Originalveröffentlichung in: Zeitschrift für germanistische Linguistik Jg. 47 (2019), Nr. 1, S. 216-248

Annelen Brunner*

Redewiedergabe – Schritte zur automatischen Erkennung

Speech, Thought and Writing Representation – Towards Automatic Detection

https://doi.org/10.1515/zgl-2019-0007

Abstract: This contribution presents a quantitative approach to speech, thought and writing representation (ST&WR) and steps towards its automatic detection. Automatic detection is necessary for studying ST&WR in a large number of texts and thus identifying developments in form and usage over time and in different types of texts. The contribution summarizes results of a pilot study: First, it describes the manual annotation of a corpus of short narrative texts in relation to linguistic descriptions of ST&WR. Then, two different techniques of automatic detection – a rule-based and a machine learning approach – are described and compared. Evaluation of the results shows success with automatic detection, especially for direct and indirect ST&WR.

1 Einführung
2 Korpusgrundlage
3 Manuelle Annotation
3.1 Formen von Redewiedergabe
3.2 Erkenntnisse durch manuelle Annotation
4 Automatische Erkennung
4.1 Regelbasierte Erkennung
4.1.1 Wiedergabewörter
4.1.3 Mustererkennung für indirekte Wiedergabe
4.1.3 Regelbasierte Erkennung von direkter, freier indirekter und erzählter Wiedergabe
4.2 Maschinelles Lernen
4.2.1 Grundprinzip
4.2.2 Attributbewertung
5 Ergebnisse der automatischen Erkennung
6 Ausblick
Literatur

***Kontaktperson: Dr. Annelen Brunner:** Institut für Deutsche Sprache, Postfach 10 16 21, D-68016 Mannheim, E-Mail: brunner@ids-mannheim.de

Publikationsserver des Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-87699

1 Einführung

Während die anderen Beiträge dieses Themenheftes das Phänomen Redewiedergabe in seinen reichen Facetten dargestellt haben, tritt dieser Beitrag einen Schritt zurück. Das methodische Paradigma, in dem er sich bewegt, ist das der Korpuslinguistik, die auf Muster in großen Datenmengen blickt, und das der Digitalen Geisteswissenschaften, die das technische Hilfsmittel Computer fruchtbar machen, um philologische Fragestellungen zu beantworten. Schlagwörter sind *distant reading* (Moretti 2005) oder auch *macroanalysis* (Jockers 2013): Anstatt einzelne Texte und Beispiele detailliert zu studieren, soll eine große Menge von Texten gleichzeitig analysiert werden, um Regelhaftigkeiten und Entwicklungen aufzuspüren, die im Einzelfall unsichtbar bleiben. Der Computer spielt hier insofern eine entscheidende Rolle, als solche Vergleiche im großen Stil nur mit automatischer Unterstützung möglich sind.

Ziel ist am Ende eine quantitative Untersuchung von Redewiedergabe über große Textmengen hinweg, welche es erlaubt, Entwicklungslinien in Form und Verwendung im diachronen Vergleich sowie im Vergleich zwischen Textsorten aufzuzeigen. Die automatische Identifizierung von Redewiedergabe kann daneben auch auf andere Arten genutzt werden, im literaturwissenschaftlichen Bereich etwa, um die Redeanteile von Figuren und Erzähler zu trennen und Analysen zur Figurendarstellung durchzuführen. In diesem Beitrag wird jedoch noch keine solche ‚Makroanalyse' vorgestellt, sondern Schritte auf dem Weg dorthin. Wie kann das Phänomen Redewiedergabe in Regeln gefasst werden? Welche Techniken können angewendet werden, um es automatisch zu identifizieren und wie gut gelingt dies? Der Beitrag richtet sich explizit an Leser, die mit der Herangehensweise der quantitativen Sprach- und Literaturwissenschaft noch wenig vertraut sind. Dargestellt werden Ergebnisse einer Pilotstudie an 13 Erzähltexten aus den Jahren 1787–1913 (Brunner 2015), die es erlaubt, mehrere Aspekte an einem einzigen Beispiel zu illustrieren.[1] Die hier beschriebenen Methoden sind erste Schritte und recht basal. Aktuell läuft ein DFG-Projekt, das auf den vorgestellten Ergebnissen aufsetzt und die Forschungen auf deutlich größerer Datengrundlage und auch textsortenübergreifend fortführt. An verschiedenen Stellen wird auf Verbesserungsmöglichkeiten und aktuellere Entwicklungen hingewiesen werden.

Der erste Schritt hin zur automatischen Erkennung ist ein manuell annotiertes Referenz- und Trainingskorpus. Das Korpus selbst, das verwendete Annota-

1 Alle Materialien der Studie (insbesondere das annotierte Korpus und die Erkenner-Prototypen), sind frei verfügbar unter http://hdl.handle.net/10932/00-027B-9E8A-9300-0B01-E.

tionssystem und erste quantitative Auswertungen werden in den Abschnitten 2 und 3 vorgestellt. Anschließend geht der Beitrag auf Strategien der automatischen Erkennung ein und vergleicht regelbasierte Ansätze und maschinelles Lernen (Abschnitt 4). Den Abschluss bilden eine Auswertung der Ergebnisse, die mit den Erkenner-Prototypen erzielt werden konnten (Abschnitt 5) sowie ein Ausblick auf weitere Forschungen (Abschnitt 6). Anhand der Erfahrungen aus der Studie werden auch zwei Fragen aufgegriffen, die für die quantitativen und digitalen Geisteswissenschaften typisch sind: Wie können vorhandene Erkenntnisse über die Struktur eines Phänomens genutzt und formalisiert werden? Und: Wie generieren die quantitative Betrachtung und die automatischen Methoden selbst neue Erkenntnisse über das Phänomen?

2 Korpusgrundlage

Das Erzähltextkorpus, die Grundlage der vorgestellten Studie, setzt sich aus 13 Erzählungen zusammen, die den Zeitraum vom späten 18. bis zum frühen 20. Jahrhundert umspannen. Es umfasst ca. 57.000 Tokens[2] (Wörter und graphemische Satzzeichen). Die Verwendung abgeschlossener Erzählungen ermöglicht es, auch Beobachtungen über die Verteilung von Redewiedergabe über einen vollständigen Text anzustellen. Auch wenn das Korpus klein ist und modernere Beispiele fehlen, was vor allem urheberrechtliche Gründe hat, wurde bei der Auswahl darauf geachtet, Variation in Alter, Erzählhaltung und Erzählstil der Texte zu erreichen. Tabelle 1 gibt einen Überblick über die verwendeten Texte.

Tabelle 1: Das Erzähltextkorpus.

Jahr	Autor	Titel	Tokens	Kurztitel
1787	Musäus, J. K. A.	Die Entführung	5222	Musäus: Entführung
1788	Bürger, G. A.	Münchhausen (Kapitel 1)	1660	Bürger: Münchhausen
1802	Bernhardi, S.	Belinde	4696	Bernhardi: Belinde
1805	Günderrode, K. von	Geschichte eines Braminen	4393	Günderrode: Bramine
1807	Kleist, H. von	Das Erdbeben in Chili	6577	Kleist: Erdbeben
1812	Tieck, L.	Der blonde Eckbert	7593	Tieck: Eckbert

2 Die Tokenisierung und Satzendeerkennung wurden durchgeführt mit Werkzeugen aus dem Java-basierten Open-Source-Framework GATE (General Architecture for Text Engineering; https://gate.ac.uk) (Cunningham et al. 2013) und zwar dem *Gate Unicode Tokenizer* und dem *RegEx Sentence Splitter* (vgl. https://gate.ac.uk/gate/doc/plugins.html).

Jahr	Autor	Titel	Tokens	Kurztitel
1825	Hauff, W.	Die Geschichte von Kalif Storch	4741	Hauff: Kalif
1849	Hebbel, F.	Die Kuh	2081	Hebbel: Kuh
1878	May, K.	Die verwünschte Ziege	5831	May: Ziege
1889	Schnitzler, A.	Mein Freund Ypsilon	4976	Schnitzler: Ypsilon
1902	Janitschek, M.	Darüber kommt kein Weib hinweg	1754	Janitschek: Weib
1913	Heym, G.	Der Irre	5653	Heym: Irre
1913	Kafka, F.	Der Jäger Gracchus	2045	Kafka: Gracchus

‚Jahr' bezeichnet das Jahr des Erstdrucks der verwendeten Fassung, mit ‚Tokens' sind sowohl Wörter als auch graphemische Satzzeichen gemeint.

Textquelle war die Plattform zeno.org, die gemeinfreie Texte digitalisiert zur Verfügung stellt. Die zugrunde liegenden Textausgaben sind in unterschiedlichem Maße modernisiert. Bei einigen entspricht die vorliegende Form tatsächlich der Erstausgabe (z. B. bei *Bernhardi: Belinde*), dies wurde jedoch nicht durchgehend sichergestellt, denn der Fokus der Studie lag nicht darauf, diese speziellen Texte in ihrer historischen Form zu untersuchen, sondern möglichst allgemeingültige Verfahren zu entwickeln. Insgesamt ist die Textqualität gut, aber relativ inhomogen: Einige Texte weisen ältere Schreibvarianten auf und auch die Zeichensetzung ist nur teilweise normiert, so dass etwa die Verwendung von Anführungszeichen zur Markierung direkter Wiedergabe von Text zu Text unterschiedlich gehandhabt wird. Detaillierte Beschreibungen der Textgestalt der einzelnen Texte und deren Abweichung von der jeweiligen Erstausgabe finden sich in Brunner (2015: 18–29).[3]

Die Tatsache, dass die Texte sowohl in ihrer Form als auch in ihrer Erzählweise recht unterschiedlich sind, ist für die Entwicklung automatischer Erkennungsmethoden eine Erschwernis. Es sollte jedoch bewusst eine Bandbreite von Texten abgedeckt werden, die in ihrer Form und Qualität ein Abbild dessen darstellen, was an digitalisiertem Textmaterial zur Verfügung steht, ohne eigene Korpora aufzubauen, und damit einen realistischen Anwendungsfall simulieren.

3 Wie in den Beiträgen von Demske und Habermann in diesem Heft dargestellt, unterliegen Formen der Redewiedergabe einem diachronen Wandel. In den Texten des Erzähltextkorpus konnte allerdings keine klar zeitlich bedingte Veränderung in Form oder Funktion beobachtet werden, was schon aufgrund der geringen Textmenge nicht erstaunlich ist. Abweichungen zwischen den einzelnen Texten lassen sich eher auf Autorenstil und unterschiedliche Textgestalt zurückführen.

3 Manuelle Annotation

3.1 Formen von Redewiedergabe

In der Einleitung dieses Themenhefts wurde klar, dass sehr unterschiedliche Beschreibungssysteme für Redewiedergabe existieren. Im Folgenden wird kurz umrissen, wie das Phänomen in diesem Beitrag betrachtet wird und welche Formen die Studie unterscheidet.

In realweltlichen Kontexten wird das Phänomen Redewiedergabe zumeist als Einbettung der Originaläußerung eines ersten Sprechers in die Äußerung eines zweiten Sprechers beschrieben. Aus der Perspektive der Literaturwissenschaft ist Redewiedergabe die Modellierung der Äußerung einer Figur im Erzählertext, und damit von besonderem Interesse, wenn es um Figurendarstellung und Erzähltechnik geht. Sie wird oft im Kontext einer umfassenden Erzähltheorie betrachtet (vgl. z. B. Genette 2010, Stanzel 2008, Martínez & Scheffel 2016, Leech & Short 2013) und in Hinblick auf ihre Rolle in den Erzählstrukturen fiktionaler Texte und der erzählerischen Wirkung untersucht. Vor diesem Hintergrund war das Interesse an der Entwicklung von detaillierten Kategoriensystemen für Redewiedergabe in der Literaturwissenschaft besonders ausgeprägt. Deswegen und weil Anwendungsbereiche der quantitativen Betrachtung und automatischen Erkennung von Redewiedergabe auch gerade in der Narratologie liegen, orientiert sich das in der Studie verwendete Kategoriensystem stark an literaturwissenschaftlichen Systemen, insbesondere Semino & Short (2004). Unterschieden werden die Typen direkte Wiedergabe (Er sagte: *„Ich bin hungrig“*), indirekte Wiedergabe (Er sagte, *er sei hungrig.*), sowie erzählte Wiedergabe (*Er sprach über das Mittagessen.*) und freie indirekte Wiedergabe. Der Begriff ‚freie indirekte Wiedergabe‘ oder auch ‚erlebte Rede‘ wird in der Forschung sehr unterschiedlich gehandhabt.[4] Im vorliegenden Beitrag bezeichnet ‚freie indirekte Wiedergabe‘ unabhängige Sätze, deren besonderes Merkmal eine Vermischung von Erzähler- und Figurenperspektive ist (Er war ratlos. *Wo sollte er jetzt nur etwas zu essen herbekommen?*). Sie können sowohl Gedanken als auch Rede oder Geschriebenes wiedergeben und weisen indikativische, meist präteritale Verbformen auf. Unabhängige konjunktivische Wiedergabesätze ohne Rahmenformel (Sie stellte viele Fragen. *Wann gebe es Mittagessen? Wo sei das Restaurant?*) werden im Annotationssystem als Sonderfall behandelt und als Mischform zwischen indirekter und freier indirekter Wiedergabe klassifiziert, da sie mit beiden Formen Gemeinsamkeiten aufweisen: Sie besitzen die Unabhängigkeit von freier indirekter Wiedergabe, aber ver-

4 Vgl. hierzu die Einleitung und vor allem den Beitrag von Holler in diesem Heft.

wenden den Modus Konjunktiv, der ein typisches (wenn auch nicht zwingendes) Merkmal von indirekter Wiedergabe ist. Bei der Auswertung werden sie doppelt gezählt, als Fälle von indirekter und freier indirekter Wiedergabe. Diese Formen traten im untersuchten Korpus allerdings nur sehr marginal auf.

Zudem wird differenziert, ob es sich um die Wiedergabe von tatsächlicher Rede – also gesprochener Sprache – von Geschriebenem oder von Gedanken handelt. Damit ergeben sich 12 Hauptkategorien für Wiedergabeformen.

Das Annotationssystem sieht zudem eine Reihe von Attributen vor, die es erlauben, die Instanzen von Wiedergabe näher zu bestimmen und Sonder- und Grenzfälle zu markieren. An dieser Stelle soll nur auf die am häufigsten gebrauchten eingegangen werden. Das Attribut *level* markiert die Einbettungstiefe einer Wiedergabe in eine andere Wiedergabe (Peter sagte: *„Meine Schwester redet viel über Restaurants."* → erzählte Wiedergabe eingebettet in direkte Wiedergabe). Mit *non-fact* (‚nicht-faktisch') wird markiert, dass die wiedergegebene Rede- Schreib- oder Gedankenhandlung in der Textwelt nicht stattgefunden hat, es sich also z. B. um eine negierte oder hypothetische Wiedergabe handelt (Er sagte nicht, *dass er Hunger habe*.). Das Attribut *border* (‚Grenzfall') wird verwendet, um Fälle zu markieren, die im Grenzbereich der Definition für verschiedene Wiedergabekategorien liegen. Dies ist vor allem bei Gedankenwiedergabe relevant, da es sehr schwer abzugrenzen ist, welche kognitiven Prozesse noch als Gedanken zu werten sind und welche nicht. So werden z. B. Wissenszustände (Er weiß, *dass er hungrig ist*.) und Wahrnehmungen (Sie sah, *dass es keinen Zweck hatte, weiter zu suchen*.) als *border*-Phänomene gekennzeichnet.

Redewiedergabe ist für die Automatisierung insofern interessant, als sie einerseits einige klare Oberflächenmerkmale aufweist, andererseits aber komplex und kontextabhängig genug ist, dass eine Klassifizierung nicht trivial und auch für Menschen nicht immer ohne Weiteres zu leisten ist. Schon die manuelle Annotation steht in diesem Spannungsfeld – einerseits muss eine gewisse Formalisierung und Verbindlichkeit vorhanden sein, um den Annotatoren Anhaltspunkte zu geben und ihre Annotation über die Zeit hinweg und auch zwischen unterschiedlichen Personen konsistent zu halten, andererseits verfügen die menschlichen Annotatoren über Textverständnis, Weltwissen und Kontextbezogenheit, die sie in die Lage versetzen, adäquat auf Strukturen zu reagieren, die nicht explizit in den Annotationsrichtlinien thematisiert sind. Letzteres ist genau die Leistung, die später idealerweise auch von automatischen Erkennern erbracht werden soll. Bei der Erstellung der Annotationsrichtlinien muss also zweierlei vermittelt werden: einerseits eine Charakterisierung der jeweiligen Form in ihrer Funktion und Wirkung, zum anderen aber auch Indikatoren, die zur Identifizierung nützlich sind. Im Falle von direkter Wiedergabe sind dies z. B. Anführungszeichen, im Falle von indirekter Wiedergabe das Auftreten einer Rahmenformel

und eines unterordneten Satzes, der die wiedergegebene Proposition enthält. In der Auseinandersetzung mit der Formulierung von Regeln und der Behandlung von Zweifel- und Sonderfällen schärft sich zugleich der Blick auf das Phänomen selbst, gerade wenn diese Regeln nicht nur auf ausgewählte Beispiele, sondern auf Instanzen in einem ganzen Korpus angewendet werden müssen. Dies hat sich bereits in anderen Annotationsprojekten zu komplexen sprachlichen Phänomenen gezeigt (z. B. Metaphernannotation bei Steen et al. 2010, Kategorien der Zeit bei Gius & Jacke 2016, Redewiedergabe im Englischen bei Semino & Short 2004).[5]

Die Organisation und Qualitätssicherung manueller Annotation ist ein sehr aktuelles Thema der Korpuslinguistik (vgl. z. B. Ide & Pustejovsky 2017). Aufgrund mangelnder Ressourcen wurde in der Studie die Annotation nur von einer einzigen Person durchgeführt, was angesichts der Komplexität des Phänomens Redewiedergabe, die sich auch in diesem Heft gezeigt hat, nicht ideal ist. Im eingangs erwähnten DFG-Projekt hingegen wird jeder Textausschnitt von zwei Annotatoren unabhängig voneinander bearbeitet und auf dieser Basis von einer dritten Person eine Konsensannotation erstellt.[6] Dieser aufwendige Prozess gewährleistet eine deutlich stärkere Konsistenz und größere Verlässlichkeit.

3.2 Erkenntnisse durch manuelle Annotation

Die systematische, manuelle Annotation einer größeren Anzahl von Texten erlaubt erste quantifizierbare Erkenntnisse. Wie Abbildung 1 zeigt, sind die Anteile der Wiedergabetechniken in den verschiedenen Texten sehr unterschiedlich.

Vor allem der Anteil an direkter Wiedergabe schwankt stark zwischen äußerst dialoglastigen Texten wie *May: Ziege*, der zu etwa 70 % aus direkter Wiedergabe besteht, und solchen wie *Bürger: Münchhausen*, der fast überhaupt keine enthält. Gerade die sehr häufige und auffallende Form direkte Wiedergabe hängt also stark von Inhalt und Erzählweise des Einzeltextes ab. Die gestrichelten Linien zeigen, wie die Anteile aussehen, wenn man Binnenerzählungen, also mehrschichtiges Erzählen, bei dem eine Figur zum Erzähler wird, ebenfalls als direkte Wiedergabe

5 Die Studie von Semino & Short (2004) beschäftigt sich mit modernen, englischsprachigen Texten und stellt auch Textsortenvergleiche an. Sie diente Brunner (2015) als Vorbild, allerdings nur für den Teil, der sich mit manueller Annotation beschäftigt. Eine Automatisierung wurde bei Semino & Short (2004) nicht versucht.

6 Die genauen Annotationsrichtlinien der Studie sind nachzulesen in Brunner (2015: 51–93). Die Annotationsrichtlinien des DFG-Projekts, welche eine direkte Fortentwicklung darstellen, sind frei verfügbar unter www.redewiedergabe.de.

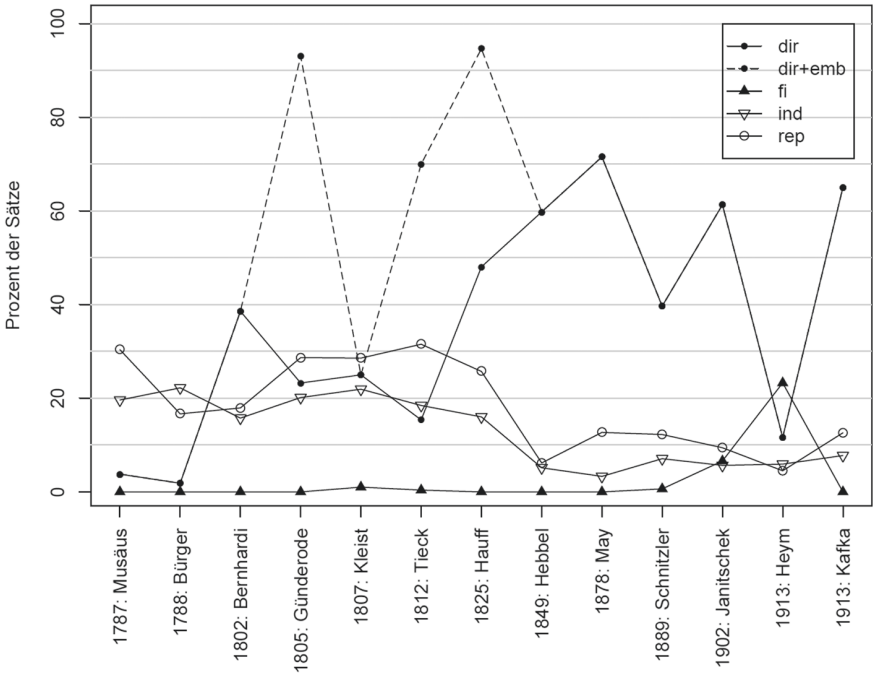

Abbildung 1: Verteilung der Wiedergabetypen in den Texten des Erzähltextkorpus. (dir = direkte Wiedergabe, fi = freie indirekte Wiedergabe, ind = indirekte Wiedergabe, rep = erzählte Wiedergabe, dir+emb = direkte Wiedergabe und Binnenerzählung zusammengefasst). Die Texte sind in der Abbildung chronologisch angeordnet, die geringe Größe des Korpus erlaubt jedoch keine verlässlichen Aussagen über zeitliche Entwicklungslinien.

wertet. Die Werke *Hauff: Kalif*, *Tieck: Eckbert* und *Günderode: Bramine* sind alle so konstruiert, dass ein wesentlicher Teil des Inhalts über die direkte Rede einer Figur vermittelt wird, womit ein Erzählerwechsel stattfindet. Diese Wiedergaben werden in der Auswertung separat behandelt und im Folgenden nicht als direkte Rede gezählt.

Freie indirekte Wiedergabe tritt nur in drei Texten auf und ist damit im Erzähltextkorpus insgesamt sehr schwach vertreten, scheint aber ähnlich textabhängig zu sein. Indirekte und erzählte Wiedergabe sind insgesamt seltener als direkte, aber deutlich gleichmäßiger verteilt.

Abbildung 2 veranschaulicht noch einmal für jeden der vier Wiedergabetypen, in wie vielen Sätzen des Erzähltextkorpus er auftritt. Enthält ein Satz mehrere unterschiedliche Wiedergabetypen (ineinander geschachtelt oder auf-

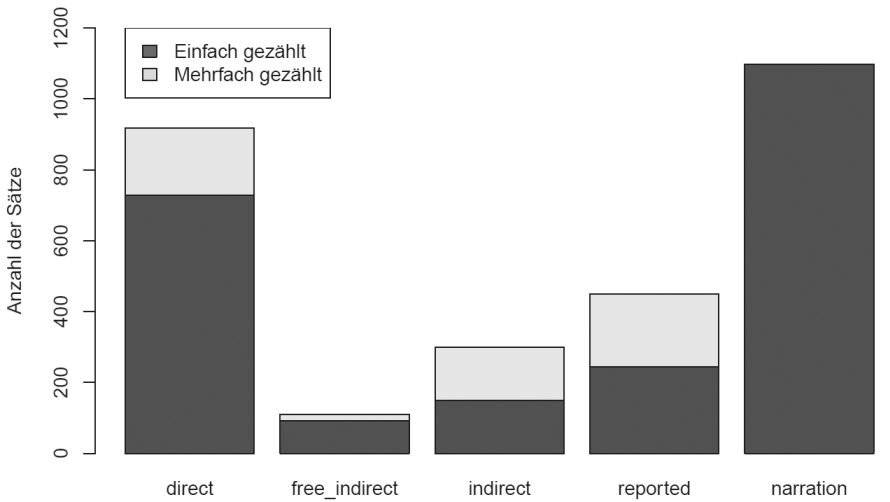

Abbildung 2: Anzahl der Sätze, die einen Wiedergabetyp enthalten (direct = direkte Wiedergabe, free_indirect = freie indirekte Wiedergabe, indirect = indirekte Wiedergabe, reported = erzählte Wiedergabe, narration = Erzählertext).

einander folgend), wird er mehrfach gezählt, so dass die Gesamtanzahl der Sätze über alle Balken in der Graphik aufsummiert größer ist als die Anzahl der Sätze im Korpus. Der Anteil an mehrfach gezählten Sätzen pro Wiedergabetyp ist hellgrau eingefärbt. Der Balken ‚narration' (Erzählertext) zeigt die Anzahl der Sätze, die gar keinen Wiedergabetyp enthalten. Es fällt auf, dass der Anteil an mehrfach gezählten Sätzen bei indirekter und erzählter Wiedergabe besonders hoch ist, an die 50 %. Ursachen sind, dass Instanzen dieser Wiedergabetypen zum einen oft kürzer sind, so dass mehrere hintereinander in einem Satz auftauchen können, zum anderen auch deutlich häufiger in andere – meist direkte – Wiedergaben eingebettet werden (vgl. Pitz & Solfjeld in diesem Heft). Eine Auswertung der Attribute ergibt zudem, dass in diesen Kategorien deutlich mehr nicht-prototypische Fälle von Wiedergabe auftreten (z. B. nicht-faktische Wiedergabe, Grenzfälle). Direkte und freie indirekte Wiedergaben dagegen umfassen meist vollständige Sätze und tendieren dazu, in mehreren Sätzen hintereinander aufzutreten.

Zudem lässt sich anhand der manuellen Annotation beobachten, dass die Wiedergabe von Rede mehr als doppelt so häufig auftritt wie die Wiedergabe von Gedanken. Wiedergabe von Geschriebenem ist im Erzähltextkorpus nur sehr gering vertreten. Es gibt Hinweise, dass dieser Trend für Erzähltexte typisch ist:

Eine sehr ähnliche Verteilung lässt sich für englischsprachige literarische Texte bei der Studie von Semino & Short (2004) beobachten (vgl. Brunner 2015: 98–99).

Um diese Beobachtungen zu untermauern und detailliertere Analysen, z. B. Entwicklungen von Wiedergabeformen über die Zeit hinweg oder Vergleiche zwischen Textsorten, zu ermöglichen, ist jedoch natürlich deutlich mehr Material notwendig als die 13 Texte des Erzähltextkorpus. Da die manuelle Annotation sehr zeitaufwendig ist, ist dies die Hauptmotivation für die Entwicklung automatischer Erkennungsmethoden.

4 Automatische Erkennung

Es lassen sich zwei grundsätzliche Herangehensweisen unterscheiden, wie man die Erkennung von Wiedergabe automatisieren kann:[7] Einerseits regelbasierte Verfahren, welche vorhandenes Wissen über die Besonderheiten der Wiedergabetechniken formalisieren und in explizite Regeln überführen. Zum anderen maschinelle Lernverfahren, welche durch die Verallgemeinerung exemplarischer Fälle Heuristiken aufbauen, wobei die manuellen Annotationen als Trainingsmaterial verwendet werden. Selbstverständlich kann man beide Vorgehensweisen kombinieren, in der Studie wurden sie jedoch getrennt voneinander implementiert und getestet, um direkte Vergleiche anstellen zu können.[8]

Für die Auswertung werden die Ergebnisse der automatischen Methoden mit den manuellen Annotationen verglichen. Letztere gelten also als der im Idealfall zu erreichende Standard (oft ‚Gold Standard' genannt). Bei den hier vorgestellten Ergebnissen ist es für ein positives Ergebnis ausreichend, wenn korrekt erkannt wird, dass ein Satz eine bestimmte Form von Wiedergabe enthält, deren genaue Abgrenzung muss nicht korrekt erfasst sein.[9] Zur Abschätzung des Erfolgs sind

7 Die Unterscheidung regelbasierte Verfahren vs. maschinelles Lernen zieht sich durch viele Bereiche der computergestützten Datenextraktion und -analyse. Für eine neuere Betrachtung welche Strategien im akademischen und kommerziellen Kontext bevorzugt werden vgl. Chiticariu, Li & Reiss (2013).

8 Die Implementierung der regelbasierten Erkenner-Prototypen erfolgte in dem Java-basierten Open-Source-Framework GATE (General Architecture for Text Engineering; https://gate.ac.uk) (vgl. Cunningham et al. 2013). Das maschinelle Lernen wurde mit der Statistik-Software R (https://cran.r-project.org) durchgeführt.

9 Die Erkenner-Prototypen in Brunner (2015) sind nicht darauf ausgelegt, die genauen Grenzen der Wiedergabe zu erfassen. Eine Ausnahme stellen die regelbasierten Methoden für die Erkennung direkter und freier indirekter Wiedergabe dar, wobei diese Formen per Definition sowieso fast immer Sätze umfassen.

zwei Maße zentral: *Precision* (Genauigkeit) und *Recall* (Trefferquote). Precision bezeichnet den Anteil von automatisch erkannten Fällen, die tatsächlich korrekt sind. Recall bezeichnet den Anteil an vorhandenen Fällen, die von den automatischen Methoden gefunden wurden (vgl. z. B. Manning, Raghavan & Schütze 2008: 142–144). Tabelle 2 zeigt eine Beispielauswertung.

Tabelle 2: Beispielauswertung.

Enthält direkte Wiedergabe?	Korrekte Antwort	Automatische Vorhersage
Satz 1	Ja	Ja
Satz 2	Ja	Nein
Satz 3	Nein	Nein
Satz 4	Nein	Ja
Satz 5	Ja	Nein

In diesem Fall wäre die Precision für die Erkennung direkter Wiedergabe 0,5 (Es wurden 2 Fälle als positiv vorhergesagt, Sätze 1 und 4, davon war eine Vorhersage korrekt) und der Recall 0,33 (Es gab 3 positive Fälle, Sätze 1, 2 und 5, davon wurde einer gefunden). Der F-Score, das harmonische Mittel von Recall und Precision, wird berechnet mit der Formel: 2 * (Precision * Recall) / (Precision + Recall) und wäre damit 0,40.

Idealerweise sollten beide dieser Werte hoch sein, oft ist es jedoch so, dass die Optimierung des einen auf Kosten des anderen geht: Eine sehr restriktive Erkennungsmethode führt zu verlässlichen Ergebnissen (hohe Precision), aber übersieht ungewöhnlichere Fälle (niedriger Recall); eine sehr großzügige Erkennungsmethode erfasst viele Vorkommen (hoher Recall) aber unter den gefundenen Fällen sind auch viele fehlerhafte (niedrige Precision). Ein Maß, das sowohl Precision als auch Recall berücksichtigt, ist der F-Score, in den die beiden Werte zu gleichen Teilen einfließen. Der Maximalwert für Precision, Recall und F-Score ist jeweils 1,0.

4.1 Regelbasierte Erkennung

Da regelbasierte Erkennung die Besonderheiten der einzelnen Wiedergabetypen kodiert, liegt es in der Natur der Sache, dass für jeden Typ ein eigenes Regelset implementiert werden muss. Der Ausgangspunkt ist immer die Textoberfläche, an der sich Indikatoren unterschiedlicher Art zeigen: lexikalische (z. B. Wörter, deren Semantik auf einen Wiedergabeakt hinweist), grammatische (z. B. Nebensatzstruktur) oder typographische (z. B. Anführungszeichen). Im Folgenden wird

dargestellt, welche Strategien in den bewusst einfach gehaltenen regelbasierten Erkenner-Prototypen angewendet werden.

4.1.1 Wiedergabewörter

Ein lexikalischer Indikator, der sich nicht nur für die regelbasierten Verfahren, sondern – wie später noch ausgeführt wird – auch für das maschinelle Lernen als zentral erwiesen hat, sind ‚Wiedergabewörter', d. h. Wörter, die auf eine Sprach-, Denk- oder Schreibhandlung hinweisen können. In linguistischen Darstellungen werden in diesem Kontext am häufigsten die *verba dicendi*, Verben des Sagens, genannt (vgl. z. B. Helbig & Buscha 2011), jedoch ist die Gruppe potentiell interessanter Verben und Nomen deutlich größer. Für die regelbasierte Erkennung wurden verschiedene Quellen verwendet, um eine Liste zu kompilieren. Den Grundstock lieferte eine linguistisch motivierte und redaktionell bearbeitete Sammlung: der Index des Handbuchs für Kommunikationsverben (Harras et al. 2004). Zur Ergänzung wurden korpuslinguistische Methoden angewendet: Mit Hilfe einer einfachen Mustersuche (Muster: Anführungszeichen – Komma – Verb) wurden aus einem Korpus literarischer Texte[10] Verben herausgefiltert, welche in der Rahmenformel von direkter Wiedergabe auftraten. Um die Liste noch durch Nomen anzureichern, welche ebenfalls als Wiedergabeindikatoren fungieren können, wurde zudem OpenThesaurus (http://openthesaurus.de) verwendet, ein online frei verfügbares, durch Crowdsourcing aufgebautes Synonymwörterbuch. Ausgehend von einigen Nomen, die klar auf Wiedergabe verweisen, (z. B. *Antwort*, *Gespräch*, *Rede*) wurden mit Hilfe der Ressource weitere, bedeutungsverwandte Nomen identifiziert. Nach Zusammenführung der drei Ergebnismengen und Entfernung von Dubletten ergab sich eine Liste von 960 Wörtern (Verben und Nomen).

Bei der Zusammenstellung wurde bewusst großzügig vorgegangen, um eine möglichst hohe Erkennungsrate zu erzielen, d. h. es sind viele Wörter vorhanden, die nur unter speziellen Umständen Indikatoren für Wiedergabe sind. Bei der Aufbereitung der Liste wurden darum den Wörtern *penalties* (‚Strafpunkte') zwischen 0 und 5 zugewiesen, die umso höher sind, je weniger prototypisch das Wort als Wiedergabeindikator ist. Wert 0 erhielten z. B. Wörter wie *sagen, denken, Nachricht*; Wert 5 Wörter, die nur sehr schwach mit Wiedergabe assoziiert sind, z. B.

10 Dieses Korpus setzte sich aus Texten aus der „Digitalen Bibliothek" zusammen, welche als Teil des „Archivs für historische Korpora (HIST)" über das Recherchesystem des Instituts für Deutsche Sprache, COSMAS II, verfügbar sind (http://www.cosmas2.de). Es umfasste ca. 45 Millionen Tokens und hatte keine Überschneidungen mit dem Erzähltextkorpus.

Demontage, *einweihen*, oder nur in speziellen Kontexten die Funktion als Wiedergabewörter übernehmen können, z. B. *sehen* (im Sinne von *erkennen*), *überbringen* (z. B. einer Nachricht). Zudem wurden Verben markiert, die typischerweise nur bei erzählter Wiedergabe verwendet werden (z. B. *bedanken, abschwören, lossagen*). Diese Zuordnungen, wie auch die Bereinigung der Liste, wurden für die Studie rein kompetenzbasiert durchgeführt und nicht weiter validiert. Es ist offensichtlich, dass es hier zahlreiche Verbesserungsmöglichkeiten gäbe: Wünschenswert wäre vor allem eine verlässlichere, am besten korpusgestützte Bewertung und Kategorisierung der Wörter, die idealerweise auch mit einbezieht, für welchen Wiedergabetyp das Wort typisch ist. Zudem erfasst die Wiedergabewortliste nur Einzelwörter, eine adäquate Abbildung von trennbaren Verben und Phrasemen fehlt. Die Berücksichtigung solcher mehrgliedrigen Strukturen setzt allerdings dann auch eine komplexere morphologische und grammatische Vorverarbeitung des zu annotierenden Textes voraus, was mit automatischen Methoden nicht trivial ist. Wie sich zeigen wird, ist die Wiedergabewortliste jedoch auch in ihrer groben Form schon sehr nützlich.[11]

4.1.2 Mustererkennung für indirekte Wiedergabe

Auf das Vorgehen bei der regelbasierten Erkennung von indirekter Wiedergabe soll etwas ausführlicher eingegangen werden, da dieser Wiedergabetyp die am stärksten strukturell definierte Form aufweist, weshalb sie in besonderem Maße für einen regelbasierten Ansatz geeignet ist. Zudem wird die indirekte Wiedergabe in linguistischen Darstellungen üblicherweise am umfassendsten behandelt, so dass sich hier gut demonstrieren lässt, wie linguistisches Wissen in die Regelformulierung einfließt.

Helbig & Buscha (2011) fassen die Indikatoren folgendermaßen zusammen:

> Zur formalen Kennzeichnung der indirekten Rede dienen:
> – der Konjunktiv
> – redeeinleitende Verben
> – die Nebensatzform
>
> Keines dieser Mittel ist obligatorisch, doch ist gewöhnlich zumindest eines vorhanden, um die indirekte Rede als solche zu kennzeichnen. (Helbig & Buscha 2011: 174)

11 Die komplette Wiedergabewortliste ist ebenfalls unter http://hdl.handle.net/10932/00-027B-9E8A-9300-0B01-E verfügbar.

Die „redeeinleitenden Verben“ wurden bereits im Kontext der Wiedergabewortliste behandelt. Der Konjunktiv hat in der Beschreibung indirekter Wiedergabe als ‚Referatskonjunktiv‘ oder ‚Indirektheitskonjunktiv‘ einen wichtigen Platz (vgl. z. B. Zifonun, Hoffmann & Strecker 1997, Wöllstein & Dudenredaktion 2016). Die Duden-Grammatik konstatiert mit Hinweis auf Korpuszählungen, dass die indirekte Wiedergabe in der Schriftsprache der wichtigste Funktionsbereich des Konjunktivs sei (vgl. Wöllstein & Dudenredaktion 2016: § 762). Weder Konjunktiv I noch Konjunktiv II sind jedoch nur für Wiedergaben reserviert (vgl. z. B. Eisenberg 2013: 107–117). Umgekehrt besteht inzwischen, im Gegensatz zu früheren, normativen Darstellungen, Konsens, dass die Verwendung des Konjunktivs in der indirekten Wiedergabe nicht obligatorisch ist (für eine kurze wissenschaftshistorische Betrachtung vgl. Kaufmann 1976: 27). Die deskriptive Untersuchung ergibt, dass in indirekten Wiedergabesätzen, die von referatseinleitenden Verben abhängen, vier unterschiedliche Formen auftreten können: Konjunktiv I, Konjunktiv II, die Ersatzform mit *würde* und Indikativ. Welche verwendet wird, ist von verschiedenen Faktoren beeinflusst: Oft genannt werden mündlicher vs. schriftlicher Sprachgebrauch und morphologische Gegebenheiten (wenn Indikativ- und Konjunktiv-I-Formen morphologisch ununterscheidbar sind, werden gerne Ersatzformen gewählt; vgl. z. B. Wöllstein & Dudenredaktion 2016: § 776 und § 777; Helbig & Buscha 2011: 177). Auch das referatseinleitende Verb sowie dessen Person und Tempus in der Rahmenformel scheinen einen gewissen Einfluss auf die Wahl der Verbform zu haben (vgl. Strecker 2009). Zusammenfassend lässt sich sagen, dass der Konjunktiv, insbesondere Konjunktiv I, bei Einleitungsverben in der dritten Person Präteritum und in geschriebener Sprache weit verbreitet ist. Schriftlichkeit, Tempus Präteritum und 3. Person sind auch Merkmale, die gerade in literarischen Texten oft vorliegen. Konjunktiv ist also ein guter, wenn auch kein uneingeschränkt verlässlicher Indikator für Wiedergabe.

Ein weiteres entscheidendes Merkmal von indirekter Wiedergabe ist das Vorhandensein der Nebensatzform. Übliche Strukturen sind Verbletztsätze mit einleitendem Element, typischerweise den Konjunktionen *dass* und *ob*, sowie W-Fragewörtern (*wer*, *wie* usw.). Der verlässlichste Indikator ist *ob* – das „Handbuch der deutschen Konnektoren“ weist sogar darauf hin, dass die „Verwendung als Interrogativausdruck in indirekten Entscheidungsfragen“ (Pasch et al. 2003: 629) die wichtigste Funktion dieses Konnektors sei. Die anderen Elemente treten hingegen auch oft in anderen Konstruktionen auf: W-Fragewörter in Relativsätzen und die Konjunktion *dass* in finaler oder konsekutiver Bedeutung.

Die zweite typische Form sind Verbzweitsätze. Diese Konstruktion ist ein sehr verlässliches Indiz für Wiedergabe, allerdings ist sie auch schwerer zu identifizieren, da es hier keine lexikalischen Indikatoren in Form eines einleitenden Kon-

nektors gibt und eine strukturelle Satzanalyse notwendig ist, um zu erkennen, dass genau ein Satzglied vor dem Verb steht. Zudem ist die Verbzweitstellung nur von der sonst üblichen Verbletztstellung zu unterscheiden, wenn es mindestens zwei Satzglieder gibt.

Eine weitere Möglichkeit, indirekte Wiedergabe zu kodieren, ist die Infinitivphrase mit *zu*. Auch diese Form findet sich in linguistischen Darstellungen, wird jedoch seltener explizit thematisiert (vgl. aber z. B. Wöllstein & Dudenredaktion 2016: § 1684). Die Wahl dieser Form wird oft durch das referatseinleitende Verb bestimmt und findet sich v. a. in Aufforderungssätzen (*Er befahl ihr, zu gehen.*). Allerdings ist die Infinitivphrase mit *zu* für sich genommen kein klarer Indikator, da es zahlreiche frequente Verben gibt, die diese Konstruktion nach sich ziehen, ohne etwas mit Wiedergabe zu tun zu haben (z. B. *anfangen*, *aufhören* usw.).

Dieser kurze Überblick zeigt, dass in der linguistischen Forschung zahlreiche Indikatoren für indirekte Wiedergabe identifiziert sind, keiner davon für sich genommen jedoch eindeutig ist und sich auch keine Regel formulieren lässt, die alle Formen von indirekter Wiedergabe erfassen würde.

An dieser Stelle noch ein paar Worte zur Definition von indirekter Wiedergabe in den Richtlinien für die manuelle Annotation. Diese ist entscheidend, da an dieser Stelle festgelegt wird, welche Phänomene eigentlich idealerweise erfasst werden sollen, was sich natürlich auf die Bewertung der automatischen Erkennung auswirkt. Die Definition von indirekter Wiedergabe in den Richtlinien ist stärker von formalen Kriterien bestimmt als bei allen anderen Definitionen für Wiedergabetechniken und verlangt explizit eine Rahmenformel und einen untergeordneten Satz. Als typische Formen werden genau die oben beschriebenen Formen genannt, also Nebensätze mit Verbletztstellung oder Verbzweistellung sowie die Infinitivphrase mit *zu* – das Annotationsystem erzwingt damit bereits eine strukturelle Beschränkung, was die Regelformulierung erleichtert. Wie bereits erwähnt, werden konjunktivische Wiedergabesätze ohne Rahmenformel als Mischform zwischen indirekter und freier indirekter Wiedergabe gekennzeichnet, zählen bei der Auswertung jedoch auch als Fälle von indirekter Wiedergabe. Im Erzähltextkorpus ist dieser Strukturtyp allerdings selten, so dass Fehler an dieser Stelle wenig ins Gewicht fallen.

Daneben werden im Kontext indirekter Wiedergabe auch die sogenannten ‚formelhaften Referatshinweise' genannt, also Sätze, die die Indirektheit der Wiedergabe durch Einschübe signalisieren, z. B. *laut X/nach Aussage von X/so X/X zufolge* (vgl. z. B. Fabricius-Hansen 2001: 22, Weinrich 2005: 900, Wöllstein & Dudenredaktion 2016: § 768). Diese sind typisch für den journalistischen Sprachgebrauch. Diese Wendungen sind vom Annotationssystem nicht abgedeckt, da im Erzähltextkorpus kein solcher Fall auftritt. Sie sind darum für die hier vorge-

stellten Auswertungen irrelevant und werden auch bei der automatischen Erkennung nicht berücksichtigt.[12]

Im Folgenden wird gezeigt, wie die beschriebenen Indikatoren für ein einfaches regelbasiertes System genutzt werden können.

Die regelbasierte Erkennung von indirekter Wiedergabe verwendet die Wörter der Wiedergabewort-Liste als Anker. Wörter, für die bei der Listenerstellung festgelegt wurde, dass sie höchstwahrscheinlich nur bei erzählter Wiedergabe auftreten, wurden komplett ausgeschlossen. Dadurch verkürzt sich die Gesamtliste von 960 auf 724 Einträge. Vom Wiedergabewort ausgehend werden Muster definiert, die unterschiedliche Nebensatzstrukturen erfassen. Tabelle 3 zeigt die verwendeten Muster.

Tabelle 3: Muster für die Erkennung von indirekter Wiedergabe. (W-Wort = Wiedergabewort)

Muster 1: Subjunktorsatz bzw. Nebensatz mit W-Fragewort						
	W-Wort	Komma/ Doppel- punkt	Subj./ W-Wort			finites Verb
er	sagte	,	daß		diese Strafe wiederkehren	sollte
die	Frage	:	was		ihr	fehle?
Muster 2a: zu + Infinitiv						
	W-Wort		optional: Komma/ Doppelpunkt		zu	Infinitiv
er	befahl	ihr	,	ein Roß	zu	besteigen
er	glaubte			ein hämisches Lächeln	zu	bemerken
Muster 2b: zu-Infinitiv (verschmolzen)						
	W-Wort			optional: Komma/ Doppelpunkt		zu-Infinitiv
sie	beschlossen		einmütig	,	hier den Abend	abzuwarten
ich	wünschte		wieder			dazubleiben

12 Im DFG-Projekt, in welchem auch nicht-fiktionale Texte bearbeitet werden, wurde das Annotationssystem dahingehend erweitert, dass diese Formen auch als indirekte Wiedergabe gelten.

Muster 3: Verbzweitsatz im Konjunktiv					
	W-Wort	Komma/ Doppelpunkt		Verb im Konjunktiv	
ich	dachte	,	jetzt	sei	es vorbei

Wie man sieht, decken diese Regeln nur die typischsten Fälle von indirekter Wiedergabe ab. Wiedergaben, bei denen die Rahmenformel eingeschoben oder nachgestellt ist, werden z. B. grundsätzlich nicht erfasst, ebenso wenig wie konjunktivische Wiedergaben ohne Rahmenformel. Bei der konkreten Implementierung kommen noch einige weitere Einschränkungen hinzu, die aus grammatischer Sicht nicht notwendig wären. So wird beispielsweise die Erkennung abgebrochen, wenn nach dem ersten gefundenen Komma ein weiteres folgt, bevor die Erkennung abgeschlossen ist. Dies verhindert, dass Nebensätze, die erst deutlich hinter dem Wiedergabewort auftreten, fälschlicherweise als indirekte Wiedergabe aufgefasst werden (z. B. Er *rief* nach ihr, als es so spät war, *dass sie gehen mussten*.). Allerdings wird damit auch verhindert, dass indirekte Wiedergaben mit eingeschobenen Relativsätzen erfasst werden (z. B. Er rief, dass die Tür, die ins Haus führte, verschlossen sei.). Selbstverständlich wäre es denkbar, exaktere Regeln zu entwickeln, um dieses Problem zu vermeiden. Naheliegend wäre an dieser Stelle beispielsweise, mit Hilfe einer automatischen Analyse der Satzstruktur (Parsing), Relativsätze zu identifizieren und als Sonderfall zu behandeln. Ein solches System kann jedoch rasch sehr komplex werden. Hinzu kommt, dass die automatische grammatische Analyse zusätzliche Werkzeuge benötigt, die rechenintensiv sind und keinesfalls immer verlässliche Ergebnisse liefern. Es ist darum bei der Entwicklung regelbasierter Systeme eine Abwägung zwischen Aufwand und Nutzen, wie weit man gehen möchte, um alle Eventualitäten abzudecken, zumal natürliche Sprache oft unerwartete Strukturen zeigt. Die hier vorgestellten Erkenner sind, wie bereits mehrfach betont, nur Prototypen, die eine erste Annäherung implementieren.

In allen Fällen fungiert das Wiedergabewort als Anker – ist kein Wiedergabewort identifiziert, so wird auch keine Struktur erkannt. Aufgrund dieses großen Einflusses wurde getestet, wie eine Veränderung der Wiedergabewort-Liste sich auf die Erkennungsrate auswirkt. Zu diesem Zweck wurde mit einer Liste nur der besten Wiedergabewörter (*penalty*-Wert 0) begonnen und diese schrittweise erweitert. In jedem Schritt wurden Precision, Recall und F-Score des Erkenners ausgewertet.

Das Grundproblem ist klar: Ist die Liste zu kurz, werden viele Fälle, die von ungewöhnlicheren Wörtern eingeleitet werden, nicht erfasst. Ist die Liste zu lang, wird die Erkennung unscharf. Der größte Verbesserungssprung lässt sich

von *penalty*-Wert 0 zu *penalty*-Wert 1 beobachten – in diesem Schritt erhöht sich die Menge Wiedergabewörter in der Liste von 89 auf 306 und der F-Score verbessert sich von 0,31 auf 0,65. Fügt man nach diesem Punkt weitere Wörter hinzu, sind sowohl die positiven als auch die negativen Veränderungen im Vergleich zu diesem Sprung sehr gering (bester F-Score mit Wörtern bis *penalty*-Wert 2: 0,68; schlechtester F-Score bei Verwendung der gesamten Liste: 0,66). Zudem ist interessant zu beobachten, dass sich die einzelnen Texte durchaus unterschiedlich verhalten: Bei einigen verbessert sich der F-Score kontinuierlich, bei anderen führt schon eine Erweiterung der Liste auf Wörter mit *penalty*-Wert 2 zu einer Verschlechterung. Für das gesamte Erzähltextkorpus ergibt sich das beste Ergebnis mit einer Wiedergabewortliste von mittlerer Länge, die Wörter bis zu einem *penalty*-Wert von 2 einschließt. Diese Wortliste wurde auch für die finale Auswertung des regelbasierten Erkenners verwendet.

4.1.3 Regelbasierte Erkennung von direkter, freier indirekter und erzählter Wiedergabe

Die Strategien zur Erkennung der restlichen drei Wiedergabetypen sollen an dieser Stelle nur kurz umrissen werden.

Zur Erkennung von direkter Wiedergabe wurde hauptsächlich ein typographisches Merkmal genutzt: das Vorhandensein von Anführungszeichen. Allerdings ist diese naheliegende Strategie nicht so verlässlich, wie man vielleicht vermuten könnte. Zum einen gibt es unterschiedliche typographische Umsetzungsmöglichkeiten für Anführungszeichen, zum anderen ist deren Gebrauch deutlich weniger stark reglementiert als etwa die Verwendung von Satzendepunkten oder Kommata (vgl. Wehde 2000: 100). Nicht nur sind die Konventionen in historischen Texten oft anders als die heutigen,[13] gerade bei literarischen Texten kommt hinzu, dass Anführungszeichen aus künstlerischen Gründen bewusst weggelassen werden, um einen Effekt der Unmittelbarkeit zu erzielen. Dies ist sogar so verbreitet, dass in einigen literaturwissenschaftlichen Darstellungen direkte Wiedergabe ohne Anführungszeichen als eigene Wiedergabekategorie beschrieben wird (vgl. z. B. Martínez & Scheffel 2016, Leech & Short 2013). Um zu verhindern, dass die Erkennung bei fehlenden oder unkonventionell gesetzten Anführungszeichen vollständig versagt, wurde zusätzlich mittels der Wiedergabewort-Liste

13 Zu verschiedenen Konventionen bei der typographischen Auszeichnung von Zitaten in wissenschaftlichen Texten des Frühneuhochdeutschen vgl. Habermann in diesem Heft.

und einer Mustererkennung versucht, die Rahmenformeln direkter Wiedergabe zu erfassen, welche ebenfalls als Indikatoren fungieren können.

Da erzählte Wiedergabe in sehr diversen Formen auftreten kann, erfolgte ihre Erkennung rein aufgrund von lexikalischen Indikatoren. Auch hier wurde die Wiedergabewort-Liste herangezogen: Alle Wiedergabewörter, die in einem Text auftreten und nicht bereits als Teile von indirekter Wiedergabe oder Teile der Rahmenformel von direkter Wiedergabe identifiziert worden sind, werden als Indikatoren für erzählte Wiedergabe interpretiert.

Am schwierigsten ist die Implementierung eines regelbasierten Ansatzes für die Erkennung von freier indirekter Wiedergabe (‚erlebter Rede'). Diese Form ist definiert als die Überlagerung von Erzähler und Figurenstimme, ihre Abgrenzung ist auch für den Menschen schwierig und sie weist keine klaren typographischen, lexikalischen oder grammatischen Merkmale auf. Zwar gibt es Indikatoren, diese sind jedoch deutlich seltener und weniger verlässlich als die für die anderen Wiedergabetypen. Es wurde darum ein Ansatz gewählt, bei dem jedem Satz eines Textes Wertungspunkte zugewiesen werden. Für jedes Merkmal, das für freie indirekte Wiedergabe spricht (z. B. emphatische Satzzeichen wie Ausrufe- oder Fragezeichen oder Wörter, die auf die Perspektive der Figur hinweisen, wie *heute, jetzt, hier*), werden Wertungspunkte hinzugezählt. Für Merkmale, die gegen freie indirekte Wiedergabe sprechen (z. B. Anzeichen für direkte Wiedergabe wie Personalpronomen der 1. oder 2. Person und Anführungszeichen) werden Punkte abgezogen. Wenn der Satz eine bestimmte Punkteschwelle überschreitet, wird er als freie indirekte Wiedergabe klassifiziert. Diese Methode erwies sich als ziemlich unbefriedigend, zumal die Festlegung der Punktewerte nur durch Versuch und Irrtum möglich war. Auch bei expliziter Anpassung an das Erzähltextkorpus waren die Ergebnisse des Erkenners deutlich schlechter als für alle anderen Wiedergabetypen. Hier zeigt sich die grundsätzliche Schwäche eines regelbasierten Ansatzes in Fällen, in denen keine klaren Regeln erkennbar sind. Gerade freie indirekte Wiedergabe ist damit ein Kandidat für das maschinelle Lernen, auf das im Folgenden eingegangen wird.

4.2 Maschinelles Lernen

4.2.1 Grundprinzip

Es gibt unterschiedliche Typen maschinellen Lernens.[14] In dieser Studie wurde sogenanntes ‚überwachtes' Lernen durchgeführt. Die Grundidee ist folgende: Dem Algorithmus wird eine Menge von Beispielen präsentiert, für die die gewünschte Klassifizierung bereits vorliegt. In unserem Fall sind dies Sätze des Erzähltextkorpus, bei denen aufgrund der manuellen Annotation klar ist, ob sie Wiedergabe enthalten und wenn ja, welchen Typ. Zusätzlich zu der Klassifizierung werden dem Algorithmus weitere Merkmale geliefert, die jeden Satz beschreiben (z. B. Satzlänge), die sogenannten ‚Attribute'. Mit diesen Informationen wird in der Trainingsphase ein Modell aufgebaut, das es dann ermöglicht, weiteren Sätzen, bei denen zwar die Attribute vorliegen, nicht aber die Kategorie bekannt ist, automatisch eine wahrscheinliche Kategorie zuzuweisen.

Die Ergebnisse werden maßgeblich beeinflusst durch den Typ von Algorithmus, der für den Aufbau des Klassifikationsmodells verwendet wird, sowie durch die Auswahl der Attribute. Für die Studie wurde der Algorithmus Random Forest gewählt (vgl. Breiman 2001) in der Implementierung von Liaw & Wiener (2002).[15] Dieser Algorithmus arbeitet gut mit relativ geringen Mengen von Trainingsbeispielen.

Für jeden Wiedergabetyp wird ein eigenes Modell trainiert, das jeweils für einen gegebenen Satz entscheidet, ob er den Wiedergabetyp enthält oder nicht. Die verwendeten Attribute sind in allen Fällen identisch, um für die Studie eine bessere Vergleichbarkeit zu erreichen. Insgesamt wurden 80 Attribute verwendet, die sich in zwei Gruppen einteilen lassen:

- Allgemeine Attribute:
  - Satzlänge
  - Anteil von Wörtern mit einer bestimmten Wortart pro Satz (also z. B. Anteil der Nomen/Verben/Adjektive etc.). Um diese Werte zu erhalten, wurden in einem Vorverarbeitungsschritt die Wörter jedes Satzes mit dem Tree Tagger (vgl. Schmid 1994, Schmid 1995) automatisch mit morphologischen Kategorien versehen. Daraus ergeben sich insgesamt 54 Attributwerte für die 54 Wortarten, die das vom Tree Tagger verwendete

14 Für einen Überblick vgl. z. B. Witten, Frank & Hall (2011) oder die Einführungskapitel in Chollet (2018).

15 Parametereinstellungen für den Random-Forest-Algorithmus waren: Anzahl der Bäume (rTreeN): 500; Attributauswahl pro Knoten: 8.

Stuttgart-Tübingen-Tagset (STTS, vgl. Schiller et al. 1999) unterscheidet. Hinzu kommen weitere 7 Attribute, die sich aus Zusammenfassungen der STTS-Tags ergeben.[16]

- Attribute, die aufgrund von Hypothesen gewählt wurden, was gute Indikatoren für bestimmte Wiedergabetypen oder Wiedergabe im Allgemeinen sein könnten:
  - Anteil von ‚emphatischen' Satzzeichen (Fragezeichen, Ausrufezeichen, Gedankenstrich)
  - Anteil von Personalpronomen der 1. und 2. Person
  - Anteil von Personalpronomen der 3. Person
  - Anteil von Verben im Konjunktiv
  - Anteil von Wiedergabewörtern (in verschiedenen Konfigurationen)
  - Anteil von Personennamen
  - Information, ob der Satz genau am Anfang oder am Ende eines Abschnitts steht

Eine Schwierigkeit beim maschinellen Lernen war, dass in allen Fällen wesentlich weniger Sätze verfügbar waren, die eine bestimmte Wiedergabeform enthielten als solche, die diese nicht enthielten. Im extremsten Fall, freier indirekter Wiedergabe, waren es 110 Sätze, die freie indirekte Wiedergabe enthielten und 2476 Sätze, die keine enthielten, ein Verhältnis von etwa 1:23 (vgl. hierzu auch Abbildung 2 aus dem Abschnitt zur Auswertung der manuellen Annotation). Solche unbalancierten Datensets sind ungünstig für das maschinelle Lernen, da der Algorithmus versucht, den Erfolg für die Gesamtmenge zu optimieren und eine Tendenz hat, im Zweifelsfall die häufigere Klasse zuzuweisen, da die Wahrscheinlichkeit, damit richtig zu liegen, höher ist. Somit werden gerade die Fälle, die eigentlich interessant sind – nämlich das Auftreten von Wiedergabe – benachteiligt. Dieses Problem ist nicht selten bei maschinellen Lernaufgaben und es gibt verschiedene Lösungsstrategien (vgl. z. B. Estabrooks, Jo & Japkowicz 2004). In der Studie wurde ‚Oversampling' gewählt. Das bedeutet, die positiven Fälle (Sätze, die Wiedergabe enthalten) wurden vervielfacht, bis ihre Menge der der negativen Fälle entsprach. Auf dieser Datenmenge wurde der Algorithmus dann trainiert.

16 Die zusammengefassten Kategorien sind: alle Verben, alle finiten Verben, alle infiniten Verben, alle Verben in Partizip Perfekt, Nomen + Eigennamen, alle Adjektive, Partikelwort ‚zu' + Infinitivformen mit eingebettetem ‚zu'.

4.2.2 Attributbewertung

Ein Nebeneffekt des maschinellen Lernens ist es, dass es bei manchen Algorithmen möglich ist, Informationen über das Modell zu gewinnen, welches aus den Trainingsdaten erstellt wurde. Im Fall von Random Forest lässt sich extrahieren, welche Attribute für die Klassifizierung der einzelnen Wiedergabetypen am relevantesten sind. Da das Modell auf der Grundlage von Sprachdaten erstellt wurde, erlaubt dies Beobachtungen, welche erwarteten oder verborgenen Regelhaftigkeiten in Sätzen auftreten, die einen bestimmten Wiedergabetyp enthalten. Maschinelles Lernen kann auf diese Weise rein deskriptiv eingesetzt werden und empirische Erkenntnisse liefern. Tabelle 4 zeigt die 10 relevantesten Attribute zur Erkennung von indirekter Wiedergabe.

Tabelle 4: Die relevantesten Attribute zur Erkennung von indirekter Wiedergabe. Die Zahlen in der Spalte ‚Wert' sind die mean decrease accuracy des Attributs.

Attribut	Wert	Erklärung
stw_word2	0,1429	Anteil Wiedergabewörter, penalty <= 2
stw_word_e1	0,1410	Anteil Wiedergabewörter, penalty = 1
stw_word1	0,1376	Anteil Wiedergabewörter, penalty <= 1
KOUS	0,1270	Anteil unterordnender Konjunktionen mit Satz: *weil, daß, damit, wenn, ob*
verb_inf	0,1194	Anteil Voll-, Hilfs- und Modalverben im Infinitiv
$,	0,1169	Anteil Kommata
stw_word4	0,1163	Anteil Wiedergabewörter, penalty <= 4
verb_konj	0,1129	Anteil Voll, Hilfs- und Modalverben im Konjunktiv[17]
stw_word3	0,1082	Anteil Wiedergabewörter, penalty <= 3
VVINF	0,0969	Anteil Vollverben im Infinitiv

Relevanz wird in diesem Zusammenhang mit einem Maß namens *mean decrease accuracy* gemessen, welches abbildet, wie sehr sich die Klassifikation verschlechtert, wenn das betrachtete Attribut zufällig verändert – also verfälscht – wird (vgl. Breiman & Cutler o. J., Abschnitt ‚Variable Importance'). Je höher die Zahl, desto wichtiger das Attribut. Hierbei wird keine Aussage über Beziehungen zwischen verschiedenen Attributen oder den Wert des Attributs selbst getroffen: Wenn der „Anteil Wiedergabewörter mit einem *penalty*-Wert kleiner oder gleich 2" hier an

17 Die Klassifizierung von Verbformen als Konjunktiv wurde mit Hilfe des RF-Taggers vorgenommen (vgl. Schmid & Laws 2008) unterstützt von einer Liste von Schreibvarianten für Konjunktivformen von ‚sein' (‚sey').

erster Stelle steht, ist unklar, ob dieser Anteil für indirekte Wiedergabe besonders hoch oder besonders niedrig ist, wir wissen nur, dass es hilfreich ist, den Wert zu kennen. Wir können allerdings aufgrund unseres Vorwissens über indirekte Wiedergabe und anhand der Betrachtung der Beispiele vermuten, dass er tatsächlich hoch ist, zumal sich dieser Indikator ja schon bei der regelbasierten Erkennung von indirekter Wiedergabe als besonders geeignet erwiesen hat. Daneben finden sich auf den ersten Plätzen Konjunktionen, Kommata (vermutlich als Signal für einen Nebensatz) und Verben im Konjunktiv – genau die Indikatoren, die auch in den linguistischen Untersuchungen hervorgehoben werden. Die Relevanz von „Vollverben im Infinitiv" könnte von der *zu*-Konstruktion herrühren.

Interessant ist im Gegenzug ein Blick auf eine Kategorie, bei der es im Vorfeld noch nicht so viele Erwartungen gab und für die die Formulierung von Regeln besonders schwerfiel: freie indirekte Wiedergabe. Tabelle 5 zeigt wiederum die 10 relevantesten Attribute.

Tabelle 5: Die relevantesten Attribute zur Erkennung von freier indirekter Wiedergabe.

Attribut	Wert	Erklärung
ADV	0,2330	Anteil Adverbien: *schon, bald, doch, …*
VVFIN	0,2092	Anteil finite Vollverben
noun	0,1641	Anteil Nomen und Eigennamen
len	0,1331	Satzlänge
VAFIN	0,1321	Anteil finite Hilfsverben
NN	0,1273	Anteil Nomen
$.	0,1080	Anteil satzbeendende Interpunktion: *.?!;:*
verb	0,1017	Anteil Verben insgesamt
per12	0,1010	Anteil Personalpronomina der 1. und 2. Person
PDS	0,0981	Anteil substituierende Demonstrativpronomina: *dieser, jener, …*

An erster Stelle steht „Anteil Adverbien". Da diese Gruppe im Stuttgart-Tübingen-Tagset Wörter wie *schon, bald* und *doch* umfasst, kann man vermuten, dass diese in den Trainingssätzen die Funktion von Modalitäts- und Subjektivitätsmarker erfüllen, welche auch in der Narratologie als Kennzeichen von freier indirekter Wiedergabe identifiziert wurden (vgl. z. B. Fludernik 1993). Allerdings ist dieses Ergebnis mit Vorsicht zu behandeln, da die Kategorie ADV bei Schiller et al. nur definiert ist als „reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen" (Schiller et al. 1999: 56), was aus linguistischer Perspektive recht unscharf ist (vgl. hierzu Hirschmann 2015: 200–203). Desweiteren scheint die Satzlänge eine Rolle zu spielen, was sich mit Vermutungen decken könnte, dass Sätze mit freier indirek-

ter Wiedergabe eher kurz sind. Die Bedeutung anderer hoch bewerteter Attribute, wie z. B. „Anteil Vollverben“ und „Anteil Nomen und Eigennamen“ ist weniger offensichtlich, könnte aber eventuell auf eine Tendenz zu einem schlichten, der Mündlichkeit angenäherten Stil hindeuten.

Es ist zu beachten, dass diese Auswertung auf einer sehr kleinen Datenmenge beruht, da das Erzähltextkorpus nur wenig freie indirekte Wiedergabe enthält, die zudem fast ausschließlich auf einen Text konzentriert ist. Dies kann dazu geführt haben, dass mehr der Autorenstil als Kennzeichen freier indirekter Wiedergabe abgebildet werden. Für tragfähige Befunde über die Oberflächenpräferenzen von freier indirekter Wiedergabe müsste deutlich mehr Datenmaterial herangezogen werden. Dennoch zeigen diese Beispiele, dass maschinelles Lernen auch jenseits des Ziels der automatischen Erkennung interessante Ansätze für empirische Studien liefern kann, gerade wenn man als Attribute sehr allgemeine Merkmale wie Wortarten wählt und damit wenig interpretatorische Vorannahmen macht.

5 Ergebnisse der automatischen Erkennung

Bei der Auswertung der automatischen Erkennung wurden für jeden Wiedergabetyp die Ergebnisse der regelbasierten Erkennung und die Ergebnisse, die mit maschinellem Lernen erzielt wurden, mit den manuellen Annotationen verglichen. Die hier wiedergegebenen Werte für das maschinelle Lernen wurden mit 10-facher Kreuzvalidierung erzielt. Das bedeutet, dass jeweils neun Teile des manuell annotierten Korpus als Trainingsmaterial verwendet wurden, um ein Modell zu erstellen, mit dem der zehnte Teil klassifiziert wurde. Dieses beim maschinellen Lernen weit verbreitete Vorgehen ermöglicht es, das vorhandene Trainingsmaterial optimal zu nutzen (jeder manuell annotierte Satz wird zum Aufbau eines Modells verwendet), aber gleichzeitig nie ein Modell auf Sätze anzuwenden, mit denen es trainiert wurde, was dazu führen würden, dass die Ergebnisse wesentlich besser wären als auf unbekanntem Material, so dass der Erfolg des Modells grob überschätzt würde (vgl. z. B. Witten, Frank & Hall 2011: 152–154).

Tabelle 6: Auswertungsbeispiel. Ergebnisse der manuellen Annotation (Gold Standard), der regelbasierten Erkennung und des maschinellen Lernens für 7 Sätze aus Bernhardi: Belinde.

	Enthält der Satz indirekte Wiedergabe?	Manuell	ML	Regel
1	Lächelnd nahm Belinde die Kleider und legte sie an; bald darauf erschien ihr Vater und **befahl ihr, ihm zu folgen**.	Ja	Nein	Ja
2	Laßt mich die Mutter nur noch einmal umarmen, sagte Belinde, daß ich ihren Segen mit mir hinweg nehme.	Nein	Nein	Ja
3	Schweig! rief der Vater, nimmer sollst du sie wieder sehen.	Nein	Nein	Nein
4	Die Thörin! sie begünstigte deine Wahl, so büße sie denn auch mit dir.	Nein	Nein	Nein
5	Ihr kennt den Jüngling nicht, den ihr verwerft, sagte Belinde.	Nein	Nein	Nein
6	Er ist meines Feindes Sohn, rief der Vater, das ist mir genug, ihn zu verwerfen; du widersetzest dich, dem Manne die Hand zu reichen, den ich für dich erwählt habe, das genügt mir, dich zu bestrafen; du sollst deinen Ungehorsam, von meinem Angesicht verbannt, mit tausend Thränen büßen.	Nein	Nein	Nein
7	Belinde mußte ihrem Vater folgen, er **befahl ihr, ein Roß zu besteigen**, und er selbst und einige Diener begleiteten sie.	Ja	Ja	Ja

Mit den Werten Precision, Recall und F-Score wird die Erkennungsqualität gemessen, die mit den unterschiedlichen Methoden erreicht werden konnte. Zur Veranschaulichung zeigt Tabelle 6 die konkreten Ergebnisse für die Erkennung indirekter Wiedergabe in einem Textausschnitt aus *Bernhardi: Belinde*. In Satz 2 klassifiziert der regelbasierte Erkenner den *dass*-Satz aufgrund der fehlenden Anführungszeichen fälschlicherweise als indirekte Wiedergabe mit Wiedergabewort *sagte*. In Satz 1 erkennt der ML-Erkenner die indirekte Wiedergabe nicht, obgleich eine ähnliche Struktur in Satz 7 korrekt erkannt wurde. Aufgrund der Natur von maschinellem Lernen ist es nicht möglich, genau zu sagen, woran dies liegt, da jedoch immer der komplette Satz betrachtet wird, fließen viele Störfaktoren mit ein. In diesem Kurzabschnitt sind die Erfolgswerte für die Erkennung indirekter Wiedergabe für die beiden Techniken fast komplementär: Regelbasierte Erkennung: Precision: 0,66 (2 von 3 gefundenen Vorkommen sind korrekt); Recall 1,0 (alle Vorkommen gefunden); ML-Erkennung: Precision 1,0 (alle gefundenen Vorkommen korrekt); Recall 0,5 (nur eines von zwei Vorkommen gefunden).

Als weiteres Maß der Erfolgsbewertung wurde zudem die Korrelation zwischen dem in einem Text vorausgesagten Anteil von Wiedergabe und dem tatsächlichen Anteil über mehrere Texte hinweg berechnet. Der ideale Wert ist auch hier 1,0. Bei dieser Auswertung wird vor allem die Stabilität der Methode gemessen: Es ergeben sich schlechte Korrelationswerte, wenn für manche Texte der Anteil über- und für andere unterschätzt wird, eine *konsistente* Über- oder Unter-

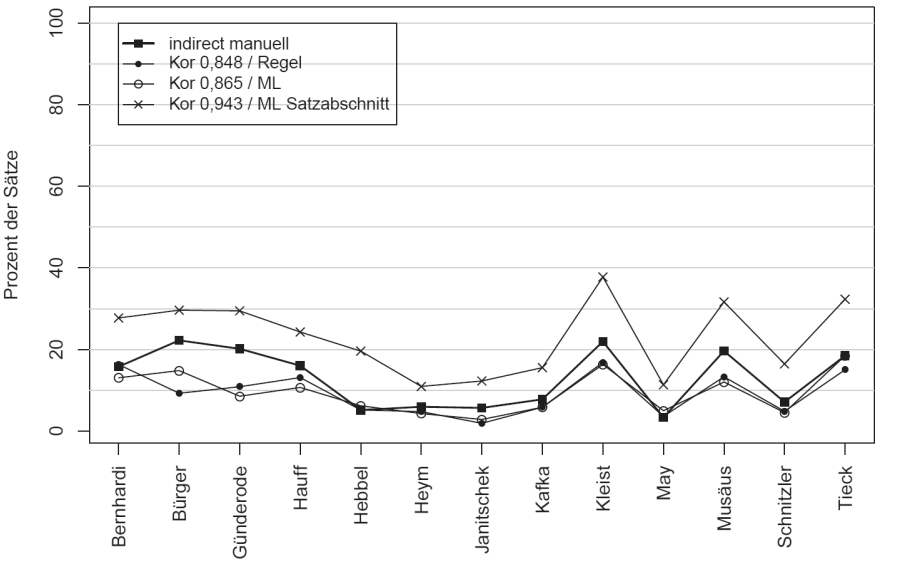

Abbildung 3: Anteile indirekter Wiedergabe pro Text – manuelle Annotation vs. verschiedene automatische Methoden.

schätzung kann jedoch durchaus gute Korrelationswerte erbringen. Zudem ist diese zweite Auswertung großzügiger als die erste, da hier die Wiedergabe nicht unbedingt im richtigen Satz erkannt werden muss. Es geht lediglich darum, eine grobe Abschätzung zu haben, wie sich die Wiedergabe-Anteile von einzelnen Texten zueinander verhalten. Dies kann nützlich sein, wenn z. B. Entwicklungslinien von Wiedergabeformen über die Zeit hinweg betrachtet werden, und es nur auf einen relativen Vergleich zwischen Texten ankommt. Abbildung 3 zeigt exemplarisch die Korrelationslinien für die Erkennung von indirekter Wiedergabe.

Die Linie mit den schwarzen Quadraten markiert die Werte der manuellen Annotation – idealerweise würden die anderen Linien dieser genau entsprechen. Wie man sieht, liegen die Voraussagen der regelbasierten Erkennung und der Erkennung mit maschinellem Lernen relativ dicht beieinander. Die vorausgesagten Anteile sind tendenziell immer etwas zu gering und für einige Texte sind die Ergebnisse deutlich schlechter als für andere. Die Korrelationswerte sind 0,85 für die regelbasierte Methode und 0,87 für maschinelles Lernen. Das ist im Vergleich zu den anderen Redewiedergabetypen ein eher schlechtes Ergebnis.

Für indirekte Wiedergabe wurde jedoch noch ein weiteres Experiment mit maschinellem Lernen angestellt. Ausgehend von der Überlegung, dass indirekte Wiedergabe oft nicht einen ganzen Satz umfasst, wurden die Sätze des Korpus

in Satzabschnitte zerlegt. Trennstellen waren Kommata, Doppelpunkte, schließende Anführungszeichen oder die Konjunktion *und*. Zudem muss jeder Satzabschnitt (außer dem letzten) mindestens eine verbale Form enthalten. Diese Satzabschnitte entsprechen also nicht Teilsätzen im grammatischen Sinne, sind jedoch eine grobe Annäherung. Die mit Kreuzen markierte Linie in Abbildung 3 („ML Satzabschnitt“) zeigt die Ergebnisse, wenn der maschinelle Lernalgorithmus auf diesen Satzabschnitten statt auf ganzen Sätzen trainiert wird. Wie man sieht, sind die Voraussagen deutlich schlechter geworden – der Anteil indirekter Wiedergabe wird nun nicht mehr unter- sondern überschätzt und zwar immer um ca. 10 Prozent. Interessanterweise ist dieser Fehler jedoch sehr konstant, so mit dieser Methode trotzdem der beste Korrelationswert erreicht wird: 0,94. Wenn man nur daran interessiert ist, wie sich die Anteile von Wiedergabe *relativ* zueinander verhalten und vor allem eine stabile Methode möchte, könnte diese Lernmethode also sogar die geeignetste sein.[18]

Tabelle 7 zeigt im Überblick die wichtigsten Ergebnisse im Vergleich zwischen regelbasiertem Ansatz und maschinellem Lernen. Die erfolgreichste Methode wird jeweils als erste genannt.

Tabelle 7: Auswertungsergebnisse für die automatische Erkennung.

	Genauigkeit				**Korrelation**	
	Technik	**Precision**	**Recall**	**F-Score**	**Technik**	**Wert**
Direkt	ML	0,88	0,85	0,87	ML	0,95
	Regelbasiert	0,81	0,87	0,84	Regelbasiert	0,76
Frei indirekt	ML	0,63	0,29	0,40	ML	0,97
	Regelbasiert	0,24	0,44	0,31	Regelbasiert	0,77
Indirekt	Regelbasiert	0,81	0,62	0,71	ML Satzabschnitte	0,94
	ML	0,62	0,47	0,53	Regelbasiert	0,85
Erzählt	Regelbasiert	0,51	0,64	0,57	ML	0,95
	ML	0,56	0,45	0,50	Regelbasiert	0,84

Am besten gelang die automatische Erkennung für direkte Wiedergabe. Bei der Bewertung, welche Methode am günstigsten ist, spielt es eine große Rolle, ob Anführungszeichen zur Markierung konsistent verwendet werden. Ist dies der Fall, kann mit Hilfe regelbasierter Methoden eine sehr genaue Erkennung mit Erfolgsraten von bis zu 100 % geleistet werden. Beim Erzähltextkorpus, das

18 Das Lernen auf Satzabschnitten wurde auch für erzählte Wiedergabe getestet, weil diese ebenfalls häufig nur Teilsätze umfasst. Hier zeigte sich dieser Effekt jedoch nicht, die Ergebnisse waren relativ ähnlich wie beim Lernen auf Sätzen.

sowohl Texte mit als auch ohne Anführungszeichen enthält, erweist sich jedoch die Erkennung mit Hilfe von maschinellem Lernen als überlegen, die fehlende Anführungszeichen besser ausgleichen kann: Sie erreicht mit 0,87 den maximalen F-Score für das Gesamtkorpus. Bei der Betrachtung der relativen Anteile von Wiedergabe kann mit maschinellem Lernen zudem eine sehr viel bessere Korrelation erreicht werden, was damit zusammenhängt, dass die Qualität der regelbasierten Erkennung abhängig von der Verwendung von Anführungszeichen stark schwankt. Es wurde auch getestet, wie die Ergebnisse sind, wenn das Vorkommen von Anführungszeichen beim maschinellen Lernen vollkommen ignoriert wird, indem das Attribut entfernt wird, das dieses abbildet. Auch dann liefert die Methode immer noch gute Ergebnisse (F-Score von 0,81), tatsächlich verbessern sich die Ergebnisse für einzelne Texte ohne Anführungszeichen sogar leicht. Maschinelles Lernen ist also auf jeden Fall bei Texten vorzuziehen, bei denen die Markierung von direkter Wiedergabe nicht vorhanden oder nicht konsistent ist.

Indirekte Wiedergabe ist der Typ, der sich am zweitbesten erfassen lässt. Hier sind die regelbasierten Methoden mit einem F-Score von 0,71[19] (gegenüber nur 0,53 beim maschinellen Lernen) klar überlegen. Bei einem stark strukturell definierten Wiedergabetyp haben diese Methoden also durchaus Vorteile. Wie oben ausgeführt, ist bei der Korrelation der relativen Anteile dennoch eine maschinelle Lernmethode, allerdings auf Basis von Satzabschnitten statt Sätzen, überlegen.

Bei erzählter Wiedergabe liefert ebenfalls der regelbasierte Ansatz die besseren Ergebnisse, allerdings ist der Abstand zum maschinellen Lernen geringer (F-Score 0,57 vs. 0,50) und die Erkennungsrate insgesamt nicht sehr gut. Bei der Korrelation liegt auch hier maschinelles Lernen vorne.

Bei freier indirekter Wiedergabe ist die Überlegenheit von maschinellem Lernen am offensichtlichsten. Zwar ist auch mit dieser Methode die Erkennungsrate nicht gut (F-Score von 0,4), jedoch eine deutliche Verbesserung gegenüber dem F-Score von 0,31 für den regelbasierten Ansatz und auch bei den Korrelationswerten liegt maschinelles Lernen deutlich vorne. Es scheint, als wäre maschinelles Lernen die Strategie, mit der man diesen Wiedergabetyp, der keine Indikatoren besitzt, die häufig, stabil und eindeutig genug sind, um gute Regeln zu formulieren, am ehesten automatisch erfassen könnte. Für bessere Ergebnisse wäre allerdings deutlich mehr Trainingsmaterial nötig als in der Studie vorhanden war.

19 Dieser Wert ist etwas höher als der F-Score, der im Abschnitt über die regelbasierte Erkennung genannt wurde (0,68). Dies liegt daran, dass die Gesamtauswertung auf Satzbasis erfolgt („Enthält dieser Satz (mindestens eine) Instanz von indirekter Wiedergabe?“), während die Auswertung zuvor das Vorkommen einzelner Instanzen verglichen hat, also strenger war.

Zusammenfassend lässt sich sagen, dass regelbasierte Methoden zwar bei direkter, erzählter und vor allem indirekter Wiedergabe Vorteile bringen können, was die Korrektheit der Erkennung angeht, jedoch maschinelle Lernverfahren bei der Korrelation der relativen Anteile immer am besten abschneiden. Es scheint, als wären diese Methoden dann zu bevorzugen, wenn man vor allem an Stabilität und groben Verlaufslinien interessiert ist. Zudem reagieren sie robuster auf unerwartete Strukturen in den Texten (z. B. fehlende Anführungszeichen).

Da sich gezeigt hat, dass die beiden Erkennertypen oft unterschiedliche Sätze annotieren, stellt sich zudem die Frage, ob es einen Vorteil bringt, beide zu kombinieren. Zu diesem Zweck wurden zum einen alle Instanzen gezählt, die sowohl von dem regelbasierten Erkenner als auch durch maschinelles Lernen identifiziert wurden (die Schnittmenge) und zum anderen alle Instanzen, die von mindestens einem von beiden Erkennern annotiert wurden (die Vereinigungsmenge). Die Auswertungen ergaben, dass auf diese Weise zwar entweder die Precision (bei der Schnittmenge) oder der Recall (bei der Vereinigungsmenge) verbessert werden können, sich in der Gesamtbetrachtung allerdings weder für den F-Score noch für die Korrelation nennenswerte Vorteile ergeben. Es ist gut möglich, dass eine Kombination von Regeln und maschinellem Lernen trotzdem fruchtbar wäre, allerdings müsste diese komplexer sein als die bloße Verrechnung der Ergebnisse beider Methoden.

6 Ausblick

Die Ausführungen haben gezeigt, dass die automatische Erkennung von Redewiedergabe zwar nicht trivial, aber auch kein hoffnungsloses Unterfangen ist. Für die Formen direkte und indirekte Wiedergabe konnten bereits mit den recht simplen Erkennern, die in der Studie implementiert wurden, brauchbare Ergebnisse erzielt werden. Es zeigen sich offensichtliche Ansätze zur Verbesserung, sei es durch eine detailliertere syntaktische Verarbeitung bei regelbasierten Ansätzen, sei es durch eine gezieltere Attributauswahl und Tests von anderen Lernalgorithmen beim maschinellen Lernen. Seit der Durchführung der Studie wurden vor allem im Bereich der automatischen Erkennung von direkter Wiedergabe mit maschinellem Lernen einige Fortschritte erzielt (z. B. Schöch et al. 2016 für französische und Jannidis et al. 2018 für deutsche Romane).

Weiterverfolgt wird das Thema insbesondere in dem laufenden DFG-Projekt (Homepage: www.redewiedergabe.de). Ein wesentlicher Unterschied zu der Studie ist, dass dort ein deutlich größeres, manuell annotiertes Redewiedergabe-Korpus aufgebaut wird. Dieses umfasst die Jahre 1840–1920 und enthält zu

gleichen Teilen Ausschnitte aus fiktionalen Erzähltexten und nicht-fiktionalen Texten (Zeitungen und Zeitschriften). Auf diese Weise kann das Phänomen Redewiedergabe auf einer allgemeineren Basis untersucht werden und auch Textsortenvergleiche sind möglich.

Wie bereits in Abschnitt 3.1 erwähnt, wird das neue Korpus auch von mehreren Personen annotiert, was nicht nur die Verlässlichkeit erhöht, sondern zudem eine Abschätzung ermöglicht, wie schwierig die Erkennung, Klassifizierung und Abgrenzung von Wiedergabe für den Menschen ist. Es zeigt sich bereits, dass die Schwierigkeit deutlich von Text zu Text und abhängig vom Wiedergabetype variiert. Menschen erzielen die beste Übereinstimmung bei der Annotation direkter Wiedergabe, gefolgt von indirekter, gefolgt von erzählter Wiedergabe – es ergibt sich also eine ganz ähnliche Rangfolge wie beim Erfolg der automatischen Annotation. Dieser Aspekt ist noch kaum erforscht, gibt aber wertvolle Hinweise, welche Erwartungen man an die Verlässlichkeit automatischer Erkennung stellen kann.

Die Annotationsrichtlinien wurden gegenüber denen in Brunner (2015) präzisiert und z. T. erweitert. So werden in Rahmenformeln der direkten und indirekten Wiedergabe nun auch der Sprecher und der Wiedergabeeinleiter annotiert. Letzteres unterstützt die Erweiterung und Verfeinerung der Wiedergabewortliste und ermöglicht gleichzeitig Studien zu der Entwicklung des Inventars von Wiedergabewörtern.

Die Wiedergabe-Erkennung soll verbessert und dann auf weitere Texte aus dem Untersuchungszeitraum angewendet werden. Dies führt zum eigentlichen Kernziel quantitativer Sprach- und Literaturwissenschaft: der Beobachtung von größeren Entwicklungslinien auf einer großen Textbasis. Verschiedene offene narratologische und linguistische Forschungsfragen sollen so untersucht werden, z. B.: Welche Entwicklungen in der Verwendung von Redewiedergabe lassen sich im Untersuchungszeitraum beobachten? Welche Rolle spielen Textsortenunterschiede bei der Entwicklung von Redewiedergabeformen? Wie kommt die Dynamik im Bestand an Verben zustande, die als Redeeinleiter in bestimmten argumentstrukturellen Mustern auftreten?

Sowohl das reich annotierte Korpus als auch die im DFG-Projekt entwickelten automatischen Erkenner werden am Ende des Projekts (2020) der Forschungsgemeinschaft zur Verfügung gestellt werden.

Literatur

Breiman, Leo (2001): Random Forests. *Machine Learning* 45(1), 5–32.

Breiman, Leo & Adele Cutler (o. J.): *Random Forests*. https://www.stat.berkeley.edu/~breiman/RandomForests (6.12.2018).

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

Brunner, Annelen, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu & Lukas Weimer (2018): Projektvorstellung: Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse. *Konferenzabstracts der DHd 2018. Kritik der digitalen Vernunft*, 458–460. Köln. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf (6.12.2018).

Chiticariu, Laura, Yunyao Li & Frederick R Reiss (2013): Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 827–832. Seattle. http://www.aclweb.org/anthology/D13-1079 (6.12.2018).

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

Cunningham, Hamish, Valentin Tablan, Angus Roberts & Kalina Bontcheva (2013): Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology* 9(2). https://doi.org/10.1371/journal.pcbi.1002854 (6.12.2018).

Eisenberg, Peter (2013): *Grundriss der deutschen Grammatik. Bd. 2: Der Satz / unter Mitarb. von Rolf Thieroff*. 4. aktualisierte und überarb. Aufl. Stuttgart, Weimar: Metzler.

Estabrooks, Andrew, Taeho Jo & Nathalie Japkowicz (2004): A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence* 20(1). 18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x (6.12.2018).

Fabricius-Hansen, Cathrine (2001): Wessen Redehintergrund? Reportive Modalität aus textorientierter kontrastiver Sicht (Deutsch – Norwegisch – Englisch). *Reports of the project Languages in Contrast (Språk in kontrast)* 14. 1–27.

Fludernik, Monika (1993): *The fictions of language and the languages of fiction. The linguisitic representation of speech and consciousness*. London, New York: Routledge.

Genette, Gérard (2010): *Die Erzählung* (UTB 8083). 3., durchges. und korrigierte Aufl. Paderborn: Fink.

Gius, Evelyn & Janina Jacke (2016): *Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets (Version 2.0)*. Hamburg. http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf (6.12.2018).

Harras, Gisela, Edeltraud Winkler, Sabine Erb & Kristel Proost (2004): *Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch*. Berlin u. a.: De Gruyter.

Helbig, Gerhard & Joachim Buscha (2011): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin u. a.: Langenscheidt.

Hirschmann, Hagen (2015): *Modifikatoren im Deutschen. Ihre Klassifizierung und varietätenspezifische Verwendung* (Studien Zur Deutschen Grammatik 89). Tübingen: Stauffenburg.

Ide, Nancy & James Pustejovsky (Hrsg.) (2017): *Handbook of Linguistic Annotation*. Vol. 1. Dodrecht: Springer.

Jannidis, Fotis, Leonard Konle, Albin Zehe, Andreas Hotho & Markus Krug (2018): Analysing Direct Speech in German Novels. *Konferenzabstracts der DHd 2018. Kritik der Digitalen*

Vernunft, 114–118. Köln. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf (6.12.2018).

Jockers, Mattew (2013): *Macroanalysis. Digital Methods and Literary History*. Champaign: University of Illinois Press.

Kaufmann, Gerhard (1976): *Die indirekte Rede und mit ihr konkurrierende Formen der Redeerwähnung* (Heutiges Deutsch. Reihe III: Linguistisch-didaktische Untersuchungen des Goethe-Instituts 1). München: Hueber.

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

Liaw, Andy & Matthew Wiener (2002): Classification and Regression by randomForest. *R News* 2/3. 18–22. https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest (6.12.2018).

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze (2008): *Introduction to Information Retrieval*. New York: Cambridge University Press.

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

Moretti, Franco (2005): *Graphs, Maps, Trees. Abstract Models for a Literary History*. London: Verso.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen (1999): *Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset)*. Institut für Maschinelle Sprachverarbeitung (Universität Stuttgart) / Seminar für Sprachwissenschaft (Universität Tübingen). http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf (6.12.2018).

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (6.12.2018)

Schmid, Helmut (1995): Improvements on Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (6.12.2018)

Schmid, Helmut & Florian Laws (2008): Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. *COLING 2008*. Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/papers/Schmid-Laws.pdf (6.12.2018)

Schöch, Christof, Daniel Schlör, Stefanie Popp, Annelen Brunner & José Calvo Tello (2016): Straight talk! Automatic Recognition of Direct Speech in Nineteenth-century French Novels. *Digital Humanities 2016*, 346–353. Jagiellonian University & Pedagogical University, Kraków. http://dh2016.adho.org/static/data/132.html (6.12.2018).

Semino, Elena & Mick Short (2004): *Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing*. London, New York: Routledge.

Stanzel, Franz K. 2008. *Theorie des Erzählens* (UTB 904). 8. Aufl. Göttingen: Vandenhoeck & Ruprecht.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr & Trijntja Pasma (2010): *A Method for Linguistic Metaphor Identification*. Amsterdam, Philadelphia: John Benjamins.

Strecker, Bruno (2009): Er sagte, dass er aus Ulm kommt, komme, käme oder kommen würde? – Mit dass eingeleitete indirekte Redewiedergabe (Teil 1). https://grammis.ids-mannheim.de/fragen/160 (6.12.2018).

Wehde, Susanne (2000): *Typographische Kultur: eine zeichentheoretische und kulturgeschichtliche Studie zur Typographie und ihrer Entwicklung* (Studien und Texte zur Sozialgeschichte der Literatur 69). Tübingen: Niemeyer.

Weinrich, Harald (2005): *Textgrammatik*. 3. Aufl. Hildesheim u. a.: Olms.

Witten, Ian A., Eibe Frank & Mark A. Hall. 2011. *Data Mining. Practical Machine Learning Tools and Techniques*. 3. Aufl. Amsterdam u. a.: Morgan Kaufmann.

Wöllstein, Angelika & Dudenredaktion (Hrsg.) (2016): *Duden. Die Grammatik*. 9. Aufl. Bd. 4. Berlin: Dudenverlag.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker (1997): *Grammatik der deutschen Sprache* (Schriften des Instituts für Deutsche Sprache 7). Bd. 3. Berlin u. a.: De Gruyter.