Das Redewiedergabe-Korpus Eine neue Ressource

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

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

Tu, Ngoc Duyen Tanja
tu@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

Einführung

In diesem Beitrag[1] wird das Redewiedergabe-Korpus (RW-Korpus) vorgestellt, ein historisches Korpus fiktionaler und nicht-fiktionaler Texte, das eine detaillierte manuelle Annotation mit Redewiedergabeformen enthält. Das Korpus entsteht im Rahmen eines laufenden DFG-Projekts und ist noch nicht endgültig abgeschlossen, jedoch ist für Frühjahr 2019 ein Beta-Release geplant, welches der Forschungsgemeinschaft zur Verfügung gestellt wird. Das endgültige Release soll im Frühjahr 2020 erfolgen. Das RW-Korpus stellt eine neuartige Ressource für die Redewiedergabe-Forschung dar, die in dieser Detailliertheit für das Deutsche bisher nicht verfügbar ist, und kann sowohl für quantitative linguistische und literaturwissenschaftliche Untersuchungen als auch als Trainingsmaterial für maschinelles Lernen dienen.

Motivation und verwandte Forschung

Redewiedergabe ist sowohl für die Linguistik als auch die Literaturwissenschaft ein interessanter Untersuchungsgegenstand. Die Repräsentation der Figurenstimme in Erzähltexten hat in der Narratologie viel Aufmerksamkeit erfahren und wurde in zahlreichen Kategoriesystemen abgebildet (vgl. z.B. Genette 2010; Martínez / Scheffel 2016). In der Linguistik besteht ein Interesse an sprachlichen Formen der Redewiedergabe, sowie an Redeeinleitungsverben (vgl. z.B. Hauser 2008, Engelberg 2015).

Detaillierte, manuell annotierte Korpora mit diesem Themenschwerpunkt sind bislang vor allem für das Deutsche sehr rar. Ein Vorbild mit detaillierter, literaturwissenschaftlich motivierter Annotation mehrere Redewiedergabetypen für das Englische ist das Korpus von Semino/Short 2004. Das ebenfalls manuell annotierte DROC-Korpus hat seinen Schwerpunkt auf Figurenreferenzen in Romanen, enthält in diesem Kontext allerdings auch Annotationen direkter Wiedergabe mit Sprecherzuordnung (Krug et al. 2018b). Unser Korpus ist eine direkte Weiterentwicklung des aus 13 Erzähltexten bestehenden Korpus aus Brunner 2015, unterscheidet sich jedoch von diesem vor allem in folgenden Aspekten: Es enthält neben fiktionalen auch nicht-fiktionale Texte, die Annotationen sind durch Mehrfachannotation wesentlich verlässlicher und es ist deutlich umfangreicher (für das Beta-Release ca. 350.000 Tokens vs. 57.000 Tokens in Brunner 2015).

Korpuszusammensetzung

Das RW-Korpus umfasst Textmaterial aus dem Zeitraum 1840-1920. Es beruht auf den folgenden drei Textquellen, aus denen jeweils nur die Texte ausgewählt wurden, die in den Untersuchungszeitraum passen:

- Erzähltexte aus der Digitalen Bibliothek, in TEI-Format konvertiert vom Projekt TextGrid (https://textgrid.de/digitale-bibliothek)
- Texte der Zeitschrift „Die Grenzboten", digitalisiert durch die Staats- und Universitätsbibliothek Bremen und in TEI-Format konvertiert durch das Deutsche Textarchiv (http://www.deutschestextarchiv.de/)
- Das Mannheimer Korpus Historischer Zeitungen und Zeitschriften (MKHZ, https://repos.ids-mannheim.de/mkhz-beschreibung.html), bereitgestellt vom Institut für Deutsche Sprache und in TEI-Format konvertiert durch das Deutsche Textarchiv

Bei der Korpuszusammenstellung sollte eine möglichst große Diversität der enthaltenen Texte erzielt werden. Um dies zu erreichen, setzt sich das Korpus aus Textausschnitten (Samples) zusammen. Diese haben mindestens 500 Wörter für fiktionale Texte bzw. 200 Wörter für nicht-fiktionale Texte – mit dieser großzügigeren Grenze war es möglich, auch kurze, abgeschlossene Artikel aufzunehmen, die für Zeitungen/Zeitschriften typisch sind. Die Samples wurden mit folgenden Besonderheiten randomisiert aus dem vorhandenen Textmaterial gezogen: Bei den Texten der Digitalen Bibliothek wurde erzwungen, dass jeder vertretene Autor innerhalb einer Dekade gleichermaßen berücksichtigt wird. Entsprechend wurde beim MKHZ

erzwungen, dass alle in einer Dekade vertretenen unterschiedlichen Zeitungen/Zeitschriften gleichermaßen berücksichtigt werden. Damit wurde verhindert, dass Autoren bzw. Zeitungen/Zeitschriften mit wenig Material beim Sampling-Prozess vollkommen herausfallen. Das Beta-Release enthält Texte von etwa 140 unterscheidbaren Autoren und aus 20 unterschiedlichen Zeitungen/ Zeitschriften.

Die Quelltexte wurden größtenteils in ihrem Ursprungszustand belassen, mit zwei Ausnahmen: Da für die Zeitschrift „Die Grenzboten" nur automatische OCR-Erkennung durchgeführt wurde, wurden die Samples aus dieser Textquelle manuell nachkorrigiert. In den Texten aus den beiden anderen Quellen wurden häufige Sonderzeichen, wie das Schaft-S, durch moderne Äquivalente ersetzt, jedoch weisen die Texte dennoch in unterschiedlichem Maße altertümliche Schreibungen und z.T. auch Sonderzeichen auf. Insgesamt ist festzuhalten, dass die Textformen im RW-Korpus sehr divers sind, so sind z.B. Texte im Dialekt enthalten, sowie Zeitungsausschnitte, die reine Listen sind. Wir haben uns bewusst dagegen entschieden, solch ungewöhnliches Material herauszufiltern, um eine realistische Repräsentation des Textmaterials aus den untersuchten Dekaden zu erhalten.

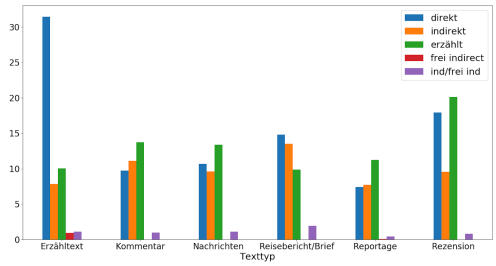
Beim RW-Korpus wurde eine Ausgewogenheit in der zeitlichen Dimension (Textmaterial pro Dekade) sowie zwischen fiktionalen und nicht-fiktionalen Texten angestrebt.

Entgegen ursprünglicher Annahmen stellte es sich als nicht sinnvoll heraus, die Trennung fiktional - nicht-fiktional rein aufgrund der Textquelle zu treffen: Es liegt in der Natur der Textsorte Zeitung/Zeitschrift, dass dort auch fiktionale Texte abgedruckt werden (im Feuilleton, als Fortsetzungsromane u.Ä.). Somit wurde das Kriterium ‚fiktional' für jedes Sample individuell festgelegt. Unsere Definition für ‚Fiktion' ist dabei angelehnt an Gabriel 2007: „Ein erfundener (‚fingierter') einzelner Sachverhalt oder eine Zusammenfügung solcher Sachverhalte zu einer erfundenen Geschichte" (Gabriel 2007: 594). Bei der Identifizierung wurde besonderer Wert auf paratextuelle Merkmale (z.B. Untertitel, Rubriken u.Ä.) gelegt. Von den Samples aus dem MKHZ und den „Grenzboten" wurden auf diese Weise ca. 12% als fiktional eingestuft.

Die folgende Tabelle zeigt die wichtigsten Metadaten des RW-Korpus, welche nach dem Sampling und der Textkorrektur vergeben werden.

Metadatum	Werte	Beschreibung
year	Zahl zwischen 1840 und 1919	Erscheinungsjahr des Textes (bei digBib-Texten: Ersterscheinungsjahr, falls verfügbar)
decade	Zahl in 10er Schritten	Erscheinungsdekade des Textes
source	digBib, grenz, mkhz	Textquelle; bei mkhz wird noch ein Kürzel für die jeweilige Zeitung/ Zeitschrift beigefügt
title	String, undefined	Titel des Textes, falls bekannt
author	String, undefined	Autor des Textes, falls bekannt
fictional	yes, no	Ist der Textausschnitt fiktional?
text_type	Erzähltext, Kommentar, Anzeige, Reportage, Nachrichten, Biographie, Rezension, Reisebericht/Brief, unsure	Texttyp; wenn ein Ausschnitt mehrere Texttypen enthält (z.B. Kommentar und Anzeigen), wird nach dem dominanten Typ klassifiziert oder ansonsten ‚unsure' vergeben

Aufgrund der Diversität der in Zeitungen/Zeitschriften vertretenen Texte wurde für jedes Sample eine nähere Klassifikation des Texttyps vorgenommen, so dass auch dessen Einfluss auf die Verteilung der Redewiedergabetypen untersucht werden kann. Die folgende Abbildung gibt einen ersten Eindruck, welch deutliche Abweichungen hier erkennbar sind. Gezeigt werden nur die Texttypen, für beim Korpusstand vom 25.09.2018 mehr als 10 Samples vorlagen. Die Y-Achse zeigt Prozent der Tokens im Text.



Annotationssystem

Wir unterscheiden die vier Typen direkte, indirekte, frei indirekte (‚erlebte') und erzählte Wiedergabe, sowie die drei Medien Rede (*speech*), Gedanke (*thought*) und Schrift (*writing*), so dass sich insgesamt zwölf Annotationsmöglichkeiten ergeben.

Außerdem annotieren wir die Rahmenformel, die eine direkte oder eine indirekte Wiedergabe einleiten kann. In den Rahmenformeln sowie den Instanzen von erzählter Wiedergabe wird das zentrale Wort markiert, das auf die Sprech-/Gedanken-/Schreibhandlung verweist (z.B. *sagte, Gedanke*). Zudem wird für alle Wiedergabetypen der Sprecher markiert, falls vorhanden.

Während die Unterscheidung der drei Medien nur in Ausnahmefällen problematisch ist, bedürfen die vier Typen genauerer Definitionen.

Die direkte Wiedergabe (*direct*) ist eine wörtlich zitierte Äußerung einer Figur. Sie kann von einer Rahmenformel eingeleitet werden und als einziger Wiedergabetyp Anführungszeichen verwenden.

Er fragte: „Wo ist das Mittagessen?"

Die indirekte Wiedergabe (*indirect*) ist eine nicht-wörtliche Wiedergabe einer Äußerung. Sie ist in unserem Annotationssystem formal definiert und besteht aus einer Rahmenformel und einem abhängigen Nebensatz, der häufig im Konjunktiv steht. Dies kann ein Nebensatz mit Verbzweitstellung sein, mit *dass*, *ob* oder *w*-Fragewort oder ein (erweiterter) Infinitivsatz.

Er fragte, wo das Mittagessen sei.

Die freie indirekte Wiedergabe (*free indirect*) – in der Literatur oft ‚erlebte Rede' genannt – definiert sich über die Überlagerung von Figuren- und Erzählerstimme und ist daher eine typische Form fiktionaler Texte. Sie weist keine Rahmenformel und keine sonstigen formalen Markierungen wie Anführungszeichen auf. Finden sich Elemente der Erzählerstimme wie das Tempus Präteritum oder Personalpronomen der dritten Person und gleichzeitig Elemente der Figurenstimme wie Deiktika, Subjektivität, Ausrufe oder figurentypischer Wortschatz, sind dies Indikatoren für freie indirekte Wiedergabe.

Woher sollte er denn jetzt bloß ein Mittagessen bekommen?

Die erzählte Wiedergabe (*reported*) ist die Erwähnung eines Sprech-, Denk oder Schreibakts, aus der man nicht auf den eigentlichen Inhalt schließen kann. Hinweise auf erzählte Wiedergabe geben Wiedergabewörter, die Thematisierung einer Wiedergabehandlung sowie der Inhalt derselben.

Er sprach über das Mittagessen.

Ein Sonderfall sind uneingeleitete Konjunktivsätze, die zur Wiedergabe verwendet werden. Diese werden als Mischform zwischen indirekter und frei indirekter Wiedergabe markiert.

Sie stellte viele Fragen. Wo sei das Mittagessen?

Darüber hinaus gibt es zusätzliche Attribute, die Besonderheiten bei der Wiedergabe markieren und in der folgenden Tabelle dargestellt werden:

level	Verschachtelungstiefe der Wiedergabe
non-fact	nicht-faktische Wiedergaben (z.B. Negationen, zukünftigen Aussagen oder Absichten)
border	Fälle, die an der Grenze von Rede-, Gedanken- oder Schriftwiedergabe liegen, also nicht alle prototypischen Kriterien der jeweiligen Wiedergabeart erfüllen
prag	sprachliche Wendung, die das Muster einer Wiedergabe aufweist, pragmatisch aber einen anderen Zweck erfüllt (z.B. Höflichkeitsfloskeln)
metaph	Metaphern in Form von Wiedergaben (z.B. *Die Klugheit riet mir davon ab.*)

Die detaillierten Annotationsrichtlinien können unter http://redewiedergabe.de/richtlinien/richtlinien.html eingesehen werden.

Annotationsprozess

Die genaue Identifizierung und Klassifizierung der Redewiedergaben auf der Grundlage des detaillierten Annotationssystems ist eine schwierige Aufgabe.

Jedes Sample des RW-Korpus durchläuft darum einen mehrschrittigen Prozess. Zunächst wird es von zwei Annotatoren unabhängig voneinander annotiert. Danach vergleicht ein weiterer Experte die Annotationen und erstellt, falls notwendig, eine Konsens-Annotation, die dann ins finale Korpus aufgenommen wird. Jedes Sample wird also von drei Personen bearbeitet, um größtmögliche Konsistenz zu gewährleisten.

Die Annotatoren arbeiten mit dem eclipse-basierten Annotationswerkzeug ATHEN (entwickelt von Markus Krug im Projekt Kallimachos, www.kallimachos.de), für das im Projekt eine spezielle Oberfläche für die Redewiedergabe-Annotation implementiert wurde (für eine detaillierte Beschreibung vgl. auch Krug et al. 2018a). Das Werkzeug ist frei verfügbar unter der Adresse http://ki.informatik.uni-wuerzburg.de/nappi/release/.

Verfügbarmachung und Ausblick

Das Beta-Release wird in einem standardisierten und dokumentierten Textformat im Langzeitarchiv des Instituts für Deutsche Sprache zur freien Nutzung zur Verfügung gestellt (https://repos.ids-mannheim.de/). Spätestens für das finale Release im Frühjahr 2020 garantieren wir ein TEI-kompatibles XML-Format. Zudem wird weiteres im Kontext des Redewiedergabe-Projekts entstandenes Material zur Verfügung gestellt, wie nur einfach annotiertes

Textmaterial und annotierte Volltexte. Im Jahr 2020 werden auch Werkzeuge fertiggestellt sein, die es erlauben, in Texten verschiedene Formen der Redewiedergabe automatisch zu erkennen.

Nutzungsszenarien für das Korpus sind vielfältig: Aus NLP-Perspektive kann es als Test- und Trainingsmaterial für automatische Redewiedergabeerkenner verwendet werden. Aus linguistischer Perspektive bieten sich Korpusstudien zu sprachlichen Eigenheiten der Redewiedergabe an, wie z.B. die laufenden Studien zu Redewiedergabeeinleitern von Tu. Aus literaturwissenschaftlicher Perspektive erlaubt das Korpus z.B. Untersuchungen zu der Häufigkeit und Form von Wiedergaben in Erzähltexten in ihrer Relation zur Figurencharakterisierung.

Fußnoten

1. Die ersten beiden Autoren haben zu gleichen Teilen an der Erstellung dieses Beitrags mitgewirkt.

Bibliographie

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

Engelberg, Stefan (2015): *„Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons“*, in: **Eichinger, Ludwig M. (eds.)**: *Sprachwissenschaft im Fokus*. Positionsbestimmungen und Perspektiven. Jahrbuch 2014 des IDS. Tübingen: Narr 205-230.

Gabriel, Gottfried (2007): *„Fiktion“*, in: **Fricke, Harald et al. (eds.)**: *Reallexikon der deutschen Literaturwissenschaft*. Bd. 1: A-G. Berlin / New York: De Gruyter 594-598.

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

Hauser, Stefan (2008): *„Beobachtungen zur Redewiedergabe in der Tagespresse. Eine kontrastive Analyse“*, in: **Lüger, Heinz-Helmut / Lenk, Hartmut E.H. (eds.)**: *Kontrastive Medienlinguistik*. Landau: Verlag Empirische Pädagogik 271-286.

Krug, Markus / Tu, Ngoc Duyen Tanja / Weimer, Lukas / Reger, Isabella / Konle, Leonard / Jannidis, Fotis / Puppe, Frank(2018a): *„Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment“*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 19-21.

Krug, Markus / Weimer, Lukas / Reger, Isabella / Macharowsky, Luisa / Feldhaus, Stephan / Puppe, Frank / Jannidis, Fotis (2018b): *Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]*. DARIAH-DE Working Papers Nr. 27, Göttingen: DARIAH-DE, 2018, URN: urn:nbn:de:gbv:7-dariah-2018-2-9.

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

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