Inter Annotator Agreement und Intersubjektivität

Ein Vorschlag zur Messbarkeit der Qualität literaturwissenschaftlicher Annotationen

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Vauth, Michael
michael.vauth@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Annotationen als etablierte Praxis der DH

Die in den Sozialwissenschaften und der Computerlinguistik schon lange etablierte Praxis der computergestützten und häufig kollaborativen manuellen Annotation ist mittlerweile auch im Zentrum der digitalen Geisteswissenschaften angekommen. Deshalb möchten wir unsere Beobachtungen zu einem zentralen Punkt teilen: dem Inter Annotator Agreement bzw. Inter Coder Agreement. Wir betrachten dieses aus der Sicht der Computational Literary Studies (CLS) anhand unseres Projekts „Evaluating Events in Narrative Theory (EvENT)“[1]. In diesem annotieren wir Ereignisse als kleinste Handlungseinheiten in Prosatexten und nutzen die Annotationen, um die Erkennung der Ereignisse zu automatisieren.[2] Diese automatisierte Ereignisanalysen können anschließend für korpusbasierte Untersuchungen zur Ereignishaftigkeit oder generell zur Ereignisstruktur literarischer Texte genutzt werden. Wie wir zeigen möchten, spielt das Inter Annotator Agreement eine vielfältige Rolle in der Arbeit an und mit manuellen Annotationen und sollte möglichst im Einklang mit der Praxis des Erkenntnisgewinns in der Literaturwissenschaft genutzt und weiterentwickelt werden.

Zum Nutzen von Inter Annotator Agreement-Messungen

Es gibt eine Vielzahl von Inter Annotator Agreement-Metriken, die als Maß eingesetzt werden, um die Verlässlichkeit manuell erstellter Annotationen zu beurteilen, die zum Überprüfen einer These oder zur Entwicklung und zum Testen computationeller Modelle genutzt werden (Artstein & Poesio 2008:556). Da bei der Betrachtung des Inter Annotator Agreements von Menschen annotierte Daten – und damit deren Analysen bestimmter Texte (oder anderer Artefakte) – miteinander verglichen werden, ist dies auch literaturwissenschaftlich interessant. Der literaturwissenschaftliche Erkenntnisgewinn basiert nämlich, in Ermangelung objektiver Fakten, ganz wesentlich auf intersubjektiver Übereinstimmung bzw. deren Abgleich.

Grundsätzlich lassen sich fünf Einsatzgebiete von Inter Annotator Agreement-Messungen unterscheiden:

1. Man kann mittels Inter Annotator Agreement-Messung feststellen, wie hoch die *Reliabilität einzelner Annotator*innen* ist. Dies wird etwa genutzt, wenn man aus Ressourcengründen größtenteils nur eine:n Annotator:in die Texte annotieren lässt und einschätzen möchte, ob diese:r die Annotationen in der gewünschten Qualität anfertigt.
2. Bei der *Entwicklung von Guidelines* können Inter Annotator Agreement-Messungen a ls Heuristik zum Aufdecken nicht übereinstimmender Annotationen genutzt werden, deren Erkenntnisse anschließend in die Überarbeitung des Annotationsworkflows und insbesondere der Guidelines einfließen.
3. Auch für eine Bewertung oder auch den Vergleich der *Qualität von Guidelines* werden Inter Annotator Agreement-Verfahren genutzt. Gute Guidelines sollten eine hohe Übereinstimmung zwischen Annotationen ermöglichen. Dies ist in Verfahren wichtig, in denen das annotierte Korpus zum Testen einer Hypothese konzipiert wird. Dort werden nämlich die Guidelines nicht im Annotationsprozess überarbeitet, sondern stehen mit Beginn der Annotation fest (vgl. dazu Artstein & Poesio 2008 bzw. Krippendorff 2004).
4. Desweiteren wird in Fällen, in denen Korpora – wie etwa Referenzkorpora – zur Nachnutzung für weitere Forschung aufgebaut werden, die *Qualität bzw. Validität der Daten* an den Inter Annotator Agreement-Angaben gemessen.
5. Schließlich können Inter Annotator Agreement-Werte auch genutzt werden, um etwas über die *annotierten Phänomene bzw. deren Operationalisierbarkeit* auszusagen. Bei einem hohen Inter Annotator Agreement-Wert kann man von einer geringen Komplexität der annotierten Phänomene bzw. einer guten Operationalisierbarkeit ihrer Bestimmung (in Form der Guidelines) ausgehen. Das Inter Annotator Agreement ist dann ein Maß, das zum einen die Schwierigkeit der Automatisierungsaufgabe beschreibt (je geringer das Agreement, desto anspruchsvoller die Aufgabe) und zum anderen die Qualität der Automatisierung evaluiert.

Für das EvENT-Projekt sind diese Einsatzbereiche unterschiedlich stark von Interesse. Wir nutzen Inter Annotator Agreement für (1) die Reliabilität von Annotator:innen und die (2) Entwicklung von Guidelines. Die (3) Bewertung der Qualität der Guideline spielt im EvENT-Projekt nur eine nachgeordnete Rolle. Im Gegensatz zur Computerlinguistik gibt es nämlich für die CLS bislang mangels Erfahrung keine Inter Annotator Agreement-Werte, an denen man sich orientieren kann. Dasselbe gilt für (4) Datenvalidität und (5) Operationalisierbarkeit.

Inter Annotator Agreement als literaturwissenschaftliches Qualitätskriterium

Literaturwissenschaftliche Befunde basieren meistens weder auf streng formalisierten Schlussfolgerungssystemen[3] noch ist mit ihnen der Anspruch verbunden, eine empirische Wahrheit abzubilden. Die Wissenschaftlichkeit der Befunde wird vielmehr durch ihre “prinzipielle intersubjektive Vermittelbarkeit – einen ‘ *sensus communis*’” garantiert.[4] Literaturwissenschaftliche Analyse

bedeutet in einem ersten Schritt, ohne Wertung "die Feststellung von allgemein beobachtbaren und intersubjektiv anerkennbaren Eigenheiten bestimmter Texte zu fixieren"[5]. Dieser Anspruch der intersubjektiven Vermittelbarkeit von beobachtbaren Texteigenschaften legt nahe, dass Inter Annotator Agreement-Maße geeignete Kandidatinnen für das 'Messen' von Intersubjektivität sind.

Mit Blick auf Intersubjektivität kann man in den fünf genannten Bereichen, in denen Inter Annotator Agreement-Messungen zum Einsatz kommen, feststellen: Im Kontext der Reliabilität von Annotator*innen (Fall 1) geht es um den Abgleich einer an sich aber *intrasubjektiven* Qualität, nämlich die Frage, wie gut Annotator:innen annotieren bzw. welche besonders gut sind. Intersubjektivität spielt hier eine untergeordnete Rolle.

Bei der Entwicklung bzw. Qualität von Guidelines (Fall 2 bzw. 3) geht es hingegen um die Frage, inwiefern eine *Guideline* ein geteiltes Verständnis von Phänomenen unterstützt. Damit geht es um die intersubjektive Übereinstimmung bei der Interpretation der Guidelines, die sich in den Annotationen niederschlägt.

Im Kontext der Qualität bzw. Validität der Daten und der Operationalisierbarkeit von Phänomenen (Fall 4 und 5) steht schließlich die intersubjektive Übereinstimmung bei der Beurteilung der Phänomene im Text im Fokus.

Aus literaturwissenschaftlicher Sicht ist die Intersubjektivität insbesondere in den letzten beiden Fällen abgebildet. Bei der Frage nach Qualität bzw. Validität der Daten und der Operationalisierbarkeit von Phänomenen wird nämlich der Grad der Übereinstimmung zwischen Annotationen auf die oben erwähnten „Eigenheiten bestimmter Texte" bezogen. Die beiden Aspekte sind auch aus computationeller Sicht wichtig, denn sie betreffen die analysierten Phänomene und damit das zentrale Forschungsinteresse vieler literaturwissenschaftlicher Ansätze in den Digital Humanities. Wie bereits angesprochen, fehlen allerdings gerade zu diesen beiden Fällen Erfahrungswerte, auf die zurückgegriffen werden kann. Da die Inter Annotator Agreement-Werte in literaturwissenschaftlichen Annotationsprojekten zudem meist deutlich unter den in anderen Disziplinen gängigen Grenzwerten liegen, können diese nicht sinnvoll genutzt werden. Stattdessen müssen Strategien entwickelt werden, die eine Beurteilung der Annotationsqualität in philologischen Forschungskontexten ermöglichen.

Wir stellen deshalb im Folgenden eine Anpassung des Verfahrens der Annotation und der Inter Annotator Agreement-Messung vor, mit der man diesem Manko in bestimmten Forschungszusammenhängen begegnen kann.

Literaturwissenschaftlich adäquate Inter Annotator Agreement-Messung

Inter Annotator Agreement-Metriken basieren auf differenzierten Formeln, die typischerweise erwartete (Nicht-)Übereinstimmungswerte berücksichtigen und z.T. auch die Gewichtung bestimmter Aspekte der Annotationen zulassen (z.B. durch das Festlegen von Ähnlichkeiten zwischen Kategorien oder die Gewichtung der Segmentierungsentscheidungen). Die Wahl der eingesetzten Metrik sollte in Abhängigkeit von den Eigenschaften der Annotationen getroffen werden. Zu diesen Eigenschaften gehören die Anzahl und Verteilung der genutzten Annotationskategorien, die Häufigkeit, mit der Annotationskategorien auftreten, die Frage, ob die Bestimmung der zu annotierenden Textsegmente Teil der Annotationsaufgabe ist und viele mehr (vgl. dazu Artstein & Poesio 2008 sowie Mathet et al. 2015). Das Problem, vor dem wir zumindest bislang stehen, ist nicht nur, dass es eine ziemliche Herausforderung ist, diese Eigenschaften zu identifizieren, sondern noch mehr, dass uns etablierte Strategien fehlen, um diese zu beurteilen.

Ein wesentlicher Grund dafür ist, dass literaturwissenschaftliche Textanalysen oft Phänomene in den Blick nehmen, die bei näherer Betrachtung keine Merkmale der Textoberfläche sind. Da diese Phänomene nicht direkt an bestimmten Texteigenschaften festgemacht werden können, muss man bei der Operationalisierung auf mit dem Phänomen mutmaßlich zusammenhängende Merkmale zurückgreifen, die sich textlich realisieren.[6] So modellieren wir im EvENT-Projekt die Ereignishaftigkeit von Texten mit der vergleichsweise granularen Annotation von Verbalphrasen, da wir diese in unseren Untersuchungen als kleinste Textspannen identifiziert haben, die auf ein Ereignis referieren können.[7] Wir annotieren also Mikrophänomene auf der Textoberfläche, um ein erzähltheoretisches Makrophänomen zu beschreiben, das sich nicht unmittelbar an der Textoberfläche manifestiert.

Eine Folge dieser indirekten Annäherung an die untersuchten Phänomene ist, dass eine Agreement-Messung mit den üblichen Metriken für bestimmte literaturwissenschaftliche Einsatzgebiete nicht sinnvoll ist, da diese für die Annotation von Textphänomenen wie etwa Wortarten oder semantische Klassen entwickelt wurden.

Nun könnte man versuchen neue, für literaturwissenschaftliche Fragestellungen passende Annotationsmetriken zu entwickeln. Ähnlich hilfreich und leichter umsetzbar ist allerdings eine Anpassung des Operationalisierungsverfahrens an das, was mit bestehenden Metriken gemessen wird.

Konkret sollte man versuchen, die genutzten Annotationskategorien so zu gestalten, dass sie:

1. einen möglichst klaren Textumfang haben sowie möglichst im ganzen Text vorkommen und
2. numerischen Werten belegt werden können.

Beim ersten Punkt ist es erstrebenswert, dass die genutzten Kategorien eine möglichst eindeutig festlegbare Texteinheit umfassen und im ganzen Text vorkommen.[8] Der zweite Punkt bedeutet, dass die genutzten Kategorien möglichst in numerische Werte überführt werden. (Dies ist übrigens in allen Fällen ein hilfreicher Schritt, in welchem es darum geht, quantifizierend mit Annotationen umzugehen.) Dies bedeutet nicht nur, dass man Annotationskategorien in numerische Werte überführt, sondern auch, dass die Werte in Bezug auf ihren Intervall bedeutungshaft sind und es zudem idealerweise auch einen absoluten Nullpunkt gibt, zu dem sie im Verhältnis stehen. Der Grad der Umsetzbarkeit dieser Vorschläge hängt natürlich von der Forschungsfrage und dem untersuchten Phänomen ab.

Eine mögliche Umsetzung dieser Punkte lässt sich an unserem Beispiel verdeutlichen.

Inter Annotator Agreement-Messung im EvENT-Projekt

Ausgehend von den erzähltheoretischen Ereigniskonzepten haben wir im EvENT-Projekt vier Annotationskategorien definiert:

- change_of_state: Die Verbalphrase referiert auf die Zustandsveränderung einer Entität in der erzählten Welt (Diegese)

- process_event: Die Verbalphrase referiert auf einen zeitlichen Vorgang in der erzählten Welt, der keine Zustandsveränderung enthält.
- stative_event: Die Verbalphrase referiert auf einen Sachverhalt in der erzählten Welt, der keine zeitliche Dimension hat.
- non_event: Die Verbalphrase oder ein elliptisches Textsegment referiert nicht auf einen Sachverhalt in der erzählten Welt.

Wir haben also eine syntaktisch weitgehend eindeutige Einheit – die Verbalphrase – identifiziert, die sich als Annotationseinheit eignet und deren Inhalt zur Bestimmung der Kategorisierung geeignet ist. Durch die Ausweitung der non_event-Kategorie auf nicht vollständige Verbalphrasen kann ein Text außerdem durchgängig mit unseren Kategorien annotiert werden kann.

Auch die Überführung der kategorialen Skalierung in eine numerische Skalierung basiert auf dem literaturwissenschaftlichen Verständnis der Kategorien. Entsprechend dem literaturwissenschaftlichen Ereignisverständnis nehmen wir an, dass diese vier Kategorien in unterschiedlichem Maß die Ereignishaftigkeit eines Textes konstituieren: Zustandsveränderungen, aber auch Bewegungs- und Kommunikationsvorgänge tun dies in stärkerem Maß als Landschafts-, Raum- oder Figurenbeschreibungen, die in vielen erzählenden Texten eher Expositionsfunktionen erfüllen.[9] Aus diesem Grund haben wir die Narrativität der Annotationskategorien mit folgenden Werten festgelegt:

- change_of_state: 7
- process_event: 5
- stative_event: 2
- non_event: 0

Doch dies war noch nicht ausreichend, um ein Agreement zu erzielen, welches aus computerlinguistischer Sicht gut ist. Hinzu kommt, dass die Agreement-Werte unsere Intuition über die Qualität der Annotationen nicht widerspiegelten (vgl. Tabelle 1).

Tab. 1: Inter Annotator Agreement-Werte für annotierte Texte

	Erdbeben in Chili	Krambambuli	Effi Briest
Cohen's κ	0.73	0.63	0.58
Krippendorf's α	0.73	0.63	0.58

Deshalb haben wir unser Vorgehen entsprechend weiterentwickelt. Der Schlüssel zu einer aussagekräftigeren Inter Annotator Agreement-Perspektive lag in der Erkenntnis, dass uns die Entwicklung von Ereignishaftigkeit im Textverlauf und entsprechend Narrativitätsverläufe interessieren.

Wir haben deshalb nicht nur die Ergebnisse der Annotationen als Verlauf visualisiert, sondern auch entschieden, die Einschätzung des Inter Annotator Agreement – ebenso wie übrigens die Qualität der automatisierten Erkennung von Ereignissen – anhand von Verläufen vorzunehmen.

Für die Darstellung des Narrativitätsverlaufs wurden die Werte der Annotationen innerhalb eines Textabschnitts anhand der Narrativitätswerte der umliegenden 50 Verbalphrasen mit einer Kosinusgewichtung geglättet. Die Kosinusgewichtung sorgt dabei dafür, dass näher liegende Textsegmente einen stärkeren Einfluss auf den Narrativitätswert des untersuchten Textsegments haben.

Auf Grundlage dieser Zuweisungen konnten wir die Narrativitätsverläufe in Einzeltexten wie in Abbildung 1 untersuchen:

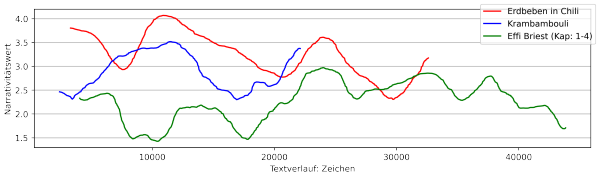

Abb. 1: Narrativitätsverläufe zu drei Prosatexten aus dem Event-Projekt.

Um die Stabilität des Verfahrens zu prüfen, haben wir mit der Zuweisung der Zahlen zu den Kategorien experimentiert, dabei aber ihre Anordnung gemäß ihrer Narrativität nicht verändert. Eine umfassende Evaluation steht noch aus, aber die bisherigen Versuche deuten darauf hin, dass die Narrativitätsverläufe dabei strukturell nicht stark variieren (vgl. Abbildung 2).

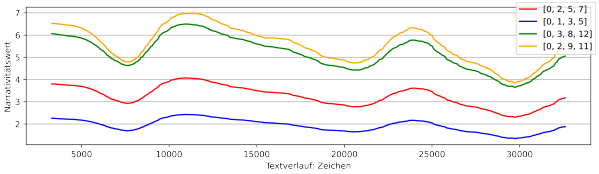

Abb. 2: Narrativitätsverläufe von Kleists *Das Erdbeben in Chili* bei variierenden Skalierungen der Ereignistypen. Die Zahlenlisten in der Legende geben die verwendeten Werte für non_events, stative_events, process_events und change_of_states in dieser Reihenfolge an.

Wir konnten also auf der Grundlage unserer Wertzuweisung für die Ereignistypen die Annotationen der unterschiedlichen Annotator:innen miteinander vergleichen (vgl. Abbildung 3).

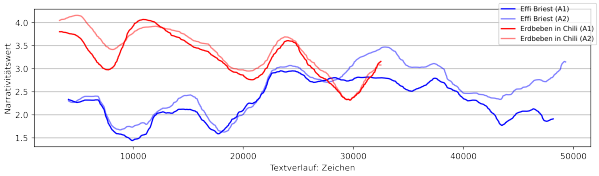

Abb. 3: Narrativitätsverläufe in Kleists *Das Erdbeben in Chili* und *Effi Briest* zum Vergleich der beiden Annotator:innen (A1 und A2).

Die Narrativitätsgraphen ähneln sich deutlich stärker als es angesichts des eher niedrigen Inter Annotator Agreement zu vermuten gewesen wäre. Das drückt sich auch in der mittleren bis starken Korrelation der Graphen aus (vgl. Tabelle 2).[10]

Tab. 2: Korrelation (Pearson) der Narrativitätsgraphen der annotierten Texte.

	Erdbeben in Chili	Krambambuli	Effi Briest
Korrelation (Pearson)	0.94	0.81	0.8

Unsere Annäherung an ein Inter Annotator Agreement, das auf die Modellierung eines literarischen Phänomens ausgerichtet ist, scheint also unsere literaturwissenschaftlich fundierte Intuition besser abzubilden als gängige Inter Annotator Agreement-Metriken.

Dafür sind zwei Aspekte entscheidend:

1. Wenn die Annotator:innen systematische Fehler machen, die sich auf ein unterschiedliches Verständnis der Annotations-

richtlinien zurückführen lassen, schlägt sich das durch die Parametrisierung nur insofern wieder, als der Narrativitätswert im Durchschnitt etwas niedriger oder höher ist. Die strukturellen Eigenschaften des Graphen berührt das kaum. Ein Beispiel: In den vier längeren Abschnitten aus *Effi Briest* zeigt sich in 3 von 4 Abschnitten eine Tendenz von Annotator:in 1 (A1) häufiger non_events und seltener stative_events zu annotieren. Annotator:in 2 (A2) identifiziert in drei Abschnitten seltener process_events (vgl. Tabelle 3).

2. Das Glättungsverfahren, das ein notwendiger Schritt ist, um die Parametrisierung der Mikrophänomene zur Modellierung von Makrophänomenen umzusetzen, nivelliert Flüchtigkeitsfehler bei der Annotation und das hinsichtlich der Segmentierung und der Klassifizierung.

	Teil 1		Teil 2		Teil 3		Teil 4	
	A1	A2	A1	A2	A1	A2	A1	A2
non_event	0.43	0.37	0.38	0.36	0.43	0.40	0.45	0.50
stative_event	0.28	0.38	0.31	0.38	0.31	0.40	0.28	0.28
process_event	0.37	0.25	0.30	0.30	0.25	0.20	0.25	0.22
change_of_state	0.02	0.01	0.01	0.00	0.00	0.00	0.01	0.00

Tab. 3: Bias der beiden Annotator:innen (A1 und A2) bei der Annotation von Ereignistypen in *Effi Briest* (vier Teile mit einem Umfang von je 4–7 Kapitel).

Durch dieses Vorgehen gelingt es uns, den Fokus auf das eigentlich untersuchte Phänomen – in unserem Fall die Ereignishaftigkeit von erzählenden Texten – zu richten. Damit lässt sich die Intersubjektivität der Analysen besser messen als anhand der Annotationen, die das Phänomen anhand von Oberflächenphänomenen (Verbalphrasen) operationalisieren und die im Kontext von gängigen Inter Annotator-Metriken entsprechend nur bedingt aussagekräftig sind. Hinzu kommt, dass es zwei wichtige Fehlerquellen bei literaturwissenschaftlichen Annotationen – nämlich einfache Fehler sowie divergierende Voranalysen (vgl. Gius & Jacke 2017) – ausgleicht.

Fußnoten

1. Das Projekt wird im DFG-Schwerpunktprogramm Computational Literary Studies (SPP 2207) gefördert und wird seit 09/2020 an der Technischen Universität Darmstadt und der Universität Hamburg durchgeführt.
2. vgl. zu den ersten Automatisierungsergebnissen Hans Ole Hatzel, Michael Vauth, Chris Biemann und Evelyn Gius (2021).
3. vgl. Danneberg und Albrecht (2016), insbesondere S. 6-8.
4. vgl. Stöckmann 2013, S. 475.
5. vgl. Fricke et al. 2000, S. 447.
6. Vgl. dazu die sogenannten‘instrumental variables’ v on Graham Sack, die eingesetzt werden, wenn keine die eigentlichen Phänomene nicht messbar sind (Moretti 2013:104).
7. Vgl. dazu unsere Guideline (Vauth & Gius 2021) unter http://doi.org/10.5281/zenodo.5078175.
8. Dadurch werden potentielle Probleme mit den Aspekten *Unitizing* und *Sporadicity* (Mathet et al. 2015) verringert. Die weiteren für Inter Annotator Agreement relevanten Aspekte *Categorization*,
9. Es gibt einige zusätzliche Ereigniseigenschaften gibt, die die Ereignishaftigkeit eines Textes determinieren und in unserem Annotationsschema Berücksichtigung erfahren haben. Auf ihnen liegt in diesem Beitrag allerdings nicht unser Fokus.
10. Für den Vergleich filtern wir die Annotationen der beiden Annotator:innen so, dass sie einen etwa gleichen Startpunkt haben (+/- 3 Zeichen). Wie bei üblichen Verfahren der Inter Annotator Agreement-Messung wird so verhindert, dass zwei Annotationen miteinander verglichen werden, die sich nicht auf die gleiche Textspanne beziehen.

Bibliographie

Artstein, Ron, und Massimo Poesio. 2008. „Inter-Coder Agreement for Computational Linguistics“. *Computational Linguistics* 34 (4): 555–96. https://doi.org/10.1162/coli.07-034-R2.

Danneberg, Lutz, und Andrea Albrecht. 2016. „Beobachtungen zu den Voraussetzungen des hypothetisch-deduktiven und des hypothetisch-induktiven Argumentierens im Rahmen einer hermeneutischen Konzeption der Textinterpretation“. *Journal of Literary Theory* 10 (1): 1–37. https://doi.org/10.1515/jlt-2016-0001.

Fricke, Harald, Klaus Grubmüller, Jan-Dirk Müller, und Klaus Weimar, Hrsg. 2000. *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. 3., Neubearb. Aufl. Berlin: De Gruyter.

Gius, Evelyn, und Janina Jacke. 2017. „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis“. *International Journal of Humanities and Arts Computing* 11 (2): 233–54. https://doi.org/10.3366/ijhac.2017.0194.

Krippendorff, Klaus. 2004. „Reliability in Content Analysis: Some Common Misconceptions and Recommendations“. *Human Communication Research* 30 (3): 411–33. https://doi.org/10.1093/hcr/30.3.411.

Mathet, Yann, Antoine Widlöcher, und Jean-Philippe Métivier. 2015. „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment“. *Computational Linguistics* 41 (3): 437–79. https://doi.org/10.1162/COLI_a_00227.

Moretti, Franco. 2013. „‘Operationalizing’. Or, the Function of Measurement in Literary Theory“. *New Left Review*, Nr. 84 (Dezember): 103–19.

Stöckmann, Ingo. 2013. „Ästhetik“. *Handbuch Literaturwissenschaft*, herausgegeben von Thomas Anz, 1:465–91. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-01271-5.

Vauth, Michael, und Gius, Evelyn. 2021. *Richtlinien für die Annotation narratologischer Ereigniskonzepte*, Juli. https://doi.org/10.5281/ZENODO.5078175.

Vauth, Michael, Hatzel, Hans Ole, Biemann, Chris, und Evelyn Gius. 2021. “Automated Event Annotation in Literary Texts” Workshop on Computational Humanities Research 2021. *CHR 2021: Computational Humanities Research Conference*, 333–45. Amsterdam, The Netherlands. http://ceur-ws.org/Vol-2989/short_paper18.pdf