Originalveröffentlichung in: Beißwenger, Michael/Lemnitzer, Lothar/Müller-Spitzer, Carolin (Hrsg.): Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium. – Paderborn: Wilhelm Fink, 2022. S. 225-245. (UTB 5711)

POSTPRINT

Annelen Brunner
Fotis Jannidis

12. Redewiedergabe in Hochliteratur und Heftromanen

Diese Fallstudie[1] untersucht die quantitative Verteilung von direkten und nicht-direkten Formen von Redewiedergabe im Vergleich zwischen zwei Literaturtypen: Hochliteratur – definiert als Werke, die auf der Auswahlliste von Literaturpreisen standen – und Heftromanen – massenproduzierten Erzählwerken, die zumeist über den Zeitschriftenhandel vertrieben werden. Die Studie geht von manuell annotierten Daten aus und überprüft daran die Verlässlichkeit automatischer Annotationswerkzeuge, die im Anschluss eingesetzt werden, um eine Untersuchung von insgesamt 250 Volltexten durchzuführen. Es kann nachgewiesen werden, dass sich die Literaturtypen sowie auch unterschiedliche Genres von Heftromanen hinsichtlich der verwendeten Wiedergabeformen unterscheiden.

1. Einleitung

Die vorgestellte Studie liegt an der Schnittstelle zwischen Linguistik und Literaturwissenschaft insofern, als dass sie eine literaturwissenschaftliche Fragestellung behandelt, die sich in der sprachlichen Form manifestiert. Der Begriff ‚Redewiedergabe' im Titel steht hier für die Art und Weise, wie in einem Erzähltext die Rede, aber auch die Gedanken und schriftlichen Äußerungen einer Figur wiedergegeben werden. Welche Form gewählt wird – das direkte Zitat? eine Paraphrasierung? vielleicht nur eine kurze Erwähnung, dass eine Rede- oder Gedankenhandlung stattgefunden hat? – verrät viel über die Figurendarstellung im Text, das Verhältnis zwischen der Erzählerinstanz und den Figuren und die Darstellung der erzählten Welt insgesamt. Wenn etwa die Äußerungen von Figuren direkt zitiert werden, so wirkt dies üblicherweise unmittelbarer und eröffnet zudem die Möglichkeit, Charakterzüge einer Figur über deren Ausdrucksweise zu kommunizieren. Werden die Figurenstimmen durch die Erzählerstimme gefiltert, so verlieren diese einen Teil ihrer individuellen Färbung. Zugleich können leichter Wertungen und Schwerpunktsetzungen – etwa durch Zusammenfassung – vorgenommen werden. Die Erzählerinstanz hat damit explizitere Kontrolle darüber, wie die Figuren wahrgenommen werden. Die verschiedenen Techniken der Wiedergabe werden deswegen in der literaturwissenschaftlichen Erzählforschung häufig auf einer Skala angeordnet zwischen unmittelbarer und mittelbarer Darstellung (vgl. Genette 1998; Martínez und Scheffel 2016) oder auch nach dem Grad der Treue zur ‚Originaläußerung' der Figur (*faithfulness*, vgl. Leech und Short 2013). Es handelt sich um einen der klassischen Aspekte, die bei der wissenschaftlichen Untersuchung von Erzähltechnik betrachtet werden, und die Frage nach der Einbindung von Figurenreden und -gedanken ist Teil von fast allen umfassenderen erzähltheoretischen Darstellungen. Der Ansatz, dieses Thema von einer stark quantitativen Perspektive zu betrachten, ist jedoch ungewöhnlich und neuartig.

Die vorgestellte Studie untersucht Redewiedergabe im Vergleich zwischen zwei Literaturtypen von gegensätzlichen Enden des Spektrums: Hochliteratur und Heftromane. Heftromane sind massenproduzierte Erzählwerke unterschiedlicher Genres, die zumeist

1 Diese Studie wurde im Rahmen des DFG-Projekts „Redewiedergabe" durchgeführt, gemeinsam mit meinen Kolleg*innen Ngoc Duyen Tanja Tu, Lukas Weimer, Stefan Engelberg und Fotis Jannidis, denen ich an dieser Stelle herzlich danke! Der Inhalt wurde auch in folgender Veröffentlichung vorgestellt: Brunner, Jannidis, Engelberg, et al. (2020).

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-110650

über den Zeitschriftenhandel vertrieben werden und früher abwertend als „Romane der Unterschicht" (vgl. Nusser 1979) bezeichnet wurden. In der neueren Literaturforschung haben diese Hefte, die bestimmt jeder schon einmal im Zeitschriftenhandel gesehen hat, wieder etwas Aufmerksamkeit erfahren und werden nicht mehr so abfällig betrachtet (vgl. z. B. Nast 2017; Stockinger 2018). Es ist jedoch nicht von der Hand zu weisen, dass sie deutlich andere Produktionsbedingungen und auch ein anderes Publikum haben als die Literatur, die wir als ‚Hochliteratur' bezeichnen – gesellschaftlich besonders geschätzte Werke, die als wertvoll und ‚anspruchsvoll' gelten.

2. Fragestellung

Unsere Forschungsfrage lautet: Gibt es zwischen Heftromanen und Hochliteratur messbare Unterschiede in den verwendeten Wiedergabeformen? Ausgehend davon betrachten wir auch folgende, weiterführende Fragen:

1. Heftromane lassen sich in unterschiedliche Genres wie Krimi, Science-Fiction oder Liebesroman untergliedern, die durch Reihen („Jerry Cotton", „Perry Rhodan", „Julia Extra" etc.) repräsentiert sind. Wie verhalten sich diese Genres zueinander und zur Hochliteratur in Hinblick auf die verwendeten Wiedergabeformen? Sind die Unterschiede zwischen den Genres untereinander vielleicht größer als zur Hochliteratur?
2. Bei der direkten Wiedergabe – also dem einfachen Zitat einer Figurenäußerung, wie z. B. *Sie sagte: „Ich lese nur Literaturnobelpreisträger!"* – liegt eine klare Trennung zwischen Figurenstimme und Erzählerstimme vor. Andere Formen weisen, wie oben erwähnt, eine Vermischung der beiden Stimmen auf und sind teilweise auch grammatisch komplexer, wie die indirekte Wiedergabe, die Nebensatzkonstruktionen und häufig auch den Modus Konjunktiv aufweist. Man kann die Hypothese aufstellen, dass direkte Wiedergabe leichter verständlich und lesbar ist als andere Formen der Wiedergabe. Zugleich gelten Heftromane als weniger komplexe Literaturform. Tritt direkte Wiedergabe folglich auch häufiger in Heftromanen auf?

3. Material, Methode und Analyse

Wir planen eine quantitative Studie, d. h., unser Ziel ist es, die Anteile verschiedener Wiedergabeformen in Texten zu messen und zu vergleichen. Solche Studien werden überzeugender, je größer die Menge von Material ist, auf der sie beruhen. Gerade bei einem Phänomen wie Redewiedergabe muss man davon ausgehen, dass die Anteile von vielen Faktoren beeinflusst sind, z. B. vom Inhalt des Textes – eine Familiensaga hat vermutlich mehr Dialog als die Geschichte eines Schiffbrüchigen auf einer einsamen Insel – sodass es umso wichtiger ist, genug Daten zu haben, um solche Störfaktoren auszugleichen. Es ist jedoch sehr zeitaufwendig, Passagen von Redewiedergabe in Texten manuell zu markieren. Deswegen verwenden wir in der Studie ein Hilfsmittel: einen Redewiedergabe-Erkenner, d. h. ein Computerprogramm, das solche Passagen automatisch erkennt und auszeichnet.

Im Bereich der maschinellen Sprachverarbeitung werden zahlreiche solche Hilfsmittel entwickelt, die dabei helfen, große Mengen von Sprachdaten zu analysieren. Die Erkennung von Redewiedergabe ist dabei ein ziemlich spezieller Anwendungsfall. Weitverbreitet und in quantitativen linguistischen Untersuchungen häufig verwendet sind vor allem *Lemmatisierer*, die Wörter auf ihre Grundform zurückführen (z. B. *liest* zu *lesen*), *Tagger*, die Wörtern morphologische Kategorien zuweisen, und *Parser*, die syntaktische Analysen durchführen. Daneben gibt es verschiedenste weitere Werkzeuge, die z. B. Wörtern semantische Kategorien zuweisen, den ‚Gefühlsgehalt' eines Textes messen (*Sentiment Analysis*) oder automatisch den Referenten für Pronomen finden (→ Kapitel 14 [Annotation] in diesem Band).

Solche Werkzeuge können außerordentlich nützlich sein. Wenn Sie damit arbeiten, gibt es jedoch einiges zu beachten: Automatische

Werkzeuge liefern niemals 100% korrekte Ergebnisse und machen oft mehr und vor allem andere Arten von Fehlern als Menschen. Zudem ist immer der Kontext zu beachten, in dem sie entwickelt wurden. Dieser bestimmt das theoretische Modell, das ihnen zugrunde liegt (z.B. welche morphologischen Kategorien unterschieden werden und wie diese definiert sind), sowie auch die Art der Sprachdaten, für die sie entwickelt wurden. Eine Übertragung auf einen anderen Typ von Sprachdaten kann sehr problematisch sein – ein morphologischer Tagger etwa, der für moderne Zeitungssprache entwickelt wurde, wird deutlich schlechtere Ergebnisse liefern, wenn man ihn auf Chat-Daten oder Texte aus dem 18. Jahrhundert anwendet. Darum ist es sehr wichtig, das Werkzeug an den eigenen Daten zu testen und seine Funktionalität zu bewerten, bevor man sich dazu entschließt, es in einer Studie zu verwenden. Wir werden dies weiter unten für unseren Anwendungsfall tun.

Da automatische Werkzeuge immer für ein bestimmtes Kategoriensystem entwickelt werden, bedingt die Auswahl des Werkzeugs auch die Kategorien, mit denen später gearbeitet werden kann. Die vorgestellte Studie ist insofern ein Sonderfall, als sie im Kontext eines größeren Forschungsprojekts entstanden ist, dem ‚Redewiedergabe-Projekt' (www.redewiedergabe.de). In diesem Projekt haben wir die automatischen Erkenner selbst entwickelt und konnten darum bestimmen, welche Fälle sie unterscheiden. Falls Sie nicht selbst Programmierer*in sind (oder eng mit jemand entsprechendem zusammenarbeiten), müssen Sie Ihre Fragestellung angesichts der verfügbaren Klassifikationen ggf. anpassen oder Wege finden, wie Sie diese adaptieren können, etwa indem Sie Kategorien zusammenfassen oder verfeinern. Solche Schritte müssen dann natürlich dokumentiert und begründet werden. Auch diesen Aspekt werden wir anhand dieser Fallstudie illustrieren.

Zur Vorbereitung unserer Studie brauchen wir also folgende Vorüberlegungen:

1. Welches Korpusmaterial verwenden wir?
2. Welche Typen von Redewiedergabe unterscheiden wir und wie sind diese definiert?
3. Welches automatische Werkzeug verwenden wir und wie verlässlich sind die Ergebnisse, die dieses liefert?

Die Punkte 2 und 3 sind, aus den oben beschriebenen Gründen, eng miteinander verknüpft.

3.1 Korpusmaterial

Der geplante Vergleich beruht auf der Unterscheidung zwischen ‚Heftroman' und ‚Hochliteratur' und im zweiten Schritt auf der Unterscheidung zwischen verschiedenen Heftroman-Genres. Was die Heftromane angeht, ist die Definition dieser Kategorien zum Glück sehr einfach: Die Literaturform ist durch ihre Publikationsform klar erkennbar und die Genres durch explizit benannte Reihen repräsentiert und unterscheidbar. Die Gleichsetzung von Genre und Heftroman-Reihe ist allerdings natürlich eine Vereinfachung und sollte bei der späteren Interpretation der Ergebnisse im Auge behalten werden.

Schwieriger ist es, zu entscheiden, was ‚Hochliteratur' sein soll. Da die Bezeichnung vor allem eine gesellschaftliche Wertung widerspiegelt, haben wir uns dafür entschieden, solche Werke in unser Korpus aufzunehmen, die entweder einen Literaturpreis erhalten haben oder auf der Auswahlliste für einen standen – also Werke, die von einem literarisch geschulten Gremium als besonders herausragend bewertet wurden. Diese Definition hat den Vorteil, dass sie sehr gut operationalisierbar ist.

Die nächste Frage ist, was der Untersuchungszeitraum sein soll. Wir wollen keine diachrone Studie durchführen; um genug Material zu bekommen, müssen wir allerdings in Kauf nehmen, dass dieses eine gewisse zeitliche Streuung aufweist. Wir beschränken uns auf die Periode von der Nachkriegszeit bis in die jüngste Vergangenheit (1947-2016), also modernere Literatur.

Ein möglicher Störfaktor, den man vor allem bei literaturwissenschaftlichen Studien im Auge behalten muss, ist zudem die Verteilung der Autor*innen. Da unterschiedliche Autor*innen einen stark unterschiedlichen

Tab. 1: Korpuszusammensetzung

Gruppe	*Anzahl Texte*	*Unterschiedliche Autor*innen(-Pseudonyme)*	*Reihe*
Hochliteratur	50	50	-
Liebesroman	50	50	Julia Extra
Science-Fiction	50	11 (2-5 Texte pro Pseudonym)	Perry Rhodan
Horror	50	4 (2-35 Texte pro Pseudonym)	John Sinclair
Krimi	50	Unbekannt	Jerry Cotton

Schreibstil haben können, besteht die Gefahr, dass andere Faktoren – z. B. die Literaturform Heftroman oder Hochliteratur, die uns ja eigentlich interessiert – von diesem Autor*innensignal überlagert werden, wir also versehentlich die Besonderheiten des Schreibstils einer Person messen statt die einer Literaturform. Wir versuchen folglich, bei unserer Korpuszusammensetzung möglichst viele unterschiedliche Autor*innen miteinzubeziehen. Bei den Heftromanen ist die Autorschaftsfrage allerdings etwas schwierig, da üblicherweise unter Pseudonym gearbeitet wird und durchaus auch mehrere Personen das gleiche Pseudonym benutzen können (vgl. Hügel 2001). Wir versuchen trotzdem, in unserem Korpus im Rahmen unserer Möglichkeiten auch bei den Heftromanen eine möglichst große Diversität bei den Autor*innen(-Pseudonymen) zu erzielen. Tabelle 1 zeigt einen Überblick über das Korpus, mit dem die Studie durchgeführt wird.

Bei den Gruppen Hochliteratur und Liebesroman konnten wir unterschiedliche Autor*innen für jeden Text auswählen, bei Science-Fiction ist die Streuung mit zwei bis fünf Texten pro Pseudonym immerhin akzeptabel. Für die Gruppe Krimi war so gut wie keine Autor*inneninformation verfügbar – wir konnten nur vier der 50 Texten einigermaßen sicher Namen zuordnen und können auf dieser Grundlage davon ausgehen, dass mindestens drei verschiedene Personen beteiligt waren. Es ist jedoch bekannt, dass über die Zeit hinweg hundert und mehr Personen an der Krimi-Reihe „Jerry Cotton" mitgearbeitet haben (vgl. Karr 2010), sodass es höchstwahrscheinlich deutlich mehr sind. Die Gruppe Horror weist leider nur wenige unterschiedliche Autor*innen-Pseudonyme auf, von denen eines sehr dominant ist. 35 unserer Horrorromane wurden unter diesem Pseudonym verfasst.

3.2 Redewiedergabe-Kategorien

Wie oben bereits erwähnt, erwächst unsere Studie aus einem weiteren Kontext, dem ‚Redewiedergabe-Projekt', dessen Ziel es ist, ein Korpus mit Redewiedergabe-Formen zu annotieren und automatische Erkenner für diese Formen zu entwickeln. Das in diesem Projekt entwickelte und erprobte Annotationssystem beruht auf Kategoriensystemen aus der Literaturwissenschaft (vgl. hierzu auch Brunner 2015) und ist ausführlich dokumentiert (vgl. Brunner, Weimer, Engelberg, et al. 2020). Wie in der Einleitung erwähnt, werden die verschiedenen Redewiedergabetypen in narratologischen Darstellungen typischerweise auf einer Skala angeordnet. Die Reihenfolge in Tabelle 2 entspricht dieser Ordnung: Direkte Wiedergabe ist die unmittelbarste oder figurennächste Form, erzählte Wiedergabe die Form, die dem Erzählertext am nächsten steht.[2] Für die Zwecke dieser Studie ist es wichtig zu wissen, dass mit „Redewiederga-

2 Verschiedene narratologische Darstellungen unterscheiden sich sowohl in der genauen Definition der Pole als auch der Granularität der Unterteilung. Unser Kategoriensystem ist vor allem angelehnt an die Systeme von Genette (1998), Martínez und Scheffel (2016) und Leech und Short (2013), entspricht keinem davon jedoch vollkommen.

Tab. 2: Hauptkategorien des Annotationssystems im Projekt „Redewiedergabe"

Form	*Beispiel*	*Beschreibung*
direkt	Er dachte: „Ich habe Hunger."	Zitat der Rede, Gedanken oder schriftlichen Äußerung einer Figur; häufig mit Einleitungsformel und typographischer Markierung (z. B. Anführungszeichen).
frei indirekt	Er war ratlos. Wo sollte er jetzt etwas zu essen finden?	Auch bekannt als ‚erlebte Rede' (vgl. z. B. Martínez und Scheffel 2016). Form, v. a. von Gedankenwiedergabe, bei der sich Merkmale der Erzählerstimme (z. B. Tempus, Personalpronomen) und der Figurenstimme (z. B. Stil, Satzform) mischen.
indirekt	Sie sagte, dass sie ein gutes Restaurant kenne.	Integration der Figurenstimme in den Erzählertext mit Einleitungsformel und abhängiger Proposition.
erzählt	Die beiden unterhielten sich über Restaurants.	Weitere mehr oder minder ausführliche Darstellungen von Rede-, Gedanken- oder Schreibhandlungen im Erzählertext.

be" nicht nur die Wiedergabe von gesprochener Sprache, sondern auch die von Gedanken und geschriebenem Text gemeint ist. Tabelle 2 illustriert die vier Haupttypen von Wiedergabe im Kategoriensystem.

3.3 Redewiedergabe-Erkenner

Die automatischen Werkzeuge, die wir verwenden werden, wurden im Rahmen des Redewiedergabe-Projekts gezielt für die Erkennung der vier oben genannten Hauptkategorien entwickelt. Sie beruhen auf maschinellem Lernen, d. h., zunächst wurde eine große Menge von Texten von Menschen annotiert, die sogenannten Trainingsdaten. Auf diese Daten wurde ein Algorithmus angewendet, der aus dem Material Regeln ableitet (‚lernt'), um in neuen, unbekannten Texten dann ähnliche Fälle finden zu können. Die Details der Implementierung übersteigen den Rahmen dieses Kapitels,[3] was man allerdings zu maschinellem Lernen wissen sollte, ist, dass so entstandene Werkzeuge grundsätzlich stark von ihren Trainingsdaten beeinflusst sind – dies sind schließlich die Beispiele, aus denen sie lernen. Das bedeutet, dass man umso vorsichtiger sein muss, was die Anwendung eines solchen Werkzeugs auf Daten angeht, die stark unterschiedlich zu denen sind, auf denen sie trainiert wurden. Tatsächlich gibt es auch eine solche Diskrepanz in unserem Anwendungsfall: Die Erkenner wurden auf dem Redewiedergabe-Korpus (vgl. Brunner, Engelberg, Jannidis et al. 2020; verfügbar unter https://github.com/redewiedergabe/corpus) trainiert, welches sich aus Ausschnitten aus Zeitungen und Zeitschriften sowie Erzählungen des 19. bis frühen 20. Jahrhunderts zusammensetzt. Dies ist auch die Textsorte, für die die Erkenner im Redewiedergabe-Projekt primär entwickelt wurden. Die Anwendung auf moderne Texte, wie in dem Untersuchungskorpus für diese Studie, war nicht von Anfang an vorgesehen und es gibt folglich Unterschiede in Orthographie und Stil zwischen den Trainingsdaten und den Anwendungsdaten. Umso wichtiger ist es, die Tauglichkeit der Werkzeuge vor der Durchführung der eigentlichen Studie zu überprüfen. Wir führen darum eine Vorstudie durch, die zum einen der Bewertung der Erkenner dient, zum anderen auch erste Einblicke in unsere Forschungsfragen bietet.

3 Einen guten Überblick zu maschinellem Lernen bietet z. B. das Einstiegskapitel in Chollet (2018), das sich auch mit DeepLearning beschäftigt, einer Unterart des maschinellen Lernens, das bei unseren Erkennern verwendet wurde. Eine aktuelle Veröffentlichung zu unseren Erkennern ist Brunner, Tu, Weimer, et al. (2020).

3.4 Vorstudie

3.4.1 Bewertung der automatischen Methoden

Für die Vorstudie wurden aus 22 Hochliteratur-Texten und 22 Heftromanen[4] zufällige Textausschnitte mit einer Länge von ca. 1000 Tokens (d.h. 1000 Wörter und Satzzeichen) gezogen. Diese Ausschnitte sind lang genug, um einen Eindruck von der Textstruktur zu bekommen, aber gleichzeitig kurz genug, um das Material in vertretbarer Zeit manuell bearbeiten zu können. Die Heftroman-Ausschnitte wurden zur Hälfte aus dem Genre Liebesroman, zur Hälfte aus dem Genre Horror gewählt, um auch Genreunterschiede beobachten zu können. Diese Ausschnitte wurden entsprechend dem oben beschriebenen Annotationssystem (vgl. Brunner, Weimer, Engelberg, et al. 2020) von zwei Personen unabhängig voneinander annotiert. Anschließend verglich eine dritte Person diese beiden Annotationen, wählte bei Unstimmigkeiten jeweils die Variante aus, die unseren Annotationsrichtlinien besser entsprach, und verbesserte offensichtliche Fehler, sodass am Ende eine möglichst verlässliche Annotation, die sogenannte *Konsens-Annotation* vorlag.

Wenn man eine Studie durchführt, die auf komplexen manuellen Kategorisierungen beruht, ist stark anzuraten, mehr als eine Person das gleiche Material annotieren zu lassen. Dann kann man die Annotationen vergleichen (*inter-annotator agreement*) und bekommt so ein besseres Verständnis dafür, wie schwierig es ist, das Kategoriensystem auf echte Sprachdaten anzuwenden (vgl. z.B. Ide und Pustejovsky 2017). Wenn man noch dabei ist, das Kategoriensystem zu entwickeln, kann dies auch helfen, Schwachstellen in der Definition zu erkennen.[5] Da sprachliche Daten sehr lebendig sind, ist es jedoch durchaus erwartbar, dass bei einem komplexen Kategoriensystem nie hundertprozentige Einigkeit erzielt werden kann (vgl. Artstein 2017; zur Beziehung zwischen Annotation und Textinterpretation auch Gius und Jacke 2017).

Nachdem die Konsens-Annotation erstellt ist, wenden wir unsere automatischen Erkenner auf die gleichen Textausschnitte an und können nun die Übereinstimmung zwischen deren Ergebnissen und der Konsens-Annotation messen. Für jede der Wiedergabeformen gibt es einen speziell trainierten Erkenner und diese werden nacheinander angewendet. So ist es möglich, dass auch ineinander verschachtelte Wiedergaben unterschiedlichen Typs erkannt werden (z.B. eine indirekte Wiedergabe in einer direkten Wiedergabe: *„Er hat mir versprochen, zum Restaurant zu kommen", erklärte sie.*).

Um die Qualität der automatischen Erkennung zu messen, verwenden wir die Maße *F1-Score*, *Precision* und *Recall*. Dies sind verbreitete Maße, die ursprünglich aus dem Bereich des *Information Retrieval* (Entwicklung von Suchmaschinen) stammen, aber in der Informatik auch zur Bewertung von Kategorisierungsaufgaben verwendet werden. Da viele Wiedergabeformen keine ganzen Sätze umfassen, werden die Berechnungen auf Tokenbasis durchgeführt, d.h. es wird gemessen, wie vielen Tokens das richtige Label (z.B. ‚Teil von direkter Wiedergabe') zugewiesen wurde. Precision (Genauigkeit) bezeichnet den Anteil von automatisch annotierten Tokens, bei denen das Label mit der manuellen Annotation übereinstimmt, die also korrekt kategorisiert sind. Recall (Trefferquote) bezeichnet den Anteil von manuell annotierten Tokens, die vom automatischen Erkenner gefunden wurden. Precision und

4 Diese Test-Texte waren nicht Teil des Untersuchungskorpus, ähnelten den Untersuchungstexten jedoch stark. Es wäre zu bevorzugen gewesen, die Test-Texte nach dem Zufallsprinzip aus dem Untersuchungskorpus selbst zu ziehen, dieses war jedoch noch nicht fertig gestellt, als wir die Vorstudie durchführten.

5 Manuelle Annotation ist eine ausgesprochen zeitaufwendige Aufgabe und erfordert auch Einarbeitungszeit in das Annotationssystem von den Annotierenden. Oft ist es darum aus Zeit- und Kostengründen bei studentischen Projekten nicht möglich, eine zweite Person für die Vergleichsannotation zu rekrutieren, ganz zu schweigen von zwei weiteren Personen wie in unserem Szenario. Hier kann es helfen, zumindest die eigene Annotation in zeitlichem Abstand noch einmal zu wiederholen und dieses Ergebnis mit dem früheren Ergebnis zu vergleichen, um die eigene Konsistenz zu messen. Man spricht dann auch von *intra-annotator agreement*.

Tab. 3: Fiktive Beispielauswertung. In diesem Fall wäre die Precision für die Erkennung direkter Wiedergabe 0,6: Es wurden 5 Fälle als positiv vorhergesagt (Token 1 bis 5), davon waren 3 Vorhersagen korrekt. Der Recall wäre 1, also perfekt: Es gab 3 positive Fälle (Token 1, 2 und 3) und diese wurden alle gefunden. Der F1-Score, das harmonische Mittel aus Precision und Recall, wird berechnet mit der Formel: 2 * (Precision * Recall) / (Precision + Recall) und wäre damit 0,75

	Ist das Token Teil von direkter Wiedergabe?	
Token	*Manuelle Annotation*	*Automatische Vorhersage*
„	Ja	Ja
Hallo	Ja	Ja
"	Ja	Ja
,	Nein	Ja
sagte	Nein	Ja
er	Nein	Nein
.	Nein	Nein

Recall stehen immer in einem Spannungsverhältnis: Eine zu restriktive Erkennungsmethode übersieht Fälle und führt zu niedrigem Recall, eine zu großzügige findet zu viele Fälle und führt zu niedriger Precision. Der F1-Score ist ein Wert, in den Precision und Recall zu gleichen Teilen verrechnet werden, und kann als Maß für den Gesamterfolg interpretiert werden. Die drei Maße haben einen Wertebereich zwischen 0 (nichts gefunden oder nur Falsches gefunden) und 1 (perfekt). Tabelle 3 zeigt eine fiktive Beispielauswertung.

In unserer Studie interessieren wir uns vor allem für die prozentualen Anteile einer Wiedergabeform in unterschiedlichen Texten. Auch diese wird auf Tokenbasis berechnet, also z. B. ‚25% der Tokens im Textausschnitt sind Teil von direkter Wiedergabe'. Um einen Eindruck zu gewinnen, wie verlässlich solche Aussagen auf der Basis der automatischen Methoden sind, zeigt Tabelle 4 auch den *durchschnittlichen absoluten Fehler* für die Prozentanteile. Dieser wird ermittelt, indem für jeden der 44 Textausschnitte zunächst der *absolute Fehler* berechnet wird. Dies ist die Abweichung zwischen dem tatsächlichen Anteil und dem automatisch vorhergesagten Anteil. Wenn also laut manueller Annotation ein Textausschnitt 25% direkte Wiedergabe enthält, aber laut automatischer Annotation nur 20%, dann wäre der absolute Fehler 5%. Es wird dabei nicht beachtet, ob zu viel oder zu wenig gefunden wurde. Wenn die automatische Annotation also 30% direkte Wiedergabe vorhergesagt hätte, wäre der absolute Fehler ebenfalls 5%. Die absoluten Fehler für die 44 Textausschnitte werden addiert und der Durchschnitt wird gebildet. Dies vermittelt einen Eindruck, um wie viele Prozentpunkte die automatischen Methoden ungefähr danebenliegen. Es ist zu betonen, dass dieser Wert keine Auskunft darüber gibt, ob der Anteil über- oder unterschätzt wurde.

Tabelle 4 zeigt die Auswertung der Erkenner für die vier unterschiedlichen Wiedergabeformen. Dabei wurden die Ergebnisse der Erkenner mit der menschlichen Konsens-Annotation verglichen und die Werte F1-Score, Precision, Recall und durchschnittlicher absoluter Fehler berechnet. Zusätzlich sind für jede Wiedergabeform die gleichen Werte auch für die Übereinstimmung zwischen den beiden Erstannotierenden angegeben. Da diese beiden Personen unabhängig voneinander gearbeitet haben, erlaubt dies eine Abschätzung, wie verlässlich die Annotation wäre, wenn man sie durch Menschen

Tab. 4: Auswertung der automatischen Erkenner an den Testdaten (44 Textausschnitte zu je ca. 1000 Tokens). Zusätzlich angegeben sich die Übereinstimmungswerte zwischen den Erstannotierenden

	F1-Score	*Precision*	*Recall*	*durchschnittlicher absoluter Fehler*
direkt				
Konsens vs. Automatisch	0,83	0,78	0,89	7,84%
zwischen Erstannotierenden	*0,98*	*0,98*	*0,98*	*0,97%*
frei-indirekt				
Konsens vs. Automatisch	0,39	0,71	0,27	6,5%
zwischen Erstannotierenden	*0,69*	*0,64*	*0,73*	*3,58%*
indirekt				
Konsens vs. Automatisch	0,77	0,81	0,75	1,61%
zwischen Erstannotierenden	*0,81*	*0,84*	*0,77*	*1,68%*
erzählt				
Konsens vs. Automatisch	0,63	0,68	0,58	2,28%
zwischen Erstannotierenden	*0,67*	*0,66*	*0,68*	*1,85%*

erledigen ließe.[6] Wie man sieht, schwankt auch dieser Wert deutlich zwischen den einzelnen Wiedergabetypen.

Bei Betrachtung der Ergebnisse stellen wir fest, dass die Erkennung von *freier indirekter Wiedergabe* mit Abstand am schlechtesten funktioniert und die Werte auch deutlich schlechter sind als das, was menschliche Annotierende leisten. Dies ist insofern nicht verwunderlich, als freie indirekte Wiedergabe ein sehr schwer zu bestimmender Wiedergabetyp ist, der wenige klare und verlässliche Oberflächenmerkmale aufweist. Das Problem ist vor allem, dass viele Fälle vom Erkenner nicht gefunden werden (schlechter Recall) und damit die Anteile massiv unterschätzt werden. Angesichts dieser Ergebnisse haben wir beschlossen, dass es nicht sinnvoll ist, eine Studie darauf aufzubauen und zu riskieren, Fehlschlüsse zu ziehen, die auf Problemen unserer automatischen Erkennung beruhen. Die Anteile freier indirekter Wiedergabe werden darum in der Hauptstudie nicht untersucht.

Glücklicherweise sind die Ergebnisse für die anderen drei Wiedergabeformen erfreulicher. Bei *direkter Wiedergabe* mag es auf den ersten Blick verwundern, wieso der Erkenner nur einen F1-Score von 0,83 erreicht. Immerhin sind die Übereinstimmungsraten zwischen Menschen hier fast perfekt. Und könnte man direkte Wiedergabe nicht einfach finden, indem man Passagen in Anführungszeichen sucht? Leider ist es nicht so einfach. Direkte Rede ist keinesfalls immer mit Anführungszeichen markiert und selbst wenn sie es ist, gibt es unterschiedliche Typen von Anführungszeichen, die von Text zu Text variieren

6 Die Angabe von F1-Score, Precision und Recall für einen Annotierendenvergleich ist eher ungewöhnlich und wurde hier gewählt, damit man die Werte direkt mit den Ergebnissen für die automatische Annotation vergleichen kann. Üblicherweise verwendet man für den Annotierendenvergleich ein zufallskorrigiertes Maß namens *Kappa*, von dem es unterschiedliche Varianten gibt (siehe z. B. Artstein 2017). Wert 1 bedeutet bei Kappa-Maßen perfekte Übereinstimmung, Wert 0 bedeutet, dass die Übereinstimmung auf zufälligem Niveau liegt. Der Wert kann im Gegensatz zu F1-Score, Precision und Recall auch negativ werden, wenn die Übereinstimmung schlechter ist, als man es bei zufälliger Verteilung erwarten würde. Die Werte für Fleiss' Kappa in unseren Fall lauten: direkt = 0,97, frei-indirekt = 0,66, indirekt= 0,79, erzählt = 0,64 (also jeweils leicht niedriger als die F1-Scores).

und nicht immer konsistent gesetzt sind. Wenn man sich die Ergebnisse für diesen Wiedergabetyp im Detail ansieht, kann man feststellen, dass die Erfolgswerte von Ausschnitt zu Ausschnitt deutlich schwanken. In einigen Ausschnitten wird die direkte Wiedergabe nahezu perfekt erkannt, in anderen liegt die Vorhersage des prozentualen Anteils an direkter Wiedergabe an allen Tokens des Texts im Extremfall bis zu 20-30% daneben. Probleme treten insbesondere bei Ich-Perspektive in Kombination mit unmarkierter Wiedergabe auf, was v.a. in der Hochliteratur häufiger vorkommt. Solche Fälle sind schwierig für den Erkenner (und teilweise auch für Menschen), weil die direkte Wiedergabe sich dann kaum noch durch Oberflächenmerkmale vom Erzählertext unterscheiden lässt. Der Erkenner neigt in solchen Fällen dazu, den Anteil an direkter Wiedergabe zu überschätzen. Trotzdem sind die mit dem maschinellen Erkenner erzielten Ergebnisse deutlich stabiler als eine Identifikation von direkter Wiedergabe anhand von Anführungszeichen gewesen wäre. Diese hätte nämlich bei unmarkierter direkter Wiedergabe vollkommen versagt, während der Erkenner zumindest noch teilweise korrekte Ergebnisse liefert (siehe Abbildung 1). Wir sollten diese Problematik im Auge behalten; die Ergebnisse sind jedoch im Schnitt gut genug, dass man mit ihnen weiterarbeiten kann.

Es mag an dieser Stelle verwundern, wieso der durchschnittliche absolute Fehler bei freier indirekter Wiedergabe, die ja deutlich schlechtere Werte für F1-Score, Precision und Recall hat, geringer ist als für direkte Wiedergabe. Dies liegt daran, dass dieser Wert davon abhängt, wie hoch die Anteile für eine Wiedergabeform insgesamt sind. Während der durchschnittliche Anteil von direkter Wiedergabe laut Konsens-Annotation in den Textausschnitten bei fast 30% liegt, sind es nur ca. 10% für freie indirekte Wiedergabe. Ein Fehler von 6,5% für freie indirekte Wiedergabe ist damit deutlich gravierender als ein Fehler von 7,84% für direkte Wiedergabe.

Bei den Formen indirekte und erzählte Wiedergabe fällt auf, dass die F1-Scores zwar schlechter sind als für direkte Wiedergabe, jedoch nicht viel unter der Übereinstimmung liegen, die die Erstannotierenden erzielt haben. Diese Formen zu annotieren ist offensichtlich auch für Menschen nicht trivial. Dies ist darauf zurückzuführen, dass durch die stärkere Integration in den Erzähltext sowohl ihre genaue Abgrenzung als auch die Entscheidung, was als Wiedergabe zu werten ist, schwieriger wird. Betrachtet man auch hier die Ergebnisse für die einzelnen Textausschnitte, stellt man fest, dass die Schwankungen deutlich geringer sind als bei direkter Wiedergabe. Die Gesamtanteile dieser Formen am Text sind jedoch auch insgesamt geringer (durchschnittlich ca. 7% Anteil indirekt und 9% Anteil erzählt). Bei beiden Formen werden die Anteile an den Tokens des Ge-

Ist sie wach? fragt der Mann ein erneutes Mal, und der lächelnde Sohn nickt, das Lächeln verschwindet nicht aus seinem Gesicht. Gut. Dann hat die Mutterseele auch deine Nachricht an deinen Sohn gehört und wird sie weiterleiten. Sei ohne Sorge. Wie kann ich das wirklich wissen? Er wird dir im Traum erscheinen, sagt der Mann. Ich sehe ihn jede Nacht in meinem Traum, sagt Fatma Hanim, deswegen bin ich nicht hergekommen. Er wird dir im Traum ein Zeichen geben, daß er verstanden hat.

Abb. 1: Ein Beispiel für die Qualität der Erkennung von unmarkierter direkter Wiedergabe im Testkorpus. Die unterstrichenen Textteile wurden von Menschen als direkte Wiedergabe identifiziert, die grau hinterlegten vom automatischen Erkenner. Wie man sieht, ist die Erkennung nicht sehr genau – der Erkenner setzt zu spät ein, hört zu früh auf und erfasst auch Rahmenformeln fälschlicherweise mit. Nur auf Basis von Anführungszeichen wäre in diesem Textausschnitt allerdings überhaupt keine direkte Wiedergabe identifiziert worden, was ein deutlich schlechteres Ergebnis wäre.

Tab. 5: Auswertung für die Erkennung der neu definierten Kategorie ‚nicht-direkt'. Als ‚nicht-direkt' zählt jedes Token, das entweder als Teil von indirekter oder als Teil von erzählter Wiedergabe markiert wurde

	F1-Score	*Precision*	*Recall*	*durchschnittlicher absoluter Fehler*
nicht-direkt				
Konsens vs. Automatisch	0,70	0,74	0,67	2,96%
zwischen Erstannotierenden	*0,78*	*0,78*	*0,78*	*2,28%*

samttexts von den Erkennern eher unter- als überschätzt.

An dieser Stelle überdenken und schärfen wir noch einmal das Design unserer Studie. Nachdem wir freie indirekte Wiedergabe aktuell nicht verlässlich mit automatischen Methoden identifizieren können, bietet sich aus literaturwissenschaftlicher Perspektive eine Gegenüberstellung von direkter und nicht-direkter Wiedergabe an, eine Dichotomie, die schon in der klassischen Rhetorik aufgemacht wurde (vgl. McHale 2014). Hier wird unterschieden zwischen dem Zitat einer Figurenstimme und deren Integration in den Erzählertext. Dies bedeutet, dass wir die Formen indirekte und erzählte Wiedergabe zu einer neuen Kategorie ‚nicht-direkt' zusammenfassen – wir leiten also aus den durch den Erkenner vorgegebenen Kategorien eine neue ab, die besser zu den Zielen unserer Studie passt. Diese umfasst alle Textpassagen, die entweder als indirekte oder als erzählte Wiedergabe ausgezeichnet wurden. Diese Zusammenfassung ist theoretisch auch insofern zu rechtfertigen, als die Grenze zwischen indirekter und erzählter Wiedergabe in der Literatur häufig unterschiedlich gezogen wird. Zur Erinnerung: ‚nicht-direkt' umfasst damit nun diejenigen Fälle, in denen Rede, Gedanken oder schriftliche Zeugnisse nicht zitiert, sondern mehr oder minder ausführlich in den Erzählertext integriert werden, d.h. sowohl die klassische indirekt-Konstruktion mit Einleitungsformel und abhängigen Nebensatz (*Sie sagte, dass sie ein gutes Restaurant kenne.*) als auch formal abweichende und häufig stärker zusammenfassende Wiedergaben (*Die beiden unterhielten sich über Restaurants.*). Da wir außerdem Gedankenwiedergabe mit untersuchen, sind z.B. auch folgende Fälle erfasst: *Er wusste nicht, was er wollte. Sein Grübeln über Restaurants war nervtötend.* Tabelle 5 zeigt die Auswertung der automatischen Erkennung für diese neue Kategorie.

Wie zu erwarten war, liegen die Messwerte für nicht-direkt etwa in der Mitte zwischen indirekt und erzählt. Der durchschnittliche absolute Fehler ist größer, allerdings werden auch die Anteile selbst durch die Zusammenfassung größer: Der Anteil von nicht-direkt in den Testdaten liegt bei durchschnittlich ca. 15%. In Relation dazu sind die Schwankungen ungefähr vergleichbar mit denen, die wir beobachten konnten, als wir die beiden Formen indirekte und erzählte Wiedergabe einzeln untersucht haben.

Wir können nun unsere Forschungsfragen noch einmal stärker operationalisieren und als Hypothesen formulieren:

1. Es gibt signifikante Unterschiede im prozentualen Anteil direkter Wiedergabe und nicht-direkter Wiedergabe
   a. zwischen den beiden Gruppen Heftromane und Hochliteratur.
   b. zwischen den einzelnen Genres von Heftromanen untereinander und zur Hochliteratur.

   Nebenhypothese: Der Anteil direkter Wiedergabe ist höher in Heftromanen.
2. Für die Wirkung eines Textes ist es auch relevant, wie das Verhältnis von direkter und nicht-direkter Wiedergabe ist: Sind beide Anteile niedrig, kann das auch einfach bedeuten, dass der Text insgesamt wenig Wiedergabe verwendet, z.B. aus inhaltlichen Gründen. Wie wir gesehen haben, sind die durchschnittlichen Anteile für die beiden Wiedergabetypen zudem unterschiedlich (ca. 30% direkt, ca. 15%

nicht-direkt). Es kann jedoch sein, dass sich in bestimmten Texten dieses Verhältnis verschiebt und damit einer der Formen ein besonderes Gewicht gegeben wird. Unsere weiterführende Hypothese ist, dass sich, wenn man die Anteile direkter und nicht-direkter Wiedergabe in den Texten in Kombination betrachtet, ebenfalls Unterschiede zwischen den untersuchten Gruppen feststellen lassen.

3.4.2 Verhalten von Redewiedergabe in den manuell annotierten Textausschnitten

Bevor wir unser Untersuchungsmaterial für die eigentliche Studie erweitern, nutzen wir die manuell annotierten Daten, um unsere Hypothesen zunächst daran zu überprüfen.

Abbildungen 2 und 3 zeigen Boxplots für die Anteile von direkter und nicht-direkter Wiedergabe in den 44 konsens-annotierten Textausschnitten. Die Aufteilung der Daten erfolgt einmal nach den übergreifenden Gruppen Heftromane und Hochliteratur und einmal mit einer zusätzlichen Aufspaltung der Heftromane nach Genres.

Ein Boxplot ist ein verbreiteter Typ von Datenvisualisierung, der einen raschen Eindruck vermittelt, in welchem Bereich Daten liegen und wie sie sich verteilen (vgl. → Kapitel 21 [Aufbereitung Untersuchungsergebnisse] in diesem Band). Auf der senkrechten Achse sind die Anteile von direkter bzw. nicht-direkter Wiedergabe auf einer Skala von 0-100% abzulesen. Der Boxplot selbst besteht aus einem Kasten (der ‚Box') und zwei davon ausgehenden Strichen, den Antennen oder ‚Whiskers'. Die Box umschließt den Bereich, in dem 50% der Daten liegen. Der Strich, der die Box teilt, markiert den Median, also den Wert, der genau in der Mitte liegt, wenn man alle Werte der Größe nach sortiert. Die Antennen markieren die weitere Streuung von Datenpunkten. Wenn zusätzlich noch Punkte über oder unter den Antennen zu sehen sind, sind dies Ausreißer, d.h. Werte, die im Vergleich zu den anderen Werten ungewöhnlich hoch oder niedrig sind. Betrachten wir als Beispiel in Abbildung 2 den Boxplot für den Anteil von nicht-direkt in Heftromanen: Der Median ist bei ca. 11% und die Werte streuen sehr wenig. Die Hälfte liegt zwischen ca. 10% und 12% (innerhalb der Box) und fast alle Werte liegen zwischen ca. 7% und 17% (innerhalb der Antennen). Es gibt allerdings vier Ausreißer, also Textausschnitte, die ungewöhnlich hohe Werte aufweisen (Maximum fast 27%). Diese werden repräsentiert durch die vier Punkte über der oberen Antenne. Insgesamt ergibt sich hier das Bild einer gleichförmigen Verteilung mit ein paar Abweichungen. Das andere Extrem ist der Boxplot für den Anteil von direkt bei Hochliteratur, wo man sieht, dass sowohl die Box selbst als auch die Antennen sich über fast die gesamte Skala ausbreiten.

Bei der Betrachtung der Boxplots in Abbildung 2 fällt auf, dass sich die Median-Werte (die Striche in der Mitte der Boxen) bei den Anteilen direkter Wiedergabe sehr ähnlich verhalten, auch wenn, wie oben bereits angemerkt, Hochliteratur deutlich mehr Streuung aufweist. Die Anteile von nicht-direkter Wiedergabe scheinen bei Hochliteratur etwas höher zu sein. Wenn man sich die Aufspaltung nach Genres in Abbildung 3 ansieht, sieht man, dass es vor allem die Horrorroman-Ausschnitte sind, die recht niedrige Anteile von nicht-direkter Wiedergabe (und eine extrem geringe Streuung dieser Anteile) aufweisen.

Um die Signifikanz dieser Beobachtungen zu überprüfen, führen wir einen Permutationstest durch, einen statistischen Test, der sich besonders gut für Korpusdaten eignet (vgl. Koplenig 2019; eine anschauliche Erklärung findet sich auch in Tu, Engelberg und Weimer 2019). Die Grundidee dabei ist folgende: Zunächst wird der durchschnittliche Unterschied eines Wertes (z.B. Anteil direkt) zwischen zwei Gruppen von Daten ermittelt (z.B. Heftromane vs. Hochliteratur). Dann durchmischt man die zugrundeliegenden Daten (also in unserem Fall die Werte von Anteil direkt für alle 44 Textausschnitte), teilt sie zufällig in neue zwei Gruppen auf und misst den Unterschied zwischen diesen Zufallsgruppen. Ist dieser größer als der Unterschied zwischen den Gruppen Heftroman vs. Hochliteratur, so ist das ein Indiz, dass der zwischen Heftroman und Hochliteratur gemessene Unterschied nur zufällig zustande

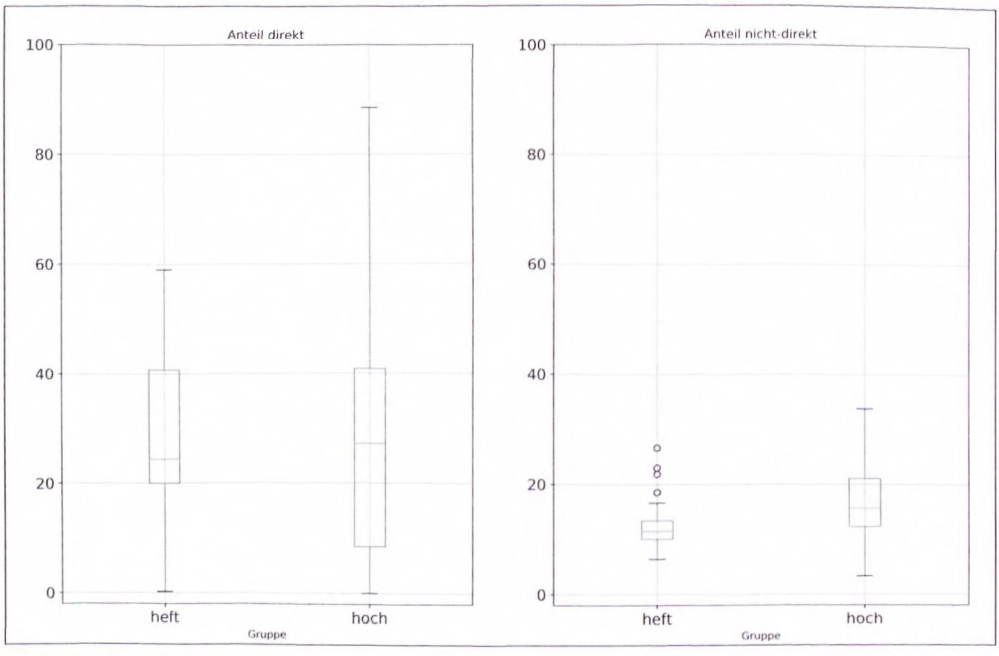

Abb. 2: Konsens-annotierte Textausschnitte: Gegenüberstellung der Anteile von direkter und nicht-direkter Wiedergabe in Heftromanen vs. Hochliteratur

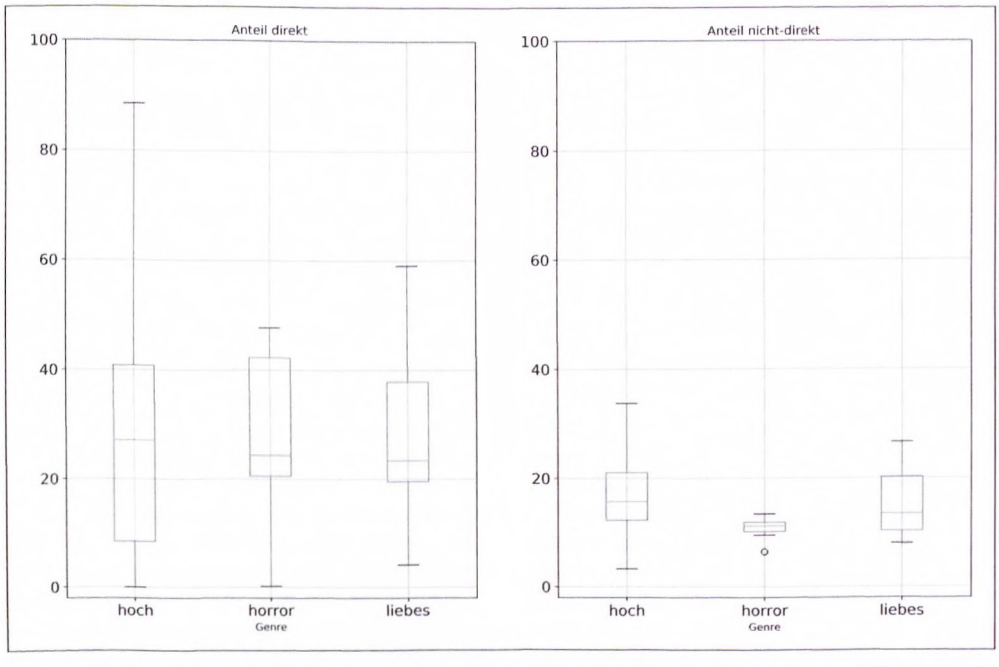

Abb. 3: Konsens-annotierte Textausschnitte: Gegenüberstellung der Anteile von direkter und nicht-direkter Wiedergabe in Hochliteratur vs. Heftroman-Genres Liebesromane und Horror

gekommen ist. Das Zufallsgruppen-Experiment wird mehrfach wiederholt, in unserem Fall 10.000 Mal. Wenn innerhalb dieser vielen Wiederholungen in weniger als 1% der Fälle (also in maximal 99 Fällen) größere Unterschiede herauskamen, kann man davon sprechen, dass der zwischen den echten Gruppen gemessene Unterschied signifikant ist auf einem Niveau von p=0,01. Es ist also sehr unwahrscheinlich, dass er zufällig zustande kam.[7]

Mit dem Permutationstest lassen sich in den konsens-annotierten Testdaten im Vergleich Heftroman vs. Hochliteratur allerdings weder für Anteil direkt noch für Anteil nicht-direkt signifikante Unterschiede nachweisen. Wenn man die Unterscheidung nach Genres mit einbezieht und drei Vergleiche durchführt (hoch vs. horror; hoch vs. liebes; liebes vs. horror) sind lediglich die Abweichungen zwischen Hochliteratur und Genre Horror im Anteil nicht-direkter Wiedergabe signifikant. Rechnet man den Permutationstest auf Basis der automatisch annotierten Daten statt auf den konsens-annotierten Daten, verschwindet auch diese Signifikanz. Letzteres ist insofern kein schlechter Befund, als dass die Ungenauigkeit, die mit der Verwendung automatischer Annotation einhergeht, hier zwar mögliche Unterschiede verwischt, aber immerhin nicht künstlich erzeugt, was deutlich problematischer wäre.

Die beiden Dimensionen interessieren uns jedoch nicht nur in Isolation, sondern wir wollen etwas über die Erzählweise erfahren. Diese wird, wie oben beschrieben, auch vom Zusammenspiel der beiden Wiedergabe-Arten innerhalb eines Textes beeinflusst. Eine

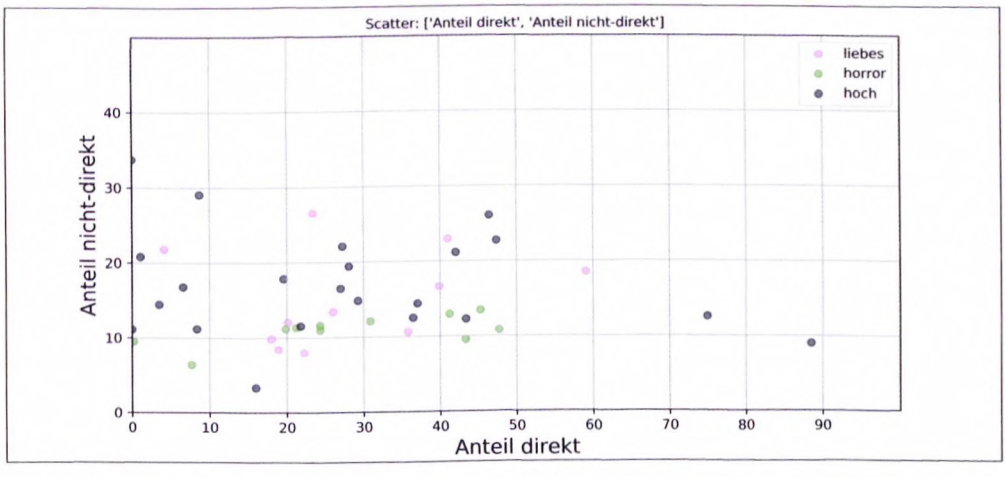

Abb. 4: Konsens-annotierte Textausschnitte: Scatterplot mit den Dimensionen Anteil direkt/Anteil nicht-direkt. Jeder Punkt repräsentiert einen Textausschnitt

7 Das Signifikanzniveau von 0,01 ist ziemlich streng gewählt. Häufig wird auch ein Signifikanzniveau von 0,05, also 5%, als akzeptabel betrachtet. Ein Grund, weshalb wir dieses strenge Niveau gewählt haben, ist, dass wir auf unseren Daten mehrere Vergleiche durchführen, nämlich nicht nur zwischen Heftromanen und Hochliteratur, sondern auch zwischen verschiedenen Genres (hoch vs. liebes, hoch vs. horror, liebes vs. horror). In der Hauptstudie werden wir noch zwei zusätzliche Genres betrachten und führen entsprechend mehr Vergleiche durch. Man geht davon aus, dass mehrere Vergleiche über die gleiche Datenmenge, auch wenn deren Aufteilung wechselt, die Gefahr von Zufallseffekten vergrößert. Deswegen ist es sinnvoll, in solchen Studien ein strengeres Signifikanzniveau anzunehmen. Eine in der Statistik häufig angewandte Strategie, mit diesem Problem umzugehen, ist auch die sogenannte Bonferroni-Korrektur (https://de.wikipedia.org/wiki/Bonferroni-Korrektur).

gute Möglichkeit, einen solchen Zusammenhang zwischen zwei Dimensionen darzustellen, ist ein Scatterplot (Abbildung 4). Bei dieser Graphik repräsentiert die horizontale X-Achse den Anteil an direkter Wiedergabe, die vertikale Y-Achse den Anteil an nicht-direkter Wiedergabe. Die 44 Textausschnitte werden in diesem Koordinatensystem als Punkte repräsentiert und die drei Gruppen farblich gekennzeichnet.

In dem Scatterplot in Abbildung 4, der auf Basis der konsens-annotierten Daten erstellt wurde, lässt sich eine Tendenz der Horror-Textausschnitte erkennen, sich mit niedrigen Werten für nicht-direkt zusammen zu gruppieren, während die Liebesroman- und vor allem die Hochliteratur-Ausschnitte gestreuter erscheinen. Dies ist ein interessanter Ersteindruck, man sollte jedoch vorsichtig sein, daraus weiterreichende Schlüsse zu ziehen, vor allem angesichts der Tatsache, dass die Permutationstests kaum signifikante Ergebnisse ergeben haben. Scatterplots sind intuitive Datenvisualisierungen, die allerdings leicht zu Überinterpretation verführen. Sie werden aussagekräftiger und verlässlicher, je mehr Datenpunkte sie enthalten. Es ist aber in jedem Fall gut, sie nicht allein zu verwenden, sondern zusätzlich Signifikanztests wie den Permutationstest durchzuführen.

Zusammenfassend stellen wir fest, dass die Vorstudie noch keine unserer Hypothesen bestätigt hat. Immerhin konnten wir jedoch nachweisen, dass die automatische Annotation nicht grob andere Ergebnisse ergibt als die manuelle, was sehr wichtig ist, da wir uns im nächsten Schritt auf diese verlassen.

3.5 Hauptstudie mit automatisch annotiertem Material

Für die Hauptstudie erweitern wir unser empirisches Material sehr stark, indem wir das unter „Korpusmaterial" (Tabelle 1) beschriebene Korpus automatisch annotieren. Es kommen nun auch zwei neue Genres von Heftromanen hinzu, die in der Vorstudie nicht untersucht wurden: Krimis und Science-Fiction. Statt 44 Ausschnitten von 1000 Tokens, betrachten wir nun insgesamt 250 vollständige Erzähltexte, deren Textlänge zwischen ca. 20.000 und 330.000 Tokens liegt.

Die großen Unterschiede in der Textlänge treten vor allem bei den Hochliteratur-Texten auf, die eine Spanne von 35.000 und 330.000 Tokens aufweisen, während die Heftromane nur zwischen 20.000 und 50.000 Tokens variieren. Um sicher zu gehen, dass diese Textlängen-Varianz nicht unsere Ergebnisse verfälscht, führen wir eine Normierung durch: Wir zerlegen jeden Volltext in 1000-Token-Abschnitte und berechnen die Anteile von direkt und nicht-direkt für jeden Abschnitt. Anschließend bilden wir die Durchschnittswerte über alle 1000-Token-Abschnitte des Textes.[8] Diese Werte liegen den folgenden Auswertungen zugrunde.

Wir führen nun mit dem erweiterten, automatisch annotierten Material die gleichen Tests durch wie zuvor mit den Textausschnitten. Betrachten wir zunächst die Boxplots (Abbildungen 5 und 6).

Anders als bei den Ausschnitten zeigen sich nun klare Unterschiede in beiden Dimensionen: Der Anteil direkter Wiedergabe ist bei Hochliteratur geringer, während der Anteil nicht-direkter Wiedergabe höher ist (Abbildung 5). Die Signifikanz beider Unterschiede lässt sich mit dem Permutationstest bestätigen.

Auch zwischen Hochliteratur und den einzelnen Heftroman-Genres lassen sich signifikante Unterschiede für viele der Vergleichspaare bestätigen (Abbildung 6): hoch/krimi, hoch/horror, horror/krimi und horror/liebes sind signifikant unterschiedlich in beiden Dimensionen; hoch/liebes, scifi/liebes und scifi/krimi sind signifikant unterschiedlich nur in Anteil direkt; horror/scifi sind signifikant unterschiedlich nur in Anteil nicht-direkt.

8 Dieses Vorgehen ist analog zu einem in der Korpuslinguistik verbreiteten Maß, der *Standardisierten Type-Token-Ratio* (STTR). Eine Beschreibung dieses Maßes findet sich z. B. in den frei verfügbaren Ergänzungsmaterialien zu dem Einführungsbuch „Korpuslinguistik" (Perkuhn, Keibel und Kupietz 2012), verfügbar unter http://corpora.ids-mannheim.de/libac/doc/libac-addOn-LexikalVielfalt.pdf.

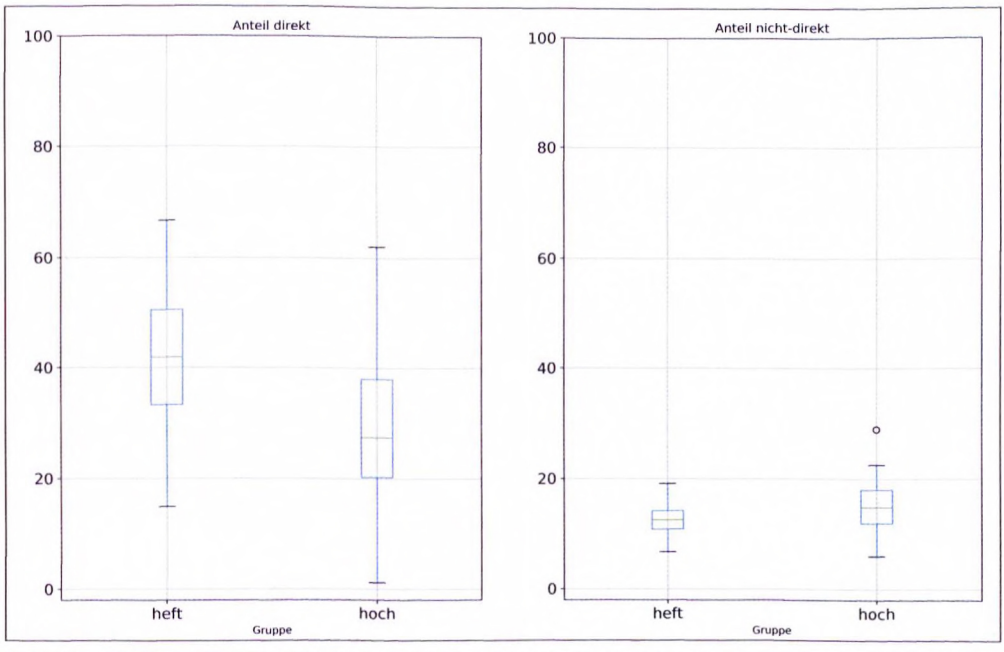

Abb. 5: Automatisch annotierte Volltexte: Gegenüberstellung der Anteile von direkter und nicht-direkter Wiedergabe in Heftromanen vs. Hochliteratur

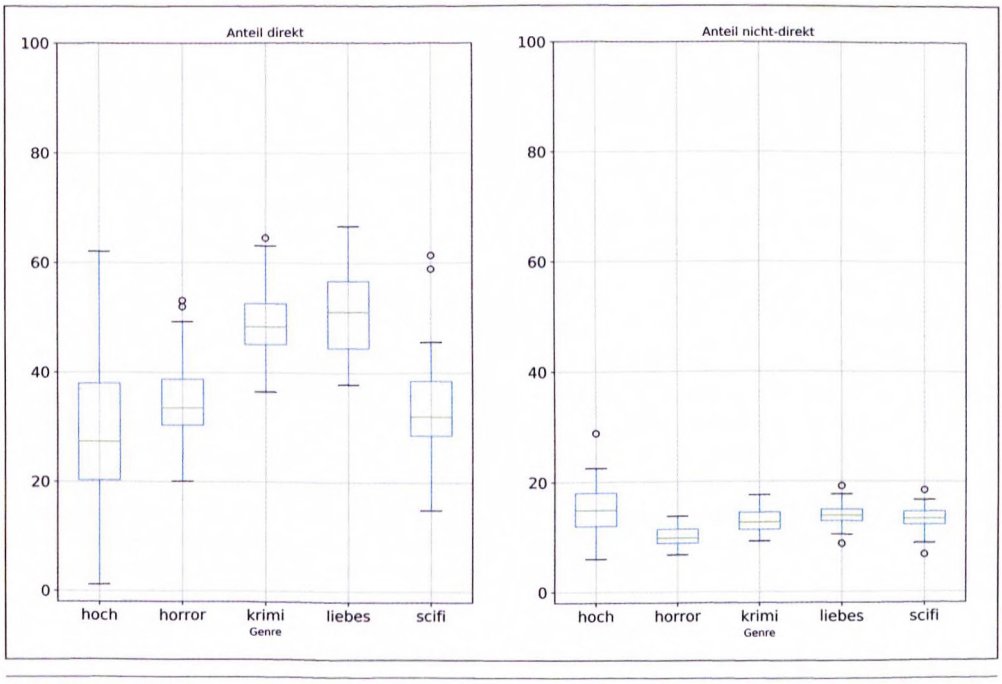

Abb. 6: Automatisch annotierte Volltexte: Gegenüberstellung der Anteile von direkter und nicht-direkter Wiedergabe in Hochliteratur vs. Heftroman-Genres Horror, Krimi, Liebesroman und Science-Fiction

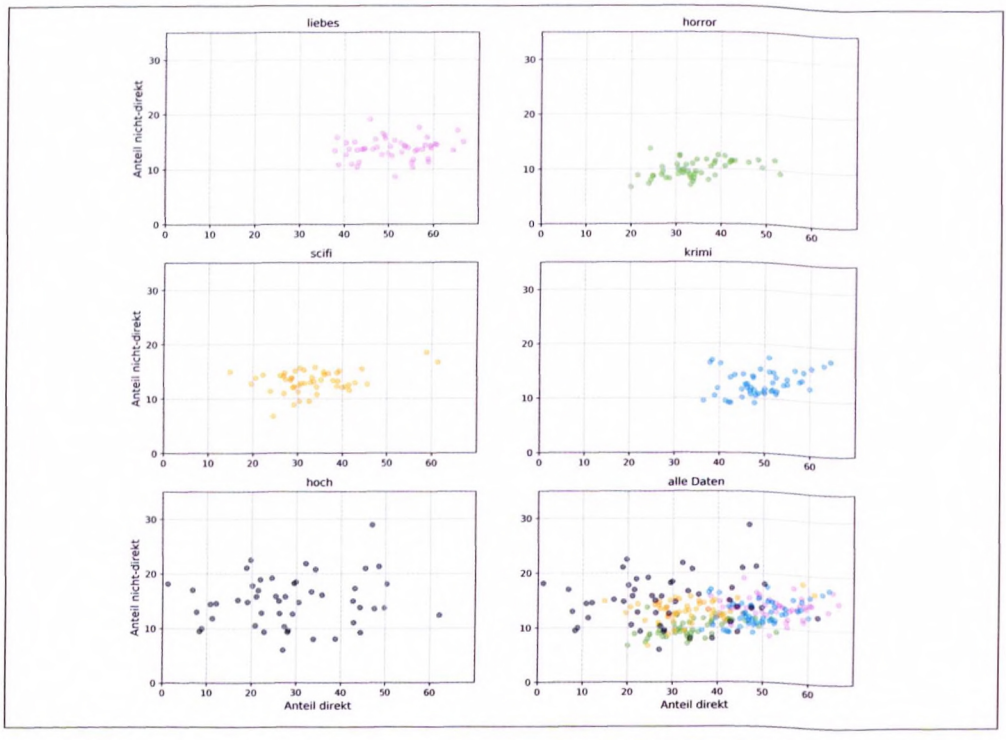

Abb. 7: Scatterplots für die automatisch annotierten Volltexte des Untersuchungskorpus. Jeder Datenpunkt repräsentiert einen Volltext

Abbildung 7 zeigt Scatterplots für die Daten der Hauptstudie. Zur besseren Übersichtlichkeit wird zunächst die Verteilung der Datenpunkte jeder Gruppe einzeln visualisiert. Der letzte Plot zeigt die Überlagerung der fünf Gruppen, also die Datenpunkte für alle 250 Volltexte unseres Untersuchungskorpus.

Es fällt sofort auf, dass die Hochliteratur-Texte eine deutliche Streuung aufweisen, während sich nicht nur die einzelnen Genres, sondern auch die Heftromane als Gruppe zusammen gruppieren. Die Hochliteratur-Texte sind zudem die einzige Gruppe, in der sogar ein ‚Übergewicht' an nicht-direkter im Vergleich zu direkter Wiedergabe auftritt. Man kann dies so interpretieren, dass die Autor*innen in der Art und Weise, wie sie Figurenstimmen in den Text einbinden, sowohl individualistischer sind als auch eher bereit, nicht das direkte Zitat zu wählen.

Innerhalb der Gruppe der Heftromane kann man für die Genres Liebesroman, Horrorroman und Krimi beobachten, dass sie sich recht deutlich entlang einer gedachten Linie gruppieren, die von links nach rechts flach ansteigt (besonders gut sichtbar ist dies in der letzten Graphik). Die Anteile von direkter und nicht-direkter Wiedergabe in diesen Texten steigen in Relation zueinander fast gleichmäßig an, wobei der Anteil direkter Wiedergabe stets höher ist. Es differenziert sich recht klar das Horror-Genre mit einem insgesamt geringeren Wiedergabeanteil, während sich die ‚kommunikativeren' Genres Liebesroman und Krimi stark überlagern. Zwischen diesen beiden Genres lassen sich auch mit dem Permutationstest keine signifikanten Unterschiede nachweisen. Science-Fiction nimmt eine Zwischenstellung ein: Die Texte sind diverser und streuen ähnlich wie Hochliteratur, wenn auch nicht so extrem. Dies ist das

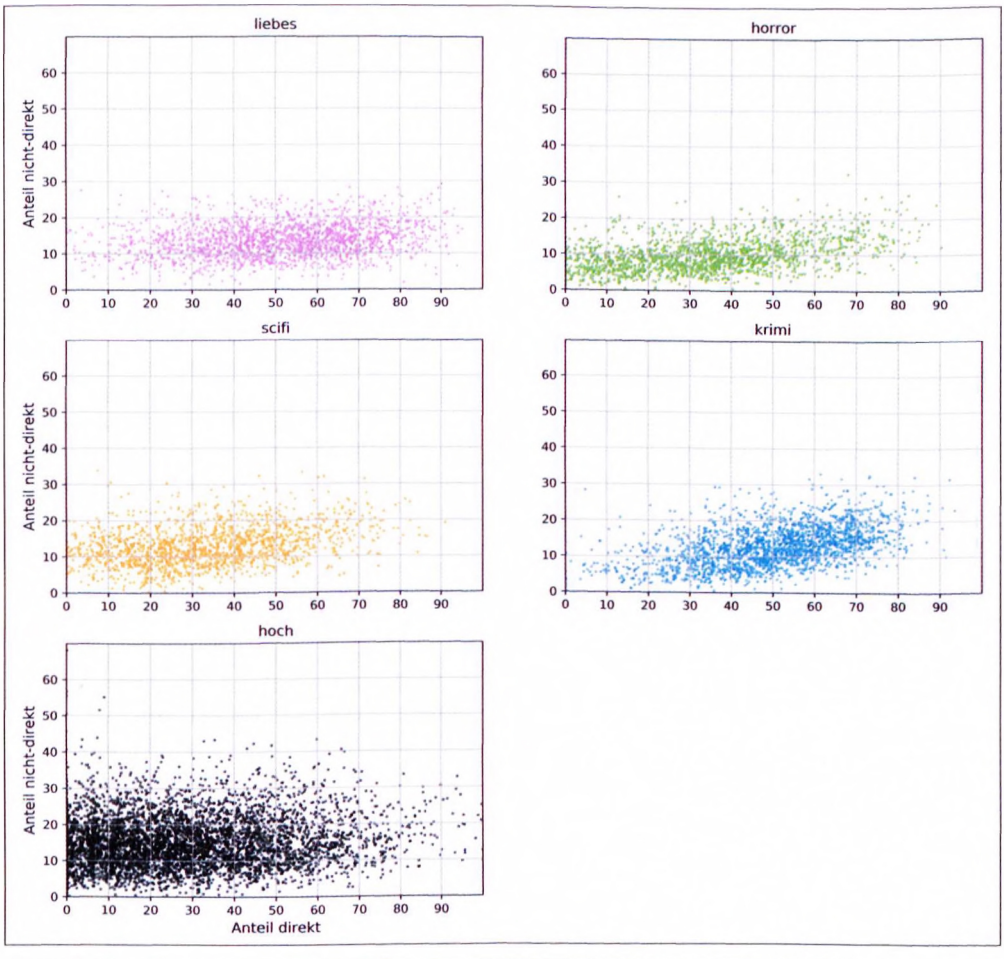

Abb. 8: Scatterplots für die automatisch annotierten Volltexte des Untersuchungskorpus, zerlegt in 1000-Token-Abschnitte. Jeder Datenpunkt repräsentiert einen 1000-Token-Abschnitt

einzige Heftroman-Genre, für das sich auf keiner der beiden Dimensionen signifikante Unterschiede zu Hochliteratur nachweisen lassen.

Warum sehen wir diese Muster – insbesondere im Scatterplot – erst in der Betrachtung der Volltexte und nicht in der Vorstudie mit den Textausschnitten? Dies hängt damit zusammen, dass die Schwankungen in den Anteilen von Wiedergabe innerhalb eines Erzähltextes sehr stark sind. Dies ist auch wenig verwunderlich, wenn man sich den Verlauf einer Geschichte vorstellt: Oft wechseln z. B. dialoglastige Passagen mit beschreibenden Passagen ab. Wie oben erwähnt haben wir für unsere Analysen die Volltexte in 1000-Token-Abschnitte zerlegt und anschließend den Durchschnittswert gebildet. Abbildung 8 zeigt nun einen Datenpunkt für jeden einzelnen dieser 1000-Token-Abschnitte. Man beachte, dass die Achsen der Plots im Gegensatz zu Abbildung 7 erweitert wurden, um die extremeren Werte, die in den kurzen Abschnitten auftreten, abbilden zu können. Auf den letzten Plot, die Überlagerung aller Datenpunkte, wurde hier verzichtet, da er nicht mehr lesbar gewesen wäre. Zwar kann man in der Gesamtheit dieser Datenpunkte die

gleichen Trends erkennen wie für die Volltexte in Abbildung 7, doch wenn man – wie bei der Vorstudie – nur wenige zufällig gezogene 1000-Token-Abschnitte aus jeder Gruppe betrachtet, ist die Wahrscheinlichkeit sehr hoch, dass die Schwankungen innerhalb der Gruppe die Gruppenunterschiede verschleiern. Man sieht, wie problematisch es ist, Aussagen über die tatsächlichen Anteile von Wiedergabe in einem Text zu treffen, wenn man ihn nicht vollständig betrachtet. Die Ausweitung auf mehr Material, die durch die Anwendung automatischer Methoden möglich wurde, führt hier also zu einem Erkenntnisgewinn, der sonst nur mit extremem Annotationsaufwand möglich gewesen wäre.

4. Ergebnisse und Diskussion

Zusammenfassend lässt sich sagen, dass sich in der Hauptstudie die anfangs aufgestellten Hypothesen für unser Untersuchungskorpus fast durchgehend bestätigen lassen: Es ergeben sich tatsächlich statistisch signifikante Unterschiede in den Anteilen sowohl von direkter als auch von nicht-direkter Wiedergabe zwischen Hochliteratur und Heftromanen. Auch die sekundäre Hypothese, dass direkte Wiedergabe in Heftromanen anteilig häufiger ist, ließ sich bestätigen.

Bei der genaueren Betrachtung der unterschiedlichen Heftroman-Genres konnten wir feststellen, dass diese sich unterschiedlich stark voneinander unterscheiden. V. a. die Genres Liebesroman und Krimi weisen starke Überschneidungen auf: Zwischen ihnen gibt es in keiner der beiden Dimensionen eine signifikante Abweichung. Insgesamt wirken die Heftromane, gerade in der Betrachtung in zwei Dimensionen, wie eine recht geschlossene Gruppe, die sich gegenüber der Hochliteratur durch eine geringere Streuung auszeichnet und in der die Anteile direkter Wiedergabe stets höher sind als die Anteile nicht-direkter Wiedergabe. Angesichts der Tatsache, dass die Heftroman-Genres bewusst reglementierte Reihen sind, während die Hochliteratur-Gruppe nur dadurch definiert ist, dass die enthaltenen Werke als literarisch hochwertig eingeschätzt wurden, mag es nicht erstaunlich scheinen, dass Hochliteratur sich diverser verhält. Es ist jedoch durchaus bemerkenswert, dass sich der Unterschied zwischen konventionalisiertem und individualistischem Erzählen auf der Dimension der verwendeten Redewiedergabeformen so deutlich quantitativ nachweisen lässt.

Eine Sonderstellung nimmt das Genre Science-Fiction ein, das eine stärkere Streuung aufweist als die anderen Heftroman-Genres und für das sich keine signifikanten Unterschiede zu Hochliteratur nachweisen lassen. Dieser Befund passt zu Beobachtungen von Jannidis, Konle und Leinen (2019), die ebenfalls Heftromane und Hochliteratur mit verschiedenen quantitativen Maßen untersuchen. Auch sie stellen fest, dass Science-Fiction sich ‚hochliteratur-ähnlicher' verhält als die anderen Heftroman-Genres und höhere Werte bei Maßen wie Wortlänge und Standardisierter Type-Token-Ratio aufweist, die mit größerer Komplexität des Erzählens in Verbindung gebracht werden können (vgl. Jannidis, Konle und Leinen 2019).

Bedauerlicherweise ist es uns nicht möglich, die annotierten Daten zur Verfügung zu stellen, da diese dem Urheberrecht unterliegen (vgl. → Kapitel 15 [Juristische Fragen] in diesem Band). Aus Gründen des Copyrights ist es grundsätzlich nicht einfach, moderne Texte in digitaler Form zu bekommen und mit ihnen zu arbeiten. Die Texte, auf denen unsere Studie beruht, wurden gekauft, per Hand gescannt und anschließend aufbereitet. Wir dürfen sie nicht veröffentlichen und weitergeben, sodass die Rohdaten unserer Studie nicht verfügbar gemacht werden können. Auf unserer Github-Seite (https://github.com/redewiedergabe/tagger) sind jedoch aktuelle Versionen unserer automatischen Erkenner zum freien Download verfügbar. Diese sind eine Weiterentwicklung der Werkzeuge, die in dieser Studie verwendet wurden (vgl. Brunner, Tu, Weimer et al. 2020).

5. Methodische Reflexion

Bei einer Studie, die sich auf automatische Methoden stützt, sind es sicherlich fast immer

die automatischen Methoden, die als erstes hinterfragt werden. Wir haben in unserer Darstellung darum viel Zeit darauf verwendet, diese auszuwerten und das Maß ihrer Verlässlichkeit zu zeigen. Es ist immer eine Abwägung zu treffen zwischen den Vorteilen und den Unsicherheiten, die eine Automatisierung bringt. Allerdings wollen wir auch nochmal ins Gedächtnis rufen, dass – selbst wenn man die Ressourcen dafür hat – auch die Annotation durch Menschen keine vollständige Sicherheit bringt, wie man an den Übereinstimmungraten für die Annotation nicht-direkter Wiedergabe zwischen unseren Erstannotierenden gesehen hat (vgl. Tabelle 5).

Die größte Herausforderung bleibt wohl die schwankende Qualität der Erkennung direkter Wiedergabe – Qualitätsschwankungen sind oft problematischer als im Schnitt schlechtere, aber stabilere Ergebnisse, da sie zu Verzerrungen in der Analyse führen können. Ein bekanntes Problem ist die Tendenz des Erkenners, für direkte Wiedergabe bei Texten in der Ich-Perspektive zu viel zu markieren. Um uns diesem Problem zumindest anzunähern, haben wir für einen großen Teil der Texte die Erzählperspektive ermittelt. Eine Mischung von Texten in Ich- und Er-Perspektive ist sowohl bei der Hochliteratur als auch bei den Heftromanen (allerdings nicht innerhalb aller Einzel-Genres) gegeben und Texte mit der gleichen Perspektive platzieren sich an unterschiedlichen Stellen. Einzig der Bereich mit sehr niedrigem Anteil von direkter Wiedergabe (<17%) ist ausschließlich durch Texte mit Er-Perspektive besetzt. Dies könnte in der Tat darauf hindeuten, dass die Anteile von direkter Wiedergabe in Ich-Erzählungen überschätzt wurden und sich einige der Texte mit Ich-Perspektive in den Scatterplots nach links verschieben sollten. Selbst wenn das jedoch der Fall wäre, sollte es für Hochliteratur und Heftromane gleichermaßen gelten. Insgesamt ist der Einfluss der Erzählperspektive ein Faktor, der in weiteren Untersuchungen genauer betrachtet werden sollte.

Eine weitere Schwäche der Studie ist, dass wir freie indirekte Wiedergabe nicht in die Analyse einbeziehen konnten. Diese Wiedergabeform ist in modernen Texten – sowohl Heftromanen als auch Hochliteratur – durchaus gängig und ein Faktor, der die Erzählweise beeinflusst. Dieses Problem kann jedoch nur gelöst werden, indem die automatische Erkennung für diese Form verbessert wird.

Zudem könnte man einwenden, dass man zwischen der Wiedergabe von Rede und von Gedanken hätte unterscheiden sollen. Wir glauben, dass es für diese erste Untersuchung gerechtfertigt ist, dies nicht zu tun, um erst einmal die allgemeineren Trends zu untersuchen; eine solche Differenzierung würde sich aber für weiterführende Studien anbieten. Ist Hochliteratur vielleicht reflektierender und hat mehr Gedankenwiedergabe? Um die Studie um diese Dimension erweitern zu können, bräuchte man allerdings ebenfalls neue, genauere automatische Erkenner.

Was die Korpuszusammenstellung angeht, so bleibt das Problem, dass wir nicht vollständig sicherstellen konnten, dass bei den Heftromanen eine gute Durchmischung von Autor*innen besteht. Da der individuelle Stil einer Person sich oft sehr stark im Textmaterial niederschlägt, ist dies ein potentieller Störfaktor. Er könnte behoben werden, indem man die Studie noch einmal durchführt und dabei nur Werke verwendet, bei denen die Autor*innen zweifelsfrei bekannt sind – das ist allerdings nicht einfach für Heftromane.

Auch die Vereinfachung, die Heftroman-Genres durch jeweils eine Reihe zu repräsentieren, ist nicht unproblematisch. Besser wäre eine Mischung von Heftromanen aus unterschiedlichen Reihen, die aber dem gleichen Genre zuzuordnen sind. Allerdings ist hier anzumerken, dass manche Heftroman-Genres extrem durch eine Reihe dominiert werden, z. B. Horrorroman durch „John Sinclair". Es ist hier nicht einfach, überhaupt an Texte anderer Reihen heranzukommen, und es stellt sich auch die Frage, ob die Prägung des Heftroman-Genres Horror durch „John Sinclair" nicht so stark ist, dass es gerechtfertigt ist, nur diese Reihe zu analysieren.

Auf einer grundsätzlicheren Ebene kann man in Frage stellen, ob der Vergleich eines Genres mit einer Gruppe wie ‚Hochliteratur' überhaupt sinnvoll ist. Wir argumentieren, dass es in unserer Studie primär um den Unterschied zwischen zwei Literaturtypen geht,

von denen die eine als besonders kreativ und innovativ, die andere als besonders konventionelle Massenware betrachtet wird. Die Genre-Unterscheidung kam hinzu, weil wir differenzierter auf die Varianz innerhalb der Gruppe Heftromane eingehen wollten – und tatsächlich konnten wir nachweisen, dass sich Science-Fiction anders zu Hochliteratur verhält als die anderen drei Genres. Ein interessanter weiterer Schritt wäre, auch bei Hochliteratur-Texten eine Einteilung in Genres durchzuführen, um zu prüfen, wie sich Genre-Unterschiede und Unterschiede zwischen den Literaturtypen zueinander verhalten. Dies wird jedoch dadurch sehr erschwert, dass gerade literarisch hochgeschätzte Werke oft bewusst Genre-Grenzen überschreiten bzw. sich nicht klar einem Genre zuordnen lassen. Immerhin gibt es ein paar Genres, die innerhalb des Literaturbetriebs noch ziemlich klar abgegrenzt sind, wie z. B. Science-Fiction, Fantasy und Kriminalroman. Eine Untersuchung von Werken, die Preise innerhalb solcher Sparten gewonnen haben, im Kontrast zu Heftromanen des gleichen Genres wäre sicherlich interessant.

Zum Weiterlesen

Überblicksdarstellungen und Ressourcen zum Forschungsfeld Digital Humanities sind das Handbuch Jannidis, Kohle und Rehbein (2017) (u. a. mit Kapiteln zu automatischer und manueller Annotation, quantitativen Methoden in der Literaturwissenschaft und Programmieren) sowie die Webressource forText: https://fortext.net (gut aufbereitete Beschreibungen zu Methoden und Werkzeugen der quantitativen Textanalyse für Einsteiger).
Zum Thema Programmieren und maschinelles Lernen bieten sich McKinney (2015) (anschauliches Einsteigerbuch mit Schwerpunkt auf Datenanalyse) und Chollet (2018) (komplexes, aber sehr gutes Einführungskapitel zur Entwicklung von maschinellem Lernen; frei im Netz verfügbar) an.

Literatur

Artstein, Ron (2017): Inter-annotator Agreement, in: *Handbook of Linguistic Annotation*, Bd. 1, 297–313.

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie* (=Narratologia 47), Berlin [u. a.]: de Gruyter.

Brunner, Annelen/Fotis Jannidis/Stefan Engelberg/Ngoc Duyen Tanja Tu/Lukas Weimer (2020): Redewiedergabe in Heftromanen und Hochliteratur, in: Schöch, Christof (Hrsg.), *Digital Humanities: Spielräume. 7. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2020)*, Paderborn: Zenodo, S. 190–194, [online] http://doi.org/10.5281/zenodo.3666690.

Brunner, Annelen/Ngoc Duyen Tanja Tu/Lukas Weimer/Fotis Jannidis (2020): To BERT or not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation, in: Ebling, Sarah/Don Tuggener/Manuela Hürlimann/Mark Cieliebak/Martin Volk (eds.), *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing*, [online] http://ceur-ws.org/Vol-2624/paper5.pdf.

Brunner, Annelen/Lukas Weimer/Stefan Engelberg/Fotis Jannidis/Ngoc Duyen Tanja Tu (2020): *Annotationsrichtlinien des Projekts „Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse" (Version 1.2)*, Zenodo, [online] http://doi.org/10.5281/zenodo.3759617.

Brunner, Annelen/Stefan Engelberg/Fotis Jannidis/Ngoc Duyen Tanja Tu/Lukas Weimer (2020): Corpus REDEWIEDERGABE, in: Calzolari, Nicoletta et al. (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, Marseille*, S. 803–812, [online] http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.100.pdf.

Chollet, François (2018): *Deep Learning with Python*, Shelter Island: Manning.

Genette, Gérard (1998): *Die Erzählung*, 2. Aufl., München: Wilhelm Fink Verlag.

Gius, Evelyn/Janina Jacke (2017): The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis, in *International Journal of Humanities and Arts Computing*, Jg. 11, H. 2, S. 233–254.

Hügel, Hans-Otto (2001): Kommunikative und ästhetische Funktion des Romanhefts, in: Leonhard, Joachim-Felix (Hrsg.), *Medienwissenschaft. Ein Handbuch*

zur Entwicklung der Medien und Kommunikationsformen. 2. Teilband, Berlin [u. a.]: De Gruyter, S. 1621–1631.

Ide, Nancy M. / James Pustejovsky (eds.) (2017): *Handbook of Linguistic Annotation*, Dodrecht: Springer.

Jannidis, Fotis / Hubertus Kohle / Malte Rehbein (Hrsg.) (2017): *Digital Humanities. Eine Einführung*, Stuttgart: J. B. Metzler.

Jannidis, Fotis / Leonard Konle / Peter Leinen (2019): Makroanalytische Untersuchung von Heftromanen, in: Sahle, Patrick (eds.), *Digital Humanities: multimedial & multimodal. Konferenzabstracts*, Frankfurt / Mainz, S. 167–173, [online] http://doi.org/10.5281/zenodo.2600812.

Karr, Hanns P. (2010): Cotton, Jerry, in Karr, Hanns P. (Hrsg.), *Lexikon der deutschen Krimi-Autoren – Internet Edition*, [online] http://www.krimilexikon.de/cotton.htm.

Koplenig, Alexander (2019): A non-parametric significance test to compare corpora, in: *PLoS ONE*, Jg. 14, H. 9, [online] https://doi.org/10.1371/journal.pone.0222703.

Leech, Geoffrey / Mick Short (2013): *Style in fiction. A linguistic introduction to English fictional prose*, 2. Aufl., London [u. a.]: Routledge.

Martínez, Matías / Michael Scheffel (2016): *Einführung in die Erzähltheorie* (=C.H. Beck Studium), 10. Aufl., München: C.H. Beck.

McHale, Brian (2014): Speech Representation, in: Hühn, Peter / John Pier / Wolf Schmid / Jörg Schönert (eds.), *The living handbook of narratology*, Hamburg: Hamburg University Press, [online] http://www.lhn.uni-hamburg.de/article/speech-representation.

McKinney, Wes (2015): *Datenanalyse mit Python*, 1. Aufl., Heidelberg: O'Reilly.

Nast, Mirjam (2017): *„Perry Rhodan" lesen. Zur Serialität der Lektürepraktiken einer Heftromanserie*, Bielefeld: transcript Verlag.

Nusser, Peter (1979): *Romane für die Unterschicht. Groschenhefte und ihre Leser*, 4., mit einem Nachwort versehene Aufl., Stuttgart: Metzler.

Perkuhn, Rainer / Holger Keibel / Marc Kupietz (2012): *Korpuslinguistik* (=LIBAC – Linguistik für Bachelor 3433), Paderborn: Fink.

Stockinger, Claudia (2018): Das Groschenheft, in: Spoerhase, Carlos / Steffen Martus (Hrsg.), *Gelesene Literatur. Populäre Lektüre im Zeichen des Medienwandels*, München: text + kritik.

Tu, Ngoc Duyen Tanja / Stefan Engelberg / Lukas Weimer (2019): „Was für Enthüllungen!", heulte die wohlgekleidete respektable Menge. Eine korpuslinguistische Untersuchung zur lexikalischen Vielfalt von Redeeinleitern, in: *Linguistische Berichte [Sonderausgabe zu Redewiedergabe]*, Bd. 27, S. 13–53.

Die Adressen aller Webseiten und Online-Ressourcen in diesem Beitrag wurden zuletzt auf Aktualität überprüft am 05. Juli 2021.