31,36	0,3
VII	11,6	9,82	9,77	31,18	0,23
I	7,83	10,39	10	28,22	0,18
III	10,29	6,48	10,95	27,72	0,07

(2002). Richtlinie VI: Füredy (1989); Nünning (2004); Ryan (1986). Richtlinie VII: Doležel (1973); Niederhoff (2013); Rimmon-Kenan (2005); Todorov (1966). Richtlinie VIII: Abbott (2002). Richtlinie III wurde nicht veröffentlicht, da sie zurückgezogen wurde.

Tab. 8: Evaluationsergebnisse in der Dimension Theorieabdeckung. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

Richtlinie	Q1	Q2	Q3	Q4
I	3 ±1,1	1 ±0	1 ±0	2,83±0,98
II	2,67±1,03	3 ±1,1	2,83±0,75	2,67±1,21
III	2,43±0,98	2,86±1,07	2,71±1,38	2,29±0,76
IV	3,71±0,49	4 ±0	3,71±0,49	3 ±0,82
V	3,71±0,49	4 ±0	3,43±0,53	3 ±1
VI	3,33±0,82	3,33±0,82	2,67±0,52	3 ±0,89
VII	2,71±1,11	3,33±0,82	2,71±0,95	2,83±1,33
VIII	2,29±1,11	1 ±0	1,67±1,21	3,14±0,69

Richtlinie	Q1	Q2
I	2,6 ±0,55	2,67±0,52
II	3,17 ±0,41	2,17 ±0,98
III	2,57±1,13	1,43±0,53
IV	3,57±0,53	2,14±0,38
V	3 ±1,15	2,29±0,95
VI	3,17±1,17	2 ±0,89
VII	2,33±0,52	1,17±0,41
VIII	2,71±1,38	3,43±0,79

Tab. 9: Evaluationsergebnisse in der Dimension Anwendbarkeit. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

griffen intensiv Gebrauch macht. Am anderen Ende des Spektrums findet sich Richtlinie I, die die wenigsten Punkte erzielt hat. Dies deckt sich mit der Selbstbeschreibung und dem primären Forschungsziel der Einreichenden, das auf das „computational understanding of stories“ abzielt (Eisenberg und Finlayson 2019, S. 1). Damit stehen in Richtlinie I klar Anwendungen im Vordergrund, eine Anbindung an narratologische Theorie ist eher zweitrangig. Außerdem werden einige nicht-erzählende Textbeispiele genannt (z. B. Drehbücher), die aus einem Annotationsprojekt der Richtlinien-Autoren stammen. Dadurch dass die narratologisch motivierten Teilnehmenden in der Überzahl waren, wurden diese Abweichungen relativ stark ‚bestraft‘, was in einem geringen Punktwert resultierte.

Die Punktzahl in der Anwendbarkeits-Dimension basiert auf der Berechnung des *inter-annotator agreement* und zwei Fragen aus dem Fragebogen (Tabelle 9). Die erste der beiden Fragen zielt darauf ab, wie gut Expert*innen für narratologische Textanalyse die Richtlinie anwenden können, die zweite stellt die gleiche Frage für Laien. Richtlinie VIII (Hammond 2020) erzielt hier die meisten Punkte, während Richtlinie III die wenigsten erhält. Für eine Interpretation dieser Ergebnisse ist es hilfreich, die einzelnen Fragen und das IAA separat zu betrach-

Tab. 10: Evaluationsergebnisse in der Dimension Nützlichkeit. Die Tabelle zeigt arithmetisches Mittel und Standardabweichung für jede Frage. Die Fragen sind in Abschnitt 4.3.1 wiedergegeben.

Richtlinie	Q1	Q2	Q3	Q4
I	3,17 ±0,75	3 ±0,71	2,17 ±0,75	1,67 ±0,52
II	3,5 ±0,55	3,4 ±0,89	3 ±0,89	2,67 ±0,52
III	3,33 ±0,82	3,17 ±0,98	2,29 ±0,76	2,17 ±0,75
IV	3,29 ±0,76	2,83 ±0,98	2,71 ±0,76	2,43 ±0,98
V	3,5 ±0,55	3,4 ±0,55	3,14 ±0,69	2,83 ±0,75
VI	3,4 ±0,55	3,4 ±0,89	2,17 ±0,75	2,4 ±0,55
VII	3 ±0,63	2,6 ±0,55	2,17 ±0,98	2 ±0,89
VIII	3,2 ±0,84	2,8 ±1,1	1,29 ±0,76	1,83 ±0,75

ten. Dass Richtlinie VIII hier die höchste Punktzahl erzielte, wirft die Frage nach dem Verhältnis aus Einfachheit des Konzeptes und Anwendbarkeit auf. Richtlinie VIII erzielt das höchste IAA, scheint also zumindest am robustesten anwendbar zu sein. Gleichzeitig hat Richtlinie VIII nur eine Position im Mittelfeld für Expert*innen erzielt und lediglich für Laien eine sehr gute Anwendbarkeit. In der Praxis scheint ein einfaches Ebenenkonzept für Laien gut anwendbar zu sein, dies scheint jedoch nicht auch (automatisch) für Expert*innen zu gelten.

Die gute Bewertung von Richtlinie V in der Dimension der Nützlichkeit lässt sich wahrscheinlich auf a) die vielen Beispiele, die die Benutzung der Kategorien illustrieren, und b) die klare Beschreibung der Forschungsziele durch den Richtlinienautor zurückführen. Ähnlich lässt sich die gute Bewertung von Richtlinie II erklären, die hier den zweiten Platz erzielt hat. Die Richtlinie erklärt, sie sei „designed for annotating analepsis, prolepsis, stream-of-consciousness, free indirect discourse, and narrative levels, with facility also for annotating instances of extended or compressed time, and for encoding the identity of the narrator“ (Kearns 2019, S. 3). Da die Dimension Nützlichkeit eine Reihe von Einsatzzwecken und letztlich das Potenzial der Richtlinie abdeckt, ist es interessant zu beobachten, dass die Angaben über konkrete Einsatzszenarien die Einschätzung der Nützlichkeit insgesamt zu erhöhen scheint.

Zum Abschluss noch ein Kommentar zur Gewinnerrichtlinie (Richtlinie V Barth 2019), für diesen Band überarbeitet und abgedruckt als Barth 2020 ab Seite 423. Richtlinie V hat keineswegs in allen Dimensionen gewonnen. Allerdings hat sie eine gute Balance zwischen den drei Evaluationsdimensionen geschaffen. Sowohl die quantitativen Ergebnisse der Evaluation als auch die Diskussion während des Workshops bestätigten, dass die Richtlinie Erzählebenen genau und präzise beschreibt. Die Richtlinie unterscheidet Erzählebenen von Erzählakten und

verwendet auch andere narratologische Konzepte (z. B. Erzähler) als Hilfsmittel zur Erkennung von Ebenen. Abstrakte Beispiele in der Form von Diagrammen und Tabellen machen die Richtlinie verständlich. Die Richtlinie enthält zudem konkrete Handlungsanweisungen und stellt heraus, wozu die Ebenenannotationen nützlich sein können.

7 Fazit

Da der *shared task* noch läuft, ist es für ein abschließendes Fazit zu früh. Einige erste Einsichten, gerade zum ersten Teil des *shared task*, lassen sich jedoch bereits formulieren.

inter-annotator agreement

Die Werte des *inter-annotator agreement* basieren auf einer relativ kleinen Zahl an Annotationen, die von Annotierenden ohne systematisches Training durchgeführt wurden. Die Annotationen wurden daher auf unterschiedlichen Niveaus durchgeführt (cf. Abbildung 3): Die studentischen Annotierenden waren mehrheitlich untrainiert (nur zwei von acht Studierenden hatten Erfahrung in anderen Annotationsaufgaben) und hatten keinerlei Vorkenntnisse in narratologischen Fragen. Die Annotierenden für die ‚foreign‘-Annotationen waren naturgemäß auf ihre eigenen Richtlinien trainiert und einige hatten offensichtliche Schwierigkeiten, sich von ihnen zu lösen. Die Teilnehmenden waren unterschiedlich zufrieden mit den Annotationen, welche ihnen schon vor dem Workshop zur Verfügung gestellt worden waren. Einige Teilnehmende empfanden gerade die ‚foreign‘-Annotationen als nicht adäquat, etwa bei narratologisch komplexen Richtlinien, die von nicht-narratologischen Teilnehmenden angewendet werden mussten. Teilweise ist dies sicher eine Folge der Interdisziplinarität des Vorhabens – für die Teilnehmenden war es schwer einzuschätzen, welche Kenntnisse sie bei den Annotierenden voraussetzen konnten.

Eine daran anschließende Beobachtung ist, dass in einigen Fällen das Agreement zwischen den ‚own‘- und ‚student‘-Annotationen höher war als zwischen den ‚own‘- und ‚foreign‘-Annotationen. Die Autor*innen der Richtlinien stimmten also eher mit den studentischen Annotierenden überein als mit den anderen Teilnehmenden am *shared task*. Es wurde daher diskutiert, ob nicht die disziplinäre Herkunft bei der Zuteilung hätte berücksichtigt werden müssen, was auf Organisationsseite wiederum nicht praktikabel erschien. Nichtsdestoweniger sind Alternativen wünschenswert, die ohne ‚foreign‘-Annotationen auskommen. Letzten Endes wurde der beschriebene Annotationsmodus gewählt, um überhaupt ein *inter-*

annotator agreement ermitteln zu können. Bei einer geeigneten Finanzierung von Annotationen unter kontrollierteren Bedingungen könnte darauf verzichtet werden.

Fragebogen

Während des Workshops wurde zudem der Fragebogen in mehreren Hinsichten diskutiert:

- Fragen zur Dimension der Theorieabdeckung zu beantworten, erfordert breites narratologisches Wissen, das nicht alle Teilnehmenden mitbrachten. Dadurch konnten letztlich nicht alle Richtlinien nach gleichen Maßstäben bewertet werden.
- In der Dimension der Anwendbarkeit wird nach der Anwendbarkeit für Laien bzw. Expert*innen. Da die teilnehmenden Gruppen sich aber aus Laien *oder* Expert*innen zusammensetzten, konnte jede Gruppe nur eine der beiden Fragen aus eigener Anschauung beantworten, während für die andere eine Reihe von Annahmen getroffen werden musste.
- Insbesondere Literaturwissenschaftler*innen äußerten Besorgnis, dass das Ergebnis der komplexen Evaluation am Ende wiederum nur Zahlen sein würden. Diese Unzufriedenheit bezog sich nicht auf die Messung des IAA oder auf den Fragebogen, sondern gründete in einer grundsätzlichen methodischen Skepsis, ob Annotationsrichtlinien auf diese Weise bewertet werden können. Dies wurde von den Teilnehmenden mit technischem Hintergrund weniger als Problematik wahrgenommen.
- Zuletzt waren auch die Fragen in der Dimension der Nützlichkeit im Einzelnen schwer zu beantworten. Die Nützlichkeit einer Richtlinie kann am besten beurteilt werden, nachdem tatsächlich ein Nutzungsversuch stattgefunden hat. Dies war aber im Rahmen des *shared task* nicht machbar, weswegen in den Antworten zum Fragebogen ein gewisses Maß an Spekulation enthalten ist.

Trotz dieser Schwierigkeiten geben die Antworten die inhaltliche, qualitative Diskussion des Workshops aus unserer Sicht angemessen wieder und zeigen eine erstaunliche Homogenität (geringe Standardabweichung, s. o.).

Evaluationsmodell

Das von uns entworfene, dreidimensionale Bewertungsmodell wurde so gestaltet, dass jede Dimension auf unterschiedliche Weise Merkmale begünstigt, die mit der disziplinären Herkunft der Teilnehmenden zusammenhängen. Die Idee

war, dadurch die disziplinär bedingte Bevorzugung bestimmter Aspekte abzumildern. Entsprechend wurden die Dimensionen und ihre Kombination so gestaltet, dass sie disziplinär bedingte Verzerrungen möglichst aufheben. Die Ergebnisse der Evaluation sprechen dafür, dass dies gelungen ist. Beispielsweise erreichte die Richtlinie IV, deren Autor*innen alle einen literaturwissenschaftlichen Hintergrund haben, die erste Position in Theorieabdeckung, eine Mittelfeldposition in Nützlichkeit und die vorletzte Position in Anwendbarkeit. Richtlinie I, die von Computerlinguisten verfasst wurde, lag bei der Theorieabdeckung an letzter Stelle, erhielt aber in den anderen beiden Dimensionen durchschnittliche Bewertungen. Dies gibt uns Grund zu der Annahme, dass disziplinäre Vor- und Nachteile durch unseren Bewertungsansatz tatsächlich ausgeglichen werden. Auch die Tatsache, dass die Richtlinien IV und VIII in den Dimensionen 1 und 2 invertierte Ränge erreichten, deutet darauf hin, dass die Dimensionen die disziplinären Vorteile neutralisieren. Die insgesamt am höchsten bewertete Richtlinie erhielt in allen Dimensionen hohe Werte, erzielte aber nur in einer Dimension den ersten Platz. Dies deutet darauf hin, dass man, um generell erfolgreich zu sein, eine gewisse Balance zwischen den Dimensionen erreichen sollte. Dies ist der Effekt, der bei der Gestaltung des Evaluationsschemas angestrebt wurde.

Funding: Der *shared task* wurde freundlicherweise von der VolkswagenStiftung unterstützt, die sowohl die Reisekosten als auch einen Teil der Kosten für die Arbeiten der Hilfskräfte übernommen hat.

Danksagung: Wir bedanken uns bei den teilnehmenden Teams für die Experimentierfreude und das Vertrauen.

Literatur

- Abbott, H. Porter (2002). *The Cambridge Introduction to Narrative*. Cambridge: Cambridge University Press.
- Abrams, Meyer Howard (1999). „Narrative and Narratology“. In: *A Glossary of Literary Terms*. Hrsg. von Meyer Howard Abrams. 7. Aufl. Boston: Heinle & Heinle.
- Barth, Florian (2019). „Annotation Guideline No. 5: Annotation Guidelines for Narrative Levels and Narrative Acts“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. DOI: 10.22148/16.056.
- Barth, Florian (2020). „Annotation narrativer Ebenen und narrativer Akte“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 423–438.

- Bauer, Matthias und Miriam Lahrsow (2020). „Annotation Guideline No. 6: SANTA 6 Collaborative Annotation as a Teaching Tool Between Theory and Practice“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/16.059.
- Buchholz, Sabine und Erwin Marsi (2006). „CoNLL-X Shared Task on Multilingual Dependency Parsing“. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City: Association for Computational Linguistics, S. 149–164. URL: <http://www.aclweb.org/anthology/W/W06/W06-2920> (besucht am 1. Juni 2020).
- Carreras, Xavier und Lluís Màrquez (2004). „Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling“. In: *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA: Association for Computational Linguistics, S. 89–97.
- Carreras, Xavier und Lluís Màrquez (2005). „Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling“. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ann Arbor, Michigan: Association for Computational Linguistics, S. 152–164. URL: <http://www.aclweb.org/anthology/W/W05/W05-0620> (besucht am 1. Juni 2020).
- Dagan, Ido, Oren Glickman und Bernardo Magnini (2006). „The PASCAL Recognising Textual Entailment Challenge“. In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Berlin/Heidelberg: Springer Berlin Heidelberg, S. 177–190.
- Doležel, Lubomír (1973). *Narrative Modes in Czech Literature*. Toronto: University of Toronto Press.
- Duyfhuizen, Bernard (2005). „Framed Narrative“. In: *The Routledge Encyclopedia of Narrative Theory*. Hrsg. von David Herman. London/New York: Routledge, S. 186–188.
- Eisenberg, Joshua und Mark Finlayson (2019). „Annotation Guideline No. 1: Cover Sheet for Narrative Boundaries Annotation Guide“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/16.051.
- Escartín, Carla Parra, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way und Chao-Hong Liu (2017). „Ethical Considerations in NLP Shared Tasks“. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, S. 66–73. doi: 10.18653/v1/W17-1608. (Besucht am 1. Juni 2020).
- Fludernik, Monika (2009). *An Introduction to Narratology*. London: Taylor & Francis.
- Füredy, Viveca (1989). „A Structural Model of Phenomena with Embedding in Literature and Other Arts“. In: *Poetics Today* 10.4, S. 745–769.
- Genette, Gérard (1980). *Narrative Discourse – An Essay in Method*. Übers. von Jane E. Lewin. Ithaca, New York: Cornell University Press.
- Genette, Gérard (1988). *Narrative Discourse Revisited*. Ithaca, New York: Cornell University Press.
- Genette, Gérard (1997). *Paratexts*. Cambridge: Cambridge University Press.
- Gius, Evelyn (2015). *Erzählen über Konflikte – Ein Beitrag zur digitalen Narratologie*. Narratologie. Berlin: De Gruyter.
- Gius, Evelyn, Nils Reiter und Marcus Willand, Hrsg. (2019). *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*.
- Hammond, Adam (2020). „Annotation Guideline No. 8: Annotation Guidelines for Narrative Levels“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/001c.11773.

- Jahn, Manfred (2017). „N2.4. Narrative Levels“. In: *Narratology: A Guide to the Theory of Narrative*. Hrsg. von Manfred Jahn. Universität zu Köln.
- Kearns, Edward (2019). „Annotation Guideline No. 2: For Annotating Anachronies and Narrative Levels in Fiction“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/16.052.
- Ketschik, Nora, André Blessing, Sandra Murr, Maximilian Overbeck und Axel Pichler (2020). „Interdisziplinäre Annotation von Entitätenreferenzen“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 204–236.
- Ketschik, Nora, Benjamin Krautter, Sandra Murr und Yvonne Zimmermann (2019). „Annotation Guideline No. 4: Annotating Narrative Levels in Literature“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/16.055.
- Ketschik, Nora, Sandra Murr, Benjamin Krautter und Yvonne Zimmermann (2020). „Zur Theorie von Erzählebenen und ihrer Annotation im digitalen Kontext“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 440–464.
- Kilgariff, Adam und Joseph Rosenzweig (2000). „Framework and Results for English SEN-SEVAL“. In: *Computers and the Humanities* 34.1, S. 15–48. doi: 10.1023/A:1002693207386. (Besucht am 1. Juni 2020).
- Kübler, Sandra (2008). „The PaGe 2008 Shared Task on Parsing German“. In: *Proceedings of the Workshop on Parsing German*. Columbus, Ohio: Association for Computational Linguistics, S. 55–63. URL: <http://www.aclweb.org/anthology/W/W08/W08-1008> (besucht am 1. Juni 2020).
- Lahn, Silke und Jan Christoph Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart, Germany: Metzler.
- Lämmert, Eberhard (1955). „Das Gefüge der Handlungsstränge“. In: *Bauformen des Erzählens*. Stuttgart: Metzler, S. 43–67.
- Mani, Inderjeet (2012). *Computational Modeling of Narrative*. Hrsg. von Graeme Hirst. Bd. 18. Synthesis Lectures on Human Language Technologies. Bristol: Morgan & Claypool. doi: 10.2200/S00459ED1V01Y201212HLT018.
- Margolin, Uri (2014). „Narrator“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.
- Martínez, Matias und Michael Scheffel (2003). *Einführung in die Erzähltheorie*. 8. Aufl. München: C.H.Beck, S. 75–80.
- Mathet, Yann, Antoine Widlöcher und Jean-Philippe Métivier (2015). „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment“. In: *Computational Linguistics* 41.3, S. 437–479. doi: 10.1162/COLI_a_00227.
- McHale, Brian (2014). „Speech Representation“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. 2014. Aufl. Hamburg: Hamburg University Press.
- Nelles, William (1997). *Frameworks: Narrative Levels and Embedded Narrative*. New York: P. Lang.
- Nelles, William (2005). „Embedding“. In: *Routledge Encyclopedia of Narrative Theory*. Hrsg. von David Herman, Manfred Jahn und Marie-Laure Ryan. London/New York: Routledge.
- Neumann, Birgit und Ansgar Nünning (2012). „Metanarration and Metafiction“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.

- Niederhoff, Burkhard (2013). „Focalization“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.
- Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel und Deniz Yuret (2007). „The CoNLL 2007 Shared Task on Dependency Parsing“. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, Czech Republic: Association for Computational Linguistics, S. 915–932. URL: <http://www.aclweb.org/anthology/D/D07/D07-1096> (besucht am 1. Juni 2020).
- Nünning, Ansgar (2004). „On Metanarrative: Towards a Definition, a Typology and an Outline of the Functions of Metanarrative Commentary“. In: *The Dynamics of Narrative Form – Studies in Anglo-American Narratology*. Hrsg. von John Pier. Narratologia. Berlin: De Gruyter, S. 11–58.
- Pagel, Janis, Nils Reiter, Ina Rösiger und Sarah Schulz (2020). „Annotation als flexibel einsetzbare Methode“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 125–141.
- Pichler, Axel und Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 43–59.
- Pier, John (2012). „Metalepsis“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Hamburg: Hamburg University Press.
- Pier, John (2014). „The Living Handbook of Narratology“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Revised version. Hamburg: Hamburg University Press.
- Pier, John (2016). „Narrative Levels“. In: *The Living Handbook of Narratology*. Hrsg. von Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert. Revised version. Hamburg: Hamburg University Press.
- Reiter, Nils (2020). „Anleitung zur Erstellung von Annotationsrichtlinien“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 193–201.
- Reiter, Nils, Gerhard Kremer, Kerstin Jung, Benjamin Krautter, Janis Pagel und Axel Pichler (2020). „Reaching out: Interdisziplinäre Kommunikation und Dissemination“. In: *Reflektierte Algorithmische Textanalyse*. Hrsg. von Nils Reiter, Axel Pichler und Jonas Kuhn. Berlin: De Gruyter, S. 467–484.
- Rickert, Edith (1923). „Some Straws in Contemporary Literature: Fiction in England and America“. In: *The English Journal* 12.8, S. 599–604.
- Rimmon-Kenan, Shlomith (2004). „Narration: Levels and Voices“. In: *Narrative fiction*. Hrsg. von Shlomith Rimmon-Kenan. 2. Aufl. London/New York: Routledge, S. 87–106.
- Rimmon-Kenan, Shlomith (2005). *Narrative Fiction: Contemporary Poetics*. London/New York: Routledge.
- Romberg, Bertil (1962). *Studies in the Narrative Technique of the First-Person Novel*. Stockholm: Almqvist & Wiksell.
- Ryan, Marie-Laure (1986). „Embedded Narratives and Tellability“. In: *Style* 20.3, S. 319–340.
- Ryan, Marie-Laure (1991). *Possible worlds, artificial intelligence, and narrative theory*. Bloomington, Indiana: Indiana University Press.
- Ryan, Marie-Laure (2001). „The Narratorial Functions: Breaking down a Theoretical Primitive“. In: *Narrative* 9.2, S. 146–152.

- Ryan, Marie-Laure (2002). „Stacks, Frames, and Boundaries“. In: *Narrative dynamics: essays on time, plot, closure, and frames*. Columbus, Ohio: Ohio State University Press, S. 366–385.
- Sang, Erik F. Tjong Kim und Sabine Buchholz (2000). „Introduction to the CoNLL-2000 Shared Task Chunking“. In: *Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, S. 127–132. URL: <http://www.aclweb.org/anthology/W00-0726> (besucht am 1. Juni 2020).
- Sang, Erik F. Tjong Kim und Hervé Déjean (2001). „Introduction to the CoNLL-2001 shared task: clause identification“. In: *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*. Toulouse, Frankreich. URL: <http://www.aclweb.org/anthology/W01-0708> (besucht am 1. Juni 2020).
- Schmid, Wolf (2014). *Elemente der Narratologie*. 3. Aufl. Berlin: De Gruyter.
- Sundheim, Beth M. (1993). „The Message Understanding Conferences“. In: *Proceedings of the TIPSTER Text Program: Phase I*. Fredericksburg, Virginia, USA: Association for Computational Linguistics, S. 5. doi: 10.3115/1119149.1119153. (Besucht am 1. Juni 2020).
- Sundheim, Beth M. und Nancy A. Chinchor (1993). „Survey of the Message Understanding Conferences“. In: *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, S. 56–60. URL: <http://www.aclweb.org/anthology/H93-1011> (besucht am 1. Juni 2020).
- Tjong Kim Sang, Erik F. und Fien De Meulder (2003). „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition“. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, S. 142–147. URL: <http://www.aclweb.org/anthology/W03-0419> (besucht am 1. Juni 2020).
- Todorov, Tzvetan (1966). „Les catégories du récit littéraire“. In: *Communications* 8, S. 125–151.
- Wirén, Mats, Adam Ek und Anna Kasaty (2020). „Annotation Guideline No. 7: Guidelines for annotation of narrative structure“. In: *Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels*. doi: 10.22148/16.060.
- Witten, I. H. und Eibe Frank (2005). *Data Mining*. 2. Aufl. Practical Machine Learning Tools and Techniques. Burlington/San Francisco: Morgan Kaufmann Publishers.