Narrativität und Handlung: Zum Verhältnis von Handlungs-zusammenfassungen und relevanten Ereignissen

Hatzel, Hans Ole
hans.ole.hatzel@uni-hamburg.de
Universität Hamburg

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

Stiemer, Haimo
stiemer@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Biemann, Chris
christian.biemann@uni-hamburg.de
Universität Hamburg

Relevante Ereignisse in Erzähltexten

Welche Ereignisse in Erzähltexten sind besonders relevant? Diese Frage wird in der Literaturwissenschaft im Kontext von verschiedenen Konzepten verhandelt. So können relevante Ereignisse identifiziert werden, indem man die für die Textinterpretation als besonders wichtig erachteten Stellen (so genannte "Schlüsselstellen") betrachtet (Arnold & Fiechter, 2022). Auf die Rezeption orientiert sind ebenfalls die empirische Leser:innenforschung (Groeben 1977; Miall & Kuiken 2001) oder die Rezeptionsästhetik (Iser 1976). Steht hingegen der Text im Fokus, kann die Frage nach der Wichtigkeit von Ereignissen in Bezug auf Ereignishaftigkeit oder die so genannte Erzählwürdigkeit untersucht werden (z. B. Hühn 2014; Baroni 2012). Allen Ansätzen gemeinsam ist, dass sie bestimmte Qualitäten von Texten bzw. Textbestandteilen betrachten.

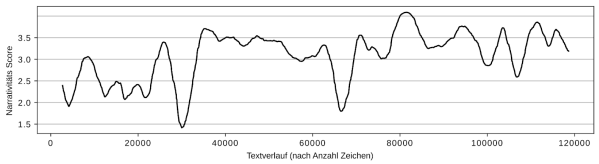

Abb. 1: Narrativitätskurve für Kafkas Die Verwandlung (Vauth et al., 2021)

In unserem EvENT-Projekt wurde bereits die Identifikation und Klassifikation von Ereignissen anhand von textuellen Merkmalen vorgenommen. Den hierbei entwickelten vier Ereignistypen haben wir Werte entsprechend ihrer Ereignishaftigkeit zugewiesen und darauf basierend Narrativitätskurven erzeugt, die den Verlauf der Ereignishaftigkeit über einen Text abbilden (siehe Abb. 1).[1]

Im Fortgang des Projektes wollen wir überprüfen, inwiefern zwischen diesen grundlegenden Ereignistypen, die auch unter dem Konzept "Event I" subsumiert werden können, und besonders erzählwürdigen Ereignissen bzw. so genannten Events II, eine Verbindung besteht.[2] Nimmt man an, dass ein Event II ein Event I mit weiteren Qualitäten ist, so ist die Bestimmung von Events II insofern komplexer, als sie zusätzliches Wissen erfordert – etwa um Regelmäßigkeiten und Auffälligkeiten in der fiktionalen Welt, aber auch um den Kontext. Mit anderen Worten, wir wollen die von uns im Vorlauf vorgenommene textoberflächenbasierte Detektion von Ereignissen dahingehend prüfen inwiefern mit ihr auch jene Textstellen erfasst werden, die in Handlungszusammenfassungen als besonders handlungsrelevant bzw. erzählwürdig gelten.

Was bereits mit unseren bestehenden Daten und ohne die weitere Operationalisierung des Event II-Konzepts möglich ist, ist der Abgleich unserer Annotationen mit als besonders handlungsrelevant markierten Textstellen. Diesen Vergleich stellen wir im vorliegenden Beitrag an, indem wir unsere Annotationen von Ereignissen in literarischen Texten mit Zusammenfassungen der entsprechenden Texte abgleichen.

Semiprofessionelle, professionelle und nutzer:innengenerierte Handlungszusammenfassungen

Wir gehen davon aus, dass Textstellen durch ihre Erwähnung in Zusammenfassungen als für die Handlung wichtig markiert werden. Für den Abgleich dieser Textstellen mit unseren Narrativitätsverläufen nutzen wir drei Typen von Zusammenfassungen: (1) semiprofessionelle Zusammenfassungen, die Studierende der Literaturwissenschaft verfasst haben, (2) professionelle Zusammenfassungen aus Kindlers Literatur Lexikon und (3) nutzer:innengenerierte Zusammenfassungen der Online-Enzyklopädie Wikipedia. Die Verwendung dieser verschiedenen Zusammenfassungstypen zielt darauf ab, zu analysieren, welche Aspekte bzw. Textstellen für die je-

weiligen Zusammenfassungstypen relevant sind und dadurch Rückschlüsse auf ihre Qualität zu ziehen.

Die (1) Zusammenfassungen der Studierenden waren Teil der Studienleistungen in einem Seminar. Sie wurden als explizit auf die Handlung bezogene Zusammenfassungen verfasst, die eine maximale Länge von 20 Sätzen haben durften. Außerdem wurden die Studierenden aufgefordert, keine Hilfsmittel (wie Zusammenfassungen auf Wikipedia oder aus Literaturlexika) zu nutzen. Für die vier genutzten Primärtexte *Das Erdbeben in Chili* (Heinrich von Kleist, 1807), *Die Judenbuche* (Annette von Droste-Hülshoff, 1842), *Krambambuli* (Marie von Ebner-Eschenbach, 1896) und *Die Verwandlung* (Franz Kafka, 1915) haben wir jeweils 11, 9, 11 und 10 Zusammenfassungen.

Aus den Beiträgen des (2) Kindler-Literaturlexikons und von (3) Wikipedia wurden nur jene Passagen verwendet, die sich auf die Handlung in den Primärtexten beziehen. Passagen, die sich der Autor:in, Rezeption oder Interpretation widmen, wurden nicht berücksichtigt. Durch eine kollaborative Annotation der einzelnen Sätze wurde deren Bezug auf die Handlung des Textes annotiert und ein entsprechender Goldstandard erstellt, der den weiteren Analysen zugrunde liegt. So wurde sichergestellt, dass alle drei Zusammenfassungstypen handlungsorientiert sind. Die Zusammenfassungen wurden dahingehend annotiert, dass jeder Satz der Zusammenfassung mit einer Referenz auf alle Spannen des Primärtextes versehen wurde, auf die er Bezug nimmt.[3] Als logische Einheiten wurden entsprechend die bereits vorliegenden Annotationen aller Verbalphrasen in Bezug auf die vier Ereignistypen genutzt.[4] Dabei wurden konkret jeweils eine oder mehrer Spannen im Originaltext als zugehörig annotiert, die Sequenzen relevanter Ereignisse enthalten; ein Satz der Zusammenfassung kann also mehrere Passagen im Originaltext referenzieren.

Evaluation der Zusammenfassungen

Um die Qualität der Zusammenfassungen und mögliche Unterschiede zwischen den drei Zusammenfassungstypen zu analysieren, evaluieren wir deren Ähnlichkeit. Dazu nutzen wir drei Metriken, mit denen implizit drei unterschiedliche Auffassungen von Ähnlichkeit verbunden sind: Eine weitgehend lexikalische (N-Gramme), eine Metrik auf Basis distributioneller Semantik (Word Embeddings) und eine, die sich weitgehend von der sprachlichen Struktur löst und inhaltsbezogene Vergleiche vornimmt (adaptierte Pyramiden-Methode).

Wir nehmen zunächst an, dass die semiprofessionellen Zusammenfassungen durchweg handlungsbezogen sind. Deshalb vergleichen wir jeweils eine semiprofessionelle Zusammenfassung mit allen anderen semiprofessionellen und jede Zusammenfassung aller anderen Typen mit allen semiprofessionellen.

N-Gramm-basierte Ähnlichkeit

Als erstes berechnen wir BLEU- (Papineni et al., 2002) und ROUGE-Scores (Lin et al., 2004), die Ähnlichkeiten unter Zusammenfassungen als N-Gramm-Ähnlichkeit abbilden.

Wir gewichten BLEU-{1,2,3} und ROUGE-{1,2,3} jeweils gleich und quantifizieren so die Überlappung von 1-, 2- und 3-Grammen zwischen den unterschiedlichen Texten und geben für BLEU die Precision und ROUGE den F1-Score an.

Anhand der Scores in Tab. 1 und Tab. 2 wird ersichtlich, dass die semiprofessionellen Zusammenfassungen nahezu durchgehend die höchsten Ähnlichkeitswerte aufweisen.

Tab. 1: BLEU-{1,2,3} Scores für Ähnlichkeiten. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,27	0,23	0,24	0,22
professionell	0,15	0,2	0,19	0,09
nutzer:innengeneriert	0,14	0,22	0,24	0,1

Tab. 2: ROUGE-{1,2,3} F-Scores. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,51	0,56	0,6	0,47
professionell	0,58	0,53	0,63	0,43
nutzer:innengeneriert	0,34	0,49	0,55	0,31

Word-embedding-basierte Ähnlichkeit

N-Gramm basierte Metriken haben den Nachteil, dass kleine Unterschiede in der Wortwahl zu einer deutlich geringeren Ähnlichkeit führen können. Um Vergleiche stärker auf die Semantik zu fokussieren, wurde mit BERTScore (Zhang et al., 2020) eine embedding-basierte Methode etabliert. Wenden wir diese auf unsere Texte an, zeigt sich ein deutlich geringerer Unterschied der Zusammenfassungstypen (siehe Tab. 3). Dies weist darauf hin, dass Unterschiede in der N-Gramm-basierten Bewertung zu einem großen Teil auf Unterschiede in der Wortwahl zurückzuführen sind.

Tab. 3: BERTScore F-Werte. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,74	0,71	0,74	0,72
professionell	0,69	0,71	0,74	0,69
nutzer:innengeneriert	0,72	0,68	0,74	0,72

Inhaltsbasierte Ähnlichkeit

Für den letzten Vergleich der Zusammenfassungen adaptieren wir die Pyramiden-Methode, die für die automatische Evaluation maschinell generierter Zusammenfassungen entwickelt wurde (Nenkova et al., 2004). Die Zusammenfassungen werden auf Basis von sogenannten Summary Content Units (SCU) mit Referenzzusammenfassungen verglichen. Eine SCU repräsentiert dabei eine semantische, inhaltliche Aussage aus dem Zusammengefassten. Die namensgebende Pyramide reprä-

sentiert dabei das Vorkommen der unterschiedlichen SCUs in der Menge der Referenzzusammenfassungen, wobei die Höhe der Pyramide n der Anzahl der Referenzzusammenfassungen entspricht. Dabei ist in der Regel eine Verteilung zu beobachten, die tatsächlich eine Pyramide aufbaut: eine SCU taucht in allen n Texten auf und bildet die Spitze, einige wenige SCUs tauchen in n-1 Texten auf usw. bis in der letzten Stufe SCUs auftauchen, die nur in einer Zusammenfassung vorkommen. Eine zu evaluierende Zusammenfassung sollte nun, um ihren Pyramiden-Score zu maximieren, SCUs aus hohen Schichten der Pyramide enthalten. Dabei erhält die oberste Schicht das Gewicht n, sodass jede SCU aus dieser Schicht n Punkte gibt. Der Score einer Zusammenfassung wird als Anteil der tatsächlichen Punkte an denen der optimalen Zusammenfassung der gleichen Länge (in SCUs) angegeben. Entsprechend sind die Pyramiden-Scores reelle Zahlen im Intervall 0 bis 1, wobei 1 eine perfekte Zusammenfassung beschreibt.

Wir passen die Pyramiden-Methode in zwei Punkten an unsere Fragestellung an. Zum einen haben wir keine Referenztexte, sondern benutzen das Verfahren zum Vergleich verschiedener Zusammenfassungen. Zum anderen enthalten unsere Daten keine SCUs, diese werden deshalb über Textspannen approximiert. Dafür nehmen wir zunächst an, dass jede Spanne des Textes eine Menge von SCUs enthält, die insofern eindeutig ist, als sie keine Schnittmenge mit den SCUs anderer, disjunkter Textabschnitte hat. Insofern kann jede Textspanne auf eine oder mehrere SCUs abgebildet werden. Textspannen werden derart in Unterspannen zerlegt, dass Spannen sich nur überlagern, wenn sie identisch sind. Somit erhalten wir Spannen, die gemäß unserer Annahme semantisch eindeutig sind (siehe Abb. 2). Eine Textspanne kann nach unseren Annahmen mehrere SCUs enthalten die wir als eine behandeln, dies entspricht einer Ereignis Modellierung in gröberer Granularität.

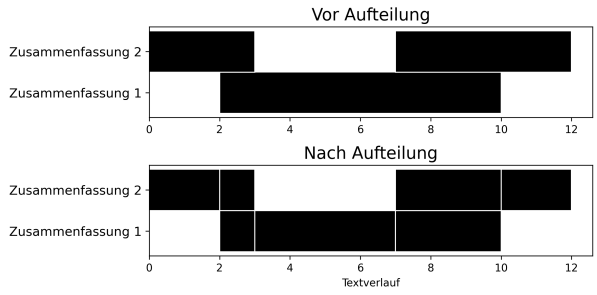

Abb. 2: Segmentierung von zwei Annotationsspannen in semantisch eindeutige SCUs. Spannen werden derart zerlegt, dass sich nur noch gleiche Spannen überhaupt überschneiden.

Die Auswertung der Zusammenfassungen mit der Pyramiden-Methode ist in Tab. 4 zu sehen. Nahezu alle semiprofessionellen Zusammenfassungen liegen dabei über dem Wert 0,70 (siehe Abb. 3), während die anderen Zusammenfassungen im Vergleich zum Mittelwert schlechter abschneiden (siehe Tab. 4).

Tab. 4: Pyramiden-Scores für unterschiedliche Zusammenfassungen. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,85	0,80	0,84	0,80
professionell	0,73	0,52	0,75	0,55
nutzer:innengeneriert	0,71	0,70	0,81	0,62

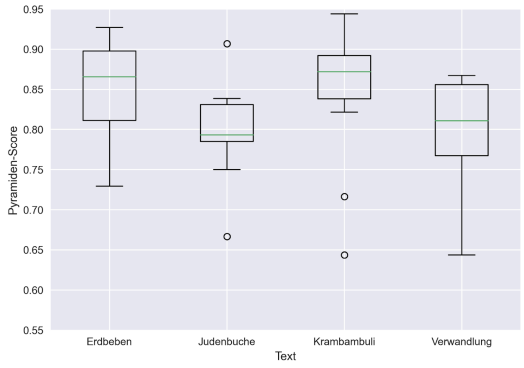

Abb. 3: Verteilung der Pyramiden-Scores für die semiprofessionellen Zusammenfassungen.

Ähnlichkeiten zwischen den Zusammenfassungen

Insgesamt wird so deutlich, dass eine Betrachtung auf Pyramiden-Ebene Unterschiede offenbart, die zwar denen der N-Gramm-Methoden ähnlich, jedoch nicht grundsätzlich anhand oberflächlichen maschinellen Textauswertungen (z.B. BERTScore) festzumachen sind.

Narrativität und Handlung

Wir wollen nun evaluieren, wie Erzählwürdigkeit, repräsentiert durch die Handlungszusammenfassungen, und Narrativität, repräsentiert durch unsere Narrativitätsgraphen, zusammenhängen. Wir überprüfen dafür, ob der Teil des Originaltextes, auf den sich die Zusammenfassungen beziehen, einen großen Narrativitätswert aufweist. Als erste Analyse berechnen wir dazu den Narrativitätswert der in der Zusammenfassung referenzierten Passagen. Wir setzen diesen ins Verhältnis zum erwarteten Gesamtscore, gegeben der Länge der in der Zusammenfassung enthaltenen Textstellen (in Ereignissen).[5] Somit ergibt sich im Mittel ein Wert von 1,0 bei zufälliger Auswahl der Passagen. Ein Wert > 1,0, hingegen heißt, dass in der Zusammenfassung referenzierte Passagen mehr Narrativität aufweisen als nicht referenzierte Passagen. Dies ist tatsächlich für alle Zusammenfassungstypen der Fall (siehe Tab. 5).

Auch ein Vergleich von Ereignissen, die in Zusammenfassungen genannt werden, mit jenen die es nicht werden, bestätigt dies anhand der Narrativitätswerte: im Mittel 3,13 für genannte und 2,86 für nicht genannte Ereignisse.

Tab. 5: Faktoren des erwarteten Narrativitätswerts. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert (inklusive der Standardabweichung) angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	1,04±0,06	1,02±0,09	1,04 ±0,07	1,06±0,10
professionell	1,00	1,03	1,02	1,05
nutzer:innengeneriert	1,06	1,12	1,07	1,08

Für den Vergleich der Ausschläge der Narrativtiätskurven verwenden wir den Gipfelprominenzfaktor. Dabei handelt es sich um ein Maß, welches die Wichtigkeit eines Ausschlags und damit seinen Wert im Vergleich zum umliegenden Kurvenverlauf quantifiziert. Für diesen Vergleich werden alle lokalen Maxima in der cosinusgeglätteten Narrativitätskurve (window size=50) berücksichtigt und für jedes lokale Maximum wird die Gipfelprominenz berechnet.[6] Jedes Ereignis erhält nun, wenn es ein lokales Maximum darstellt, den Wert der Gipfelprominenz, andernfalls den Wert 0. Nun wird wie oben verfahren und der erwartete Prominenzwert mit dem tatsächlichen verglichen. Es wird der erwartete Wert, also die durchschnittliche Gipfelprominenz des Graphen, ins Verhältnis zur tatsächlich vorgefundenen Gipfelprominenz des betrachteten Segments gesetzt. Lokale Maxima sind durch das Smoothing relativ selten, dementsprechend ist die Streuung der Werte deutlich größer. Dies erschwert die Interpretation der Werte. Interessant aber ist, dass in einigen Fällen der Faktor deutlich über 1 liegt (vgl. Tab. 7), wobei dies für die nutzer:innengenerierten Zusammenfassungen durchweg der Fall ist. Abb. 4 veranschaulicht trotz der starken Varianz erkennbare Unterschiede zwischen den Originaltexten.

Tab. 6: Gipfelprominenzfaktoren

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,91±0,52	0,68±0,65	0,90±0,57	1,45±0,78
professionell	0,09	1,15	1,42	1,38
nutzer:innengeneriert	1,23	1,75	1,28	1,1

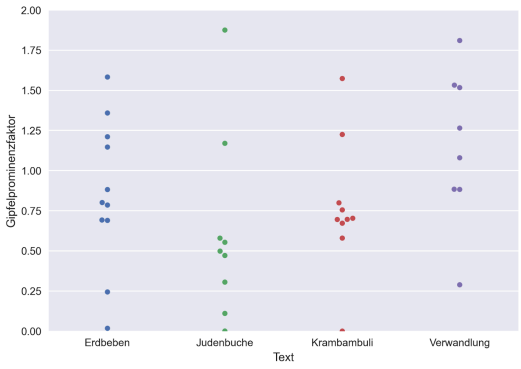

Abb. 4: Gipfelprominenzfaktoren der semiprofessionellen Zusammenfassungen

Narrativität und Handlung: Ein kurzes Fazit

Die vorgestellten Ergebnisse deuten darauf hin, dass die Nutzung von Zusammenfassungen für die weitere Arbeit mit Ereignissen und Ereignishaftigkeit produktiv ist. Hervorzuheben ist, dass wir einen Relevanzzusammenhang zwischen Ereignissen und Handlung nachweisen konnten, der auf einer Operationalisierung ersterer anhand der sprachlichen Oberfläche aufbaut. Damit kann unser Ereigniskonzept mit reduziertem Handlungsbezug anhand von handlungsbezogenen Informationen aus Zusammenfassungen weiterentwickelt werden. Die bereits umgesetzte, vergleichsweise erfolgreiche Automatisierung der Ereigniserkennung und damit der Narrativitätsverläufe wird nun in Bezug auf die handlungsbezogene Relevanz von Ereignissen erweiterbar, ohne dass Handlungsinformationen mühevoll manuell für die einzelnen Ereignisse bestimmt werden müssen. Dafür erscheint es vielversprechend, den Handlungsbezug von Zusammenfassungen weiter zu evaluieren und dabei ein Verfahren zu entwickeln, das besonders relevante Stellen identifizieren kann.

Danksagung

Dieser Beitrag entstand im von der DFG im Schwerpunktprogramm Computational Literary Studies (SPP 2207) geförderten Projekt „Evelautaing Events in Narrative Theory“ (EvENT).

Fußnoten

1. Dabei hat der Ereignistyp, der Zustandsveränderungen beschreibt, den höchsten Wert, gefolgt von weiteren weniger ereignishaften Kategorien. Vgl. zu den Kategorien und zur Bestimmung der Ereignistypen die Annotationsrichtlinien Vauth & Gius (2021) und zu den Narrativitätskurven Vauth et al. (2021), für den Classifier Hatzel (2022).
2. Zur Diskussion der beiden narratologischen Ereigniskonzepte und ihrem Verhältnis vgl. Hühn (2014).
3. Trotz des mit diesen beiden Ansätzen erreichten hohen Handlungsbezugs haben 13% der Sätze keine entsprechend annotierte Spanne im Originaltext (18% bei den professionellen und 9% bei den semiprofessionellen Zusammenfassungen).
4. Die Daten wurden in Vauth & Gius (2022) publiziert.
5. Der Gesamtscore ist also jener, der bei der zufälligen Auswahl der Events aus dem Text im Durchschnitt zustande kommt. ‘1‘ bedeutet also, dass der Score zufällig ausgewählten Events entspricht, ‘2‘ heißt, dass der Score doppelt so hoch ist wie bei zufälligen Events.
6. Für die Berechnung wurde SciPy verwendet: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.peak_prominences.html#scipy.signal.peak_prominences
7. Für die Berechnung wurde SciPy verwendet: https://docs.scipy.org/doc/scipy/reference/gene-

rated/scipy.signal.peak_prominences.html#scipy.signal.peak_prominences

Bibliographie

Arnold, Frederik, und Benjamin Fiechter. 2022. „Lesen, was wirklich wichtig ist. Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse". In DHd2022. Potsdam, Deutschland. https://doi.org/10.5281/zenodo.6327917

Arnold, Heinz Ludwig, Hrsg. 2020. Kindlers Literatur Lexikon (KLL). Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05728-0.

Baroni, Raphaël. 2012. „Tellability". In the living handbook of narratology, herausgegeben von Peter Hühn, John Pier, Wolf Schmid, und Jörg Schönert. Hamburg: Hamburg University Press. http://hup.sub.uni-hamburg.de/lhn/index.php?title=Tellability&oldid=1577.

Gius, Evelyn, und Michael Vauth. 2022. „Inter Annotator Agreement und Intersubjektivität". In DHd2022, 147-151. Potsdam, Deutschland.

Groeben, Norbert. 1977. Rezeptionsforschung als empirische Literaturwissenschaft: Paradigma- durch Methodendiskussion an Untersuchungsbeispielen. Empirische Literaturwissenschaft. Bd. 1. Königstein/Ts.: Athenäum.

Hatzel, Hans Ole. 2022. Event Narrativity Classifier. Zenodo. https://doi.org/10.5281/zenodo.6821142.

Hühn, Peter. 2014. „Event and Eventfulness". In the living handbook of narratology, herausgegeben von Peter Hühn, John Pier, Wolf Schmid, und Jörg Schönert. Hamburg: Hamburg University Press. https://www.lhn.uni-hamburg.de/node/39.html.

Iser, Wolfgang. 1976. Der Akt des Lesens. Theorie ästhetischer Wirkung. München: Fink.

Lin, Chin-Yew. 2004. „ROUGE: A Package for Automatic Evaluation of Summaries". In Text Summarization Branches Out, 74–81. Barcelona, Spanien: Association for Computational Linguistics. https://aclanthology.org/W04-1013.

Miall, David, und Don Kuiken. 2001. "Shifting perspectives: Readers' feelings and literary response". In New Perspectives on Narrative Perspective, herausgegeben von Willi Van Peer und Seymour Chatman, 289-301. Albany: SUNY Press.

Nenkova, Ani, und Rebecca Passonneau. 2004. „Evaluating Content Selection in Summarization: The Pyramid Method". In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 145–52. Boston, Massachusetts, USA: Association for Computational Linguistics. https://aclanthology.org/N04-1019.

Papineni, Kishore, Salim Roukos, Todd Ward, und Wei-Jing Zhu. 2002. „BLEU: a Method for Automatic Evaluation of Machine Translation". In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–18. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Vauth, Michael, und Gius, Evelyn. 2021. „Richtlinien für die Annotation narratologischer Ereigniskonzepte". Zenodo. https://doi.org/10.5281/zenodo.5078174.

Vauth, Michael, und Evelyn Gius. 2022. forTEXT/EvENT_Dataset: v.1.1 (Version v.1.1). Zenodo. https://doi.org/10.5281/ZENODO.6406568.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, und Chris Biemann. 2021. „Automated Event Annotation in Literary Texts". In CHR 2021: Computational Humanities Research Conference, 333–45. Amsterdam, Niederlande. http://ceur-ws.org/Vol-2989/short_paper18.pdf.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, und Yoav Artzi. „BERTScore: Evaluating Text Generation with BERT." In International Conference on Learning Representations. Online, 2020. https://openreview.net/forum?id=SkeHuCVFDr.