# Projektvorstellung – Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse

**Brunner, Annelen**
brunner@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

**Engelberg, Stefan**
engelberg@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

**Jannidis, Fotis**
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Tu, Ngoc Duyen Tanja**
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

**Weimer, Lukas**
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

## Einführung

Das laufende DFG-Projekt „Redewiedergabe“ stellt einen Anwendungsfall quantitativer Sprach- und Literaturwissenschaft dar und beschäftigt sich mit dem Phänomen „Redewiedergabe“ auf der Grundlage großer Datenmengen. Zu diesem Zweck wird zum einen ein Korpus manuell mit Redewiedergabeformen annotiert, zum anderen werden Verfahren zur automatischen Erkennung des Phänomens entwickelt. Ziel ist es, Forschungsfragen nach der Entwicklung von Redewiedergabe vor allem im 19. Jahrhundert zu beantworten.

Das Poster präsentiert einen Überblick über das Gesamtprojekt sowie erste Projektergebnisse.

## Stand der Forschung

Sowohl aus linguistischer als auch aus literaturwissenschaftlicher Perspektive ist Redewiedergabe ein interessantes Phänomen. Die Art und Weise, wie die Figurenstimme in die Erzählung eingebunden ist, steht in engem Zusammenhang mit Erzählweise und -haltung, sowie der Konstruktion der erzählten Welt. Folglich wird dem Phänomen in der Erzählforschung viel Aufmerksamkeit geschenkt und es liegen zahlreiche systematische Analysen vor (vgl. z.B. Genette 1998; Martínez / Scheffel 2007). Zu Phänomenen wie der erlebten Rede, dem Bewusstseinsstrom usw. gibt es eine umfangreiche Spezialforschung (Überblick bei McHale 2014). Aus linguistischer Perspektive ist Redewiedergabe vor allem in Bezug auf den Funktionswandel des Konjunktivs im Zusammenhang mit seinem Auftreten in indirekter Rede untersucht worden (vgl. z.B. Übersicht in Ágel 2000). In geringem Umfang sind auch Redewiedergabeverben und ihr Verhältnis zur wiedergegebenen Rede in das Blickfeld der Forschung gerückt (eine kurze Synopse bei Fritz 2005).

Ein Vorbild für die ausführliche, manuelle Annotation von Redewiedergabe ist v.a. Semino / Short 2004. Implementierungen der automatischen Erkennung stammen vor allem aus dem Bereich der Computerlinguistik und werden oft als Vorverarbeitungsschritt für andere Anwendungen durchgeführt (z.B. Wissensextraktion, Sprechererkennung oder dem Aufbau von sozialen Netzwerken literarischer Figuren, vgl. z.B. Krestel / Bergler / Witte 2008; Elson / Dames / McKeown 2010; Iosif / Mishra 2014). Eine literaturwissenschaftlich motivierte Anwendung ist die Untersuchung von Schöch et al. 2016 zur Erkennung von direkter Wiedergabe in französischen Romantexten. Die wichtigste Vorarbeit für das vorgestellte Projekt ist die Studie Brunner 2015, auf deren Ergebnissen es aufbaut. In dieser Studie wurde ein Korpus von 13 Erzähltexten manuell annotiert und Prototypen für die automatische Erkennung (sowohl regelbasiert als auch mit Hilfe von maschinellem Lernen) wurden entwickelt und ausgewertet.

## Datengrundlage, Methodik und Ziele

Das Untersuchungskorpus umfasst die Jahre 1840-1920 und enthält sowohl fiktionale als auch nicht-fiktionale Texte. Der nicht-fiktionale Teil setzt sich zusammen aus Texten des „Mannheimer Korpus Historischer Zeitungen und Zeitschriften“ und der Zeitschrift „Die Grenzboten“ (digitalisiert durch die Staats- und Universitätsbibliothek Bremen), der fiktionale Teil aus Erzählungen der Sammlung der Digitalen Bibliothek (textgrid). So sind sowohl Beobachtungen von Entwicklungen über die Zeit hinweg als auch Vergleiche zwischen Textsorten möglich.

Auszüge aus den Texten werden manuell annotiert. Das in Brunner 2015 vorgestellte und an Kategoriensystemen der Literaturwissenschaft orientierte Annotationssystem wurde für das Projekt erweitert und präzisiert. Es unterscheidet zwischen Wiedergabe von gesprochener Sprache, von Schrift und von Gedanken sowie den Typen direkte Wiedergabe ( *Er sagte: "Ich bin hungrig."*), indirekte Wiedergabe ( *Er sagte, er sei hungrig.*), erzählte

Wiedergabe ( *Er sprach über das Mittagessen.*) und freie indirekte Wiedergabe ('erlebte Rede') ( *Wo sollte er jetzt nur etwas zu essen bekommen?*). Attribute spezifizieren die Annotation (z.B. Verschachtelungstiefe) und markieren Sonderfälle (z.B. nicht-faktische Wiedergabe). Zusätzlich werden Sprecher, Rahmenformeln und die redeeinleitenden Verben bzw. Nomen markiert.

Die Annotatoren arbeiten mit dem im Projekt Kallimachos ( www.kallimachos.de ) von Markus Krug entwickelten Eclipse-basierten Annotationswerkzeug ATHEN ( https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen ), für welches eine spezielle Annotationsoberfläche für Redewiedergabeformen implementiert wurde. Es werden, zumindest in Teilen, Mehrfachannotationen durchgeführt und Annotatorenvergleiche angestellt.

Die zweite Projektphase, welche zum Einreichungszeitpunkt dieses Posters gerade beginnt, umfasst die Entwicklung eines automatischen Erkenners für Redewiedergabeformen. Hierbei dient das manuell annotierte Material als Test- und Trainingsmaterial. Die in Brunner 2015 implementierten Prototypen dienen als Ausgangspunkt, die Implementierung erfolgt unter Nutzung des UIMA-Frameworks sowie in Python. Geplant ist eine Verbesserung des maschinellen Lernens durch Optimierung der Attributauswahl sowie Tests mit verschiedenen Lernalgorithmen (RandomForest, SVM, eventuell Conditional Random Fields und Deep Learning) und verschiedenen Parametereinstellungen. Auch regelbasierte Ansätze sollen weiter verfolgt werden, eventuell auf Grundlage einer aufwendigeren Vorverarbeitung (z.B. Parsing). Zudem ist eine Ergänzung und Verfeinerung einer Liste von Wörtern geplant, die auf Redewiedergabe hinweisen, welche sich bereits in Brunner 2015 als wertvolles Werkzeug bei der automatischen Erkennung erwiesen hat.

Der Redewiedergabe-Erkenner wird dann auf weitere Texte in unserem Untersuchungszeitraum angewendet, um größere Entwicklungslinien beobachten zu können und verschiedene offene narratologische und linguistische Forschungsfragen auf quantitativer Basis zu untersuchen, z.B.: Welche Entwicklungen in der Verwendung und Form 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 gebraucht werden?

Sowohl das manuell annotierte Korpus als auch der automatische Erkenner werden am Ende des Projekts der Forschungsgemeinschaft zur Verfügung gestellt. Es werden dafür sowohl das CLARIN-D-Forschungsdatenrepositorium des Instituts für Deutsche Sprache als auch das DARIAH-DE-Repository genutzt.

# Bibliographie

**Ágel, Vilmos** (2000): "Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts", in: Besch, Werner / Betten, Anne / Reichmann, Oskar (eds.): *Sprachgeschichte.* Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Berlin / Boston: de Gruyter 1855-1903.

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

**Elson, David K. / Dames, Nicholas J. / McKeown, Kathleen** (2010): "Extracting Social Networks from Literary Fiction", in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* 138-147.

**Fritz, Gerd** (2005): *Einführung in die historische Semantik.* Tübingen: M. Niemeyer.

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

**Iosif, Elias / Mishra, Taniya** (2014): "From Speaker Identification to Affective Analysis: A Multi-Step System from Analyzing Children's Stories", in: *Proceedings of the Third Workshop on Computational Linguistics for Literature* 40-49.

**Krestel, Ralf / Bergler, Sabine / Witte, René** (2008): "Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles", in: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)* 2823-2828.

**Martínez, Matías / Scheffel, Michael** (2007): *Einführung in die Erzähltheorie.* München: C. H. Beck.

**McHale, Brian** (2014): "Speech Representation" in: Hühn, Peter / Pier, John / Schmid, Wolf / Schönert, Jörg (eds.): *The living handbook of narratology.* Hamburg: Hamburg University Press 434-446 http://www.lhn.uni-hamburg.de/article/speech-representation [letzter Zugriff 18. September 2017].

**Schöch, Christof / Schlör, Daniel / Popp, Stefanie / Brunner, Annelen / Henny, Ulrike / Calvo Tello, José** (2016): "Straight talk! Automatic Recognition of Direct Speech in Nineteenth-century French Novels", in: *Conference Abstracts. Jagiellonian University & Pedagogical University* 346-353.

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