

Beitrag aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft

Autor*in:
Sabine Bartsch

Kontakt: sabine.bartsch@tu-darmstadt.de
Institution: Technische Universität Darmstadt
GND: [122348839X](#) ORCID: [0000-0001-7379-2158](#)

Autor*in:
Evelyn Gius

Kontakt: evelyn.gius@tu-darmstadt.de
Institution: Technische Universität Darmstadt
GND: [1084241307](#) ORCID: [0000-0001-8888-8419](#)

Autor*in:
Marcus Müller

Kontakt: marcus.mueller@tu-darmstadt.de
Institution: Technische Universität Darmstadt
GND: [133933482](#) ORCID: [0000-0003-4921-4512](#)

Autor*in:
Andrea Rapp

Kontakt: andrea.rapp@tu-darmstadt.de
Institution: Technische Universität Darmstadt
GND: [118015915](#) ORCID: [0000-0003-4933-4397](#)

Autor*in:
Thomas Weitin

Kontakt: thomas.weitin@tu-darmstadt.de
Institution: Technische Universität Darmstadt
GND: [128442433](#) ORCID: [0000-0002-9003-5746](#)

DOI des Artikels:
[10.17175/2023_003](https://doi.org/10.17175/2023_003)

Nachweis im OPAC der Herzog August Bibliothek:
[1830041150](#)

Erstveröffentlichung:
01.06.2023

Lizenz:

Sofern nicht anders angegeben



Medienlizenzen:
Medienrechte liegen bei den Autor*innen

Letzte Überprüfung aller Verweise:
16.05.2023

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
[Linguistik](#) | [Literaturwissenschaft](#) | [Philologie](#) | [Segmentierung](#) | [Textanalyse](#) |

Empfohlene Zitierweise:
Sabine Bartsch / Evelyn Gius / Marcus Müller / Andrea Rapp / Thomas Weitin: Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft. In: Zeitschrift für digitale Geisteswissenschaften 8 (2023). HTML / XML / PDF. DOI: [10.17175/2023_003](https://doi.org/10.17175/2023_003).

Sabine Bartsch, Evelyn Gius, Marcus Müller, Andrea Rapp, Thomas Weitin

Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft

Abstracts

Zu einer nachhaltigen Entwicklung der Digital Humanities (DH) gehören Ergebnisse, die auch für Geisteswissenschaftler:innen jenseits der DH-Communities relevant sind, weil sie auf existierende Forschungsfragen Antworten geben. Als ein möglicher Weg, um in dieser Hinsicht ins Gespräch zu kommen, erscheint uns das Nachdenken darüber, wie digitale Analysen Grundbegriffe der Sprach- und Literaturwissenschaft problematisch und damit deutlicher werden lassen. Wir zeigen das exemplarisch an der *Segmentierung* von Text, die für viele Analysen eine Basis darstellt, die gleichermaßen, aber auf je unterschiedliche Art und Weise zum Lesen und zum Rechnen im Sinne einer intellektuellen und computationellen Verarbeitung von Text genutzt werden kann. Vor dem Hintergrund von konkreten Problemen aus der sprach- und literaturwissenschaftlichen Segmentierungspraxis diskutieren wir Ansätze aus den Teildisziplinen der Philologie.

Part of a sustainable development of the Digital Humanities (DH) are findings whose relevance extends beyond DH communities because they are providing answers to existing research questions. A possible way of entering into this conversation seems to be a reflection on questions of how digital analyses critically question basic concepts of linguistics and literary studies and thus sharpen their definition. In this paper, we are going to address such questions based on the example of *text segmentation* which forms the basis for many analyses, and which can be used, albeit in different ways, as a foundation for both reading and computational processing of text. We are taking exemplary issues from linguistic and literary segmentation practice as a vantage point for discussion in this paper.

1. Die epistemische Bedeutung digitaler Zugänge

Zu einer nachhaltigen Entwicklung der Digital Humanities gehören Erkenntnisse, die auch für Geisteswissenschaftler:innen jenseits der DH-Communities relevant sind, weil sie auf existierende Forschungsfragen Antworten geben oder auf neue, mit den Forschungsfragen zusammenhängende Problemfelder verweisen.¹ Entsprechend ist auch der an unserem Institut² praktizierte Zugang zur *Digital Philology* so ausgerichtet, dass bestimmte Forschungsaspekte als gemeinsame Herausforderung und als Diskussionsgegenstand zwischen den Teilfächern der Philologie – Linguistik und Literaturwissenschaft (einschließlich der Mediävistik) – betrachtet werden. Das Digitale wird im Kontext der Digital Philology nicht als vom Nichtdigitalen abgesetzt, sondern als damit zusammenhängend betrachtet. Der digitale Zugang zu philologischen Fragen wird entsprechend auch für eine Auseinandersetzung über die philologischen Teilfächer hinweg genutzt, der sich häufig als erkenntnisproduktiver erweist.³

Das Nachdenken darüber, wie digitale Analysen Grundbegriffe in den Philologien problematisch und damit deutlicher werden lassen, ist ein möglicher Weg, um in dieser Hinsicht ins Gespräch zu kommen. Im Folgenden wollen wir anhand von Forschungsproblemen aus der Linguistik und Literaturwissenschaft zeigen, wie eine digitale Betrachtung von fachspezifischen Problemen übergeordnete Konzepte in den Blick bringen und so zu einem produktiven Austausch zwischen den Disziplinen beitragen kann. Dies geschieht exemplarisch anhand eines Verfahrens, das für viele Textanalysen eine Basis darstellt: die *Segmentierung* von Text. Denn jede Textanalyse bezieht sich auf bestimmte Textsegmente, welche wiederum auf je unterschiedliche Art und Weise zum Lesen und auch zum Rechnen genutzt werden. In den verschiedenen Zugängen unterscheidet sich nicht nur die Auffassung von Text als einer aus Textsegmenten zusammengesetzte Einheit. Je nach disziplinärer Verortung und theoretischem Hintergrund eines Forschungszugangs wird auch die Segmentierung als Verfahren nicht immer explizit gemacht, genauso wenig wie die methodologischen Konsequenzen des genutzten Konzepts von Segment bzw. des angewendeten Verfahrens der Segmentierung thematisiert werden.

Das Konzept der Segmentierung, das alle in diesem Beitrag diskutierten Ansätze verbindet, steht dabei im weitesten Sinne für eine Aufteilung des zu untersuchenden Textmaterials. Während eine solche Aufteilung in der Linguistik selbstverständlich ist – deren Grundvoraussetzung ist die Segmentierung des sprachlichen Kontinuums in diskrete Einheiten wie Phonem, Morphem, Wort, Phrase oder Satz (siehe Kapitel 2.1) –, ist das Konzept in der literaturwissenschaftlichen Terminologie wenig verankert.

¹ Dieser Beitrag basiert auf einem gleichnamigen Vortrag, den wir im Dezember 2019 auf der Tagung ›Wozu Digitale Geisteswissenschaften? Innovationen, Revisionen, Binnenkonflikte‹ an der Leuphana Universität Lüneburg gehalten haben.

² Es handelt sich um das Institut für Sprach- und Literaturwissenschaft an der Technischen Universität Darmstadt.

³ Vgl. Adler et al. 2020.

Mehr noch: Selbst die Entwicklung von computationellen Zugängen im Kontext der literaturwissenschaftlichen Digital Humanities, die häufig auf Textsegmenten basiert, führte bislang nicht zu einer disziplinär begründeten Auseinandersetzung mit der Frage der Textsegmentierung.

Trotz dieser unterschiedlichen Ausgangslage in den Teildisziplinen wird an den im Folgenden diskutierten fünf Zugängen eine sie verbindende Herausforderung sichtbar. Die Segmentierung von Texten ist eng verbunden mit der Kategorisierung von Phänomenen und der Interpretation dieser Phänomene in den Texten, also der kontrollierten Praxis der Herstellung von Sinn. Mit ›Sinn‹, dessen reiche Begriffsgeschichte in Linguistik, Hermeneutik und Sprachphilosophie wir an dieser Stelle nicht diskutieren können,⁴ meinen wir hier die durch Interpretation hergestellte Situationsbedeutung einer textuellen und diskursiven Einheit (im Gegensatz etwa zur Systembedeutung oder dem referenziellen Potenzial sprachlicher Zeichen). Hilfreich mag es sein, darauf hinzuweisen, dass ›Sinn‹ etymologisch auf die indogermanische Wurzel *sent- ›eine Richtung nehmen, gehen‹ zurückführbar ist⁵ und dem Begriff, wie Donatus Thürnaus ausführte,⁶ metaphorisch die Idee der Richtungsgebung innewohnt. Bei der Herstellung von sprachlichem Sinn geht es darum, das Sprachverstehen in eine von typischerweise mehreren möglichen Richtungen zu treiben. Dazu braucht es Kontextsignale. Wir wollen in diesem Beitrag zeigen, dass dieser Prozess der Kontextualisierung sprachlicher Bedeutung unmittelbar und ursächlich mit der Größe der dafür in Betracht gezogenen Textsegmente sowie den gewählten Segmentierungsverfahren in Zusammenhang steht. Eine weitere Ebene ergibt sich, wenn man auch den ›materiellen‹ Textträger in den Blick nimmt und dessen texttopografische Einheiten analytisch segmentiert. Das Layout wird in diesem Verständnis als weiterer Bedeutungsträger gesehen.

2. Segmentierung

2.1 Segmentierung in der Linguistik

Das älteste Segmentierungsverfahren ist nach dem Artikel von Wolf Thümmel im *Metzler Lexikon Sprache*⁷ die Schrift, da im Prozess der Verschriftung das Lautkontinuum der mündlichen Sprache notwendigerweise in diskrete Einheiten zerlegt und so der Kategorisierung und Reflexion zugänglich gemacht wird. Auch Schriften und Texte können verschiedene Stufen von Segmentierung durchlaufen: Das moderne Schrift- und Layoutbild mitteleuropäischer Schrift- und Textzeugen z. B. entwickelt sich erst im Mittelalter nach einer Phase der *Scriptio Continua* ohne Segmentierung der Wort- und Texteinheiten. Segmentierung nennt man in der Linguistik die Zerlegung komplexer Einheiten in ihre Elemente zum Zweck ihrer Klassifikation. Es geht also darum, in der *chaîne parlée*⁸ Einheiten voneinander abzugrenzen und kategorial zu bestimmen, z. B. einen Laut (*Phon*) einer Lautklasse (*Phonem*) zuzuordnen. Segmentierung hat sich im Paradigma der strukturalistischen Sprachbeschreibung zu einem Standardverfahren entwickelt, das traditionellerweise auf die linguistischen Einheiten unterhalb der Satzebene angewendet wird: Phonem (Lautklasse), Morphem, Wort und Phrase. Auch außerhalb der Linguistik bekannte Segmentierungsverfahren sind beispielsweise die im Schulunterricht eingeübten Praktiken zur Erkennung von Satzgliedern wie die Umstell- und Ersetzungsprobe.

⁴ Für die hermeneutische Diskussion vgl. Angehrn 2010; für die linguistische Semantik Katz 2004; für die Diskursforschung Keller 2015.

⁵ Vgl. Pfeifer et al. 1993.

⁶ Donatus Thürnaus 2017.

⁷ Thümmel 2016, S. 602.

⁸ Vgl. Saussure 1916, S. 16.

Einheiten	Beispiel	Klasse	Beispiel
Phon Graph	[dax] Dach	Phonem Graphem	/d/ - /a/ - /x/ <D> - <a> - <ch>
Morph	Bett - <i>en</i> Kind - <i>er</i>	Morphem	Flexionsmorphem, Nominativ Plural
Wort / Token	Klang Wunder	Wortart	Nomen appellativum
Phrase	die kleine Biene unser schöner Garten	Phrasentyp	Nominalphrase
Satz	Der Klügere gibt nach. Gertrud, deren Nachbarin Kakteen züchtet, steht am Fenster.	Satztyp	Aussagesatz
Text	Liebe Heidrun, mir geht es gut. Dein Volker	Textsorte	Brief

Tab. 1: Linguistische Einheiten der Segmentierung.

Seit die Linguistik sich der Einheit *Text* zugewandt hat, werden Segmentierungsverfahren auch oberhalb der Satzebene, etwa zur Erkennung von Layout- und Textstrukturen oberhalb des Satzes, also Absätzen, Kapiteln etc., angewendet. Solche Verfahren der Textsegmentierung spielen insbesondere bei der Arbeit mit digitalen Korpora eine große Rolle. In der digitalen Linguistik bezeichnet man Segmentierungsverfahren auf Wort- und Satzebene als *Tokenisierung*. Das beinhaltet die Identifikation von Satzgrenzen (Satztokenisierung oder *sentence splitting*) und die Bestimmung von Datumsangaben (z. B. 9. Nov. 1989) und sogenannten Alphabetismen (z. B. U.S.A.) ebenso wie die Tokenisierung der Einheiten auf der Ebene des Wortes (z. B. die Identifikation von einfachen *Lexemen* wie *Haus* sowie von Komposita wie *Hausboot* oder *Vitamin-B-haltig* als Tokens). Segmentierungsentscheidungen im Sinne der Tokenisierung werden zumeist in unmittelbarer Abhängigkeit der Identifizierbarkeit von Zeichenketten als Einheiten implementiert. So wird als einfachste Annahme beispielsweise das Leerzeichen als Tokengrenze angenommen und durch das verbundene Auftreten von Interpunktionszeichen als Tokengrenze ergänzt. Ein Komma oder Punkt am Ende einer Sequenz alphabetischer Zeichen, die ein linguistisches Wort konstituieren, sind also ebenso wie ein Leerzeichen als Tokenisierungsgrenze definiert. Ausnahmen zu dieser Regel, wie im Falle von Abkürzungswörtern, Titeln, wie *Dr.*, oder Alphabetismen, wie *R.E.M.*, werden anhand von Regelerweiterungen definiert. Computationell implementiert werden Tokenisierungsalgorithmen beispielsweise auf der Grundlage sogenannter regulärer Ausdrücke und durch Regeln für spezifische Anwendungsszenarien ergänzt. Ein Beispiel hierfür ist die Tokenisierung von Diskursen mit Elementen konzeptioneller Mündlichkeit, wie *Social-Media-Kommunikate* (z. B. Bildungen aus Vollverb + Personalpronomen wie *schreibste* und sogenannter Aktionswörter wie **grins** oder *beidirseinwill* sowie Bildungen mit Sonderzeichen als Wortbestandteil, wie Formen mit Hashtags *#Urlaub* oder E-Mail-Adressen und URLs). Die automatische Implementierung solcher Tokenisierungsalgorithmen wird häufig in regelbasierten Tokenisierungsprozessen eingesetzt und durch Verfahren des maschinellen Lernens ergänzt und kann im Fall einer sehr guten Passung zwischen Modell bzw. Trainingsdaten und der Komplexität der zu tokenisierenden Daten hohe Genauigkeiten von über 99 % (*F1 score*) erreichen,⁹ bei schlechterer Passung aber auch deutlich darunter liegen. Der Prozess der Tokenisierung steht in direkter Abhängigkeit zu den Eigenschaften der zu tokenisierenden Textdaten sowie der Qualität bzw. der Passung der Modellierung des Tokenisierungsprozesses.¹⁰

2.2 Segmentierung in der Literaturwissenschaft

In der literaturwissenschaftlichen Textanalyse geht es zumeist um wesentlich größere Texteinheiten als in der Linguistik. In der Analyse von Erzähltexten umfasst eine Einheit typischerweise sogar den ganzen Text, wobei nicht selten eine ganze Reihe von Texten im Fokus steht, etwa das Gesamtwerk einer Autor:in oder eine Sammlung zumeist als exemplarisch geltender Texte z. B. einer Strömung oder Gattung. Wenn auch im Einzelfall durchaus Kapitel und ähnliche strukturelle Einheiten in Analysen mit einbezogen werden, gibt es kaum konventionalisierte Unterteilungen von Texten, die in Analysen normalerweise genutzt werden.

⁹ Beißwenger et al. 2016.

¹⁰ Ortman et al. 2019.

Das wird bei der Betrachtung von literaturwissenschaftlichen Grundlagenwerken offensichtlich. So findet sich im *Handbuch Literaturwissenschaft* mit seinen drei Bänden zu Gegenständen und Grundbegriffen (Band 1), Methoden und Theorien (Band 2) und Institutionen und Praxisfeldern (Band 3) nichts zu Segmentierung als Verfahren oder zu Segmenten als Texteinheiten.¹¹ Zwar wird in den Bänden 1 und 2 auf Segmentierung im Kontext von Dramen und von Lyrik eingegangen und es werden jeweils typische Segmente genannt.¹² Dies geschieht aber ohne Explizierung des Konzepts bzw. des Verfahrens. Für Erzähltexte hingegen werden die in der Analyse genutzten Segmente erst gar nicht bestimmt, allerdings wird zumindest implizit Text als aus Segmenten zusammengesetzt dargestellt. Außerdem klingt, wie auch bei der Beschreibung lyrischer Texte, eine semantische Segmentierung – etwa nach Handlung oder Figuren – an, die jedoch nicht systematisch in Bezug auf das Segmentieren der Texte beschrieben wird.¹³ Dieser geringen Relevanz von Segmenten und Segmentierung in der Literaturwissenschaft entspricht auch, dass sich im Standardnachschlagewerk der germanistischen Literaturwissenschaft, dem *Reallexikon der deutschen Literaturwissenschaft*¹⁴, kein eigener Eintrag dazu findet. In Bezug auf Prosatexte – und zum Teil auch auf alle Texte – ist der literaturwissenschaftliche Zugang zu Text also insofern vorwiegend ein holistischer, als er keine allgemein akzeptierte Segmentierung kennt.

Für die Automatisierung im Bereich der computationellen Literaturwissenschaft ist die fehlende Segmentierung von Prosatexten problematisch. Anders als in den meisten linguistischen Fragestellungen kann hier nicht auf standardisierte Segmente zurückgegriffen werden, dabei basieren viele Verfahren auf Segmenten. Sei es bei der Bestimmung von Autor:innenschaft, bei der Berechnung von als eine Art thematische Struktur aufgefassten *Topics*, bei *Sentiment-Analysen* oder bei Verfahren zur Bestimmung semantischer Ähnlichkeiten, wie sie in der distributionellen Semantik gerade populär sind: Die Verfahren basieren in verschiedenem Umfang auf einer Unterteilung der Texte in kleinere Einheiten. Dabei segmentieren sie je nachdem, ob sie dem Paradigma der Suche in oder der abstrakten Repräsentation von Texten verpflichtet sind, auf ganz unterschiedlichen Ebenen. Für Verfahren wie die Sentiment-Analyse ist zu vermuten, dass eine literaturwissenschaftliche Bestimmung der genutzten Einheiten – etwa als figurenbezogene Segmente oder eine Segmentierung der Handlung – zu einer wesentlichen Verbesserung der Verfahren führen würde. Andere Verfahren nehmen zwar bereits behelfsmäßig eine Unterteilung der Texte in Segmente gleichen Wortumfangs vor, können damit aber wahrscheinlich nicht ihre vollen Möglichkeiten ausschöpfen. Das betrifft etwa das *Topic Modeling*, welches insbesondere für auch in ihrer Struktur vergleichbare Texte bzw. Texteinheiten entworfen wurde. Da bislang mit von der Wortzahl bestimmten Segmenten gearbeitet wird, wird die Textstruktur nicht berücksichtigt.

3. Fünf Beispiele

3.1 Segmentierung von Layout und Textstruktur

Neben der inhaltsbezogenen Bildrecherche,¹⁵ die auf u. a. Segmentierungsverfahren aufsetzt, ist die Segmentierung von Layoutelementen und entsprechenden Textstrukturen für Texterkennungsverfahren wie *Optical Character Recognition* (OCR) und *Handwriting Recognition* (HWR) bzw. *Handwritten Text Recognition* (HTR) von zentraler Bedeutung. Darüber hinaus erlaubt eine solche Segmentierung jedoch auch die Analyse von Schreibprozessen¹⁶ und gibt neue Einsichten in Textüberlieferungsprozesse auch sehr komplexer Natur.¹⁷ Die mathematischen und informatisch-methodischen Grundlagen der verschiedenen Verfahren beschreibt Rainer Herzog.¹⁸ Aus dieser Perspektive ist die Segmentierung von Texteinheiten auf der Bildoberfläche eine seit langem erkannte Herausforderung, die intensiv erforscht wird. Die Analyse von modernen oder historischen Drucken sowie von Handschriften aller Art bringt dabei aufgrund der Beschaffenheit und des Erhaltungszustands unterschiedliche Anforderungen mit sich.

Das Beispiel der als so »kontaminiert« geltenden Überlieferung der Aristoteles-Schrift *de interpretatione*, dass sie in vertretbarer bzw. verfügbarer Zeit eines Forscher:innenlebens nicht entschlüsselt werden könne,¹⁹ soll zeigen, wie eine Segmentierung von Layout- und Textstrukturen in Grundtext und Paratexte neue Einsichten in die Überlieferungs- und Rezeptionsgeschichte geben

¹¹ Vgl. Anz (Hg.) 2013.

¹² Für Dramen wird etwa festgestellt: »Den sichtbar markierten Segmentierungen von Texten in »Auftritte«, »Szenen«, »Akte«, »Kapitel« oder den unmarkierten Segmentierungen liegt vielfach ein Zeit- und Raumwechsel zugleich zugrunde« (vgl. Anz 2013, S. 118). Bei lyrischen Texten werden darüber hinaus semantische Segmente thematisiert: »In einem ersten Schritt kann der Interpret eine erste Gliederung seines Textes in syntaktische Einheiten (etwa Strophen, Kapitel, Abschnitte) und semantische Segmente (etwa Orte, Figuren, Figurencharakteristika, Handlungselemente) vornehmen« (vgl. Köppe / Winko 2013, S. 296).

¹³ Vgl. dazu den Abschnitt »Erzähltexttheorie« im Handbuch Literaturwissenschaft (Schmid 2013, S. 89–120).

¹⁴ Vgl. Fricke et al. (Hg.) 1997–2003.

¹⁵ Vgl. Bullin / Henrich 2020.

¹⁶ Vgl. Gabler 2007.

¹⁷ Vgl. Krewet 2015.

¹⁸ Vgl. Herzog 2018.

¹⁹ Vgl. Krewet et al. 2019, S. 77.

kann. Dabei werden sowohl Veränderungen der Texteinheiten und ihrer Layoutgestalt(ung) als äußerer Transfer (materielles Resultat) wie auch die sich daraus ergebende Veränderung der Konstitution des Grundtexts als innerer Transfer (epistemischer Prozess) in den Blick genommen, denn beides gehört untrennbar zusammen. Segmentierung erlaubt also Analyse textueller Dynamiken an der Oberfläche (Layout) sowie in der Tiefe (Textverständnis).

Solche Segmentierungsfragen an Texte bzw. Textüberlieferungen sind im Zuge des *Material Turns* in den Blick geraten,²⁰ erfahren vor allem aber auch durch digitale Möglichkeiten weitere Aufmerksamkeit. Zum einen lassen sich Segmente umfangreicher Überlieferungen (semi-)automatisch bestimmen, zum anderen solche Segmente und ihre Schichtungen dann anschaulich visualisieren. Darüber hinaus eröffnet gerade ein gemischter Ansatz von automatischer (Vor-)Segmentierung und intellektueller (Tiefen-)Annotation sowie quantitativer und qualitativer Verfahren bessere Analysemöglichkeiten, wie unsere Arbeiten im Infrastruktur- sowie im Gastprojekt des SFB 980 ›Episteme in Bewegung‹ zeigen konnten.²¹ Die strukturelle und materielle Organisation der Wissensbestände in *de interpretatione* verändert sich im Laufe der Überlieferungsgeschichte beständig: Verschiedenste Arten und Formen von Interlinear- oder Marginal-Glossen, Scholien, Diagrammen und Kommentaren können in gleicher, ähnlicher oder stark veränderter Form bis hin zur Aufnahme in den Grundtext in eine neu entstehende Abschrift aufgenommen werden. Die Segmentierung dieser Einheiten und ihre eindeutige Kategorisierung führt dazu, dass Versionen dieser Einheiten vergleichbar gemacht und damit Kontaminationen aufgespürt werden können.²²

Auch im Bereich der materiellen Textoberfläche wirken also digitale Modellierungen auf Forschungsfragen und Forschungsergebnisse zurück, indem in einem ersten Schritt der Segmentierung relevante Eigenschaften der Forschungsgegenstände kategorisiert und festgehalten werden. Idealerweise geschieht dies in einem Aushandlungsprozess zwischen Text- und Informationswissenschaftler:innen, damit die gemeinsame Arbeit am Modell zu Erkenntnisfortschritt am Gegenstand und zu Interoperabilität der Kategorien führt.

3.2 Segmentierung jenseits der Textoberfläche: Mehrwortausdrücke

Segmentierung kann verschiedene Formen und Rollen im analytischen Workflow annehmen. Neben ihrer Rolle als Teil der Vorverarbeitung von Korpora spielt die Segmentierung auch als Teil der Heuristik zur Identifikation und Extraktion lexikalischer *Mehrwortausdrücke* (MWA) eine zentrale Rolle. In diesem Abschnitt soll am Beispiel der Identifikation und Extraktion lexikalischer Mehrwortausdrücke in Korpora, hier anhand von *Kollokationen*, gezeigt werden, auf welchen Segmentierungen der Prozess beruht und wie diese das Extraktionsergebnis beeinflussen. Die Herausforderung ist im Falle solcher MWA, dass sie keine feste identifizierbare Oberflächenform aufweisen, also keinem festen Muster folgen, sondern neben kontinuierlichen, ununterbrochenen Wortfolgen, wie z. B. Kollokationen zwischen prädikativem Adjektiv und Substantiv (wie *blondes Haar*) auch diskontinuierliche MWA, wie z. B. Kollokationen zwischen Prädikatsverb und Substantiv in der Nominalphrase (wie *ein Verbrechen begehen – jemand begeht schreckliche Verbrechen*) umfassen.

Segmentierungsentscheidungen beruhen auf möglichst gut beschriebenen Merkmalen zur Identifikation der zu segmentierenden Einheiten an der sprachlichen Oberfläche. Typischerweise werden entsprechende Entscheidungskriterien in Segmentierungs- und Annotationsrichtlinien festgehalten sowie bei automatischen Verfahren in entsprechenden Segmentierungsalgorithmen implementiert. Dabei kommt es auf eine Balance zwischen der Treffsicherheit und Zuverlässigkeit der Segmentierung, der möglichst exhaustiven Identifikation der zu untersuchenden linguistischen Einheiten und der Implementierungs- bzw. Anwendungsanforderungen an. Das heißt, dass sich unter Umständen nicht alle aus linguistischer Perspektive interessanten Einheiten mit der angestrebten Zuverlässigkeit in Form manuell oder automatisch anwendbarer Segmentierungsanweisungen umsetzen lassen. Entscheidend sind hierfür eine Reihe von Faktoren, denen es systematisch beizukommen gilt. Auf der ersten Ebene geht es um die eindeutige Beschreibung der Merkmale in Form von Segmentierungsrichtlinien. Diese müssen auf der nächsten Ebene entweder manuell von menschlichen Segmentierer:innen möglichst gut verstanden und so bei der Segmentierung befolgt und bei automatisierten Verfahren in Form von Regeln implementiert werden. Grundvoraussetzung ist im Fall der manuellen Anwendung der Segmentierung ein intellektuelles Verstehen der Eigenschaften der zu segmentierenden sprachlichen Einheiten im Sinne ihrer diskreten Identifikation in den Daten. Dieser Prozess beinhaltet als elementaren Bestandteil eine Beschreibung der Merkmale, an denen diskrete Segmentierungseinheiten an der Oberfläche der Daten identifizierbar sind. Eine solche Modellierung von Segmentierungseinheiten und deren Kennzeichen an der sprachlichen Oberfläche ist elementar für die algorithmische Implementierung der Segmentierung. Die relative Einfachheit der Identifikation von Segmentierungsmerkmalen an der Oberfläche ist zugleich die größte Stärke und Schwäche des Verfahrens, denn zuverlässig lassen sich so zunächst am besten solche Segmentierungsentscheidungen modellieren, die sich anhand oberflächenstruktureller

²⁰ Vgl. Schubert (Hg.) 2010.

²¹ Vgl. Krewet et al. 2019; Krewet / Hegel 2020.

²² Vgl. dazu ausführlich Krewet / Hegel 2020.

Merkmale möglichst eindeutig, z. B. anhand regulärer Ausdrücke, abbilden lassen. Anhand solcher Verfahren gut zu modellierende Segmentierungen sind Wortgrenzen auf der Basis von Leerzeichen und Interpunktionszeichen sowie regelkonforme Sätze.

Korpora mit auf diese Weise vorgenommenen Segmentierungen, wie die Tokenisierung auf Wortebene sowie die Satztokenisierung, bilden die Grundlage für die linguistische Kategorisierung durch Verfahren wie die automatische Wortartenannotation (*Part-of-Speech-Tagging*). Das Part-of-Speech-Tagging ist so eng mit der Tokenisierung auf Wort- und Satzebene verwoben und von der Tokenisierungsqualität abhängig, dass beide Prozesse in der Regel in einem Workflow miteinander verbunden sind. So implementieren gängige Part-of-Speech-Tagger, wie der *Stanford Log-Linear Part-of-Speech Tagger*²³ und der *TreeTagger*²⁴ eigene Tokenisierungsprozesse, die dem Part-of-Speech-Tagging innerhalb des Workflows vorgeschaltet werden. Es handelt sich bei der Tokenisierung also um Vorverarbeitungsschritte, die Voraussetzung und Grundlage für automatische Annotationsprozesse, wie das Part-of-Speech-Tagging, sind. Die meisten aktuellen Part-of-Speech-Tagger bringen eine eigene Implementierung der Tokenisierung mit, da die Modellierung der Wortartenannotation eine spezifische Tokenisierung voraussetzt und eine gute Tokenisierung auch entscheidenden Einfluß auf die Qualität der Wortartenannotation hat.

Doch Segmente als Grundlage linguistischer Analysen konstituieren sich im Prozess der Analyse auch auf andere Weise, nämlich dann, wenn die zu untersuchenden linguistischen Phänomene zwar auf tokenisierten Elementen beruhen, sich jedoch darüber hinaus strukturell auf weiteren Ebenen der linguistischen Organisation konstituieren. Ein Beispiel hierfür sind Kollokationen und andere komplexe lexikalische Gruppen, die sich aus zwei oder mehr kontinuierlichen oder diskontinuierlichen lexikalischen Einheiten konstituieren. Die folgenden Beispiele sollen dies illustrieren:

Mehrwortausdruck	Beispiel
Stützverbkonstruktionen, z. B. Rede halten, Bad nehmen	Diese Rede hielt die Kanzlerin anlässlich des ... Die Rede , die die Kanzlerin anlässlich des ... hielt .
trennbare Verben, z. B. stoßen auf, binden an	Im Verlaufe der Untersuchung stießen die Ärzte auf neue Symptome.
Kollokationen, z. B. Verbrechen begehen	Von dieser Gruppierung wurden über Jahre schwerste Verbrechen begangen . Die Verbrechen wurden von dieser Gruppierung über mehrere Jahre begangen .

Tab. 2: Beispiele lexikalischer Mehrwortausdrücke.

Lexikalische Mehrwortausdrücke sind phraseologisch also als relativ feste Kombinationen im Inventar der Sprache etabliert. Manche davon sprichwörtlich und relativ unveränderlich (»Der Spatz in der Hand ist besser als die Taube auf dem Dach«), die meisten aber – wie die hier adressierten Beispiele – flexibler im Rahmen des Sprachsystems. Ihr habituelles kombiniertes Auftreten, das J. R. Firth mit dem DiKtum »You shall know a word by the company it keeps«²⁵ beschreibt – hat in der Forschung der vergangenen Jahrzehnte, vor allem aber seit der Einführung digitaler Daten und Verfahren in der Linguistik eine Reihe von Ansätzen mit dem Ziel hervorgebracht, relevante und reproduzierbare Ergebnismengen lexikalischer Mehrwortausdrücke aus Korpora zu extrahieren. Dies geschieht üblicherweise auf der Grundlage statistischer Verfahren, wie zum Beispiel anhand von Assoziationsmaßen, unter denen *Log-Likelihood-Ratio*, *t-Score*, der *Dice-Koeffizient* und *Mutual Information Score* (MI) zu den am häufigsten verwendeten zählen.²⁶ Neben der relativ etablierten Kookkurrenz von zwei oder mehr Konstituenten der lexikalischen Ebene weisen lexikalische Mehrwortausdrücke aber auch ein gewisses Maß an Flexibilität bezüglich struktureller Permutationen im Rahmen der regulären Grammatik auf, können also je nach Kontext modifiziert und anhand der Regeln der Grammatik verändert und umgestellt werden (siehe Beispiel zu Kollokationen in Tabelle 2). Kollokationen sind zudem häufig zu einem gewissen Grad semantisch transparent, ihre Bedeutung lässt sich also vollständig oder teilweise aus den Einzelbedeutungen ihrer Konstituenten ableiten.

Im Folgenden soll am Beispiel der Extraktion von Kollokationen im Firth'schen Sinne²⁷ habitueller Kookkurrenzen lexikalischeblor Einheiten – also des wiederkehrenden gemeinsamen Auftretens als Wortverbindungen – und einer um syntaktische Relationen erweiterten Definition diskutiert werden, welchen Einfluss die Segmentierung im Sinne von Tokenisierung und

²³ Vgl. Toutanova / Manning 2000.

²⁴ Vgl. Schmid 1994; Schmid 1995.

²⁵ Firth 1964 [1957].

²⁶ Vgl. Evert 2008; Bartsch / Evert 2014.

²⁷ Vgl. Firth 1964 [1957].

Satzsegmentierung auf die Extraktion von Kollokationen aus Korpora anhand statistischer Assoziationsmaße hat und wie die Veränderung von Parametern mit Bezug auf den Suchraum – die gleichfalls eine Form der Segmentierung darstellt –, die Ergebnismenge bei der Identifikation von Kollokationen beeinflusst. Anhand dieses Beispiels lässt sich zeigen, dass die Operationalisierung einer bereits in prä-computationaler Zeit formulierten Definition mittels digitaler Verfahren und statistischer Assoziationsmaße auf Segmentierungsentscheidungen fußt und auf dieser Grundlage im Digitalen geschärft und weiterentwickelt werden konnte.

Bei Kollokationen handelt es sich um ein Phänomen, das Stefan Evert als »Epiphänomen«²⁸ bezeichnet hat, da diese im klassischen Firth'schen Sinne zunächst ein Effekt der habituellen Kookkurrenz ihrer Konstituenten sind, denen man auch nach Firth eine gegenseitige Erwartbarkeit (»mutual expectancy between certain words«²⁹) zwischen bestimmten Wörtern zuschreibt. Diese habituell kookkurrierenden, wechselseitig erwartbaren lexikalischen Konstituenten bilden von kompetenten Muttersprachler:innen im Sprachgebrauch beherrschte, relativ feste, aber auch im Rahmen der Grammatik flexible, wiederkehrende Einheiten, die zumeist semantisch relativ transparent sind, aber auch anhand der etablierten Kookkurrenz zusätzliche Bedeutung tragen können. Jedenfalls sind Kollokationen aber derart etabliert, dass sie im Sprachgebrauch als Einheiten empfunden werden und entsprechend funktional sind. Ein Verstoß gegen die Verwendung der etablierten Wortkombinationen wird unter Umständen verstanden, aber auch als den Konventionen widersprechend erkannt.

Eine der zentralen Herausforderungen der Identifikation von Kollokationen in Korpora liegt nun darin, dass sie einerseits relativ fest etablierte, wiederkehrende Verbindungen aus zwei oder mehr Konstituenten darstellen, dass diese Konstituenten aber erstens, wie oben ausgeführt, den im Rahmen der Grammatik gestatteten Permutationen ihrer relativen Reihenfolge und Formenbildung unterliegen (z. B. ein Verbrechen begehen und Verbrechen werden begangen), und dass die Konstituenten von Kollokationen im Gegensatz zu n-Grammen nicht notwendigerweise konsekutiv aufeinanderfolgen. Weiterhin wird in den meisten Definitionen des Phänomens die Kookkurrenz innerhalb der Satzgrenzen, entweder innerhalb von Phrasen (z. B. in der Nominalphrase blondes Haar) oder über Phrasengrenzen hinweg zwischen Satzkonstituenten definiert (z. B. die Kollokation aus dem Prädikats-Verb und einem Substantiv als Kopf der Nominalphrase in Objektposition, X begeht ein Verbrechen). Die Identifikation von Kollokationen fußt so auf der Tokenisierung der lexikalischen Ebene, also der Identifikation von Lexemen des untersuchten Korpus, und wird begrenzt durch die Satztokenisierung als äußerer Grenze. Kollokationen reichen in den meisten Definitionen nicht über die Satzgrenze hinaus,³⁰ wiewohl Konstituenten selbstverständlich im Folgesatz erneut aufgegriffen werden können. Es ist bei der Identifikation von Kollokationen in Korpora authentischer Sprache also einerseits, wie bei vielen Analysen, die Tokenisierung Grundvoraussetzung. Andererseits ist auf Grundlage der Tokenisierung im Korpus der Gegenstand trotzdem nicht sicher erfassbar, weil eben die Konstituenten von Kollokationen nicht notwendigerweise konsekutiv aufeinander folgen. Kollokationen müssen also innerhalb eines definierten Kontexts (oder Fensters) als an der sprachlichen Oberfläche zunächst scheinbar unverbundene – und für uninformierte Sprachteilnehmer:innen nicht unmittelbar erkennbare – Einheiten durch statistische Verfahren ermittelt werden.

Umfang und Qualität der Korpusdaten und hier vor allem der Segmentierung als Teil der linguistischen Vorverarbeitung sind von entscheidender Bedeutung für die Identifikation von Kollokationen, da korrekt identifizierte Tokens und Satzgrenzen die Einhaltung der Parameter, die Identifikation der Konstituenten und damit auch von Kollokationen beeinflussen. Gerade in sehr großen Korpora, deren Vorverarbeitung aufgrund des Datenumfangs nicht manuell qualitätsgesichert werden kann und wird, verbleiben häufig Artefakte der Digitalisierung, unverbundene Zeichen etc., die die Vorverarbeitung erschweren und damit auch die Qualität der Kollokationsextraktion negativ beeinflussen. So konnte in der erwähnten Studie von Sabine Bartsch und Stefan Evert³¹ gezeigt werden, dass bei der Extraktion von Kollokationen entgegen quantitativer Vorannahmen, die große Korpora als zentrale Grundlage für die Kollokationsforschung annehmen, das Kriterium der Korpusgröße immer auch gegen die Korpusqualität abgewogen werden muss und dass kleinere, aber sehr sauber vorverarbeitete Korpora durchaus bessere Extraktionsergebnisse liefern können als sehr große Korpora, die aufgrund ihrer Größe unter Umständen weniger saubere Daten enthalten.

Bei der Kollokationsextraktion kommt darüber hinaus eine zweite Ebene der Segmentierung zusätzlich zur im Rahmen des *Pre-Processings* erfolgten Tokenisierung zum Einsatz, durch die der Suchraum, innerhalb dessen Konstituenten von Kollokationen erwartbar auftreten, eingegrenzt und für die Statistik beherrschbar wird, indem das Rauschen in den Daten reduziert und damit die Sicherheit der Identifikation relevanter Kollokationen erhöht wird. Die hier vorgenommene Segmentierung ist eine heuristische, sie modelliert und grenzt den potenziellen Suchraum ein, ohne dabei das gesuchte Phänomen direkt auszuwählen. Sie nähert sich sozusagen dem gesuchten Phänomen durch Begrenzung des Suchraumes an, ohne direkt Kollokationskandidaten zu adressieren. In extensiven Untersuchungen konnte gezeigt werden, dass die Parameter für die Begrenzung des Suchraums

²⁸ Evert 2008.

²⁹ Firth 1968, S. 181.

³⁰ Vgl. jedoch die abweichende Definition von Halliday / Hasan 1976, S. 284–286.

³¹ Vgl. Bartsch / Evert 2014.

sowohl Einfluss auf die Menge der identifizierten Kollokationskandidaten (*Recall*), also auch auf die der tatsächlich relevanten Kollokationskandidaten (*Precision*) haben, die in Tests als tatsächliche Kollokationen identifiziert werden konnten. Das Verhältnis zwischen Precision und Recall gibt schließlich Aufschluss über die Güte der Extraktion im Sinne des Umfangs der mit einer gewissen Konfidenz (im nichttechnischen Sinne) identifizierbaren Ergebnismenge.

Die von Evert und Bartsch³² getesteten Parameter-Settings für den Suchraum umfassen neben den etablierten wort-basierten Suchfenstern³³ 3:3, 5:5 und 10:10 Wörter als je linker und rechter Kontext auch die Satzgrenze, also den kompletten Satz als delimitierenden Kontext. Weiterhin werden einer z. B. von Bartsch (2004) vorgeschlagenen Definition folgend und über die Firth'sche Definition hinausgehend Kollokationskandidaten auf der Grundlage einer direkten syntaktischen Relation zwischen den Konstituenten extrahiert.³⁴ Dies erfordert wiederum eine weitere Segmentierung und Annotation im Sinne der Identifikation syntaktischer Einheiten und Relationen. Es kann so unter anderem der Einfluss unterschiedlicher Parametersettings für den Suchraum systematisch evaluiert werden, zum anderen kann aufgezeigt werden, dass unter Hinzuziehung unterschiedlicher linguistischer Parameter, wie zum Beispiel Lemmatisierung, lexiko-grammatischer Wortartenklassifikation und grammatischer Dependenz, die Ergebnisse der Kollokationsextraktion beeinflusst und die Ergebnismengen verändert und verbessert werden können.

Anhand dieses Beispiels konnte gezeigt werden, dass Segmentierungsentscheidungen zum einen als Teil des linguistischen Pre-Processings Einfluss auf die Qualität linguistischer Analysen haben, indem sie den Zugriff auf lexikalische Einheiten als Konstituenten von Kollokationen ermöglichen. Zum anderen ist eine Segmentierung im Sinne der Auswahl begrenzter Suchräume innerhalb der Korpusdaten elementarer Bestandteil von linguistischen und statistischen Verfahren zur Identifikation von Kollokationskandidaten. So werden auf der Grundlage wortbasierter Suchfenster (3–5 Wörter als linker und rechter Kontext eines Suchwortes) oder grammatisch definierter Segmente, wie Sätze oder Konstituenten in direkter syntaktischer Relation, Kollokationskandidaten anhand statistischer Maße für die Bedeutung des gemeinsamen Auftretens ermittelt.

3.3 Analyse heuristischer Textpraktiken

An der Diskussion von Mehrworteinheiten ist schon deutlich geworden, dass sprachliche Segmente nur funktional zu bestimmen sind: Es gehört zusammen, was gemeinsam eine Funktion in der Kommunikation übernimmt. Während Mehrworteinheiten funktional auf den Satz bezogen sind, konstituiert sich die Einheit des Satzes dadurch, dass mit ihr sprachliche Praktiken vollzogen werden. Bei deren Analyse wiederum ist das Zusammenspiel von Handlungsinterpretation, kategorialer Subsumption und struktureller Segmentierung entscheidend. Was das bedeutet, soll im Folgenden anhand von Segmentierungsfragen beim Erstellen eines Tagsets zur Analyse heuristischer Textpraktiken erläutert werden. Unter ›heuristischen Textpraktiken‹ verstehen wir Formulierungsverfahren, mit denen in institutionell verankerten Routinen neues Wissen generiert und an vorhandenes Wissen angeschlossen wird, z. B. ›die Relevanz eines Forschungsthemas markieren‹, ›einen Begriff definieren‹ oder ›eine Aussage argumentativ stützen‹. Dabei interessiert uns, wie solche Formulierungsverfahren in Dissertationen verschiedener wissenschaftlicher Disziplinen zum Einsatz kommen, an welcher Stelle im Text sie verwendet werden, wie sie miteinander kombiniert werden, welche Textfunktion sie haben und was man aus all dem über die epistemischen Praktiken der jeweiligen Disziplin lernen kann. Dieser Forschungsansatz ist ausführlich in einer Pilotstudie dokumentiert.³⁵ Darin wurde in einem abduktiven Verfahren ein Tagset entwickelt und ein Pilotkorpus von 65 Einleitungen zu Dissertationen aus den 13 an der Technischen Universität Darmstadt vertretenen Fachbereichen händisch kollaborativ annotiert. Es ergeben sich als übergeordnete Textpraktiken die expositorischen Verfahren der Relevanzmarkierung, der Zielsetzung und der Thesenstellung sowie Praktiken der Definition und der Stützung von Assertionen. Dazu gibt es jeweils Subkategorien (vgl. Abbildung 1).³⁶

³² Bartsch / Evert 2014.

³³ Vgl. Sinclair 1991.

³⁴ Vgl. Bartsch 2004, S. 76.

³⁵ Vgl. Bender / Müller 2020.

³⁶ Ausführlich in Bender / Müller 2020, S. 22–24.

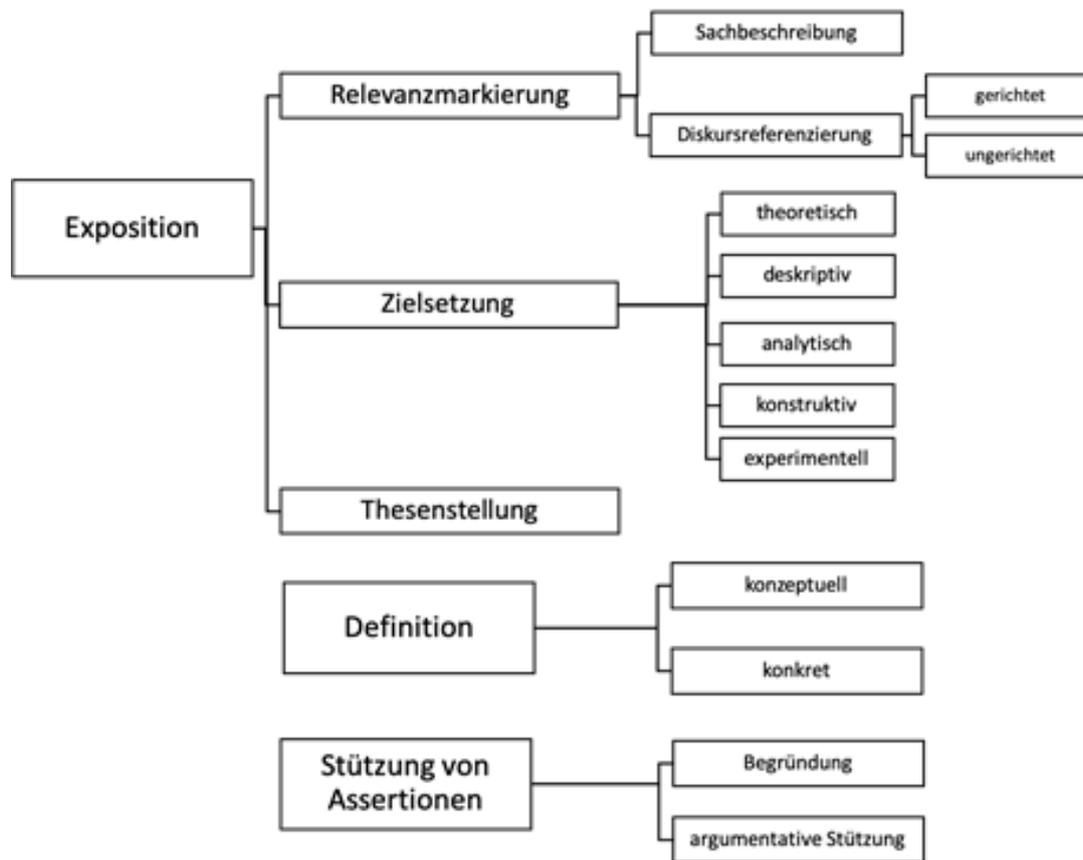


Abb. 1: Das taxonomische Annotationsschema HeuTex. [Bender / Müller 2020, S. 23]

Wie bei allen pragmatischen Untersuchungen in der Linguistik handelt es sich dabei um ein x-als-y-Verfahren: Da Formulierungsverfahren nicht sinnvoll in actu beobachtet werden können (zumindest nicht großflächig), gilt es, die Textsegmente zu ermitteln, mit denen eine bestimmte Textpraktik vollzogen wurde. Nun kann zum Beispiel die Praktik des Argumentierens – je nach Argumentationsbegriff – mit einem Wort, einem Satz oder einer Buchreihe vollzogen werden. Würde man aber ad hoc für jede sprachliche Praktik das Textsegment festlegen, mit dem sie vollzogen wird, könnte man die Segmente nicht miteinander vergleichen, zumindest eine quantitative Auswertung wäre unmöglich. Daher haben wir Annotations-Kategorien grundsätzlich der Einheit ›Satz‹ zugewiesen. Das ist auch die Basiskategorie der linguistischen Pragmatik: Sprachliche Handlungen werden typischerweise mit Sätzen vollzogen. Da andererseits Textpraktiken, wie man sich am Argumentieren gut verdeutlichen kann, oft über die Satzgrenze hinausreichen, ist bei der Analyse mit ›Zonen‹ aufeinanderfolgender diskreter heuristischer Textpraktiken zu rechnen.

Die Festlegung einer linguistischen Einheit zur Segmentierung war auch aus einem anderen Grund wichtig: Das Projekt zielt darauf ab, heuristische Textpraktiken automatisch zu klassifizieren. Hierzu liegen vielversprechende Ergebnisse einer Pilotstudie vor, in der wir eine *Deep-Learning*-Architektur, genauer: ein *Recurrent Neural Network* (RNN), zur Klassifikation auf der Basis unseres Tagsets (vgl. Abbildung 1) verwendet haben.³⁷ Wir konnten feststellen, dass unser Modell auf allen Annotationsebenen recht gut funktioniert und Genauigkeitswerte von bis zu 93 % (Level 0) erreicht. Auf Level 0 ist die Basisebene der Kategorisierung angesiedelt, die Levels 1 und 2 sind jeweils Unterkategorien. Wir sehen auch ein ausgewogenes Verhältnis zwischen *Precision*- und *Recall*-Scores auf allen Ebenen, mit dem größten Unterschied auf Level 0, wo der *Recall* 8 Prozentpunkte über dem *Precision*-Score liegt, was darauf hindeutet, dass das Modell hier (etwas) besser darin ist, alle relevanten Instanzen jedes Labels innerhalb unseres Datensatzes zu finden, als korrekte Vorhersagen der Labels für die Instanzen zu machen. Um diese Ergebnisse zu interpretieren, muss die *Majority Baseline* – d. h. die Menge der häufigsten Instanzen innerhalb des Datensatzes – berücksichtigt

³⁷ Vgl. Becker et al. 2020.

werden. Wir haben die Majority Baseline auf allen drei Ebenen übertroffen, wobei die auffälligste Verbesserung auf Ebene 2 zu verzeichnen war (vgl. Tabelle 3). Da es sich um eine komplexe Klassifikationsaufgabe auf der Basis eines extrem kleinen Trainingsdatensatzes (2.689 Sätze) handelt, haben sich Tagset und Segmentierungspraxis als robust erwiesen.

	Level 0	Level 1	Level 2
Numb. of Labels	5	11	2
Accuracy	0.8302	0.7548	0.9292
F1	0.8071	0.7546	0.9291
Precision	0.7661	0.7541	0.9309
Recall	0.8537	0.7549	0.9295
Majority Baseline	0.7164 (Relevanzmarkierung)	0.6219 (Sachbeschreibung)	0.6023 (gerichtet)

Tab. 3: Ergebnis der RNN-Klassifizierung auf verschiedenen Ebenen. [Aus: Becker et al. 2020]

Nun mag man annehmen, die Satzsegmentierung sei ein mechanischer Schritt des Pre-Processings ohne weitere semantische Implikationen. Hier soll aber an zwei Fällen demonstriert werden, dass die Segmentierungsentscheidung unmittelbaren Einfluss auf kategoriale Zuweisungen hat, und zwar weil sie die Tiefe des hermeneutischen Zugriffs delimitiert, die wiederum die Kategorisierung bestimmt. Unsere Segmentierung ist grundsätzlich der syntaktisch autonome Satz (*Sentence*), da untergeordnete Teilsätze funktional dem Matrixsatz beigeordnet sind und daher auch keine eigenständige Texthandlung repräsentieren. Wird zum Beispiel der folgende Teilsatz (*Clause*) als Segment der Kategorisierungsentscheidung zugrundegelegt (a) und kategorisiert man flach, d. h. ohne Einbeziehung von Kontextwissen, dann ist ein assertiver Sprechakt, in unserem Schema eine Relevanzmarkierung durch Sachbeschreibung, zu kategorisieren. Ist das Segment aber der Satz im Sinne von *Sentence*, in diesem Fall also ein Satzgefüge, dann ergibt sich eine Argumentation (b).

- a) Ihr Lebenslauf ist für die bürgerlichen Frauen ihrer Epoche keineswegs exemplarisch ³⁸
- b) Ihr Lebenslauf ist für die bürgerlichen Frauen ihrer Epoche keineswegs exemplarisch, denn diese wurden weiterhin als Hausfrauen und Mütter definiert. ³⁹

Ein etwas anders gelagerter Fall findet sich in den Beispielen (c) und (d). Betrachtet man Satz (c) isoliert und ohne Kontextwissen einzubeziehen, dann wäre er als deontisch modalisierte Proposition zu interpretieren und pragmatisch als direkter Sprechakt einzuordnen, konkret als Handlungsempfehlung. Berücksichtigt man aber den unmittelbaren Textzusammenhang (d), linguistisch gesprochen: den *Kotext*, dann ergibt sich nach unserem Kategorienschema die heuristische Textpraktik einer Zielsetzung.

- c) Die Arbeit soll in diesen [sic] Zusammenhang Aspekte darstellen, die bei der Erstellung eines solchen Verfahrens grundsätzlich zu beachten sind, und Wege aufzeigen, wie diese im konkreten Anwendungsfall zu einem anwendungsfähigen Verfahren konkretisiert werden können. ⁴⁰
- d) 1.2 Zielsetzung der Arbeit. Ziel der Arbeit ist die Erarbeitung von allgemeingültigen Hinweisen für die Entwicklung von Entscheidungsverfahren, [...]. ⁴¹ i Die Arbeit soll in diesen [sic] Zusammenhang Aspekte darstellen, die bei der Erstellung eines solchen Verfahrens grundsätzlich zu beachten sind, ii und Wege aufzeigen, wie diese im konkreten Anwendungsfall zu einem anwendungsfähigen Verfahren konkretisiert werden können. ii ⁴¹

An dem Beispiel sieht man, dass das Segment ›Satz‹ auch im konkretisierten Sinne keineswegs eindeutig ist. Das Automatisierungsvorhaben unseres Projektes bringt es mit sich, dass Sätze formal im Sinne von ›mit satzabschließendem Interpunktionszeichen abgeschlossene Einheit‹ bestimmt sein müssen. Im entscheidenden Satz der Beispiele (c) und (d) sind aber zwei Satzgefüge miteinander koordiniert, die textpragmatisch nach unserem Schema jeweils unterschiedlich zu kategorisieren wären, nämlich als deskriptive Zielsetzung (Teilsatz i) bzw. als konstruktive Zielsetzung (Teilsatz ii). In unserem Ansatz muss nun aber eine Kategorisierungsentscheidung getroffen und das Satzgefüge damit als pragmatisch subordinativ interpretiert werden: in diesem Fall als deskriptive Zielsetzung.

³⁸ Vgl. Siegel 2002, S. 241.

³⁹ Vgl. Siegel 2002, S. 241.

⁴⁰ Vgl. Dieleman 2016, S. 3.

⁴¹ Vgl. Dieleman 2016, S. 3.

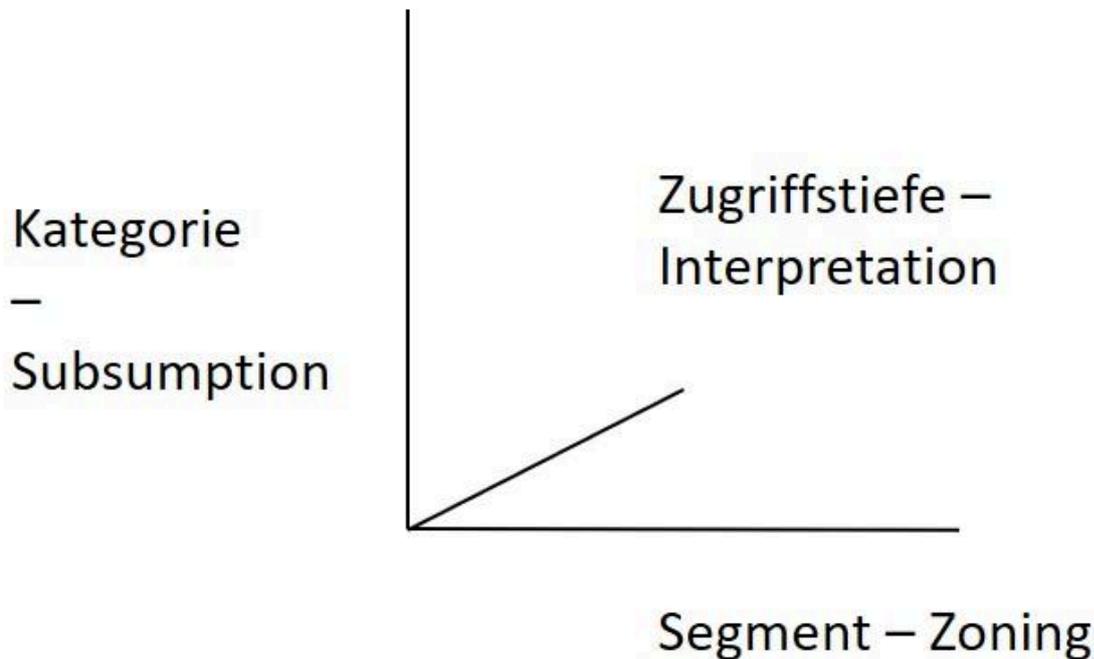


Abb. 2: Dimensionen der Segmentierung. [Eigene Darstellung]

Segmentierung ist also keineswegs eine rein formale Angelegenheit, sondern muss immer als hermeneutische Praktik gedacht werden, in der drei Teilprozesse untrennbar ineinandergreifen und aufeinander bezogen sind (vgl. Abbildung 2): das Ausschneiden eines Segments aus dem Textkontinuum (*Zoning*), die Zuweisung des Segments zu einer analytischen Kategorie (*Subsumption*) und die Festlegung der Tiefe des gedanklichen Zugriffs auf der Basis von mehr oder weniger Kontextinformation (*Interpretation*). An den hier diskutierten Beispielen heuristischer Textpraktiken konnten wir zeigen, dass diese drei Teilprozesse bei jeder Segmentierungsentscheidung eine zentrale Rolle spielen und dementsprechend zu reflektieren sind.

3.4 Die Sequenzierbarkeit der Novelle

Dass Segmentierung heute in der Literaturwissenschaft systematisch unterreflektiert erscheint, erstaunt nicht nur wegen ihrer großen Bedeutung bei der Operationalisierung von Textanalysen. Es erstaunt auch historisch im Hinblick auf alle generischen Ansätze in der Gattungspoetik, die seit dem 19. Jahrhundert zum festen Bestandteil der Literaturtheorie gehören. Vor allem bei den an der Schwelle zur Moderne viel theoretisierten Prosagattungen Märchen und Novelle hat die Segmentierbarkeit definitorischen Charakter. Vladimir Propps legendäre *Morphologie des Märchens* leitet aus der Analyse eines einschlägigen Korpus ein festes Set von Handlungssequenzen ab, aus dem sich jedes Märchen (re)produzieren lassen soll.⁴² Dieser Ansatz hat heute in der Filmbranche wieder Konjunktur, um Plots automatisch herzustellen.⁴³ Während Formalismus und früher Strukturalismus die generische Funktion von Segmenten und Sequenzen vor allem texttheoretisch ausgebeutet haben,⁴⁴ sind die früheren Versuche im Zeichen des literarischen Realismus für die aktuelle Forschung besonders wertvoll, weil sie noch stark auf Operationalisierungen in einem hermeneutischen Rahmen ausgelegt sind. Textmerkmale werden dabei nicht isoliert, sondern rezeptionsästhetisch im Hinblick auf ihre kognitive Wirkung beim Lesen betrachtet, was Wolfgang Iser später als das Zusammenwirken von Text- und Aktstruktur bezeichnet hat.⁴⁵

Im **Darmstädter LitLab** sind schon seit einiger Zeit verschiedene korpusbasierte Untersuchungen zu Novellen des 19. Jahrhunderts durchgeführt worden. Im Sommersemester 2019 wurde erstmals eine empirische Wirkungsstudie durchgeführt, um das gattungspoetisch bedeutsame Kriterium der Sequenzierbarkeit zu testen. Theoretische Grundlage dafür war neben Propp die realistische Gattungspoetik, die Paul Heyse in der Einleitung zum *Deutschen Novellenschatz* 1871 formuliert hat. Sie geht davon aus, dass die zeitgenössische Hochkonjunktur der Novelle mit der im literarischen Massenmarkt unvermeidlichen

⁴² Vgl. Propp 1975.

⁴³ Vgl. Batty 2014.

⁴⁴ Vgl. etwa Barthes 1971.

⁴⁵ Vgl. Iser 1976, S. 101–102.

Auseinandersetzung um die knappe Ressource Aufmerksamkeit erklärt werden kann und die daraus resultierenden Eigenschaften zugleich das zentrale Merkmal darstellen, das Novellen von anderen Erzähltexten unterscheidet. Eine Novelle zeichnet sich demnach dadurch aus, dass sie ein handlungsleitendes ›Grundmotiv‹ besitzt, das Leser:innen in die Lage versetzt, »den Inhalt in wenige Zeilen zusammenzufassen«. Heyse fordert dazu auf, die »Probe« zu machen, »ob der Versuch gelingt«.⁴⁶

Bei unserem Test kam es uns darauf an, von den insgesamt 86 Novellen der Anthologie eine auszuwählen, die der Einschätzung der historischen Herausgeber nach ihre Novellen-Definition gut erfüllt (den Einleitungen zufolge gab es bei der Aufnahme von Texten zum Teil auch andere Rücksichten) und zugleich unter pragmatischen Gesichtspunkten für ein Leseexperiment geeignet ist, vor allem was die Textlänge angeht. Die Wahl fiel auf Hieronymus Lorms *Ein adeliges Fräulein* (1867). Der Text hat ein klares Grundmotiv, das wie in den meisten Fällen als Dingsymbol angelegt ist. Auf der Suche nach einem bestimmten Gemälde trifft ein Kunstexperte auf die Besitzerin (Rahmenhandlung), die ihm die Geschichte des Bildes erzählt (Binnenhandlung). Es stammt aus dem Geschenkfundus einer gescheiterten Brautwerbung. Der Vater der adligen Titelheldin hatte sie einem vermögenden Bürgerlichen versprochen, fühlt sich daran jedoch nicht mehr gebunden, als ein adliger Mitbewerber auftaucht. Nachdem dieser sich jedoch als mittellos entpuppt, will der Vater die ursprünglich angebahnte Ehe schließen. Obwohl das ihren aufrichtigen Gefühlen entspricht, weigert sich die Tochter des gebrochenen Versprechens wegen.

Unser Experiment wurde am 20. Mai 2019 in einem Hörsaal der Technischen Universität Darmstadt mit 85 Proband:innen durchgeführt, die dafür eine Aufwandsentschädigung von 10 Euro erhielten. Den Teilnehmenden wurde neben einem Ausdruck der Novelle ein Fragebogen mit fünf Rubriken vorgelegt. An Metadaten wurden neben Alter, Geschlecht und Muttersprache die Lesehäufigkeit im Alltag und die Art der gelesenen Texte erhoben (Nachrichten, Sachbücher, Literatur, Social Media). Außerdem wurde gefragt, ob die Novelle vorab bekannt war. Im zweiten Schritt wurde darum gebeten, den Text aufmerksam durchzulesen. Die drei restlichen Aufgaben wurden nach der Lektüre absolviert. Wir fragten zunächst, was vom Inhalt des Textes als Erstes in den Sinn kommt, dann baten wir, das Geschehen in ganzen Sätzen zusammenzufassen, wozu unnummerierte Zeilenkästen vorgegeben wurden. Schließlich fragten wir nach der wesentlichen Textaussage.

Bei der Studie, die aus dem Experiment hervorgehen soll, arbeiten wir mit der Arbeitsgruppe ›Soziale Netzwerke‹ von Ulrik Brandes an der ETH Zürich zusammen.⁴⁷ Ziel der Studie ist es, herauszufinden, ob sich die als Gattungsmerkmal postulierte leichte Zusammenfassbarkeit der Novelle empirisch nachweisen lässt. Eine starke Übereinstimmung in den Zusammenfassungen der Teilnehmenden unseres Experiments wäre ein Beleg dafür. Grundlage der Auswertung sind die Transkriptionen der Fragebögen, die die Studierenden des Seminars ›Empirische Textanalysen‹ im Sommersemester 2019 an der Technischen Universität Darmstadt angefertigt haben. Da sich die Proband:innen bei ihren Zusammenfassungen weniger klar an die von uns vorgegebenen Zeilenkästen gehalten haben als erhofft, haben wir für die Auswertung jeden einzelnen Satz als Segment aufgefasst. Im Pre-Processing wurden die Stoppwörter entfernt und mit Hilfe der Levenshtein-Distanz⁴⁸ automatische Vereinheitlichungen vorgenommen (z. B. ›adelig‹ – ›adlig‹). Die Auswertung der Daten dauert derzeit noch an. In den bisherigen Berechnungen wurde für die Wortebene eines jeden Segments der *term frequency - inverse document frequency* (tf-idf)-Score ermittelt, um zu bestimmen, wie charakteristisch es sich im Korpus sämtlicher Segmente ausnimmt. Auf der Basis dieser Formalisierung kann die Distanz zu einer von uns selbst stammenden Muster-Zusammenfassung berechnet werden. Ein erstes *Clustering* der Ergebnisse schien unter Berücksichtigung unserer Metadaten die Tendenz zu zeigen, dass im Subset derjenigen Proband:innen, die häufig lesen, eine größere Übereinstimmung herrscht als im Subset der deutschen Muttersprachler:innen. Allerdings sind bei der Auswertung eine Reihe von Schwierigkeiten aufgetaucht, die wir erst bewältigen müssen, bevor wir solchen Ergebnissen Erklärungslasten aufbürden. Im Vergleich des *Goldstandards* unserer eigenen Muster-Zusammenfassung und den Zusammenfassungen der Fragebögen fiel uns auf, dass die Teilnehmenden vor allem Schwierigkeiten hatten, das Verhältnis von Rahmen- und Binnenhandlung in die verlangte Folge von ganzen Sätzen einzugliedern. Wir haben daher die Studierenden des Seminars ›Empirische Textanalysen‹ gebeten, aus den Satzfolgen der Fragebögen eine weitere Muster-Zusammenfassung zu erstellen. Wir wollen auf dieser Basis unsere bisherigen Ergebnisse und den Ansatz des Experiments kritisch hinterfragen.

3.5 Inter-Annotator-Agreement-Parameter als Heuristik für die Segmentierung literarischer Texte

Als letztes Beispiel wird ein Segmentierungsproblem vorgestellt, das in einem weiteren Projekt aus dem Bereich der *Computational Literary Studies* offensichtlich wurde. Dieses Projekt ist genauso wie der oben vorgestellte Zugang zur Sequenzierung datengetrieben, hat aber im Unterschied dazu eine automatisierte Textanalyse zum Ziel, die auf in der

⁴⁶ Vgl. Heyse / Kurz 1871, S. XIX.

⁴⁷ Autor:innen der Studie: Thomas Weitin, Katharina Herget, Anastasia Glawion, Simon Pöpcke, Ulrik Brandes.

⁴⁸ Levenshtein 1966 [1965].

Computerlinguistik bzw. der automatischen Sprachverarbeitung etablierten Verfahren zur Automatisierung basiert. Ziel des noch laufenden Projektes ist die Automatisierung der Erkennung szenenhafter Passagen in Prosatexten.⁴⁹ Als szenenhaft werden dabei jene Abschnitte erzählender Texte verstanden, in denen die Figurenkonstellation und der Raum der Erzählung weitgehend unverändert sind und die Geschehnisse chronologisch, zusammenhängend und weitgehend zeitdeckend erzählt werden.⁵⁰ Textgrundlage sind sogenannte Heftrömane, die in ihrer Struktur weniger komplex sind als Höhenkammliteratur und insbesondere weniger Varianz aufweisen. Im skizzierten Projekt ist das Auffinden der Szenen zwar das Ziel, allerdings sind diese Szenen als Vorverarbeitungsschritt für eine Reihe zukünftiger computationeller Analysen angelegt. Die automatisierte Annotation szenenhafter Passagen stellt nämlich eine für literaturwissenschaftliche computationelle Textanalysen geeignete Segmentierung literarischer Texte zur Verfügung.

Im ›Szenen-Projekt‹ wird ein für viele Analyseansätze in den Computational Literary Studies typischer Zugang umgesetzt: Als *Input* für die Automatisierung der Analyse literarischer Texte werden sogenannte *Golddaten* bzw. ein sogenannter Goldstandard erstellt. Dafür annotieren mehrere Annotator:innen anhand von Annotationsrichtlinien dieselben Texte. Auf Basis der so erstellten Daten – also der Texte und ihrer Annotationen – wird dann an der Automatisierung der Erkennung der annotierten Phänomene gearbeitet. Diese Praxis ist für die ansonsten mit exemplarischen Textstellen in Definitionen und Analysen arbeitende Literaturwissenschaft ungewöhnlich. In der literaturwissenschaftlichen Textanalyse werden traditionell nur einzelne Textstellen genutzt – um sie als besonders typische Textpassagen zu analysieren oder um die Definition von Phänomenen an ihnen zu veranschaulichen. Dahinter steht zumindest implizit eine behauptete Exemplarität und damit Repräsentativität der ausgewählten Beispiele.

Während in der Literaturwissenschaft also die Repräsentativität durch die Feststellung von Expert:innen gewährleistet wird, nähert man sich in Automatisierungszugängen einer Repräsentativität an, indem man mit einer großen Anzahl an Beispielen arbeitet und außerdem Annotationen mehrfach anfertigt, welche anschließend in eine konsolidierte Annotation im Sinne des Goldstandards überführt werden. Im computationellen Zugang ist die Repräsentativität der genutzten Textstellen fundamental, da aus diesen Textstellen zum Teil unüberwacht gelernt wird und entsprechend nach der Annotation keine weiteren menschlichen Analysen in den Prozess mit einfließen. Der literaturwissenschaftliche Umgang mit Beispielen und der Zugang zu Annotationen im Bereich der computationellen Literaturwissenschaft unterscheiden sich damit recht deutlich. Will man jedoch sicherstellen, dass ein Annotationsverfahren im Kontext von Automatisierungsaufgaben auch literaturwissenschaftlich adäquat ist, so muss man sich fragen, inwiefern ein Goldstandard auch eine literaturwissenschaftlich gute Textanalyse abbildet bzw. abbilden kann.

Eine naheliegende Begründung der Adäquatheit ist, dass in den Annotationen, die zur Erstellung des Goldstandards genutzt werden, intersubjektiv gültige Analysen der entsprechenden Textpassagen abgebildet werden. In der Literaturwissenschaft wird nämlich Intersubjektivität, also die Übereinstimmung mehrerer Subjekte in Bezug auf Urteile über literarische Texte, als geeignete Alternative zu einer – als nicht vorhanden bzw. zugänglich angenommenen – Realität oder objektiven Wahrheit betrachtet. Um wissenschaftlich zu sein, müssen literaturwissenschaftliche Befunde demnach eine »prinzipielle intersubjektive Vermittelbarkeit – einen ›sensus communis‹ [als von Kant für Geschmacksurteile angenommene Basis]« aufweisen.⁵¹ Literaturwissenschaftliche Analyse hat die Aufgabe, vorerst ohne Wertung »die Feststellung von allgemein beobachtbaren und intersubjektiv anerkehbaren Eigenheiten bestimmter Texte zu fixieren«⁵², wobei Ansätze wie die systemtheoretische oder die strukturalistische Literaturwissenschaft »die Möglichkeit rationaler, intersubjektiver Analysierbarkeit und Theoriebildung auch gegenüber Objekten wie der Literatur« postulieren.⁵³ Das bedeutet insbesondere, dass sie »prinzipiell explizierbare, rationale, intersubjektiv diskutierbare Methodologien und Theoriebildungen für den Objektbereich der Literaturwissenschaft« anstreben.⁵⁴

Vor diesem Hintergrund erscheinen *Inter-Annotator-Agreement-Maße* (IAA-Maße), die den Grad der Übereinstimmung zwischen Annotationen angeben, als Möglichkeit, um Intersubjektivität zu messen. Die Tatsache, dass bei typischen Annotationsaufgaben der Sprachverarbeitung wie etwa der *Part-of-Speech-Bestimmung* Übereinstimmungen von über 95 % durchaus möglich sind, eine solche hohe Übereinstimmung bei komplexeren Textphänomenen aber nicht erreicht werden kann, spricht ebenfalls dafür, dass IAA-Maße Intersubjektivität abbilden können.

⁴⁹ Vgl. dazu Gius et al. 2019; Zehe et al. 2021.

⁵⁰ Szene wird dabei verstanden als »segment of the discours (presentation) of a narrative which presents a part of the histoire (chronologically ordered, causally connected events in the narrated world) in such a way that a) time is more or less equal in discours and histoire, b) place stays – more or less – the same c) it centers around a particular action, and d) the character configuration is – again: more or less – equal« (Gius et al. 2019, Abs. 3). Für eine weitere Diskussion und Beispiele vgl. Zehe et al. 2021.

⁵¹ Vgl. Stöckmann 2013, S. 475.

⁵² Vgl. Fricke et al. (Hg.) 1997–2003, S. 447.

⁵³ Vgl. Fricke et al. (Hg.) 1997–2003, S. 535.

⁵⁴ Vgl. Fricke et al. (Hg.) 1997–2003, S. 536.

Nun war es aber so, dass im ›Szenen-Projekt‹ Szenengrenzen aus literaturwissenschaftlicher Sicht erstaunlich übereinstimmend annotiert wurden, dieser Umstand sich aber nicht in einem entsprechend hohen IAA-Maß niederschlug. Damit scheint der angenommene Zusammenhang zwischen dem Maß der Übereinstimmung zwischen zwei Annotator:innen und dem Grad der intersubjektiven Gültigkeit ihrer Analysen fraglich. Dies könnte auch daran liegen, dass IAA-Maße für sehr unterschiedliche Zwecke genutzt werden und diese nicht immer einen Bezug zu Intersubjektivität haben – etwa wenn sie eingesetzt werden, um die Konsistenz von Annotationen einzelner Annotator:innen zu überprüfen, die Qualität von Guidelines zu evaluieren oder Automatisierungsverfahren zu bewerten.⁵⁵ Konzentriert man sich aber auf die Messung von Übereinstimmung verschiedener Annotator:innen, also auf den Aspekt, der literaturwissenschaftlich als Intersubjektivität gefasst werden kann, müssen für die Erklärung der schlechten IAA-Werte im ›Szenen-Projekt‹ die in den IAA-Metriken abgebildeten Prinzipien betrachtet werden. IAA-Metriken beinhalten sehr differenzierte Berechnungen, die für eine gewisse Vergleichbarkeit der berechneten Werte sorgen sollen. Diese Berechnung im Hinblick auf Vergleichbarkeit kann man entsprechend als eine Operationalisierung von Intersubjektivität auffassen, anhand derer man die oben geforderte literaturwissenschaftlich adäquate Umsetzung von Annotationsverfahren bzw. des Goldstandards erreichen könnte. Für die Wahl einer geeigneten IAA-Metrik muss deshalb die jeweilige Operationalisierung von Vergleichbarkeit oder Übereinstimmung der IAA-Metrik – in Form von (Nicht-)Einbezug der erwarteten Übereinstimmung, Gewichtung der Nicht-Übereinstimmung in Abhängigkeit der betroffenen Kategorien usw. – berücksichtigt und gegebenenfalls geeignete Einstellungen der zur Verfügung gestellten Parameter gefunden werden.⁵⁷ Im Fall des ›Szenen-Projekts‹ hat sich bei der Beschäftigung mit der Vergleichbarkeit von IAA-Werten und Wahrnehmung der intersubjektiven Übereinstimmung neben dem Testen verschiedener Parametrisierungen schnell herausgestellt, dass die gewählte IAA-Metrik berücksichtigen muss, dass es sich um eine sogenannte Segmentierungsaufgabe handelt. Während etwa eine Part-of-Speech-Annotation auf vorausgewählten Segmenten (nämlich Wörtern) stattfindet, müssen Annotator:innen bei Segmentannotationen wie den oben beschriebenen Textpraktiken⁵⁸ die zu annotierende Textspanne selbst auswählen. Nutzt man für die Evaluation der manuellen Annotationen klassische IAA-Metriken wie Fleiss' π ,⁵⁹ Cohens κ ⁶⁰ oder Krippendorffs α ⁶¹, so erhält man meist schlechte Werte, weil diese Metriken nicht berücksichtigen, dass die zu annotierenden Texteinheiten nicht vorgegeben sind. Deshalb wurden für die Annotation von Segmenten eigene Metriken entwickelt.⁶² Mathet et al. geben einen Überblick über einige Segmentierungsmetriken und schlagen für die Darstellung bestehender Metriken und die Entwicklung ihres eigenen Vorschlags γ sechs Parameter vor, die Segmentierungsmetriken berücksichtigen sollten.⁶³ Diese berücksichtigen, dass eine Annotationsaufgabe (i) die Zuweisung von Kategorien (*Categorization*) und / oder (ii) die Bestimmung von Texteinheiten (*Unitizing*) beinhalten kann, die zu annotierenden Textphänomene im Text (iii) ineinander verschachtelt (*Embedding*) oder (iv) sich anderweitig überlappend (*Free Overlap*) vorkommen können sowie (v) nicht durchgehend vorhanden sein müssen (*Sporadicity*) und außerdem (vi) zwei aufeinanderfolgende Einheiten gegebenenfalls zu einer zusammengefasst werden können (*Aggregatable*).⁶⁴ Damit wird deutlich, wie wichtig die Wahl einer zur Annotationsaufgabe passenden IAA-Metrik ist – ein Umstand, der zumindest in der computationellen Literaturwissenschaft nur selten reflektiert wird. Die sechs Kategorien von Mathet et al. sind aber auch jenseits von Annotation und Automatisierung für die Arbeit mit literaturwissenschaftlichen Phänomenen hilfreich, da sie die Schärfung der genutzten Analysekonzepte unterstützen. Im Folgenden wird eine Auseinandersetzung mit diesen Kategorien anhand des Szenenkonzepts und der dort bestehenden Rahmenbedingungen skizziert.

Die *Categorization* ist grundlegend für jede Textanalyse: Man möchte dem Text bzw. seinen Teilen Kategorien zuweisen, die die analysierten Phänomene benennen. Selbst in dem diesbezüglich einfachen Fall im ›Szenen-Projekt‹, das nur auf einer Kategorie, nämlich ›Szene‹, aufbaut, kann man von der Zuweisung von zwei Kategorien ausgehen, da in der Annotation zwischen szenenhaften Passagen und nichtszenenhaften Textteilen unterschieden wird. Problematisch ist in solchen Fällen aber fast immer das *Unitizing*, welches die Segmentierung des Textes betrifft und das, wie bereits erläutert, in der Literaturwissenschaft nur bedingt standardisiert ist. Im ›Szenen-Projekt‹ wurde diskutiert, ob die oben besprochene grundlegende Diskurseinheit Satz als Basiseinheit gewählt werden soll oder eine größere, wie etwa der Absatz. Aufgrund der absehbaren Probleme bei der Bestimmung von Absätzen in Rohertexten wurde eine satzbasierte Annotation gewählt, wobei eine Szene normalerweise eine

⁵⁵ Vgl. dazu auch Gius / Vauth 2022.

⁵⁶ In der Computerlinguistik wird zum Teil auch keine Unterscheidung zwischen Intersubjektivität und Objektivität gemacht. So wird bei der Anwendung von IAA-Maßen für die Evaluation von Algorithmen nicht problematisiert, dass auch ein Goldstandard keine objektive Tatsache ist, sondern eben intersubjektiv erstellt. Pevzner und Hearst weisen zum Beispiel zwar auf das Problem der Bestimmung von Segmentgrenzen bzw. der Referenz für die Bewertung von Segmentierungsalgorithmen hin – »human judges do not always agree where boundaries should be placed and how fine-grained an analysis should be«. Der folgende Verweis auf eine der praktizierten Lösungen – »others have several human judges make ratings to produce a »gold standard« wird dann gemacht, ohne zu thematisieren, dass diese gegebenenfalls anhand derselben Maße gemessen wird (vgl. Pevzner / Hearst 2002, S. 2).

⁵⁷ Für eine zusammenfassende Darstellung der gängigen Koeffizienten-Berechnungen vgl. Artstein / Poesio 2008, S. 560–570, sowie den aktuellen Überblick von Reiter / Konle 2022.

⁵⁸ Vgl. Bender / Müller 2020; Teufel 1999.

⁵⁹ Vgl. Fleiss 1971.

⁶⁰ Vgl. Cohen 1960.

⁶¹ Vgl. Krippendorff 1980; Krippendorff 2004.

⁶² Gängige IAA-Metriken für Segmentierung sind Pk (Beeferman et al. 1997), WindowDiff (Pevzner / Hearst 2002; Lamprier et al. 2007), GHD (Bestgen 2009), Boundary Edit Distance (Fournier / Inkpen 2012), Boundary Similarity (B) (Fournier 2013) und γ (Mathet et al. 2015). Für eine Diskussion der Metriken bis 2009 vgl. Fournier 2013; Mathet et al. 2015.

⁶³ Vgl. Mathet et al. 2015.

⁶⁴ Vgl. Mathet et al. 2015, S. 440.

ganze Reihe von Sätzen umfasst. Diese vermeintlich triviale, durch technische Einschränkungen motivierte Entscheidung hat aber auch konzeptuelle Konsequenzen, da sie die Granularität des Szenenbegriffs beeinflusst. Eine zusätzliche Annäherung an Segmentgrößen in Form der Angabe des typischen Textumfangs, den ein Phänomen umfasst, hat sich außerdem als hilfreiche Heuristik herausgestellt.⁶⁵

	<i>Szene als Textphänomen</i>	<i>Operationalisierung von Szene</i>	<i>Schwierigkeit</i>
Categorization	X	X	– (ist Voraussetzung für Textanalyse)
Unitizing	X	X	literaturwissenschaftliche Grundlage fehlt
Embedding	X	-	zur Vereinfachung der Annotationsaufgabe nicht berücksichtigt
Free Overlap	-	-	–
Sporadicity	X	X	erschwert die Annotationsaufgabe
Aggregatable	x	?	nicht generalisierbar

Tab. 4: Parameter von Mathet et al. 2015 angewendet auf Szenen (Phänomen und Operationalisierung in den Guidelines).

Die Frage, ob Phänomene verschachtelt (*Embedding*) oder sich anderweitig überlappend (*Free Overlap*) auftreten können, ist wiederum verbunden mit ihrer Definition. Können die Phänomene in eine Begriffstaxonomie eingeordnet werden, die hierarchisch organisiert ist oder auch darüber hinausgehende, nicht hierarchische Verbindungen zwischen den Kategorien aufweist? Dazu gehört auch die Frage von über- und untergeordneten Phänomenen, in die die Untersuchungsbegriffe gegebenenfalls aufgeteilt werden können. Für Szenen kann man annehmen, dass sie verschachtelt auftreten können, da sie – ähnlich wie Erzählebenen – von anderen Szenen unterbrochen und anschließend weitergeführt werden können. Nicht zuletzt aufgrund der relativ wenig komplexen Handlungsabläufe in den untersuchten Hefromanen haben wir auf Verschachtelungen verzichtet. Diese Entscheidung war zusätzlich dadurch motiviert, dass eine hierarchische Szenenstruktur für das maschinelle Lernen eine ungleich größere Herausforderung bedeutet. Unabhängig von der Entscheidung für oder gegen Verschachtelung und Überschneidung muss geprüft werden, welche Konsequenzen die jeweilige Ausprägung für die Analysekonzepte hat. Die daraus resultierende Taxonomie führt nämlich dazu, dass die genutzten Phänomene sich gegenseitig ausschließen, über- oder unterordnen etc. – oder eben nicht.

Die Frage nach der durchgehenden Präsenz von Phänomenen im Text (*Sporadicity*) betrifft u. a. das Textkonzept. Ausschlaggebend ist hier, inwiefern die untersuchten Phänomene als für Texte konstitutiv und ob Texte gegebenenfalls als noch aus weiteren Phänomenen zusammengesetzt gesehen werden. Für die Szenenanalyse kann man davon ausgehen, dass es Textabschnitte gibt, die nicht als Szenen gelten sollten. Dies liegt daran, dass das Zeitkriterium – im Gegensatz zu den Kriterien Wechsel von Raum, Figuren oder Handlung – nicht für einen Wechsel, sondern für die Szenenhaftigkeit an sich steht und entsprechend Abschnitte, die es nicht erfüllen, nicht szenenhaft sind. Im Gegensatz dazu gehen etwa die meisten Konzepte von Erzählebenen davon aus, dass jeder Abschnitt eines Textes (mindestens) einer Erzählebene zugeordnet werden kann. Dies ist in der Analysepraxis der einfachere Fall. Das Erkennen von Nicht-Szenen hat sich nämlich als durchaus problematisch erwiesen, da das zugrunde liegende Kriterium, wie in der Textanalyse häufig, graduell ist.

Schließlich ist die Frage der Aggregierbarkeit (*Aggregatable*) im Falle der Szenen problematisch. Während direkt aufeinanderfolgende Szenen potenziell zu einer einzelnen Szene zusammengefasst werden können, wenn man die Szenen als kleinere Unterteilungen einer großen Szene interpretieren kann, sollten aufeinanderfolgende Szenen, die sich sehr deutlich in Raum, Zeit, Figuren oder Handlung unterscheiden, nicht zu einer längeren Szene zusammengefasst werden. An dieser Stelle wird offensichtlich, dass die aufgrund der damit verbundenen Probleme nicht vorgenommene Vorsegmentierung der Texte nun ihrerseits zu Problemen führt.

Auch wenn die sechs vorgestellten Parameter für Analysen der Übereinstimmung von Annotationen auf der Textoberfläche vorgeschlagen werden, soll die Darstellung der einzelnen Punkte zeigen, dass sie eine darüber hinausgehende Relevanz für die literaturwissenschaftliche Beschreibung und Definition von Textphänomenen haben. Ihre Explizierung kann erheblich zur

⁶⁵ Vgl. Gius 2016, S. 12.

intersubjektiven Verständlichkeit von Phänomenen und ihren Definitionen beitragen. Auch die aufgezeigten Zusammenhänge zwischen den Parametern können bei der Entwicklung intersubjektiv besser nachvollziehbarer und damit stabilerer Definitionen nützlich sein.

4. Sinn und Segment

Die in diesem Beitrag diskutierten Aspekte von Segmentierung reichen von der Layoutanalyse über die Identifizierung lexikalischer Mehrwortausdrücke, die Auseinandersetzung mit heuristischen Textpraktiken und die Frage nach der Sequenzierbarkeit von Novellen bis hin zu Segmenten als intersubjektiv identifizierbare Einheiten in Prosatexten. Segmentierung wurde also im Kontext (zumindest scheinbar) standardisierter Segmentierungspraktiken wie dem *Unitizing*, der Tokenisierung oder der Bestimmung von Diskurs- und Layouteinheiten, aber auch in Bezug auf den Zugang zu nicht ohne Weiteres linguistisch und / oder am Schriftbild bzw. Layout bestimmbar Segmentkonzepten diskutiert. Die Bandbreite der Ansätze deckt damit weite Teile der philologischen Analysepraxis ab und verdeutlicht, dass Segmentierung im philologischen Kontext wesentliche Konsequenzen hat. Das gilt unabhängig davon, ob im Forschungszugang die Bestimmung der Segmente im Zentrum steht oder auf Segmentierung aufbauende Forschung betrieben wird. Für alle Zugänge lässt sich beobachten: Für eine geeignete Segmentierung des Untersuchungsgegenstands muss das relevante Wissen zu theoretischen Konzepten mit der Gestaltung des Analyseprozesses und Eigenschaften des untersuchten Mediums in Einklang gebracht werden. Segmentierung kann man entsprechend als eine Art Vorverarbeitung des zu analysierenden Textmaterials sehen, während erst das Ergebnis der darauf aufbauenden Analyse und Interpretation als die anvisierte Erkenntnis betrachtet werden kann. Aufgrund der skizzierten Zusammenhänge besteht allerdings eine Interaktion von Vorverarbeitung und Analyse, die zu einer Kreisbewegung führt. Entsprechend fehlt der gesamten Tätigkeit ein offensichtlicher Anfangspunkt. Die epistemische – oder auch hermeneutische – Zwickmühle wird hier schnell offensichtlich: Vorverarbeitung und Analyse sollten getrennt stattfinden, aber wir können nicht interpretieren, ohne zu segmentieren, und wir können nicht segmentieren, ohne zu interpretieren. Wünschenswert ist daher in erster Linie die Transparenz der Methoden und ihres Ineinandergreifens bzw. die konsequente Reflexion und Explizierung des Forschungsprozesses.

Dies gilt für Segmentierung unabhängig vom digitalen Zugang. Wie jede Textanalyseaufgabe ist auch Segmentierung nicht trivial oder zumindest ist die Aufgabe in vielen Fällen nicht eindeutig lösbar. Im analogen Analysemodus können und werden diese Unbestimmtheiten allerdings regelmäßig nicht expliziert, jedenfalls aber häufig hingenommen. Der digitale Zugang wirkt hingegen – wie bei jeder anderen Textanalyseaufgabe – problemverstärkend. Die Bestimmung von für eine Analyse adäquaten Textsegmenten wird in digitalen Zugängen dadurch erschwert, dass diese den Zwang des Diskreten mit sich bringen. Die Operationalisierung der Segmentierung resultiert in binären Entscheidungen – etwas ist ein Segment oder nicht. Hier liegt der wesentliche Unterschied zwischen digitalen und analogen Zugängen: Der digitale Zugang erzwingt gewissermaßen die Offenlegung von Segmentierungsentscheidungen und fördert stark die Festlegung und Explizierung von Kriterien für diese Entscheidungen. Die algorithmische Formulierung sowie deren Implementierung als eindeutige und diskrete Segmentierungsentscheidung schärft damit unsere Begriffe.

Die diskutierten Segmentierungsprobleme verdeutlichen außerdem, dass Segmentierungsentscheidungen den Forschungsprozess wesentlich beeinflussen. Dieser Zusammenhang ist ebenfalls unabhängig vom digitalen Zugang, allerdings wird er durch diesen leichter offensichtlich. So wird die Menge der Untersuchungseinheiten sowie deren Granularität durch die Segmentierungsentscheidungen bestimmt. Man kann auch sagen: Nur was als Segment einer Analyseebene identifiziert und segmentiert wird, wird auch analysiert; aber auch: Was analysiert werden soll, also Gegenstand der Untersuchung sein soll, muss auch segmentiert, also als diskrete Einheit identifiziert werden (können). Hinzu kommt: Aufgrund der beschriebenen Interaktion von Segmentierung und Interpretation werden – unabhängig vom digitalen Zugang – bereits im Korpusaufbau Segmentierungsentscheidungen getroffen, die sich auf die späteren Ergebnisse auswirken. Im automatisierten Zugang wird es insbesondere dann problematisch, wenn die für die ursprüngliche Segmentbestimmung relevanten Unsicherheiten und Entscheidungen nicht mehr ohne Weiteres zugänglich oder aufgrund ihrer schieren Menge nicht mehr erfassbar sind.

Bibliografische Angaben

- Marc Adler / Sabine Bartsch / Maria Becker / Michael Bender / Luise Borek / Cindy Brinkmann / Friedrich Michael Dimpel / Rotraut Fischer / Anastasia Glawion / Svenja A. Gülden / Canan Hastik / Philipp Hegel / Katharina Herget / Franziska Horn / Celia Krause / Marcus Müller / Alexandra Núñez / Andrea Rapp / Lisa Scharrer / Oliver Schmid / Jörn Stegmeier / Beate Thull / Thomas Weitin: *Digitale Philologie: Das Darmstädter Modell*. Darmstadt 2020. (= Digital Philology. Working Papers in Digital Philology, 1). DOI: 10.25534/tuprints-00012476
- Emil Angehrn: *Sinn und Nicht-Sinn. Das Verstehen des Menschen*. Tübingen 2010. (= Philosophische Untersuchungen, 25). [Nachweis im GVK]
- Thomas Anz: *Textwelten*. In: *Handbuch Literaturwissenschaft*. Hg. von Thomas Anz. 3 Bde. Stuttgart 2013. Bd. 1: *Gegenstände und Grundbegriffe*, S. 111–130. [Nachweis im GVK]
- Ron Artstein / Massimo Poesio: *Inter-Coder Agreement for Computational Linguistics*. In: *Computational Linguistics* 34 (2008), H. 4, S. 555–596. DOI: 10.1162/coli.07-034-R2 [Nachweis im GVK]
- Roland Barthes: *Action Sequences*. In: *Patterns of Literary Style*. Hg. von Joseph Strelka. University Park, PA u. a. 1971, S. 5–14. (= Yearbook of Comparative Criticism, 3). [Nachweis im GVK]
- Craig Batty: *„Show Me Your Slugline and I'll Let You Have the Firstlook“: Some Thoughts on Today's Digital Screenwriting Tools and Aprs*. In: *Media International Australia* 153 (2014), H. 1, S. 118–127. DOI: 10.1177/1329878X1415300114 [Nachweis im GVK]
- Sabine Bartsch: *Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints in Lexical Co-occurrence*. Tübingen 2004. [Nachweis im GVK]
- Sabine Bartsch / Stefan Evert: *Towards a Firthian Notion of Collocation*. In: *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*. Hg. von Andrea Abel / Lothar Lemnitzer. Mannheim 2014, S. 48–61. (= OPAL - Online publizierte Arbeiten zur Linguistik, 2014.2). URN: urn:nbn:de:bsz:mh39-24029
- Maria Becker / Michael Bender / Marcus Müller: *Classifying Heuristic Textual Practices in Academic Discourse. A Deep Learning Approach to Pragmatics*. In: *International Journal of Corpus Linguistics* 25 (2020), H. 4, S. 426–460. 11.11.2020. DOI: 10.1075/ijcl.19097.bec [Nachweis im GVK]
- Doug Beeferman / Adam Berger / John Lafferty: *Text Segmentation Using Exponential Models*. In: *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP: 2, Providence, 01.–02.08.1997)*. Somerset, NJ u. a. 1997, S. 35–46. [online] [Nachweis im GVK]
- Michael Beißwenger / Sabine Bartsch / Stefan Evert / Kay-Michael Würzner: *EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-mediated Communication and Web Corpora*. In: *Proceedings of the 10th Web as Corpus Workshop and the EmpiriST Shared Task*. Hg. von Association for Computational Linguistics. (WAC-X: 10, Berlin, 12.08.2016). Stroudsburg, PA 2016, S. 44–56. DOI: 10.18653/v1/W16-2606
- Michael Bender / Marcus Müller: *Heuristische Textpraktiken. Eine kollaborative Annotationsstudie zum akademischen Diskurs*. In: *Zeitschrift für Germanistische Linguistik* 48 (2020), H. 1, S. 1–46. DOI: 10.1515/zgl-2020-0001 [Nachweis im GVK]
- Yves Bestgen: *Quel indice pour mesurer l'efficacité en segmentation de textes? In: Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*. Hg. von Adeline Nazarenko / Thierry Poibeau. Senlis 2009, S. 171–180. [online]
- Martin Bullin / Andreas Henrich: *Die inhaltsbasierte Bildsuche und Bilderschließung. Ansätze und Problemfelder*. In: *Bilddaten in den digitalen Geisteswissenschaften*. Hg. von Canan Hastik / Philipp Hegel. Wiesbaden 2020, S. 11–33 (= Episteme in Bewegung, 16). DOI: 10.13173/9783447114608 [Nachweis im GVK]
- Jacob Cohen: *A Coefficient of Agreement for Nominal Scales*. In: *Educational and Psychological Measurement* 20 (1960), H. 1, S. 37–46. DOI: 10.1177/001316446002000104 [Nachweis im GVK]
- Owen Dieleman: *Hinweise für die Entwicklung von Verfahren zur maßnahmenartübergreifenden Dringlichkeitsbewertung von Straßenbaumaßnahmen*. Darmstadt 2016. (= Schriftenreihe des Instituts für Verkehr, Darmstadt, Technische Universität Darmstadt, 34). [online] [Nachweis im GVK]
- Stefan Evert: *Corpora and Collocations*. In: *Corpus Linguistics: An International Handbook*. Hg. von Anke Lüdeling / Merja Kytö. 2. Bde. Berlin u. a. 2008. Bd. 2, S. 1212–1248. (= Handbooks of Linguistics and Communication Science, 29). DOI: 10.1515/9783110213881.2.1212
- John Rupert Firth: *Papers in Linguistics 1934–1951*. London u. a. 1964 [1957]. URN: urn:oclc:record:1150956406 [Nachweis im GVK]
- John Rupert Firth: *Selected Papers of J. R. Firth. 1952–59*. Hg. von Frank Robert Palmer. London 1968. [Nachweis im GVK]
- Joseph L. Fleiss: *Measuring Nominal Scale Agreement among Many Raters*. In: *Psychological Bulletin* 76 (1971), H. 5, S. 378–382. DOI: 10.1037/h0031619 [Nachweis im GVK]
- Chris Fournier: *Evaluating Text Segmentation Using Boundary Edit Distance*. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Long Papers*. Hg. von Hinrich Schuetze / Pascale Fung / Massimo Poesio. (ACL 51: Sofia, 04.–09.08.2013). Stroudsburg, PA 2013, S. 1702–1712. PDF. [online]
- Chris Fournier / Diana Inkpen: *Segmentation Similarity and Agreement*. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Hg. von Eric Fosler-Lussier / Ellen Riloff / Srinivas Bangalore. (NAACL: Montréal, 03.–08.06.2012). Stroudsburg, PA 2012, S. 152–161. [online] [Nachweis im GVK]
- Reallexikon der deutschen Literaturwissenschaft. *Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Hg. von Harald Fricke / Klaus Grubmüller / Jan-Dirk Müller / Klaus Weimar. 3 Bde. Berlin 1997–2003. [Nachweis im GVK]
- Hans-Walter Gabler: *The Primacy of the Document in Editing*. In: *Ecdotica* 4 (2007), S. 197–207. [Nachweis im GVK]
- Evelyn Gius: *Narration and Escalation. An Empirical Study of Conflict Narratives*. In: *Diegesis* 5 (2016), H. 1, S. 4–25. [online] [Nachweis im GVK]
- Evelyn Gius / Fotis Jannidis / Markus Krug / Albin Zehe / Andreas Hotho / Frank Puppe / Jonathan Krebs / Nils Reiter / Natalie Wiedmer / Leonard Konle: *Detection of Scenes in Fiction*. In: *Digital Humanities 2019 Conference papers. Book of Abstracts*. (DH 2019: Utrecht, 09.–12.07.2019). Utrecht 2019.
- Evelyn Gius / Michael Vauth: *Inter Annotator Agreement and Intersubjektivität*. In: *DHD2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*. (DHD 8: Potsdam, 07.–11.03.2022). Potsdam 2022. DOI: 10.5281/zenodo.6328208
- Michael Alexander Kirkwood Halliday / Ruqaiya Hasan: *Cohesion in English*. London 1976. [Nachweis im GVK]
- Handbuch Literaturwissenschaft*. Hg. von Thomas Anz. 3 Bde. Stuttgart 2013. [Nachweis im GVK]
- Rainer Herzog: *Ein generischer Ansatz zur digitalen Layoutanalyse von Manuskripten*. Hamburg 2018. PDF. [online] [Nachweis im GVK]
- Paul Heyse / Hermann Kurz: *Einleitung*. In: *Deutscher Novellenschatz*. Hg. von Paul Heyse / Hermann Kurz. 24 Bde. München 1871–1876, Bd. 1 (1871), S. V–XXIV. [Nachweis im GVK]
- Wolfgang Iser: *Der Akt des Lesens. Theorie ästhetischer Wirkung*. München 1976. [Nachweis im GVK]
- Jerrold Jacob Katz: *Sense, Reference, and Philosophy*. Oxford u. a. 2004. [Nachweis im GVK]
- Diskurs – Interpretation – Hermeneutik*. Hg. von Reiner Keller. Weinheim 2015. (= Zeitschrift für Diskursforschung / Beihefte, 1). [Nachweis im GVK]
- Tilman Köppe / Simone Winko: *Theorien und Methoden der Literaturwissenschaft*. In: *Handbuch Literaturwissenschaft*. Hg. von Thomas Anz. 3 Bde. Stuttgart 2013. Bd. 2: *Methoden und Theorien*, S. 285–371. [Nachweis im GVK]

- Michael Krewet: Wissenstransfer in Scholien. Zur Präsenz Platons in den Marginalien von de interpretatione-Handschriften. Berlin 2015. (= Working Paper des SFB 980 Episteme in Bewegung, 6). PDF. [\[online\]](#) [\[Nachweis im GVK\]](#)
- Michael Krewet / Philipp Hegel / Germaine Götzelmann / Sibylle Söring / Danah Tonne: Aristoteles auf Reisen: Handschriftenforschung in der digitalen Infrastruktur des SFBs 980 »Episteme in Bewegung«. In: Forschungsinfrastrukturen in den digitalen Geisteswissenschaften. Hg. von Martin Huber / Sybille Krämer / Claus Pias: Fachinformationsdienst Allgemeine und Vergleichende Literaturwissenschaft. (DFG-Symposienreihe Digitalität in den Geisteswissenschaften, Bayreuth, 26.–28.09.2018). Frankfurt / Main 2019, S. 77–87. PDF. [\[online\]](#) [\[Nachweis im GVK\]](#)
- Michael Krewet / Philipp Hegel: Diagramme in Bewegung: Scholien und Glossen zu »de interpretatione«. In: Bilddaten in den digitalen Geisteswissenschaften. Hg. von Canan Hastik / Philipp Hegel. Wiesbaden 2020, S. 199–216. (= Episteme in Bewegung, 16). [\[Nachweis im GVK\]](#)
- Klaus Krippendorff: Content Analysis: an Introduction to Its Methodology. Beverly Hills, CA 1980. (= The Sage Commtext Series, 5). [\[Nachweis im GVK\]](#)
- Klaus Krippendorff: Reliability in Content Analysis: Some Common Misconceptions and Recommendations. In: Human Communication Research 30 (2004), H. 3, S. 411–433. [\[Nachweis im GVK\]](#)
- Sylvain Lamprier / Tassadit Amghar / Bernard Levrat / Frederic Saubion: On Evaluation Methodologies for Text Segmentation Algorithms. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence. (ICTAI 19: Patras, 29.–31.10.2007). Los Alamitos, CA 2007, S. 19–26. DOI: 10.1109/ICTAI.2007.142 [\[Nachweis im GVK\]](#)
- Vladimir I. Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In: Soviet Physics Doklady 10 (1966) [1965], H. 8, S. 707–710. [\[Nachweis im GVK\]](#)
- Materialität in der Editionswissenschaft. Hg. von Martin Schubert. Berlin 2010. (= Editio / Beihefte, 32). DOI: 10.1515/9783110231311
- Yann Mathet / Antoine Widlöcher / Jean-Philippe Métivier: The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. In: Computational Linguistics 41 (2015), H. 3, S. 437–479. DOI: 10.1162/COLI_a_00227 [\[Nachweis im GVK\]](#)
- Metzler Lexikon Sprache. Hg. von Helmut Glück / Michael Rödel. 5. aktualisierte und überarbeitete Auflage. Stuttgart 2016. [\[Nachweis im GVK\]](#)
- Katrin Ortman / Adam Roussel / Stefanie Dipper: Evaluating Off-the-Shelf NLP Tools for German. In Proceedings of the 15th Conference on Natural Language Processing. Hg. von German Society for Computational Linguistics & Language Technology. (KONVENS 15: Erlangen, 09.–11.10.2019). 2019, S. 212–222. [\[Nachweis im GVK\]](#)
- Lev Pevzner / Marti Alice Hearst: A Critique and Improvement of an Evaluation Metric for Text Segmentation. In: Computational Linguistics 28 (2002), H. 1, S. 19–36. [\[Nachweis im GVK\]](#)
- Wolfgang Pfeifer et al.: Etymologisches Wörterbuch des Deutschen, digitalisierte und von Wolfgang Pfeifer überarbeitete Version im Digitalen Wörterbuch der deutschen Sprache. Berlin 1993. HTML. [\[online\]](#)
- Vladimir Propp: Morphologie des Märchens. Frankfurt / Main 1975. [\[Nachweis im GVK\]](#)
- Nils Reiter / Leonard Konle: Messverfahren zum Inter-annotator-agreement (IAA): Eine Übersicht. Göttingen 2022. (= DARIAH-DE Working Papers, 44). PDF. DOI: 10.47952/gro-publ-103
- Ferdinand de Saussure: Cours de linguistique générale. Hg. von Charles Bally / Albert Sechehaye. Lausanne u. a. 1916. [\[Nachweis im GVK\]](#)
- Helmut Schmid: Probabilistic Part-of-Speech Tagging Using Decision Trees. (International Conference on New Methods in Language Processing, Manchester, 06.07–08.07.1994). Manchester 1994. PDF. [\[online\]](#)
- Helmut Schmid: Improvements in Part-of-Speech Tagging with an Application to German. (ACL SIGDAT-Workshop, Dublin 1995, 27.03.1995). PDF. [\[online\]](#)
- Wolf Schmid: Erzähltextanalyse. In: Handbuch Literaturwissenschaft. Hg. von Thomas Anz. 3 Bde. Stuttgart 2013. Bd. 2: Methoden und Theorien, S. 98–120. [\[Nachweis im GVK\]](#)
- Monika Siegel: Ich hatte einen Hang zur Schwaermerey ... Das Leben der Schriftstellerin und Übersetzerin Meta Forkel-Liebeskind im Spiegel ihrer Zeit. Darmstadt 2002. PDF. [\[online\]](#)
- John Sinclair: Corpus, Concordance, Collocation. Oxford 1991. (= Describing English Language). [\[Nachweis im GVK\]](#)
- Ingo Stöckmann: Ästhetik. In: Handbuch Literaturwissenschaft. Hg. von Thomas Anz. 3 Bde., Stuttgart 2013. Bd. 1: Gegenstände und Grundbegriffe, S. 465–491. [\[Nachweis im GVK\]](#)
- Simone Teufel: Argumentative Zoning: Information Extraction from Scientific Text. Dissertation, University of Edinburgh. 1999. PDF. [\[online\]](#)
- Wolf Thümmel: Segmentierung. In: Metzler Lexikon Sprache. Hg. von Helmut Glück / Michael Rödel. 5. aktualisierte und überarbeitete Auflage. Stuttgart 2016, S. 602. [\[Nachweis im GVK\]](#)
- Donatus Thürnau: Sinn/Bedeutung. In: Historisches Wörterbuch der Philosophie. Onlineversion. Hg. von Joachim Ritter, Karlfried Gründer und Gottfried Gabriel. 2017. DOI: 10.24894/HWPh.3901
- Kristina Toutanova / Christopher David Manning: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (EMNLP/VLC-2000: Hong Kong, Oktober 2000) S. 63–70. PDF. DOI: 10.3115/1117794.1117802
- Albin Zehe / Leonard Konle / Lea Katharina Dümpelmann / Evelyn Gius / Andreas Hotho / Fotis Jannidis / Lucas Kaufmann / Markus Krug / Frank Puppe / Nils Reiter / Anneke Schreiber / Natalie Widmer: Detecting Scenes in Fiction. A New Segmentation Task. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Hg. von Association for Computational Linguistics. (EACL 16: online, 19.–23.04.2021) 2021, S. 3167–3177. PDF. DOI: 10.18653/v1/2021.eacl-main.276

Abbildungs- und Tabellenverzeichnis

Tab. 1: Linguistische Einheiten der Segmentierung.

Tab. 2: Beispiele lexikalischer Mehrwortausdrücke.

Abb. 1: Das taxonomische Annotationsschema HeuTex. [Bender / Müller 2020, S. 23]

Tab. 3: Ergebnis der RNN-Klassifizierung auf verschiedenen Ebenen. [Aus: Becker et al. 2020]

Abb. 2: Dimensionen der Segmentierung. [Eigene Darstellung]

Tab. 4: Parameter von Mathet et al. (2015) angewendet auf Szenen (Phänomen und Operationalisierung in den Guidelines).